ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELOS ESTADÍSTICOS PARA LA ESTIMACIÓN DE LA CAPACIDAD DE PAGO DE PERSONAS NATURALES CON CUOTA ESTIMADA CON INFORMACIÓN EN EL SISTEMA DE REGISTRO DE DATOS CREDITICIOS

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

JAIME RAÚL TOAQUIZA CUYO

jaimerault.jrt@outlook.com

jaimerault.jrt@gmail.com

DIRECTOR: MSC.DIEGO PAÚL HUARACA SAGÑAY

diego.huaracas@epn.edu.ec

CODIRECTOR: MSC.MENTHOR OSWALDO URVINA MAYORGA

menthor.urvina@epn.edu.ec

DMQ, AGOSTO 2023

CERTIFICACIONES

Yo, JAIME RAÚL TOAQUIZA CUYO, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

Jaime Raúl Toaquiza Cuyo

Certifico que el presente trabajo de integración curricular fue desarrollado por Jaime Raúl Toaquiza Cuyo, bajo mi supervisión.

MSc.Diego Paúl Huaraca Sagñay
DIRECTOR

MSc.Menthor Oswaldo Urvina Mayorga
CODIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

Jaime Raúl Toaquiza Cuyo

MSc.Diego Paúl Huaraca Sagñay

MSc.Menthor Oswaldo Urvina Mayorga

DEDICATORIA

Con una reverencia que trasciende el tiempo y las edades, dedico este pergamino de conocimiento a aquellos cuyas manos sostuvieron mis pasos en este intrincado sendero, con una bondad que desafía toda medida y un respaldo que nunca conoce declive.

A mis progenitores venerados, Daniel y Esther, cuyo amor inquebrantable y sacrificio silente han sido la savia que nutre este logro, consagro estas líneas. Vuestro legado de abnegación y los valores tejidos en mi ser han trazado la ruta hacia la excelencia.

A mis nobles hermanos, cuyos ánimos y risas se convirtieron en mi refugio en las encrucijadas, dedico este canto de esfuerzo. Vuestra presencia constante ha sido el faro que ha disipado las sombras del desafío, insuflando mi ser de valentía.

A mis amigos entrañables, cuya lealtad ha sido el lazo inalterable tejido en las travesías compartidas, les ofrezco estas páginas. Vuestras risas y palabras de aliento han acompañado mi camino, otorgando solaz en los momentos más enrevesados.

A aquellos almas inquietas y apasionadas por los misterios de la matemática, cuyas disquisiciones y búsquedas incansables por el saber ha enriquecido mi pasión por esta disciplina abstracta.

Que este Trabajo de Integración Curricular sea la partitura donde se inscribe vuestro afecto, la melodía donde se entretejen vuestro respaldo y dedicación. En cada trazo de tinta, en cada algoritmo desvelado, perdurará la esencia de vuestra influencia. Con gratitud que trasciende las palabras, dedico este tributo a vosotros, mis adorados seres que han dado sentido y luz a mi travesía matemática.

AGRADECIMIENTO

Con la gratitud que desborda del corazón, elevo mis palabras hacia aquellos cuyos nombres están inscritos en las páginas más profundas de mi ser, con la certeza de que ningún gesto, por más sutil que sea, pasa desapercibido ante los ojos del destino.

A mis padres, Daniel y Esther, pilares inquebrantables de amor y sabiduría, mis palabras de agradecimiento son un eco de la devoción que han derramado sobre mi camino. Vuestra guía y sacrificio han sido faro y brújula en este viaje de conocimiento.

A Rafael y Soraya, amigos en cuyas compañías encuentro refugio y consuelo, vuestras palabras han sido bálsamo en las horas de tribulación. Vuestras amistades son un tesoro incalculable que resplandece en las noches cósmicas.

A Farhad, colega y amigo que con tenacidad y camaradería logramos formular robustos modelos y algoritmos.

A mi respetado y laudable tutor y director de este Trabajo de Integración Curricular, Diego, mi agradecimiento se entrelaza con mi admiración. Vuestra ayuda, paciencia y guía han iluminado los senderos de este proyecto, y vuestra amistad es una fortuna cuyo valor no puede medirse.

Que estas palabras, humildes como son, resplandezcan con la profunda gratitud que albergan. No son letras al azar, sino testimonio sincero de mi reconocimiento por los tesoros de predilección, amistad y conocimiento que habéis derramado en mi vida.

RESUMEN

Este proyecto tiene como objetivo estimar la capacidad de pago de personas naturales bancarizadas que al momento de la estimación cuentan con cuota estimada mediante metodologías o modelos estadísticos paramétricos y no paramétricos (Regresión Lineal Múltiple, Random Forest, Gradient Boosting Machine y XGBoost) con información en el sistema de registro de datos crediticios.

El entrenamiento de los modelos estuvo precedido de una rigurosa exploración y tratamiento de base de datos que consta de 950821 registros y 1172 variables; para capturar el comportamiento y patrón característico de la población se vio que es adecuado dividir en tres grupos de estudio y para cada grupo se formuló las cuatro metodologías o modelos indicados; el primer grupo consta de aquellos individuos que tienen un cuota estimada (monto de amortización en USD de crédito) menor o igual a 107 USD, el segundo grupo con aquellos cuya cuota estimada es mayor a 107 y menor igual a 435 USD y el tercer grupo con aquellos con más de 435 USD en cuota estimada.

Para mejorar el poder predictivo de los modelos en las colas de las distribuciones (sujetos con ingresos muy bajos o muy altos) se empleó la técnica del remuestreo o balanceo con el que efectivamente se obtuvo mejores resultados en comparación a los modelos base (sin remuestreo). De entre todos los modelos implementados, se eligió al Modelo Gradient Boosting Machine como el mejor modelo por su nivel de predicción y rendimiento computacional en las tres sub poblaciones.

La elección de las variables se realizó usando la metodología de Kolmogorov Smirnov (KS) para variables cuantitativas y el Valor de Información (VI) para variables categóricas; para el primer grupo o sub población (que en este trabajo serán sinónimos) se seleccionó 18, para el segundo 35 y para el tercero 28 variables.

Los resultados obtenidos fueron aceptables, dado que el poder predictivo para los tres grupos o sub poblaciones estuvo sobre el 70% de acierto, sin embargo para el grupo 1 y 2 los resultados no fueron los esperados en las colas de la distribución, se esperaba mejores predicciones, sin embargo, para el grupo 3 se logró una alta tasa de predicción; este poder predictivo de los modelos se validó con una nueva base de datos que consta de 84877 individuos y 1177 variables; las predicciones de ingresos para el grupo 1 y 2 no fueron los esperados (como se preveía); sin embargo, para el grupo 3 se obtuvo un muy buen nivel de predicción en especial para sujetos con ingresos muy altos.

Palabras clave: **Estimación de Ingresos, Random Forest (RF), Gradient Boosting Machine (GBM), XGBoost (XGB), Regresión Lineal Múltiple (RLM), Test KS, Test VI, Remuestreo, Función de Balanceo**

ABSTRACT

The objective of this project is to estimate the payment capacity of banked individuals who at the time of the estimation have an estimated quota using parametric and non-parametric statistical methodologies or models (Multiple Linear Regression, Random Forest, Gradient Boosting Machine and XGBoost) with information in the credit data registry system.

The training of the models was preceded by a rigorous exploration and treatment of the database consisting of 950821 records and 1172 variables; in order to capture the behavior and characteristic pattern of the population, it was found to be appropriate to divide it into three study groups and for each group the four methodologies or models indicated were formulated; The first group consists of those individuals who have an estimated quota (amount of amortization in USD of credit) less than or equal to 107 USD, the second group with those whose estimated quota is greater than 107 and less than 435 USD and the third group with those with more than 435 USD in estimated quota.

To improve the predictive power of the models in the tails of the distributions (subjects with very low or very high incomes), the resampling or balancing technique was used, which effectively obtained better results compared to the base models (without resampling). Among all the models implemented, the Gradient Boosting Machine Model was chosen as the best model for its level of prediction and computational performance in the three sub-populations.

The choice of variables was made using the Kolmogorov Smirnov (KS) methodology for quantitative variables and the Value of Information (VI) for categorical variables; 18 variables were selected for the first group or sub-population (which in this work will be synonymous), 35 for the second and 28 for the third.

The results obtained were acceptable, given that the predictive power for the three groups or sub populations was over 70% correct, however for group 1 and 2 the results were not as expected in the tails of the distribution, better predictions were expected, however, for group 3 a high prediction rate was achieved; This predictive power of the models was validated with a new database consisting of 84877 individuals and 1177 variables; the income predictions for group 1 and 2 were not as expected (as expected); however, for group 3 a very good level of prediction was obtained, especially for subjects with very high incomes.

Keywords: **Key words: Income Estimation, Random Forest (RF), Gradient Boosting Machine (GBM), XGBoost (XGB), Multiple Linear Regression (MLR), KS Test, VI Test, Resampling, Balancing Function.**

Contents

List of Figures

List of Tables

Capítulo 1

Descripción del componente desarrollado

1.1 Descripción del proyecto

La estimación de la capacidad de pago como una medida de liquidez y solvencia que puede presentar una persona natural para hacer frente a sus obligaciones crediticias, constituye un soporte técnico importante para las entidades financieras a la hora de asignar los montos y los cupos en los diferentes tipos de crédito a colocar en el mercado, debido a que una estimación adecuada permitirá mitigar la posible existencia de escenarios de sobreendeudamiento, lo cual ocurre cuando el monto de la deuda es superior al patrimonio del cliente, por otra parte, en el área de cobranzas la estimación de la capacidad de pago permitirá priorizar las acciones aumentando los índices de recupero.

El estimador de la capacidad de pago es un modelo analítico que se construye mediante la utilización de técnicas estadísticas de regresión como son los modelos lineales generalizados (GLM) y con técnicas de Machine Learning (ML), dichas técnicas permiten estimar la capacidad de pago de una persona natural a partir de los conjuntos de información referentes a datos crediticios actuales e históricos, información socio-demográfica e información socio-económica.

Considerando las fuentes de información, en particular la información crediticia que es la de mayor importancia en este proyecto, se hace nece-

saria la segmentación de la población en tres grupos distintos:

1. **Población Tarjetahabiente:** Formada por los sujetos que a la fecha de consulta cuentan con un cupo asignado de Tarjeta de Crédito y además cuentan con información crediticia histórica (al menos 3 meses anteriores a la fecha de consulta).
2. **Población Bancarizada:** Formada por los sujetos que a la fecha de consulta no cuentan con Tarjetas de Crédito pero tienen información crediticia histórica (al menos 3 meses anteriores a la fecha de consulta).
3. **Población No Bancarizada:** Formada por los sujetos que presentan menos de 3 meses de información crediticia histórica.

En este proyecto nos centraremos en estimar modelos estadísticos que se ajusten lo mejor posible a la capacidad de pago de las personas naturales en la Población Bancarizada.

1.2 Objetivo general

Construir modelos analíticos que permitan estimar la capacidad de pago de una persona natural para hacer frente a sus obligaciones crediticias.

1.3 Objetivos específicos

1. Construir indicadores que permitan estimar la liquidez disponible de las personas naturales a fin de determinar los cupos de crédito adecuados, mitigando la probabilidad de existencia de sobre endeudamiento.
2. Obtener modelos analíticos que minimicen la sobreestimación de ingresos en clientes con capacidades de pago bajas y la subestimación de ingresos en clientes con capacidades de pago altas.
3. Evaluar modelos paramétricos y no paramétricos dependiendo de la distribución de los ingresos en las subpoblaciones y del conjunto de

variables crediticias disponibles, con el fin de maximizar la exactitud y minimizar el error.

1.4 Alcance

Para alcanzar nuestro objetivo principal, es necesario adquirir un conocimiento en modelos lineales generalizados y modelos de regresión no paramétricos, así que se comenzará por estudiar estos tipos de modelos. Se consolidará, se analizará y se depurará la información crediticia, sociodemográfica y socioeconómica disponible para el desarrollo del proyecto (información con fecha de corte diciembre de 2021).

Se calcularán al menos 2 medidas de divergencia en función de cada tipo de variable de manera que se pueda generar un ranking entre las variables candidatas a formar parte de cada uno de los modelos.

Posteriormente, se entrenarán al menos 3 diferentes tipos de modelos que permitan estimar la capacidad de pago de una persona natural y se evaluará el poder predictivo y el error de ajuste de cada uno de los modelos candidatos en la población de estudio.

Finalmente, se realizará una ejecución del modelo ganador sobre una base de clientes más actualizada con la finalidad de medir la estabilidad del modelo.

Capítulo 2

Marco Teórico

En este capítulo se explica la base matemática de la metodología utilizada en el proyecto. Primero se explican los algoritmos de modelamiento predictivo: Regresión Lineal Múltiple, Random Forest, Gradient Boosting Machine, xgBoost. Después se explican los test estadísticos, junto con los estadísticos más útiles, que se usan para evaluar la bondad de ajuste del mejor modelo seleccionado.

2.1 Modelos estadísticos

2.1.1 Modelos de Aprendizaje Supervisado

El aprendizaje supervisado es una categoría de algoritmos de aprendizaje automático donde el modelo se entrena utilizando un conjunto de datos que posee una variable de interés, Y.

Cada muestra del conjunto de entrenamiento consta de variables explicativas y por lo menos una variable dependiente. El objetivo es que el modelo asocie a los individuos de la muestra, hacia los valores observados de la muestra para la variable dependiente y así poder hacer predicciones precisas en individuos que no formaron parte de la muestra para calcular el modelo. Ejemplos de algoritmos:

1. Regresión Lineal Múltiple

2. Regresión Logística
3. Árboles de decisión
4. Random Forest
5. Gradient Boosting Machine
6. xgBoost

2.1.2 Modelos de Aprendizaje No Supervisado

El aprendizaje no supervisado es otra categoría de algoritmos de aprendizaje automático donde el modelo se entrena con un conjunto de datos no etiquetado. En este caso, el modelo intenta encontrar patrones o estructuras ocultas en los datos sin tener información previa sobre las salidas esperadas, Y. El objetivo principal del aprendizaje no supervisado es explorar la estructura inherente en los datos para obtener información valiosa. Ejemplos de algoritmos:

1. K-Means
2. Análisis de Componentes Principales (PCA)

2.1.3 Modelos Paramétricos

Los modelos paramétricos son aquellos que tienen un número fijo de parámetros y hacen suposiciones específicas sobre la distribución de los errores en un modelo poblacional. Una vez que se han estimado los parámetros utilizando el conjunto de entrenamiento, el modelo está completamente definido y puede ser utilizado para hacer predicciones en nuevos datos. Ejemplos:

1. Regresión Lineal Múltiple
2. Regresión logística

2.1.4 Modelos No Paramétricos

Los modelos no paramétricos, no hacen suposiciones específicas sobre la distribución de los datos ni de los errores, y por tanto no tienen parámetros que se relacionen con la distribución de los datos ni de los errores. Esto les permite ser más flexibles y capaces de ajustarse mejor a datos complejos y no lineales. Ejemplos:

1. Árboles de decisión
2. Random Forest
3. Gradient Boosting Machine
4. xgBoost

2.2 Modelo de Regresión Múltiple

Este es el único modelo paramétrico que se utiliza para estimar los resultados del proyecto. El propósito de implementar dicho modelo fue establecer una comparación entre sus resultados y los obtenidos mediante modelos no paramétricos.

Es importante tener en cuenta que al ser un modelo paramétrico es necesario que se cumplan los siete supuestos (hipótesis) del modelo lineal clásico, no solo para los datos de la muestra, sino también con respecto a la distribución de los errores. En esta sección no se explicitan los detalles matemáticos de este modelo, dado que se lo estudia ampliamente a lo largo de la carrera y es fácil de encontrar en cualquier bibliografía.

En general, algunas de las desventajas que tiene el modelo de regresión múltiple son:

1. *Restricciones en la forma funcional:* El modelo de regresión múltiple asume una relación lineal entre las variables independientes y la variable dependiente. Sin embargo, en la realidad, las relaciones son en su mayoría no lineales. Esto puede llevar a un mal ajuste cuando los datos siguen patrones no son lineales.

2. *Sensibilidad a outliers (valores atípicos):* Los modelos de regresión múltiple pueden ser sensibles a valores atípicos en los datos. Los outliers pueden influir significativamente en la estimación de los coeficientes y afectar la calidad del ajuste del modelo.

3. *Multicolinealidad:* Si existe una alta correlación entre algunas de las variables independientes, se puede presentar un problema de multicolinealidad. La multicolinealidad puede hacer que las estimaciones de los coeficientes sean inestables y difíciles de interpretar correctamente.

4. *Sobreajuste (overfitting) o subajuste (underfitting):* Los modelos de regresión múltiple paramétricos pueden tener una alta tendencia a sobreajustar los datos de entrenamiento. Es decir, el modelo puede capturar ruido y características irrelevantes en los datos, lo que lleva a una mala generalización a nuevos datos no utilizados en la base de entrenamiento.

5. *Dificultad para modelar relaciones no lineales:* Los modelos de regresión múltiple asumen una relación lineal entre las variables. Modelar relaciones no lineales requiere transformar manualmente las variables, lo que puede ser complicado y consumir mucho tiempo. Además, puede ser que no se capture información sobre patrones que no se pueden ver fácilmente en los gráficos.

6. *No apto para datos complejos:* En problemas con muchas variables independientes y relaciones no lineales complejas, los modelos de regresión múltiple paramétricos pueden no ser lo suficientemente flexibles como para capturar la información relevante de los datos.

7. *Requiere supuestos sobre los errores:* Los modelos de regresión múltiple paramétricos asumen que los errores tienen una distribución normal con media cero y varianza constante. Si estos supuestos no se cumplen, las inferencias y predicciones del modelo pueden ser incorrectas.

Aunque los modelos de regresión múltiple paramétricos tienen estas desventajas, también tienen ventajas, como su interpretabilidad (que es ampliamente estudiado en la Econometría).

Sin embargo, en situaciones donde los datos son altamente no lineales o cuando se desconoce la verdadera forma funcional de la relación entre las variables, los modelos no paramétricos pueden ser más adecuados.

2.3 Random Forest

El Random Forest es un algoritmo de aprendizaje supervisado no paramétrico, que genera varios árboles de decisión sobre un conjunto de datos de entrenamiento, y luego los resultados obtenidos para cada árbol se combinan para obtener una estimación más cercana al valor verdadero.

Para entender de mejor forma lo que realiza este algoritmo, a continuación se explican algunos conceptos relacionados, ya que fueron de importancia para después utilizar correctamente las librerías en R.

2.3.1 Algoritmo de Árbol de decisión (AD):

Es de mucha importancia conocer el concepto de *Árbol de Decisión*, ya que este modelo es la base para definir los tres métodos no paramétricos que se utilizaron en este proyecto: Random Forest, Gradient Boosting Machine y xgBoost.

Definición 2.1 (Árbol de decisión)**.** *Un árbol de decisión, es un grafo dirigido con estructura de árbol (no necesariamente binario), en el que también usa estadísticos provenientes de la* ***Teoría de información*** *para ir generando, a partir del nodo padre, sus nodos hijos hasta llegar a los nodos hojas.*

En un árbol de decisión, cada nodo interno representa una pregunta o condición sobre una variable explicativa del conjunto de datos. A medida que se sigue el árbol desde el nodo raíz hasta las hojas, las respuestas a estas preguntas guían el camino de la predicción. Cada nodo hoja representa una clase o valor de predicción final.

Existen dos tipos de árboles de decisión: **árboles de clasificación** y **árboles de regresión**. En el proyecto se utilizó la idea de un árbol de regresión ya que la variable dependiente que se desea predecir, es continua y

no discreta.

Por otro lado, el algoritmo general para crear un solo árbol de decisión está dado en el Algoritmo(1).

Algorithm 1 Generación de *árbol de decisión*

Input: Base de datos de aprendizaje, $\mathcal{L}$.
Output: Árbol de decisión, φ.

1: Crear un árbol de decisión, φ, con nodo raíz t_0
2: Crear una pila vacía S de nodos de la forma $(t, \mathcal{L}_t)$
3: $S.push((t, \mathcal{L}_t))$
4: **while** $S \neq \emptyset$ **do**
5: $t, \mathcal{L}_t = S.pop()$
6: **if** Se cumple el criterio de parada **then**
7: $\overline{y}_t$ =constante
8: **else**
9: Hallar la partición de $\mathcal{L}_t$ que maximiza la ganancia de información:

$$s_t^* = \max_{A \in \text{Var-Explicativas}} \triangle i(A, t)$$

10: Particionar $\mathcal{L}_t$ en $\bigcup_{v \in \text{niveles(A)}} \mathcal{L}_{t_v}$
11: Crear los v nodos hijo de t
12: $S.push((t, \mathcal{L}_{t_v})), \forall v \in$ niveles(A)
13: **return** φ

2.3.2 Explicación detallada del algoritmo AD y Ejemplo

Un árbol de decisión sirve para predecir valores de una variable dependiente. Se lo realiza individuo por individuo (fila a fila) y en cada nodo intermedio se van realizando preguntas sobre las variables de estudio, de esta manera se va particionando la base de entrenamiento de forma que los nodos hojas contengan individuos de la muestra con características similares.

Viéndolo de forma más gráfica, el algoritmo de árbol de decisión está particionando individuos en regiones lo más homogéneas posible. Ver Figura(2.2).

Dentro del algoritmo, se pueden identificar tres acciones importantes que se deben analizar y realizar para el correcto funcionamiento del mismo:

1. *Definir un* ***criterio de parada*** *para la generación de nodos hojas del*

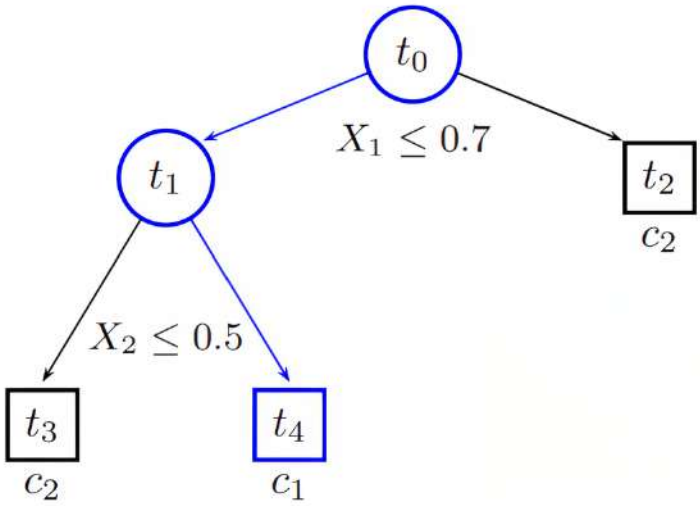


Figura 2.1: *Ejemplo: Árbol de decisión para variable binaria.* [8]

árbol.

El porcentaje de individuos en cada nodo hoja debe ser de tal forma que no caiga en una sobre-estimación o en una sub-estimación en el modelo final. Este porcentaje es muy importante y va a ser el criterio de parada de los algoritmos, ya que si no se fija un criterio de parada, el algoritmo va a crear nodos hijos hasta que en cada hoja solamente exista 1 individuo.

2. *Calcular una posible* ***estimación de la variable objetivo*** *que se está tratando de modelar.*

   De forma general, se utiliza el promedio muestral (en cada nodo hoja) para estimar el valor de los ingresos reales de las personas. El promedio es muy común por las propiedades estadísticas que posee (converge con probabilidad 1 al parámetro poblacional), y ha sido implementado en las librerías de R que se utilizaron para calcular los modelos de prueba.

3. *Hallar la* ***partición del conjunto de entrenamiento*** *(*$\mathcal{L}$*), que maximice la* ***ganancia de información*** *en cada nodo del árbol, hasta llegar a tener nodos hojas.*

   Como se muestra en la figura (2.2), en cada nodo del árbol de decisión se realiza una partición de los individuos, hasta que en los

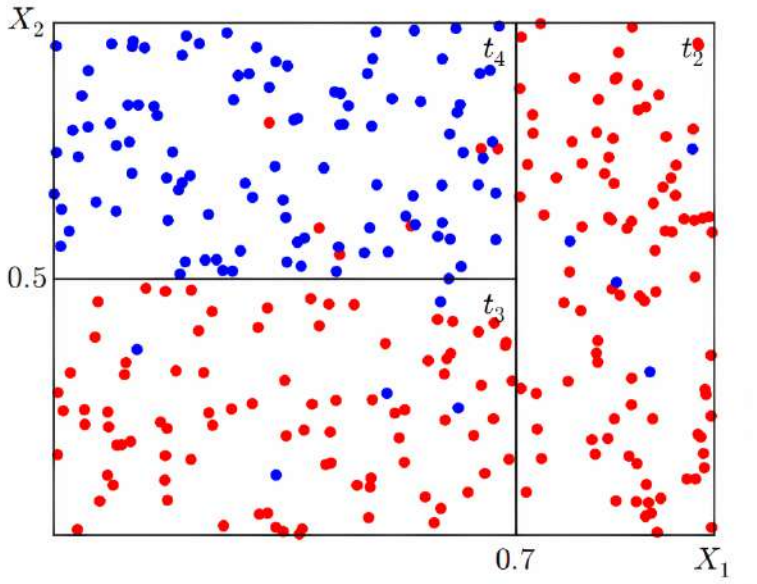


Figura 2.2: *Clasificación de árbol de la* Figura(2.1).

nodos hoja se llega a tener un porcentaje fijo de individuos del conjunto de entrenamiento.

Para saber de qué forma ir creando cada nodo del árbol, se utiliza un estadístico llamado ganancia de información, dado por la fórmula general:

$$\triangle i(A,t) = i(t) - \sum_{v \in \text{Niveles(A)}} \frac{|S_v|}{|S|} \cdot i(t_v) \tag{2.1}$$

en donde:

(a) A es una variable explicativa cualquiera,

(b) t es el nodo al que se quiere calcular la ganancia de información,

(c) $|S|$ es la cardinalidad del conjunto de individuos que actualmente se tiene en el nodo t,

(d) $|S_v|$ es la cardinalidad del subconjunto de S que contiene los individuos en el nivel v de la variable A,

(e) Finalmente, $i(t)$ es una medida de impureza de cada nodo, que ayuda a medir la homogeneidad de los niveles de una variable. En este sentido, $i(t_v)$ mide la impureza de cada nodo hijo del nodo actual t (en el supuesto caso que se decida usar la variable A para particionar el nodo t).

En árboles de clasificación es muy común utilizar las funciones de

impureza de: *Entropía de Shannon* o también de *Índice de Gini*. Por otro lado, en árboles de regresión se utiliza una función de impureza relacionada con el error cuadrático medio en cada nodo:

$$i(t) = \frac{1}{N_t} \sum_{\mathbf{X}, y \in \mathcal{L}_t} (y - \overline{y}_t)^2 \tag{2.2}$$

en donde:

(a) N_t es el número de individuos que hay actualmente en el nodo t,

(b) $\mathcal{L}_t$ es el subconjunto, del conjunto de entrenamiento, que se encuentra actualmente en el nodo t,

(c) $\mathbf{X}$ es la matriz de información muestral de las variables explicativas, que se obtiene de $\mathcal{L}_t$, de forma similar, y es el vector muestral de la variable dependiente que se estudia y se obtiene de $\mathcal{L}_t$,

(d) y es el valor real de la variable objetivo, y $\overline{y}_t$ es el valor estimado de la variable, que se calcula en el nodo t.

Es importante resaltar también que existe una relación entre la función de impureza de Gini (clasificación) y la función de impureza dada antes (regresión):

$$i(t) = \frac{1}{2} \cdot i_{gini}(t) \tag{2.3}$$

por lo que se puede implementar un árbol de clasificación o de regresión usando un mismo criterio para ambos.

Para finalizar el cálculo de la maximización de la ganancia de información, se debe calcular, para cada variable que no ha sido considerada hasta ese momento en la creación del nodo t:

$$s_t^* = \max_{A \in \text{Var-Explicativas}} \triangle i(A, t) \tag{2.4}$$

donde s_t^* es la partición en $\mathcal{L}_t$ que maximiza la ganancia de información, en el nodo actual t.

Si se modifican los estadísticos de los tres pasos anteriores y se los utiliza

en el algoritmo indicado, Alg.(1), va a cambiar la forma de generar el árbol de decisión.

2.3.3 Algoritmo de Random Forest:

Como ya se explicó antes, los bosques aleatorios (Random Forest) consisten en construir un conjunto de varios árboles de decisión, que se calculan a partir de una modificación en la aleatoriedad del algoritmo (1), presentado antes para árboles de decisión.

Existen varios métodos para calcular el bosque aleatorio, y estos se diferencian en la forma de introducir la perturbación aleatoria que va a influir en el cálculo de cada árbol de decisión por separado.

En el presente proyecto, se utiliza la función *ranger* de la librería de mismo nombre, del lenguaje R. Este algoritmo para calcular el bosque aleatorio, se resume en los siguientes pasos:

1. **Conjunto de datos:** El algoritmo de Bosques Aleatorios parte de un conjunto de datos de entrenamiento que incluye variables independientes, junto con la variable objetivo que se desea predecir.

   Cada fila en el conjunto de datos representa una observación.

2. **Muestreo bootstrap:** Se inicia el proceso generando múltiples conjuntos de datos de entrenamiento mediante el muestreo bootstrap con reemplazo. Esto implica seleccionar al azar filas del conjunto de datos original para formar una nueva muestra, permitiendo que una misma observación aparezca varias veces o no aparezca en absoluto.

   Cada conjunto de datos de entrenamiento generado de esta manera es del mismo tamaño que el conjunto de datos de entrenamiento original.

3. **Construcción de árboles:** Para cada conjunto de datos de entrenamiento, se construye un árbol de decisión.

4. **Votación y predicción:** Una vez que se han construido todos los árboles, para realizar una predicción sobre una nueva observación, se evalúa esa observación por cada árbol del bosque.

En el caso de clasificación, cada árbol emite una predicción para la clase de la observación. Luego, se realiza una votación entre todos los árboles para determinar la clase final más frecuente. En el caso de regresión, cada árbol emite una predicción numérica, y el resultado final es el *promedio* de las predicciones de todos los árboles.

5. **Evaluación y ajuste:** Una vez que se ha construido el bosque y se han realizado las predicciones, se evalúa la precisión del modelo utilizando un conjunto de datos de prueba. Este conjunto de prueba contiene individuos no utilizados en el proceso de entrenamiento y permite evaluar cómo se desempeña el modelo frente a datos no vistos previamente.

   Se pueden ajustar hiperparámetros del algoritmo, como el número de árboles en el bosque o la profundidad máxima de los árboles, mediante técnicas de validación cruzada para obtener un modelo final con un rendimiento óptimo.

Lo mencionado en el paso 5 forma parte del trabajo realizado en el proyecto. Específicamente, después de efectuar las predicciones iniciales, se procedió a ajustar los hiperparámetros de la función *ranger*. El objetivo fue la minimización del error cuadrático medio (MSE), la reducción de la cantidad de individuos en cada nodo hoja y la limitación del número de árboles generados por el algoritmo. Estos ajustes se llevaron a cabo con el propósito de evitar tanto el sobreajuste como el subajuste en las estimaciones posteriores, al mismo tiempo que se buscó mantener la eficiencia computacional.

Cabe señalar que se tomó esta medida considerando que las computadoras empleadas en los experimentos poseían recursos limitados en términos de capacidad de procesamiento.

2.4 Gradient Boosting Machine (GBM)

El algoritmo de *GBM* es el segundo algoritmo de aprendizaje supervisado no paramétrico que se utilizó en el proyecto. A diferencia del algoritmo de Random forest, este genera árboles de decisión de forma secuencial sobre un conjunto de datos de entrenamiento; es decir, cada árbol de decisión

que se genera, depende del anterior árbol que se creó para ser generado (ver más en [5] y [10]).

El algoritmo de Gradient Boosting Machine se resume en el Algoritmo(2). Los parámetros que se observan en este algoritmo son:

1. y_i es el valor real de la observación de la variable dependiente para el individuo i,
2. La función de pérdida es: $L(y_i, \gamma) = (y_i - \gamma)^2$,
3. *M* denota el número de árboles que se van a crear, mientras que m es el m-ésimo árbol generado,
4. F_{m-1} es la predicción del paso anterior, empieza con F_0,
5. j representa el nodo hoja, mientras que J_m representa el total de hojas en el árbol m,
6. $\gamma_{j,m}$ representa la estimación del error en la hoja j del árbol m
7. $n_{j,m}$ representa el número de individuos en la hoja j del árbol m,
8. $r_{i,m}$ representa el error en el individuo i del árbol m, del subconjunto de individuos $R_{j,m}$,
9. el parámetro η representa el coeficiente de aprendizaje del algoritmo,
10. y por último se tiene la función indicadora $\mathbb{1}(x \in R_{j,m})$, esta asegura que el individuo x se vincule apropiadamente al único nodo hoja que le corresponde, ya que cada individuo está asociado a un único nodo hoja del árbol de decisión.

2.4.1 Explicación detallada de algoritmo y Ejemplo

A continuación se explican los pasos que se realiza en el algoritmo para construir el modelo predictivo, junto con un ejemplo del algoritmo:

1. *Inicialización del algoritmo con el promedio muestral:*

   Para predecir la variable objetivo Y en función de las variables explicativas $X_1, \ldots, X_k$, el algoritmo comienza utilizando el promedio

Algorithm 2 Algoritmo de Gradient Boosting Machine. [9]

Input: Base de datos de aprendizaje, $\mathcal{L}$.
Output: Estimación de variable dependiente $Y = F_M(x)$.

1: Calcular el valor constante:

$$F_0(x) = \min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) = \overline{Y_n} \tag{2.5}$$

2: **for** $m = 1$ hasta M **do**

3: Calcular los residuales, $\forall i = 1, \ldots, n$:

$$r_{i,m} = -\frac{\delta L(y_i, \gamma)}{\delta \gamma}\Big|_{\gamma = F_{m-1}(x)} = y_i - F_{m-1} \tag{2.6}$$

4: Entrenar el árbol de decisión, $r_m \sim X_1, \ldots, X_k$, y hallar las regiones de individuos en cada hoja, $R_{j,m}$, para $j = 1, \ldots, J_m$

5: Calcular, $\forall j = 1, \ldots, J_m$:

$$\gamma_{j,m} = \min_{\gamma} \sum_{x_i \in R_{j,m}} L(y_i, F_{m-1}(x_i) + \gamma) = \frac{1}{n_{j,m}} \sum_{x_i \in R_{j,m}} r_{i,m} \tag{2.7}$$

6: Actualizar las estimaciones del modelo:

$$F_m(x) = F_{m-1}(x) + \eta \cdot \sum_{j=1}^{J_m} \gamma_{j,m} \cdot \mathbb{1}(x \in R_{j,m}) \tag{2.8}$$

7: **return** $F_M(x)$

muestral como el primer predictor para dicha variable. Es decir, se calcula el promedio de todos los valores de Y en el conjunto de entrenamiento y se utiliza este valor como la predicción inicial para todas las instancias de los datos de entrenamiento. Ver figura (2.3).

$$F_0 = \overline{Y}_0$$

2. *Generación de nuevos modelos débiles para reducir errores:*

   El residuo se calcula restando las predicciones del primer árbol, de los valores reales Y en el conjunto de entrenamiento. Ver figura (2.4).

$$r_0 = Y - \overline{Y}_0$$

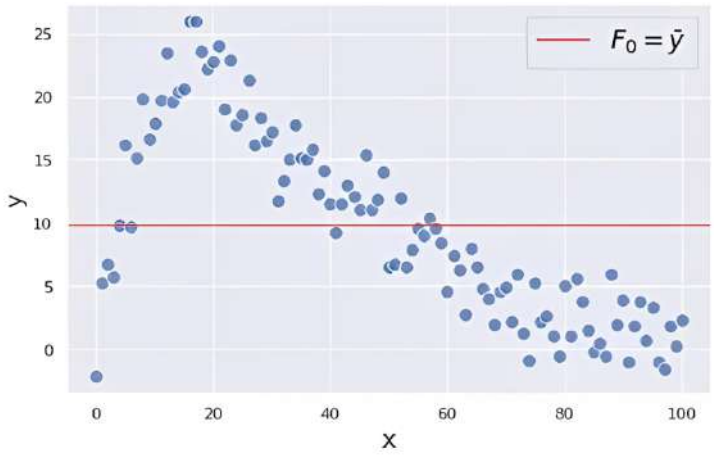


Figura 2.3: *Muestra y promedio muestral. [9]*

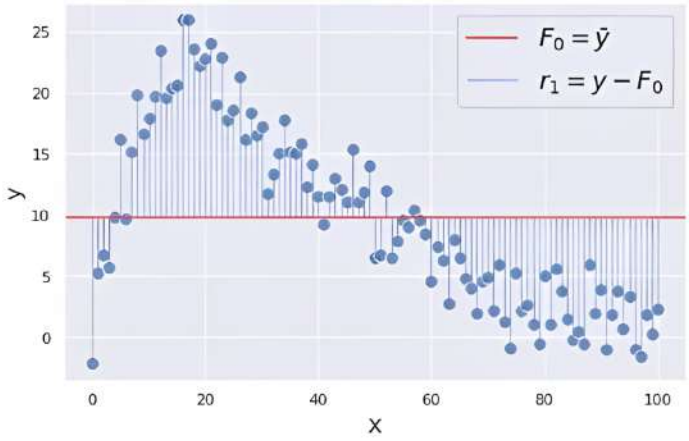


Figura 2.4: *Errores para F_0. [9]*

Con el fin de reducir los errores de predicción causados por la primera estimación (media muestral), se genera un primer árbol de decisión. En este caso, el árbol se construye de manera que la variable dependiente sea el residuo de la primera estimación, r_0, mientras que las variables explicativas iniciales se mantienen, es decir, $r_0 \sim X_1, \ldots, X_k$. Ver figura (2.5).

Las predicciones brindadas por este primer árbol se las denota como γ_1. Luego, la nueva predicción de la variable Y ahora se obtiene sumando el promedio inicial (calculado en el primer paso) con el valor que se obtiene en la hoja correspondiente del árbol de decisión resultante. Sin embargo, para controlar la contribución de

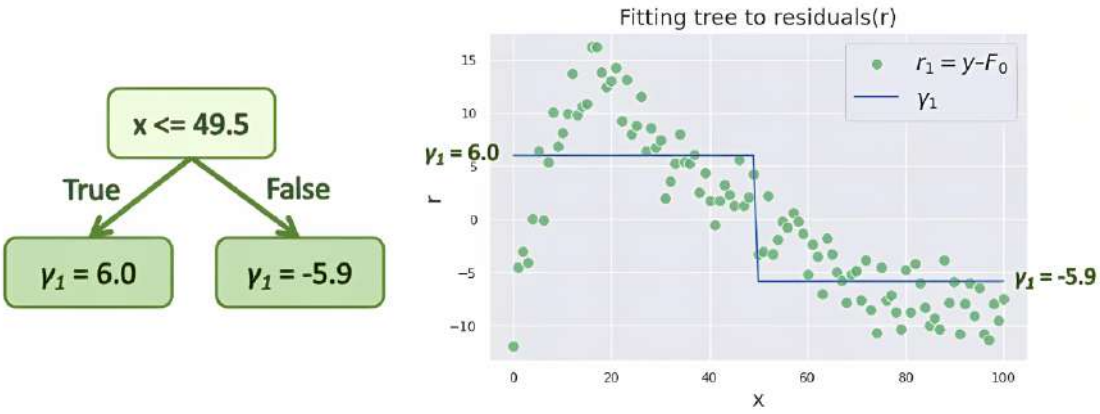


Figura 2.5: *Árbol generado para los errores y gráfico de errores. [9]*

este nuevo modelo al ensamble final, se multiplica por un coeficiente de aprendizaje que está en el rango de 0 a 1. Ver figura (2.6).

$$F_1 = F_0 + \eta \cdot \gamma_1$$

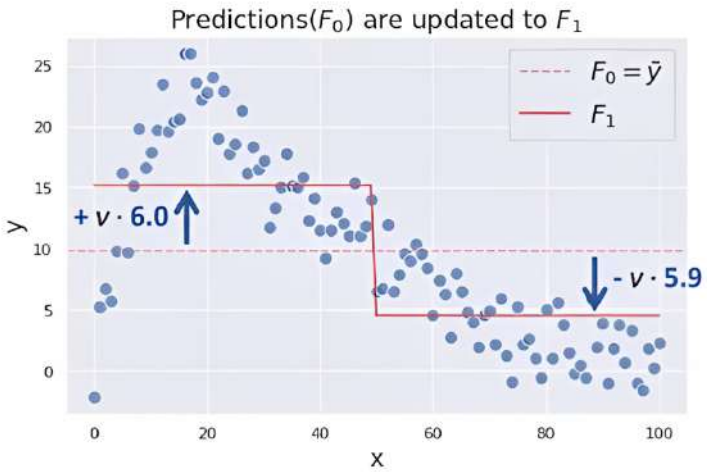


Figura 2.6: *Actualización de estimación de Y. [9]*

Este coeficiente de aprendizaje, η_0, es un hiperparámetro del algoritmo y determina cuánto impacto tiene el nuevo modelo en la predicción final, pero nunca es 1 para evitar sobre-ajuste o sub-ajuste de información que no fué usada para crear el modelo. En el presente proyecto, se tomó $\eta = 0.03$ para todas las iteraciones.

3. *Repetición del proceso hasta convergencia*

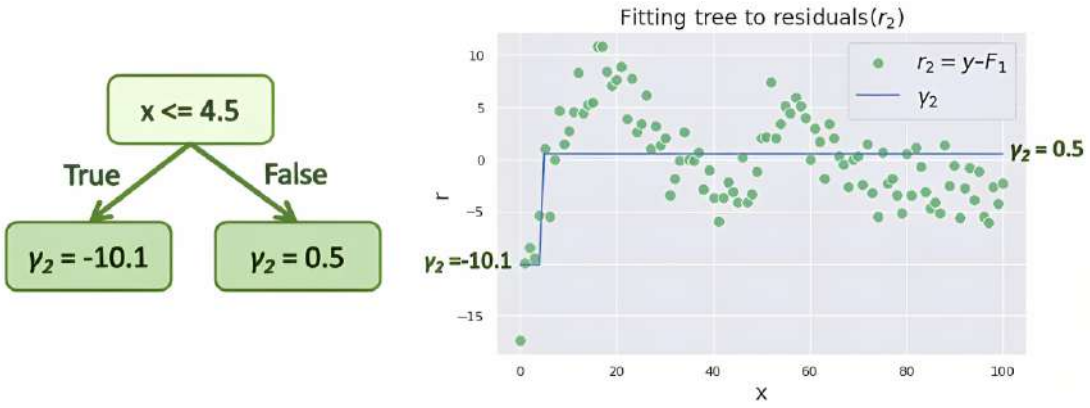


Figura 2.7: *Se calculan nuevamente los errores.* [9]

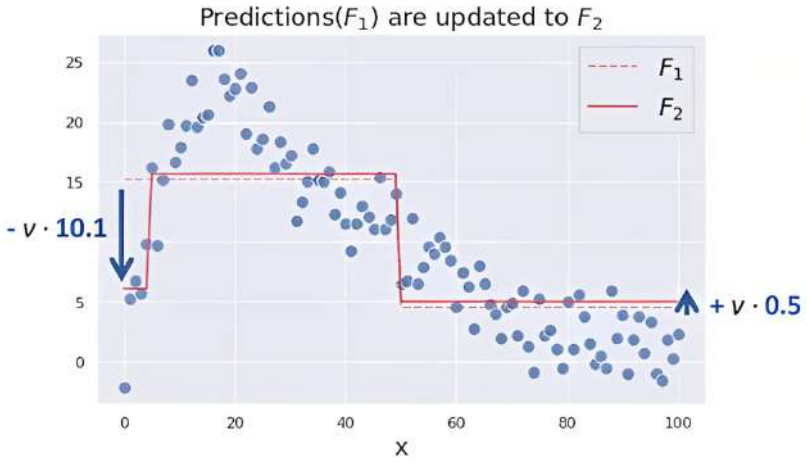


Figura 2.8: *Segunda actualización de* Y. [9]

El paso 2 se repite varias veces para construir más modelos débiles y mejorar el rendimiento del ensamble. En cada iteración, se construye un nuevo árbol de decisión que se enfoca en estimar los residuos de los árboles anteriores:

$$\begin{aligned}
&\text{Iteración 1: } r_1 = r_0 - \gamma_1 \implies \text{estimar: } \gamma_2 \implies F_2 = F_1 + \eta \cdot \gamma_2 \\
&\text{Iteración 2: } r_2 = r_1 - \gamma_2 \implies \text{estimar: } \gamma_3 \implies F_3 = F_2 + \eta \cdot \gamma_3 \\
&\text{Iteración 3: } r_3 = r_2 - \gamma_3 \implies \text{estimar: } \gamma_4 \implies F_4 = F_3 + \eta \cdot \gamma_4 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots
\end{aligned}$$

Por lo tanto, las predicciones del algoritmo se actualizan mediante la suma de los resultados obtenidos en las hojas de los nuevos árboles, creados para los errores en cada iteración.

El proceso de construcción de nuevos árboles de decisión, y la actualización de las predicciones, se repite iterativamente hasta que los errores de predicción, r_k, ya no cambian drásticamente de un árbol a otro, o hasta que se alcance un número predefinido de árboles de decisión en el ensamble.

Cuando se estudia al algoritmo graficando cada iteración sobre un diagrama de puntos, como se realiza desde la Figura(2.3) hasta la Figura(2.8), se observa que la estimación de la variable dependiente se va dando poco a poco, de forma que variables indicadoras se van aproximando más a la tendencia de los puntos. Ejemplo tomado de [9].

2.5 Extreme Gradient Boosting (xgBoost)

El algoritmo de *xgBoost* es el tercer algoritmo de aprendizaje supervisado no paramétrico que se utilizó en el proyecto. De forma similar que el algoritmo de *GBM*, este método utiliza árboles de decisión y una técnica de gradiente descendente.

En este caso se minimiza la suma entre la pérdida de entrenamiento y la regularización del árbol generado. Además, para cada árbol que se genera, se trata de crecer el árbol hasta una altura en la que no vaya a existir sub o sobreestimación. Luego, la estimación final es la sumatoria de la predicción inicial y la predicción de cada árbol siguiente.

La diferencia entre GBM y xgBoost se puede ver en la figura(2.9). Es decir que, a veces, el algoritmo de *GBM* genera muchas particiones a la base de datos (ver [3]).

Aún así, este algoritmo tiene mejoras en los modelos matemáticos que se calculan en cada iteración en conjunto con una optimización del sistema computacional. La diferencia en la minimización de la función objetivo, es que se agrega un término de regularización que tiene que ver con la creación del árbol:

$$\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma(x_i)) + \Omega(\gamma) \tag{2.9}$$

donde γ representa todos los valores de los nodos hoja del árbol generado,

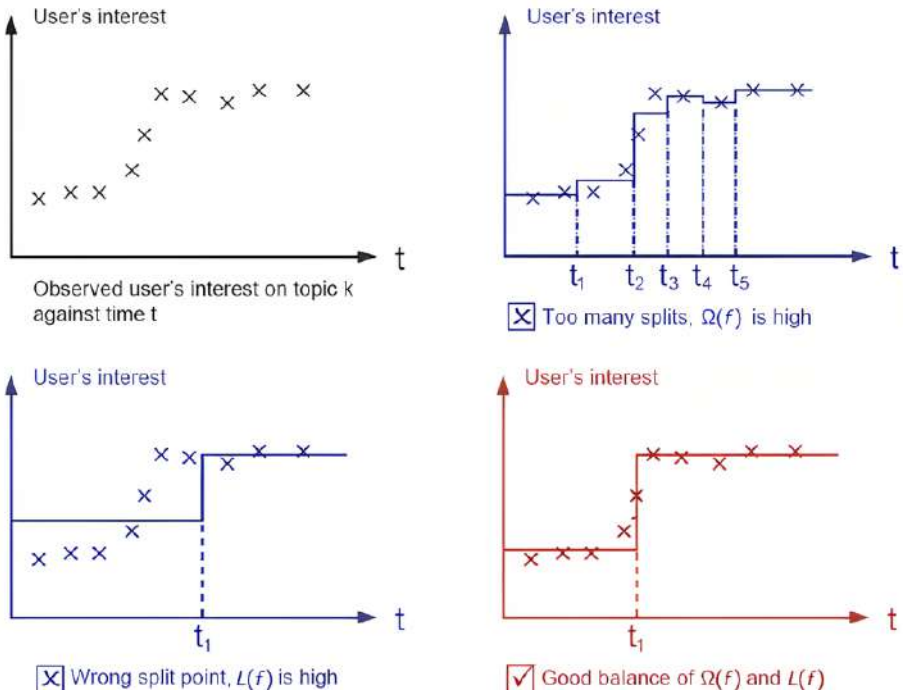


Figura 2.9: *Idea gráfica del algoritmo xgBoost. [2]*

F_{m-1} representa lo mismo pero para el árbol generado en el paso anterior.

El primer término se lo conoce como *pérdida de entrenamiento (training loss en inglés)*, y mide qué tan bien se incorpora el modelo a los datos de entrenamiento, y trabaja de forma similar que en *GBM*.

El término $\Omega(\gamma)$ es la regularización del algoritmo, que está relacionado con el número de regiones(o particiones), $R_{j,m}$, de individuos que se crean en la hoja j del árbol m. Este término de regularización mide la complejidad de los árboles que se van creando en cada iteración. En este algoritmo se considera:

$$\Omega(\gamma) = \eta \cdot T + \frac{1}{2}\lambda \cdot \sum_{j=1}^{T} \gamma_j^2 \tag{2.10}$$

En donde η es el coeficiente de aprendizaje para cada nodo hoja *(en el caso de GBM se utilizó $\eta = 0.03$)*, T es el número de nodos hoja en el árbol, λ es un coeficiente de penalización para el tamaño del árbol generado, y γ_j es la estimación de la variable dependiente en cada hoja del árbol.

Si no se controla la partición de individuos, los árboles que se generan en cada iteración crecen hasta su máxima extención, lo que causa que el

árbol creado *"pierda generalidad"*, y no realice buenas predicciones con individuos que no pertenecieron a la base de entrenamiento.

Tomando en cuenta el primer término en la función objetivo, y aplicando *Series de Taylor de orden 2, sobre la segunda componente de la pérdida L*, se obtiene:

$$\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i)+\gamma(x_i)) \approx \sum_{i=1}^{n} \left[L(y_i, F_{m-1}(x_i)) + g_i \cdot \gamma(x_i) + \frac{1}{2} h_i \gamma^2(x_i) \right] \quad (2.11)$$

en donde: $g_i = \delta_{F_{m-1}}(F_{m-1}(x_i)-y_i)^2$ y $h_i = \delta^2_{F_{m-1}}(y_i-F_{m-1}(x_i))^2$, son la primera y segunda derivada respectivamente, con respecto a $F_{m-1}(x_i)$ (ver [12]).

Ahora, sumando los dos términos de la función objetivo, y simplificando la notación, se obtiene:

$$\min_{\gamma} \sum_{j=1}^{T} \left[G_j \cdot \gamma_j + \frac{1}{2}(H_j + \lambda)\gamma_j^2 \right] + \eta \cdot T \quad (2.12)$$

en donde se considera $L(y_i, F_{m-1}(x_i)) \approx 0$, pues es lo que se espera de la estimación; además: $G_j = \sum_{i \in I_j} g_i$ y $H_j = \sum_{i \in I_j} h_i$, donde $I_j = \{i \in 1, \ldots, n : x_i$ está en la hoja $j\}$.

Observar que esta última sumatoria es la suma de T funciones cuadráticas respecto a γ_j cada una, por lo tanto al derivar e igualar a cero, se obtiene el estimador de los valores en cada hoja del árbol:

$$\widehat{\gamma_j} = -\frac{G_j}{H_j + \lambda} \quad \forall j = 1, \ldots, T \quad (2.13)$$

Por lo tanto, el mínimo de esta función objetivo es:

$$-\frac{1}{2} \cdot \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \eta \cdot T.$$

Ahora, con los cálculos realizados antes, es posible definir una nueva forma de generar un árbol de decisión, denominada *método "greedy"*. Tomar en cuenta que al mismo tiempo, en la creación de nodos considera que pueden existir valores perdidos (valores NA) en la base de

entrenamiento, por lo que también se introduce una técnica de *"sparsity aware split finding"*[1], la cual solo se menciona pero no se la estudia a profundidad ya que no es el objetivo de este proyecto. Ver Algoritmo(3).

Algorithm 3 Generación "Greedy" de *árbol de decisión* (caso Y binaria)

Input: Base de datos de aprendizaje, $\mathcal{L}$.
Output: Árbol de decisión, φ.

1: Crear un árbol de decisión, φ, con nodo raíz t_0
2: Particionar t_0 en nodo izquierda y derecha si:

$$Gain = \frac{1}{2} \cdot \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \eta \geq 0 \tag{2.14}$$

   En el caso que $Gain < 0$, se deja de particionar el nodo (técnica de podado de árboles).
3: Realizar el paso anterior para cada nodo que se vaya creando pero de forma ordenada, es decir, de izquierda a derecha.
4: **return** φ

Finalmente, el algoritmo para el *Modelo xgBoost* está dado por el Algoritmo(4). Aquí, M denota el número de árboles que a generar.

La metodología del xgBoost es mucho más eficiente computacionalmente que las otras metodologías como Random Forest, Gradient Boosting Machine y Regresión Lineal Múltiple. En este proyecto no se estudia estas mejoras computacionales pues no se halla dentro del alcance de conocimientos de la carrera.

2.6 Test de Kolmogorov-Smirnov (KS)

El test de Kolmogorov-Smirnov, o también conocido como el test KS, es un test estadístico que permite comparar dos distribuciones diferentes.

Este test se puede utilizar para comparar una distribución muestral con una distribución teórica muestral (test KS de una muestra), o también dos distribuciones muestrales (test KS de dos muestras).

[1]Esta técnica está diseñada para manejar conjuntos de datos donde muchas de las características tienen un gran número de valores nulos o cero, durante el proceso de construcción del árbol de decisión.

Algorithm 4 Algoritmo de Modelo xgBoost

Input: Base de datos de aprendizaje, $\mathcal{L}$.
Output: Predicción, F_M, de la variable dependiente.

1: Empezar con un predictor "débil" de la variable: $F_0 = \overline{Y}$.
2: **for** $m = 1, \ldots, M$ **do**
3: Calcular las primeras y segundas derivadas:

$$\forall i = 1, \ldots, n: \quad g_i = \delta_{F_{m-1}}(F_{m-1}(x_i) - y_i)^2$$
$$h_i = \delta^2_{F_{m-1}}(y_i - F_{m-1}(x_i))^2$$

4: Entrenar el árbol "greedly" de decisión. Al final se denota al árbol como:

$$E_m(x_i) = \sum_{j=1}^{T} \widehat{\gamma}_j \cdot \mathbb{1}(x \in R_{j,m}),$$

en donde $\{R_{j,m}\}_{j=1}^{T}$ es la mejor partición de individuos en los nodos en el árbol m. Además: $\widehat{\gamma}_j = -\frac{G_j}{H_j + \lambda}$.
5: Actualizar: $F_m = F_{m-1} + E_m$
6: Calcular:

$$F_M = F_0 + \sum_{m=1}^{M} E_m$$

7: **return** F_M

La prueba de hipótesis que se estudia con este estadístico es:

H_0 : las muestras vienen de una población con la misma distribución.

H_a : las muestras vienen de una población con diferente distribución.

De manera general, el estadístico de prueba que se utiliza para realizar un test KS de dos muestras, es (ver [1]):

$$D_i = \max_z (F_{X_i}(z) - F_Y(z))$$

Por otro lado, si la variable X tiene una muestra de n individuos y Y tiene una muestra de m individuos, el estadístico crítico para aceptar o rechazar H_0 es:

$$D_{crit} = c(\alpha) \cdot \sqrt{\frac{n+m}{n \cdot m}},$$

en donde $c(\alpha) = \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot 0.5}$ y α es el nivel de confianza. Así, se rechaza H_0 si $D_i > D_{crit}$ (ver [14]).

Por ejemplo, se compara los siguientes vectores de resultados:

$$X : (4, 7, 9, 12, 21, 23, 23, 24, 28)$$
$$Y : (1, 5, 7, 13, 13, 18, 20, 20, 25)$$

Se calculan las funciones de distribución acumuladas:

$$CDF(X) = \begin{cases} 0 & x < 4 \\ \frac{1}{9} & 4 \leq x < 7 \\ \frac{2}{9} & 7 \leq x < 9 \\ \frac{1}{3} & 9 \leq x < 12 \\ \frac{4}{9} & 12 \leq x < 21 \\ \frac{5}{9} & 21 \leq x < 23 \\ \frac{7}{9} & 23 \leq x < 24 \\ \frac{8}{9} & 24 \leq x < 28 \\ 1 & 28 \leq x \end{cases} \qquad CDF(Y) = \begin{cases} 0 & x < 1 \\ \frac{1}{9} & 1 \leq x < 5 \\ \frac{2}{9} & 5 \leq x < 7 \\ \frac{1}{3} & 7 \leq x < 13 \\ \frac{5}{9} & 13 \leq x < 18 \\ \frac{2}{3} & 18 \leq x < 20 \\ \frac{8}{9} & 20 \leq x < 25 \\ 1 & 25 \leq x \end{cases}$$

Se grafica las distribuciones acumuladas para visualizar:

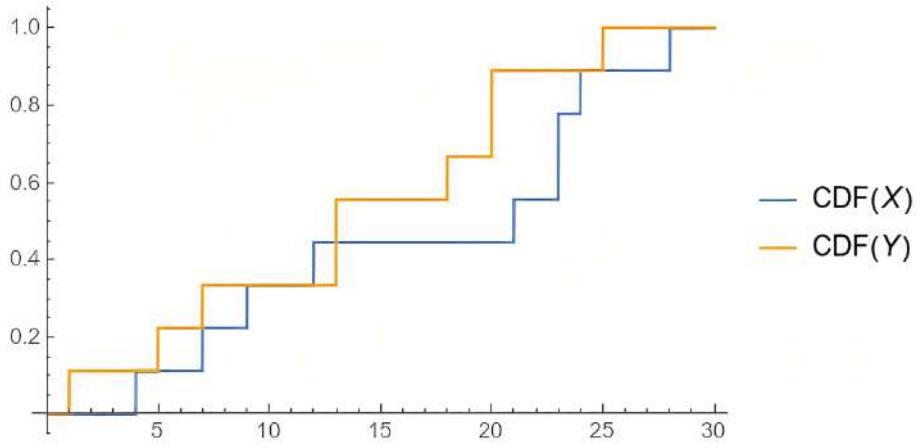


Figura 2.10: *Gráficas de las distribuciones acumuladas de X e Y.* [11]

Ahora, se calcula el estadístico de prueba: $D = \max_x(F_X(x) - F_Y(x))$:

Por lo tanto, el estadístico de prueba tiene el valor $\frac{4}{9} = 0.\overline{4}$; mientras que el estadístico crítico, con confianza de $1 - \alpha = 0.95$, es $0.64021\ldots$. Como el estadístico de prueba es menor que el valor crítico, no se rechaza la hipótesis nula. Ejemplo tomado de [13].

$$D = \begin{cases} 0 & x < 1 \\ \frac{1}{9} & 1 \leq x < 4 \\ 0 & 4 \leq x < 5 \\ \frac{1}{9} & 5 \leq x < 9 \\ 0 & 9 \leq x < 12 \\ \frac{1}{9} & 12 \leq x < 18 \\ \frac{2}{9} & 18 \leq x < 20 \\ \frac{4}{9}* & 20 \leq x < 21 \\ \frac{1}{3} & 21 \leq x < 23 \\ \frac{1}{9} & 23 \leq x < 24 \\ 0 & 24 \leq x < 25 \\ \frac{1}{9} & 25 \leq x < 28 \\ 0 & 28 \leq x \end{cases}$$

En el proyecto se utilizó el test KS de dos muestras, para comparar las distribuciones muestrales de cada variable explicativa con la variable dependiente que es la **Capacidad de pago**. Solamente para el cálculo del test KS y del test VI (que se explica en la siguiente sección), se discretizó la variable dependiente en una variable categórica con 3 niveles diferentes.

También, para realizar los cálculos del estadístico de prueba en el proyecto, se lo realizó de forma ponderada. Es decir, se calculó el estadístico de prueba de una variable explicativa con respecto a los posibles niveles de la variable dependiente; y este cálculo se lo realiza para cada combinación posible de dos niveles en la variable dependiente.

Después, para combinar todos los resultados, se realiza una suma ponderada:

$$estad_{KS} = \sum_{i,j=1}^{k} D_{i,j} \cdot pond_{i,j}$$

en donde $pond_{i,j} = \left(\frac{frec_{nivel[i]} + frec_{nivel[j]}}{n(k-1)} \right)$, k es el número de niveles distintos de la variable dependiente Y, y n es el número de observaciones de la variable X. También, se tiene que $k \leq n$, pues como máximo existen n valores diferentes para Y.

En este caso, al ser Y una variable numérica, se genera una nueva variable que recoge la información de la variable Y en tres niveles. La nueva variable se trata del *Rango de ingresos*. Además, este cálculo ponderado se lo realiza para cada una de las variables explicativas.

De esta manera, escogiendo aquellas variables que cumplen $estad_{KS} \geq 0.20$, se logra obtener las variables que más aportan para describir a la variable dependiente. Hay que tomar en cuenta que, mientras más se acerca el valor del estadístico de prueba a 1, mayor **poder predictivo** tendrá la variable comparada.

2.7 Valor de Información (VI)

El Valor de Información (VI) es una medida utilizada en el contexto del análisis de tablas de contingencia para evaluar la asociación entre dos variables categóricas. Es una medida de la ganancia o pérdida de información proporcionada por una variable categórica al predecir otra variable categórica.

Cuando se trabaja con **tablas de contingencia**, se presentan las frecuencias conjuntas de dos variables categóricas. Estas tablas se organizan en filas y columnas, donde cada celda representa el recuento de observaciones que pertenecen a una combinación particular de categorías de ambas variables. A partir de esta información, se pueden calcular diversas medidas de asociación, y el Valor de Información es una de ellas. Ver Tabla(2.1).

$X \setminus Y$	d_1	$\cdots$	d_k	$\cdots$	d_s	total
c_1	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1s}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_h	n_{h1}	$\cdots$	n_{hk}	$\cdots$	n_{hs}	$n_{h\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_r	n_{r1}	$\cdots$	n_{rk}	$\cdots$	n_{rs}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$\cdots$	$n_{\bullet k}$	$\cdots$	$n_{\bullet s}$	n

Tabla 2.1: *Tabla de contingencia de VI.*

El cálculo del Valor de Información implica comparar la distribución con-

junta de las dos variables con la distribución que se esperaría si fueran independientes. La fórmula para calcular el Valor de Información es la siguiente:

$$VI = \sum_{i=1}^{r} \sum_{j=1}^{s} p_{ij} \log \left(\frac{p_{ij}}{p_i \cdot p_j} \right)$$

Donde:

- p_{ij}: Frecuencia conjunta observada en la celda (i, j) de la tabla de contingencia.
- p_i: Frecuencia marginal de la fila i, es decir, la suma de las frecuencias en la fila i.
- p_j: Frecuencia marginal de la columna j, es decir, la suma de las frecuencias en la columna j.
- r: Número de filas de la tabla de contingencia, es decir, el número de categorías de la variable 1.
- s: Número de columnas de la tabla de contingencia, es decir, el número de categorías de la variable 2.

El Valor de Información es una medida que varía entre 0 y ∞, donde un valor cercano a 0 indica que las dos variables son independientes, mientras que un valor mayor indica una mayor asociación entre ellas. Cuanto mayor sea el Valor de Información, mayor es la ganancia de información y mayor es la asociación entre las variables categóricas (ver [6] y [7]).

El Valor de Información se utiliza en el análisis de datos, la minería de datos y la inteligencia artificial para medir la relevancia de una variable categórica al predecir otra variable categórica, y es especialmente útil en el contexto de selección de características y análisis de atributos en modelos predictivos.

2.8 Prueba Chi-Cuadrado para tablas de contingencia

Al momento de realizar predicciones sobre la muestra de validación, es necesario medir las frecuencias de las predicciones para saber si nuestro modelo predice de la misma forma a la muestra de test y la muestra de validación. En otras palabras, lo que se quiere y se espera es que, la distribución de las predicciones en la muestra de test sea muy similar a la distribución de las predicciones con la muestra de validación.

Para poder saber si lo último mencionado se está cumpliendo, se utilizó el estadístico para la prueba de Chi-Cuadrado en una tabla de contingencia generada por las variables *Rango de Ingresos Real* y *Rango de Ingresos Estimados*. Esta tabla de contingencia es calculada para la muestra de test y la muestra de validación.

El estadístico de prueba es:

$$\chi^2 = \sum_{i,j=1}^{k} \frac{(t_{ij} - v_{ij})^2}{t_{ij}},$$

en donde k es el número de niveles en los rangos de ingresos, t_{ij} es la frecuencia de las estimaciones con la muestra test en la posición (i, j) en la primera tabla de contingencia, v_{ij} es la frecuencia de las estimaciones con la muestra de validación en la posición (i, j) en la segunda tabla de contingencia. Se calcularon los rangos de tal forma que haya un número igual de rangos reales y estimados.

El estadístico crítico está dado por una variable aleatoria χ^2 con $(k-1)^2$ grados de libertad y una confianza del 95%. Además, la hipótesis nula es *"Las distribuciones en modelamiento y test son similares."*; la cual se rechaza si $\chi^2 > \chi^2_{crit}$ (ver [4]).

Capítulo 3

Metodología

En este capítulo, se describen los métodos y enfoques utilizados para llevar a cabo la investigación y abordar las preguntas de investigación planteadas. La elección de la metodología es fundamental para garantizar la validez y confiabilidad de los resultados obtenidos. A lo largo de este capítulo, se presentará el enfoque general y particular de investigación, se justificará la elección de la metodología y se describirán los procedimientos específicos implementados.

Se presenta un esquema que describe el camino, la vía y la forma como se abordó la investigación, desde la exploración, el tratamiento y reducción de la base de datos, la especificación de la población de estudio en tres sub poblaciones, la base de datos de entrenamiento y validación, el remuestreo, el entrenamiento de los distintos modelos paramétricos y no paramétricos, la obtención de resultados de la base de datos de entrenamiento y de la base de datos de validación, la elección del mejor modelo y la creación de indicadores de liquidez tanto de ingresos reales como la de los ingresos estimados para la población en general y para cada sub población tanto para la base de datos de modelamiento y comprobación.

Es imperante especificar las bases de datos de trabajo:

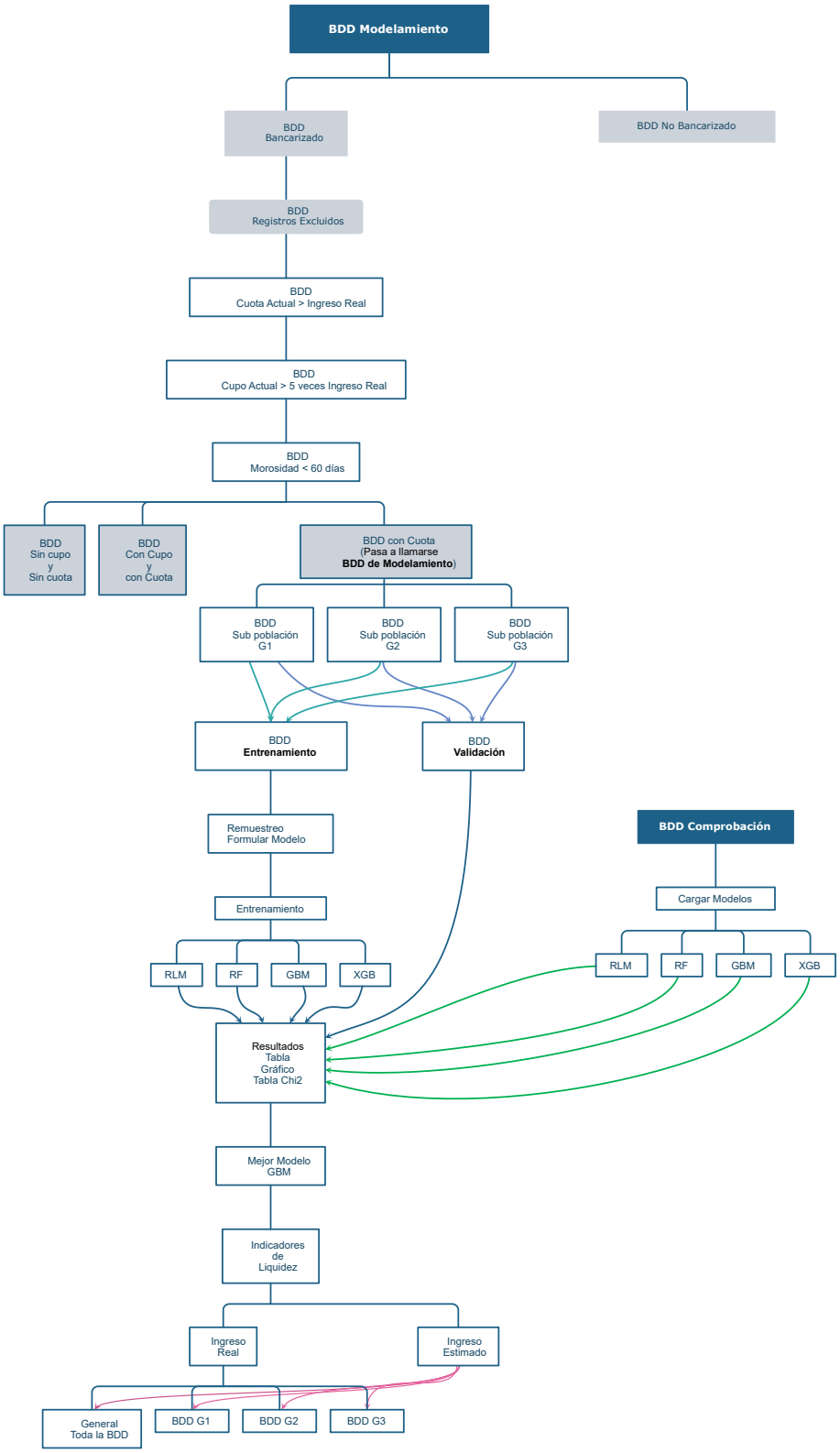
- **Base de datos de modelamiento**: aquella que consta de 950821 registros y 1172 variables más las otras variables que se generar o crean como combinación (sección de elección de variables). Más

tarde, esta base de datos constará de 298449 registros (solamente con individuos que tienen cuota estimada).

- **Base de datos entrenamiento**: aquella que consta con la mitad de registros, según la sub población específica, de la base de datos de modelamiento para entrenar al modelo.
- **Base de datos validación**: aquella que consta con la mitad de registros, la sub población específica, de la base de datos de modelamiento para validar al modelo.
- **Base de datos de comprobación**: aquella nueva base de datos con el que se valida al mejor modelo elegido, en cada sub población, y también se calcula los indicadores de liquidez.

Se expondrá de forma detallada, para la sub población G1, las implementaciones y/o explicaciones de los distintos procedimientos empleados en esta investigación ; más, para las sub poblaciones G2 y G3 se presentarán solamente los resultados, dado que para la obtención de estos se procedió de manera similar que en la sub población G1.

3.1 Esquema metodológico y de resultados



3.2 Exploración y descripción de la Base de Datos

3.2.1 Población de modelamiento

La base de datos está conformado por una muestra aleatoria y representativa de sujetos del Sistema Crediticio Ecuatoriano con ingreso real reportado a una entidad financiera.

Consta de 950821 individuos y 1172 variables (tabla 3.1).

BDD	Casos	Variables
General	950,821	1172
%	100.00	

Tabla 3.1: Población de Modelamiento

3.2.2 Identificación de Bancarizado

Dado que la evaluación de la capacidad de pago se llevará a cabo utilizando datos históricos de Buró, se procede a identificar a aquellos individuos que carecen de esta información (es decir, que no tienen historial crediticio). Se origina la categoría de <NO BANCARIZADO> basada en la presencia o ausencia de la estimación de sus ingresos; esta variable que viene por defecto en la base de datos se llama <Estimación Actual> y contrasta a aquellos individuos que cuentan con información histórica.

Nótese que: La estimación actual ha sido calculada previamente por la institución para sus fines.

Así, tenemos: (tabla 3.2)

BANCARIZADO_36M	N	%
BANCARIZADO	927,647	97.56%
NO BANCARIZADO	23,174	2.44%
Total	**950,821**	**100.00%**

Tabla 3.2: Bancarizado y No Bancarizado

De donde, la población de interés es: <Bancarizado> que consta de 927,647 registros.

3.2.3 Descripción del ingreso real y estimado actual

Se presentan algunas estadísticas relacionadas con el ingreso real y estimado actual (tabla 3.3).

Ingreso Real

Decil	Mínimo	10%	20%	30%	40%	50%	60%	70%	80%	90%	Máximo
Valor en USD	400	483	540	600	710	815	996	1214	1720	2937	35000

Ingreso Estimado

Decil	Mínimo	10%	20%	30%	40%	50%	60%	70%	80%	90%	Máximo
Valor en USD	88	478	509	534	568	599	676	808	955	1249	7737

Tabla 3.3: Deciles de Ingreso Real y Estimado Actual

De la tabla 3.3, se tiene las siguientes interpretaciones, así como la identificación clara de la problemática que se intenta solventar con este estudio.

El modelo que generó la <estimación actual> presentaría restricciones, dado que su enfoque de predicción fue para individuos que poseían tarjetas de crédito. Individuos desprovistos de esta modalidad de financiamiento carecían de una evaluación estimada de ingresos, lo que los excluía como candidatos para la obtención de tarjetas de crédito. El enfoque primordial de esta investigación se dirige precisamente hacia aquellos individuos que carecen de un cupo registrado registrado en tarjeta de créditos, pero sí presentan un registro de cuota, independientemente de su naturaleza crediticia, en una entidad financiera.

Se observa que el modelo subestimó los ingresos, evidenciado al comparar los deciles del ingreso real con los estimados. Por ejemplo, el ingreso mínimo real fue de 400 USD en 2018, mientras que el modelo estimó solo 88 USD. Similarmente, el ingreso máximo real de 35,000 USD fue estimado en 7,737 USD por el modelo.

Es importante notar que más del 90% de los ingresos reales están por debajo de 2,937 USD, sin embargo, el modelo indica que ese porcentaje está por debajo de 1,249 USD. La discrepancia es más notoria en el hecho

de que en el ingreso real, el 70% de los ingresos se encuentra por debajo de aproximadamente 1,200 USD.

En resumen, el modelo presenta una tendencia a subestimar los ingresos respecto a lo que la población efectivamente recibe o percibe. Esta discrepancia tiene implicaciones significativas, ya que los cálculos de los límites de crédito para tarjetas se basan en los ingresos declarados. Cuando los ingresos son subestimados, el resultado es que se otorgan límites de crédito inferiores a lo que los clientes pueden realmente manejar.

Esta situación lleva a consecuencias notables: los titulares de tarjetas con ingresos subestimados pueden recibir límites de crédito que no reflejan adecuadamente su capacidad real. Esto puede resultar en un menor atractivo para el uso de la tarjeta, ya que los límites otorgados no son proporcionales a los ingresos.

En un escenario inverso, donde los ingresos son sobre estimados, existe el riesgo de otorgar límites de crédito que excedan la capacidad de pago del usuario. Este puede acumular una deuda significativa que no puede pagar, generando pérdidas tanto para la institución financiera como para el cliente.

Para evitar estas situaciones, se propone otorgar límites de crédito en función de reglas claras. Por ejemplo, establecer un límite máximo de 4.5 veces el ingreso real del individuo, con excepciones para aquellos con ingresos más altos o perfiles de riesgo específicos.

Gestión de Riesgo Crediticio y Provisión de Fondos

En el ámbito financiero, el cálculo de la pérdida esperada es un proceso fundamental para evaluar y mitigar el riesgo crediticio. Esta pérdida esperada se calcula a través de un conjunto de factores clave que reflejan la realidad de los préstamos y las deudas.

La fórmula para calcular la pérdida esperada implica considerar varios elementos esenciales. En primer lugar, se toma en cuenta la exposición, que corresponde a la deuda pendiente del prestatario en un momento específico. Luego, el factor LGD (Loss Given Default), que representa

la cantidad de pérdida asumida en caso de incumplimiento, se calcula restando la tasa de recuperación del valor 1. La tasa de recuperación, en este contexto, se refiere a la proporción de la deuda que se recupera después de un incumplimiento.

Sin embargo, la probabilidad de default es un aspecto crucial en esta ecuación. Es la estimación de la probabilidad de que un individuo no cumpla con los pagos en el siguiente año. Si esta probabilidad de default es alta y la tasa de recuperación es baja, se traduce en una pérdida esperada significativa. Esto señala un nivel elevado de riesgo en la relación crediticia.

$$PerdidaEsperada = Exposicion * (1 - TasaRecuperacion) * Prob_{default}$$

En situaciones donde el perfil de riesgo del prestatario es particularmente alto, las instituciones financieras se ven en la necesidad de realizar provisiones sobre el monto otorgado como préstamo. Imaginemos que se otorgaron 5,000 dólares; en este caso, se podría realizar una provisión de 100 o 200 dólares, o incluso más. Esta provisión funciona como una estimación de la porción de la deuda que es probable que el prestatario no cubra.

Esta suma provisionada se deposita en una cuenta controlada por el Banco Central (BC) y se le remunera con una tasa pasiva, aproximadamente alrededor del 4%. Sin embargo, este enfoque puede resultar desventajoso para el banco, ya que el dinero podría haber sido utilizado para otorgar nuevos créditos, que generan tasas de interés más altas, oscilando entre el 16% y el 24%. En este sentido, existe una pérdida de oportunidad al no aprovechar este capital de manera más lucrativa.

En un contexto de regulación y supervisión financiera, tanto la Superintendencia de Bancos (SB) como la Superintendencia de Economía Popular y Solidaria (SEPS) desempeñan un papel fundamental. Estas instituciones establecen pautas y regulaciones para las provisiones y la gestión del riesgo crediticio. Las provisiones se ajustan en función del riesgo y del perfil del cliente, basándose en varios modelos de riesgo que integran diversos factores para determinar la cantidad adecuada a ser provisionada.

En conclusión, la precisión en la estimación de ingresos es esencial para el funcionamiento óptimo de las operaciones crediticias. Un modelo subestimado o sobreestimado puede resultar en pérdida de oportunidades para las instituciones financieras y en consecuencias adversas para los clientes. El uso de modelos adecuados y reglas claras es crucial para una gestión crediticia efectiva y equitativa.

Hasta ahora, se ha observado que la exclusión de 23,174 (tabla 3.2) individuos con ingresos superiores a 7,500 USD afecta significativamente la precisión de las estimaciones de ingresos reales. Por tanto, es esencial ajustar el modelo de predicción para considerar estos casos y mejorar la exactitud en las predicciones, especialmente en los extremos de la distribución. Se planea desarrollar un nuevo modelo que simule valores de ingresos reales utilizando la técnica de remuestreo bootstrap para obtener estimaciones más robustas en las colas de la distribución y corregir la subestimación actual. Este enfoque permitirá tomar decisiones financieras más informadas y mejorar la calidad de las estimaciones de ingresos.

3.2.4 Registros excluidos

Para el modelamiento se van excluyendo ciertas casuisticas, que en principio tienen una relación limitada con el análisis en cuestión. Este enfoque de refinamiento garantiza la calidad de los datos y contribuye a generar resultados más fiables y coherentes en el estudio.

Con la finalidad de depurar los registros, se emplearon 3 reglas:

1. Se excluyen los registros cuya cuota estimada actual es superior al ingreso.

El primer criterio de exclusión se basa en la relación entre la cuota estimada actual y el ingreso del individuo. La cuota estimada es la suma de los pagos que una persona debe realizar por las deudas adquiridas, como la cuota A y la cuota B en el caso de tener dos créditos. Esto plantea una pregunta relevante: ¿qué sucede si alguien gana solo la mitad de su cuota total estimada? ¿Cómo es posible que los bancos hayan otorgado créditos

a personas cuyos ingresos son solo una fracción de sus obligaciones de pago?

En el proceso de otorgamiento de crédito, los bancos evalúan minuciosamente la capacidad de pago de sus clientes. Esto se conecta directamente con el modelo estimador de ingresos. Por regulación, una persona no debe destinar más del 50% de sus ingresos al pago de deudas, reservando la otra mitad para necesidades esenciales como educación, salud y alimentación.

Mediante un análisis detallado, se identifican los casos en los cuales la cuota estimada excede el ingreso del individuo. Se observa que esta situación se presenta en aproximadamente el 23% (tabla 3.4) de los registros. Este hallazgo subraya la importancia de la evaluación precisa de la capacidad de pago al otorgar créditos y resalta la necesidad de ajustar los criterios para garantizar la sostenibilidad financiera tanto de los clientes como de las instituciones financieras.

CUOTA ESTIMADA SUPERIOR AL INGRESO		
MARCA	**N**	**%**
NO	711,320	76.68%
SI	216,327	23.32%
Total	**927,647**	**100.00%**

Tabla 3.4: Registros cuya cuota estimada actual es superior al ingreso

Cuota Estimada

Decil	Mínimo	10%	20%	30%	40%	50%	60%	70%	80%	90%	Máximo
Valor en USD	0	0	56	123	191	270	361	473	645	1015	31852

Tabla 3.5: Deciles de Cuota Estimada

Se puede colegir que: el 90% de la población tiene una cuota estimada (monto de amortización de créditos) menor a 1,015 USD (tabla 3.5).

2. Se excluyen los registros cuyo Máximo cupo de TC es superior en 5 veces al ingreso

Se excluyen los registros cuyo Máximo cupo de TC es superior en 5 veces al ingreso (Regla aplica únicamente a sujetos con ingresos inferiores a 1,000 dólares)

Es importante destacar que la misma regla no se aplica a individuos con ingresos más altos. En este caso, a medida que los ingresos aumentan, se permite un mayor nivel de endeudamiento, ya que se supone una mayor capacidad de pago disponible.

Al realizar este análisis en la base de datos, se observa que alrededor del 2.73% (tabla 3.6) de los registros se ve afectado por esta consideración. Estos resultados refuerzan la idea de que la evaluación de la capacidad de pago debe ser adaptable a diferentes perfiles de ingresos, asegurando una política crediticia equitativa y sostenible.

CUPO TC SUPERIOR EN 5 VECES AL INGRESO		
MARCA	**N**	**%**
NO	691,920	97.27%
SI	19,400	2.73%
Total	**711,320**	**100.00%**

Tabla 3.6: Cupo de TC 5 veces mayor al Ingreso Real

Máximo Cupo TC

Decil	Mínimo	10%	20%	30%	40%	50%	60%	70%	80%	90%	Máximo
Valor en USD	0	0	0	0	0	0	0	800	1510	3400	200000

Tabla 3.7: Máximo Cupo TC

Se puede colegir que: el 70% de la población tiene cupo máximo de 800 USD; el 90% de la población tiene cupo máximo de 3,400 USD (tabla 3.7).

3. Se excluyen los registros que presentan morosidad al punto de observación mayor a 60 días

Se excluyen los registros que presentan morosidad al punto de observación mayor a 60 días (y que cupo máximo no sea 5 veces mayor al ingreso real).

Es fundamental mencionar que se excluyó a los individuos con deudas vencidas en el análisis. Los modelos que se aplican se enfocan exclusivamente en personas que mantienen sus pagos al día, es decir, aquellos cuya morosidad no excede los 60 días. Sin embargo, surge la interrogante de cómo abordar a aquellos individuos que no están al día en sus pagos

pero que, por diversas circunstancias, podrían regularizar su situación. Esta perspectiva podría abrir la puerta a un estudio más profundo y específico en el futuro.

Al solicitar una tarjeta de crédito, uno de los primeros criterios evaluados es si el solicitante mantiene una conducta de pago puntual. Este aspecto es esencial, ya que las personas con retrasos de pago en otros bancos (superiores a 2 meses) podrían repetir ese comportamiento y no cumplir con los pagos de una nueva tarjeta.

Al aplicar estas restricciones, se observa que aproximadamente el 12% (tabla 3.8) de los registros son excluidos del proceso de construcción del modelo. Estos hallazgos subrayan la importancia de la solidez crediticia en la solicitud y otorgación de tarjetas de crédito, con el objetivo de minimizar los riesgos asociados a impagos y proteger tanto a los clientes como a las instituciones financieras.

MOROSIDAD MAYOR A 60 DIAS		
MARCA	**N**	**%**
NO	607,457	87.79%
SI	84,463	12.21%
Total	**691,920**	**100.00%**

Tabla 3.8: Morosidad mayor a 60 días

Morosidad Actual

Decil	Mínimo	10%	20%	30%	40%	50%	60%	70%	80%	90%	Máximo
Num de Días	0	0	0	0	0	0	0	0	0	0	60

Tabla 3.9: Deciles del número de días de Morosidad

Podemos colegir que: el 90% (tabla 3.9) de la población ya filtrada tiene 0 días de morosidad; existe un 10% restante que tiene 60 días de morosidad como máximo

3.2.5 Poblaciones independientes para el estudio

Adicionalmente, se lleva a cabo un análisis cruzado entre los individuos que cuentan con una cuota estimada actual y un cupo de tarjeta de crédito actual.

Este estudio se basa en poblaciones bien definidas, cada una con sus propias características y comportamientos particulares. Por esta razón, el enfoque de estudio es independiente para cada población. Como resultado, este trabajo de integración curricular (TIC) se realiza en colaboración de tres estudiantes.

Nuestra población de interés se compone de individuos sin un cupo de TC asignado pero cuentan con una cuota estimada. Esta base de datos consta de 298,449 individuos (tabla 3.10). El segundo estudiante se centrará en analizar la población que tiene tanto un cupo asignado como una cuota estimada, mientras que el tercer estudiante abordará la población que no tiene cupo ni cuota estimada. Esta división nos permitirá obtener una comprensión más completa y específica de cada segmento de la población en estudio.

	TIENE_CUOTA (crédito)		
TIENE_TC (cupo)	**NO**	**SI**	**Total**
NO	77,819	298,449	**376,268**
SI	3	231,186	**231,189**
Total	**77,822**	**529,635**	**607,457**

Tabla 3.10: Tres poblaciones independientes de estudio

3.2.6 Especificación de la población de estudio

A la población que Tiene Cuota: SI y Tiene Cupo: No, los 298,449 individuos (tabla 3.10) los denominaremos población Cuota (o base de datos Cuota). Se explora el comportamiento de los deciles de ingresos real y estimado actual.

Se puede colegir que (tabla 3.11): el 80% de la población percibe un ingreso menor a 1,680 USD y la estimación actual indica que perciben un ingreso menor a 595 USD. Se tiene como Ingreso Real máximo 35,000 USD, el modelo sub estima e indica ingreso máximo de 2,098 USD. Este resultado hace énfasis nuevamente en que debemos formular modelos adecuados que eviten la sub y la sobre estimación y realicen predicciones aceptables en los extremos de la colas de las distribuciones de ingresos.

Ingreso real

Quintil	Mínimo	20%	40%	60%	80%	Maximo
Valor en USD	400	520	700	952	1680	35000

Ingreso estimado

Quintil	Mínimo	20%	40%	60%	80%	Maximo
Valor en USD	92	481	510	552	595	2098

Tabla 3.11: Quintiles de Ingreso Real y Estimado Actual - Población Cuota

3.2.7 Relación entre Ingresos y Cuota Estimada Actual

Con la finalidad de excluir del modelamiento a los registros que no guardan relación entre el ingreso y la cuota estimada actual, se generan matrices duales (tabla 3.12).

Al comparar los ingresos reales con las cuotas estimadas, se identifican ciertos escenarios que no tienen sentido incluir en el modelo.

Se procede a excluir a aquellos individuos cuyos ingresos sean inferiores a 450 USD. Estas personas presentan ingresos muy bajos y carecerían de la capacidad financiera necesaria para cumplir con las obligaciones de pago, ya que sus ingresos apenas alcanzan para cubrir necesidades básicas como alimentación, educación y vestimenta.

También se eliminarán los sujetos con ingresos significativamente altos pero cuotas estimadas muy bajas. Estos individuos optan por pagos en efectivo en lugar de acumular deudas. Aunque parecerían menos riesgosos, ya que tienen altos ingresos y bajas deudas, si estas características se reflejan en el modelo, éste tendería a subestimar los ingresos.

Identificación de sub poblaciones G1, G2 y G3

A partir de estas exclusiones se obtiene la base de datos final para modelamiento (desde aquí se denomina base de datos de modelamiento), que consta de 278,452 individuos (tabla 3.13); a partir de la base de datos de modelamiento se establecen tres sub poblaciones según el valor de la cuota estimada:

Cuota Maxima

Quintil	Mínimo	25%	75%	Max
Valor	0	107	438	16338

Sujetos que no tienen TC

	CUOTA ESTIMADA			
INGRESOS	<= 107	107 - 435	>435	
<= 450 usd	6,634	8,981	46	
451 - 1000	55,166	96,232	22,311	
1001 - 2500	10,149	34,756	29,767	
2501- 5000	1,656	6,853	14,532	
>5000	784	1,896	8,686	**298,449**
	74,389	**148,718**	**75,342**	

	CUOTA ESTIMADA		
INGRESOS	<= 107	107 - 435	>435
<= 450 usd	2.2%	3.0%	0.0%
451 - 1000	18.5%	32.2%	7.5%
1001 - 2500	3.4%	11.6%	10.0%
2501- 5000	0.6%	2.3%	4.9%
>5000	0.3%	0.6%	2.9%

EXCLUIDO	6.7%

Tabla 3.12: Relación entre Ingresos Reales y Cuota Estimada

- **Grupo 1:** Cuota estimada menor igual a 107 USD.
- **Grupo 2:** Cuota estimada entre 107 y 450 USD.
- **Grupo 3:** Cuota estimada mayor a 450 USD.

Sujetos que no tienen TC pero si tienen Cuota

CUOTA ESTIMADA (deuda total en créditos)	GRUPO	Sujetos	%	Ingreso REAL		Ingreso ESTIMADO	
				Media	Mediana	Media	Mediana
Menor a 107 usd	G1	65,315	23.5%	764	627	564	551
De 107 a 435 usd	G2	137,841	49.5%	1,013	760	540	523
Más de 435 usd	G3	75,296	27.0%	2,548	1,572	542	513
Total		**278,452**	**100.0%**				

Tabla 3.13: Sub poblaciones G1, G2 y G3

3.3 Elección de variables

Creación de bases de datos de entrenamiento y validación

En el contexto del desarrollo de modelos matemáticos y algoritmos predictivos, surge la necesidad imperante de establecer una adecuada partición de la base de datos general en sub conjuntos de entrenamiento y validación. Esta práctica, que guarda estrecha relación con los conceptos previamente abordados, se fundamenta en la búsqueda constante de una evaluación precisa y confiable del rendimiento y la capacidad de generalización de los modelos frente a datos desconocidos.

Desde una perspectiva coherente con las discusiones anteriores, esta división estratégica adquiere un significado crucial. En la construcción de modelos de estimación de ingresos, por ejemplo, la exclusión de registros con ingresos atípicamente bajos o altos pretende asegurar que el modelo se centre en patrones y relaciones realistas. Sin embargo, al trabajar con una única base de datos, existe el riesgo inherente de que el modelo pueda capturar particularidades y ruido, sesgando así su capacidad de generalización a nuevos datos.

En este contexto, la partición de los datos en conjuntos de entrenamiento y validación emerge como una salvaguarda esencial contra el sobreajuste. Al entrenar el modelo únicamente en los datos de entrenamiento, se facilita el aprendizaje de patrones y tendencias fundamentales, y al evaluar su desempeño en el conjunto de validación, se obtiene una evaluación imparcial de su capacidad para hacer predicciones en escenarios del mundo real.

La partición adecuada de la base de datos en subconjuntos de entrenamiento y validación se erige como un pilar fundamental en el desarrollo de modelos matemáticos. Esto se alinea perfectamente con las decisiones anteriores de excluir ciertas casuísticas para lograr una estimación precisa de ingresos y asegurar que el modelo capture adecuadamente los patrones. A través de esta práctica, se logra reducir el riesgo de sobreajuste, se facilita la optimización del modelo y se garantiza que el algoritmo pueda ofrecer predicciones confiables y generalizables en situaciones del mundo real.

Así, de la base de datos modelamiento se toma el 50% para la base de batos entrenamiento y el otro 50% para la base de batos validación; los individuos son seleccionados empleando muestreo aleatorio simple. Esta proporción del tamaño para cada sub base de datos es adecuada, dado que tenemos una población bastante grande y robusta.

Creación y combinación de variables

La creación y combinación de variables desempeña un papel crucial en la construcción de modelos analíticos y predictivos, este procedimiento permite enriquecer y refinar el conjunto de datos original, transformándolo en una fuente aún más valiosa de información.

En el caso presentado, se inició con un conjunto de 1186 variables. Sin embargo, al ejecutar las operaciones de acumulación, generación de ratios y combinación entre variables, se logró aumentar la riqueza informativa del conjunto de datos. La dimensión final del conjunto, que comprende 1882 variables.

La creación de variables acumuladas, que resultó en 73 nuevas variables, permite condensar la información relevante de varias variables en una sola, simplificando la estructura del conjunto de datos y proporcionando una perspectiva agregada; es decir, recoge la información de las distintas variables que se hallan en varias temporalidades: 3, 6, 12, 18, 24 y 36 meses.

La generación de ratios, que sumó 565 variables adicionales, aporta un enfoque relacional y comparativo entre variables. Estos ratios pueden ofrecer una comprensión más profunda de las relaciones existentes en los datos. Además, los ratios normalizados pueden ser particularmente útiles para minimizar efectos de escala y hacer que las variables sean más comparables.

La combinación de variables, que resultó en 58 nuevas variables, agrega aún más capas de complejidad y conocimiento al conjunto de datos y pueden capturar de manera más efectiva la influencia de una variable sobre otra, recogiendo así las características de los diferentes sistemas: bancario, cooperativa y sistema comercial. Esta práctica puede desvelar conexiones no evidentes entre variables, lo que potencialmente permite

a los modelos identificar patrones ocultos y mejorar la precisión de las predicciones.

En síntesis, el proceso de creación y combinación de variables resultó en un total de 696 nuevas variables (73 + 565 + 58), transformando un conjunto de datos inicialmente grande y complejo en una forma más procesable y significativa. Esta práctica permite capturar relaciones, tendencias y patrones que podrían ser cruciales para el rendimiento de modelos matemáticos.

Rango de ingreso real como variable categórica

Para tener una mejor comprensión del Ingreso Real, se efectúa comparaciones entre los percentiles 30 y 70 de cada grupo: G1, G2 y G3, para así crear una nueva variable categórica que se denomina Rango de Ingresos por grupos, pues para entrenar y validar a los modelos, no hace sentido contrastar dato a dato.

Estos resultados, que representan los valores en dólares, se resumen en la siguiente tabla 3.14:

Grupo	Percentil 30	Percentil 70
G1	530	800
G2	600	1006
G3	1023	2598

Tabla 3.14: Percentil 30 y 70 del Ingreso Real por grupos

Es importante recordar que: estos cálculos se efectúan en la base de datos de entrenamiento.

Este proceso de asignación de categorías basadas en los percentiles previamente calculados permite tener una nueva perspectiva sobre la distribución de los ingresos en cada grupo, facilitando así un análisis más detallado y diferenciado de los datos (tabla 3.15).

Grupo G1 (entrenamiento):

Rango de Ingresos	Categoría	Individuos
<= 530	1	9867
531 - 800	2	13512
>800	3	9307
		32686

Grupo G2 (entrenamiento):

Rango de Ingresos	Categoría	Individuos
<= 600	1	22631
601 - 1005	2	25668
>1005	3	20707
		69006

Grupo G3 (entrenamiento):

Rango de Ingresos	Categoría	Individuos
<= 1023	1	11251
1024 - 2597	2	15020
>2597	3	11261
		37532

Tabla 3.15: Rangos de Ingreso Real Por grupos

Selección de variables cuantitativas (Test KS)

Se realizó un análisis del test de Kolmogorov-Smirnov para evaluar la capacidad discriminativa de las variables independientes en los tres grupos de modelado.

Aquí se presenta un algoritmo para la selección de variables resultantes después de aplicar el método KS en el análisis:

1. Se identificó las variables con un valor de KS (Kolmogorov-Smirnov) mayor igual a 0.20. Estas variables se consideran significativas para el modelo.

2. De las variables significativas, se realizó una selección cuidadosa para evitar redundancia. Dado que una misma variable puede estar presente en diferentes temporalidades, elegimos la variable con el valor de KS más alto entre las que comparten el mismo nombre y temporalidades.

3. Se marcó en color celeste las variables que hemos seleccionado para incluir en el modelo A.1.

4. Finalmente, se incorporó al modelo las variables que han pasado todas las etapas de selección, es decir, las últimas variables seleccionadas que cumplen con los criterios de KS alto, no redundancia y baja correlación.

Ver tabla A.1 de variables cuantitativas seleccionadas.

Para el grupo 1 se seleccionó 18 variables; para el grupo 2 se seleccionó 35 variables y para el grupo 3 se seleccionó 28 variables.

Selección de variables cualitativas (Test VI)

En el proceso de selección de variables cualitativas, también se aplicó un enfoque similar utilizando el Test de Importancia de Variable (VI), que nos ayuda a medir la relevancia de las variables en el modelo. Aquí se detalla el procedimiento:

1. Se calcula el Valor de Importancia de Variable (VI) para cada variable. Este valor puede variar desde un 0% hasta un 100%.

    - VI entre 0% y 3%: Se considera que la variable tiene una influencia muy débil en el modelo.
    - VI entre 3% y 10%: La variable aporta una influencia débil en el modelo.
    - VI entre 10% y 30%: La variable muestra un aporte de influencia moderada en el modelo.
    - VI entre 30% y 50%: La variable tiene una influencia fuerte en el modelo.
    - VI mayor a 50%: La variable tiene una influencia muy fuerte en el modelo.

2. Para el Grupo 1, Se seleccionó las variables que tienen un VI entre 10% y 50%. En total, elegimos 3 variables que cumplen con este criterio.

3. En el caso del Grupo 2, se optó por las variables con un VI entre 10% y 50%. Similar al Grupo 1, Se seleccionó 3 variables en total.

4. Para el Grupo 3, se aplicó un enfoque diferente. Se seleccionó aquellas variables cuyo VI sea mayor al 15%. Esto nos brinda un total de 31 variables.

El proceso de selección basado en el Test de Importancia de Variable (VI) nos permite elegir de manera fundamentada las variables más relevantes para cada grupo, considerando su influencia en el modelo y su capacidad para aportar información significativa en el análisis.

Ver tabla A.2 de variable cualitativas.

Las variables cualitativas en consideración han sido sometidas a una evaluación exhaustiva y a un análisis detenido en colaboración con el profesor tutor de esta investigación. Según su orientación, se determinó que estas variables no poseen relevancia en el contexto de un modelo predictivo de ingresos. Es importante destacar que, si se estuviera desarrollando un modelo de scoring crediticio, estas mismas variables podrían proporcionar información de gran valor. Sin embargo, para los objetivos de este estudio en particular, su contribución resulta insuficiente y, por lo tanto, no se incluirán en el modelo final.

3.4 Representatividad en los nodos hoja

Los modelos de Random Forest (RF), Gradient Boosting Machine (GBM) y Extreme Gradient Boosting (XGBoost) son algoritmos de aprendizaje automático que se fundamentan en el concepto de árboles de decisión. Estos algoritmos son eficaces para resolver problemas de clasificación y regresión al construir una serie de árboles de decisión interconectados.

En estos modelos, los árboles de decisión se generan de tal manera que se busca minimizar la impureza o el error en las hojas del árbol. Un aspecto crucial es el número mínimo de individuos que se permite en una hoja. Esta restricción garantiza que las decisiones tomadas por cada árbol se basen en un subconjunto de la población que sea lo suficientemente representativo y no esté sesgado por individuos atípicos. Es interesante notar que estas hojas pueden corresponder a porcentajes específicos de la población, como las colas de la distribución.

Es así, que es imperante determinar de manera clara el porcentaje de pesos en las colas de la distribución para cada grupo, así como el número de arboles aleatorios adecuados para el modelo y el número variables predictoras que se seleccionan aleatoriamente en cada división de nodo.

1. El porcentaje o representatividad de las colas se explora entre 0% al 10%. Este porcentaje específico permite controlar de manera más adecuada y eficiente los porcentajes de remuestreos en las colas de las distribuciones; pues se recuerda, que uno de los objetivos de este estudio es mejorar las predicciones de ingresos en las colas de las distribución; es decir, para personas con ingresos muy altos y muy bajos.

2. El número adecuado de árboles de decisión se explora entre 300, 400 y 500.

3. El número de variables predictoras para modelos de predicción está dado por: numero de variables que ingresan al modelo entre 3.

En el contexto de esta investigación, no se aborda la optimización de los restantes parámetros asociados a los modelos considerados: Bosque

Aleatorio (RF), Gradient Boosting (GBM) y Extreme Gradient Boosting (XGBoost). Esta elección se fundamenta en la complejidad y profundidad que implica abordar la optimización exhaustiva de estos parámetros, lo que requeriría proyectos separados y extensos en sí mismos. El enfoque central de este estudio se concentra en la metodología y los pasos esenciales para el desarrollo de los modelos predictivos en poblaciones con altos ingresos y bajos ingresos.

Algoritmo (proceso iterativo)

para encontrar el equilibrio entre la representatividad y la precisión del modelo en relación al porcentaje de individuos en los nodos hojas (min.node.size):

1. Establecer un rango inicial de valores para min.node.size que se ajuste al problema y conjuntos de datos. 0% al 10% del total de la base (en cada grupo G1, G2 y G3)

2. Para cada valor de min.node.size en el rango, entrenar un modelo de Random Forest utilizando ese valor y evaluar su desempeño en términos de precisión. Esto implica dividir tus datos en conjuntos de entrenamiento y prueba, ajustar el modelo en el conjunto de entrenamiento y evaluar su capacidad de predicción en el conjunto de prueba.

3. Comparar los resultados de precisión obtenidos para cada valor de min.node.size y analizar cómo varía la representatividad de las colas. Observa si hay un punto en el que la precisión comienza a disminuir significativamente o si la representatividad de las colas es demasiado baja o alta.

4. Ajustar el rango de valores de min.node.size en función de los resultados obtenidos en el paso anterior. Si encuentras un rango más estrecho donde los valores de min.node.size producen un buen equilibrio entre precisión y representatividad, se puede reducir el rango de búsqueda.

5. Repetir los pasos 2 a 4 varias veces, refinando el rango de búsqueda y evaluando los modelos con diferentes valores de min.node.size.

Esto te permitirá acercarte gradualmente al valor óptimo que equilibre la representatividad y la precisión del modelo.

6. Finalmente, seleccionar el valor de min.node.size que proporcione el mejor equilibrio entre la representatividad y la precisión del modelo, teniendo en cuenta también otros factores como la interpretabilidad y el costo computacional.

3.4.1 Grid de hiper parámetros G1

Luego de rigurosas simulaciones, se lograron determinar los valores óptimos para los tres parámetros mencionados anteriormente (tabla A.3).

Número de árboles	300
Número de variables predictoras	5
Nodos finales al	3.6%
Min nod size	1177
N nodos	28

Tabla 3.16: Resumen Grid de hiperparámetros del Grupo 1

3.4.2 Grid de hiperámetros G2

Luego de rigurosas simulaciones, se lograron determinar los valores óptimos para los tres parámetros mencionados anteriormente (tabla A.4).

Número de árboles	300
Número de variables predictoras	11
Nodos finales al	3.6%
Min nod size	2484
N nodos	28

Tabla 3.17: Resumen Grid de hiperparámetros del Grupo 2

3.4.3 Grid de hiperámetros G3

Luego de rigurosas simulaciones, se lograron determinar los valores óptimos para los tres parámetros mencionados anteriormente (tabla A.5).

Número de árboles	300
Número de variables predictoras	9
Nodos finales al	3.6%
Min nod size	1351
N nodos	28

Tabla 3.18: Resumen Grid de hiperparámetros del Grupo 3

3.5 Función de balanceo para remuestreo

La función de balanceo está diseñada para realizar un proceso de remuestreo que equilibra las colas de las distribuciones en función de los cortes y porcentajes definidos. Esto resulta ser útil para abordar el desbalance en los datos cuando se trabaja con modelos predictivos, como los que se ha estado discutiendo anteriormente.

Primero, la función divide los datos en tres conjuntos: **nb** para aquellos individuos con ingresos menores o iguales al **corte1**, **na** para aquellos con ingresos mayores o iguales al **corte2**, y **nc** para aquellos con ingresos entre **corte1** y **corte2**. Luego, calcula el peso inicial de las colas izquierda (**perc1**) y derecha (**perc2**) en relación con el tamaño total del conjunto de datos.

Luego, viene la parte de remuestreo. Se calcula el número de individuos que se deben muestrear en cada cola (**nperc1** y **nperc2**) utilizando los porcentajes proporcionados (**porc1** y **porc2**) y los pesos iniciales de las colas. Se utiliza la función *sample* para realizar el muestreo aleatorio con reemplazo de los individuos de las colas.

Después del remuestreo, se crea una nueva base de datos combinando los individuos remuestreados de las colas con los individuos que estaban entre los cortes (**nc**).

Finalmente, la función calcula y muestra los pesos finales de las colas izquierda y derecha en la nueva base de datos resultante, ayudando a visualizar el efecto del remuestreo en el equilibrio de los datos.

Es decir, esta función busca abordar el desbalance en los datos mediante el remuestreo de los individuos en las colas de las distribuciones de ingresos. El proceso ayuda a igualar la representación de los grupos con ingresos extremadamente bajos o altos, permitiendo una mejor representación en el entrenamiento de los modelos predictivos. Esta técnica se integra en el proceso general de modelado para mejorar la precisión y el rendimiento de los modelos.

En el proceso de entrenamiento de los modelos, en cada grupo se deberá insertar los siguientes insumos o parámetros a la función de balanceo:

A modo de ilustración, se tiene una sentencia de remuestreo para el grupo 1.

```
G1_corte1 <- 601  #Percentil 40% Ingreso Real
G1_corte2 <- 950  #Percentil 80% Ingreso Real
G1_porc1 <- 0.65  #Nuevo percentil para 601 USD
G1_porc2 <- 0.13  #Nuevo percentil para 950 USD
mod_g1 <- BDD_G1  #Base de datos train del grupo 1

rmod_g1 <- boots_2tail(mod_g1, corte1 =  G1_corte1, corte2 = G1_corte2,
    porc1 = G1_porc1, porc2 = G1_porc2)
#rmod_g1: nueva base de datos remuestreado
```

Código 3.1: Parámetros para la función de Balanceo

Los desarrollos explícitos de estos códigos se halla en la sección de Anexos.

3.6 Modelos para población G1

3.6.1 Exploración y descripción de la base de datos

Esta población cuenta con sujetos con cuota estimada actual menor o igual a 107 dólares.

Se comentó en la sección de <entrenamiento y validación> que para entrenar a los modelos de Machine Learning, como estándar se exige que haya al menos 4 mil registros; dado que tenemos más de 4 mil registros optamos por tomar la mitad para entrenamiento y la otra mitad para validación (tabla 3.19).

Muestra	N	%
Entrenamiento	32,686	50.04%
Validación	32,629	49.96%
Total	**65,315**	**100.00%**

Tabla 3.19: Población del Grupo 1

Distribución del Ingreso Real del Grupo 1

Percentil	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Valor en USD	450.01	481.13	483.23	483.27	500.00	500.00	528.18	560.00	600.00	600.00	626.54

Percentil		55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
Valor en USD		670.979	708.746	753.00	800.00	856.21	920.00	1029.719	1206.00	1552.65	2500.00

Tabla 3.20: Percentil del Ingreso Real del Grupo 1

Esta tabla 3.20 de percentil será de mucha ayuda para el entrenamiento del modelo, dado que ayuda a determinar los cortes y proporciones adecuados para los parámetros del remuestreo.

3.6.2 Modelo RF para G1

En las próximas secciones, se brindará una descripción exhaustiva del proceso de entrenamiento del modelo de Bosques Aleatorios (Random Forest, RF), así como de la metodología utilizada para la creación y presentación de las tablas de resultados. Se abordarán los pasos clave en la configuración y ajuste del modelo RF, junto con una explicación detallada de las métricas utilizadas para evaluar su rendimiento. Asimismo, se proporcionará un análisis profundo de las tablas generadas, las cuales capturan de manera concisa y clara los resultados obtenidos a partir de los experimentos y validaciones realizadas. Estas secciones permitirán al lector comprender en detalle el enfoque adoptado y los criterios de evaluación aplicados en el desarrollo de este estudio.

En el caso de los modelos RLM, GBM y XGB aplicados en los grupos G1, G2 y G3, se sigue un enfoque análogo al presentado para el modelo RF en el grupo G1. Sin embargo, para evitar redundancias y mantener el enfoque en los aspectos más relevantes, se optará por una presentación más sintética en estas secciones.

Las etapas de selección y ajuste de hiper parámetros se llevarán a cabo de manera similar a como se describirá en esta sección. No obstante, en lugar de entrar en detalles repetitivos, se expondrán los resultados y observaciones más destacados de cada modelo y grupo. Esto permitirá resaltar los aspectos esenciales del análisis y mantener la presentación del documento de manera eficiente.

Es importante resaltar que, a pesar de la concisión en la presentación, se mantiene la integridad metodológica y la rigurosidad en la aplicación de los modelos. El objetivo principal es comunicar de manera efectiva las conclusiones y resultados clave sin abrumar al lector con información redundante. Este enfoque permitirá una comprensión clara y precisa de la eficacia de los modelos RF, RLM, GBM y XGB en los diferentes grupos de estudio.

Con el propósito de proporcionar una guía comprensible y ejemplar, se procederá a detallar la explicación correspondiente a los modelos más destacados elegidos en cada categoría y bajo cada algoritmo de modelado.

Parámetros para remuestreo y función de balanceo

El objetivo principal del proceso de remuestreo es aumentar el tamaño de las colas en las distribuciones de ingresos para lograr un mejor equilibrio en los datos. Para lograr esto, se considera un valor menor o mayor al percentil real.

Luego, se aplica la función 'boots_2tail' (función de balanceo o remuestreo) a la base de datos 'mod_g1', con los parámetros 'corte1', 'corte2', 'porc1' y 'porc2' específicos para el grupo G1. Esta función realiza un proceso de remuestreo donde se seleccionan individuos de las colas izquierda y derecha de la distribución de ingresos, buscando aumentar su tamaño en función de los valores de los percentiles deseados.

Nota: 'mod_g1': es la base de datos de entrenamiento del grupo G1.

El resultado de este proceso es una nueva base de datos 'rmod_g1', en la cual se ha logrado incrementar el número de individuos en las colas de la distribución. Este enfoque tiene un efecto particular en el rango del centro de la distribución, donde el modelo suele tener un rendimiento óptimo en sus estimaciones. En esencia, el remuestreo está redirigiendo individuos desde el centro hacia las colas, permitiendo evaluar cómo las modificaciones en la distribución afectan el rendimiento del modelo en diferentes rangos de ingresos.

Nota: mod_g1: es la base de datos original del grupo 1; rmod_g1: es la base de datos resultante luego del remuestreo para el grupo 1; estas dos bases de datos se usa al correr el modelo del RF.

Se presenta la instancia del ejemplo:

```
G1_corte1 = 601  #40%
G1_corte2 = 950  #80%
G1_porc1 = 0.65
G1_porc2 = 0.13

rmod_g1 <- boots_2tail(mod_g1, corte1 =  G1_corte1, corte2 = G1_corte2,
    porc1 = G1_porc1, porc2 = G1_porc2)
```

Código 3.2: Parámetros para remuestreo y Remuestreo de G1

Representatividad en los nodos hojas en cada árbol

En la sección del grid de hiperparámetros, se determinó que una representatividad adecuada para los nodos hoja es del 3.6% con respecto a la base de datos de entrenamiento que se utilizará en el modelo. Ahora, en esta etapa, se debe calcular cuántos individuos corresponden al 3.6% en la base de datos 'mod_g1' o 'rmod_g1'.

Para hacerlo, primero se obtiene el número total de individuos en la base de datos 'rmod_g1' (o 'mod_g1') y se almacena en la variable 'n1'. Luego, se utiliza la función 'quantile' para calcular los percentiles correspondientes a los valores del 3% al 4%, con un incremento de 0.2%. Esto ayuda a entender cómo varía el tamaño del conjunto de datos a medida que se aumenta el porcentaje desde el 3% hasta el 4%.

Los resultados obtenidos se presentan en forma de vector, donde se muestra la cantidad de individuos necesarios para alcanzar los porcentajes mencionados. En particular, el valor que nos interesa es el correspondiente al 3.6%, que es 1304 individuos para esta instancia (ejemplo). Esto significa que, para lograr una representatividad del 3.6% en los nodos hoja del modelo, se requerirían 1304 individuos de la base de datos 'rmod_g1' (o 'mod_g1') para entrenar y evaluar el modelo de forma adecuada.

```
n1 <- nrow(rmod_g1) #rmod_g1  o  mod_g1

round(quantile(0:n1, probs = seq(0.03,0.04,by=0.002)),0)
# 3%  3.2% 3.4% 3.6% 3.8%   4%
#1086 1159 1231 1304 1376 1449

#Vector con el porcentaje de presentatividad del 1% - 10%
presen_colas_g1 <- round(quantile(0:n1, probs = seq(0.03,0.04,by=0.002)
    ),0)
presen_colas_g1 <- as.data.frame(presen_colas_g1)
presen_colas_g1 <- presen_colas_g1[,1]
presen_colas_g1 <- presen_colas_g1[1:6]

#Numero del individuos que representan el 3.6% en los nodos hoja
presen_colas_g1[4]
#1304
```

Código 3.3: Representatividad de los nodos hoja

Creación del modelo con la BDD original o remuestreada

Habiendo determinado previamente los valores adecuados para los parámetros del modelo, el proceso procede con la ejecución del modelo RF (Random Forest). En este caso, se toma el valor de 'G1_min_nodo' que fue calculado anteriormente y corresponde al número de individuos necesarios para alcanzar una representatividad del 3.6% en los nodos hoja del modelo para el grupo G1.

La ejecución del modelo RF se realiza mediante la función 'ranger', que es parte de la librería de R para el entrenamiento de modelos de Random Forest. En esta función se definen varios argumentos importantes:

- **formula**: Representa la fórmula que define la relación entre las variables predictoras y la variable objetivo.
- **data**: Corresponde a la base de datos utilizada para entrenar el modelo ('rmod_g1' o 'mod_g1').
- **num.trees**: Indica el número de árboles que compondrán el Random Forest (para G1 se usan 300 árboles).
- **mtry**: Representa el número de variables predictoras seleccionadas aleatoriamente en cada división de un árbol (para G1 se usa 5).
- **min.node.size**: Este es el parámetro crítico que determina el número mínimo de observaciones requeridas en un nodo hoja del árbol. Se establece en 'G1_min_nodo'.
- **importance**: Se define como 'impurity' para calcular la importancia de las variables basándose en la medida de impureza.
- **write.forest**: Se establece en 'TRUE' para guardar los detalles del bosque (forest) entrenado.
- **seed**: Define la semilla para la generación de números aleatorios, asegurando la reproducibilidad.

```
1 G1_min_nodo = presen_colas_g1[4]
2
```

```r
rf_g1_5x100 <- ranger(formula = as.formula(formula_g1), data = rmod_g1,
    num.trees = 300, mtry = 5, min.node.size = G1_min_nodo, importance
    = 'impurity', write.forest = TRUE, seed = 1234)
```

Código 3.4: Hiper parámetros aplicados al entrenamiento del modelo RF para G1

Este proceso resulta en el entrenamiento del modelo RF 'rf_g1_5x100' para el grupo G1 con los hiper parámetros y la base de datos especificados. El modelo está listo para ser evaluado y utilizado para realizar predicciones en nuevos datos.

Así, tenemos el modelo entrenado:

```
Ranger result

Call:
 ranger(formula = as.formula(formula_g1), data = rmod_g1, num.trees =
    300,      mtry = 5, min.node.size = G1_min_nodo, importance = "
    impurity",      write.forest = TRUE, seed = 1234)

Type:                             Regression
Number of trees:                  300
Sample size:                      36213
Number of independent variables:  17
Mtry:                             5
Target node size:                 1304
Variable importance mode:         impurity
Splitrule:                        variance
OOB prediction error (MSE):       75686.51
R squared (OOB):                  0.1858215
```

Código 3.5: Modelo entrenado RF para G1

Se guarda el modelo:

```r
setwd(dir.m)
saveRDS(object = rf_g1_5x100, file = "modelo_RF_g1.rds")
```

Código 3.6: Guarda el modelo RF para G1

Es crucial enfatizar que el ejemplo presentado corresponde al mejor modelo de Random Forest (RF) seleccionado específicamente para el grupo G1. Estos resultados no fueron obtenidos de manera trivial, sino a través de un proceso que involucró múltiples simulaciones y pruebas exhaustivas.

La elección de los valores óptimos de los hiperparámetros, como el número de árboles (‘num.trees‘), la cantidad de variables predictoras seleccionadas (‘mtry‘), y el tamaño mínimo de los nodos hoja (‘min.node.size‘), requirió una evaluación rigurosa. Además, se llevaron a cabo iteraciones y ajustes en estos hiper parámetros para lograr un modelo con un rendimiento óptimo.

En cada paso del proceso, se consideraron diferentes combinaciones de hiper parámetros y se evaluaron mediante métricas relevantes para la calidad del modelo. Finalmente, después de numerosas simulaciones y pruebas, se identificó el modelo RF que demostró tener el mejor rendimiento para el grupo G1 en términos de predicción de ingresos.

Este enfoque metódico y basado en pruebas garantiza que el modelo seleccionado sea lo más eficaz posible para la tarea de predicción de ingresos en el contexto de este estudio.

Predicción del grupo G1 sobre toda la base

En esta sección, se realiza la predicción sobre toda la base de datos de Modelamiento. Esto se realiza mediante la creación de una nueva variable llamada <INGRESO_EST_G1_5X100>, en la cual se almacenan las predicciones generadas por el modelo RF previamente seleccionado: rf_g1_5x100. El proceso de predicción se lleva a cabo a través de la línea de código siguiente:

```
runtime <- system.time({
info[, INGRESO_EST_G1_5x100 := predict(object=rf_g1_5x100, data=info)$
    predictions]
})

runtime_df <- data.frame(user = runtime[1],
                         system = runtime[2],
                         elapsed = runtime[3])
```

Código 3.7: Predicción con RF de G1 sobre toda la BDD de Modelamiento

Una vez completada la predicción, se calcula el tiempo total de ejecución utilizando la función "system.time()". Este tiempo se desglosa en categorías de tiempo del usuario, del sistema y el tiempo total de ejecución y se almacena en el dataframe <runtime_df> para futuras referencias y

análisis detallado. En conjunto, este proceso permite obtener las predicciones estimadas y evaluar la eficiencia del procedimiento en términos de tiempo.

- **user**: Es el tiempo de CPU utilizado por el proceso en modo de usuario, es decir, el tiempo que la CPU dedicó a ejecutar el código del usuario.
- **system**: Es el tiempo de CPU utilizado por el sistema operativo para ejecutar el código del kernel y otras tareas del sistema.
- **elapsed**: Es el tiempo total transcurrido desde el inicio de la ejecución hasta su finalización, incluyendo el tiempo de CPU utilizado y cualquier tiempo de espera o bloqueo que pueda haber ocurrido.

El elapsed incluye el tiempo de CPU utilizado tanto por el usuario (user) como por el sistema operativo (system); ambos tiempos sumados dan como resultado el tiempo total de ejecución (elapsed).

Es importante tener en cuenta que los tiempos de ejecución pueden variar dependiendo del hardware y del sistema operativo en el que se esté ejecutando el código; también pueden variar si hay otras tareas en ejecución en la computadora que estén utilizando recursos

Rangos reales y estimados

Se analizan los rangos reales y estimados para el grupo G1 de ingresos. Mediante el uso de la función <quantile>, se calculan los intervalos que representan los diferentes percentiles de la distribución real de ingresos.

Con el fin de comparar y visualizar los rangos reales y estimados, se procede a dividir el grupo G1 en cinco subgrupos de ingresos. Estos subgrupos se establecen basándose en los percentiles de 20, 40, 60, 80 y 100. Esto garantiza que haya un número similar de individuos en cada intervalo de ingresos.

En esencia, esta sección permiten analizar los rangos de ingresos reales y estimados en el grupo G1, y además, crear subgrupos para una comparación más detallada entre los valores reales y las estimaciones proporcionadas por el modelo RF seleccionado.

Nótese que: estos rangos real y estimado se guardan en las variables <RANGO_REALG1> y <RANGO_EST_G1_5X100>

```
quantile(info[GRUPO_CUOTA=="G1"]$INGRESO_REAL, probs=seq(0,1,by=0.20))
# 0%      20%      40%      60%      80%     100%
#450.010  500.000  600.000  708.746  920.000 2500.000

#Crea Rango Real
info[, RANGO_REALG1 := cut(INGRESO_REAL, breaks = c(450, 500, 600, 750,
     950, 2500), labels = c("[450-500]", "(500-600]", "(600-750]", "
    (750-950]", "(950-250]"))]

#Crea Rango Estimado
info[, RANGO_EST_G1_5x100 := cut(INGRESO_EST_G1_5x100, breaks = c(450,
    500, 600, 750, 950, 2500), labels = c("[450-500]", "(500-600]", "
    (600-750]", "(750-950]", "(950-2500]"))]
```

Código 3.8: Rangos Reales y Estimados del G1

Matriz de coincidencias

En esta sección se indica como se genera la matriz de coincidencia, una herramienta que permite evaluar la similitud entre la distribución de rangos reales y estimados sin necesidad de recurrir a la representación gráfica o histogramas.

La matriz de coincidencia nos muestra cómo se distribuyen los datos en los diferentes rangos reales (filas) en comparación con los rangos estimados por el modelo RF (columnas). Cada valor en la matriz representa la cantidad de individuos que coinciden en un rango real y estimado específico.

```
G1M_train_R_E_G1_5x100 <- info[ModVal == 0 & GRUPO_CUOTA == "G1"][,
    table(RANGO_REALG1, RANGO_EST_G1_5x100)]

G1M_val_R_E_G1_5x100 <- info[ModVal == 1 & GRUPO_CUOTA == "G1"][,table(
    RANGO_REALG1, RANGO_EST_G1_5x100)]
```

Código 3.9: Matriz de coincidencia de G1

Matrices de coincidencia de entrenamiento y validación.

```
> G1M_train_R_E_G1_5x100
            RANGO_EST_G1_5x100
```

```
RANGO_REALG1 [450-500] (500-600] (600-750] (750-950] (950-2500]
   [450-500]          0       4031       3366       1042          86
   (500-600]          0       2250       3158       1534         120
   (600-750]          0       1312       2520       1570         148
   (750-950]          0        625       2435       1992         246
   (950-250]          0        128       1852       3438         833
> G1M_val_R_E_G1_5x100
             RANGO_EST_G1_5x100
RANGO_REALG1 [450-500] (500-600] (600-750] (750-950] (950-2500]
   [450-500]          0       3876       3505       1156          93
   (500-600]          0       2209       3118       1570         107
   (600-750]          0       1320       2609       1580         150
   (750-950]          0        751       2305       1901         224
   (950-250]          0        145       1908       3313         789
```

Código 3.10: Resultados Matriz de coincidencia de G1

En la sección de Resultados, se presentan los resultados detallados de las matrices de coincidencia para los conjuntos de entrenamiento y validación.

Métricas

Se calcula el Error Cuadrático Medio (MSE) y el Error Absoluto Medio (MAE) para el grupo G1 en el modelo RF seleccionado. El MSE se obtiene al elevar al cuadrado la diferencia entre los valores reales y estimados del ingreso, promediando luego estos valores. De manera similar, el MAE se obtiene calculando el valor absoluto de la diferencia entre los valores reales y estimados del ingreso, y luego promediando esos valores.

Estos cálculos se realizan por separado para los conjuntos de entrenamiento (ModVal == 0) y validación (ModVal == 1), y se agrupan por la variable ModVal para proporcionar una comparación entre ambos conjuntos. La ordenación por ModVal asegura que los resultados estén presentados de manera coherente. Estos valores de MSE y MAE son indicadores clave para evaluar la precisión del modelo en la predicción de los ingresos en el grupo G1.

```
1
2 #Error cuadratico medio
3 MSE_G1_5x100_train_val <- info[, MSE_G1_5x100 := (INGRESO_REAL -
     INGRESO_EST_G1_5x100)^2][GRUPO_CUOTA == "G1"][,list(MSE_G1_5x100 =
```

```
    mean(MSE_G1_5x100)), by = ModVal][order(ModVal)]

# Error absoluto medio (MAE)
MAE_G1_5x100_train_val <- info[, MAE_G1_5x100 := abs(INGRESO_REAL -
    INGRESO_EST_G1_5x100)][GRUPO_CUOTA == "G1"][,list(MAE_G1_5x100 =
    mean(MAE_G1_5x100)), by = ModVal][order(ModVal)]
```

Código 3.11: Métricas del Modelo RF en G1

Resultados de Métricas para entrenamiento y validación.

```
#MSE para Train y Test
> MSE_G1_5x100_train_val
   ModVal MSE_G1_5x100
1:      0     111004.8
2:      1     108794.2

#MAE para Train y Test
> MAE_G1_5x100_train_val
   ModVal MAE_G1_5x100
1:      0     211.7396
2:      1     211.2356
```

Código 3.12: Resultados Métricas del Modelo RF en G1

Además, se calcula el Error Cuadrático Medio (MSE), dividido en los conjuntos de entrenamiento y validación, por rangos de ingresos reales (RANGO_REALG1) y se ordenan de acuerdo a estos rangos para presentar los resultados de manera coherente. El MSE por rangos permite entender cómo el modelo está funcionando en diferentes intervalos de ingresos, tanto en el conjunto de entrenamiento como en el de validación, lo que proporciona una visión más detallada de su desempeño en distintos segmentos de la población.

```
#MSE por Rango de Ingreso Real para Train
train_MSE_G1_5x100_Rreal <- info[GRUPO_CUOTA == "G1" & ModVal == 0][,
    list(MSE_G1_5x100 = mean(MSE_G1_5x100)), by = RANGO_REALG1][order(
    RANGO_REALG1)]

#MSE por Rango de Ingreso Real para Test
val_MSE_G1_5x100_Rreal <- info[GRUPO_CUOTA == "G1" & ModVal == 1][,list
    (MSE_G1_5x100 = mean(MSE_G1_5x100)), by = RANGO_REALG1][order(RANGO
    _REALG1)]
```

Código 3.13: MSE por Rangos de Ingresos del Modelo RF en G1

Resultados del MSE por rangos de Ingresos Reales para entrenamiento y validación.

```
#MSE por Rangos de Ingreso Real para Train
> train_MSE_G1_5x100_Rreal
   RANGO_REALG1 MSE_G1_5x100
1:    [450-500]     32058.90
2:    (500-600]     23240.11
3:    (600-750]     14125.91
4:    (750-950]     25067.36
5:    (950-250]    476671.58

#MSE por Rangos de Ingreso Real para Test
> val_MSE_G1_5x100_Rreal
   RANGO_REALG1 MSE_G1_5x100
1:    [450-500]     33807.78
2:    (500-600]     23297.02
3:    (600-750]     14343.95
4:    (750-950]     26603.96
5:    (950-250]    467246.88
```

Código 3.14: Resultados MSE por Rangos de Ingresos del Modelo RF en G1

En la sección de resultados, se presentan los resultados detallados de las matrices de métricas para los conjuntos de entrenamiento y validación.

Resultados

En esta sección se presentan los resultados obtenidos mediante la implementación del mejor modelo de Random Forest (RF) para el grupo G1, tanto en la base de datos original como en la base de datos sometida a remuestreo. Este enfoque permitirá comparar y evaluar el impacto del proceso de remuestreo en la precisión y rendimiento del modelo.

Los resultados se examinarán considerando diferentes métricas de evaluación, como el Error Cuadrático Medio (MSE) y el Error Absoluto Medio (MAE). Además, se analizará la distribución de las predicciones realizadas por el modelo en relación con los ingresos reales de los individuos. Este análisis detallado permitirá entender cómo el modelo se comporta en distintos rangos de ingresos y cómo el remuestreo puede influir en su capacidad predictiva.

Se busca así proporcionar una visión integral de cómo el proceso de remuestreo puede mejorar o afectar las predicciones del modelo, lo que resulta crucial para tomar decisiones informadas sobre si implementar o no esta técnica en la construcción del modelo predictivo para el grupo G1.

Además, se exhibe la matriz de chi cuadrado de Pearson, una herramienta estadística esencial para el análisis de tablas de contingencia. En este contexto, la hipótesis nula (Ho) establece que las distribuciones entre el conjunto de entrenamiento y el conjunto de validación son similares, lo que se traduce en un porcentaje de coincidencia entre ambos conjuntos. Por otro lado, la hipótesis alternativa (Ha) sugiere que las distribuciones no son similares y, en consecuencia, el porcentaje de coincidencia entre el conjunto de entrenamiento y el conjunto de prueba es significativamente diferente.

La matriz de chi cuadrado de Pearson se utiliza para evaluar si las diferencias observadas entre los conjuntos de entrenamiento y prueba son estadísticamente significativas. Esto permite determinar si las diferencias en los porcentajes de coincidencia pueden atribuirse al azar o si indican una discrepancia genuina entre las distribuciones de los datos en ambos conjuntos.

Métricas de MSE y MAE:

El Error Cuadrático Medio (MSE) y el Error Absoluto Medio (MAE) son métricas clave para evaluar la precisión de un modelo de predicción. Observamos que tanto en el conjunto de entrenamiento como en el conjunto de validación, el MSE es similar en ambos modelos tanto en la BDD de entrenamiento y validación, lo que sugiere que no hay una mejora significativa en la precisión al aplicar el remuestreo. Esto puede indicar que el modelo base ya está capturando la variabilidad de los datos (tabla: 3.21 y 3.22).

Cruces y Sub/Sobre Estimaciones:

Los cruces indican qué porcentaje de estimaciones se encuentra cerca de los valores reales, mientras que las subestimaciones y sobreestimaciones indican en qué medida el modelo predice valores por debajo o por encima de los reales. En ambos modelos y en ambas muestras (entrenamiento

y validación, sin remuestreo y con remuestreo), los cruces son relativamente consistentes (tabla: 3.21 y 3.22).

En modelo base sin remuestreo; el cruce en entrenamiento y validación son similares, así como la sub estimación y sobre estimación; esto nos sugiere que el modelo es consistente. El cruce está bajo el 70%, tenemos una sobre estimación cerca del 28% y una sub estimación cercana del 3%.

En el modelo con remuestreo; los cruces, sobre y sub estimación también son consistentes; confirmando que el modelo efectivamente es consistente en las predicciones. El cruce mejora y está sobre el 70%, se reduce la sobre estimación al 20% y la subestimación al 8% aproximadamente.

Tiempos de ejecución:

Se puede observar que los tiempos de ejecución para el modelo sin remuestreo y el modelo con remuestreo son bastante similares. En ambos casos, los tiempos de ejecución son relativamente bajos, lo que indica que ambos modelos son eficientes en términos de procesamiento computacional.

Dado que los tiempos de ejecución son muy similares en ambos casos, no parece haber una diferencia significativa en términos de eficiencia entre los dos modelos. Esto sugiere que la aplicación del remuestreo no ha tenido un impacto adverso sustancial en el tiempo de procesamiento. En otras palabras, el proceso adicional de remuestreo no ha aumentado significativamente la carga computacional, a lugar el modelo con remuestreo tiene un tiempo mejor de ejecución (tabla: 3.21 y 3.22).

Chi Cuadrado de Pearson:

La tabla de chi cuadrado proporciona información valiosa sobre la relación entre las categorías reales y estimadas en el modelo. Sin embargo, como se mencionó, el chi cuadrado puede ser sensible a pequeños cambios en los individuos, lo que puede llevar a interpretaciones limitadas o incluso a conclusiones erróneas si se depende únicamente de este valor.

En este caso, al observar la tabla de chi cuadrado y los valores esperados teóricos, podemos notar algunas discrepancias entre las categorías reales y estimadas. Las decisiones de rechazo o aceptación de la hipóte-

sis nula basadas en el valor del chi cuadrado deben ser interpretadas con precaución. Dado que este valor puede ser influenciado por la cantidad de individuos en cada categoría, es importante considerar con base a las métricas más estables y robustas como el MSE y el MAE, no se rechaza la hipótesis nula: las distribuciones de modelamiento y validación son similares; si vemos las gráficas de las distribuciones veremos que afectivamente son similares (tabla: 3.21 y 3.22, figura: 3.5).

Decisión:

Con base a los criterios expuestos, basándonos en la mejora sustancial en la precisión y la consistencia de las métricas de evaluación, así como en la comparación del tiempo de ejecución, el modelo con remuestreo es la elección recomendada para futuras aplicaciones.

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Random Forest
VARIABLES: 18 **G1**

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0	2929	3269	2074	253	8525.00	26%
(500-600]	0	1491	2607	2599	365	7062.00	22%
(600-750]	0	829	1873	2354	494	5550.00	17%
(750-950]	0	331	1479	2780	708	5298.00	16%
(950-250]	0	50	644	3646	1911	6251.00	19%
						32686.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	34%	38%	24%	3%	53133.45
(500-600]	0%	21%	37%	37%	5%	44444.31
(600-750]	0%	15%	34%	42%	9%	27597.35
(750-950]	0%	6%	28%	52%	13%	23203.67
(950-2500]	0%	1%	10%	58%	31%	380675.87

Entrenamiento: 0	Métricas		
MSE	104709.40		
MAE	221.73		
Cruce	25%	SubEstima	3%
Cruce +/-	69%	SobreEstim	28%

Muestra de Validación

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0	2809	3303	2246	272	8630.00	26%
(500-600]	0	1451	2607	2560	386	7004.00	21%
(600-750]	0	803	1944	2443	469	5659.00	17%
(750-950]	0	423	1439	2672	647	5181.00	16%
(950-250]	0	69	723	3532	1831	6155.00	19%
						32629.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	33%	38%	26%	3%	55691.45
(500-600]	0%	21%	37%	37%	6%	44625.13
(600-750]	0%	14%	34%	43%	8%	28065.35
(750-950]	0%	8%	28%	52%	12%	24293.78
(950-2500]	0%	1%	12%	57%	30%	376632.66

Validación: 1	Métricas		
MSE	104080.22		
MAE	222.27		
Cruce	24%	SubEstima	4%
Cruce +/-	68%	SobreEstim	28%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	χ^2	Teórico
[450-500]	0.0	4.9	0.4	14.3	1.4	**91.9**	**26.3**
(500-600]	0.0	1.1	0.0	0.6	1.2	**Decisión**	
(600-750]	0.0	0.8	2.7	3.4	1.3	Se rechaza H0	
(750-950]	0.0	25.6	1.1	4.2	5.3		
(950-2500]	0.0	7.2	9.7	3.6	3.3		

Tiempo de Ejecución		
user	system	elapsed
11.92	0.41	3.71

Tabla 3.21: Resultados RF sin remuestreo para G1

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G1		
MODELO: Random Forest	G1_corte1 = 601 #40%	G1_porc1 = 0.65
VARIABLES: 18	G1_corte2 = 950 #80%	G1_porc2 = 0.13

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0.00	4031.00	3366.00	1042.00	86.00	8525.00	26%
(500-600]	0.00	2250.00	3158.00	1534.00	120.00	7062.00	22%
(600-750]	0.00	1312.00	2520.00	1570.00	148.00	5550.00	17%
(750-950]	0.00	625.00	2435.00	1992.00	246.00	5298.00	16%
(950-250]	0.00	128.00	1852.00	3438.00	833.00	6251.00	19%
						32686.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	47%	39%	12%	1%	32058.90
(500-600]	0%	32%	45%	22%	2%	23240.11
(600-750]	0%	24%	45%	28%	3%	14125.91
(750-950]	0%	12%	46%	38%	5%	25067.36
(950-250]	0%	2%	30%	55%	13%	476671.58

Entrenamiento: 0	Métricas		
MSE	111004.82		
MAE	211.74		
Cruce	23%	SubEstima	8%
Cruce +/-	73%	SobreEstim	19%

Muestra de Validación

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0	3876	3505	1156	93	8630.00	26%
(500-600]	0	2209	3118	1570	107	7004.00	21%
(600-750]	0	1320	2609	1580	150	5659.00	17%
(750-950]	0	751	2305	1901	224	5181.00	16%
(950-250]	0	145	1908	3313	789	6155.00	19%
						32629.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	45%	41%	14%	1%	33807.78
(500-600]	0%	31%	44%	22%	2%	23297.02
(600-750]	0%	24%	47%	28%	3%	14343.95
(750-950]	0%	14%	44%	36%	4%	26603.96
(950-250]	0%	2%	31%	53%	13%	467246.88

Validación: 1	Métricas		
MSE	108794.19		
MAE	211.24		
Cruce	23%	SubEstima	9%
Cruce +/-	71%	SobreEstim	20%

Chi Cuadrado de Pearson

Real	Estimado					χ^2	Teórico
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0.0	6.0	5.7	12.5	0.6	**80.8**	**26.3**
(500-600]	0.0	0.7	0.5	0.8	1.4	**Decisión**	
(600-750]	0.0	0.0	3.1	0.1	0.0	Se rechaza H0	
(750-950]	0.0	25.4	6.9	4.2	2.0		
(950-250]	0.0	2.3	1.7	4.5	2.3		

Tiempo de Ejecución		
user	system	elapsed
18.5	0.05	18.5

Tabla 3.22: Resultados RF con remuestreo para G1

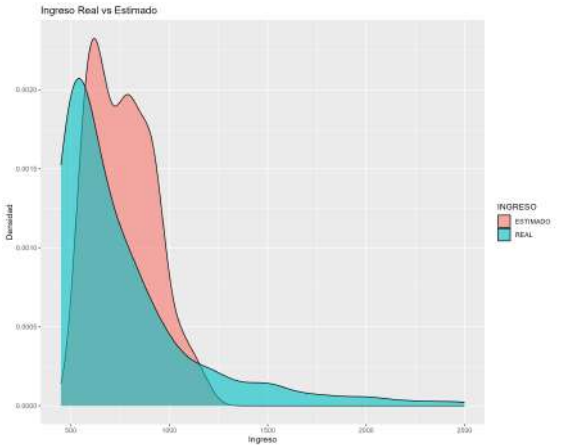


Figura 3.1: Sin remuestreo entrenamiento

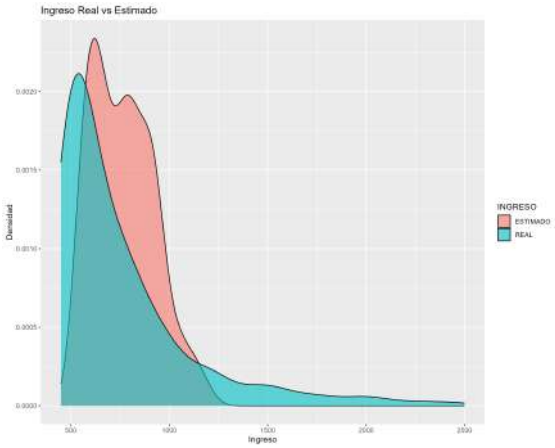


Figura 3.2: Sin remuestreo validación

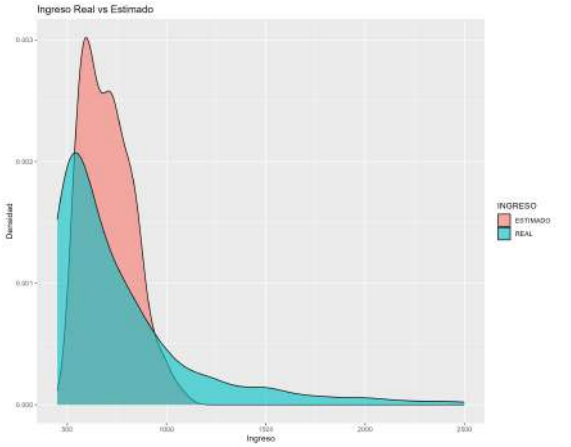


Figura 3.3: Con remuestreo entrenamiento

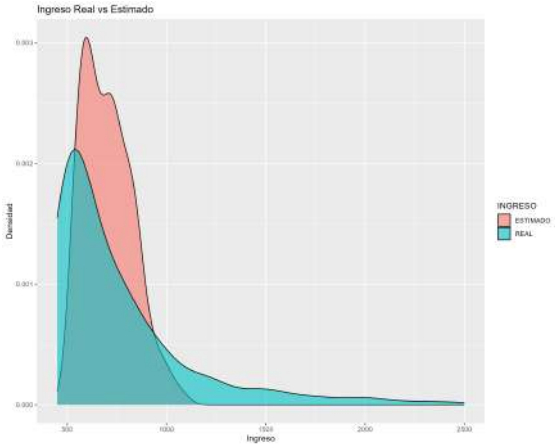


Figura 3.4: Con remuestreo validación

Figura 3.5: Distribuciones de Ingreso Real y Estimado del Modelo RF en BDD entrenamiento y validación sin remuestreo y con remuestreo para G1

3.6.3 Modelo GBM para G1

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos de entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G1_corte1 = 501; G1_corte2 = 1300   # 15%   # 80%
G1_porc1 = 0.45; G1_porc2 = 0.11

rmod_g1 <- boots_2tail(mod_g1, corte1 =  G1_corte1, corte2 = G1_corte2,
    porc1 = G1_porc1, porc2 = G1_porc2)
G1_min_nodo = presen_colas_g1[4]
rgbm_g1 <- gbm(formula = as.formula(formula_g1), data = rmod_g1, n.
   trees = 300, n.minobsinnode = G1_min_nodo, shrinkage = 0.03,
   distribution = "laplace")
runtime <- system.time({info[, INGRESO_EST_G1_5x100 := predict(rgbm_g1,
    n.trees = rgbm_g1$n.trees, info)]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[3])
info[, RANGO_REALG1 := cut(INGRESO_REAL, breaks = c(450, 500, 600, 750,
    950, 2500), labels = c("[450-500]", "(500-600]", "(600-750]", "
   (750-950]", "(950-250]"))]
info[, RANGO_EST_G1_5x100 := cut(INGRESO_EST_G1_5x100, breaks = c(450,
   500, 600, 750, 950, 2500), labels = c("[450-500]", "(500-600]", "
   (600-750]", "(750-950]", "(950-2500]"))]
```

Código 3.15: Estracto esencial de código para obtener los Resultados del modelo GBM para G1

Métricas de MSE y MAE:

En el modelo base el MSE es aproximadamente de 120000 unidades y el

MAE es aproximadamente de 215 unidades; en el modelo con remuestreo el MSE disminuye a 113000 unidades aproximadamente y el MAE aumenta a 227 unidades aproximadamente.

Estos resultados nos sugiere que el modelo es consistente controlando los errores en el modelo tanto en entrenamiento como en validación; en el modelo con remuestreo existe mejor MSE (tabla: 3.23 y 3.24).

Cruces y Sub/Sobre Estimaciones:

En el modelo base el cruce es aproximadamente de 72%, la sobre estimación es de 15% aproximadamente y la sub estimación es de 13% aproximadamente. En el modelo con remuestreo el cruce disminuye al 69% aproximadamente, la sobre estimación aumenta al 20% y la sub estimación disminuye al 10% aproximadamente (tabla: 3.23 y 3.24).

Nótese que en el modelo base no se tiene una matriz banda, la columna de ingresos estimado entre 450 a 550 y 950 a 2500 es muy mala; en el modelo con remuestreo se tiene predicciones sobre estas columnas, formando una mejor distribución en las predicciones, es por ello, que a pesar que la predicción disminuye, se controla la estimación en varios grupos de ingresos (figura 3.10).

Tiempos de ejecución:

En el modelo base, el tiempo de ejecución total es de 2.34 segundos aproximadamente; en el modelo con remuestreo es de 2.54 segundos aproximadamente; no existe un costo computacional significativo al entrenar el modelo con la base de datos remuestreada (tabla: 3.23 y 3.24).

Chi Cuadrado de Pearson:

Dado que el Chi cuadrado de Pearson el muy sensible al numero de individuos en la matriz de cruces, se apoya en las métricas para determinar que la distribución de entrenamiento y validación son similares (tabla: 3.23 y 3.24, figura: 3.10).

Decisión:

Con base a los criterios expuestos; se decide como modelo adecuado al modelo con remuestreo; dado que se tiene predicciones en las colas de la distribución, es decir, para sujetos con ingresos muy bajos o muy altos en el grupo 1.

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Gradient Boosting Machine
VARIABLES: 18 **G1**

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	721	4308	2590	906	0	8525.00	26%
(500-600]	317	2900	2463	1382	0	7062.00	22%
(600-750]	176	1815	2117	1442	0	5550.00	17%
(750-950]	76	1163	2303	1756	0	5298.00	16%
(950-250]	14	444	2275	3518	0	6251.00	19%
						32686.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	8%	51%	30%	11%	0%	24549.06
(500-600]	4%	41%	35%	20%	0%	17645.79
(600-750]	3%	33%	38%	26%	0%	13104.85
(750-950]	1%	22%	43%	33%	0%	32408.59
(950-250]	0%	7%	36%	56%	0%	549892.42

Entrenamiento: 0	Métricas		
MSE	122857.05		
MAE	215.40		
Cruce	23%	SubEstima	13%
Cruce +/-	72%	SobreEstim	15%

Muestra de Validación

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	744	4178	2760	948	0	8630.00	26%
(500-600]	308	2910	2378	1408	0	7004.00	21%
(600-750]	160	1877	2170	1452	0	5659.00	17%
(750-950]	67	1244	2192	1678	0	5181.00	16%
(950-250]	16	442	2220	3477	0	6155.00	19%
						32629.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	9%	48%	32%	11%	0%	25508.91
(500-600]	4%	42%	34%	20%	0%	17742.74
(600-750]	3%	33%	38%	26%	0%	13110.99
(750-950]	1%	24%	42%	32%	0%	33739.67
(950-250]	0%	7%	36%	56%	0%	532683.73

Validación: 1	Métricas		
MSE	118669.92		
MAE	213.32		
Cruce	23%	SubEstima	13%
Cruce +/-	72%	SobreEstim	16%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	χ^2	Teórico
[450-500]	0.7	3.9	11.2	1.9	0.0	**44.1**	**26.3**
(500-600]	0.3	0.0	2.9	0.5	0.0	**Decisión**	
(600-750]	1.5	2.1	1.3	0.1	0.0	Se rechaza H0	
(750-950]	1.1	5.6	5.3	3.5	0.0		
(950-250]	0.3	0.0	1.3	0.5	0.0		

Tiempo de Ejecución		
user	system	elapsed
1.95	0	2.34

Tabla 3.23: Resultados GBM sin remuestreo para G1

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G1		
MODELO: Gradient Boosting Machine	G1_corte1 = 484 #17%	G1_porc1 = 0.40
VARIABLES: 18	G1_corte2 = 1250 #91%	G1_porc2 = 0.30

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	2967	2140	1140	1894	384	8525	26%
(500-600]	1588	1718	927	2258	571	7062	22%
(600-750]	946	1219	690	2020	675	5550	17%
(750-950]	455	902	746	2335	860	5298	16%
(950-250]	98	419	551	3189	1994	6251	19%
						32686.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	35%	25%	13%	22%	5%	47700.96
(500-600]	22%	24%	13%	32%	8%	48637.35
(600-750]	17%	22%	12%	36%	12%	40059.16
(750-950]	9%	17%	14%	44%	16%	40834.49
(950-250]	2%	7%	9%	51%	32%	415221.85

Entrenamiento: 0	Métricas		
MSE	115778.89		
MAE	227.87		
Cruce	30%	SubEstima	10%
Cruce +/-	69%	SobreEstim	21%

Muestra de Validación

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	2921	2125	1153	2029	402	8630.00	26%
(500-600]	1577	1746	894	2220	567	7004.00	21%
(600-750]	930	1200	766	2085	678	5659.00	17%
(750-950]	530	923	695	2208	825	5181.00	16%
(950-250]	124	430	540	3121	1940	6155.00	19%
						32629.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	34%	25%	13%	24%	5%	49890.21
(500-600]	23%	25%	13%	32%	8%	48675.49
(600-750]	16%	21%	14%	37%	12%	39964.27
(750-950]	10%	18%	13%	43%	16%	42564.03
(950-250]	2%	7%	9%	51%	32%	402982.51

Validación: 1	Métricas		
MSE	113350.55		
MAE	227.06		
Cruce	29%	SubEstima	11%
Cruce +/-	68%	SobreEstim	22%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	χ^2	Teórico
[450-500]	0.7	0.1	0.1	9.6	0.8	**59.8**	**26.3**
(500-600]	0.1	0.5	1.2	0.6	0.0	**Decisión**	
(600-750]	0.3	0.3	8.4	2.1	0.0	Se rechaza H0	
(750-950]	12.4	0.5	3.5	6.9	1.4		
(950-250]	6.9	0.3	0.2	1.4	1.5		

Tiempo de Ejecución		
user	system	elapsed
1.97	0.01	2.54

Tabla 3.24: Resultados GBM con remuestreo para G1

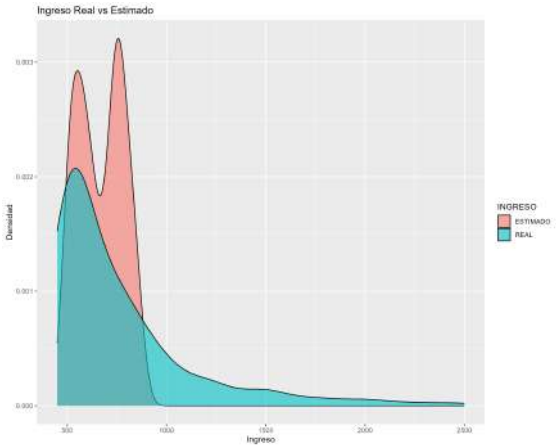


Figura 3.6: Sin remuestreo entrenamiento

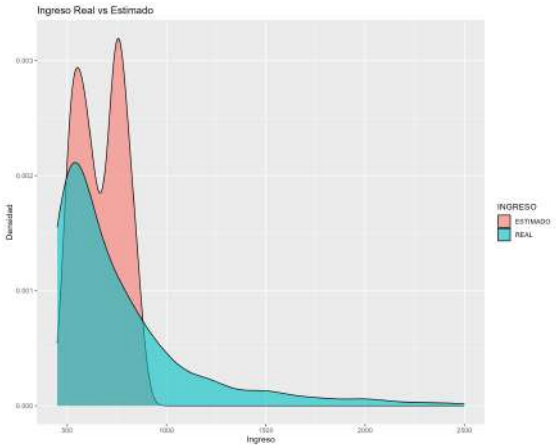


Figura 3.7: Sin remuestreo validación

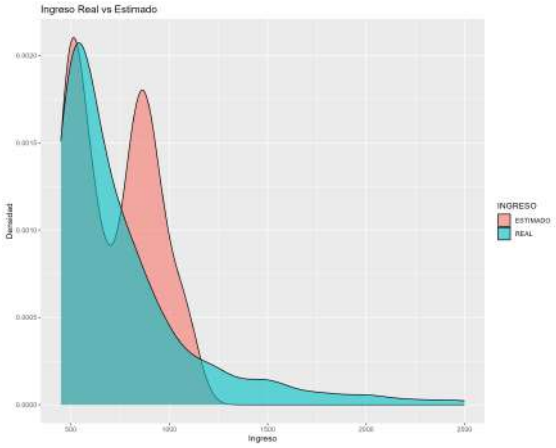


Figura 3.8: Con remuestreo entrenamiento

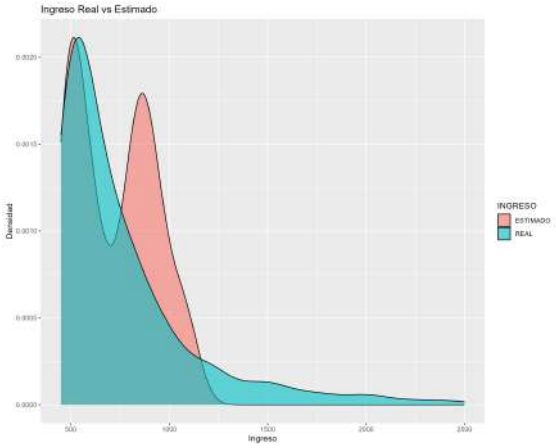


Figura 3.9: Con remuestreo validación

Figura 3.10: Distribuciones de Ingreso Real y Estimado del Modelo GBM en BDD entrenamiento y validación sin remuestreo y con remuestreo para G1

3.6.4 Modelo XGB para G1

La aplicación de la metodología XGBoost se lleva a cabo a través de la plataforma H2O, una herramienta altamente especializada en modelos de aprendizaje automático. H2O ofrece una variedad de modelos tanto supervisados como no supervisados. Optamos por utilizar H2O debido a la ineficiente implementación de XGBoost en R, que presenta problemas de rendimiento y consumo de recursos. XGBoost es conocido por su eficacia tanto en términos matemáticos como algorítmicos, por lo que nuestra elección fue probar este modelo en H2O, que se destaca por su eficiencia computacional en comparación con otros modelos como Random Forest (RF), Gradient Boosting Machine (GBM), Regresión Lineal Múltiple (RLM), entre otros.

Es importante señalar que H2O no se ejecuta directamente en R, sino que se utiliza a través del entorno RStudio. La ejecución del modelo se realiza mediante la activación de un clúster H2O, lo que nos permite tener un control absoluto sobre la potencia computacional, la memoria RAM, los núcleos de procesador y otros recursos que asignemos al proceso.

En este trabajo, desarrollamos una integración de H2O y R, donde transformamos las bases de datos de entrenamiento y prueba al formato requerido por H2O para entrenar el modelo. Una vez entrenado, el modelo realiza predicciones en el clúster y los resultados se pasan a un DataFrame en R para llevar a cabo cálculos adicionales como métricas y matriz de coincidencias.

El enfoque algorítmico y de programación es similar al empleado en Random Forest para el Grupo 1 (G1), con la única diferencia en las transformaciones necesarias para adaptar los datos a H2O y en el proceso de entrenamiento y predicción en el clúster. Estas peculiaridades se describen detalladamente aquí, mientras que para los otros grupos y modelos XGBoost (G2 y G3), ya no se describen en detalle.

Es importante destacar que XGBoost no es compatible con el sistema operativo Windows y funciona correctamente en sistemas operativos Mac y Linux. Para este proyecto, implementamos el modelo en Linux, lo que puede limitar su uso para usuarios menos familiarizados con sistemas operativos. Esta restricción fue una consideración significativa, ya que a

pesar de su eficacia, la implementación de XGBoost en H2O en sistemas operativos distintos a Linux puede ser un desafío.

3.6.5 XGB en H2O

1. **Activación del Cluster H2O**: Es necesario instalar el paquete H2O con todas sus dependencia y luego activarlo indicando la potencia computacional requerida.

```
install.packages("h2o",dependencies = TRUE)

library(h2o)
localH2O = h2o.init(ip = "localhost", nthreads = -1, max_mem_size
    = "32G")
h2o.getVersion()

h2o.removeAll()
```

Código 3.16: Activación del cluster H2O

2. **Transformación de la BDD a formato H2O**: Se recuerda que: en todo el entrenamiento continuamente estamos trabajando con 3 bases de datos. La base de datos de modelamiento, es decir la base de datos general con individuos que tienen cuota (obtenida a partir de la suma de las cuotas de amortizaciones de todos los créditos); la segunda es la base de datos de entrenamiento y la tercera es la base de datos de validación. Estos se debe transformar al tipo H2O.

```
info_names <- names(info)
info_em <- as.h2o(x = setDT(info)[, info_names, with=FALSE])

# rmod_g1 o mod_g1
rmod_g1_em <- as.h2o(x = setDT(mod_g1)[, formula_g1, with=FALSE])
    # Train
rval_g1_em <- as.h2o(x = setDT(val_g1)[, formula_g1, with=FALSE])
    # Test
```

Código 3.17: Transformación de las BDD al tipo H2O

$formula\_g1$ es un vector que contiene a las variables que ingresarán al modelo, más la variable a predecir.

3. **Entrenamiento del modelo**: El entrenamiento resulta ser muy similar a los ya expuesto para los otros modelos.

```
my_xgb_g1 <- h2o.xgboost(x = x_g1_em,
                         y = y_em,
                         model_id = "XGB",
                         training_frame = rmod_g1_em,
                         ntrees = 300,
                         learn_rate = 0.03,
                         max_depth = 6,
                         min_rows = g1_min_nodo,
                         nfolds = nfolds,
                         fold_assignment = "Auto",
                         keep_cross_validation_predictions = TRUE,
                         seed = 12345,
                         stopping_rounds = 50,
                         stopping_metric = "RMSE",
                         stopping_tolerance = 0)
```

Código 3.18: Entrenamiento del Modelo en H2O

4. **Predicción**: La predicción lo hacemos sobre el cluster y con BDD de H2O e inmediatamente se lo transforma a una dataframe y el resto de sentencias se lo hace con código R.

```
runtime <- system.time({
info[, INGRESO_EST_G1_5x100 := setDT(as.data.frame(h2o.predict(my_
    xgb_g1, newdata =info_em)))]
})

runtime_df <- data.frame(user = runtime[1],system = runtime[2],
    elapsed = runtime[3])
```

Código 3.19: Predicción sobre H2O

3.6.6 Resultados XGB

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado,

que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G1_corte1 = 475  #3%
G1_corte2 = 790  #67%
G1_porc1 = 0.39
G1_porc2 = 0.15

rmod_g1 <- boots_2tail(mod_g1, corte1 =  G1_corte1, corte2 = G1_corte2,
    porc1 = G1_porc1, porc2 = G1_porc2)

g1_min_nodo = presen_colas_g1[4]

#Transformacion de BDD Modelamiento, Train y Test a H2O
#entrenamiento del Modelo en H20

runtime <- system.time({
info[, INGRESO_EST_G1_5x100 := setDT(as.data.frame(h2o.predict(my_xgb_
   g1, newdata =info_em)))]
})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[1])

info[, RANGO_REALG1 := cut(INGRESO_REAL, breaks = c(450, 500, 600, 750,
    950, 2500), labels = c("[450-500]", "(500-600]", "(600-750]", "
   (750-950]", "(950-250]"))]
info[, RANGO_EST_G1_5x100 := cut(INGRESO_EST_G1_5x100, breaks = c(450,
   500, 600, 750, 950, 2500), labels = c("[450-500]", "(500-600]", "
   (600-750]", "(750-950]", "(950-2500]"))]
```

Código 3.20: Estracto esencial de código para obtener los resultados del modelo XGB para G1

Métricas de MSE y MAE:

En el modelo base, el MSE es aproximadamente de 104 mil unidades y el MAE es aproximadamente de 221 unidades; en el modelo con remuestreo

el MSE disminuye a 124 mil unidades aproximadamente y el MAE disminuye a 210 unidades.

El MSE por rango de ingreso real son similares en entrenamiento y validación tanto para el modelo base como para el remuestreado.

Esto nos sugiere que el modelo es consistente controlando los errores en las predicciones; y esta consistencia se mantiene en el modelo remuestreado (tabla: 3.25 y 3.26).

Cruces y Sub/Sobre Estimaciones:

En el modelo base el cruce es aproximadamente de 70%, la sobre estimación es aproximadamente de 27% y la sub estimación es aproximadamente de 3%; en el modelo remuestreado el cruce aumenta al 73%, la sobre estimación disminuye al 10% y la sub estimación aumenta al 17% (tabla: 3.25 y 3.26).

Al analizar la matriz de coincidencia se ve que se tiene algo de predicciones para individuos con ingresos estimado entre 40 a 550 USD.

Tiempos de ejecución:

El tiempo de ejecución en el modelo base es de 0.209 segundos aproximadamente; en el modelo remuestreado es de 0.30 aproximadamente.

El costo computacional de este modelo es formidable, no se tiene aumento de costos computacional significativo al remuestrear (tabla: 3.25 y 3.26).

Chi Cuadrado de Pearson:

Como es conocido, nos apoyamos de las métricas para determinar que la distribución en entrenamiento y validación son similares tanto en el modelo base como en el remuestreado (tabla: 3.25 y 3.26, figura: 3.15).

Decisión:

Con base a los criterios expuestos, se decide por el modelo remuestreado para el grupo 1, aun cuando no existe ganancia significativa.

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: XGBOOST
VARIABLES: 18 **G1**

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0	3310	2931	1961	323	8525.00	26%
(500-600]	0	1723	2411	2442	486	7062.00	22%
(600-750]	0	978	1739	2249	584	5550.00	17%
(750-950]	0	445	1356	2636	861	5298.00	16%
(950-2500]	0	75	600	3425	2151	6251.00	19%
						32686.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	39%	34%	23%	4%	52726.41
(500-600]	0%	24%	34%	35%	7%	45862.75
(600-750]	0%	18%	31%	41%	11%	29660.63
(750-950]	0%	8%	26%	50%	16%	26206.75
(950-2500]	0%	1%	10%	55%	34%	376124.52

Entrenamiento: 0	Métricas		
MSE	104876.39		
MAE	221.18		
Cruce	25%	SubEstima	3%
Cruce +/-	70%	SobreEstim	27%

Muestra de Validación

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0	3289	2880	2105	356	8630.00	26%
(500-600]	0	1708	2416	2406	474	7004.00	21%
(600-750]	0	970	1823	2289	577	5659.00	17%
(750-950]	0	527	1320	2546	788	5181.00	16%
(950-2500]	0	97	651	3354	2053	6155.00	19%
						32629.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	38%	33%	24%	4%	55393.94
(500-600]	0%	24%	34%	34%	7%	45367.23
(600-750]	0%	17%	32%	40%	10%	30325.89
(750-950]	0%	10%	25%	49%	15%	26795.90
(950-250]	0%	2%	11%	54%	33%	369489.17

Validación: 1	Métricas		
MSE	103602.67		
MAE	221.11		
Cruce	25%	SubEstima	4%
Cruce +/-	69%	SobreEstim	27%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	χ^2	Teórico
[450-500]	0.0	0.1	0.9	10.6	3.4	**62.9**	**26.3**
(500-600]	0.0	0.1	0.0	0.5	0.3	**Decisión**	
(600-750]	0.0	0.1	4.1	0.7	0.1	Se rechaza H0	
(750-950]	0.0	15.1	1.0	3.1	6.2		
(950-250]	0.0	6.5	4.3	1.5	4.5		

Tiempo de Ejecución		
user	system	elapsed
0.209	0	0.209

Tabla 3.25: Resultados XGB sin remuestreo para G1

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G1		
MODELO: XGBOOST	G1_corte1 = 475 #3%	G1_porc1 = 0.39
VARIABLES: 35	G1_corte2 = 790 #67%	G1_porc2 = 0.15

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-500]	328	5619	2164	352	62	8525.00	26%
(500-600]	131	3677	2640	533	81	7062.00	22%
(600-750]	70	2351	2401	642	86	5550.00	17%
(750-950]	32	1513	2724	866	163	5298.00	16%
(950-2500]	7	560	3320	1816	548	6251.00	19%
						32686.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-500]	4%	66%	25%	4%	1%	17075.92
(500-600]	2%	52%	37%	8%	1%	12617.40
(600-750]	1%	42%	43%	12%	2%	13993.89
(750-950]	1%	29%	51%	16%	3%	41517.06
(950-2500]	0%	9%	53%	29%	9%	564243.20

Entrenamiento: 0	Métricas		
MSE	124193.36		
MAE	212.24		
Cruce	24%	SubEstima	17%
Cruce +/-	73%	SobreEstim	10%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-500]	327	5518	2331	393	61	8630.00	26%
(500-600]	136	3655	2598	552	63	7004.00	21%
(600-750]	77	2447	2401	618	116	5659.00	17%
(750-950]	36	1570	2637	780	158	5181.00	16%
(950-2500]	5	625	3216	1771	538	6155.00	19%
						32629.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	4%	65%	27%	5%	1%	18348.30
(550-700]	2%	52%	37%	8%	1%	12344.51
(700-900]	1%	44%	43%	11%	2%	14414.72
(900-1300]	1%	30%	50%	15%	3%	42891.07
(1300-5000]	0%	10%	51%	28%	9%	550724.27

Validación: 1	Métricas		
MSE	120699.57		
MAE	210.42		
Cruce	24%	SubEstima	17%
Cruce +/-	72%	SobreEstim	11%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	χ^2	Teórico
[450-550]	0.0	1.8	12.9	4.8	0.0	**67.9**	**26.3**
(550-700]	0.2	0.1	0.7	0.7	4.0	**Decisión**	
(700-900]	0.7	3.9	0.0	0.9	10.5	Se rechaza H0	
(900-1300]	0.5	2.1	2.8	8.5	0.2		
(1300-5000]	0.6	7.5	3.3	1.1	0.2		

Tiempo de Ejecución		
user	system	elapsed
0.232	0.07	0.302

Tabla 3.26: Resultados XGB con remuestreo para G1

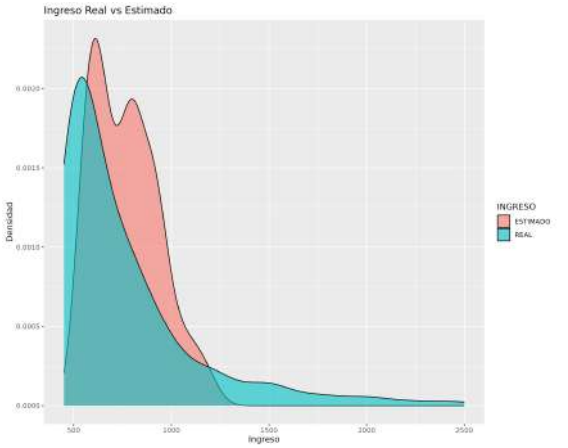


Figura 3.11: Sin remuestreo entrenamiento

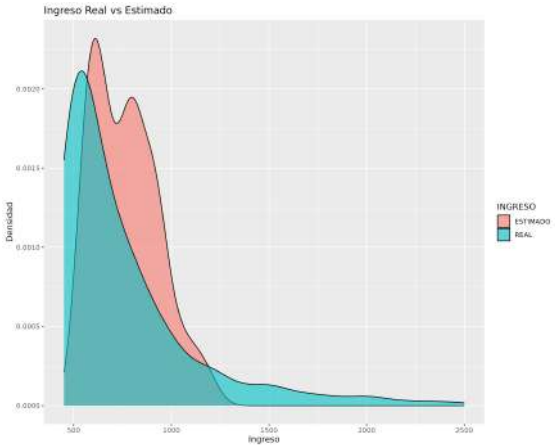


Figura 3.12: Sin remuestreo validación

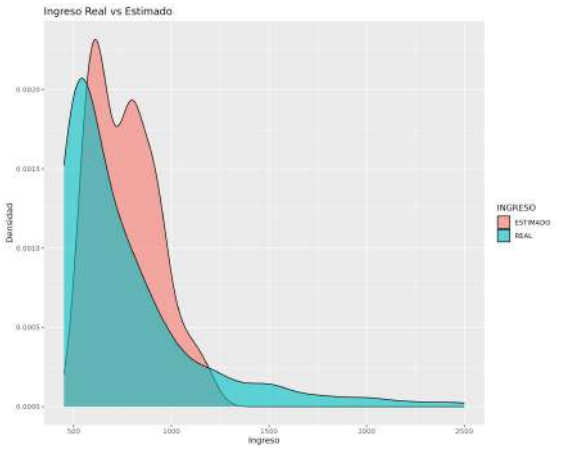


Figura 3.13: Con remuestreo entrenamiento

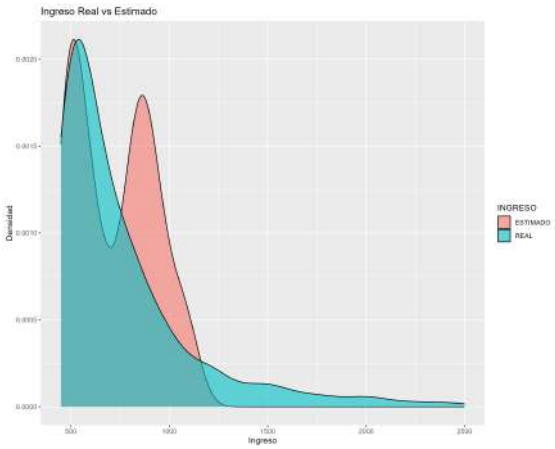


Figura 3.14: Con remuestreo validación

Figura 3.15: Distribuciones de Ingreso Real y Estimado del Modelo XGB en BDD entrenamiento y validación sin remuestreo y con remuestreo para G1

3.6.7 Modelo RLM para G1

El modelo de Regresión Lineal Múltiple (RLM) es un enfoque paramétrico que se basa en suposiciones específicas sobre la relación lineal entre las variables predictoras y la variable de respuesta. Para su validez, se deben cumplir una serie de hipótesis, como la linealidad, la independencia de los errores, la homocedasticidad, la normalidad de los errores y la ausencia de multicolinealidad. Sin embargo, en la práctica, estos supuestos son difíciles de verificar y a menudo pueden no cumplirse. Esto hace que el modelo RLM sea vulnerable a violaciones de estas hipótesis, lo que afecta su precisión y capacidad para manejar relaciones no lineales y datos complejos.

En contraste, los modelos no paramétricos como RF, GBM y XGB no imponen restricciones estrictas sobre la forma funcional de la relación entre las variables predictoras y la variable de respuesta. Estos modelos son capaces de capturar relaciones no lineales, interacciones complejas entre variables y patrones subyacentes en los datos de manera más efectiva. Además, son menos sensibles a las violaciones de las hipótesis de normalidad y homogeneidad, lo que los hace más robustos en situaciones del mundo real donde los datos pueden ser inherentemente ruidosos y no cumplir con todas las suposiciones.

Debido a la necesidad de realizar una comparación exhaustiva entre las predicciones de modelos paramétricos y no paramétricos, se procederá con la implementación de este modelo en particular sin verificar las hipótesis mencionadas. Verificarlas requeriría un trabajo independiente y sólido por sí mismo. Se realizan pruebas similares: el cálculo de métricas de rendimiento como el Error Cuadrático Medio (MSE) y el Error Absoluto Medio (MAE) en datos de entrenamiento y validación, tiempos de ejecución, etc.

Resultado RLM En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entre-

namiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G1_corte1 = 481  #5%
G1_corte2 = 950  #80%
G1_porc1 = 0.25
G1_porc2 = 0.18

rmod_g1 <- boots_2tail(mod_g1, corte1 =  G1_corte1, corte2 = G1_corte2,
    porc1 = G1_porc1, porc2 = G1_porc2)

rrlm_g1 <-lm(INGRESO_REAL ~ r_PROM_DEUDA_TOTAL_SICOM_OP_24s36M+
       r_DEUDA_TOTAL_SICOMsSCE_24M+
       DEUDA_TOTAL_SICOM_OP_24M+
       PROM_DEUDA_TOTAL_SICOM_OP_24M+
       r_PROM_XVEN_SICOM_OP_24s36M+
       PROM_XVEN_SICOM_OP_24M+
       r_DEUDA_TOTAL_SICOM_OP_12s24M+
       r_NOPE_APERT_SICOMsSCE_OP_36M+
       NOPE_APERT_SICOM_OP_36M+
       salTotOpCom037+
       salOpDiaCom005+
       NumAcreedoresDDCom404+
       numMesesInfoCredBanCoopD36M421+
       DEUDA_TOTAL_OP_OTROS+
       r_NOPE_APERT_SICOM_OP_24s36M+
       DEUDA_TOTAL_SBS_SC_24M+
       NOPE_TOTAL_OP_M,data = mod_g1)

runtime <- system.time({
info[, INGRESO_EST_G1_5x100 := predict(rrlm_g1, newdata = info)]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[3])

info[, RANGO_REALG1 := cut(INGRESO_REAL, breaks = c(450, 500, 600, 750,
    950, 2500), labels = c("[450-500]", "(500-600]", "(600-750]", "
   (750-950]", "(950-250]"))]
```

```
31
32 info[, RANGO_EST_G1_5x100 := cut(INGRESO_EST_G1_5x100, breaks = c(450,
       500, 600, 750, 950, 2500), labels = c("[450-500]", "(500-600]", "
       (600-750]", "(750-950]", "(950-2500]"))]
```

Código 3.21: Estracto esencial de código para obtener los Resultados del modelo RLM para G1

Métricas de MSE y MAE:

En el modelo base, el MSE es de 113 mil unidades y el MAE es de 232 unidades; en el modelo con remuestreo el MSE 116 mil unidades y el MAE disminuye a 222 unidades. (valores de comparación aproximados)

El MSE por rango de Ingreso Real son similares en entrenamiento y validación tanto para el modelo base como para el remuestreado.

Esto nos indica que el porcentaje de error en las predicciones se mantiene en entrenamiento y validación, y esto también se refleja en el modelo remuestreado (tabla: 3.27 y 3.28).

Cruces y Sub/Sobre Estimaciones:

En el modelo base, el cruce es de 64%, la sobre estimación de 32% y la sub estimación de 4%; en el modelo remuestreado el cruce aumenta al 71%, la sobre estimación disminuye al 23% y la sub estimación aumenta al 7%.

Al analizar la matriz de coincidencias se que que no existe mejoras con el modelo remuestreado, más aún no se tiene predicciones muy malas para individuos con ingresos bajo y altos a pesar de que las métricas de los errores estén controlados en proporción de error.

Ninguno de los dos modelos tiene buenas predicciones para los grupos de interés con ingresos bajos y altos (tabla: 3.27 y 3.28).

Tiempos de ejecución:

El tiempo de ejecución es 0.17 segundos para el modelo base y de 0.41 segundos para el modelo remuestreado; el costo computacional de estos modelos son relativamente bajos; pero tienen muy malas predicciones en los grupos de interés (tabla: 3.27 y 3.28).

Chi Cuadrado de Pearson:

Como se conoce, se apoya en la métricas para determinar que la distribución de entrenamiento y validación no son similares tanto en el modelo Base como en el remuestreado (gráficas 3.20). Las distribuciones son notablemente diferentes (tabla: 3.27 y 3.28, figura: 3.20).

Decisión:

Bajo los criterios expuestos, no se decide por ninguno de los dos modelos para describir el comportamiento de los ingresos para el grupo 1.

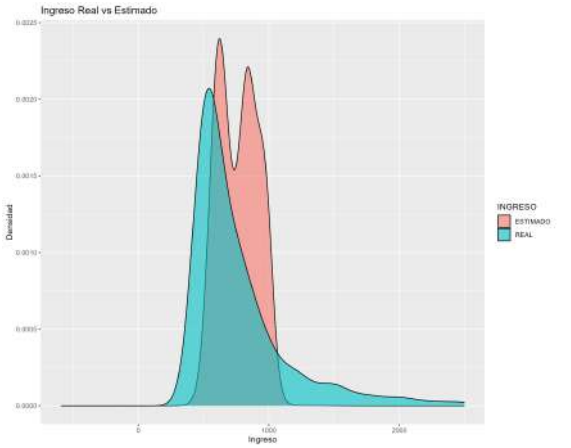


Figura 3.16: Sin remuestreo entrenamiento

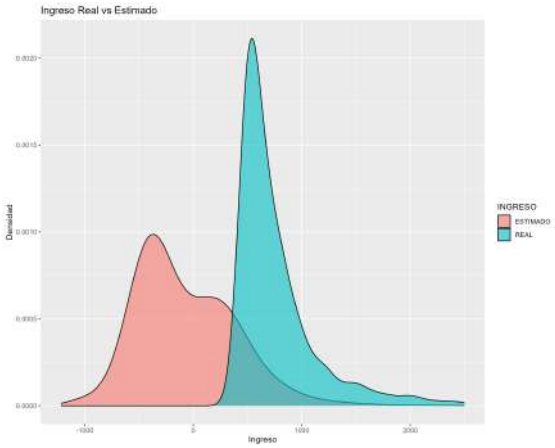


Figura 3.17: Sin remuestreo validación

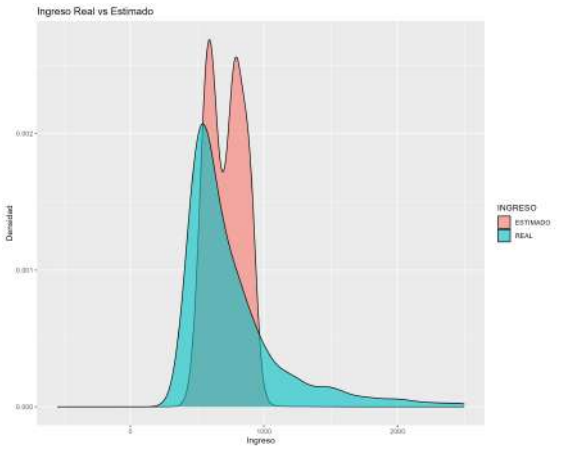


Figura 3.18: Con remuestreo entrenamiento

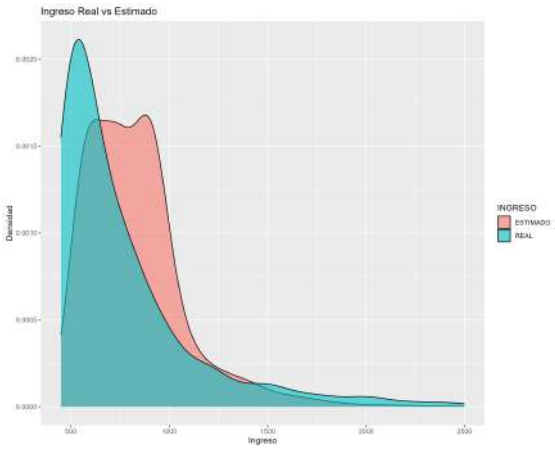


Figura 3.19: Con remuestreo validación

Figura 3.20: Distribuciones de Ingreso Real y Estimado del Modelo RLM en BDD entrenamiento y validación sin remuestreo y con remuestreo para G1

CONSIDERACIONES: Nodos finales al ∼∼∼% - Sin aplicación de remuestreo
MODELO: Regresión Lineal Múltiple
VARIABLES: 18 **G1**

Muestra de Modelamiento

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	64.00	2107.00	3718.00	2214.00	415.00	8518.00	26%
(500-600]	53.00	1342.00	2538.00	2499.00	617.00	7049.00	22%
(600-750]	37.00	812.00	1741.00	2224.00	725.00	5539.00	17%
(750-950]	38.00	451.00	1316.00	2571.00	918.00	5294.00	16%
(950-2500]	6.00	101.00	750.00	3340.00	2051.00	6248.00	19%
						32648.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	1%	25%	44%	26%	5%	59359.98
(500-600]	1%	19%	36%	35%	9%	47558.99
(600-750]	1%	15%	31%	40%	13%	28247.60
(750-950]	1%	9%	25%	49%	17%	22608.04
(950-2500]	0%	2%	12%	53%	33%	413815.96

Entrenamiento: 0	Métricas		
MSE	113358.03		
MAE	232.72		
Cruce	24%	SubEstima	4%
Cruce +/-	65%	SobreEstim	31%

Muestra de Validación

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	35	2125	3683	2321	454	8618.00	26%
(500-600]	51	1303	2538	2492	611	6995.00	21%
(600-750]	42	784	1821	2294	709	5650.00	17%
(750-950]	29	472	1327	2487	856	5171.00	16%
(950-2500]	5	125	763	3220	2039	6152.00	19%
						32586.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	25%	43%	27%	5%	61023.30
(500-600]	1%	19%	36%	36%	9%	47636.98
(600-750]	1%	14%	32%	41%	13%	58589.01
(750-950]	1%	9%	26%	48%	17%	23367.31
(950-2500]	0%	2%	12%	52%	33%	398333.64

Validación: 1	Métricas		
MSE	115377.28		
MAE	231.36		
Cruce	24%	SubEstima	4%
Cruce +/-	64%	SobreEstim	32%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	χ^2	Teórico
[450-500]	13.1	0.2	0.3	5.2	3.7	**52.2**	**26.3**
(500-600]	0.1	1.1	0.0	0.0	0.1	**Decisión**	
(600-750]	0.7	1.0	3.7	2.2	0.4	Se rechaza H0	
(750-950]	2.1	1.0	0.1	2.7	4.2		
(950-2500]	0.2	5.7	0.2	4.3	0.1		

Tiempo de Ejecución		
user	system	elapsed
0.12	0.05	0.17

Tabla 3.27: Resultados RLM sin remuestreo para G1

CONSIDERACIONES: Nodos finales al ~~~% - Con aplicación de remuestreo G1		
MODELO: Regresión Lineal Múltiple	G1_corte1 = 481 #5%	G1_porc1 = 0.25
VARIABLES: 18	G1_corte2 = 950 #80%	G1_porc2 = 0.18

Muestra de Modelamiento

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	161.00	3521.00	2812.00	1987.00	25.00	8506.00	26%
(500-600]	147.00	2183.00	2095.00	2595.00	22.00	7042.00	22%
(600-750]	90.00	1324.00	1594.00	2506.00	21.00	5535.00	17%
(750-950]	80.00	736.00	1381.00	3055.00	37.00	5289.00	16%
(950-2500]	10.00	226.00	1005.00	4890.00	116.00	6247.00	19%
						32619.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	2%	41%	33%	23%	0%	40208.20
(500-600]	2%	31%	30%	37%	0%	30002.48
(600-750]	2%	24%	29%	45%	0%	17993.92
(750-950]	2%	14%	26%	58%	1%	25300.71
(950-2500]	0%	4%	16%	78%	2%	481953.79

Entrenamiento: 0	Métricas		
MSE	116296.12		
MAE	222.10		
Cruce	22%	SubEstima	7%
Cruce +/-	71%	SobreEstim	23%

Muestra de Validación

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	140	3528	2774	2142	27	8611.00	26%
(500-600]	144	2141	2102	2580	18	6985.00	21%
(600-750]	102	1314	1624	2580	24	5644.00	17%
(750-950]	64	794	1378	2894	33	5163.00	16%
(950-2500]	14	232	987	4790	131	6154.00	19%
						32557.00	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	2%	41%	32%	25%	0%	41462.08
(500-600]	2%	31%	30%	37%	0%	30079.79
(600-750]	2%	23%	29%	46%	0%	50357.15
(750-950]	1%	15%	27%	56%	1%	26505.83
(950-2500]	0%	4%	16%	78%	2%	464778.72

Validación: 1	Métricas		
MSE	118039.39		
MAE	220.30		
Cruce	21%	SubEstima	7%
Cruce +/-	70%	SobreEstim	23%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	χ^2	Teórico
[450-500]	2.7	0.0	0.5	12.1	0.2	**44.8**	**26.3**
(500-600]	0.1	0.8	0.0	0.1	0.7	**Decisión**	
(600-750]	1.6	0.1	0.6	2.2	0.4	Se rechaza H0	
(750-950]	3.2	4.6	0.0	8.5	0.4		
(950-2500]	1.6	0.2	0.3	2.0	1.9		

Tiempo de Ejecución		
user	system	elapsed
0.33	0.08	0.41

Tabla 3.28: Resultados RLM con remuestreo para G1

3.6.8 Elección del mejor modelo entre RLM, RF, GBM y XGB para G1

De los resultados expuestos, se elige como mejor modelo al Gradient Boosting Machine (BGM); dado que el costo computacional es relativamente bajo, mejora las predicciones para individuos con ingresos muy bajos y muy altos; en el proceso de las simulaciones y/o remuestreos lograr esa suerte de matriz banda llevó menos tiempo; es decir, el proceso de entrenamiento fue más rápido.

Ciertamente, el modelo más eficiente tanto matemáticamente como computacionalmente es el XGB; pero este modelo tiene la limitante de que no está disponible para el sistema operativo Windows, este detallé causo muchas dificultades para su implementación; es por ello que no se decide por este modelo y se inclina por el GBM.

3.7 Modelos para población G2

3.7.1 Exploración y descripción de la base de datos

Esta población cuenta con sujetos con cuota estimada actual mayor a 107 dólares y menor o igual a 435 dólares.

Razonando de forma similar al grupo 1, dado que tenemos más de 4 mil registros optamos por tomar la mitad para entrenamiento y la otra mitad para validación (tabla 3.29).

Muestra	N	%
Entrenamiento	69,006	50.06%
Validación	68,835	49.94%
Total	**137,841**	**100.00%**

Tabla 3.29: Población del Grupo 2

Distribución del Ingreso Real del Grupo 2

Percentil	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Valor en USD	450.03	483	493	500	541	589	600	627	675	720	760

Percentil		55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
Valor en USD		800	873	930	1,008	1,152	1,300	1,515	1,880	2,500	5,000

Tabla 3.30: Percentil del Ingreso Real del Grupo 2

La tabla 3.30 de percentil será de mucha ayuda para el entrenamiento del modelo, dado que ayuda a determinar los cortes y proporciones adecuados para los parámetros del remuestreo.

3.7.2 Modelo RF para G2

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G2_corte1 = 501; G2_corte2 = 1300   # 15%   # 80%
G2_porc1 = 0.45; G2_porc2 = 0.11

rmod_g2 <- boots_2tail(mod_g2, corte1 =  G2_corte1, corte2 = G2_corte2,
    porc1 = G2_porc1, porc2 = G2_porc2)
G2_min_nodo = presen_colas_g2[4]
rf_g2_5x100 <- ranger(formula = as.formula(formula_g2), data = rmod_g2,
    num.trees = 300, mtry = 11, min.node.size = G2_min_nodo,
   importance = 'impurity', write.forest = TRUE, seed = 1234)
runtime <- system.time({info[, INGRESO_EST_G2_5x100 := predict(object=
   rf_g2_5x100, data=info)$predictions]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[3])
info[, RANGO_REALG2 := cut(INGRESO_REAL, breaks = c(450, 550, 700, 900,
    1300, 5000), labels = c("[450-550]", "(550-700]", "(700-900]", "
   (900-1300]", "(1300-5000]"))]
info[, RANGO_EST_G2_5x100 := cut(INGRESO_EST_G2_5x100, breaks = c(450,
   550, 700, 900, 1300, 5000), labels = c("[450-550]", "(550-700]", "
   (700-900]", "(900-1300]", "(1300-5000]"))]
```

Código 3.22: Estracto esencial de código para obtener los Resultados del modelo RF para G2

Métricas de MSE y MAE:

En el modelo Base, frisa al rededor de los 390000 unidades tanto en en-

trenamiento y validación; el MAE está aproximadamente en 400 unidades. Por otro lado, en el modelo con remuestreo, el MSE aumenta a 425000 unidades aproximadamente y el MAE disminuye a 370 unidades aproximadamente. Estos resultados nos sugieren que el modelo es consistente en las predicciones para entrenamiento y validación (tanto en el modelo base y remuestreada), pues los valores de error son similares; en el caso del modelo con remuestreo la disminución del MAE indica que existe una mejora en la precisión de las predicciones.

Estos resultados confirman la coherencia intrínseca del modelo en sus predicciones tanto en el conjunto de entrenamiento como en el de validación. Adicionalmente, esta estabilidad se conserva al entrenar el modelo con la base de datos de remuestreo. La convergencia en las métricas de evaluación respalda la capacidad del modelo para proporcionar estimaciones precisas y consistentes del ingreso real (tabla: 3.31 y 3.32).

Cruces y Sub/Sobre Estimaciones:

En el modelo base, el porcentaje de cruce en entrenamiento y validación es aproximadamente del 71%; la sobre estimación es aproximadamente del 26% y la sub estimación aproximadamente del 2%. Vemos que el modelo es consistente en estas métricas al tener valores similares. En el modelo con remuestreo, Se mantiene esta consistencia entre entrenamiento y validación; más aún, se mejora el porcentaje del cruce al 79% aproximadamente, se reduce la sobre estimación al 14% aproximadamente y la sub estimación se incrementa al 8%.

Estos resultados indica que el modelo es consistente en los porcentajes de predicción: acierto, sobre estimación y sub estimación. Y se tiene una mejora en los porcentajes de predicción con la base remuestreada (tabla: 3.31 y 3.32).

Tiempos de ejecución:

En el modelo base, el tiempo de ejecución total es aproximadamente de 15.9 segundos, mientras que en el modelo con remuestreo es de 17.08 segundos; este incremento en el tiempo de ejecución no es significativa dado que tenemos precisión en las predicciones (tabla: 3.31 y 3.32).

Chi Cuadrado de Pearson:

Dado que el Chi cuadrado de Pearson es muy sensible a pequeños cambios se apoya en las métricas MSE y MAE que dice que el modelo conserva el porcentaje de errores y también el porcentaje de cruces es aceptable; por tanto no se rechaza la hipótesis nula: las distribuciones de ingreso en entrenamiento y validación son similares (tabla: 3.31 y 3.32, figura: 3.25).

Decisión:

Basándonos en los criterios presentados y las comparaciones detalladas, se propone que el modelo con remuestreo sea considerado como el modelo más apropiado para futuros análisis y cálculos.

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Random Forest
VARIABLES: 35 **G2**

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	0	5811	4859	3893	563	15126.00	22%
(550-700]	0	3998	4417	5850	1098	15363.00	22%
(700-900]	0	2136	3442	6258	1779	13615.00	20%
(900-1300]	0	357	1752	6210	2937	11256.00	16%
(1300-5000]	0	95	1054	6479	6018	13646.00	20%
						69006.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	0%	38%	32%	26%	4%	150229.68
(550-700]	0%	26%	29%	38%	7%	138962.73
(700-900]	0%	16%	25%	46%	13%	104553.39
(900-1300]	0%	3%	16%	55%	26%	76808.71
(1300-5000]	0%	1%	8%	47%	44%	1482024.86

Entrenamiento: 0	Métricas		
MSE	390096.8		
MAE	400.96		
Cruce	29%	SubEstima	2%
Cruce +/-	72%	SobreEstim	26%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	0	5672	4836	3825	550	14883.00	22%
(550-700]	0	3921	4366	6065	1092	15444.00	22%
(700-900]	0	2190	3371	6189	1838	13588.00	20%
(900-1300]	0	378	1729	6119	3023	11249.00	16%
(1300-5000]	0	118	1208	6514	5831	13671.00	20%
						68835.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	0%	38%	32%	26%	4%	150224.35
(550-700]	0%	25%	28%	39%	7%	141347.61
(700-900]	0%	16%	25%	46%	14%	107488.54
(900-1300]	0%	3%	15%	54%	27%	80761.72
(1300-5000]	0%	1%	9%	48%	43%	1491991.99

Validación: 1	Métricas		
MSE	394927.4		
MAE	405.38		
Cruce	28%	SubEstima	2%
Cruce +/-	71%	SobreEstim	26%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	χ^2	Teórico
[450-550]	0.0	3.3	0.1	1.2	0.3	**59.9**	**26.3**
(550-700]	0.0	1.5	0.6	7.9	0.0	**Decisión**	
(700-900]	0.0	1.4	1.5	0.8	2.0	Se rechaza H0	
(900-1300]	0.0	1.2	0.3	1.3	2.5		
(1300-5000]	0.0	5.6	22.5	0.2	5.8		

Tiempo de Ejecución		
user	system	elapsed
15.29	0.61	15.9

Tabla 3.31: Resultados RF sin remuestreo para G2

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G2		
MODELO: Random Forest	G2_corte1 = 501 # 15%	G2_porc1 = 0.45
VARIABLES: 35	G2_corte2 = 1300 # 80%	G2_porc2 = 0.11

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	1905	7714	3688	1705	114	15126.00	22%
(550-700]	907	6175	4840	3206	235	15363.00	22%
(700-900]	294	4285	4235	4496	305	13615.00	20%
(900-1300]	23	1369	3383	5900	581	11256.00	16%
(1300-5000]	6	636	3086	7858	2060	13646.00	20%
						69006.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	13%	51%	24%	11%	1%	66497.65
(550-700]	6%	40%	32%	21%	2%	57505.46
(700-900]	2%	31%	31%	33%	2%	46878.09
(900-1300]	0%	12%	30%	52%	5%	70292.72
(1300-5000]	0%	5%	23%	58%	15%	1905668.41

Entrenamiento: 0	Métricas		
MSE	424941.5		
MAE	370.55		
Cruce	29%	SubEstima	8%
Cruce +/-	79%	SobreEstim	13%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	1866	7505	3732	1652	128	14883.00	22%
(550-700]	860	6238	4736	3380	230	15444.00	22%
(700-900]	384	4224	4089	4570	321	13588.00	20%
(900-1300]	45	1382	3349	5813	660	11249.00	16%
(1300-5000]	13	803	3193	7658	2004	13671.00	20%
						68835.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	13%	50%	25%	11%	1%	67886.27
(550-700]	6%	40%	31%	22%	1%	58372.46
(700-900]	3%	31%	30%	34%	2%	48432.36
(900-1300]	0%	12%	30%	52%	6%	71918.97
(1300-5000]	0%	6%	23%	56%	15%	1908361.95

Validación: 1	Métricas		
MSE	428098.9		
MAE	374.42		
Cruce	29%	SubEstima	8%
Cruce +/-	78%	SobreEstim	14%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	χ^2	Teórico
[450-550]	0.8	5.7	0.5	1.6	1.7	**156.6**	**26.3**
(550-700]	2.4	0.6	2.2	9.4	0.1	**Decisión**	
(700-900]	27.6	0.9	5.0	1.2	0.8	Se rechaza H0	
(900-1300]	21.0	0.1	0.3	1.3	10.7		
(1300-5000]	8.2	43.9	3.7	5.1	1.5		

Tiempo de Ejecución		
user	system	elapsed
16.83	0.25	17.80

Tabla 3.32: Resultados RF con remuestreo para G2

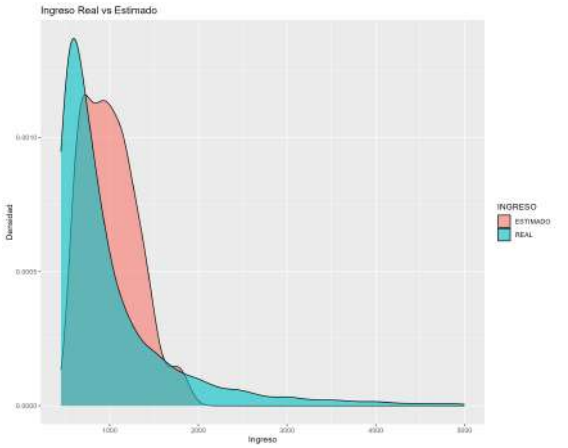


Figura 3.21: Sin remuestreo entrenamiento

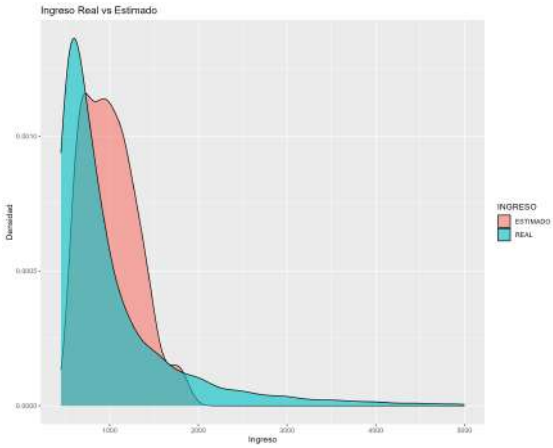


Figura 3.22: Sin remuestreo validación

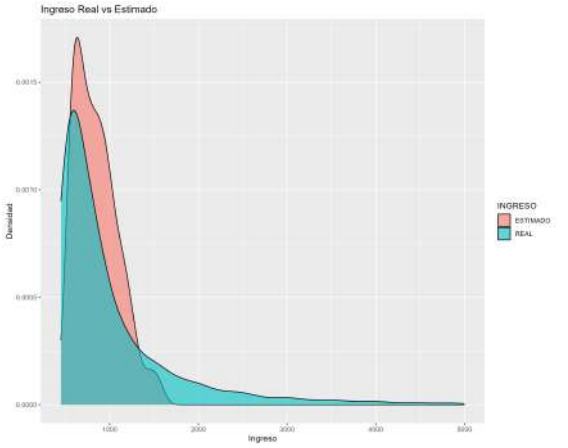


Figura 3.23: Con remuestreo entrenamiento

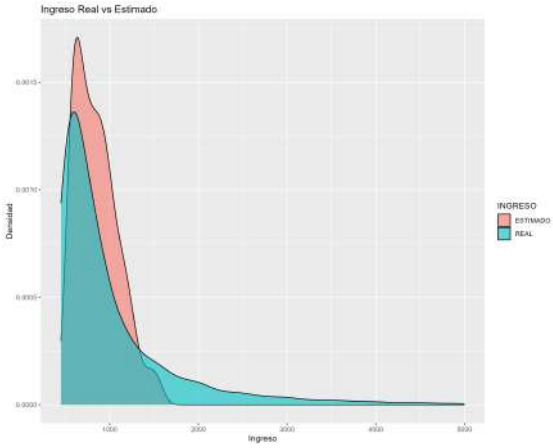


Figura 3.24: Con remuestreo validación

Figura 3.25: Distribuciones de Ingreso Real y Estimado del Modelo RF en BDD entrenamiento y validación sin remuestreo y con remuestreo para G2

3.7.3 Modelo GBM para G2

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G2_corte1 = 501; G2_corte2 = 1300   # 15%   # 80%
G2_porc1 = 0.45; G2_porc2 = 0.11

rmod_g2 <- boots_2tail(mod_g2, corte1 =  G2_corte1, corte2 = G2_corte2,
    porc1 = G2_porc1, porc2 = G2_porc2)
G2_min_nodo = presen_colas_g2[4]
rgbm_g2 <- gbm(formula = as.formula(formula_g2), data = rmod_g2, n.
   trees = 300, n.minobsinnode = G2_min_nodo, shrinkage = 0.03,
   distribution = "laplace")
runtime <- system.time({info[, INGRESO_EST_G2_5x100 := predict(rgbm_g2,
    n.trees = rgbm_g2$n.trees, info)]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[3])
info[, RANGO_REALG2 := cut(INGRESO_REAL, breaks = c(450, 550, 700, 900,
    1300, 5000), labels = c("[450-550]", "(550-700]", "(700-900]", "
   (900-1300]", "(1300-5000]"))]
info[, RANGO_EST_G2_5x100 := cut(INGRESO_EST_G2_5x100, breaks = c(450,
   550, 700, 900, 1300, 5000), labels = c("[450-550]", "(550-700]", "
   (700-900]", "(900-1300]", "(1300-5000]"))]
```

Código 3.23: Estracto esencial de código para obtener los Resultados del modelo GBM para G2

Métricas de MSE y MAE:

En el modelo base, el MSE es aproximadamente de 465000 unidades y el

MAE es aproximadamente de 382 unidades; en el modelo con remuestreo el MSE reduce a 433000 unidades aproximadamente y el MAE aumenta a 296 unidades aproximadamente.

El MSE por rango de Ingreso Real son también consistentes, pues son muy similares en entrenamiento y validación tanto en el modelo base como en el modelo con remuestreo.

Estos resultados no sugieren que el modelo es consistente controlando los errores en entrenamiento y validación; esta consistencia se conserva en el modelo con remuestreo; más aún existe mejora en el MSE y el MAE no crece significativamente (tabla: 3.33 y 3.34).

Cruces y Sub/Sobre Estimaciones:

En el modelo base, el cruce es aproximadamente del 75%, la sobre estimación es de 14% aproximadamente y la sub estimación es de 10% aproximadamente. En el modelo con remuestreo el cruce disminuye al 72%, la sobre estimación aumenta al 18% y la sub estimación se mantiene al 10% aproximadamente.

Nótese que en el modelo con remuestreo mejora las predicciones en las colas, es decir en la columna 1 y 5 de la matriz de coincidencias, pues en la columna de ingresos estimados de 1300 a 5000 USD no se tiene estimación, pero con el remuestreo si se logra capturar predicciones para este intervalo (tabla: 3.33 y 3.34).

Tiempos de ejecución:

En el modelo base, el tiempo de ejecución es aproximadamente de 1.4 segundos, en el modelo con remuestreo es de 2.02 aproximadamente; no se tiene un incremento significativo en el costo computación al modelar con remuestreo de las bases, y se gana predicción en las colas de la distribución de ingresos (tabla: 3.33 y 3.34).

Chi Cuadrado de Pearson:

Dado que el chi cuadrado es sensible al numero de individuos en los cruces, se apoya en las métricas para decidir si la distribución de ingresos en entrenamiento y validación son similares; así, no se rechaza la hipótesis nula (tabla: 3.33 y 3.34, figura: 3.30).

Decisión:

Con base a los criterios expuestos, se decide como mejor modelo al modelo con remuestreo para el grupo 2 dado que se tiene predicciones para los sujetos con ingresos muy bajos o muy altos.

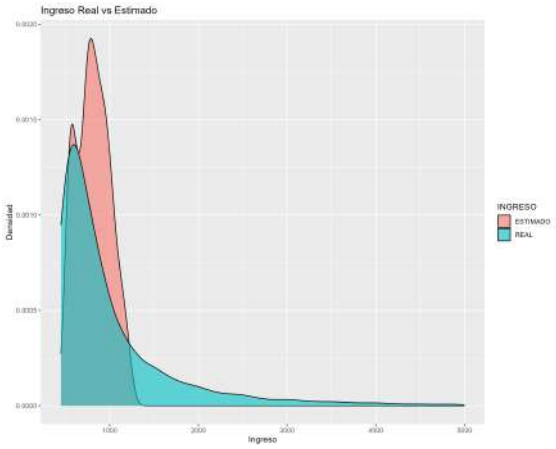


Figura 3.26: Sin remuestreo entrenamiento

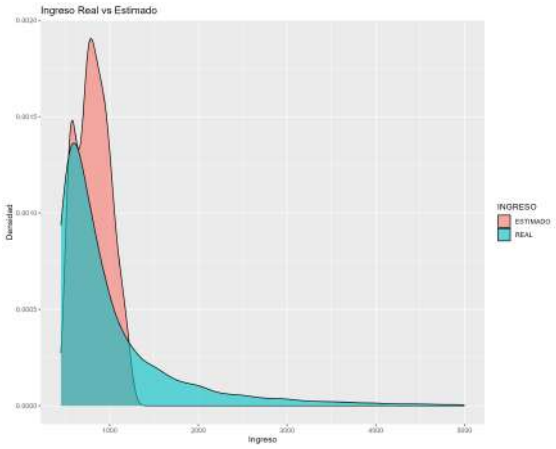


Figura 3.27: Sin remuestreo validación

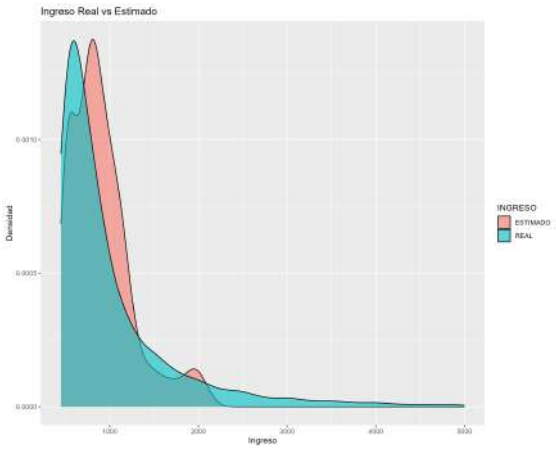


Figura 3.28: Con remuestreo entrenamiento

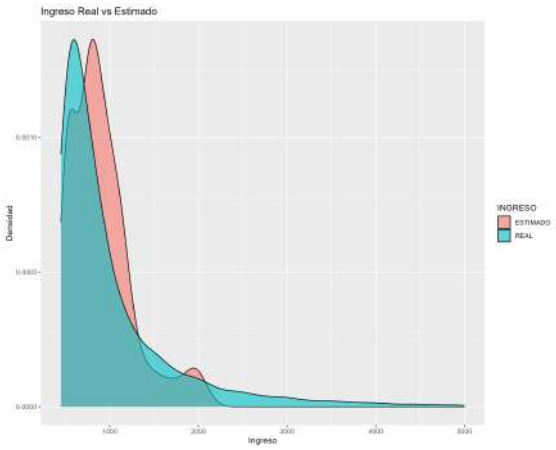


Figura 3.29: Con remuestreo validación

Figura 3.30: Distribuciones de Ingreso Real y Estimado del Modelo GBM en BDD entrenamiento y validación sin remuestreo y con remuestreo para G2

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Gradient Boosting Machine
VARIABLES: 35 **G2**

Muestra de Modelamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	2885	5175	5438	1628	0	15126.00	22%
(550-700]	1663	4403	6379	2918	0	15363.00	22%
(700-900]	757	2929	5831	4098	0	13615.00	20%
(900-1300]	144	1004	4527	5581	0	11256.00	16%
(1300-5000]	47	563	4344	8692	0	13646.00	20%
						69006.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	19%	34%	36%	11%	0%	62468.47
(550-700]	11%	29%	42%	19%	0%	43118.79
(700-900]	6%	22%	43%	30%	0%	30606.60
(900-1300]	1%	9%	40%	50%	0%	70176.44
(1300-5000]	0%	4%	32%	64%	0%	2145314.07

Entrenamiento: 0	Métricas		
MSE	465016.12		
MAE	382.27		
Cruce	27%	SubEstima	10%
Cruce +/-	76%	SobreEstim	14%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	2922	5050	5362	1549	0	14883.00	22%
(550-700]	1625	4443	6447	2929	0	15444.00	22%
(700-900]	821	2960	5614	4193	0	13588.00	20%
(900-1300]	157	1010	4446	5636	0	11249.00	16%
(1300-5000]	67	578	4452	8574	0	13671.00	20%
						68835.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	20%	34%	36%	10%	0%	61906.84
(550-700]	11%	29%	42%	19%	0%	43179.73
(700-900]	6%	22%	41%	31%	0%	31276.85
(900-1300]	1%	9%	40%	50%	0%	70182.28
(1300-5000]	0%	4%	33%	63%	0%	2123519.86

Validación: 1	Métricas		
MSE	462458.6		
MAE	383.32		
Cruce	27%	SubEstima	10%
Cruce +/-	75%	SobreEstim	14%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	χ^2	Teórico
[450-550]	0.5	3.0	1.1	3.8	0.0	**42.8**	**26.3**
(550-700]	0.9	0.4	0.7	0.0	0.0	**Decisión**	
(700-900]	5.4	0.3	8.1	2.2	0.0	Se rechaza H0	
(900-1300]	1.2	0.0	1.4	0.5	0.0		
(1300-5000]	8.5	0.4	2.7	1.6	0.0		

Tiempo de Ejecución		
user	system	elapsed
1.06	0	1.4

Tabla 3.33: Resultados GBM sin remuestreo para G2

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G2		
MODELO: Gradient Boosting Machine	G2_corte1 = 601 # 33%	G2_porc1 = 0.61
VARIABLES: 35	G2_corte2 = 1871 # 90%	G2_porc2 = 0.35

Muestra de Modelamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	5121	2768	4622	2130	485	15126.00	22%
(550-700]	3474	2486	5199	3433	771	15363.00	22%
(700-900]	1857	1734	4577	4389	1058	13615.00	20%
(900-1300]	433	713	3378	5083	1649	11256.00	16%
(1300-5000]	193	420	3102	6308	3623	13646.00	20%
						69006.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	34%	18%	31%	14%	3%	113404.92
(550-700]	23%	16%	34%	22%	5%	110094.60
(700-900]	14%	13%	34%	32%	8%	102199.35
(900-1300]	4%	6%	30%	45%	15%	120617.82
(1300-5000]	1%	3%	23%	46%	27%	1743407.38

Entrenamiento: 0	Métricas		
MSE	433968.08		
MAE	396.36		
Cruce	30%	SubEstima	10%
Cruce +/-	72%	SobreEstim	18%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	5113	2691	4575	2022	482	14883.00	22%
(550-700]	3403	2515	5228	3559	739	15444.00	22%
(700-900]	1902	1784	4429	4398	1075	13588.00	20%
(900-1300]	438	723	3335	5023	1730	11249.00	16%
(1300-5000]	208	448	3128	6370	3517	13671.00	20%
						68835.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	34%	18%	30%	13%	3%	112247.00
(550-700]	22%	16%	34%	23%	5%	109896.38
(700-900]	14%	13%	33%	32%	8%	104232.36
(900-1300]	4%	6%	30%	45%	15%	126375.03
(1300-5000]	2%	3%	23%	47%	26%	1725852.81

Validación: 1	Métricas		
MSE	432917.08		
MAE	398.69		
Cruce	30%	SubEstima	10%
Cruce +/-	72%	SobreEstim	18%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	χ^2	Teórico
[450-550]	0.0	2.1	0.5	5.5	0.0	**36.0**	**26.3**
(550-700]	1.5	0.3	0.2	4.6	1.3	**Decisión**	
(700-900]	1.1	1.4	4.8	0.0	0.3	Se rechaza H0	
(900-1300]	0.1	0.1	0.5	0.7	4.0		
(1300-5000]	1.2	1.9	0.2	0.6	3.1		

Tiempo de Ejecución		
user	system	elapsed
1.48	0	2.02

Tabla 3.34: Resultados GBM con remuestreo para G2

3.7.4 Modelo XGB para G2

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G2_corte1 = 499   #10%
G2_corte2 = 1000  #68%
G2_porc1 = 0.35
G2_porc2 = 0.15
rmod_g2 <- boots_2tail(mod_g2, corte1 =  G2_corte1, corte2 = G2_corte2,
    porc1 = G2_porc1, porc2 = G2_porc2)
g2_min_nodo = presen_colas_g2[4]
#Transformacion de BDD Modelamiento, Train y Test a H2O
#entrenamiento del Modelo en H20
runtime <- system.time({
info[, INGRESO_EST_G2_5x100 := setDT(as.data.frame(h2o.predict(my_xgb_
   g2, newdata =info_em)))] })
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[1])
info[, RANGO_REALG2 := cut(INGRESO_REAL, breaks = c(450, 550, 700, 900,
    1300, 5000), labels = c("[450-550]", "(550-700]", "(700-900]", "
   (900-1300]", "(1300-5000]"))]
info[, RANGO_EST_G2_5x100 := cut(INGRESO_EST_G2_5x100, breaks = c(450,
   550, 700, 900, 1300, 5000), labels = c("[450-550]", "(550-700]", "
   (700-900]", "(900-1300]", "(1300-5000]"))]
```

Código 3.24: Estracto esencial de código para obtener los Resultados del modelo XGB para G2

Métricas de MSE y MAE:

En el modelo base, el MSE es aproximadamente de 385 mil unidades y el MAE de 396 unidades aproximadamente; en el modelo remuestreado el MSE a 424 mil unidades aproximadamente y el MAE disminuye a 370 unidades.

El MSE por rango de Ingreso Real son similares para entrenamiento y validación tanto en el modelo base como en el remuestreado.

Esto nos dice que el modelo controla adecuadamente los errores de predicción, y esta consistencia se mantiene para el modelo remuestreado (tabla: 3.35 y 3.36).

Cruces y Sub/Sobre Estimaciones:

En el modelo base, el cruce es de 72%, la sobre estimación es 25% y la sub estimación de 3%; en el modelo con remuestreo el cruce al 78%, la sobre estimación disminuye al 12% y la sub estimación aumenta al 10%; todos estos valores son aproximados.

Esto nos sugiere que el modelo el consistente en la predicciones: acierto, sobre y sub estimación en entrenamiento y validación; esta consistencia se conserva en el modelo remuestreado (tabla: 3.35 y 3.36).

Tiempos de ejecución:

En el modelo base el tiempo de ejecución es de 0.22 segundos aproximadamente, en el modelo remuestreado es de 0.27 aproximadamente; se nota que no existe aumento significativo en el costo computacional al remuestrear (tabla: 3.35 y 3.36).

Chi Cuadrado de Pearson:

Como es conocido, se apoya en la métricas para determinar que la distribución de entrenamiento y validación son similares (tabla: 3.35 y 3.36, figura: 3.35).

Decisión:

Con base a los criterios expuestos, se decide por el modelo remuestreado para el grupo 2.

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: XGBOOST
VARIABLES: 35 **G2**

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	413	5816	4868	3436	593	15126.00	22%
(550-700]	164	4105	4577	5372	1144	15362.00	22%
(700-900]	73	2284	3565	5786	1907	13615.00	20%
(900-1300]	4	377	1963	5822	3090	11256.00	16%
(1300-5000]	0	117	1185	5924	6420	13646.00	20%
						69005.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	3%	38%	32%	23%	4%	144037.95
(550-700]	1%	27%	30%	35%	7%	138605.05
(700-900]	1%	17%	26%	42%	14%	113900.93
(900-1300]	0%	3%	17%	52%	27%	93120.65
(1300-5000]	0%	1%	9%	43%	47%	1441193.80

Entrenamiento: 0	Métricas		
MSE	385090.65		
MAE	396.59		
Cruce	29%	SubEstima	3%
Cruce +/-	72%	SobreEstim	25%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	388	5761	4756	3399	579	14883.00	22%
(550-700]	176	4009	4583	5454	1222	15444.00	22%
(700-900]	82	2325	3476	5738	1967	13588.00	20%
(900-1300]	6	428	1859	5716	3240	11249.00	16%
(1300-5000]	5	142	1280	6079	6165	13671.00	20%
						68835.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	3%	39%	32%	23%	4%	143530.32
(550-700]	1%	26%	30%	35%	8%	140739.85
(700-900]	1%	17%	26%	42%	14%	116082.92
(900-1300]	0%	4%	17%	51%	29%	98655.96
(1300-5000]	0%	1%	9%	44%	45%	1447105.53

Validación: 1	Métricas		
MSE	389049.8		
MAE	400.86		
Cruce	29%	SubEstima	3%
Cruce +/-	72%	SobreEstim	25%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	χ^2	Teórico
[450-550]	1.5	0.5	2.6	0.4	0.3	**71.2**	**26.3**
(550-700]	0.9	2.2	0.0	1.3	5.3	**Decisión**	
(700-900]	1.1	0.7	2.2	0.4	1.9	Se rechaza H0	
(900-1300]	1.0	6.9	5.5	1.9	7.3		
(1300-5000]	0.0	5.3	7.6	4.1	10.1		

Tiempo de Ejecución		
user	system	elapsed
0.226	0.016	0.226

Tabla 3.35: Resultados XGB sin remuestreo para G2

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G2		
MODELO: XGBOOST	G2_corte1 = 499 #10%	G2_porc1 = 0.35
VARIABLES: 35	G2_corte2 = 1000 #68%	G2_porc2 = 0.15

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	3166	7026	3362	1379	183	15116.00	22%
(550-700]	1693	6101	4550	2639	379	15362.00	22%
(700-900]	782	4286	4201	3734	610	13613.00	20%
(900-1300]	105	1651	3529	4823	1148	11256.00	16%
(1300-5000]	34	952	3252	6383	3025	13646.00	20%
						68993.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	21%	46%	22%	9%	1%	62415.79
(550-700]	11%	40%	30%	17%	2%	57804.75
(700-900]	6%	31%	31%	27%	4%	55101.04
(900-1300]	1%	15%	31%	43%	10%	83412.35
(1300-5000]	0%	7%	24%	47%	22%	1887304.18

Entrenamiento: 0	Métricas		
MSE	424244.25		
MAE	370.30		
Cruce	31%	SubEstima	10%
Cruce +/-	78%	SobreEstim	12%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	3157	6867	3327	1354	175	14880.00	22%
(550-700]	1667	6090	4548	2774	360	15439.00	22%
(700-900]	820	4265	4123	3741	637	13586.00	20%
(900-1300]	104	1602	3582	4722	1238	11248.00	16%
(1300-5000]	54	989	3311	6329	2988	13671.00	20%
						68824.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	21%	46%	22%	9%	1%	62468.30
(550-700]	11%	39%	29%	18%	2%	58432.31
(700-900]	6%	31%	30%	28%	5%	56080.73
(900-1300]	1%	14%	32%	42%	11%	84964.48
(1300-5000]	0%	7%	24%	46%	22%	1881868.88

Validación: 1	Métricas		
MSE	425320.9		
MAE	373.14		
Cruce	31%	SubEstima	10%
Cruce +/-	77%	SobreEstim	13%

Chi Cuadrado de Pearson

Real	Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[450-550]	0.0	3.6	0.4	0.5	0.3
(550-700]	0.4	0.0	0.0	6.9	1.0
(700-900]	1.8	0.1	1.4	0.0	1.2
(900-1300]	0.0	1.5	0.8	2.1	7.1
(1300-5000]	11.8	1.4	1.1	0.5	0.5

χ^2	Teórico
44.3	**26.3**
Decisión	
Se rechaza H0	

Tiempo de Ejecución		
user	system	elapsed
0.27	0.012	0.27

Tabla 3.36: Resultados XGB con remuestreo para G2

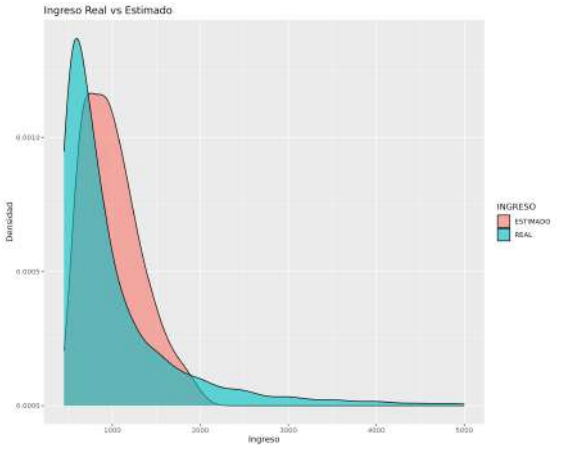


Figura 3.31: Sin remuestreo entrenamiento

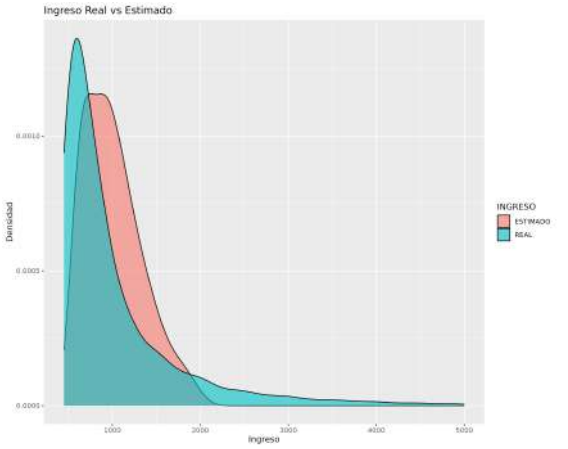


Figura 3.32: Sin remuestreo validación

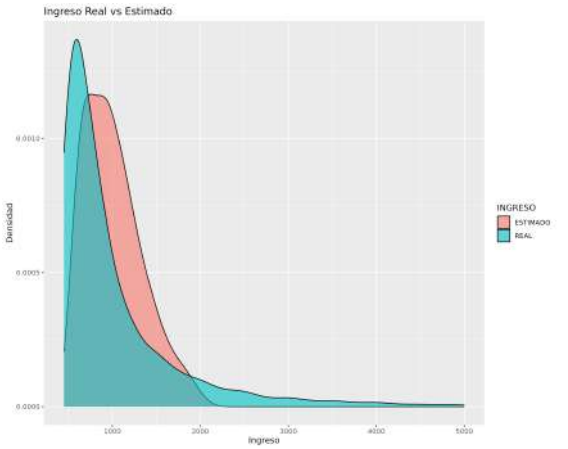


Figura 3.33: Con remuestreo entrenamiento

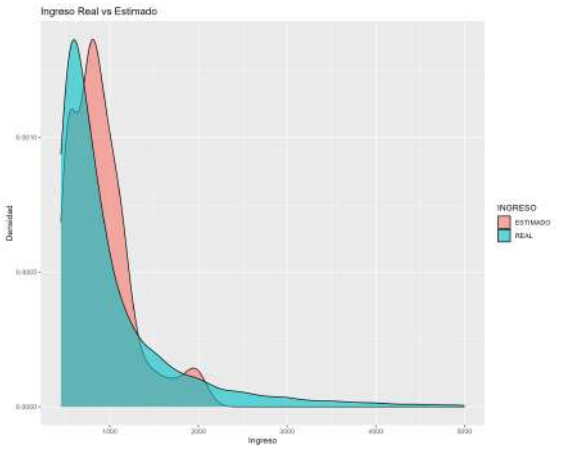


Figura 3.34: Con remuestreo validación

Figura 3.35: Distribuciones de Ingreso Real y Estimado del Modelo XGB en BDD entrenamiento y validación sin remuestreo y con remuestreo para G2

3.7.5 Modelo RLM para G2

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G2_corte1 = 499   # 11%
G2_corte2 = 1000  # 68%
G2_porc1 = 0.35
G2_porc2 = 0.20

rmod_g2 <- boots_2tail(mod_g2, corte1 =  G2_corte1, corte2 = G2_corte2,
    porc1 = G2_porc1, porc2 = G2_porc2)

rrlm_g2 <- lm(INGRESO_REAL ~ ANTIGUEDAD_OP_SICOM+
                  CUOTA_EST_OP+
                  DEUDA_TOTAL_OP_OTROS+
                  DEUDA_TOTAL_SBS_SC_24M+
                  DEUDA_TOTAL_SCE_24M+
                  DEUDA_TOTAL_SICOM_OP_24M+
                  LN_DEUDA_TOTAL_SCE_24M+
                  MaxMontoOpD24M417+
                  NOPE_APERT_SICOM_OP_36M+
                  NOPE_TOTAL_OP_OTROS+
                  numMesesInfoCredBanCoopD36M421+
                  PROM_DEUDA_TOTAL_SICOM_OP_24M+
                  PROM_XVEN_SICOM_OP_24M+
                  r_DEUDA_TOTAL_SICOM_OP_12s24M+
                  r_DEUDA_TOTAL_SICOMsSCE_24M+
                  r_NOPE_APERT_SICOM_OP_24s36M+
                  r_NOPE_APERT_SICOMsSCE_OP_36M+
```

```
                    r_PROM_DEUDA_TOTAL_SICOM_OP_24s36M+
                    r_PROM_XVEN_SICOM_OP_24s36M+
                    r_NOPE_APERT_SICOMsSCE_OP_24M+
                    NOPE_APERT_SICOM_OP_24M+
                    NumAcreedoresDDCom404+
                    cuotaCom056+
                    salTotOpCom037+
                    DEUDA_TOTAL_SCE_6M+
                    LN_DEUDA_TOTAL_SCE_6M+
                    salOpDiaCom005+
                    maxMontoOp096+
                    salPromD36M319+
                    DEUDA_TOTAL_SCE_3M+
                    LN_DEUDA_TOTAL_SCE_3M+
                    cuotaEstimadaD24M416+
                    salProm36M303+
                    SalTotOpD383, data = rmod_g2)

runtime <- system.time({
info[, INGRESO_EST_G2_5x100 := predict(rrlm_g2, newdata=info)]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
    = runtime[3])

info[, RANGO_REALG2 := cut(INGRESO_REAL, breaks = c(450, 550, 700, 900,
     1300, 5000), labels = c("[450-550]", "(550-700]", "(700-900]", "
    (900-1300]", "(1300-5000]"))]

info[, RANGO_EST_G2_5x100 := cut(INGRESO_EST_G2_5x100, breaks = c(450,
    550, 700, 900, 1300, 5000), labels = c("[450-550]", "(550-700]", "
    (700-900]", "(900-1300]", "(1300-5000]"))]
```

Código 3.25: Estracto esencial de código para obtener los Resultados del modelo RLM para G2

Métricas de MSE y MAE:

En el modelo base, el MSE es de 417 mil unidades y el MAE de 420 unidades, en el modelo con remuestreo el MSE aumenta a 438 mil unidades y el MAE disminuye a 390 unidades; estos valores de comparación son aproximados.

El MSE por rango de Ingreso Real, se mantiene similar para entrenamiento y validación.

La proporción de error de predicción se mantiene en entrenamiento y validación (tabla: 3.37 y 3.38).

Cruces y Sub/Sobre Estimaciones:

En el modelo base, el cruce es de 69%, la sobre estimación de 28% y la sub estimación de 3%; en el modelo remuestreado el cruce aumenta al 75%, la sobre estimación disminuye al 18% y la sub estimación aumenta al 7%. (estos valores de comparación son aproximados).

Nótese que, a pesar de tener cruce cerca y sobre el 70%, no tenemos buenas predicciones en los rangos de ingresos bajos y altos; en el modelo base no se tiene predicciones malas para rango de ingreso 450 a 550; en el modelo remuestreado se tiene malas predicciones para el rango de ingreso estimado entre 1300 a 5000 (tabla: 3.37 y 3.38).

Tiempos de ejecución:

Los tiempos de ejecución son significativamente pequeños; ambos modelos cerca de 0.22 segundos aproximadamente (tabla: 3.37 y 3.38).

Chi Cuadrado de Pearson:

Respecto al chi cuadrado, con ayuda de las métricas no se rechaza la hipótesis nula de que las distribuciones en entrenamiento y validación son similares (tabla: 3.37 y 3.38, figura: 3.40).

Decisión:

Con base a los criterios expuestos, no se decide por ninguno de los dos modelos, dado que uno de nuestros objetivos es mejorar las predicciones en las colas de la distribución de ingresos (individuos con ingresos muy bajo y muy altos)

CONSIDERACIONES: Nodos finales al ~~~% - Sin aplicación de remuestreo
MODELO: Regresión Lineal Múltiple
VARIABLES: 35 **G2**

Muestra de Modelamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	518	4819	4364	4844	478	15023.00	22%
(550-700]	454	3464	3572	6757	1018	15265.00	22%
(700-900]	287	1959	2759	6990	1544	13539.00	20%
(900-1300]	98	454	1302	7082	2298	11234.00	16%
(1300-5000]	36	230	880	7897	4588	13631.00	20%
						68692.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	3%	32%	29%	32%	3%	171704.44
(550-700]	3%	23%	23%	44%	7%	156860.70
(700-900]	2%	14%	20%	52%	11%	111394.47
(900-1300]	1%	4%	12%	63%	20%	69200.69
(1300-5000]	0%	2%	6%	58%	34%	1573771.66

Entrenamiento: 0	Métricas		
MSE	417040.53		
MAE	420.62		
Cruce	27%	SubEstima	3%
Cruce +/-	69%	SobreEstim	28%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	538	4835	4253	4702	449	14777.00	22%
(550-700]	472	3374	3635	6876	988	15345.00	22%
(700-900]	338	1935	2774	6891	1569	13507.00	20%
(900-1300]	85	486	1297	6876	2478	11222.00	16%
(1300-5000]	36	242	923	7917	4531	13649.00	20%
						68500.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	4%	33%	29%	32%	3%	166073.81
(550-700]	3%	22%	24%	45%	6%	155323.41
(700-900]	3%	14%	21%	51%	12%	218809.36
(900-1300]	1%	4%	12%	61%	22%	72460.65
(1300-5000]	0%	2%	7%	58%	33%	1555598.18

Validación: 1	Métricas		
MSE	434740.55		
MAE	421.52		
Cruce	26%	SubEstima	3%
Cruce +/-	69%	SobreEstim	27%

Chi Cuadrado de Pearson

Real	Estimado					χ^2	Teórico
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	**55.5**	**26.3**
[450-550]	0.8	0.1	2.8	4.2	1.8	**Decisión**	
(550-700]	0.7	2.3	1.1	2.1	0.9	Se rechaza H0	
(700-900]	9.1	0.3	0.1	1.4	0.4		
(900-1300]	1.7	2.3	0.0	6.0	14.1		
(1300-5000]	0.0	0.6	2.1	0.1	0.7		

Tiempo de Ejecución		
user	system	elapsed
0.16	0.06	0.22

Tabla 3.37: Resultados RLM sin remuestreo para G2

CONSIDERACIONES: Nodos finales al ~~~% - Con aplicación de remuestreo G2		
MODELO: Regresión Lineal Múltiple	G2_corte1 = 499 # 11%	G2_porc1 = 0.35
VARIABLES: 35	G2_corte2 = 1000 # 68%	G2_porc2 = 0.20

Muestra de Modelamiento

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	2365	6151	3118	3166	98	14898.00	22%
(550-700]	1707	4659	3353	5220	208	15147.00	22%
(700-900]	1029	3049	2995	6121	273	13467.00	20%
(900-1300]	259	1060	2181	7153	538	11191.00	16%
(1300-5000]	126	622	2011	9302	1548	13609.00	20%
						68312.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	16%	41%	21%	21%	1%	90267.30
(550-700]	11%	31%	22%	34%	1%	79942.79
(700-900]	8%	23%	22%	45%	2%	59518.79
(900-1300]	2%	9%	19%	64%	5%	68803.35
(1300-5000]	1%	5%	15%	68%	11%	1912084.05

Entrenamiento: 0	Métricas		
MSE	438666.81		
MAE	390.89		
Cruce	27%	SubEstima	7%
Cruce +/-	75%	SobreEstim	18%

Muestra de Validación

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	2264	6120	3101	3043	86	14614.00	21%
(550-700]	1656	4698	3309	5368	184	15215.00	22%
(700-900]	1093	2994	3007	6074	267	13435.00	20%
(900-1300]	251	1069	2233	7040	588	11181.00	16%
(1300-5000]	128	677	2026	9325	1475	13631.00	20%
						68076.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	15%	42%	21%	21%	1%	87648.28
(550-700]	11%	31%	22%	35%	1%	78268.33
(700-900]	8%	22%	22%	45%	2%	196533.85
(900-1300]	2%	10%	20%	63%	5%	70819.14
(1300-5000]	1%	5%	15%	68%	11%	1894463.01

Validación: 1	Métricas		
MSE	463130.61		
MAE	392.17		
Cruce	27%	SubEstima	8%
Cruce +/-	75%	SobreEstim	18%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	χ^2	Teórico
[450-550]	4.3	0.2	0.1	4.8	1.5	**42.2**	**26.3**
(550-700]	1.5	0.3	0.6	4.2	2.8	**Decisión**	
(700-900]	4.0	1.0	0.0	0.4	0.1	Se rechaza H0	
(900-1300]	0.2	0.1	1.2	1.8	4.6		
(1300-5000]	0.0	4.9	0.1	0.1	3.4		

Tiempo de Ejecución		
user	system	elapsed
0.15	0.12	0.27

Tabla 3.38: Resultados RLM con remuestreo para G2

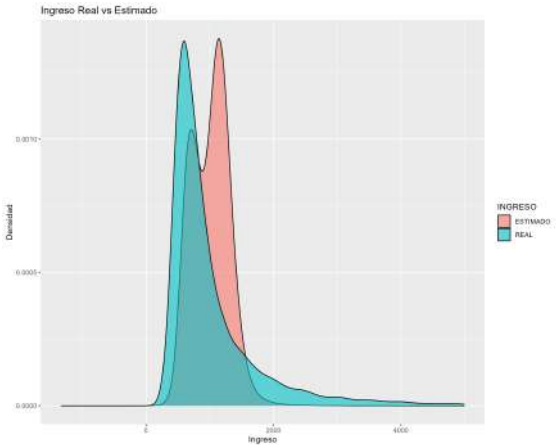


Figura 3.36: Sin remuestreo entrenamiento

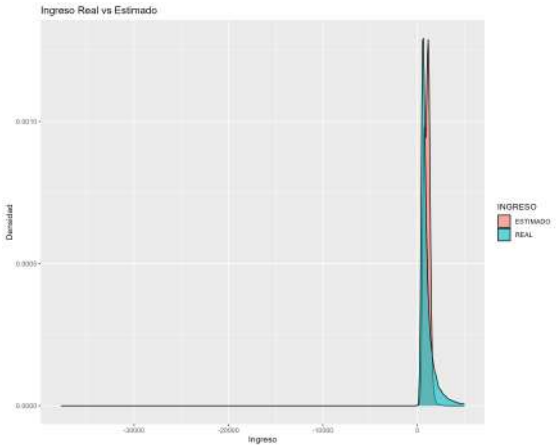


Figura 3.37: Sin remuestreo validación

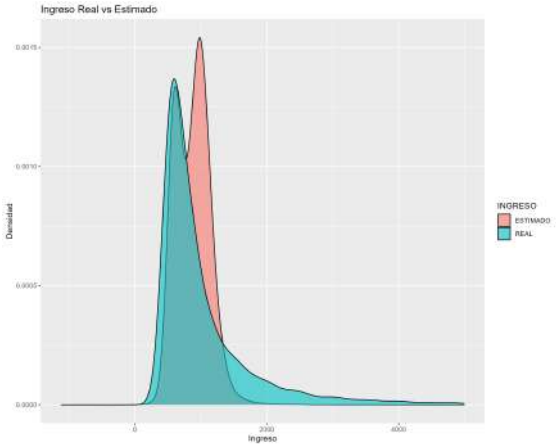


Figura 3.38: Con remuestreo entrenamiento

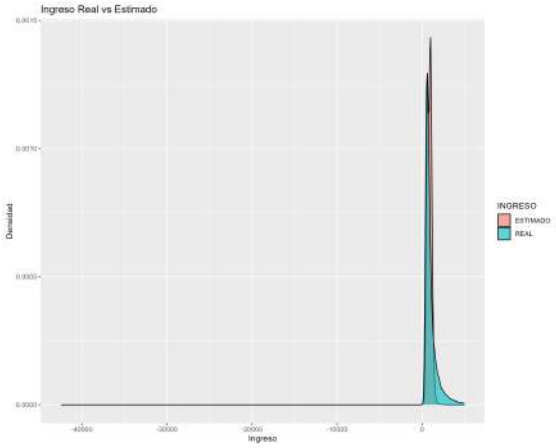


Figura 3.39: Con remuestreo validación

Figura 3.40: Distribuciones de Ingreso Real y Estimado del Modelo RLM en BDD entrenamiento y validación sin remuestreo y con remuestreo para G2

3.7.6 Elección del mejor modelo entre RLM, RF, GBM y XGB para G2

Se elige como mejor Modelo al GBM para esta sub población o grupo G2, debido a su costo computacional relativamente bajo y su capacidad de mejorar las predicciones tanto para individuos con ingresos muy bajos como muy altos. Además, durante las simulaciones y remuestreos, se obtuvo una suerte de matriz banda en las matrices de coincidencia, esto indica una mejora en las predicciones tanto en las colas y el cuerpo de la distribución de ingresos.

Se reconoce que el modelo XGB es mucho más eficiente desde el punto de vista matemático y computacional; pero su dificultad para la implementación nos persuade a no seleccionarlo, pues no está disponible para windows.

3.8 Modelos para población G3

3.8.1 Exploración y descripción de la base de datos

Esta población cuenta con sujetos con cuota estimada actual mayor a 435 dólares.

Siguiendo el mismo razonamiento, dado que tenemos más de 4 mil registros optamos por tomar la mitad para entrenamiento y la otra mitad para entrenamiento (tabla 3.39).

Muestra	N	%
Entrenamiento	37,532	49.85%
Validación	37,764	50.15%
Total	**75,296**	**100.00%**

Tabla 3.39: Población del Grupo 3

Percentil	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Valor en USD	452	600	700	800	852	930	1012	1152	1260	1436	1572

Percentil		55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
Valor en USD		1,786	2,000	2,250	2,583	3,000	3,591	4,333	5,481	7,700	35000

Tabla 3.40: Percentil del Ingreso Real del Grupo 3

De igual forma tabla 3.40 de percentiles será de mucha ayuda para el entrenamiento del modelo, dado que ayuda a determinar los cortes y proporciones adecuados para los parámetros del remuestreo.

3.8.2 Modelo RF para G3

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G3_corte1 = 501; G3_corte2 = 1300   # 15%   # 80%
G3_porc1 = 0.45; G3_porc2 = 0.11

rmod_g3 <- boots_2tail(mod_g3, corte1 =  G3_corte1, corte2 = G3_corte2,
    porc1 = G3_porc1, porc2 = G3_porc2)
G3_min_nodo = presen_colas_g3[4]
rf_g3_5x100 <- ranger(formula = as.formula(formula_g3), data = rmod_g3,
    num.trees = 300, mtry = 11, min.node.size = G3_min_nodo,
   importance = 'impurity', write.forest = TRUE, seed = 1234)
runtime <- system.time({info[, INGRESO_EST_G3_5x100 := predict(object=
   rf_g3_5x100, data=info)$predictions]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[3])
info[, RANGO_REALG3 := cut(INGRESO_REAL, breaks = c(450, 850, 1250,
   2000, 4000, 35000), labels = c("[450-850]", "(850-1250]", "
   (1250-2000]", "(2000-4000]", "(4000-35000]"))]
info[, RANGO_EST_G3_5x100 := cut(INGRESO_EST_G3_5x100, breaks = c(450,
   850, 1250, 2000, 4000, 35000), labels = c("[450-850]", "(850-1250]"
   , "(1250-2000]", "(2000-4000]", "(4000-35000]"))]
```

Código 3.26: Estracto esencial de código para obtener los Resultados del modelo RF para G3

Métricas de MSE y MAE:

En el modelo base, el MSE el aproximadamente del 5100000 unidades en

entrenamiento y validación, el MAE es aproximadamente 1250 unidades; en el modelo con remuestreo el MSE aumenta a 5700000 unidades y el MAE disminuye a 1190 aproximadamente.

Estos resultados nos sugieren que el modelo el consistente controlando los errores en entrenamiento y validación. Se tiene una mejora en el MAE en el modelo con remuestreo dado que predice mejor en los extremos de las colas de la distribución de ingresos (tabla: 3.41 y 3.42, figura: 3.45)

Cruces y Sub/Sobre Estimaciones:

En el modelo base se tiene un cruce de 79%, una sobre estimación del 18% y una su estimación del 2% aproximadamente. En el modelo con remuestreo el cruce aumenta al 80% , la sobre estimación disminuye al 7% y la sub estimación aumenta al 7% aproximadamente.

Estos resultados indica que tenemos una mejora en: aciertos y sobre estimación; una perdida en la sub estimación (tabla: 3.41 y 3.42).

Tiempos de ejecución:

En el modelo base, el tiempo de ejecución es de 13.56 segundo, mientras que en el modelo con remuestreo el tiempo de ejecución aumenta a 16.96 segundos. El aumento de estos tiempos es aceptable dado que mejoramos en predicciones (tabla: 3.41 y 3.42).

Chi Cuadrado de Pearson:

Se recuerda que el Chi cuadro es sensible a valores por cruces, por ello tenemos valores calculados mayores a los teóricos; por esta razón se recurre al apoyo de las métricas para decidir que la distribución de entrenamiento y validación son similares (tabla: 3.41 y 3.42, figura: 3.45).

Decisión:

Con base a los criterios expuesto, se elige al modelo con remuestro como el modelo adecuado para el grupo G3.

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Random Forest
VARIABLES: 28 **G3**

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1054	2378	3219	739	20	7410.00	20%
(850-1250]	156	979	3945	2287	56	7423.00	20%
(1250-2000]	37	329	2960	4256	403	7985.00	21%
(2000-4000]	6	131	1828	4699	1737	8401.00	22%
(4000-35000]	1	18	450	2673	3171	6313.00	17%
						37532.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	14%	32%	43%	10%	0%	720823.20
(850-1250]	2%	13%	53%	31%	1%	990753.97
(1250-2000]	0%	4%	37%	53%	5%	1230095.94
(2000-4000]	0%	2%	22%	56%	21%	1828788.17
(4000-35000]	0%	0%	7%	42%	50%	24377607.00

Entrenamiento: 0	Métricas		
MSE	5109705		
MAE	1256.5		
Cruce	34%	SubEstima	2%
Cruce +/-	80%	SobreEstim	18%

Muestra de Validación

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1088	2397	3315	813	18	7631.00	20%
(850-1250]	185	957	3928	2354	88	7512.00	20%
(1250-2000]	49	316	2934	4268	428	7995.00	21%
(2000-4000]	12	127	1751	4724	1752	8366.00	22%
(4000-35000]	1	35	568	2599	3057	6260.00	17%
						37764.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	14%	31%	43%	11%	0%	744259.98
(850-1250]	2%	13%	52%	31%	1%	1067525.951
(1250-2000]	1%	4%	37%	53%	5%	1295034.529
(2000-4000]	0%	2%	21%	56%	21%	1856590.695
(4000-35000]	0%	1%	9%	42%	49%	27705781.87

Validación: 1	Métricas		
MSE	5640900		
MAE	1291.9		
Cruce	34%	SubEstima	2%
Cruce +/-	79%	SobreEstim	19%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	χ^2	Teórico
[450-850]	1.1	0.2	2.9	7.4	0.2	**106.9**	**26.3**
(850-1250]	5.4	0.5	0.1	2.0	18.3	**Decisión**	
(1250-2000]	3.9	0.5	0.2	0.0	1.6	Se rechaza H0	
(2000-4000]	6.0	0.1	3.2	0.1	0.1		
(4000-35000]	0.0	16.1	30.9	2.0	4.1		

Tiempo de Ejecución		
user	system	elapsed
13.09	0.47	13.56

Tabla 3.41: Resultados RF sin remuestreo para G3

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G3		
MODELO: Random Forest	G3_corte1 = 780 #14%	G3_porc1 = 0.35
VARIABLES: 28	G3_corte2 = 2200 #64%	G3_porc2 = 0.13

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	2533	3422	1354	100	1	7410.00	20%
(850-1250]	562	2758	3338	759	6	7423.00	20%
(1250-2000]	156	1602	3770	2317	140	7985.00	21%
(2000-4000]	59	958	2986	3293	1105	8401.00	22%
(4000-35000]	14	221	1212	2479	2387	6313.00	17%
						37532.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	34%	46%	18%	1%	0%	203269.41
(850-1250]	8%	37%	45%	10%	0%	332229.01
(1250-2000]	2%	20%	47%	29%	2%	578577.09
(2000-4000]	1%	11%	36%	39%	13%	1920285.30
(4000-35000]	0%	4%	19%	39%	38%	29995527.20

Entrenamiento: 0	Métricas		
MSE	5704103		
MAE	1170.5		
Cruce	39%	SubEstima	7%
Cruce +/-	87%	SobreEstim	6%

Muestra de Validación

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	2557	3482	1476	116	0	7631.00	20%
(850-1250]	584	2723	3401	799	5	7512.00	20%
(1250-2000]	180	1627	3688	2345	155	7995.00	21%
(2000-4000]	63	932	2984	3282	1105	8366.00	22%
(4000-35000]	17	279	1222	2409	2333	6260.00	17%
						37764.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	34%	46%	19%	2%	0%	216750.64
(850-1250]	8%	36%	45%	11%	0%	366836.05
(1250-2000]	2%	20%	46%	29%	2%	609238.58
(2000-4000]	1%	11%	36%	39%	13%	1916768.11
(4000-35000]	0%	4%	20%	38%	37%	32831916.30

Validación: 1	Métricas		
MSE	6112807		
MAE	1193.9		
Cruce	39%	SubEstima	7%
Cruce +/-	86%	SobreEstim	7%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	χ^2	Teórico
[450-850]	0.2	1.1	11.0	2.6	1.0	**48.6**	**26.3**
(850-1250]	0.9	0.4	1.2	2.1	0.2	**Decisión**	
(1250-2000]	3.7	0.4	1.8	0.3	1.6	Se rechaza H0	
(2000-4000]	0.3	0.7	0.0	0.0	0.0		
(4000-35000]	0.6	15.2	0.1	2.0	1.2		

Tiempo de Ejecución		
user	system	elapsed
16.58	0.38	16.96

Tabla 3.42: Resultados RF con remuestreo para G3

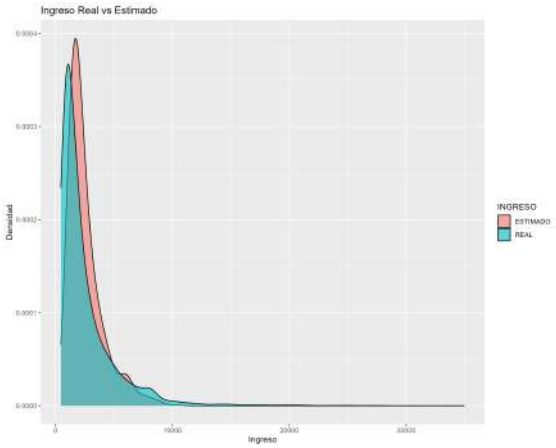


Figura 3.41: Sin remuestreo entrenamiento

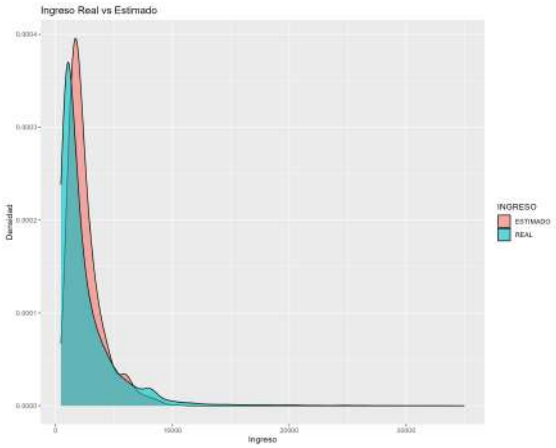


Figura 3.42: Sin remuestreo validación

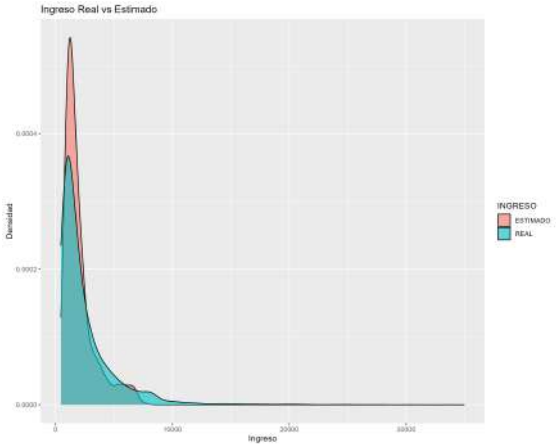


Figura 3.43: Con remuestreo entrenamiento

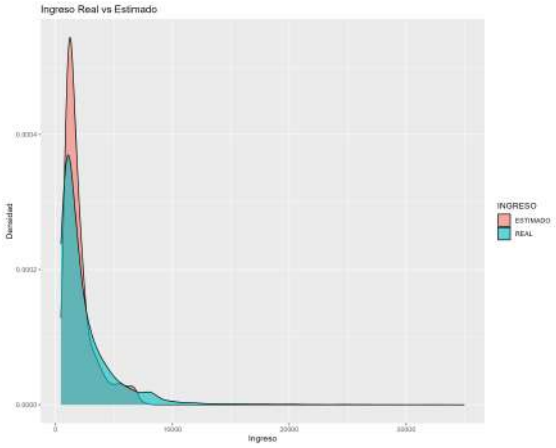


Figura 3.44: Con remuestreo validación

Figura 3.45: Distribuciones de Ingreso Real y Estimado del Modelo RF en BDD entrenamiento y validación sin remuestreo y con remuestreo para G3

3.8.3 Modelo GBM para G3

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G3_corte1 = 501; G3_corte2 = 1300   # 15%   # 80%
G3_porc1 = 0.45; G3_porc2 = 0.11

rmod_g3 <- boots_2tail(mod_g3, corte1 =  G3_corte1, corte2 = G3_corte2,
    porc1 = G3_porc1, porc2 = G3_porc2)
G3_min_nodo = presen_colas_g3[4]
rgbm_g3 <- gbm(formula = as.formula(formula_g3), data = rmod_g3, n.
   trees = 300, n.minobsinnode = G3_min_nodo, shrinkage = 0.03,
   distribution = "laplace")
runtime <- system.time({info[, INGRESO_EST_G3_5x100 := predict(rgbm_g3,
    n.trees = rgbm_g3$n.trees, info)]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[3])
info[, RANGO_REALG3 := cut(INGRESO_REAL, breaks = c(450, 850, 1250,
   2000, 4000, 35000), labels = c("[450-850]", "(850-1250]", "
   (1250-2000]", "(2000-4000]", "(4000-35000]"))]
info[, RANGO_EST_G3_5x100 := cut(INGRESO_EST_G3_5x100, breaks = c(450,
   850, 1250, 2000, 4000, 35000), labels = c("[450-850]", "(850-1250]"
   , "(1250-2000]", "(2000-4000]", "(4000-35000]"))]
```

Código 3.27: Estracto esencial de código para obtener los Resultados del modelo GBM para G3

Métricas de MSE y MAE:

En el modelo base, el MSE es de 10 millones unidades, el MAE es de 1590

unidades aproximadamente; en el modelo con remuestreo el MSE disminuye a 5 millones y el MAE disminuye a 1300 unidades aproximadamente.

El MSE por rango de Ingreso Real son similares en entrenamiento y validación tanto en el modelo base como en el remuestreado.

Estos resultados indica que el modelo es consistente en entrenamiento y validación, que mantiene controlado a los errores en las predicciones; se nota claramente una mejora significativa en el modelo con remuestreo (tabla: 3.43 y 3.44).

Cruces y Sub/Sobre Estimaciones:

En el modelo base, el cruce es de 63%, la sobre estimación es del 0% y la sub estimación es del 36% aproximadamente. En el modelo con remuestreo, el cruce aumenta al 83%, la sobre estimación al 9% y la sub estimación disminuye al 8% aproximadamente.

Al analizar los cruces de las matrices del modelo base y del modelo con remuestreo, vemos que existe una ganancia significativa con el modelo con remuestreo; pues ya se tiene predicciones para los rangos de ingreso 2000 a 4000 y 4000 a 35000 (tabla: 3.43 y 3.44).

Tiempos de ejecución:

El modelo base tiene una tiempo de ejecución de 1.20 segundos aproximadamente; en el modelo con remuestreo el tiempo es de 1.35 segundos aproximadamente. Este incremento en el tiempo no es significativo, más aún cuando el modelo tiene mejoras significativas en las predicciones (tabla: 3.43 y 3.44).

Chi Cuadrado de Pearson:

Dado que el Chi cuadrado el sensible al numero de individuos por cruce, nos apoyamos en las métricas para decidir que la distribución de ingresos en entrenamiento y validación son similares (gráficos 3.50).

Es importante notar que, en el modelo base, el valor de chi cuadrado es 20.3, considerablemente menor que el valor teórico de 26.3. Esto sugiere que no deberíamos rechazar la hipótesis nula en cuanto a la similitud de distribuciones. Sin embargo, a pesar de este resultado, el modelo base demostró un rendimiento deficiente en términos de predicciones. Aquí es donde entra en juego el análisis de las métricas, que nos proporciona una

comprensión más sólida de la similitud de las distribuciones. Este enfoque se aplica de manera similar en el caso del modelo con remuestreo, donde el chi cuadrado calculado es 36.6, mayor que el valor teórico de 26.3 (esta justificación es aplicable para todos los modelos respecto al Chi cuadrado) (tabla: 3.43 y 3.44, figura: 3.50).

Decisión:

Con base a los criterios expuestos, se selecciona como mejor modelo al modelo con remuestreo para el grupo 3.

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Gradient Boosting Machine
VARIABLES: 29 **G3**

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	3602	3721	87	0	0	7410.00	20%
(850-1250]	1478	5682	263	0	0	7423.00	20%
(1250-2000]	731	6392	862	0	0	7985.00	21%
(2000-4000]	332	6261	1808	0	0	8401.00	22%
(4000-35000]	91	3706	2516	0	0	6313.00	17%
						37532.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	49%	50%	1%	0%	0%	74200.45
(850-1250]	20%	77%	4%	0%	0%	38328.77
(1250-2000]	9%	80%	11%	0%	0%	341488.47
(2000-4000]	4%	75%	22%	0%	0%	3261180.24
(4000-35000]	1%	59%	40%	0%	0%	53866898.60

Entrenamiento: 0	Métricas		
MSE	9885431.9		
MAE	1593.7		
Cruce	27%	SubEstima	36%
Cruce +/-	63%	SobreEstim	0%

Muestra de Validación

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	3733	3810	88	0	0	7631.00	20%
(850-1250]	1519	5701	292	0	0	7512.00	20%
(1250-2000]	716	6470	809	0	0	7995.00	21%
(2000-4000]	310	6297	1759	0	0	8366.00	22%
(4000-35000]	91	3636	2533	0	0	6260.00	17%
						37764.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	49%	50%	1%	0%	0%	73074.69
(850-1250]	20%	76%	4%	0%	0%	38779.6062
(1250-2000]	9%	81%	10%	0%	0%	342245.611
(2000-4000]	4%	75%	21%	0%	0%	3300806.54
(4000-35000]	1%	58%	40%	0%	0%	56984273.5

Validación: 1	Métricas		
MSE	10272251		
MAE	1596.4		
Cruce	27%	SubEstima	36%
Cruce +/-	64%	SobreEstim	0%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	χ^2	Teórico
[450-850]	4.8	2.1	0.0	0.0	0.0	**20.3**	**26.3**
(850-1250]	1.1	0.1	3.2	0.0	0.0	**Decisión**	
(1250-2000]	0.3	1.0	3.3	0.0	0.0	No se rechaza H0	
(2000-4000]	1.5	0.2	1.3	0.0	0.0		
(4000-35000]	0.0	1.3	0.1	0.0	0.0		

Tiempo de Ejecución		
user	system	elapsed
0.97	0.02	1.2

Tabla 3.43: Resultados GBM sin remuestreo para G3

CONSIDERACIONES: Nodos finales al 3.6% -Con aplicación de remuestreo G3		
MODELO: Gradient Boosting Machine	G3_corte1 = 840 #19%	G3_porc1 = 0.30
VARIABLES: 29	G3_corte2 = 7060 #94%	G3_porc2 = 0.30

Muestra de Entrenamiento

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	3490	2447	1141	282	45	7405.00	20%
(850-1250]	1271	2597	2280	1133	142	7423.00	20%
(1250-2000]	469	1840	2542	2393	741	7985.00	21%
(2000-4000]	202	1206	1988	2628	2377	8401.00	22%
(4000-35000]	57	351	756	1546	3603	6313.00	17%
						37527.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	47%	33%	15%	4%	1%	391330.67
(850-1250]	17%	35%	31%	15%	2%	789138.19
(1250-2000]	6%	23%	32%	30%	9%	1794643.46
(2000-4000]	2%	14%	24%	31%	28%	3845927.87
(4000-35000]	1%	6%	12%	24%	57%	24857036.60

Entrenamiento: 0	Métricas		
MSE	5657035.9		
MAE	1293.4		
Cruce	40%	SubEstima	8%
Cruce +/-	83%	SobreEstim	9%

Muestra de Validación

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	3576	2487	1208	313	46	7630.00	20%
(850-1250]	1296	2600	2315	1121	180	7512.00	20%
(1250-2000]	476	1897	2445	2429	748	7995.00	21%
(2000-4000]	218	1202	1934	2629	2383	8366.00	22%
(4000-35000]	61	372	794	1498	3535	6260.00	17%
						37763.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	48%	34%	16%	4%	1%	397277.84
(850-1250]	17%	35%	31%	15%	2%	878634.596
(1250-2000]	6%	24%	31%	30%	9%	1859754.71
(2000-4000]	3%	14%	23%	31%	28%	3807787.58
(4000-35000]	1%	6%	13%	24%	56%	27082861.1

Validación: 1	Métricas		
MSE	5981763.9		
MAE	1314.4		
Cruce	39%	SubEstima	8%
Cruce +/-	82%	SobreEstim	10%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	χ^2	Teórico
[450-850]	2.1	0.7	3.9	3.4	0.0	**36.6**	**26.3**
(850-1250]	0.5	0.0	0.5	0.1	10.2	**Decisión**	
(1250-2000]	0.1	1.8	3.7	0.5	0.1	Se rechaza H0	
(2000-4000]	1.3	0.0	1.5	0.0	0.0		
(4000-35000]	0.3	1.3	1.9	1.5	1.3		

Tiempo de Ejecución		
user	system	elapsed
0.94	0	1.35

Tabla 3.44: Resultados GBM con remuestreo para G3

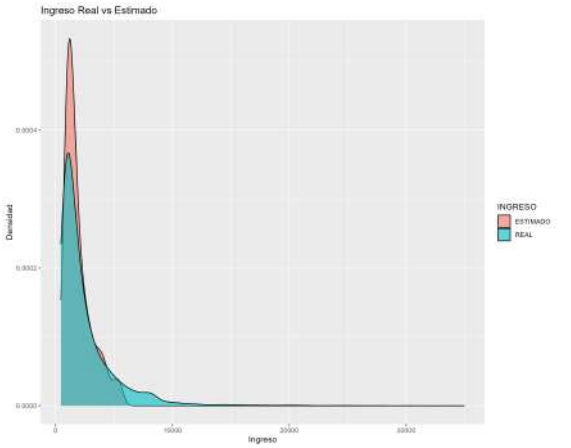


Figura 3.46: Sin remuestreo entrenamiento

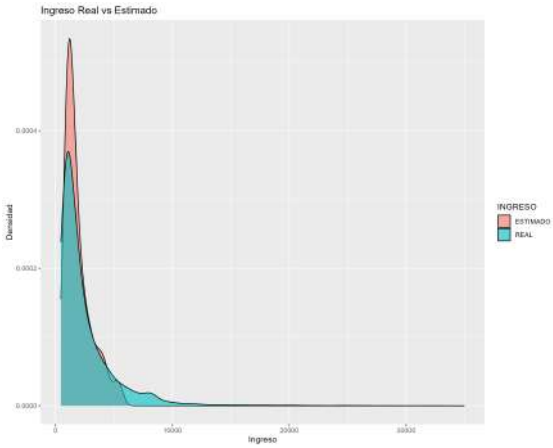


Figura 3.47: Sin remuestreo validación

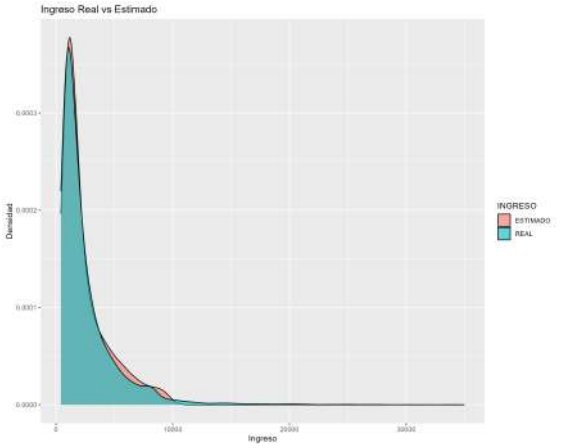


Figura 3.48: Con remuestreo entrenamiento

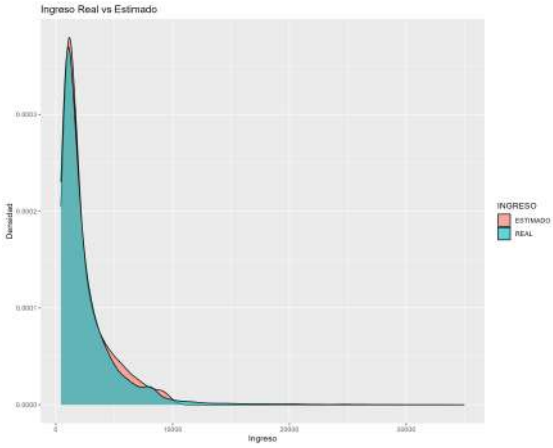


Figura 3.49: Con remuestreo validación

Figura 3.50: Distribuciones de Ingreso Real y Estimado del Modelo GBM en BDD entrenamiento y validación sin remuestreo y con remuestreo para G3

3.8.4 Modelo XGB para G3

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G3_corte1 = 780  #14%
G3_corte2 = 4050 #83%
G3_porc1 = 0.40
G3_porc2 = 0.10
rmod_g3 <- boots_2tail(mod_g3, corte1 =  G3_corte1, corte2 = G3_corte2,
    porc1 = G3_porc1, porc2 = G3_porc2)
g3_min_nodo = presen_colas_g3[4]
#Transformacion de BDD Modelamiento, Train y Test a H2O
#entrenamiento del Modelo en H20
runtime <- system.time({
info[, INGRESO_EST_G3_5x100 := setDT(as.data.frame(h2o.predict(my_xgb_
   g3, newdata =info_em)))]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
   = runtime[1])
info[, RANGO_REALG3 := cut(INGRESO_REAL, breaks = c(450, 850, 1250,
   2000, 4000, 35000), labels = c("[450-850]", "(850-1250]", "
   (1250-2000]", "(2000-4000]", "(4000-35000]"))]
info[, RANGO_EST_G3_5x100 := cut(INGRESO_EST_G3_5x100, breaks = c(450,
   850, 1250, 2000, 4000, 35000), labels = c("[450-850]", "(850-1250]"
   , "(1250-2000]", "(2000-4000]", "(4000-35000]"))]
```

Código 3.28: Estracto esencial de código para obtener los Resultados del modelo XGB para G3

Métricas de MSE y MAE:

En el modelo base, el MSE es de 5 millones de unidades y el MAE de 1257 unidades; en el modelo remuestreado, el MSE en 5 millones de unidades y el MAE disminuye a 1164. Estos comparaciones son aproximadas.

El MSE por rango de Ingreso Real se mantiene similar para entrenamiento y validación tanto para el modelo base como para el remuestreado.

Esto nos dice que el modelo es consistente en cuanto al control de los errores de predicción en entrenamiento y validación, y esta consistencia se mantiene en el modelo remuestreado (tabla: 3.45 y 3.46).

Cruces y Sub/Sobre Estimaciones:

En el modelo base se tiene un cruce de 82%, una sobre estimación de 16% y la sub estimación de 2%; en el modelo remuestreado, el cruce aumenta al 86%, la sobre estimación disminuye al 10% y la sub estimación aumenta al 5%. Estas comparaciones son aproximadas (tabla: 3.45 y 3.46).

Tiempos de ejecución:

El tiempo de ejecución del modelo base es de 0.21 segundo, en el modelo remuestreado es de 0.23 segundos aproximadamente. No existe aumento significativo en el costo computacional al remuestrear (tabla: 3.45 y 3.46).

Chi Cuadrado de Pearson:

Con es conocido, se apoya en las métricas para determinar que la distribución de entrenamiento y validación son similares tanto en modelo base como en el remuestreado (tabla: 3.45 y 3.46, 3.55).

Decisión:

Con base a los criterios expuestos, se decide como mejor modelo al modelo remuestreado para el grupo 3.

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: XGBOOST
VARIABLES: 29 **G3**

Muestra de entrenamiento

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1599	2502	2621	660	27	7409.00	20%
(850-1250]	332	1438	3595	1952	105	7422.00	20%
(1250-2000]	91	529	2979	3972	414	7985.00	21%
(2000-4000]	22	222	1824	4525	1808	8401.00	22%
(4000-35000]	3	49	463	2436	3362	6313.00	17%
						37530.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	22%	34%	35%	9%	0%	660611.64
(850-1250]	4%	19%	48%	26%	1%	970639.14
(1250-2000]	1%	7%	37%	50%	5%	1262668.46
(2000-4000]	0%	3%	22%	54%	22%	2057807.84
(4000-35000]	0%	1%	7%	39%	53%	23559835.10

Entrenamiento: 0	Métricas		
MSE	5014480.4		
MAE	1226.7		
Cruce	37%	SubEstima	2%
Cruce +/-	82%	SobreEstim	15%

Muestra de Validación

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1653	2612	2629	714	23	7631.00	20%
(850-1250]	349	1415	3621	2019	108	7512.00	20%
(1250-2000]	85	567	2940	3944	459	7995.00	21%
(2000-4000]	24	225	1761	4568	1788	8366.00	22%
(4000-35000]	6	53	548	2385	3268	6260.00	17%
						37764.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	22%	34%	34%	9%	0%	665181.60
(850-1250]	5%	19%	48%	27%	1%	1043690.708
(1250-2000]	1%	7%	37%	49%	6%	1361669.283
(2000-4000]	0%	3%	21%	55%	21%	2091068.406
(4000-35000]	0%	1%	9%	38%	52%	26385216.76

Validación: 1	Métricas		
MSE	5467325.7		
MAE	1257.6		
Cruce	37%	SubEstima	2%
Cruce +/-	82%	SobreEstim	16%

Chi Cuadrado de Pearson

Real	Estimado					χ^2	Teórico
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1.8	4.8	0.0	4.4	0.6	**49.9**	**26.3**
(850-1250]	0.9	0.4	0.2	2.3	0.1	**Decisión**	
(1250-2000]	0.4	2.7	0.5	0.2	4.9	Se rechaza H0	
(2000-4000]	0.2	0.0	2.2	0.4	0.2		
(4000-35000]	3.0	0.3	15.6	1.1	2.6		

Tiempo de Ejecución		
user	system	elapsed
0.213	0.011	0.213

Tabla 3.45: Resultados XGB sin remuestreo para G3

CONSIDERACIONES: Nodos finales al 3.6% -Con aplicación de remuestreo G3		
MODELO: XGBOOST	G3_corte1 = 780 #14%	G3_porc1 = 0.40
VARIABLES: 29	G3_corte2 = 4050 #83%	G3_porc2 = 0.10

Muestra de entrenamiento

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	2848	2502	1767	288	3	7408.00	20%
(850-1250]	792	2047	3267	1288	29	7423.00	20%
(1250-2000]	226	1068	3292	3176	223	7985.00	21%
(2000-4000]	85	527	2294	4062	1433	8401.00	22%
(4000-35000]	21	125	714	2548	2905	6313.00	17%
						37530.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	38%	34%	24%	4%	0%	336072.34
(850-1250]	11%	28%	44%	17%	0%	584312.43
(1250-2000]	3%	13%	41%	40%	3%	857751.94
(2000-4000]	1%	6%	27%	48%	17%	1835289.50
(4000-35000]	0%	2%	11%	40%	46%	26907179.28

Entrenamiento: 0	Métricas		
MSE	5301078.7		
MAE	1164.4		
Cruce	40%	SubEstima	5%
Cruce +/-	86%	SobreEstim	10%

Muestra de Validación

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	2964	2519	1815	330	2	7630.00	20%
(850-1250]	809	2016	3316	1326	44	7511.00	20%
(1250-2000]	248	1087	3293	3080	286	7994.00	21%
(2000-4000]	88	539	2259	4067	1413	8366.00	22%
(4000-35000]	17	160	758	2461	2864	6260.00	17%
						37761.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	39%	33%	24%	4%	0%	346923.37
(850-1250]	11%	27%	44%	18%	1%	626527.82
(1250-2000]	3%	14%	41%	39%	4%	939304.03
(2000-4000]	1%	6%	27%	49%	17%	1835421.70
(4000-35000]	0%	3%	12%	39%	46%	29534586.96

Validación: 1	Métricas		
MSE	5696039.5		
MAE	1187.6		
Cruce	40%	SubEstima	5%
Cruce +/-	85%	SobreEstim	10%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	χ^2	Teórico
[450-850]	4.7	0.1	1.3	6.1	0.3	**64.3**	**26.3**
(850-1250]	0.4	0.5	0.7	1.1	7.8	**Decisión**	
(1250-2000]	2.1	0.3	0.0	2.9	17.8	Se rechaza H0	
(2000-4000]	0.1	0.3	0.5	0.0	0.3		
(4000-35000]	0.8	9.8	2.7	3.0	0.6		

Tiempo de Ejecución		
user	system	elapsed
0.238	0.008	0.238

Tabla 3.46: Resultados XGB con remuestreo para G3

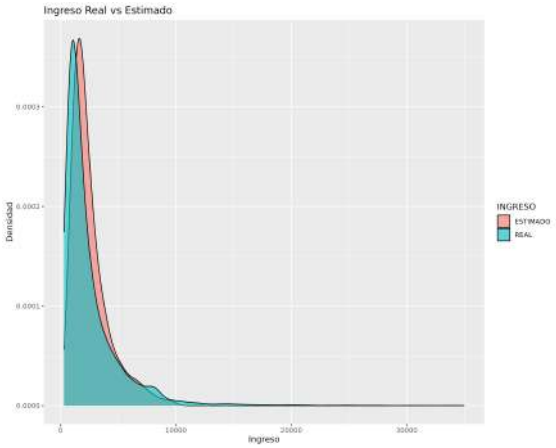


Figura 3.51: Sin remuestreo entrenamiento

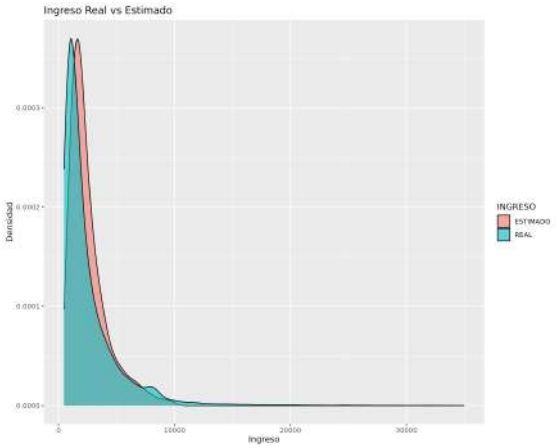


Figura 3.52: Sin remuestreo validación

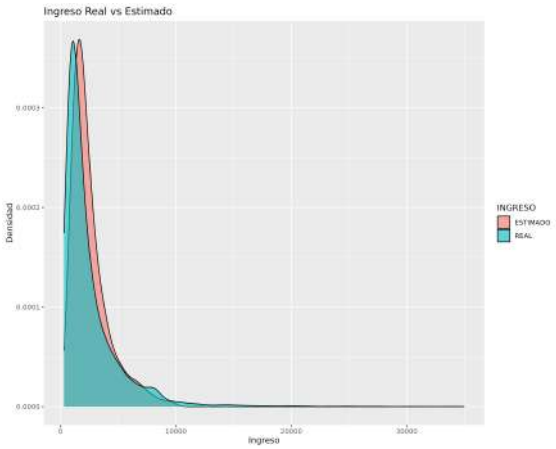


Figura 3.53: Con remuestreo entrenamiento

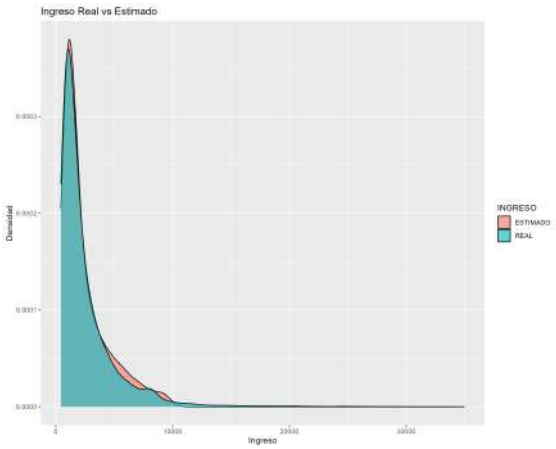


Figura 3.54: Con remuestreo validación

Figura 3.55: Distribuciones de Ingreso Real y Estimado del Modelo XGB en BDD entrenamiento y validación sin remuestreo y con remuestreo para G3

3.8.5 Modelo RLM para G3

En la presente sección, se exponen los resultados obtenidos tanto del modelo base como del modelo remuestreado. Se incluyen la matriz de coincidencias, las métricas aplicadas en los grupos entrenamiento y validación, los valores del error cuadrático medio (MSE) correspondientes a diferentes rangos de ingreso real, así como la matriz Chi cuadrado, que evidencia la semejanza en la distribución entre los conjuntos entrenamiento y validación. Además, se presentan los tiempos de ejecución correspondientes a los modelos base y remuestreado, proporcionando una visión completa de los resultados y desempeño de ambos enfoques.

Se proporciona aquí un extracto esencial del código utilizado para obtener los resultados mencionados previamente. Para acceder a la versión completa y reproducible de todo el proyecto de integración curricular, se encuentra disponible en la sección de anexos al final del documento.

```
G3_corte1 = 840  #19%
G3_corte2 = 4000 #83%
G3_porc1 = 0.35
G3_porc2 = 0.13

rmod_g3 <- boots_2tail(mod_g3, corte1 =  G3_corte1, corte2 = G3_corte2,
    porc1 = G3_porc1, porc2 = G3_porc2)

rrlm_g3 <- lm(INGRESO_REAL ~ CUOTA_EST_OP+
                  salOpDiaCoo004+
                  cuotaD053+
                  DEUDA_TOTAL_OP_OTROS+
                  salTotOpBCoo036+
                  DEUDA_TOTAL_SCE_24M+
                  PROM_DEUDA_TOTAL_SBS_OP_36M+
                  cuota052+
                  DEUDA_TOTAL_SBS_SC_24M+
                  PROM_DEUDA_TOTAL_SC_OP_36M+
                  r_DEUDA_TOTAL_SICOMsSCE_24M+
                  salPromD36M319+
                  salProm36M303+
                  LN_DEUDA_TOTAL_SCE_24M+
                  PROM_XVEN_SC_OP_36M+
                  cuotaCoo055+
                  salTotOp040+
```

```
                    maxMontoOp096+
                    DEUDA_TOTAL_OP_M+
                    DEUDA_TOTAL_SC_OP_24M+
                    PROM_XVEN_SBS_OP_36M+
                    MaxMontoOpD24M417+
                    salOpDia008+
                    SalTotOpD383+
                    cuotaEstimadaD24M416+
                    cuotaTotOp059,data = rmod_g3)

runtime <- system.time({
info[, INGRESO_EST_G3_5x100 := predict(rrlm_g3, newdata = info)]})
runtime_df <- data.frame(user = runtime[1],system = runtime[2],elapsed
    = runtime[3])

info[, RANGO_REALG3 := cut(INGRESO_REAL, breaks = c(450, 850, 1250,
    2000, 4000, 35000), labels = c("[450-850]", "(850-1250]", "
    (1250-2000]", "(2000-4000]", "(4000-35000]"))]

info[, RANGO_EST_G3_5x100 := cut(INGRESO_EST_G3_5x100, breaks = c(450,
    850, 1250, 2000, 4000, 35000), labels = c("[450-850]", "(850-1250]"
    , "(1250-2000]", "(2000-4000]", "(4000-35000]"))]
```

Código 3.29: Estracto esencial de código para obtener los Resultados del modelo RLM para G3

Métricas de MSE y MAE:

El MSE es aproximadamente de 5 millones de unidades en el modelo base como en el modelo remuestreado, el MAE es de 1300 en el modelo base y es de 1250 unidades en el remuestreado aproximadamente.

Los MSE de los rangos de Ingresos Real son muy similares en entrenamiento y validación.

Este modelo controla bien el porcentaje de error de predicción (tabla: 3.47 y 3.48).

Cruces y Sub/Sobre Estimaciones:

En el modelo base, el cruce es de 76%, la sobre estimación es de 21% y la sub estimación de 2%. En el modelo remuestreado el cruce aumenta al 81%, la sobre estimación disminuye al 16% y la sub estimación aumenta al 3%, valores aproximados.

Esto nos dice que el modelo conserva el porcentaje de: acierto, sobre y sub estimación.

Al analizar la amtriz de coincidencia, vemos que el modelo base no da una suerte de banda, pero con el remuestreo ya tenemos una matriz banda bastante buena (tabla: 3.47 y 3.48).

Tiempos de ejecución:

El tiempo de ejecución del modelo base es de 0.17 segundos; el tiempo del modelo remuestreado es de 0.36 segundo; no existe aumento del costo computacional al remuestrear (tabla: 3.47 y 3.48).

Chi Cuadrado de Pearson:

Como es conocido, se apoya en las métricas para colegir que no se rechaza la Hipótesis nula respecto a la similitud de la distribución de entrenamiento y validación (tabla: 3.47 y 3.48, figura 3.60).

Decisión:

Con base a lo expuesto, se decide por el modelo remuestreado para el grupo 3, dado que mejora el porcentaje de predicciones y el costo computacional es bajo.

CONSIDERACIONES: Nodos finales al ∼∼∼% - Sin aplicación de remuestreo
MODELO: Regresión Lineal Múltiple
VARIABLES: 29 **G3**

Muestra de entrenamiento

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1186	1223	3608	1065	7	7089.00	19%
(850-1250]	278	568	3492	2990	39	7367.00	20%
(1250-2000]	86	227	2699	4711	252	7975.00	21%
(2000-4000]	23	96	1738	5144	1399	8400.00	23%
(4000-35000]	10	18	528	2919	2834	6309.00	17%
						37140.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	17%	17%	51%	15%	0%	865426.26
(850-1250]	4%	8%	47%	41%	1%	1119023.33
(1250-2000]	1%	3%	34%	59%	3%	1138363.61
(2000-4000]	0%	1%	21%	61%	17%	1625098.75
(4000-35000]	0%	0%	8%	46%	45%	25988098.20

Entrenamiento: 0	Métricas		
MSE	5369404.0		
MAE	1305.2		
Cruce	33%	SubEstima	2%
Cruce +/-	77%	SobreEstim	21%

Muestra de Validación

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1215	1183	3767	1117	6	7288.00	20%
(850-1250]	300	557	3463	3090	47	7457.00	20%
(1250-2000]	77	231	2686	4712	279	7985.00	21%
(2000-4000]	19	102	1671	5186	1381	8359.00	22%
(4000-35000]	4	20	551	2877	2808	6260.00	17%
						37349.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	17%	16%	52%	15%	0%	887584.57
(850-1250]	4%	7%	46%	41%	1%	1169696.63
(1250-2000]	1%	3%	34%	59%	3%	1154660.93
(2000-4000]	0%	1%	20%	62%	17%	1683341.35
(4000-35000]	0%	0%	9%	46%	45%	28091312

Validación: 1	Métricas		
MSE	5685994.1		
MAE	1321.4		
Cruce	33%	SubEstima	2%
Cruce +/-	76%	SobreEstim	22%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	χ^2	Teórico
[450-850]	0.7	1.3	7.0	2.5	0.1	**32.7**	**26.3**
(850-1250]	1.7	0.2	0.2	3.3	1.6	**Decisión**	
(1250-2000]	0.9	0.1	0.1	0.0	2.9	Se rechaza H0	
(2000-4000]	0.7	0.4	2.6	0.3	0.2		
(4000-35000]	3.6	0.2	1.0	0.6	0.2		

Tiempo de Ejecución		
user	system	elapsed
0.17	0	0.17

Tabla 3.47: Resultados RLM sin remuestreo para G3

CONSIDERACIONES: Nodos finales al ~~~% -Con aplicación de remuestreo G3		
MODELO: Regresión Lineal Múltiple	G3_corte1 = 840 #19%	G3_porc1 = 0.35
VARIABLES: 29	G3_corte2 = 4000 #83%	G3_porc2 = 0.13

Muestra de entrenamiento

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1422	1849	3359	514	2	7146.00	19%
(850-1250]	341	972	4179	1863	21	7376.00	20%
(1250-2000]	96	502	3508	3716	153	7975.00	21%
(2000-4000]	31	270	2440	4531	1129	8401.00	23%
(4000-35000]	11	60	803	2903	2535	6312.00	17%
						37210.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	20%	26%	47%	7%	0%	571009.82
(850-1250]	5%	13%	57%	25%	0%	747037.30
(1250-2000]	1%	6%	44%	47%	2%	818078.10
(2000-4000]	0%	3%	29%	54%	13%	1573598.44
(4000-35000]	0%	1%	13%	46%	40%	27604269.30

Entrenamiento: 0	Métricas		
MSE	5429882.4		
MAE	1237.4		
Cruce	35%	SubEstima	3%
Cruce +/-	81%	SobreEstim	16%

Muestra de Validación

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	1461	1894	3472	518	2	7347.00	20%
(850-1250]	361	988	4164	1930	29	7472.00	20%
(1250-2000]	97	499	3476	3743	171	7986.00	21%
(2000-4000]	23	232	2409	4579	1118	8361.00	22%
(4000-35000]	9	69	818	2840	2523	6259.00	17%
						37425.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	20%	26%	47%	7%	0%	587552.77
(850-1250]	5%	13%	56%	26%	0%	788217.337
(1250-2000]	1%	6%	44%	47%	2%	821098.224
(2000-4000]	0%	3%	29%	55%	13%	1630324.74
(4000-35000]	0%	1%	13%	45%	40%	29759725.3

Validación: 1	Métricas		
MSE	5743686.1		
MAE	1250.8		
Cruce	35%	SubEstima	3%
Cruce +/-	80%	SobreEstim	16%

Chi Cuadrado de Pearson

Real	Estimado						
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	χ^2	Teórico
[450-850]	1.1	1.1	3.8	0.0	0.0	**27.4**	**26.3**
(850-1250]	1.2	0.3	0.1	2.4	3.0	**Decisión**	
(1250-2000]	0.0	0.0	0.3	0.2	2.1	Se rechaza H0	
(2000-4000]	2.1	5.3	0.4	0.5	0.1		
(4000-35000]	0.4	1.4	0.3	1.4	0.1		

Tiempo de Ejecución		
user	system	elapsed
0.17	0.19	0.36

Tabla 3.48: Resultados RLM con remuestreo para G3

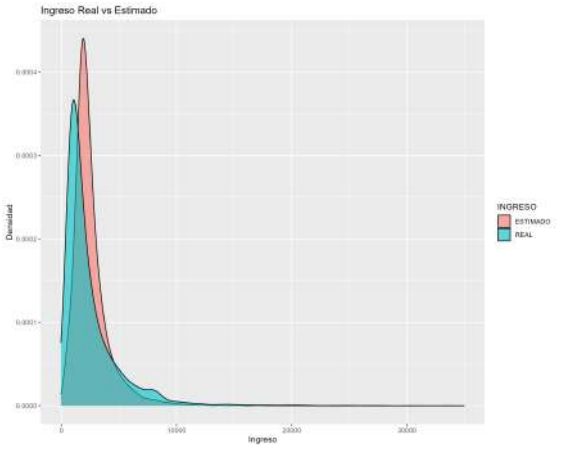


Figura 3.56: Sin remuestreo entrenamiento

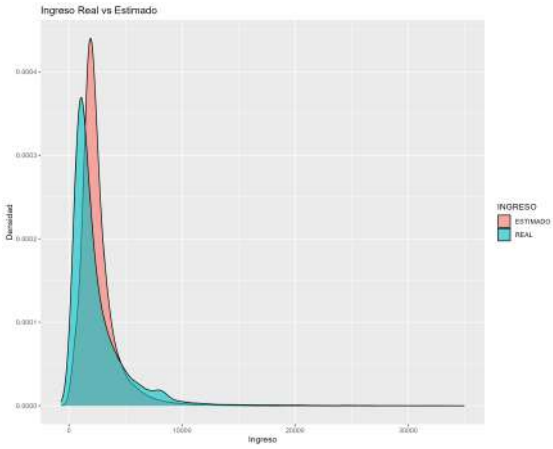


Figura 3.57: Sin remuestreo validación

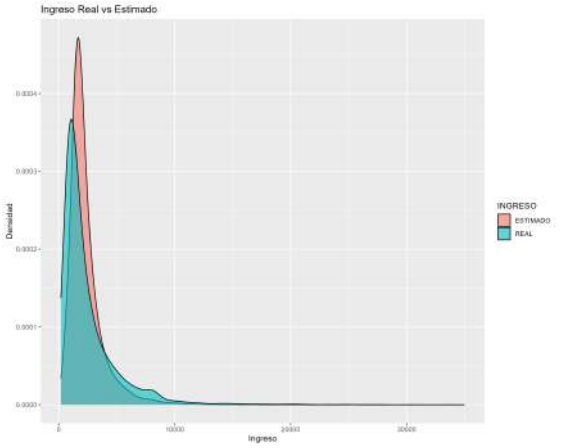


Figura 3.58: Con remuestreo entrenamiento

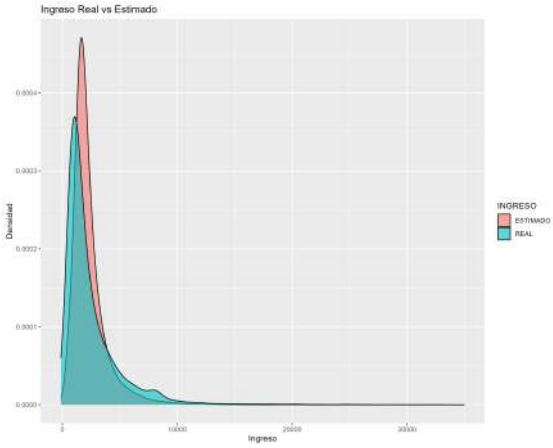


Figura 3.59: Con remuestreo validación

Figura 3.60: Distribuciones de Ingreso Real y Estimado del Modelo RLM en BDD entrenamiento y validación sin remuestreo y con remuestreo para G3

3.8.6 Elección del mejor modelo entre RLM, RF, GBM y XGB para G3

Para la sub población o grupo 3, también se selecciona como mejor modelo al GBM bajo las mismas justificaciones que para el grupo 1 y 2.

En este caso también el modelo de XGB es mucho más eficiente en predicción, rendimiento computacional, etc. pero su limitante de no estar disponible para windows nos persuade a no seleccionarlo como mejor modelo.

Capítulo 4

Discusión de resultados

En este capítulo, se expondrán las predicciones generadas por los modelos seleccionados: el Gradient Boosting Machine (GBM), para los grupos G1, G2 y G3. Estas predicciones se aplicarán sobre una nueva base de datos denominada <base de datos de comprobación> que consta de 84,877 individuos y 1,177 variables, con el objetivo de evaluar la aceptabilidad de las predicciones resultantes. Para ello, se llevará a cabo un análisis detallado de los resultados obtenidos.

Junto con las predicciones, se presentarán la matriz de coincidencias y la matriz de porcentajes correspondientes a las predicciones realizadas. Estos elementos permitirán una evaluación más profunda de la concordancia entre los valores de ingreso real y estimado, proporcionando una visión completa sobre la eficacia del modelo.

Además, se examinarán los indicadores de liquidez utilizando el modelo GBM, que fue elegido como el más adecuado contrastando los ingresos reales vs los ingresos estimados. Los indicadores de liquidez serán analizados con respecto a diversos criterios, incluyendo edad, estado civil, género, región y las provincias más grandes. Este análisis proporcionará una comprensión más completa de cómo el modelo GBM se comporta en diferentes segmentos de la población, ayudando a identificar posibles patrones y tendencias.

4.1 Evaluación del mejor modelo con BDD de Comprobación para grupo G1

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Gradient Boosting Machine
VARIABLES: 18 **G1**

Comprobación del modelo sin Remuestreo

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0	4	135	142	0	281	11%
(500-600]	2	7	140	201	0	350	14%
(600-750]	0	8	122	189	0	319	13%
(750-950]	1	11	142	253	0	407	16%
(950-250]	2	43	370	743	0	1158	46%
						2515	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	1%	48%	51%	0%	72317.31
(500-600]	1%	2%	40%	57%	0%	40072.37
(600-750]	0%	3%	38%	59%	0%	10639.82
(750-950]	0%	3%	35%	62%	0%	13480.81
(950-250]	0%	4%	32%	64%	0%	717502.23

Entrenamiento: 0	Métricas		
MSE	347552.6		
MAE	415.8		
Cruce	15%	SubEstima	17%
Cruce +/-	64%	SobreEstim	19%

Tiempo de Ejecución		
user	system	elapsed
0.1	0	0.1

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G1
MODELO: Gradient Boosting Machine
VARIABLES: 18
G1_corte1 = 484 #17% G1_porc1 = 0.40
G1_corte2 = 1250 #91% G1_porc2 = 0.30

Comprobación del modelo con Remuestreo

Real	Estimado					Total	%
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]		
[450-500]	0	4	15	213	49	281	11%
(500-600]	2	6	15	264	63	350	14%
(600-750]	1	9	17	228	64	319	13%
(750-950]	1	7	26	259	114	407	16%
(950-250]	6	31	66	619	436	1158	46%
						2515	

Real	Estimado					MSE
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]	
[450-500]	0%	1%	5%	76%	17%	166851
(500-600]	1%	2%	4%	75%	18%	117189
(600-750]	0%	3%	5%	71%	20%	56249
(750-950]	0%	2%	6%	64%	28%	20691
(950-250]	1%	3%	6%	53%	38%	537412

Comprobación: 1	Métricas		
MSE	292879		
MAE	404		
Cruce	29%	SubEstima	4%
Cruce +/-	69%	SobreEstim	27%

Tiempo de Ejecución		
user	system	elapsed
0.08	0	0.08

Tabla 4.1: Evaluación del modelo GBM en BDD de Comprobación para G1

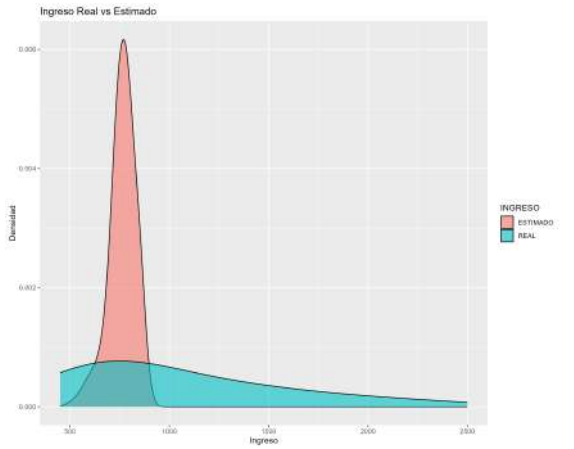


Figura 4.1: Modelo sin remuestreo

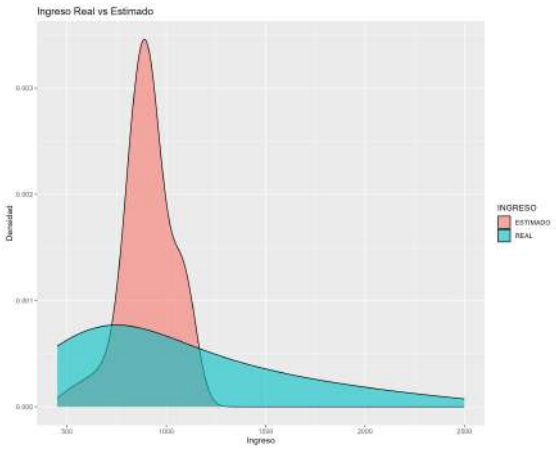


Figura 4.2: Modelo con remuestreo

Figura 4.3: Distribuciones de Ingreso Real y Estimado del Modelo GBM en BDD de Comprobación para G1

De las simulaciones se conocía que el grupo 1 tiene un compartiendo bastante difícil de capturar en el modelo. Si se hubiere modelo sin remuestreo, el porcentaje de predicción de aciertos sería del 64%, la sobre estimación 19% y la sub estimación del 17% con un costo computacional relativamente bajo. Al observar los resultados del modelo con remuestreo se tiene una cierta mejora, la predicción de cruces aumenta al 69%, la sobre estimación aumenta al 27% y la sub estimación disminuye al 4%, de igual forma con un costo computacional relativamente bajo.

Con el remuestreo se mejora la predicción en la cola derecha, es decir para individuos con ingresos altos; pero, se sobre estima para individuos con ingresos bajos.

El modelo tiene este comportamiento en las predicciones, dado que el comportamiento de los individuos no es homogenio por rango de ingreso real como se entrenó al modelo. En el rango de ingreso 950-2,500 se concentra el 46% de los individuos, por ello es que no se tiene predicciones adecuadas para individuos con ingresos bajos.

Estos resultados nos sugiere el que modelo para el grupo 3 tendrá mejores predicciones y una suerte de matriz banda en la matriz de coincidencias; es decir, se podría pensar que esta base de datos es adecuado predecirla con el modelo del grupo 3.

4.2 Evaluación del mejor modelo con BDD de Comprobación para grupo G2

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Gradient Boosting Machine
VARIABLES: 35 **G2**

Comprobación del modelo sin Remuestreo

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	13	70	737	385	0	1205.00	9%
(550-700]	13	106	1130	523	0	1772.00	13%
(700-900]	9	95	1367	897	0	2368.00	17%
(900-1300]	12	133	1426	1839	0	3410.00	25%
(1300-5000]	16	156	1566	3148	0	4886.00	36%
						13641.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	1%	6%	61%	32%	0%	144961.05
(550-700]	1%	6%	64%	30%	0%	68411.13
(700-900]	0%	4%	58%	38%	0%	22318.07
(900-1300]	0%	4%	42%	54%	0%	59930.63
(1300-5000]	0%	3%	32%	64%	0%	2205093.35

Entrenamiento: 0	Métricas		
MSE	830379.1		
MAE	563.47		
Cruce	24%	SubEstima	14%
Cruce +/-	74%	SobreEstim	12%

Tiempo de Ejecución		
user	system	elapsed
0.12	0	0.12

CONSIDERACIONES: Nodos finales al 3.6% - Con aplicación de remuestreo G2
MODELO: Gradient Boosting Machine
VARIABLES: 35
G2_corte1 = 601 # 33% G2_porc1 = 0.61
G2_corte2 = 1871 # 90% G2_porc2 = 0.35

Comprobación del modelo con Remuestreo

Real	Estimado					Total	%
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]		
[450-550]	29	60	618	329	169	1205.00	9%
(550-700]	41	77	927	555	172	1772.00	13%
(700-900]	37	74	1011	1049	197	2368.00	17%
(900-1300]	41	116	1024	1689	540	3410.00	25%
(1300-5000]	60	122	1133	1989	1582	4886.00	36%
						13641.00	

Real	Estimado					MSE
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]	
[450-550]	2%	5%	51%	27%	14%	337844.52
(550-700]	2%	4%	52%	31%	10%	192343.94
(700-900]	2%	3%	43%	44%	8%	96793.56
(900-1300]	1%	3%	30%	50%	16%	109911.25
(1300-5000]	1%	2%	23%	41%	32%	1770278.24

Comprobación: 1	Métricas		
MSE	733195.52		
MAE	547.59		
Cruce	32%	SubEstima	11%
Cruce +/-	74%	SobreEstim	15%

Tiempo de Ejecución		
user	system	elapsed
0.06	0	0.06

Tabla 4.2: Evaluación del modelo GBM en BDD de Comprobación para G2

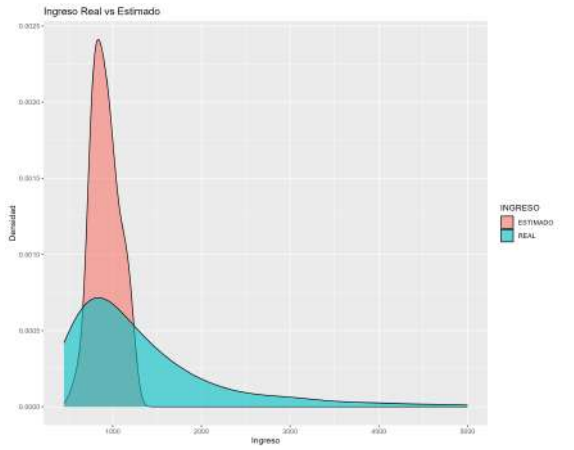


Figura 4.4: Modelo sin remuestreo

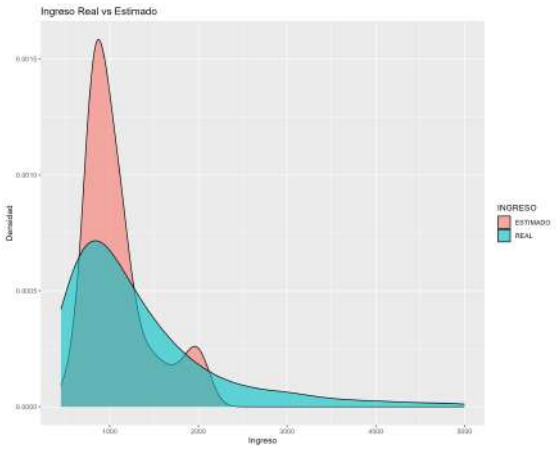


Figura 4.5: Modelo con remuestreo

Figura 4.6: Distribuciones de Ingreso Real y Estimado del Modelo GBM en BDD de Comprobación para G2

La Comprobación con el modelo base resulta un cruce de 75% de aciertos, una sobre estimación del 12% y una sub estimación del 14% con un costo computacional relativamente bajo. En el modelo con remuestreo el cruce se mantiene, la sobre estimación aumenta al 15% y la sub estimación disminuye al 11% con un costo computación también relativamente bajo.

Al analizar la matriz de coincidencia se ve que se tiene predicciones para la columna 5 de ingresos estimados, lo cual no se tenia con el modelo base.

Estos resultados de porcentajes de aciertos, sobre y sub estimación similares y mejora en la estimación de la cola derecha dice que si fue una buena estrategia el remuestrear las base de datos. Ciertamente no se obtiene la matriz banda deseada, pero esto se debe al comportamiento peculiar de esta nueva base de datos que concentra el mayor número de individuos en el grupo de ingreso real 1,300 al 5,000 USD.

4.3 Evaluación del mejor modelo con BDD de Comprobación para grupo G3

CONSIDERACIONES: Nodos finales al 3.6% - Sin aplicación de remuestreo
MODELO: Gradient Boosting Machine
VARIABLES: 29 **G3**

Comprobación del modelo sin Remuestreo

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	81	631	335	28	0	1075.00	7%
(850-1250]	61	899	954	174	0	2088.00	14%
(1250-2000]	88	1000	1976	1281	67	4412.00	29%
(2000-4000]	38	501	1359	2135	649	4682.00	30%
(4000-35000]	15	188	568	1208	1141	3120.00	20%
						15377.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	8%	59%	31%	3%	0%	370042.62
(850-1250]	3%	43%	46%	8%	0%	268334.19
(1250-2000]	2%	23%	45%	29%	2%	564124.35
(2000-4000]	1%	11%	29%	46%	14%	1541701.66
(4000-35000]	0%	6%	18%	39%	37%	40777462.50

Entrenamiento: 0	Métricas		
MSE	8967348.9		
MAE	1464.9		
Cruce	41%	SubEstima	9%
Cruce +/-	87%	SobreEstim	4%

Tiempo de Ejecución		
user	system	elapsed
0.13	0	0.13

CONSIDERACIONES: Nodos finales al 3.6% -Con aplicación de remuestreo G3
MODELO: Gradient Boosting Machine G3_corte1 = 840 #19% G3_porc1 = 0.30
VARIABLES: 29 G3_corte2 = 7060 #94% G3_porc2 = 0.30

Comprobación del modelo con Remuestreo

Real	Estimado					Total	%
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]		
[450-850]	76	552	414	33	0	1075.00	7%
(850-1250]	54	763	1021	249	1	2088.00	14%
(1250-2000]	78	807	1937	1505	85	4412.00	29%
(2000-4000]	37	392	1230	2291	732	4682.00	30%
(4000-35000]	12	147	502	1268	1191	3120.00	20%
						15377.00	

Real	Estimado					MSE
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]	
[450-850]	7%	51%	39%	3%	0%	466340.78
(850-1250]	3%	37%	49%	12%	0%	353502.922
(1250-2000]	2%	18%	44%	34%	2%	608147.046
(2000-4000]	1%	8%	26%	49%	16%	1525379.87
(4000-35000]	0%	5%	16%	41%	38%	39298741.3

Comprobación: 1	Métricas		
MSE	8693274.2		
MAE	1453.4		
Cruce	41%	SubEstima	8%
Cruce +/-	87%	SobreEstim	5%

Tiempo de Ejecución		
user	system	elapsed
0.11	0	0.11

Tabla 4.3: Evaluación del modelo GBM en BDD de Comprobación para G3

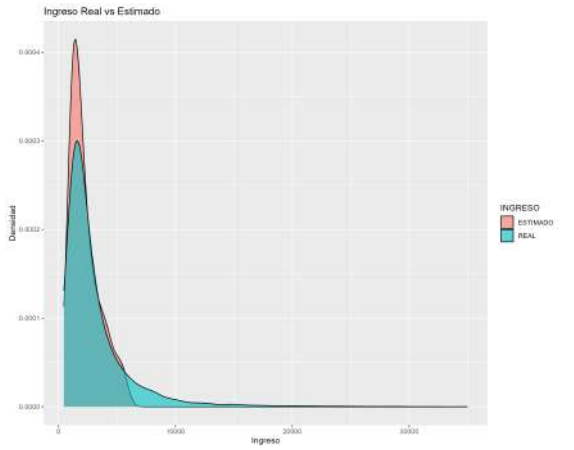


Figura 4.7: Modelo sin remuestreo

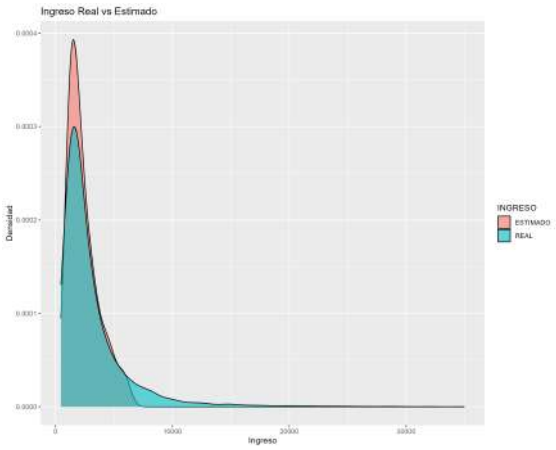


Figura 4.8: Modelo con remuestreo

Figura 4.9: Distribuciones de Ingreso Real y Estimado del Modelo GBM en BDD de Comprobación para G3

Como se esperaba, por los pésimos resultados de predicción en el grupo 1 y 2, este modelo resulta con mejores predicciones, pues, se puede ver que la individuos son más homogenios por rango de ingreso real.

El modelo base tiene un cruce del 87%, una sobre estimación del 4% y una sub estimación del 9% con un costo computacional relativamente bajo; estos resultados logra conservarse con el modelo con remuestreo, mejoran el por porcentaje de sub estimación al 8% y el aumento de la sobre estimación en 1%, de igual forma el costo computacional es relativamente bajo.

En esta evaluación, observamos una especie de "matriz banda" que refleja predicciones altamente precisas, donde la estimación del ingreso se alinea estrechamente con los valores reales. Sin embargo, en la columna correspondiente a un rango de ingresos estimados de 450 a 550 USD, notamos una presencia mínima de predicciones. Esto se debe a que esta base de datos contiene principalmente individuos con ingresos muy elevados en esa categoría. Al analizar la gráfica de la función de densidad, vemos que son bastante similares entre el ingreso real y estimado.

Los resultados nos indican que si fue buena estrategia el remuestrear la base de datos.

4.4 Indicadores de Liquidez

General				
Edad	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
[18,30)	11%	11%	13%	13%
[30,40)	28%	28%	25%	25%
[40,55)	36%	36%	34%	34%
[55,65)	16%	16%	16%	16%
[65,100)	9%	9%	11%	11%

G1				
Edad	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
[18,30)	20%	20%	16%	16%
[30,40)	29%	29%	21%	21%
[40,55)	31%	31%	30%	30%
[55,65)	12%	12%	18%	18%
[65,100)	8%	8%	17%	17%

E Civil	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Soltero	42%	42%	34%	34%
Casado	48%	48%	55%	55%
Divorciado	7%	7%	8%	8%
Viudo	3%	3%	3%	3%
En union de hecho	0%	0%	0%	0%

E Civil	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Soltero	60%	60%	37%	37%
Casado	33%	33%	51%	51%
Divorciado	5%	5%	8%	8%
Viudo	2%	2%	4%	4%
En union de hecho	0%	0%	0%	0%

Genero	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Masculino	57%	57%	55%	55%
Femenino	43%	43%	45%	45%

Genero	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Masculino	55%	55%	52%	52%
Femenino	45%	45%	48%	48%

Region	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Amazonia	11%	11%	11%	11%
Costa	28%	28%	8%	8%
Sierra	61%	61%	81%	81%

Region	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Amazonia	8%	8%	9%	9%
Costa	53%	53%	10%	10%
Sierra	39%	39%	81%	81%

Provincia	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Bolivar	2%	2%	3%	3%
Cotopaxi	7%	7%	9%	9%
Guayas	10%	10%	5%	5%
Pichincha	64%	64%	18%	18%
Tungurahua	13%	13%	13%	13%
Otros	4%	4%	52%	52%

Provincia	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Bolivar	1%	1%	3%	3%
Cotopaxi	3%	3%	5%	5%
Guayas	23%	23%	7%	7%
Pichincha	60%	60%	14%	14%
Tungurahua	10%	10%	17%	17%
Otros	3%	3%	54%	54%

G2				
Edad	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
[18,30)	20%	20%	16%	16%
[30,40)	29%	29%	26%	26%
[40,55)	31%	31%	31%	31%
[55,65)	13%	13%	15%	15%
[65,100)	8%	8%	11%	11%

G3				
Edad	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
[18,30)	11%	11%	9%	9%
[30,40)	28%	28%	27%	27%
[40,55)	36%	36%	38%	38%
[55,65)	16%	16%	17%	17%
[65,100)	9%	9%	10%	10%

E Civil	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Soltero	55%	55%	39%	39%
Casado	37%	37%	51%	51%
Divorciado	6%	6%	8%	8%
Viudo	2%	2%	3%	3%
En union de hecho	0%	0%	0%	0%

E Civil	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Soltero	42%	42%	29%	29%
Casado	48%	48%	60%	60%
Divorciado	7%	7%	8%	8%
Viudo	3%	3%	3%	3%
En union de hecho	0%	0%	0%	0%

Genero	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Masculino	55%	55%	53%	53%
Femenino	45%	45%	47%	47%

Genero	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Masculino	57%	57%	58%	58%
Femenino	43%	43%	42%	42%

Region	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Amazonia	9%	9%	11%	11%
Costa	39%	39%	7%	7%
Sierra	53%	53%	82%	82%

Region	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Amazonia	11%	11%	11%	11%
Costa	28%	28%	9%	9%
Sierra	61%	61%	80%	80%

Provincia	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Bolivar	2%	2%	4%	4%
Cotopaxi	5%	5%	9%	9%
Guayas	14%	14%	5%	5%
Pichincha	64%	64%	15%	15%
Tungurahua	11%	11%	12%	12%
Otros	3%	3%	56%	56%

Provincia	Modelamiento		Validación	
	I Real	I Estimado	I Real	I Estimado
Bolivar	2%	2%	3%	3%
Cotopaxi	7%	7%	10%	10%
Guayas	10%	10%	6%	6%
Pichincha	64%	64%	20%	20%
Tungurahua	13%	13%	11%	11%
Otros	4%	4%	50%	50%

Tabla 4.4: Indicadores de Liquidez General, G1, G2 y G3

Figura 4.10: Indicadores de Liquidez General, G1, G2 y G3

En esta sección, a la base de datos de comprobación lo denominaremos validación, esto no causa confusión.

Edad:

Se puede ver que la Base de datos de Modelamiento consta de 13.15% de individuos entre 18-30 años, 25.45% de individuos entre 30-40 años, 33.98% de individuos entre 40-55 años, 16.33% de individuos entre 55-65 años y 11.09% de individuos mayor a 65 años. Siendo los que frisan entre las edad 30-55 años los que más ganan; se puede ver que a mayor edad se percibe mayor ingreso, esto es coherente dado que a mayor edad mayor experiencia y por tanto puestos más remunerados. Este mismo comportamiento capta el modelo y se tiene el comportamiento similar.

En la base de datos de Validación se tiene el mismo comportamiento tanto para Ingreso Real como para el ingreso estimado.

En los grupos 1, 2 y 3 se tiene el mismo comportamiento (Figura de Estadísticas 4.10).

Estado Civil:

Los individuos cuyo estado civil es casado, son quienes más ingresos real perciben, seguido de los solteros, los de la categorías divorciado, viudo y en unión de hecho perciben ingresos muy bajos. Esta característica es capturada por el modelo y en los ingresos estimados se observa el mismo comportamiento.

La base de datos de Validación sigue este mismo comportamiento, tanto para el ingreso real y estimado.

En los grupos 1, 2 y 3 se tiene el mismo comportamiento (Figura de Estadísticas 4.10).

Género:

Los individuos cuyo género es masculino son quienes más ingresos perciben; la diferencia de ingresos es de aproximadamente 14%. El modelo logra capturar esta característica y el ingreso estimado tiene el mismo comportamiento.

En la base de datos de Validación se tiene similar comportamiento, tanto para el ingreso real y estimado.

En los grupos 1, 2 y 3 se tiene el mismo comportamiento (Figura de

Estadísticas 4.10).

Región:

Los individuos de la región Sierra son quienes más ingresos perciben, estando sobre el 60% de los ingresos más altos, seguido de la región costa con 28% y la Amazonia con 11%. Este comportamiento es capturado por el modelo y se observa el mismo patrón en los ingresos estimados.

En la base de datos de Validación se tiene un comportamiento similar, siendo la región siendo con 80% de ingresos más altos, la región Costa con 8% y la Amazonia con 11%.

En el Grupo 1 no se tiene este comportamiento, los individuos de la región Costa son quienes más Ingreso Perciben, con 53%, seguido por la región Sierra con el 39% y la Amazonia con 8%. El modelo captura este comportamiento y la estimación sigue el mismo patrón; en base de datos de validación los individuos de la región sierra son quienes más ingresos perciben y el modelo estima adecuadamente el comportamiento del ingreso.

El grupo 2 y 3 tienen un comportamiento similar al análisis del comportamiento general, tanto en Modelamiento y Validación (Figura de Estadísticas 4.10).

Provincia:

En la base de datos de Modelamiento, los individuos que residen en la provincia de Pichincha son quienes más ingresos perciben siendo 64% de todo el ingreso, seguido por por Tunguarahua con el 13%, Guayas con el 10%, etc. Este comportamiento es capturado por el modelo y la estimación del ingreso sigue este patrón. Se puede ver que para los grupos 1, 2 y 3 se tiene resultados y comportamientos similares.

En la base de datos de Validación, los individuos que residen en otras provincias que no sean Pichincha, Tungurahua, Guayas, Bolivar y Cotopaxi son quienes más ingresos perciben, siendo el 52% de todo el ingreso, seguido por Pichincha con 18%, Tungurahua 13%; el modelo estima correctamente este comportamiento en los ingresos; en los grupos 1, 2 y 3 se tiene un patrón similar (Figura de Estadísticas 4.10).

Capítulo 5

Conclusiones y recomendaciones

5.1 Conclusiones

1. Para alcanzar la estimación exitosa de la capacidad de pago de individuos bancarizados con cuota estimada, se procedió a formular y desarrollar diversos modelos, incluyendo tanto enfoques no paramétricos como un modelo paramétrico. Esta formulación y desarrollo se basó en el objetivo general de estimar la capacidad de pago de personas naturales bancarizadas con cuota estimada a través de métodos estadísticos. Las iteraciones y simulaciones desplegadas abordaron diferentes perspectivas y consideraron variados enfoques de estudio, siendo crucial el adecuado tratamiento aplicado a la base de datos original.

2. Un objetivo específico de este estudio era desarrollar modelos estadísticos no paramétricos basados en el árboles de decisión que abordará tanto la sobreestimación como la subestimación de la capacidad de pago en personas naturales bancarizadas con diferentes niveles de capacidad financiera. Se basó en la información recopilada en el sistema de registro crediticio.

   Para lograrlo, se construyeron los siguientes modelos: Random Forest (RF), Gradient Boosting Machine (GBM), XGBoost (XGB) y Regresión Lineal Multivariante (RLM). Entre ellos, los modelos no paramétri-

cos (RF, GBM y XGB) demostraron una capacidad de estimación de ingresos superior al modelo paramétrico RLM.

La población de estudio se dividió en tres grupos según el monto de la cuota de crédito individual: el primer grupo con cuotas de hasta 107 USD, el segundo grupo con cuotas superiores a 107 USD pero no más de 435 USD, y el tercer grupo con cuotas superiores a 435 USD.

En el análisis, se observó que el primer grupo presentaba un comportamiento difícil de capturar mediante los modelos. Sin embargo, con la aplicación de técnicas de remuestreo, fue posible mejorar las predicciones en cierta medida. En el segundo grupo, cuyo comportamiento era menos complicado, las predicciones mejoraron significativamente con el modelo remuestreado. El tercer grupo, caracterizado por atributos más definidos, permitió que el modelo capturara patrones efectivamente, resultando en pronósticos precisos tanto para ingresos bajos como altos.

Al evaluar y validar los modelos utilizando una base de datos independiente, se observaron predicciones menos precisas para los grupos 1 y 2. Esto se debió a que los patrones de esta base de datos eran más similares a los del tercer grupo. En consecuencia, el modelo desarrollado para este último grupo logró destacarse en sus pronósticos.

Es decir, este estudio cumplió su objetivo específico de desarrollar un modelo predictivo basado en árboles de decisión capaz de mejorar la estimación de la capacidad de pago en personas naturales bancarizadas con diferentes niveles de ingresos y cuotas de crédito.

3. Uno de los objetivos específicos trazados en este contexto fue el de obtener modelos analíticos capaces de abordar de manera efectiva la sobre-estimación de ingresos en clientes con capacidades de pago bajas, así como la sub-estimación de ingresos en aquellos con capacidades de pago altas.

   Para cumplir con esta meta, se procedió a implementar diversos modelos no paramétricos, los cuales mostraron resultados prometedores en cuanto a la precisión en las estimaciones de ingresos. En-

tre los modelos no paramétricos explorados, se destacó el Gradient Boosting Machine (GBM) como el más eficiente en la predicción.

La técnica del remuestreo desempeñó un papel crucial en la mejora de los niveles de predicción, especialmente en las colas de la distribución de ingresos. Es decir, este enfoque resultó fundamental para lograr una mejor identificación de individuos con ingresos extremadamente bajos o extremadamente altos.

4. El modelo que demostró ser más eficiente tanto desde una perspectiva matemática como en términos de su implementación algorítmica fue el XGBoost (XGB). Sin embargo, nos encontramos con una complicación en su implementación en el entorno R, donde su rendimiento resultó ser notablemente ineficiente. Debido a esta limitación, optamos por emplear el cluster H2O para lograr una emulación satisfactoria del modelo XGBoost.

   Además, nos tropezamos con otra dificultad; aunque H2O ofrecía varios modelos debidamente optimizados, lamentablemente el XGBoost no estaba disponible para el sistema operativo Windows, a diferencia de las versiones compatibles con macOS y Linux. Esta particularidad generó una serie de obstáculos, ya que nos vimos obligados a llevar a cabo su implementación en un entorno Linux, junto con sus complementos necesarios, para asegurar su funcionamiento sin contratiempos. A pesar de ser el modelo más sobresaliente con un rendimiento computacional excepcional, esta limitante representó un inconveniente significativo, lo cual nos llevó a descartarlo como la elección principal.

5. El remuestreo se ha mostrado como una herramienta valiosa para mejorar la precisión de los modelos predictivos. Sin embargo, esta ventaja se hace realmente notable cuando el investigador crea y ejecuta su propio método de remuestreo equilibrado. Simplemente aplicar la técnica de bootstrap sin una estructura clara no produce resultados significativos. En nuestro caso, desarrollar y aplicar nuestra propia técnica de remuestreo, específicamente diseñada para abordar ambos extremos de los datos, resultó ser una estrategia altamente efectiva y beneficiosa.

Es precisamente esta razón la que respalda la idea de que los modelos construidos a partir de datos remuestreados arrojaron resultados considerablemente mejores en comparación con los modelos creados sin utilizar remuestreo. Esta ventaja se mantuvo constante tanto para modelos que asumen ciertas distribuciones (paramétricos) como para aquellos que no lo hacen (no paramétricos).

Además, es relevante destacar que el proceso de remuestreo no aumentó significativamente los costos computacionales. Este logro subraya aún más la viabilidad y eficacia de esta estrategia en la mejora de la precisión de los modelos predictivos.

6. La aplicación precisa y eficiente del proceso de remuestreo fue precedida por una etapa crucial: la determinación de la representatividad o peso adecuado asignado a las colas de la distribución. Esta determinación, fijada en un 3.6%, se reveló como una decisión fundamental. Al establecer este peso específico, se logra un control efectivo del proceso de remuestreo al definir cortes y proporciones óptimas. Esto a su vez permite regular de manera precisa la migración de individuos desde el centro de la distribución hacia sus extremos.

   En este contexto, el empleo de un grid de hiperparámetros demostró ser una herramienta excepcional para la obtención del valor del 3.6%. Esta técnica no solo facilitó la identificación de la representatividad óptima, sino que también contribuyó en la determinación del número de árboles de decisión más adecuado para los modelos. La conjunción de estas estrategias permitió no solo mejorar la calidad predictiva de los modelos, sino también optimizar el rendimiento global del proceso.

7. Se formuló los Indicadores de Liquidez con las características de edad, estado civil, género, región y provincia asumiendo que están fuertemente relacionadas con los niveles de ingresos de los individuos. Los modelos construidos capturan estos patrones y reflejan adecuadamente estas relaciones en las estimaciones de ingresos tanto en la fase de modelamiento como en la validación. Estos patrones son consistentes en los tres grupos analizados en términos de cuota de crédito.

- Edad: La edad se correlaciona positivamente con los ingresos. Las personas en el rango de edad de 30 a 55 años presentan los ingresos más altos, lo que es coherente con su experiencia laboral y posiciones mejor remuneradas.
- Estado Civil: Los individuos con estado civil casado tienden a tener los ingresos más altos, seguidos por aquellos que son solteros. En cambio, las categorías divorciado, viudo y en union de hecho muestran ingresos más bajos.
- Género: Los individuos de género masculino tienden a ganar más que otros géneros, con una diferencia de alrededor del 14%.
- Región: Los individuos en la región Sierra presentan los ingresos más altos, seguidos por la región Costa y luego la región Amazonia.
- Provincia: En el modelamiento, los individuos de la provincia de Pichincha presentan los ingresos más altos, seguidos por Tungurahua y Guayas.

5.2 Recomendaciones

1. Para mejorar la estimaciones en los grupos 1 y 2 se recomienda estudiar a profundidad, como proyectos independientes debido a su complejidad, los ajustes óptimos de los hiper parámetros de los modelos no paramétricos; aunque suena simple, esto es realmente importante para hacer que los modelos capturen de mejor manera el comportamiento de los individuos en dichos grupos.

   Encontrar estos ajustes ideales no solo hace que las estimaciones sean más precisas, sino que también ayuda a entender mejor cómo funcionan realmente los grupos 1 y 2.

2. Para optimizar el costo computacional de los modelos, una estrategia recomendable es migrar su implementación al clúster H2O. Este enfoque se justifica porque H2O está diseñado específicamente para tareas de machine learning y el manejo eficiente de conjuntos de datos extensos. En contraste, la construcción de modelos en R a menudo implica considerables demandas computacionales, además

de limitaciones en términos de escalabilidad para su implementación en entornos de producción.

El aprovechamiento de la plataforma H2O ofrece ventajas significativas en términos de rendimiento y escalabilidad. Al estar concebida para el procesamiento de datos a gran escala, permite ejecutar modelos de manera más eficiente, lo que resulta en un costo computacional reducido y tiempos de respuesta más rápidos. Esta elección también simplifica la transición de los modelos desde la etapa de desarrollo hasta la implementación en producción, lo que contribuye a un flujo de trabajo más fluido y coherente.

3. Se recomienda evitar una percepción dicotómica de superioridad o inferioridad entre los modelos estadísticos no paramétricos y los paramétricos, ya que su eficacia depende directamente de los objetivos específicos del estudio en cuestión. En nuestra investigación, donde las relaciones entre las variables no seguían patrones lineales, los modelos no paramétricos demostraron ser más adecuados para capturar la complejidad de los datos.

   Los modelos no paramétricos demostraron su idoneidad al lidiar con relaciones no lineales y variables predictoras altamente interdependientes. En contraste, el modelo paramétrico de Regresión Lineal Múltiple (RLM) no pudo ajustarse de manera eficaz a estas condiciones. Sin embargo, es relevante señalar que el RLM desempeña un papel crucial en contextos como los modelos econométricos, donde las suposiciones de linealidad y relaciones específicas son fundamentales.

Referencias bibliográficas

[1] Vance W. Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Encyclopedia of Statistics in Behavioral Science, © John Wiley Sons, Ltd*, 2005. Tomado de: https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781118445112.stat06558.

[2] Tianqi Chen. Introduction to boosted trees, 2014. Tomado de: https://web.njit.edu/~usman/courses/cs675_fall16/BoostedTree.pdf.

[3] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Arxiv*, 2016. Tomado de: https://arxiv.org/abs/1603.02754.

[4] Santiago de la Fuente Fernández. Aplicaciones de la chi-cuadrado: Tablas de contingencia, homogeneidad, dependencia e independencia. *Universidad Autónoma de Madrid*, 2016. Tomado de: https://www.fuenterrebollo.com/Aeronautica2016/contingencia.pdf.

[5] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *JSTOR*, 1999. Tomado de: https://www.jstor.org/stable/2699986.

[6] Elshaddai Harris. Understanding weight of evidence and information value, 2022. Tomado de: https://www.linkedin.com/pulse/understanding-weight-evidence-information-value-elshaddai-harris/.

[7] Kruthika Kulkarni. Understand weight of evidence and information value, 2021. Tomado de: https://www.analyticsvidhya.com/blog/2021/06/understand-weight-of-evidence-and-information-value/.

[8] Gilles Louppe. Understanding random forests: From theory to practice. ResearchGate; SourcearXiv, October 2014. Thesis for: PhDAdvisor: Pierre Geurts, URL: https://www.researchgate.net/publication/264312332_Understanding_Random_Forests_From_Theory_to_Practice.

[9] Tomonori Masui. All you need to know about gradient boosting algorithm part 1. regression. Medium; Towards Data Science, January 2022. URL: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502.

[10] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers*, 2013. Tomado de: https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial.

[11] User418251 Noah. Two-sample kolmogorov-smirnov test. Mathematics Stack Exchange, 2020. URL: https://math.stackexchange.com/questions/3577453/two-sample-kolmogorov-smirnov-test.

[12] Ajit Samudrala. Unveiling mathematics behind xgboost, 2018. Tomado de: https://medium.com/@samudralaajit/unveiling-mathematics-behind-xgboost-c7f1b8201e2a.

[13] Eric Towers. Two-sample kolmogorov-smirnov test, 2020. Tomado de: https://math.stackexchange.com/questions/3577453/two-sample-kolmogorov-smirnov-test.

[14] Wikipedia. Kolmogorov–smirnov test, 2023. Tomado de: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test.

Capítulo A

Anexos

En esta sección se presentan resultados relevantes que no se colocó en el cuerpo principal del documento debido a su extensión, pero que desempeñan un papel fundamental en la formulación de conclusiones sólidas. Estos resultados, que abarcan diversas facetas del estudio, han contribuido de manera significativa a la obtención de conclusiones claras y enriquecedoras.

Entre los aspectos destacados se encuentran las Tablas de KS y VI, los grids de hiperparámetros aplicados a los grupos G1, G2 y G3, los Indicadores de Liquidez reales y estimados, que aportan un análisis cuantitativo y detallado de la situación financiera en las bases de datos de Modelamiento y Validación para los grupos mencionados; la sección también incluye Gráficas de Indicadores de liquidez que enriquecen la comprensión visual de los resultados.

Para aquellos interesados en explorar más allá, se proporciona un enlace al repositorio del proyecto en GitLab: https://gitlab.com/jaimerault.jrt/tic. Aquí, se encuentra el código necesario para replicar integralmente el proyecto y obtener los resultados presentados en este trabajo. Esta fuente adicional de información permite una verificación transparente y una exploración en profundidad de los procesos subyacentes.

A.1 Test KS y VI

Variables Cuantitativas: Test KS

	Grupo N1 Variable	KS	Grupo N2 Variable	KS	Grupo N3 Variable	KS
1	DEUDA_TOTAL_OP_OTROS	22.6%	DEUDA_TOTAL_SICOM_OP_24M	29.5%	cuotaD053	38.4%
2	DEUDA_TOTAL_SICOM_OP_24M	30.0%	PROM_DEUDA_TOTAL_SICOM_OP_24M	29.3%	DEUDA_TOTAL_SCE_24M	36.8%
3	NumAcreedoresDDCom404	24.2%	r_PROM_XVEN_SICOM_OP_24s36M	28.2%	CUOTA_EST_OP	38.8%
4	numMesesInfoCredBanCoopD36M421	23.0%	r_PROM_DEUDA_TOTAL_SICOM_OP_24s36M	29.7%	cuota052	39.5%
5	PROM_DEUDA_TOTAL_SICOM_OP_24M	29.7%	r_DEUDA_TOTAL_SICOMsSCE_24M	30.5%	DEUDA_TOTAL_SBS_SC_24M	37.1%
6	PROM_XVEN_SICOM_OP_24M	28.4%	ANTIGUEDAD_OP_SICOM	28.5%	PROM_DEUDA_TOTAL_SC_OP_36M	27.8%
7	r_DEUDA_TOTAL_SICOMsSCE_24M	30.0%	r_NOPE_APERT_SICOMsSCE_OP_36M	26.1%	salPromD36M319	32.4%
8	r_NOPE_APERT_SICOM_OP_24s36M	22.3%	r_NOPE_APERT_SICOM_OP_24s36M	22.0%	salProm36M303	33.0%
9	r_PROM_DEUDA_TOTAL_SICOM_OP_24s36M	30.2%	R_INGRESOS	65.0%	LN_DEUDA_TOTAL_SCE_24M	36.8%
10	r_PROM_XVEN_SICOM_OP_24s36M	29.1%	CUOTA_EST_OP	22.4%	MaxMontoOpD24M417	34.4%
11	DEUDA_TOTAL_SBS_SC_24M	21.7%	numMesesInfoCredBanCoopD36M421	22.6%	cuotaEstimadaD24M416	32.0%
12	NOPE_APERT_SICOM_OP_36M	26.3%	MaxMontoOpD24M417	24.1%	salOpDiaCoo004	21.9%
13	NOPE_TOTAL_OP_M	20.1%	DEUDA_TOTAL_SCE_24M	25.9%	DEUDA_TOTAL_OP_OTROS	22.3%
14	r_DEUDA_TOTAL_SICOM_OP_12s24M	27.2%	NOPE_APERT_SICOM_OP_36M	26.0%	salTotOpBCoo036	21.9%
15	R_INGRESOS	35.6%	PROM_XVEN_SICOM_OP_24M	27.7%	PROM_DEUDA_TOTAL_SBS_OP_36M	20.6%
16	r_NOPE_APERT_SICOMsSCE_OP_36M	26.9%	DEUDA_TOTAL_OP_OTROS	26.7%	r_DEUDA_TOTAL_SICOMsSCE_24M	20.2%
17	salOpDiaCom005	24.3%	LN_DEUDA_TOTAL_SCE_24M	25.9%	PROM_XVEN_SC_OP_36M	27.8%
18	salTotOpCom037	24.5%	r_DEUDA_TOTAL_SICOM_OP_12s24M	26.4%	cuotaCoo055	27.4%
19	DEUDA_TOTAL_SICOM_OP_3M	25.3%	DEUDA_TOTAL_SBS_SC_24M	27.3%	salTotOp040	29.8%
20	NOPE_APERT_SICOM_OP_24M	22.3%	NOPE_TOTAL_OP_OTROS	23.6%	maxMontoOp096	33.3%
21	PROM_DEUDA_TOTAL_SICOM_OP_6M	25.9%	r_NOPE_APERT_SICOMsSCE_OP_24M	22.0%	DEUDA_TOTAL_OP_M	23.0%
22	PROM_XVEN_SICOM_OP_12M	26.3%	NOPE_APERT_SICOM_OP_24M	21.9%	DEUDA_TOTAL_SC_OP_24M	26.3%
23	PROM_XVEN_SICOM_OP_36M	28.0%	NumAcreedoresDDCom404	21.9%	PROM_XVEN_SBS_OP_36M	20.7%
24	PROM_XVEN_SICOM_OP_3M	25.2%	cuotaCom056	21.9%	salOpDia008	29.7%
25	PROM_XVEN_SICOM_OP_6M	25.2%	salTotOpCom037	21.9%	SalTotOpD383	31.6%
26	r_DEUDA_TOTAL_SICOM_OP_6s12M	25.8%	DEUDA_TOTAL_SCE_6M	21.9%	MaxCalTC	26.0%
27	r_DEUDA_TOTAL_SICOMsSCE_12M	26.9%	LN_DEUDA_TOTAL_SCE_6M	21.9%	R_INGRESOS	98.2%
28	r_DEUDA_TOTAL_SICOMsSCE_3M	24.2%	salOpDiaCom005	21.8%	cuotaTotOp059	40.2%
29	r_DEUDA_TOTAL_SICOMsSCE_6M	25.2%	maxMontoOp096	20.8%	cuotaTCS374	38.4%
30	r_NOPE_APERT_SICOMsSCE_OP_24M	22.6%	salPromD36M319	20.8%	LN_DEUDA_TOTAL_SCE_3M	33.2%
31	r_PROM_DEUDA_TOTAL_SICOM_OP_12s24M	27.2%	DEUDA_TOTAL_SCE_3M	20.7%	RANGO_INGRESO_G2	63.8%
32	r_PROM_DEUDA_TOTAL_SICOM_OP_6s12M	25.8%	LN_DEUDA_TOTAL_SCE_3M	20.7%	PROM_XVEN_SC_OP_3M	23.9%
33	r_PROM_XVEN_SICOM_OP_12s24M	25.9%	cuotaEstimadaD24M416	20.7%	DEUDA_TOTAL_SC_OP_3M	23.9%
34	r_PROM_XVEN_SICOM_OP_12s36M	25.9%	salProm36M303	20.5%	DEUDA_TOTAL_SCE_3M	33.2%
35	r_PROM_XVEN_SICOM_OP_6s24M	25.2%	SalTotOpD383	20.0%	DEUDA_TOTAL_SCE_12M	35.9%
36	RANGO_INGRESO_G2	69.2%	r_PROM_DEUDA_TOTAL_SICOM_OP_3s12M	22.3%	PROM_XVEN_SC_OP_12M	25.5%
37	RANGO_INGRESO_G3	34.5%	DEUDA_TOTAL_SICOM_OP_3M	22.4%	DEUDA_TOTAL_SCE_6M	34.1%
38	ANTIGUEDAD_OP_SICOM	27.5%	r_PROM_DEUDA_TOTAL_SICOM_OP_3s6M	22.3%	LN_DEUDA_TOTAL_SCE_6M	34.1%
39	cuotaCom056	24.5%	r_DEUDA_TOTAL_SICOM_OP_3s6M	22.3%	LN_DEUDA_TOTAL_SCE_12M	35.9%
40	DEUDA_TOTAL_OP_M	20.1%	r_PROM_XVEN_SICOM_OP_3s12M	22.2%	PROM_DEUDA_TOTAL_SC_OP_3M	23.9%
41	DEUDA_TOTAL_SICOM_OP_12M	27.6%	PROM_DEUDA_TOTAL_SICOM_OP_12M	26.5%	DEUDA_TOTAL_SBS_SC_12M	36.2%
42	DEUDA_TOTAL_SICOM_OP_6M	26.3%	DEUDA_TOTAL_SICOM_OP_12M	26.5%	INGRESO_REAL	100.0%
43	INGRESO_REAL	100.0%	PROM_XVEN_SICOM_OP_3M	22.2%	RANGO_INGRESO_G1	37.3%
44	PROM_DEUDA_TOTAL_SICOM_OP_12M	27.5%	PROM_DEUDA_TOTAL_SICOM_OP_6M	24.0%	PROM_XVEN_SC_OP_6M	24.4%
45	PROM_DEUDA_TOTAL_SICOM_OP_36M	29.3%	DEUDA_TOTAL_SICOM_OP_6M	24.0%	DEUDA_TOTAL_SC_OP_12M	25.6%
46	PROM_DEUDA_TOTAL_SICOM_OP_3M	25.0%	PROM_XVEN_SICOM_OP_12M	25.0%	salProm6M095	29.4%
47	r_DEUDA_TOTAL_SICOM_OP_3s12M	25.0%	r_PROM_DEUDA_TOTAL_SICOM_OP_12s24M	26.5%	PROM_DEUDA_TOTAL_SC_OP_6M	24.5%
48	r_DEUDA_TOTAL_SICOM_OP_3s6M	25.1%	PROM_DEUDA_TOTAL_SICOM_OP_3M	22.4%	DEUDA_TOTAL_SC_OP_6M	24.5%
49	r_DEUDA_TOTAL_SICOM_OP_6s24M	25.8%	r_PROM_DEUDA_TOTAL_SICOM_OP_6s24M	24.0%	salPromD6M316	29.4%
50	r_PROM_DEUDA_TOTAL_SICOM_OP_3s12M	25.0%	r_DEUDA_TOTAL_SICOM_OP_3s12M	22.3%	PROM_XVEN_SC_OP_24M	26.8%
51	r_PROM_DEUDA_TOTAL_SICOM_OP_3s6M	25.0%	r_PROM_XVEN_SICOM_OP_12s24M	25.0%	PROM_DEUDA_TOTAL_SC_OP_24M	26.9%
52	r_PROM_DEUDA_TOTAL_SICOM_OP_6s24M	25.8%	DEUDA_TOTAL_SBS_SC_12M	25.4%	PROM_DEUDA_TOTAL_SC_OP_12M	25.7%
53	r_PROM_XVEN_SICOM_OP_3s12M	25.2%	r_DEUDA_TOTAL_SICOM_OP_6s24M	24.0%	RANGO_INGRESO_G3	100.0%
54	r_PROM_XVEN_SICOM_OP_3s6M	25.2%	RANGO_INGRESO_G1	72.2%		
55	r_PROM_XVEN_SICOM_OP_6s12M	25.2%	RANGO_INGRESO_G3	63.8%		
56	RANGO_INGRESO_G1	100.0%	r_DEUDA_TOTAL_SICOMsSCE_12M	27.0%		
57			r_DEUDA_TOTAL_SICOMsSCE_3M	22.9%		
58			r_PROM_XVEN_SICOM_OP_3s6M	22.2%		
59			PROM_XVEN_SICOM_OP_36M	27.4%		
60			INGRESO_REAL	100.0%		
61			PROM_DEUDA_TOTAL_SICOM_OP_36M	29.0%		
62			r_PROM_XVEN_SICOM_OP_6s24M	23.1%		
63			LN_DEUDA_TOTAL_SCE_12M	23.9%		
64			r_PROM_XVEN_SICOM_OP_12s36M	25.0%		
65			PROM_XVEN_SICOM_OP_6M	23.1%		
66			r_PROM_XVEN_SICOM_OP_6s12M	23.1%		
67			r_DEUDA_TOTAL_SICOMsSCE_6M	24.4%		
68			r_PROM_DEUDA_TOTAL_SICOM_OP_6s12M	24.0%		
69			r_DEUDA_TOTAL_SICOM_OP_6s12M	24.0%		
70			DEUDA_TOTAL_SCE_12M	23.9%		
71			RANGO_INGRESO_G2	100.0%		

Tabla A.1: Variables del Test KS

Variables Cualitativas: Test VI

	Grupo N1 Variable	VI	Grupo N2 Variable	VI	Grupo N3 Variable	VI
1	ENTIDAD	36.8%	ENTIDAD	43.2%	descripcionProvincia	15.8%
2	PeorCalHis5N36MDDCom393	17.1%	PeorCalHis5N36MDDCom393	17.7%	FECHA_VENC_OTROS_OP_12M	15.3%
3	PeorCal5NDDCom400	15.3%	PeorCal5NDDCom400	13.9%	fechaPeorEdadVenD36M317	17.4%
4					numAcreedoresOPyTC386	16.4%
5					descripcionCanton	19.6%
6					ENTIDAD	111.2%
7					fecUltVencido097	21.2%
8					fecUltVencidoD341	19.2%
9					PeorCal5NDDSerCob401	17.0%
10					PeorCalHis5N36MDDCom393	24.6%
11					peorEdadVen066	19.9%
12					peorEdadVen24M257	21.8%
13					peorEdadVen36M281	21.2%
14					peorEdadVenComD24M260	15.3%
15					peorEdadVenD072	19.3%
16					peorEdadVenD24M263	21.3%
17					peorEdadVenPicSceAct370	16.4%
18					peorEdadVenPicSfnrComAct373	18.1%
19					peorNivelRiesgo5N350	20.8%
20					peorNivelRiesgoTitularSfr369	21.5%
21					provinciaDescripcion	16.1%
22					fechaPeorEdadVenD6M313	16.5%
23					PeorCalHis5N36MDDCoop392	15.6%
24					PeorCalHis5N36MDDSerCob397	15.1%
25					peorEdadVen12M233	20.3%
26					peorEdadVen3M209	19.3%
27					peorEdadVenComD36M284	15.1%
28					peorEdadVenD12M239	19.6%
29					peorEdadVenD36M287	20.6%
30					peorEdadVenD3M215	18.7%
31					peorEdadVenD6M312	19.3%

Tabla A.2: Variables del Test VI

A.2 Grid de hiperámetros para grupos G1, G2 y G3

	num_trees	mtry	min_node	error	%Represen	#Nodos	Total
1	300	6	981	327.343925	3.00%	33	32686
2	400	6	981	327.356751	3.00%	33	
3	500	6	981	327.363581	3.00%	33	
4	500	7	981	327.434589	3.00%	33	
5	500	5	981	327.436905	3.00%	33	
6	300	6	1046	327.452822	3.20%	31	
7	400	5	981	327.460093	3.00%	33	
8	400	7	981	327.463121	3.00%	33	
9	500	6	1046	327.470441	3.20%	31	
10	400	6	1046	327.471085	3.20%	31	
11	300	5	981	327.481258	3.00%	33	
12	300	7	981	327.50228	3.00%	33	
13	400	5	1046	327.5292	3.20%	31	
14	500	5	1046	327.535669	3.20%	31	
15	500	7	1046	327.536227	3.20%	31	
16	300	5	1046	327.544466	3.20%	31	
17	300	6	1111	327.545545	3.40%	29	
18	400	6	1111	327.550757	3.40%	29	
19	500	6	1111	327.55843	3.40%	29	
20	400	7	1046	327.558998	3.20%	31	
21	300	7	1046	327.584599	3.20%	31	
22	400	5	1111	327.619921	3.40%	29	
23	500	5	1111	327.626494	3.40%	29	
24	500	7	1111	327.63218	3.40%	29	
25	300	5	1111	327.651761	3.40%	29	
26	400	7	1111	327.652135	3.40%	29	
27	300	7	1111	327.685771	3.40%	29	
28	500	6	1177	327.704881	3.60%	28	
29	300	6	1177	327.721544	3.60%	28	
30	400	6	1177	327.721893	3.60%	28	
31	400	5	1177	327.72868	3.60%	28	
32	500	7	1177	327.729974	3.60%	28	
33	500	5	1177	327.7405	3.60%	28	
34	**300**	**5**	**1177**	**327.747305**	**3.60%**	**28**	
35	400	7	1177	327.75221	3.60%	28	
36	300	7	1177	327.792709	3.60%	28	
37	500	6	1242	327.832002	3.80%	26	
38	500	5	1242	327.837383	3.80%	26	
39	400	5	1242	327.840907	3.80%	26	
40	300	6	1242	327.842281	3.80%	26	
41	400	6	1242	327.844886	3.80%	26	
42	500	7	1242	327.866105	3.80%	26	
43	300	5	1242	327.876751	3.80%	26	
44	400	7	1242	327.892968	3.80%	26	
45	300	7	1242	327.910875	3.80%	26	
46	400	5	1307	327.91587	4.00%	25	
47	500	5	1307	327.93016	4.00%	25	
48	500	7	1307	327.941694	4.00%	25	
49	500	6	1307	327.943788	4.00%	25	
50	300	5	1307	327.945201	4.00%	25	
51	400	7	1307	327.948947	4.00%	25	
52	400	6	1307	327.94919	4.00%	25	
53	300	6	1307	327.951505	4.00%	25	
54	300	7	1307	327.96936	4.00%	25	

Tabla A.3: Grid de Hiperparámetros del Grupo 1

	num_trees	mtry	min_node	error	%Represen	#Nodos	Total
1	300	6	981	327.343925	3.00%	33	32686
2	400	6	981	327.356751	3.00%	33	
3	500	6	981	327.363581	3.00%	33	
4	500	7	981	327.434589	3.00%	33	
5	500	5	981	327.436905	3.00%	33	
6	300	6	1046	327.452822	3.20%	31	
7	400	5	981	327.460093	3.00%	33	
8	400	7	981	327.463121	3.00%	33	
9	500	6	1046	327.470441	3.20%	31	
10	400	6	1046	327.471085	3.20%	31	
11	300	5	981	327.481258	3.00%	33	
12	300	7	981	327.50228	3.00%	33	
13	400	5	1046	327.5292	3.20%	31	
14	500	5	1046	327.535669	3.20%	31	
15	500	7	1046	327.536227	3.20%	31	
16	300	5	1046	327.544466	3.20%	31	
17	300	6	1111	327.545545	3.40%	29	
18	400	6	1111	327.550757	3.40%	29	
19	500	6	1111	327.55843	3.40%	29	
20	400	7	1046	327.558998	3.20%	31	
21	300	7	1046	327.584599	3.20%	31	
22	400	5	1111	327.619921	3.40%	29	
23	500	5	1111	327.626494	3.40%	29	
24	500	7	1111	327.63218	3.40%	29	
25	300	5	1111	327.651761	3.40%	29	
26	400	7	1111	327.652135	3.40%	29	
27	300	7	1111	327.685771	3.40%	29	
28	500	6	1177	327.704881	3.60%	28	
29	300	6	1177	327.721544	3.60%	28	
30	400	6	1177	327.721893	3.60%	28	
31	400	5	1177	327.72868	3.60%	28	
32	500	7	1177	327.729974	3.60%	28	
33	500	5	1177	327.7405	3.60%	28	
34	**300**	**5**	**1177**	**327.747305**	**3.60%**	**28**	
35	400	7	1177	327.75221	3.60%	28	
36	300	7	1177	327.792709	3.60%	28	
37	500	6	1242	327.832002	3.80%	26	
38	500	5	1242	327.837383	3.80%	26	
39	400	5	1242	327.840907	3.80%	26	
40	300	6	1242	327.842281	3.80%	26	
41	400	6	1242	327.844886	3.80%	26	
42	500	7	1242	327.866105	3.80%	26	
43	300	5	1242	327.876751	3.80%	26	
44	400	7	1242	327.892968	3.80%	26	
45	300	7	1242	327.910875	3.80%	26	
46	400	5	1307	327.91587	4.00%	25	
47	500	5	1307	327.93016	4.00%	25	
48	500	7	1307	327.941694	4.00%	25	
49	500	6	1307	327.943788	4.00%	25	
50	300	5	1307	327.945201	4.00%	25	
51	400	7	1307	327.948947	4.00%	25	
52	400	6	1307	327.94919	4.00%	25	
53	300	6	1307	327.951505	4.00%	25	
54	300	7	1307	327.96936	4.00%	25	

Tabla A.4: Hiperparámetros del Grupo 2

	num_trees	mtry	min_node	error	%Represen	#Nodos	Total
1	500	11	1126	2297.31053	3.00%	33	37532
2	300	11	1126	2297.32321	3.00%	33	
3	400	11	1126	2297.73867	3.00%	33	
4	300	10	1126	2297.9462	3.00%	33	
5	500	10	1126	2297.95472	3.00%	33	
6	400	10	1126	2298.13606	3.00%	33	
7	300	9	1126	2299.23941	3.00%	33	
8	400	9	1126	2299.25886	3.00%	33	
9	500	9	1126	2299.38816	3.00%	33	
10	500	11	1201	2300.03442	3.20%	31	
11	500	10	1201	2300.22652	3.20%	31	
12	300	10	1201	2300.38367	3.20%	31	
13	400	10	1201	2300.41439	3.20%	31	
14	300	11	1201	2300.49309	3.20%	31	
15	400	11	1201	2300.57514	3.20%	31	
16	300	9	1201	2301.7933	3.20%	31	
17	500	9	1201	2301.91002	3.20%	31	
18	400	9	1201	2302.06073	3.20%	31	
19	500	11	1276	2302.1697	3.40%	29	
20	400	10	1276	2302.45241	3.40%	29	
21	300	10	1276	2302.45856	3.40%	29	
22	500	10	1276	2302.47018	3.40%	29	
23	300	11	1276	2302.74597	3.40%	29	
24	400	11	1276	2302.86854	3.40%	29	
25	300	9	1276	2303.70252	3.40%	29	
26	500	9	1276	2303.88379	3.40%	29	
27	400	9	1276	2304.13699	3.40%	29	
28	500	11	1351	2304.61564	3.60%	28	
29	500	10	1351	2304.77252	3.60%	28	
30	400	10	1351	2304.87562	3.60%	28	
31	300	10	1351	2304.89409	3.60%	28	
32	300	11	1351	2305.14779	3.60%	28	
33	400	11	1351	2305.1559	3.60%	28	
34	**300**	**9**	**1351**	**2306.3196**	**3.60%**	**28**	
35	500	9	1351	2306.34516	3.60%	28	
36	400	9	1351	2306.58753	3.60%	28	
37	500	11	1426	2306.90858	3.80%	26	
38	400	11	1426	2307.38861	3.80%	26	
39	300	11	1426	2307.49646	3.80%	26	
40	500	10	1426	2307.68816	3.80%	26	
41	400	10	1426	2307.96395	3.80%	26	
42	300	10	1426	2308.07805	3.80%	26	
43	300	9	1426	2308.69453	3.80%	26	
44	500	9	1426	2309.2606	3.80%	26	
45	400	9	1426	2309.38745	3.80%	26	
46	500	11	1501	2310.09148	4.00%	25	
47	500	10	1501	2310.24988	4.00%	25	
48	400	11	1501	2310.33132	4.00%	25	
49	300	11	1501	2310.41805	4.00%	25	
50	300	10	1501	2310.51282	4.00%	25	
51	400	10	1501	2310.52007	4.00%	25	
52	300	9	1501	2311.71288	4.00%	25	
53	500	9	1501	2311.95175	4.00%	25	
54	400	9	1501	2312.19993	4.00%	25	

Tabla A.5: Hiperparámetros del Grupo 3

A.3 IL reales y estimados con BDD de modelamiento y validación General

BDD Modelamiento

Edad	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
[18,30)	2744	2197	1554	1264	673
[30,40)	4950	4934	4337	4176	2982
[40,55)	4687	4609	5668	6497	5371
[55,65)	1801	2009	2738	3127	2368
[65,100)	859	1186	1683	1703	1179
					75296

Edad	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
[18,30)	1970	2654	2001	1229	577
[30,40)	3667	5225	4978	4229	3275
[40,55)	3518	5397	5976	6055	5886
[55,65)	1281	2360	2846	2835	2721
[65,100)	680	1363	1602	1624	1341
					75290

Edad	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
[18,30)	3.64%	2.92%	2.06%	1.68%	0.89%
[30,40)	6.57%	6.55%	5.76%	5.55%	3.96%
[40,55)	6.22%	6.12%	7.53%	8.63%	7.13%
[55,65)	2.39%	2.67%	3.64%	4.15%	3.14%
[65,100)	1.14%	1.58%	2.24%	2.26%	1.57%
					100%

Edad	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
[18,30)	2.62%	3.53%	2.66%	1.63%	0.77%
[30,40)	4.87%	6.94%	6.61%	5.62%	4.35%
[40,55)	4.67%	7.17%	7.94%	8.04%	7.82%
[55,65)	1.70%	3.13%	3.78%	3.77%	3.61%
[65,100)	0.90%	1.81%	2.13%	2.16%	1.78%
					100%

E Civil	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
100%	8509	6935	6076	5889	4016
200%	5206	6426	8045	9120	7248
300%	951	1075	1284	1295	1013
400%	303	404	506	413	259
500%	47	75	53	40	27
					75215

E Civil	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
100%	6672	8176	7195	5376	4003
200%	3454	6981	8392	8944	8272
300%	714	1292	1262	1193	1157
400%	203	487	477	395	323
500%	35	53	60	55	38
					75209

E Civil	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	11.31%	9.22%	8.08%	7.83%	5.34%
2	6.92%	8.54%	10.70%	12.13%	9.64%
3	1.26%	1.43%	1.71%	1.72%	1.35%
4	0.40%	0.54%	0.67%	0.55%	0.34%
5	0.06%	0.10%	0.07%	0.05%	0.04%
					100%

E Civil	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	8.87%	10.87%	9.57%	7.15%	5.32%
2	4.59%	9.28%	11.16%	11.89%	11.00%
3	0.95%	1.72%	1.68%	1.59%	1.54%
4	0.27%	0.65%	0.63%	0.53%	0.43%
5	0.05%	0.07%	0.08%	0.07%	0.05%
					100%

Genero	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	8644	8953	8704	9053	7473
2	6372	5962	7260	7704	5090
					75215

Genero	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	6486	9707	9536	8833	8259
2	4592	7282	7850	7130	5534
					75209

Genero	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	11.49%	11.90%	11.57%	12.04%	9.94%
2	8.47%	7.93%	9.65%	10.24%	6.77%
					100%

Genero	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	8.62%	12.91%	12.68%	11.74%	10.98%
2	6.11%	9.68%	10.44%	9.48%	7.36%
					100%

Region	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Amazonia	1512	1641	1561	1832	1377
Costa	6237	4274	3975	3893	3019
Sierra	7292	9020	10444	11042	8177
					75296

Region	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Amazonia	1003	1761	1683	1777	1698
Costa	5974	4574	4439	3599	2810
Sierra	4139	10664	11281	10596	9292
					75290

Region	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Amazonia	2.01%	2.18%	2.07%	2.43%	1.83%
Costa	8.28%	5.68%	5.28%	5.17%	4.01%
Sierra	9.68%	11.98%	13.87%	14.66%	10.86%
					100%

Region	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Amazonia	1.33%	2.34%	2.24%	2.36%	2.26%
Costa	7.93%	6.07%	5.90%	4.78%	3.73%
Sierra	5.50%	14.16%	14.98%	14.07%	12.34%
					100%

Provincia	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Bolivar	293	378	335	242	116
Cotopaxi	706	1254	1367	949	485
Guayas	2409	1546	1255	1179	1111
Pichincha	8329	8017	9444	11458	9074
Tungurahua	1960	2447	2359	1849	1071
Otros	607	556	613	583	451
					72443

Provincia	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Bolivar	124	299	412	364	165
Cotopaxi	261	1245	1406	1172	677
Guayas	2607	1562	1297	1038	994
Pichincha	5804	10057	10652	10276	9530
Tungurahua	1418	2507	2342	1939	1479
Otros	336	589	635	652	598
					72437

Provincia	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Bolivar	0.40%	0.52%	0.46%	0.33%	0.16%
Cotopaxi	0.97%	1.73%	1.89%	1.31%	0.67%
Guayas	3.33%	2.13%	1.73%	1.63%	1.53%
Pichincha	11.50%	11.07%	13.04%	15.82%	12.53%
Tungurahua	2.71%	3.38%	3.26%	2.55%	1.48%
Otros	0.84%	0.77%	0.85%	0.80%	0.62%
					100%

Provincia	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Bolivar	0.17%	0.41%	0.57%	0.50%	0.23%
Cotopaxi	0.36%	1.72%	1.94%	1.62%	0.93%
Guayas	3.60%	2.16%	1.79%	1.43%	1.37%
Pichincha	8.01%	13.88%	14.70%	14.18%	13.16%
Tungurahua	1.96%	3.46%	3.23%	2.68%	2.04%
Otros	0.46%	0.81%	0.88%	0.90%	0.83%
					100%

Tabla A.6: IL Real y Estimado con BDD Modelamiento General

BDD Validación

Edad	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
[18,30)	1496	1108	624	689	495
[30,40)	1573	2021	1663	1846	1438
[40,55)	1773	2152	2167	2691	2618
[55,65)	1051	1092	1020	1244	1071
[65,100)	822	911	673	700	616
					33554

Edad	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
[18,30)	810	2597	613	302	90
[30,40)	704	3944	2000	1423	470
[40,55)	731	4647	2769	2310	944
[55,65)	312	2423	1399	1000	344
[65,100)	261	1936	819	560	146
					33554

Edad	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
[18,30)	4.46%	3.30%	1.86%	2.05%	1.48%
[30,40)	4.69%	6.02%	4.96%	5.50%	4.29%
[40,55)	5.28%	6.41%	6.46%	8.02%	7.80%
[55,65)	3.13%	3.25%	3.04%	3.71%	3.19%
[65,100)	2.45%	2.72%	2.01%	2.09%	1.84%
					100%

Edad	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
[18,30)	2.41%	7.74%	1.83%	0.90%	0.27%
[30,40)	2.10%	11.75%	5.96%	4.24%	1.40%
[40,55)	2.18%	13.85%	8.25%	6.88%	2.81%
[55,65)	0.93%	7.22%	4.17%	2.98%	1.03%
[65,100)	0.78%	5.77%	2.44%	1.67%	0.44%
					100%

E Civil	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
100%	3364	2822	1784	1946	1549
200%	2502	3554	3734	4585	4119
300%	613	623	464	451	405
400%	218	255	142	156	134
500%	9	13	9	14	17
					33482

E Civil	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
100%	1578	6355	1928	1192	412
200%	929	7291	4946	3909	1419
300%	223	1319	548	346	120
400%	48	534	157	130	36
500%	7	19	14	16	6
					33482

E Civil	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	10.05%	8.43%	5.33%	5.81%	4.63%
2	7.47%	10.61%	11.15%	13.69%	12.30%
3	1.83%	1.86%	1.39%	1.35%	1.21%
4	0.65%	0.76%	0.42%	0.47%	0.40%
5	0.03%	0.04%	0.03%	0.04%	0.05%
					100%

E Civil	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	4.71%	18.98%	5.76%	3.56%	1.23%
2	2.77%	21.78%	14.77%	11.67%	4.24%
3	0.67%	3.94%	1.64%	1.03%	0.36%
4	0.14%	1.59%	0.47%	0.39%	0.11%
5	0.02%	0.06%	0.04%	0.05%	0.02%
					100%

Genero	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	3280	4049	3416	4076	3616
2	3426	3218	2717	3076	2608
					33482

Genero	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	1330	8446	4142	3204	1315
2	1455	7072	3451	2389	678
					33482

Genero	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	9.80%	12.09%	10.20%	12.17%	10.80%
2	10.23%	9.61%	8.11%	9.19%	7.79%
					100%

Genero	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
1	3.97%	25.23%	12.37%	9.57%	3.93%
2	4.35%	21.12%	10.31%	7.14%	2.02%
					100%

Region	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Amazonia	856	791	639	714	596
Costa	202	425	530	843	768
Sierra	5657	6068	4978	5613	4874
					33554

Region	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Amazonia	306	1741	758	589	202
Costa	524	1056	567	436	185
Sierra	1988	12750	6275	4570	1607
					33554

Region	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Amazonia	2.55%	2.36%	1.90%	2.13%	1.78%
Costa	0.60%	1.27%	1.58%	2.51%	2.29%
Sierra	16.86%	18.08%	14.84%	16.73%	14.53%
					100%

Region	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Amazonia	0.91%	5.19%	2.26%	1.76%	0.60%
Costa	1.56%	3.15%	1.69%	1.30%	0.55%
Sierra	5.92%	38.00%	18.70%	13.62%	4.79%
					100%

Provincia	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Bolivar	267	283	207	230	114
Cotopaxi	524	733	526	550	499
Guayas	111	277	341	496	515
Pichincha	979	1023	942	1401	1357
Tungurahua	1596	758	491	627	563
Otros	2820	3864	3375	3653	3060
					32182

Provincia	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Bolivar	72	533	269	172	55
Cotopaxi	112	1281	714	543	182
Guayas	373	685	337	236	109
Pichincha	510	2313	1263	1127	489
Tungurahua	608	1702	985	541	199
Otros	999	8306	3737	2802	928
					32182

Provincia	Ingreso Real				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Bolivar	0.83%	0.88%	0.64%	0.71%	0.35%
Cotopaxi	1.63%	2.28%	1.63%	1.71%	1.55%
Guayas	0.34%	0.86%	1.06%	1.54%	1.60%
Pichincha	3.04%	3.18%	2.93%	4.35%	4.22%
Tungurahua	4.96%	2.36%	1.53%	1.95%	1.75%
Otros	8.76%	12.01%	10.49%	11.35%	9.51%
					100%

Provincia	Ingreso Estimado				
	[400,791]	(791,1200]	(1200,1740]	(1740,3000]	(3000,35000]
Bolivar	0.22%	1.66%	0.84%	0.53%	0.17%
Cotopaxi	0.35%	3.98%	2.22%	1.69%	0.57%
Guayas	1.16%	2.13%	1.05%	0.73%	0.34%
Pichincha	1.58%	7.19%	3.92%	3.50%	1.52%
Tungurahua	1.89%	5.29%	3.06%	1.68%	0.62%
Otros	3.10%	25.81%	11.61%	8.71%	2.88%
					100%

Tabla A.7: IL Real y Estimado con BDD Validación General

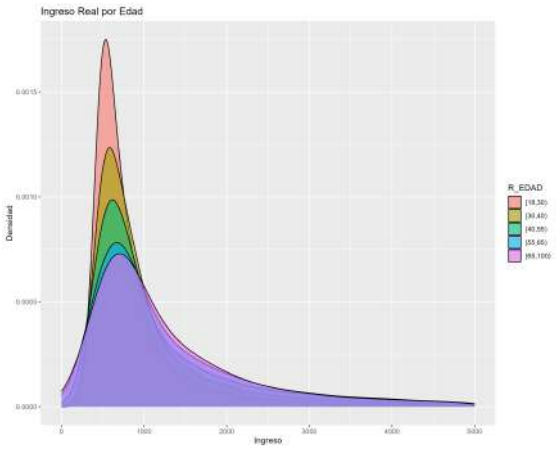


Figura A.1: BDD Modelamiento: I Real

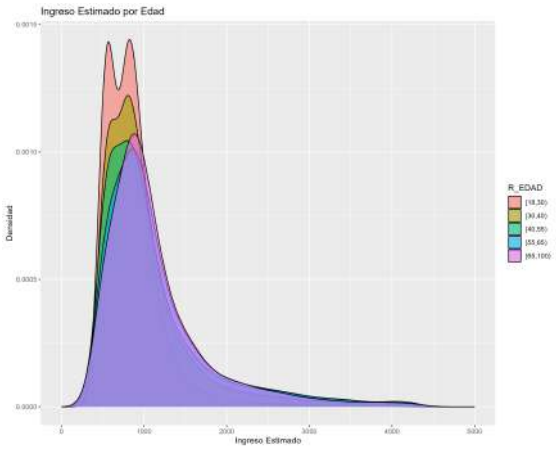


Figura A.2: BDD Modelamiento: I Estimado

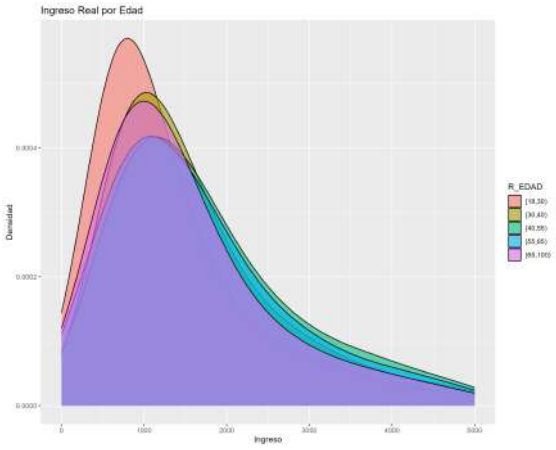


Figura A.3: BDD Validación: I Real

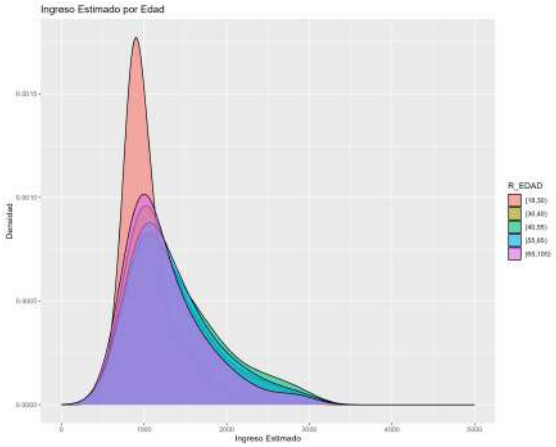


Figura A.4: BDD Validación:I Estimado

Figura A.5: Indicadores de Liquidez por Edad para BDD Modelamiento y Validación según Ingreso Real y Estimado - General

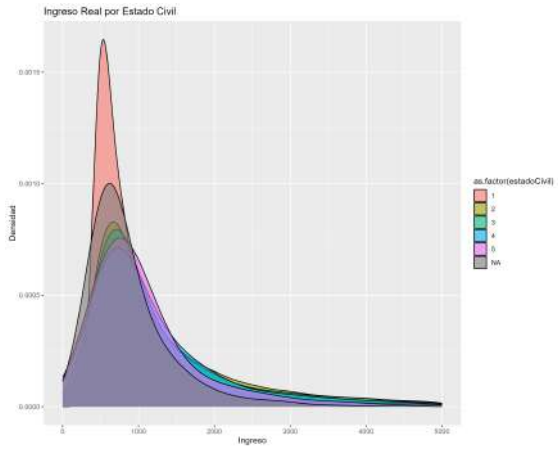


Figura A.6: BDD Modelamiento: I Real

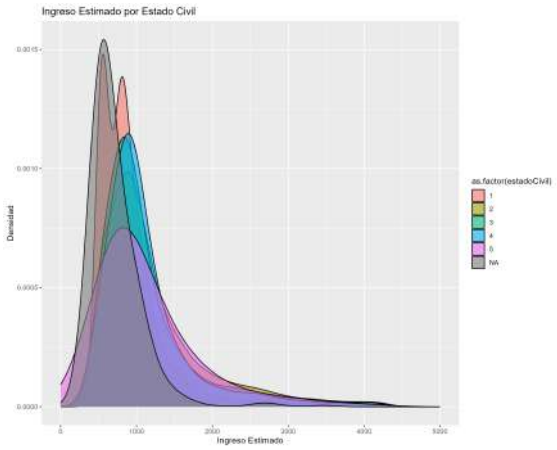


Figura A.7: BDD Modelamiento: I Estimado

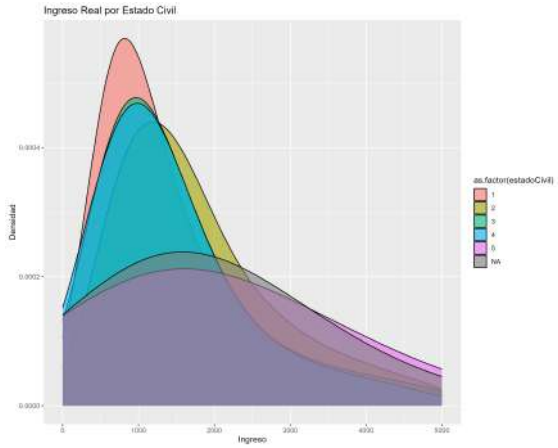


Figura A.8: BDD Validación: I Real

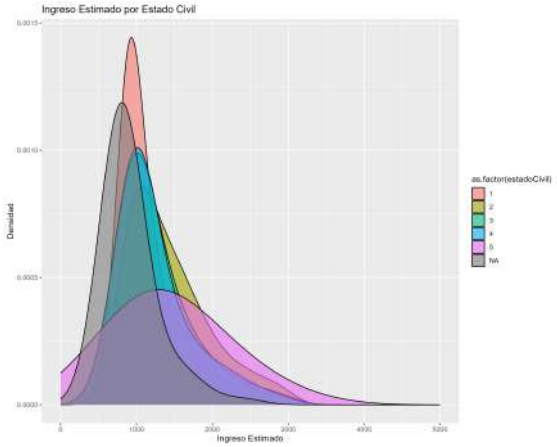


Figura A.9: BDD Validación:I Estimado

Figura A.10: Indicadores de Liquidez por Estado Civil para BDD Modelamiento y Validación según Ingreso Real y Estimado - General

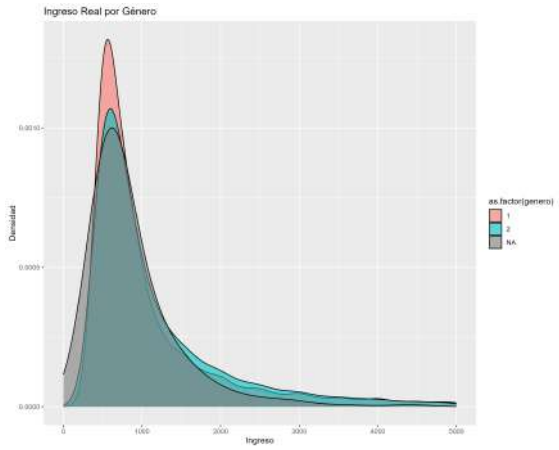


Figura A.11: BDD Modelamiento: I Real

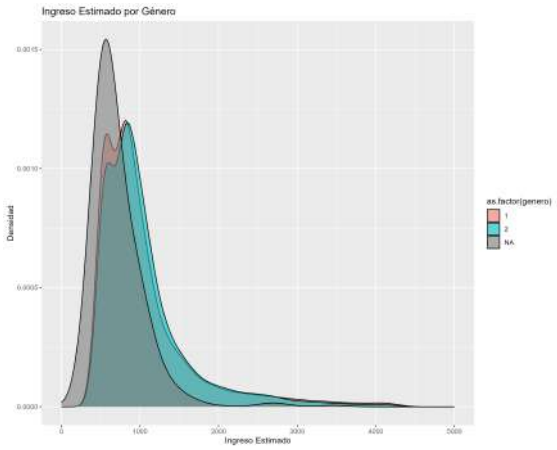


Figura A.12: BDD Modelamiento: I Estimado

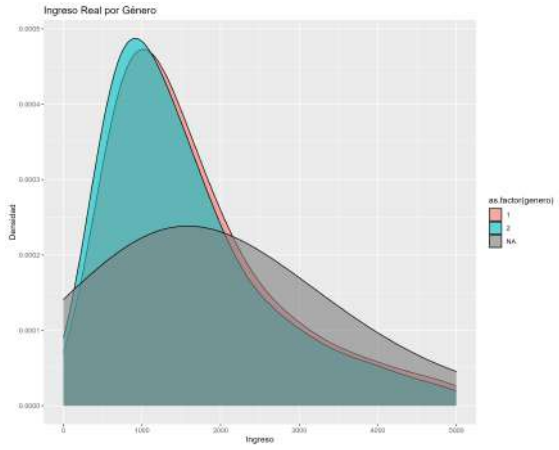


Figura A.13: BDD Validación: I Real

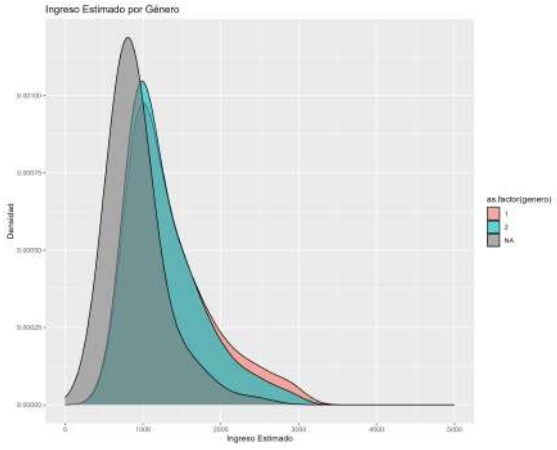


Figura A.14: BDD Validación:I Estimado

Figura A.15: Indicadores de Liquidez por Genero para BDD Modelamiento y Validación según Ingreso Real y Estimado - General

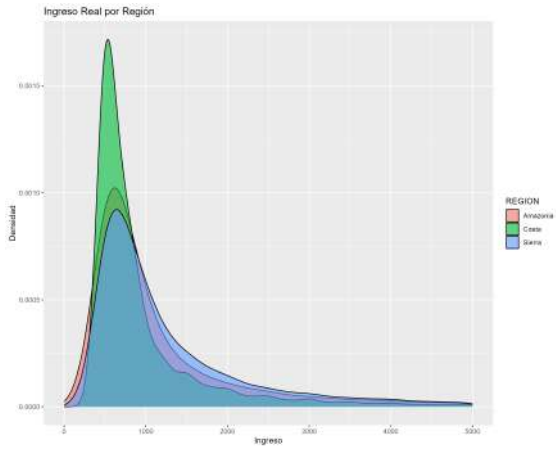


Figura A.16: BDD Modelamiento: I Real

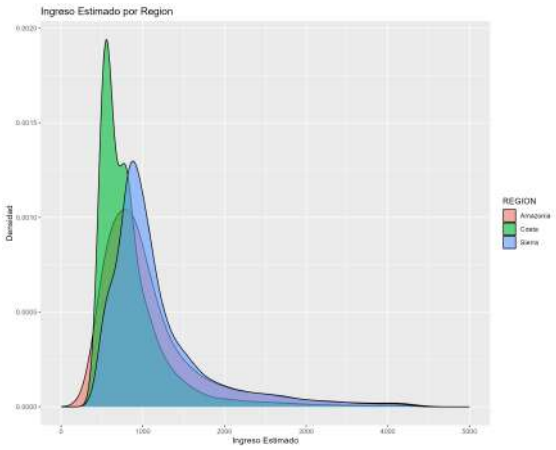


Figura A.17: BDD Modelamiento: I Estimado

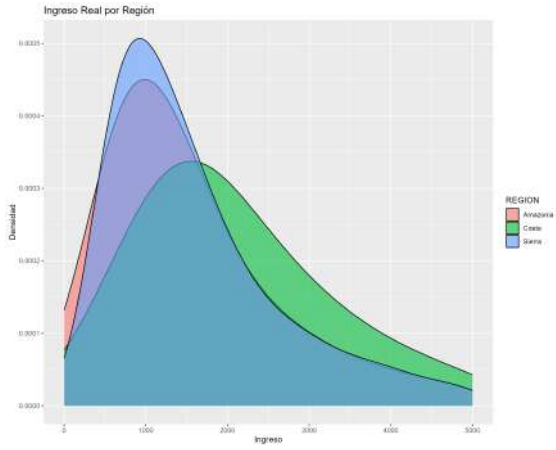


Figura A.18: BDD Validación: I Real

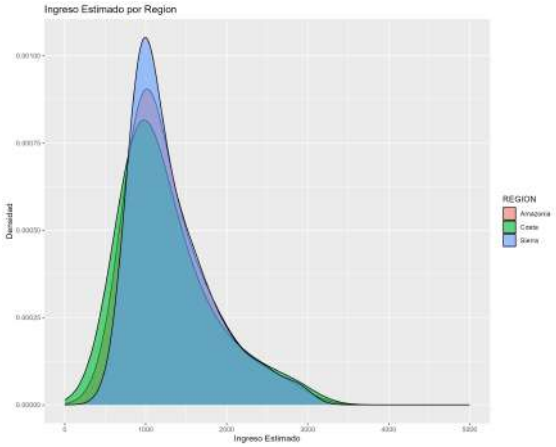


Figura A.19: BDD Validación:I Estimado

Figura A.20: Indicadores de Liquidez por Region para BDD Modelamiento y Validación según Ingreso Real y Estimado - General

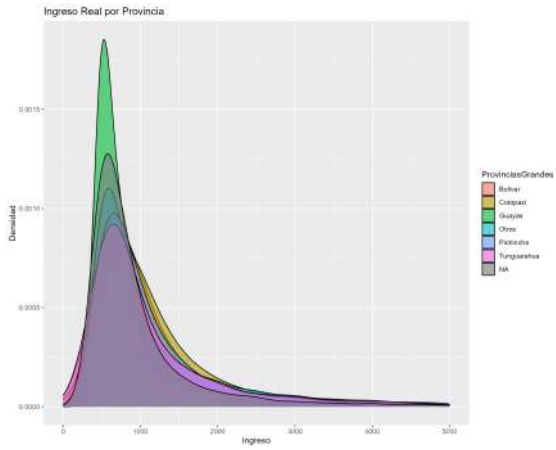


Figura A.21: BDD Modelamiento: I Real

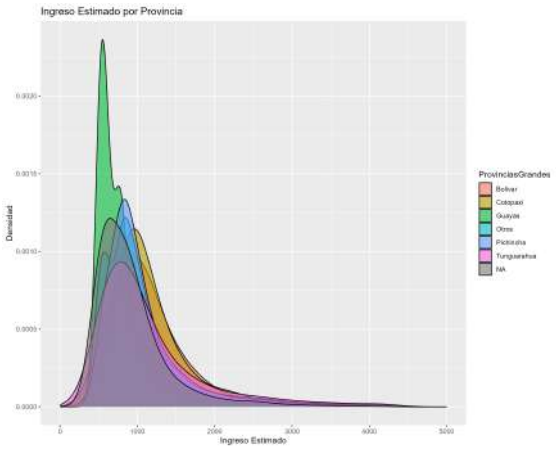


Figura A.22: BDD Modelamiento: I Estimado

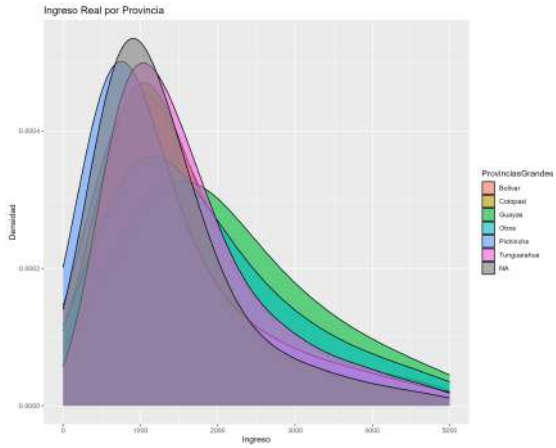


Figura A.23: BDD Validación: I Real

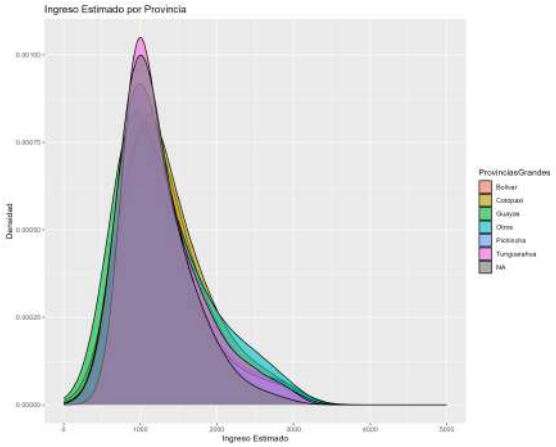


Figura A.24: BDD Validación:I Estimado

Figura A.25: Indicadores de Liquidez por Provincias para BDD Modelamiento y Validación según Ingreso Real y Estimado - General

A.4 IL reales y estimados con BDD de modelamiento y validación para grupo G1

BDD Modelamiento

Edad	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
[18,30)	3750	3318	2480	1818	1883
[30,40)	4882	4198	3303	3212	3398
[40,55)	5407	4222	3358	3287	4017
[55,65)	1944	1460	1293	1286	1826
[65,100)	1172	868	775	876	1282
					65315

Edad	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
[18,30)	2722	2649	1593	4586	1699
[30,40)	3011	4036	2531	7156	2259
[40,55)	3909	4099	2662	6864	2757
[55,65)	1566	1356	850	2792	1245
[65,100)	928	682	466	1961	936

Edad	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
[18,30)	5.74%	5.08%	3.80%	2.78%	2.88%
[30,40)	7.47%	6.43%	5.06%	4.92%	5.20%
[40,55)	8.28%	6.46%	5.14%	5.03%	6.15%
[55,65)	2.98%	2.24%	1.98%	1.97%	2.80%
[65,100)	1.79%	1.33%	1.19%	1.34%	1.96%
					100%

Edad	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
[18,30)	4.17%	4.06%	2.44%	7.02%	2.60%
[30,40)	4.61%	6.18%	3.88%	10.96%	3.46%
[40,55)	5.98%	6.28%	4.08%	10.51%	4.22%
[55,65)	2.40%	2.08%	1.30%	4.27%	1.91%
[65,100)	1.42%	1.04%	0.71%	3.00%	1.43%
					100%

E Civil	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	11601	8912	6899	5835	5680
2	4688	4276	3502	3768	5560
3	567	614	545	615	828
4	238	176	193	204	257
5	25	39	28	34	36
					65120

E Civil	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	8511	8509	4797	13333	3777
2	3014	3549	2740	8003	4488
3	358	576	405	1446	384
4	171	125	114	459	199
5	22	28	18	70	24

E Civil	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	17.81%	13.69%	10.59%	8.96%	8.72%
2	7.20%	6.57%	5.38%	5.79%	8.54%
3	0.87%	0.94%	0.84%	0.94%	1.27%
4	0.37%	0.27%	0.30%	0.31%	0.39%
5	0.04%	0.06%	0.04%	0.05%	0.06%
					100%

E Civil	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	13.07%	13.07%	7.37%	20.47%	5.80%
2	4.63%	5.45%	4.21%	12.29%	6.89%
3	0.55%	0.88%	0.62%	2.22%	0.59%
4	0.26%	0.19%	0.18%	0.70%	0.31%
5	0.03%	0.04%	0.03%	0.11%	0.04%
					100%

Genero	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	9729	8222	6546	5473	6074
2	7390	5795	4621	4983	6287
					65120

Genero	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	7266	7990	4211	12769	3808
2	4810	4797	3863	10542	5064

Genero	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	14.94%	12.63%	10.05%	8.40%	9.33%
2	11.35%	8.90%	7.10%	7.65%	9.65%
					100%

Genero	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	11.16%	12.27%	6.47%	19.61%	5.85%
2	7.39%	7.37%	5.93%	16.19%	7.78%
					100%

Region	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Amazonia	1619	1076	933	827	912
Costa	10594	8038	5945	5343	4813
Sierra	4942	4952	4331	4309	6681
					65315

Region	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Amazonia	1089	1074	631	1917	656
Costa	8826	8553	4721	10317	2316
Sierra	2221	3195	2750	11125	5924

Region	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Amazonia	2.48%	1.65%	1.43%	1.27%	1.40%
Costa	16.22%	12.31%	9.10%	8.18%	7.37%
Sierra	7.57%	7.58%	6.63%	6.60%	10.23%
					100%

Region	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Amazonia	1.67%	1.64%	0.97%	2.94%	1.00%
Costa	13.51%	13.10%	7.23%	15.80%	3.55%
Sierra	3.40%	4.89%	4.21%	17.03%	9.07%
					100%

Provincia	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Bolivar	244	149	167	162	200
Cotopaxi	296	309	292	314	349
Guayas	4140	3311	2369	2221	1904
Pichincha	9779	7976	6309	5808	7448
Tungurahua	1140	1286	1060	1207	1714
Otros	415	290	372	228	268
					61727

Provincia	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Bolivar	82	149	88	306	297
Cotopaxi	117	131	146	608	558
Guayas	3273	3612	2080	4293	687
Pichincha	7225	6975	4076	13442	5602
Tungurahua	304	860	995	3078	1170
Otros	264	327	265	522	195

Provincia	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Bolivar	0.40%	0.24%	0.27%	0.26%	0.32%
Cotopaxi	0.48%	0.50%	0.47%	0.51%	0.57%
Guayas	6.71%	5.36%	3.84%	3.60%	3.08%
Pichincha	15.84%	12.92%	10.22%	9.41%	12.07%
Tungurahua	1.85%	2.08%	1.72%	1.96%	2.78%
Otros	0.67%	0.47%	0.60%	0.37%	0.43%
					100%

Provincia	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Bolivar	0.13%	0.24%	0.14%	0.50%	0.48%
Cotopaxi	0.19%	0.21%	0.24%	0.98%	0.90%
Guayas	5.30%	5.85%	3.37%	6.95%	1.11%
Pichincha	11.70%	11.30%	6.60%	21.78%	9.08%
Tungurahua	0.49%	1.39%	1.61%	4.99%	1.90%
Otros	0.43%	0.53%	0.43%	0.85%	0.32%
					100%

Tabla A.8: IL Real y Estimado con BDD Modelamiento para G1

BDD Validación

Edad	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
[18,30)	72	55	58	65	141
[30,40)	47	65	72	76	258
[40,55)	71	100	77	110	385
[55,65)	39	64	61	78	203
[65,100)	52	66	51	78	171
					2515

Edad	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
[18,30)	2	6	31	287	65
[30,40)	3	21	33	303	158
[40,55)	2	13	51	418	259
[55,65)	1	12	12	276	144
[65,100)	2	5	12	299	100

Edad	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
[18,30)	2.86%	2.19%	2.31%	2.58%	5.61%
[30,40)	1.87%	2.58%	2.86%	3.02%	10.26%
[40,55)	2.82%	3.98%	3.06%	4.37%	15.31%
[55,65)	1.55%	2.54%	2.43%	3.10%	8.07%
[65,100)	2.07%	2.62%	2.03%	3.10%	6.80%
					100%

Edad	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
[18,30)	0.08%	0.24%	1.23%	11.41%	2.58%
[30,40)	0.12%	0.83%	1.31%	12.05%	6.28%
[40,55)	0.08%	0.52%	2.03%	16.62%	10.30%
[55,65)	0.04%	0.48%	0.48%	10.97%	5.73%
[65,100)	0.08%	0.20%	0.48%	11.89%	3.98%
					100%

E Civil	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	138	154	140	146	352
2	109	149	138	191	688
3	21	32	31	44	70
4	13	14	10	25	35
5	0	0	0	1	3
					2504

E Civil	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	8	25	68	669	160
2	2	26	52	693	502
3	0	4	11	144	39
4	0	2	4	67	24
5	0	0	1	2	1

E Civil	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	5.51%	6.15%	5.59%	5.83%	14.06%
2	4.35%	5.95%	5.51%	7.63%	27.48%
3	0.84%	1.28%	1.24%	1.76%	2.80%
4	0.52%	0.56%	0.40%	1.00%	1.40%
5	0.00%	0.00%	0.00%	0.04%	0.12%
					100%

E Civil	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	0.32%	1.00%	2.72%	26.72%	6.39%
2	0.08%	1.04%	2.08%	27.68%	20.05%
3	0.00%	0.16%	0.44%	5.75%	1.56%
4	0.00%	0.08%	0.16%	2.68%	0.96%
5	0.00%	0.00%	0.04%	0.08%	0.04%
					100%

Genero	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	128	165	165	217	620
2	153	184	154	190	528
					2504

Genero	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	6	29	48	837	375
2	4	28	88	738	351

Genero	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	5.11%	6.59%	6.59%	8.67%	24.76%
2	6.11%	7.35%	6.15%	7.59%	21.09%
					100%

Genero	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
1	0.24%	1.16%	1.92%	33.43%	14.98%
2	0.16%	1.12%	3.51%	29.47%	14.02%
					100%

Region	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Amazonia	24	34	32	35	103
Costa	5	8	13	32	181
Sierra	252	308	274	340	874
					2515

Region	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Amazonia	0	4	11	145	68
Costa	1	15	23	175	25
Sierra	9	38	105	1263	633

Region	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Amazonia	0.95%	1.35%	1.27%	1.39%	4.10%
Costa	0.20%	0.32%	0.52%	1.27%	7.20%
Sierra	10.02%	12.25%	10.89%	13.52%	34.75%
					100%

Region	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Amazonia	0.00%	0.16%	0.44%	5.77%	2.70%
Costa	0.04%	0.60%	0.91%	6.96%	0.99%
Sierra	0.36%	1.51%	4.17%	50.22%	25.17%
					100%

Provincia	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Bolivar	13	11	13	14	27
Cotopaxi	7	20	15	15	57
Guayas	3	5	7	24	137
Pichincha	40	43	40	51	171
Tungurahua	63	74	55	53	164
Otros	147	179	176	235	555
					2414

Provincia	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Bolivar	0	0	5	46	27
Cotopaxi	0	3	1	55	55
Guayas	1	15	17	124	19
Pichincha	0	6	29	219	91
Tungurahua	5	18	41	260	85
Otros	4	12	42	817	417

Provincia	Ingreso Real				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Bolivar	0.54%	0.46%	0.54%	0.58%	1.12%
Cotopaxi	0.29%	0.83%	0.62%	0.62%	2.36%
Guayas	0.12%	0.21%	0.29%	0.99%	5.68%
Pichincha	1.66%	1.78%	1.66%	2.11%	7.08%
Tungurahua	2.61%	3.07%	2.28%	2.20%	6.79%
Otros	6.09%	7.42%	7.29%	9.73%	22.99%
					100%

Provincia	Ingreso Estimado				
	[450-500]	(500-600]	(600-750]	(750-950]	(950-2500]
Bolivar	0.00%	0.00%	0.21%	1.91%	1.12%
Cotopaxi	0.00%	0.12%	0.04%	2.28%	2.28%
Guayas	0.04%	0.62%	0.70%	5.14%	0.79%
Pichincha	0.00%	0.25%	1.20%	9.07%	3.77%
Tungurahua	0.21%	0.75%	1.70%	10.77%	3.52%
Otros	0.17%	0.50%	1.74%	33.84%	17.27%
					100%

Tabla A.9: IL Real y Estimado con BDD Validación para G1

A.5 IL reales y estimados con BDD de modelamiento y validación para grupo G2

BDD Modelamiento

Edad	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[18,30)	7704	7253	5147	3881	3271
[30,40)	8577	9302	8127	6726	6754
[40,55)	8731	8823	8385	6965	9734
[55,65)	3271	3441	3462	3007	4644
[65,100)	1726	1988	2082	1926	2914
					137841

Edad	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[18,30)	5350	3748	9551	7446	1161
[30,40)	6293	4839	12108	12070	4176
[40,55)	7030	5073	11817	13137	5581
[55,65)	2326	1782	4982	6132	2603
[65,100)	1143	840	3115	3930	1608

Edad	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[18,30)	5.59%	5.26%	3.73%	2.82%	2.37%
[30,40)	6.22%	6.75%	5.90%	4.88%	4.90%
[40,55)	6.33%	6.40%	6.08%	5.05%	7.06%
[55,65)	2.37%	2.50%	2.51%	2.18%	3.37%
[65,100)	1.25%	1.44%	1.51%	1.40%	2.11%
					100%

Edad	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[18,30)	3.88%	2.72%	6.93%	5.40%	0.84%
[30,40)	4.57%	3.51%	8.78%	8.76%	3.03%
[40,55)	5.10%	3.68%	8.57%	9.53%	4.05%
[55,65)	1.69%	1.29%	3.61%	4.45%	1.89%
[65,100)	0.83%	0.61%	2.26%	2.85%	1.17%
					100%

E Civil	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	20182	18792	15183	10765	11248
2	7901	9462	9574	9788	13553
3	1282	1784	1721	1317	1813
4	493	607	596	505	595
5	58	65	69	66	46
					137465

E Civil	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	15228	10398	24955	21236	4353
2	5499	4669	13007	17617	9486
3	931	844	2563	2608	971
4	270	254	906	1095	271
5	41	48	85	85	45

E Civil	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	14.68%	13.67%	11.04%	7.83%	8.18%
2	5.75%	6.88%	6.96%	7.12%	9.86%
3	0.93%	1.30%	1.25%	0.96%	1.32%
4	0.36%	0.44%	0.43%	0.37%	0.43%
5	0.04%	0.05%	0.05%	0.05%	0.03%
					100%

E Civil	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	11.08%	7.56%	18.15%	15.45%	3.17%
2	4.00%	3.40%	9.46%	12.82%	6.90%
3	0.68%	0.61%	1.86%	1.90%	0.71%
4	0.20%	0.18%	0.66%	0.80%	0.20%
5	0.03%	0.03%	0.06%	0.06%	0.03%
					100%

Genero	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	17152	18421	15117	11538	12976
2	12764	12289	12026	10903	14279
					137465

Genero	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	12959	9018	23810	21850	7567
2	9010	7195	17706	20791	7559

Genero	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	12.48%	13.40%	11.00%	8.39%	9.44%
2	9.29%	8.94%	8.75%	7.93%	10.39%
					100%

Genero	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	9.43%	6.56%	17.32%	15.89%	5.50%
2	6.55%	5.23%	12.88%	15.12%	5.50%
					100%

Region	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Amazonia	2806	2837	2543	1903	2123
Costa	15177	13012	10679	6445	7900
Sierra	12026	14958	13981	14157	17294
					137841

Region	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Amazonia	1955	1338	3863	3591	1465
Costa	14713	9315	14962	11224	2999
Sierra	5474	5629	22748	27900	10665

Region	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Amazonia	2.04%	2.06%	1.84%	1.38%	1.54%
Costa	11.01%	9.44%	7.75%	4.68%	5.73%
Sierra	8.72%	10.85%	10.14%	10.27%	12.55%
					100%

Region	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Amazonia	1.42%	0.97%	2.80%	2.61%	1.06%
Costa	10.67%	6.76%	10.85%	8.14%	2.18%
Sierra	3.97%	4.08%	16.50%	20.24%	7.74%
					100%

Provincia	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Bolivar	459	494	579	498	434
Cotopaxi	956	1407	1552	1686	1335
Guayas	5883	4934	3594	1994	2110
Pichincha	17460	18184	15944	13716	19160
Tungurahua	2584	3327	3220	2968	2733
Otros	883	913	825	588	645
					131065

Provincia	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Bolivar	217	197	582	1006	462
Cotopaxi	320	355	2129	3070	1062
Guayas	6165	3841	5113	2645	751
Pichincha	12196	8758	25583	27939	9988
Tungurahua	1179	1702	5102	5122	1727
Otros	600	543	1051	1134	526

Provincia	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Bolivar	0.35%	0.38%	0.44%	0.38%	0.33%
Cotopaxi	0.73%	1.07%	1.18%	1.29%	1.02%
Guayas	4.49%	3.76%	2.74%	1.52%	1.61%
Pichincha	13.32%	13.87%	12.16%	10.47%	14.62%
Tungurahua	1.97%	2.54%	2.46%	2.26%	2.09%
Otros	0.67%	0.70%	0.63%	0.45%	0.49%
					100%

Provincia	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Bolivar	0.17%	0.15%	0.44%	0.77%	0.35%
Cotopaxi	0.24%	0.27%	1.62%	2.34%	0.81%
Guayas	4.70%	2.93%	3.90%	2.02%	0.57%
Pichincha	9.31%	6.68%	19.52%	21.32%	7.62%
Tungurahua	0.90%	1.30%	3.89%	3.91%	1.32%
Otros	0.46%	0.41%	0.80%	0.87%	0.40%
					100%

Tabla A.10: IL Real y Estimado con BDD Modelamiento para G2

BDD Validación

Edad	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[18,30)	310	401	465	470	576
[30,40)	272	459	684	933	1167
[40,55)	318	459	624	1082	1790
[55,65)	185	247	325	494	831
[65,100)	120	206	270	431	522
					13641

Edad	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[18,30)	61	101	1170	735	155
[30,40)	63	141	1168	1491	652
[40,55)	56	122	1222	1821	1052
[55,65)	20	49	592	901	520
[65,100)	8	36	561	663	281

Edad	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[18,30)	2.27%	2.94%	3.41%	3.45%	4.22%
[30,40)	1.99%	3.36%	5.01%	6.84%	8.56%
[40,55)	2.33%	3.36%	4.57%	7.93%	13.12%
[55,65)	1.36%	1.81%	2.38%	3.62%	6.09%
[65,100)	0.88%	1.51%	1.98%	3.16%	3.83%
					100%

Edad	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
[18,30)	0.45%	0.74%	8.58%	5.39%	1.14%
[30,40)	0.46%	1.03%	8.56%	10.93%	4.78%
[40,55)	0.41%	0.89%	8.96%	13.35%	7.71%
[55,65)	0.15%	0.36%	4.34%	6.61%	3.81%
[65,100)	0.06%	0.26%	4.11%	4.86%	2.06%
					100%

E Civil	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	687	912	1081	1186	1376
2	381	618	983	1861	3121
3	97	172	225	250	289
4	36	60	76	99	78
5	3	4	0	4	5
					13604

E Civil	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	119	241	2400	2004	478
2	68	141	1764	2990	2001
3	15	49	379	446	144
4	2	9	150	157	31
5	0	3	5	3	5

E Civil	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	5.05%	6.70%	7.95%	8.72%	10.11%
2	2.80%	4.54%	7.23%	13.68%	22.94%
3	0.71%	1.26%	1.65%	1.84%	2.12%
4	0.26%	0.44%	0.56%	0.73%	0.57%
5	0.02%	0.03%	0.00%	0.03%	0.04%
					100%

E Civil	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	0.87%	1.77%	17.64%	14.73%	3.51%
2	0.50%	1.04%	12.97%	21.98%	14.71%
3	0.11%	0.36%	2.79%	3.28%	1.06%
4	0.01%	0.07%	1.10%	1.15%	0.23%
5	0.00%	0.02%	0.04%	0.02%	0.04%
					100%

Genero	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	571	943	1336	1848	2544
2	633	823	1029	1552	2325
					13604

Genero	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	96	197	2536	3100	1313
2	108	246	2162	2500	1346

Genero	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	4.20%	6.93%	9.82%	13.58%	18.70%
2	4.65%	6.05%	7.56%	11.41%	17.09%
					100%

Genero	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
1	0.71%	1.45%	18.64%	22.79%	9.65%
2	0.79%	1.81%	15.89%	18.38%	9.89%
					100%

Region	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Amazonia	157	254	325	320	452
Costa	28	58	78	208	554
Sierra	1020	1460	1965	2882	3880
					13641

Region	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Amazonia	27	44	564	656	217
Costa	47	84	314	344	137
Sierra	134	321	3835	4611	2306

Region	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Amazonia	1.15%	1.86%	2.38%	2.35%	3.31%
Costa	0.21%	0.43%	0.57%	1.52%	4.06%
Sierra	7.48%	10.70%	14.41%	21.13%	28.44%
					100%

Region	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Amazonia	0.20%	0.32%	4.13%	4.81%	1.59%
Costa	0.34%	0.62%	2.30%	2.52%	1.00%
Sierra	0.98%	2.35%	28.11%	33.80%	16.90%
					100%

Region	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Bolivar	39	89	110	98	119
Cotopaxi	100	156	244	343	302
Guayas	13	33	46	131	369
Pichincha	205	255	332	426	786
Tungurahua	336	316	217	287	373
Otros	436	807	1273	1980	2770
					12991

Region	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Bolivar	4	10	165	186	90
Cotopaxi	11	24	335	522	253
Guayas	32	58	211	207	84
Pichincha	37	84	681	840	362
Tungurahua	42	101	464	532	390
Otros	68	153	2615	3037	1393

Provincia	Ingreso Real				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Bolivar	0.30%	0.69%	0.85%	0.75%	0.92%
Cotopaxi	0.77%	1.20%	1.88%	2.64%	2.32%
Guayas	0.10%	0.25%	0.35%	1.01%	2.84%
Pichincha	1.58%	1.96%	2.56%	3.28%	6.05%
Tungurahua	2.59%	2.43%	1.67%	2.21%	2.87%
Otros	3.36%	6.21%	9.80%	15.24%	21.32%
					100%

Provincia	Ingreso Estimado				
	[450-550]	(550-700]	(700-900]	(900-1300]	(1300-5000]
Bolivar	0.03%	0.08%	1.27%	1.43%	0.69%
Cotopaxi	0.08%	0.18%	2.58%	4.02%	1.95%
Guayas	0.25%	0.45%	1.62%	1.59%	0.65%
Pichincha	0.28%	0.65%	5.24%	6.47%	2.79%
Tungurahua	0.32%	0.78%	3.57%	4.10%	3.00%
Otros	0.52%	1.18%	20.13%	23.38%	10.72%
					100%

Tabla A.11: IL Real y Estimado con BDD Validación para G2

A.6 IL reales y estimados con BDD de modelamiento y validación para grupo G3

BDD Modelamiento

Edad	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
[18,30)	2744	2197	1554	1264	673
[30,40)	4950	4934	4337	4176	2982
[40,55)	4687	4609	5668	6497	5371
[55,65)	1801	2009	2738	3127	2368
[65,100)	859	1186	1683	1703	1179
					75296

Edad	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
[18,30)	1970	2654	2001	1229	577
[30,40)	3667	5225	4978	4229	3275
[40,55)	3518	5397	5976	6055	5886
[55,65)	1281	2360	2846	2835	2721
[65,100)	680	1363	1602	1624	1341

Edad	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
[18,30)	3.64%	2.92%	2.06%	1.68%	0.89%
[30,40)	6.57%	6.55%	5.76%	5.55%	3.96%
[40,55)	6.22%	6.12%	7.53%	8.63%	7.13%
[55,65)	2.39%	2.67%	3.64%	4.15%	3.14%
[65,100)	1.14%	1.58%	2.24%	2.26%	1.57%
					100%

Edad	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
[18,30)	2.62%	3.52%	2.66%	1.63%	0.77%
[30,40)	4.87%	6.94%	6.61%	5.62%	4.35%
[40,55)	4.67%	7.17%	7.94%	8.04%	7.82%
[55,65)	1.70%	3.13%	3.78%	3.77%	3.61%
[65,100)	0.90%	1.81%	2.13%	2.16%	1.78%
					100%

E Civil	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	8509	6935	6076	5889	4016
2	5206	6426	8045	9120	7248
3	951	1075	1284	1295	1013
4	303	404	506	413	259
5	47	75	53	40	27
					75215

E Civil	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	6672	8176	7195	5376	4003
2	3454	6981	8392	8944	8272
3	714	1292	1262	1193	1157
4	203	487	477	395	323
5	35	53	60	55	38

E Civil	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	11.31%	9.22%	8.08%	7.83%	5.34%
2	6.92%	8.54%	10.70%	12.13%	9.64%
3	1.26%	1.43%	1.71%	1.72%	1.35%
4	0.40%	0.54%	0.67%	0.55%	0.34%
5	0.06%	0.10%	0.07%	0.05%	0.04%
					100%

E Civil	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	8.87%	10.87%	9.57%	7.15%	5.32%
2	4.59%	9.28%	11.16%	11.89%	11.00%
3	0.95%	1.72%	1.68%	1.59%	1.54%
4	0.27%	0.65%	0.63%	0.53%	0.43%
5	0.05%	0.07%	0.08%	0.07%	0.05%
					100%

Genero	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	8644	8953	8704	9053	7473
2	6372	5962	7260	7704	5090
					75215

Genero	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	6486	9707	9536	8833	8259
2	4592	7282	7850	7130	5534

Genero	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	11.49%	11.90%	11.57%	12.04%	9.94%
2	8.47%	7.93%	9.65%	10.24%	6.77%
					100%

Genero	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	8.62%	12.91%	12.68%	11.74%	10.98%
2	6.11%	9.68%	10.44%	9.48%	7.36%
					100%

Region	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Amazonia	1512	1641	1561	1832	1377
Costa	6237	4274	3975	3893	3019
Sierra	7292	9020	10444	11042	8177
					75296

Region	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Amazonia	1003	1761	1683	1777	1698
Costa	5974	4574	4439	3599	2810
Sierra	4139	10664	11281	10596	9292

Region	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Amazonia	2.01%	2.18%	2.07%	2.43%	1.83%
Costa	8.28%	5.68%	5.28%	5.17%	4.01%
Sierra	9.68%	11.98%	13.87%	14.66%	10.86%
					100%

Region	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Amazonia	1.33%	2.34%	2.24%	2.36%	2.26%
Costa	7.93%	6.07%	5.90%	4.78%	3.73%
Sierra	5.50%	14.16%	14.98%	14.07%	12.34%
					100%

Provincia	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Bolivar	293	378	335	242	116
Cotopaxi	706	1254	1367	949	485
Guayas	2409	1546	1255	1179	1111
Pichincha	8329	8017	9444	11458	9074
Tungurahua	1960	2447	2359	1849	1071
Otros	607	556	613	583	451
					72443

Provincia	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Bolivar	124	299	412	364	165
Cotopaxi	261	1245	1406	1172	677
Guayas	2607	1562	1297	1038	994
Pichincha	5804	10057	10652	10276	9530
Tungurahua	1418	2507	2342	1939	1479
Otros	336	589	635	652	598

Provincia	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Bolivar	0.40%	0.52%	0.46%	0.33%	0.16%
Cotopaxi	0.97%	1.73%	1.89%	1.31%	0.67%
Guayas	3.33%	2.13%	1.73%	1.63%	1.53%
Pichincha	11.50%	11.07%	13.04%	15.82%	12.53%
Tungurahua	2.71%	3.38%	3.26%	2.55%	1.48%
Otros	0.84%	0.77%	0.85%	0.80%	0.62%
					100%

Provincia	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Bolivar	0.17%	0.41%	0.57%	0.50%	0.23%
Cotopaxi	0.36%	1.72%	1.94%	1.62%	0.93%
Guayas	3.60%	2.16%	1.79%	1.43%	1.37%
Pichincha	8.01%	13.88%	14.70%	14.18%	13.16%
Tungurahua	1.96%	3.46%	3.23%	2.68%	2.04%
Otros	0.46%	0.81%	0.88%	0.90%	0.83%
					100%

Tabla A.12: IL Real y Estimado con BDD Modelamiento para G3

BDD Validación

Edad	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
[18,30)	134	302	401	368	194
[30,40)	274	623	1244	1237	719
[40,55)	325	644	1553	1872	1376
[55,65)	219	295	766	781	556
[65,100)	123	224	448	424	275
					15377

Edad	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
[18,30)	111	403	401	296	188
[30,40)	190	929	1027	1002	948
[40,55)	173	1036	1309	1544	1707
[55,65)	89	532	624	671	701
[65,100)	66	335	390	385	318

Edad	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
[18,30)	0.87%	1.96%	2.61%	2.39%	1.26%
[30,40)	1.78%	4.05%	8.09%	8.04%	4.68%
[40,55)	2.11%	4.19%	10.10%	12.17%	8.95%
[55,65)	1.42%	1.92%	4.98%	5.08%	3.62%
[65,100)	0.80%	1.46%	2.91%	2.76%	1.79%
					100%

Edad	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
[18,30)	0.72%	2.62%	2.61%	1.92%	1.22%
[30,40)	1.24%	6.04%	6.68%	6.52%	6.17%
[40,55)	1.13%	6.74%	8.51%	10.04%	11.10%
[55,65)	0.58%	3.46%	4.06%	4.36%	4.56%
[65,100)	0.43%	2.18%	2.54%	2.50%	2.07%
					100%

E Civil	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	430	810	1317	1200	725
2	491	1016	2592	3060	2119
3	109	182	372	297	200
4	44	71	121	105	62
5	1	8	5	13	10
					15360

E Civil	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	302	1226	1122	997	835
2	232	1588	2257	2515	2684
3	66	298	273	282	241
4	23	113	88	87	92
5	2	7	8	11	9

E Civil	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	2.80%	5.27%	8.57%	7.81%	4.72%
2	3.20%	6.61%	16.88%	19.92%	13.80%
3	0.71%	1.18%	2.42%	1.93%	1.30%
4	0.29%	0.46%	0.79%	0.68%	0.40%
5	0.01%	0.05%	0.03%	0.08%	0.07%
					100%

E Civil	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	1.97%	7.98%	7.30%	6.49%	5.44%
2	1.51%	10.34%	14.69%	16.37%	17.47%
3	0.43%	1.94%	1.78%	1.84%	1.57%
4	0.15%	0.74%	0.57%	0.57%	0.60%
5	0.01%	0.05%	0.05%	0.07%	0.06%
					100%

Genero	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	496	1213	2578	2753	1876
2	579	874	1829	1922	1240
					15360

Genero	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	311	1820	2127	2255	2403
2	314	1412	1621	1637	1458

Genero	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	3.23%	7.90%	16.78%	17.92%	12.21%
2	3.77%	5.69%	11.91%	12.51%	8.07%
					100%

Genero	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
1	2.02%	11.85%	13.85%	14.68%	15.64%
2	2.04%	9.19%	10.55%	10.66%	9.49%
					100%

Region	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Amazonia	122	260	487	474	305
Costa	42	105	400	492	354
Sierra	911	1723	3525	3716	2461
					15377

Region	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Amazonia	74	364	398	422	390
Costa	101	255	331	353	352
Sierra	454	2616	3022	3123	3120

Region	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Amazonia	0.79%	1.69%	3.17%	3.08%	1.98%
Costa	0.27%	0.68%	2.60%	3.20%	2.30%
Sierra	5.92%	11.21%	22.92%	24.17%	16.00%
					100%

Region	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Amazonia	0.48%	2.37%	2.59%	2.74%	2.54%
Costa	0.66%	1.66%	2.15%	2.30%	2.29%
Sierra	2.95%	17.01%	19.65%	20.31%	20.29%
					100%

Provincia	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Bolivar	30	99	178	146	54
Cotopaxi	114	254	432	397	274
Guayas	25	66	234	266	240
Pichincha	139	345	798	1002	718
Tungurahua	373	237	358	437	265
Otros	323	979	2234	2308	1518
					14843

Provincia	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Bolivar	15	124	123	139	106
Cotopaxi	50	312	376	370	363
Guayas	75	153	199	207	197
Pichincha	101	543	673	784	900
Tungurahua	112	361	452	366	378
Otros	245	1595	1782	1906	1834

Provincia	Ingreso Real				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Bolivar	0.20%	0.67%	1.20%	0.98%	0.36%
Cotopaxi	0.77%	1.71%	2.91%	2.67%	1.85%
Guayas	0.17%	0.44%	1.58%	1.79%	1.62%
Pichincha	0.94%	2.32%	5.38%	6.75%	4.84%
Tungurahua	2.51%	1.60%	2.41%	2.94%	1.79%
Otros	2.18%	6.60%	15.05%	15.55%	10.23%
					100%

Provincia	Ingreso Estimado				
	[450-850]	(850-1250]	(1250-2000]	(2000-4000]	(4000-35000]
Bolivar	0.10%	0.84%	0.83%	0.94%	0.71%
Cotopaxi	0.34%	2.10%	2.53%	2.49%	2.45%
Guayas	0.51%	1.03%	1.34%	1.39%	1.33%
Pichincha	0.68%	3.66%	4.53%	5.28%	6.06%
Tungurahua	0.75%	2.43%	3.05%	2.47%	2.55%
Otros	1.65%	10.75%	12.01%	12.84%	12.36%
					100%

Tabla A.13: IL Real y Estimado con BDD Validación para G3