# ESCUELA POLITÉCNICA NACIONAL

## FACULTAD DE CIENCIAS

# OPTIMIZACIÓN DE GENERACIÓN DE CONTACTOS TELEFÓNICOS EN GESTIÓN DE COBRANZAS MEDIANTE UN MODELO MULTINOMIAL DE MEJOR HORARIO DE LLAMADA (BEST TIME TO CALL)

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

ANDRÉS SEBASTIÁN CARRERA SÁNCHEZ
andres09_seb@hotmail.com

DIRECTORA: PHD. ADRIANA UQUILLAS ANDRADE
adriana.uquillas@epn.edu.ec

QUITO, JULIO 2017

## DECLARACIÓN

Yo ANDRÉS SEBASTIÁN CARRERA SÁNCHEZ, declaro bajo juramento que el trabajo aquí escrito 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.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

Andrés Sebastián Carrera Sánchez

## CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por
ANDRÉS SEBASTIÁN CARRERA SÁNCHEZ, bajo mi supervisión.

PhD. Adriana Uquillas Andrade
Directora del Proyecto

# Índice de contenido

# Índice de figuras

# Índice de tablas

# Resumen

En los últimos años, se han desarrollado industrias en torno a la facilidad para la gestión de procesos para atención a los clientes y sus interacciones con sus proveedores. Una de estas son los Centros de Contacto Telefónico, llamados también Contact Center o Call Center, que ya son toda una industria consolidada a nivel mundial y que mueve miles de millones de dólares en el mundo. Una vez que un Call Center ya posee lo último en tecnología (discador predictivo, ACD, grabador, todo en tecnología IP, etc.) ¿Qué queda por hacer para diferenciarse de los pares con igual nivel tecnológico?. La respuesta está en aplicar Inteligencia de Negocios, a través de prácticas analíticas, estadística aplicada, big data, data mining entre otras.

El objetivo principal de todos los proyectos desarrollados bajo esta filosofía es la de optimizar los procesos de un contact center, apuntando a mejorar la eficiencia y la efectividad, en un entorno de negocios exigente, competitivo y que requiere de mucha flexibilidad.

Las soluciones más conocidas en el ámbito de las plataformas de generación y recepción de contactos, son las de Best Time To Call y las de gestión de fuerza de trabajo. Ambas soluciones apuntan principalmente a ahorro de recursos, y a la automatización de servicios. El presente trabajo se centra en soluciones de Best Time To Call, que buscan discar cada registro a la hora de mayor probabilidad de contacto y/o en busca de un comportamiento determinado.

Estimar el mejor momento para ponerse en contacto con un cliente incluye la estimación de un modelo estadístico que calcula una puntuación para determinar el éxito de un contacto con el cliente en distintos horarios durante el día, basado en un conjunto de datos históricos de contacto y de comportamiento. El querer estimar el mejor momento u horario para ponerse en contacto con un cliente lleva directamente a considerar más de dos posibilidades, es decir nos enfrentamos a un problema de respuesta multicategórca o multinomial.

El modelo consigue relacionar los efectos de las diferentes variables o factores con la probabilidad de ponerse en contacto con un cliente.

# Abstract

In recent years, industries have developed around the ease of managing both customers and their interactions with their suppliers. One of these is the Telephone Contact Centers, also called Contact Center or Call Center, which are already an industry consolidated worldwide and that moves billions of dollars in the world. Once a Call Center already has the latest technology (predictive dialer, ACD, recorder, everything in IP technology, etc.), what remains to be done to differentiate from the peers with the same technological level ?. The answer lies in applying Business Intelligence, through analytical practices, applied statistics, big data, data mining among others.

The main objective of all the projects developed under this philosophy is to optimize the processes of a contact center, aiming at improving efficiency and effectiveness, in a demanding, competitive and demanding business environment.

The best known solutions in the field of contact generation and reception platforms are Best Time To Call and workforce management. Both solutions aim mainly to save resources, and automate services. The present work focuses on solutions of Best Time To Call, which seek to dial each register at the time of greater probability of contact and / or in search of a determined behavior.

Estimating the best time to contact a customer may include estimating a statistical model that calculates a score to determine the success of a customer contact at different times during the day based on a set of historical contact data and Of behavior. Wanting to estimate the best time or time to get in touch with a client leads directly to considering more than two possibilities, that is to say, we face a multi-homogeneous or multinomial response problem.

The model manages to relate the effects of different variables or factors to the likelihood of contacting a customer.

# Capítulo 1

# Introducción: Aspectos generales en la gestión telefónica de cobranzas, estrategias usuales y optimización.

En los últimos años, se han desarrollado industrias en torno a la facilidad de gestionar tanto a los clientes como sus interacciones con sus proveedores. Una de estas son los Centros de Contacto Telefónico (Contact Center o Call Center) que ya son toda una industria consolidada a nivel mundial y que mueve miles de millones de dólares en el mundo.

La gestión de cobranza es una de las actividades más solicitadas por los clientes de un contact center ya que es una etapa fundamental en la administración de créditos masivos, por tanto, si no se cuenta con las herramientas que permitan un proceso efectivo y ágil se pueden generar desincentivos, por parte de los clientes a quienes se les asignaron los créditos, para el pago de sus obligaciones. Debido a la gran cantidad de endeudamiento en el Ecuador, los contact center enfrentan un desafío común y constante: cobrar más rápido y mejor sin gastar más, es decir existe una necesidad de generar estrategias que den mejores resultados en la cobranza y que permitan adelantarse al resto de acreedores. Las estrategias usuales en gestión de cobranza se basan en segmentación de carteras por días mora y tipo de producto y en utilización de canales de contacto como llamadas telefónicas, visitas de campo, envío de SMS o mails con el fin de obtener una respuesta positiva por parte del cliente para el cumplimiento de sus obligaciones.

A medida que la edad de mora avanza, la probabilidad de recuperación disminuye, porque entre otras cosas, los clientes no son contactados facilmente y en muchos casos son inubicables.

La gestión de cobranza está relacionada con las estrategias, los canales de contacto y la segmentación por tanto se deben definir claramente las acciones sobre estas tres variables para obtener resultados. Al analizar los canales de contacto, se deben considerar el nivel de impacto, el costo y la interacción con el cliente ya que existen varias alternativas. En base a la experiencia y el uso de distintos canales de contacto durante varios años se establecen los siguientes canales con sus impactos y costos.

**Tabla 1.1:** CANALES DE CONTACTO
Fuente: Sic Contac Center S.A

| **Canal de Contacto** | **Costo** | **Interacción** | **Impacto** |
|---|---|---|---|
| Telefonía Manual | Medio | Alta | Alto |
| Telefonía Marcado Automático | Bajo | Baja | Bajo |
| Visita | Alto | Alta | Alto |
| e-mail | Bajo | Media | Bajo |
| SMS | Medio | Baja | Medio |
| Chat | Bajo | Media | Alto |

Dado que la gestión telefónica manual tiene un costo medio, impacto e interacción altos, la gestión de cobranza se realiza com mayor prioridad por este canal y para ello, se definen los denominados árboles de gestión que son parametrizables por tipo de empresa. El árbol de gestión telefónica, por ejemplo, incorpora acciones desde el proceso de marcado hasta el manejo de las objeciones por parte del gestor, con esto se logra una estandarización tanto en los procesos de llamadas como en los resultados de la gestión y se obtienen indicadores para cada acción. A continuación un ejemplo de árbol de gestión para telefonía.

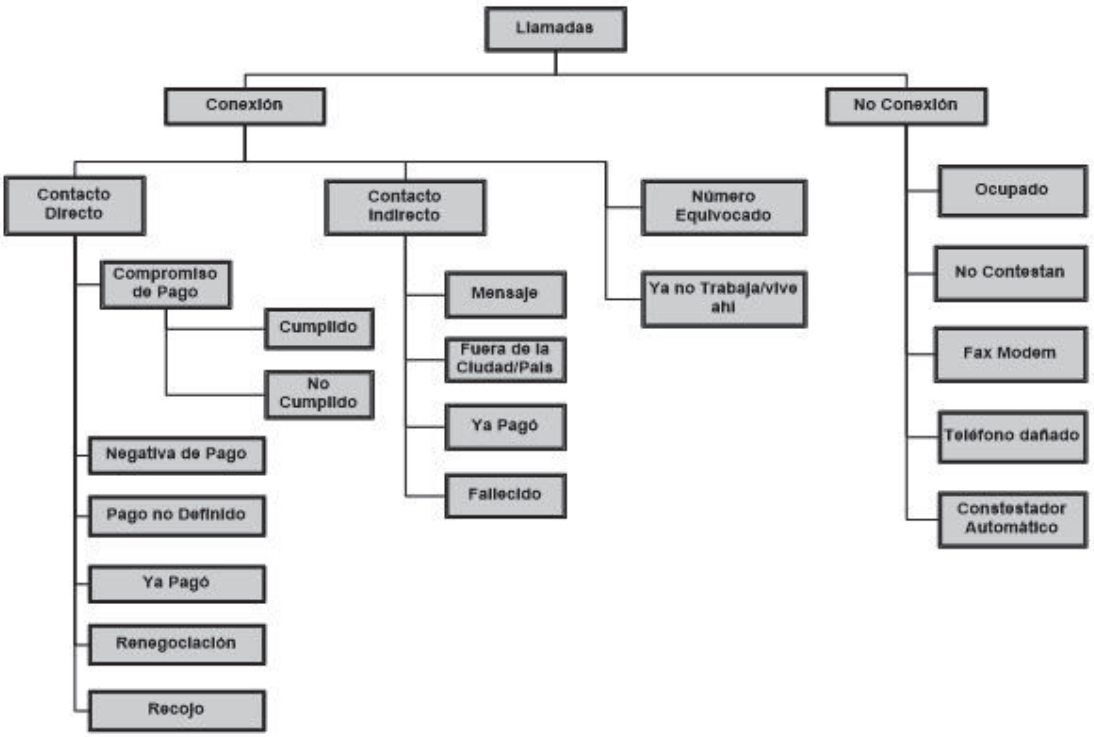


**Figura 1.1:** Árbol de Gestión

## 1.1. Indicadores de gestión y recuperación

Los indicadores de gestión se relacionan con el proceso realizado en el canal correspondiente, por lo que su definición se basa en las respuestas del árbol construido. Para el caso de gestión telefónica, los indicadores nacen del resultado de la llamada, sin embargo, es de vital importancia tener estadísticas sobre el proceso global, es decir ir desde lo macro a lo micro. Lo primero que se debe garantizar, es que toda la cartera sea gestionada, por tanto, el primer indicador se refiere a la cantidad de clientes gestionados y se lo conoce como **barrido de cartera** y se lo calcula como el número de casos gestionados sobre el número de asignados.

$$\%barrido = 100 * \frac{\sharp gestionados}{\sharp asignados}$$

Basados en el árbol de gestión, se define la **tasa de conexión** que permite analizar la estrategia de mejor hora, mejor día, manejo de agenda, actualización de información o reporte de avería en la línea, se lo calcula como el número de llamadas conectadas

sobre el número de llamadas totales.

$$\%tasa\_de\_conexion = 100 * \frac{\sharp conexiones}{\sharp llamadas\_totales}$$

Luego, se puede determinar la **tasa de contactos efectivos**, que no es más que el nivel de contacto con el titular de la deuda, esta tasa es una de las variables más importantes, ya que a mayor nivel de contacto efectivo mayor negociación y por tanto mayor recuperación. Se lo obtiene como el número de contactos efectivos sobre el total de conexiones.

$$\%tasa\_de\_contactos\_efectivos = 100 * \frac{\sharp contactos\_efectivos}{\sharp conexiones}$$

Al continuar con el proceso, se evalúa la capacidad del gestor para llegar a un compromiso de pago, aquí se mide el **porcentaje de compromisos de pago** realizados, calculado como la cantidad de promesas de pago sobre el número de contactos efectivos realizados.

$$\%compromisos\_de\_pago = 100 * \frac{\sharp compromisos\_de\_pago}{\sharp contactos\_efectivos}$$

Finalmente, el resultado del árbol se cierra con la recuperación mediante el **cumplimiento de los compromisos de pago** y el análisis de las causas de los compromisos incumplidos. El indicador de cumplimiento permite mejorar las capacidades del gestor y planificar las estrategias en función de los datos históricos, este indicador se lo obtiene de la razón entre el total de los pagos realizados y los compromisos registrados.

$$\%cumplimiento\_compromisos\_pago = 100 * \frac{\sharp pagos\_realizados}{\sharp compromisos\_de\_pago}$$

Los indicadores de recuperación están más ligados a definir las estrategias para generar el resultado esperado por edad de mora.

La industria de cobranzas mantiene algunos estándares de medición dependiendo del producto y el canal de gestión.

## 1.2. Optimización de gestión telefónica

Uno de los factores claves en el éxito de generación de contactos, es la estrategia de gestión y mejoramiento continuo de esta, basándose en resultados previos de contactabilidad. Tanto en venta como en cobranza, el aumento de la contactabilidad entrega beneficios directos al negocio, simplemente por los ahorros en recursos que es posible generar y por el aumento neto de negocios que se pueden lograr.

La contactabilidad se mide directamente como contactos efectivos sobre los contactos realizados y depende de varios factores. Algunos de estos son:

- **La calidad de los datos:** Este componente es el más importante para una buena contactabilidad. Una buena base de datos, donde cada registro posee uno o más números telefónicos, todos correctos y bien calificados (casa, trabajo, celular, etc.), debería dar como resultado de gestión, una contactabilidad efectiva alta, es decir, sólo con al persona que se busca y no a un familiar o referente.

- **La estrategia de discado o marcación:** Esto es parte de la gestión que el supervisor o el grupo de trabajo a cargo de la estrategia de negocios deben definir. Generalmente, las estrategias se aplican basadas en el perfil que presenten los registros. Si un registro indica que el cliente (o prospecto) es empleado de una compañía, lo lógico sería llamarlo a media mañana al número de su trabajo. Si el cliente indica que no posee trabajo, el llamado debe ser al teléfono de la casa o/y a su teléfono celular personal. Al mismo tiempo, esta estrategia debe considerar factores locales, como el caso de ciudades pequeñas donde las personas acostumbran almorzar en casa y el reingreso al trabajo ocurre a las 15 o 16 horas. Asumiendo que se poseen sistemas automatizados o predictivos, estrategias de discado deben ser discutidas en conjunto por el equipo de estrategia de negocios, y esperar que los sistemas de discado permitan programar la estrategia definida previamente.

  Otro factor que debe ser definido, es la política de intensidad de llamadas, que de manera automatizada o manual, se deben aplicar a los casos no exitosos de contacto. Hay que considerar que en algunos países esta intesidad de llamado están regulados, lo que debe tenerse en consideración. La política de intensidad, determina la cantidad y frecuencia de llamadas que se realizan ante resultados que difieren de aquellos que se consideran como caso cerrado(compromiso de pago cumplido o venta realizada). Al mismo tiempo, algunos casos determinan si un determinado número telefónico debe ser descartado luego de un número específico de intentos fallidos.

- **La tecnología de marcado disponible:** Existen casos donde no existe, o inclusive, no se requiere automatización tecnológica y otros casos donde es una norma de la industria. Los sistemas de marcado automático o marcado predictivo, generalmente permiten incorporar en sus campañas, las estrategias de marcado definidas previamente.

  Por otro lado, algunos sistemas permiten tener visibilidad del progreso de las campañas en tiempo real (o cercano al tiempo real) para identificar si una campaña está cumpliendo con las espectativas o si sus indicadores de desempeño son adecuados. Una característica de algunos sistemas de marcado predictivo, es que permiten realizar cambios en las estrategias en tiempo real, lo que facilita adaptar las estrategias al resultado esperado y al mismo tiempo, alinearlas con las campañas y obtener mejores resultados.

- **La adhesión a los procedimientos de los agentes:** Este tema, siempre ha sido una situación bastante complicada en todos los Call Center. El hecho de que un agente ingrese exactamente y siempre el resultado correcto de los llamados sería el resultado óptimo, sin embargo existen algunos casos en que los agentes están motivados a mentir en el resultado al finalizar una gestión, producto de los incentivos a los cuales están expuestos. Un ejemplo, es cuando un agente toma un contacto, que tiene como resultado real, 'Contacto con un Tercero' y el agente ingresa en el sistema 'Indeciso'. En este caso, el actuar de los agentes está motivado con el hecho que si el resultado es 'Indeciso', este prospecto queda reservado para él, y no vuelve a entrar a la lista de operaciones para gestionar (donde generalmente cada operación se distribuye al agente disponible en el momento, buscando la eficiencia del discador y no la conveniencia del agente). La motivación del agente, es el factor que puede determinar, si en un nuevo intento el contacto es efectivo, la posibilidad de cerrar la venta o la promesa de pago. Casos como este hay muchos y la razón más común va de la mano con las motivaciones y/o incentivos que tienen los agentes.

- **La tecnología de apoyo a la gestión disponible:** Cuando aparecieron los primeros sistemas de Call Center y se masificaron, entregaban muy poca información respecto al desempeño de las campañas. Algunos indicadores básicos y sobretodo, fuera de tiempo real, era lo que se disponía para las áreas de gestión, además de sistemas ineficientes de almacenamiento y procesamiento de información a pesar de tener carteras pequeñas,

  Hoy en día es muy difícil pensar que un sistema de gestión de marcado no posea herramientas de gestión y reporting. Actualmente, la información de gestión disponible es abundante. Tanto así que muchas veces no es posible aprovecharla

al 100 %. Una plataforma de marcado predictivo, es capaz de generar varios Terabytes (1 Terabyte = 1024 Gigabytes) de información durante un corto periodo de funcionamiento. Esta información requiere de herramientas de reporting avanzadas capaces de organizar la información en cubos multidimensionales para su despliegue y, sobretodo, para su entendimiento con el fin de mejorar la gestión.

- **La capacidad de aplicar inteligencia de negocios a la distribución de llamados:** En el punto anterior, se hace mención de las herramientas de reporting, capaces de generar cubos multidimensionales. Éstas son la base para la aplicación de Inteligencia de Negocios, con el objetivo de provocar mejoras significativas en las tasas de contactabilidad y penetración de las campañas. Existen soluciones de marcado que utilizan esta tecnología, como las de Best Time To Call, que buscan organizar los contactos durante el día, con el objetivo de aumentar la posibilidad de contacto efectivo, basado en estadísticas previas de contacto ya sea del mismo cliente o de su perfil. Estas soluciones tienen la capacidad de utilizar el conocimiento previo del contacto o del comportamiento de los clientes para aumentar la probabilidad de obtener un resultado de negocios deseado.

Intuitivamente, es natural aplicar la lógica de llamar a un cliente a la misma hora que lo he contactado en otras ocasiones, sin embargo, ¿cómo aplicarlo en una lista de 100.000 clientes?. ¿Qué pasa con aquellos que nunca he contactado previamente?. Las respuestas a estas interrogantes están basadas en modelos estadísticos y segmentación aplicada y son capaces de provocar mejoras importantes en la contactabilidad y en los resultados de negocios.

De los elementos mencionados anteriormente, existen algunos que no requieren de una gran inversión, aplicando políticas simples y apoyándose en la tecnología disponible. Sin embargo, los últimos dos puntos (Sistemas de Apoyo de Gestión e Inteligencia de Negocios), generalmente no forman parte de las soluciones estándares disponibles en el mercado. Estas soluciones son de proveedores distintos y forman toda una industria en la actualidad.

### 1.2.1. Best Time to Call (BTTC)

Una vez que un call center ya posee lo último en tecnología (discador predictivo, ACD, grabador, todo en tecnología IP, etc.), cabe hacerse la siguiente pregunta: ¿qué queda por hacer para diferenciarse de otros call center que poseen igual nivel tecnológico? La respuesta está en aplicar Inteligencia de Negocios, a través de prácticas analíticas, estadística aplicada, big data, entre otras.

El objetivo principal de todos los proyectos desarrollados bajo esta filosofía es la

de optimizar los procesos de un contact center, apuntando a mejorar la eficiencia y la efectividad, en un entorno de negocios cada vez más exigente, más competitivo y que requiere de mucha flexibilidad.

Según Douglas Conley [1],existen cuatro grandes grupos de ámbitos principales de aplicación:

- Optimización de Plataformas de generación de contactos
- Optimización de plataformas de recepción de contactos
- Optimización de plataformas híbridas (generación y recepción)
- Planeación, Gestión y Certificación de procesos

Las soluciones más conocidas en el ámbito de las plataformas de generación y recepción de contactos, son las de Best Time To Call y las de gestión de fuerza de trabajo. Ambas son soluciones que apuntan principalmente al ahorro de recursos y a la automatización de servicios.

El presente trabajo se centra en la optimización de las plataformas de generación de contactos, ya que estas plataformas, tienen su mejor desarrollo en negocios de venta, cobranza, promoción y en menor medida como apoyo operativo (mesas de ayuda, gestión de despacho, entre otros).

Estimar el mejor momento para ponerse en contacto con un cliente incluye la estimación de un modelo estadístico que calcula una puntuación para determinar el éxito de un contacto con el cliente para un período de tiempo, basado en un conjunto de datos históricos de contacto y de comportamiento.

Un modelo predictivo que estimaría la puntuación del mejor momento para ponerse en contacto con un cliente, es construido basado en no sólo estas características, sino también características adicionales que pueden mejorar el desempeño del modelo, construidas mediante un riguroso estudio de Data Mining que incluye la agregación, la transformación (log, raíz, identidad, etc.), la categorización o combinaciones de estas.

Esto puede incluir el número total de intentos de contacto y el número total de contactos de éxito por un periodo de tiempo, la suma de contacto ponderada en el tiempo, intentos y la suma de los contactos exitosos ponderada en el tiempo, el perfil de riesgo de un cliente, la ocupación, el estilo de vida del cliente, su trabajo, el tiempo de espera hasta el abandono de una llamada entrante, y otros.

El modelo consigue relacionar los efectos de las diferentes variables o factores con la probabilidad de ponerse en contacto con un cliente.

[1] DC. Ingeniero Electrónico de la Universidad Técnica Federico Santa María (Valparaiso. Chile), con 16 años de experiencia en la comercialización, Consultoría, Diseño, implementación y Mantención de Call Center y sistemas de gestión de clientes (CRM). Actualmente es Director de BORU Ltda. Empresa de Consultoría para Contact Center.

El querer estimar el mejor momento u horario para ponerse en contacto con un cliente lleva directamente a considerar más de dos posibilidades, es decir nos enfrentamos a un problema de respuesta multicategórica o multinomial. Por ejemplo, en el trabajo realizado por Halil BayraK [1], cada hora desde las 7:00 horas hasta 19:00 horas se considera una categoría o posible hora de contacto, por tanto en total se tendrán 12 categorías.

Tradicionalmente las variables dependientes con más de dos categorías han sido modeladas mediante análisis discriminante, pero gracias al desarrollo de las técnicas estadísticas es cada vez más habitual el uso de técnicas de regresión multinomial, debido a la mejor interpretabilidad de resultados que proporciona [2].

Los ingredientes de un modelo multinomial son los objetos de la elección, el conjunto de alternativas disponibles, las características observadas de los individuos y de las alternativas y el modelo de comportamiento o elección individual. El proceso de decisión del individuo $i$ puede representarse mediante una variable categórica $Y_i$ tal que $Y_i = j$ si el individuo elige la alternativa $j$.

Las alternativas deben ser mutuamente excluyentes y, además, exhaustivas, es decir, el conjunto de alternativas especificadas debe recoger todas las opciones posibles. Una vez especificada la variable dependiente, la probabilidad de que el individuo $i$ elija la alternativa $j$, $P(Y_i = j)$ , podrá expresarse como una función de un conjunto de factores, que pueden ser tanto características propias del individuo como características específicas de cada alternativa.

Para la ejecución del presente trabajo, se tomará información de gestión de cobranza y de los portafolios que administra SICCONTAC CENTER S.A.

SICONTAC CENTER S.A. es una empresa especialista en gestión de cobranza, que brinda sus servicios a entidades bancarias, aseguradoras, etc. Se usarán datos correspondientes a los tres productos más representativos de la empresa, estos son: **tarjeta crédito**, **microcrédito** y **prestamos de olla de oro**.

# Capítulo 2

# Marco Teórico: Regresión Logística Multinomial, Medidas de Separación o divergencia, Medidas de Asociación, Árboles de decisión

En este capítulo se explican todos los conceptos teóricos para comprender la metodología utilizada en el desarrollo del modelo de BTTC, se empezará describiendo los modelos regresión multinomial, la formulación, interpretación de parámetros, selección de variables y finalmente su validación. En la selección de variables se describen ciertos estadísticos e índices que miden la divergencia entre las distribuciones de probabilidad y la técnica de árboles de decisión.

Desde el punto de vista del individuo, formular un modelo de elección discreta en el que la probabilidad de elegir una alternativa se define como la probabilidad de que dicha alternativa tenga el mayor beneficio entre el conjunto de alternativas posibles.

Suponiendo que el proceso de decisión implica elegir entre $J+1$ alternativas, es decir, la variable dependiente $Y_i$ toma valores desde 0 hasta $J$, de modo que interesa evaluar $P(Y_i = j)$, $j = 0, ..., J$. El beneficio que obtiene el individuo $i$ de la alternativa j puede representarse mediante $U_{i,j}$, $j = 0, ..., J$. El beneficio de cada una de las alternativas no es observable por el investigador; sin embargo, depende de un conjunto de características del individuo, $x_i$ , y un conjunto de atributos propios de cada una de las alternativas, $s_j$, que sí son observables. Así, el componente determinístico o sistemático del beneficio de la alternativa $j$ para el individuo $i$ puede especificarse como $V_{i,j} = v_{i,j}(x_i, s_j)$.

La forma funcional que se utiliza generalmente para expresar el componente deter-

minístico es lineal en los parámetros, de modo que $V_{i,j} = \beta_j^x x_i + \beta^s s_j$, donde $\beta_j^x$ es un vector de parámetros que determina diferentes probabilidades para cada alternativa $j$ en función de las características del individuo $i$, mientras que $\beta^s$ es otro vector de parámetros que introduce diferencias en las probabilidades de elegir cada alternativa como función de los atributos propios de ésta.

Además, aparte de errores de medición, hay factores no observables por el analista que influyen en la utilidad y que no están incluidos en $V_{i,j}$; por ello, además del componente determinístico, debe considerarse un término aleatorio, cuya inclusión permite tener en cuenta que individuos aparentemente idénticos puedan escoger alternativas diferentes. Así, el beneficio del individuo para cada una de las alternativas queda recogida mediante la expresión $U_{i,j} = V_{i,j} + \varepsilon_{i,j}$. El individuo elegirá aquella alternativa que le proporcione el máximo beneficio. Así, se elegirá la alternativa $j$ si y sólo si $U_{i,j} > U_{i,k}$, $\forall k \neq j$. Entonces, la probabilidad de que el individuo $i$ elija la alternativa $j$ puede expresarse como

$$P(Y_i = j) = P(U_{i,j} > U_{i,k}, \forall k \neq j) = P(\varepsilon_{i,k} - \varepsilon_{i,j} < V_{i,j} - V_{j,k}, \forall k \neq j) \tag{2.1}$$

Así pues, el modelo finalmente especificado depende de la distribución del vector de términos de error $(\varepsilon_{i,0}, ..., \varepsilon_{i,J})$. Si $\varepsilon_i = (\varepsilon_{i,0}, ..., \varepsilon_{i,J}) \approx N(0, \Sigma)$, se obtiene el modelo *probit multinomial*.

La especificación del modelo probit multinomial ofrece un amplio rango de posibilidades en función de la estructura de correlación especificada para los términos de perturbación del vector $\varepsilon_i$. En particular, la especificación apropiada de la matriz $\Sigma$ permite considerar que el patrón de sustitución entre las alternativas $j$ y $k$, definido como cociente de las probabilidades de elección respectivas, se modifique cuando cambia alguno de los atributos que define una tercera alternativa m.

En muchas situaciones prácticas, es difícil admitir que dicho cociente de probabilidades se mantiene constante cuando cambia el beneficio que proporciona otra alternativa. Este supuesto rígido, poco frecuente en la práctica, se impone a priori en especificaciones menos flexibles como, por ejemplo, el modelo *logit multinomial*, que, sin embargo, presenta menos dificultades computacionales y, además, ofrece posibilidades de interpretación más ricas de la compleja red de efectos a menudo presente en los procesos de decisión.

De hecho, la aplicabilidad del modelo probit multinomial se ve restringida cuando el número de alternativas es elevado, dada la complejidad de las integrales que resultan y que deben ser evaluadas numéricamente. Esta complejidad ha limitado fuertemente el uso de estos modelos. Sin embargo, en los últimos años se han desarrollado métodos de aproximación basados en la simulación que amplían significativamente las posibilidades

de aplicación de esta especificación en situaciones reales.

En el modelo logit multinomial se exige que los elementos del vector de perturbaciones $\varepsilon_i = (\varepsilon_{i,0}, ..., \varepsilon_{i,J})$ sigan distribuciones de Gumbel independientes, es decir, la función de densidad de cada término de perturbación $\varepsilon_{i,j}$ es

$$f(\varepsilon_{i,j}) = \exp(-\varepsilon_{i,j}) \exp(-\exp(-\varepsilon_{i,j})) \tag{2.2}$$

Por tanto de (2.1) se obtiene que

$$P(Y_i = j) = \frac{\exp(V_{i,j})}{\sum_{k=0}^{J} \exp(V_{i,k})}, j = 0, ..., J \tag{2.3}$$

Esta especificación permite advertir uno de sus principales inconvenientes. Y es que el patrón de sustitución entre las alternativas $j$ y $k$ , definido como cociente de las probabilidades respectivas, sólo depende de la diferencia de beneficios entre dichas alternativas, pero no de los beneficios que proporcionan las demás.

Desde el punto de vista del enfoque de beneficio aleatorio, esta propiedad, conocida como independencia de alternativas irrelevantes (IIA), se deriva del supuesto distribucional sobre el vector $(\varepsilon_{i,0}, ..., \varepsilon_{i,J})$ y, más en concreto, del supuesto de igualdad de varianza e independencia para cada uno de sus elementos. Sin embargo, en muchas ocasiones la descripción adecuada del proceso de decisión exige considerar la presencia de estructuras de correlación entre los elementos del vector.

Desde este punto de vista, el modelo probit multinomial supone, como ya se comentó, una especificación más flexible, pero también se han desarrollado otras especificaciones. Ahora bien, la propiedad de IIA puede proporcionar una representación adecuada de la realidad en determinados casos y, entonces, la formulación del modelo logit supone ventajas prácticas. [3]

## 2.1. Regresión Logística Multinomial

Los modelos de regresión logística son modelos estadísticos en los que se pretende conocer la relación entre una variable dependiente cualitativa y variables explicativas independientes, que pueden ser cualitativas o cuantitativas. Cuando la variable dependiente es dicotómica (con dos categorías), se acostumbra usar la regresión logística binaria clásica, en el caso de que la variable dependiente sea politómica (con más de dos categorías) la opción que ha venido dando mejores resultados, además de ser menos complicada en su interpretación, es la regresión logística multinomial.

Para el caso en que las covariables cualitativas sean dicotómicas, se aconseja codificarlas dando valores de 0, para una de las categorías o para su ausencia y 1 para la otra categoría o para su presencia, esta codificación es importante, ya que cualquier otra codificación podría provocar modificaciones en la interpretación del modelo.

Si la covariable cualitativa tiene más de dos categorías, se realiza una transformación para incluirla en el modelo. Esta transformación consiste en crear varias variables dummies, es decir, crear variables cualitativas dicotómicas ficticias, de esta manera una de las variables se tomaría como categoría de referencia y cada una de las variables creadas entraría en el modelo de forma individual. De manera general, si la covariable cualitativa posee n categorías, habrá que realizar $n-1$ covariables dummies. También se pueden agrupar las categorías, de tal forma que las categorías que pertenezcan al grupo 1 serán marcadas como 0 y las que pertenezcan al grupo 2 se marcarán como 1, así se tendrá una nueva variable dicotómica dummy, en lugar de varias para una misma variable. Para agrupar las categorías se debe realizar un análisis bivariado con respecto a la variable dependiente. [4]

La regresión logística multinomial es usada para estimar modelos con variable dependiente cualitativa con más de dos categorías y es una extensión multivariante de la regresión logística binaria clásica. Las variables independientes pueden ser tanto continuas (covariables) como categóricas (factores).

Tradicionalmente las variables dependientes con más de dos categorías han sido modeladas mediante técnicas multivariantes como el análisis discriminante pero, gracias al creciente desarrollo de las técnicas de cálculo, se ha hecho cada vez es más habitual el uso de modelos de regresión logística multinomial que actualmente ya están implementados en paquetes estadísticos como SPSS o R, debido a la mejor interpretabilidad de los resultados que proporciona [2].

Los modelos multinomiales se analizan tomando una de las categorías de la variable dependiente como categoría de referencia o categoría pivote, y se modelan varias ecuaciones simultáneamente, una para cada una de las restantes categorías respecto a la de referencia.

### 2.1.1. Formulación del Modelo Multinomial

Empecemos recordando la regresión logística binaria, si tenemos una variable dependiente $Y$ de dos categorías, que toma valores $Y = 1$ (presencia de una característica u otra categoría de la variable) y $Y = 0$ (ausencia de la característica o la otra categoría

de la variable), la ecuación de partida del modelo viene dada por:

$$P[Y = 1|X] = \frac{\exp\left(b_0 + \sum_{s=1}^{n} b_s x_s\right)}{1 + \exp\left(b_0 + \sum_{s=1}^{n} b_s x_s\right)} \tag{2.4}$$

donde $P[Y = 1|X]$ es la probabilidad de que $Y$ tome le valor de 1, dado el conjunto covariables X, y se la denota por $p(X)$.

$X$ es el conjunto de n covariables $x_1, x_2, ..., x_n$ que forman parte del modelo, $b_0$ es el intercepto o la constante del modelo y los $b_s$ son los coeficientes de las covariables.

Esta ecuación inicial es de tipo exponencial pero se la puede expresar en forma logarítmica:

$$\ln\left[\frac{p(X)}{1 - p(X)}\right] = b_0 + \sum_{s=1}^{n} b_s x_s \tag{2.5}$$

De esta forma, el modelo es lineal y es de más fácil interpretación.

Para el caso en que la variable dependiente tenga más de dos categorías, como es en este caso, se utiliza el modelo de regresión logística multinomial, que se modela mediante varios modelos logits simultáneamente, uno para cada una de las restantes categorías respecto a la categoría de referencia que se haya considerado de la variable dependiente.

A continuación se presenta la formulación de estos modelos de forma general.

Sea $Y$ una variable de respuesta politómica con más de dos categorías de respuesta $Y_1, Y_2, \ldots, Y_k$.

Para explicar la probabilidad de cada categoría de respuesta en función de un conjunto de covariables $X = \{x_1, x_2, \ldots, x_n\}$ observadas, se debe ajustar un modelo de la forma

$$p_j(x) = P[Y = Y_j|X = x] = f_j(x), \quad \forall j = 1, ..., k \tag{2.6}$$

para cada vector x de valores observados de las variables explicativas X.

En el caso de una variable de respuesta binaria, su distribución condicionada a cada combinación de valores observados sigue una Bernouilli, mientras que si la variable de respuesta es politómica, la distribución de Bernoulli se convierte en una distribución multinomial cuyos parámetros son las probabilidades de cada una de las categorías de respuesta, es decir,

$$(Y|X = x) \to M(1; p_1(x), p_2(x), ..., p_k(x))$$

siendo $\sum_{j=1}^{k} p_j(x) = 1$.

Por tanto, para obtener un modelo lineal, se tendrán $\binom{k}{2}$ transformaciones logit para comparar cada par de categorías de la variable respuesta, que sería del tipo

$$\ln\left[\frac{\frac{p_i(x)}{p_i(x)+p_j(x)}}{\frac{p_j(x)}{p_i(x)+p_j(x)}}\right] = \ln\left[\frac{p_i(x)}{p_j(x)}\right], \qquad \forall i,j = 1,...,k \quad (i \neq j) \tag{2.7}$$

esta ecuación representa el logaritmo de la ventaja que tiene la categoría $Y_i$ frente a la categoría $Y_j$ sujeto a las observaciones de las variables independientes que caen en una de ambas categorías. Sin embargo, para construir el modelo logit de respuesta multinomial basta considerar $k-1$ transformaciones logit al tomar una categoría como referencia.

Sea $Y_k$ la categoría pivote, luego las transformaciones logit se definen como

$$L_j(x) = \ln\left[\frac{p_j(x)}{p_k(x)}\right] \quad \forall j = 1,...,k-1 \tag{2.8}$$

donde $L_j(x)$ es el logaritmo de la ventaja que tiene la categoría $Y_j$ dado que las observaciones de las variables independientes caen el la categoría $Y_j$ o en la categoría $Y_k$.

El modelo lineal para cada una de las transformaciones logit, para n variables independientes, está dado por

$$L_j(x) = \sum_{s=0}^{n} b_{sj}x_s = x'b_j \quad \forall j = 1,...,k-1 \tag{2.9}$$

para cada vector $x = (x_0, x_2, ..., x_n)'$ de las variables independientes con $x_0 = 1$ y $b_j = (b_{0j}, b_{1j}, ..., b_{nj})$ el vector de parámetros asociado a la categoría $Y_j$.

Luego, como

$$\frac{p_j(x)}{p_k(x)} = \exp(x'b_j) \tag{2.10}$$

entonces

$$\sum_{j=1}^{k-1} \frac{p_j(x)}{p_k(x)} = \sum_{j=1}^{k-1} \exp(x'b_j) \tag{2.11}$$

y como $\sum_{j=1}^{k-1} p_j + p_k = 1$ entonces

$$\frac{1 - p_k(x)}{p_k(x)} = \sum_{j=1}^{k-1} \exp(x' b_j) \tag{2.12}$$

y por tanto

$$p_k = \frac{1}{1 + \sum_{j=1}^{k-1} \exp\left(\sum_{s=0}^{n} b_{sj} x_s\right)} \tag{2.13}$$

y finalmente, una vez más de (2.10) se tiene

$$p_j(x) = \frac{\exp\left(\sum_{s=0}^{n} b_{sj} x_s\right)}{1 + \sum_{j=1}^{k-1} \exp\left(\sum_{s=0}^{n} b_{sj} x_s\right)} \quad \forall j = 1, ..., k-1 \tag{2.14}$$

### 2.1.2. Interpretación de parámetros

Para la interpretación de los parámetros del modelo se deben distinguir dos casos según el tipo de variables explicativas que contenga el modelo, sean estas cuantitativas o cualitativas.

- **Variables explicativas cuantitativas:** Suponiendo que se tiene una sola variable explicativa cuantitativa $X$, para cada valor observado $x \in X$ el modelo viene dado por

  $$L_j(x) = a_j + b_j x, \quad \forall j = 1, ..., k-1 \tag{2.15}$$

  Los parámetros $b_j$ asociados a las categorías $Y_j$ de la variable dependiente $Y$, se interpretan en términos de los cocientes de ventajas, llamados también ODDS RATIO.

  $$\theta_j(\Delta X = 1) = \frac{\frac{p_j(x+1)}{p_k(x+1)}}{\frac{p_j(x)}{p_k(x)}} = \frac{\exp(a_j + b_j(x+1))}{\exp(a_j + b_j x)} = \exp(b_j) \quad \forall j = 1, ..., k-1 \tag{2.16}$$

  $\theta_j(\Delta X = 1)$ es el cociente de ventajas de respuesta $Y_j$ frente a la categoría de referencia $Y_k$ cuando aumenta la variable $X$ en una unidad.

- **Varias variables explicativas cuantitativas:** Para el modelo de regresión logit multinomial los ODDS RATIO se definen incrementando una de las variables y

controlando fijas las demás.

$$\theta_j(\Delta X_r = 1|X_s = x_s, s \neq r) = \frac{\frac{P[Y=Y_j|X_r=x_r+1,X_s=x_s,s\neq r]}{P[Y=Y_k|X_r=x_r+1,X_s=x_s,s\neq r]}}{\frac{P[Y=Y_j|X_r=x_r,X_s=x_s,s\neq r]}{P[Y_k|X_r=x_r,X_s=x_s,s\neq r]}} = \exp(b_{rj}) \quad \forall j = 1, ..., k-1 \tag{2.17}$$

- **Variables explicativas cualitativas:** Como se había mencionado en la sección anterior, si el modelo incluye variables independientes categóricas, se introduce como variables dummy.

  Sea $A$ la variable categórica con $A_1, ..., A_p$ categorías, si se crean variables dummy asignando un uno a la variable asociada a cada categoría y un cero al resto y tomando como categoría de referencia la primera se obtienen $p-1$ variables que se las denota como $X_m^A$ $(m = 2, ..., p)$.

  Por tanto, como en los casos anteriores, el modelo de regresión logística multinomial sigue siendo lineal para cada logit en función de las variables dummy procedentes de la variable $A$, y está dado por

$$L_{j/l} = \ln\left[\frac{p_{j/l}}{p_{k/l}}\right] = b_{0j} + \sum_{m=2}^{p} \tau_{mj}^A X_{lm}^A \quad l = 1, ...p; \;\; j = 1, ..., k-1 \tag{2.18}$$

  donde $p_{j/l}$ es la probabilidad de respuesta $Y_j$ en la categoría $A_l$.

  O se lo puede definir de la siguiente manera

$$L_{j/l} = b_{0j} + \tau jl \quad l = 1, ...p; \;\; j = 1, ..., k-1 \tag{2.19}$$

  donde $\tau_{j1} = 0$, $\forall j = 1, ..., k-1$ que en términos de los ODDS RATIOS viene dado por

$$\theta_{j/l1} = \frac{\frac{p_{j/l}}{p_{k/l}}}{\frac{p_{j/1}}{p_{k/1}}} = \frac{\exp(b_{0j} + \tau_{lj})}{\exp(b_{0j})} = \exp(\tau_{lj}) \quad \forall l = 2, ...p; \;\; \forall j = 1, ..., k-1 \tag{2.20}$$

  que no es más que el cociente de ventajas de la respuesta $Y_j$ frente a la respuesta $Y_k$ para la categoría $A_l$ de $A$ respecto a la primera categoría $A_1$.

## 2.2. Selección de Variables

Para el desarrollo de cualquier modelo estadístico, es muy importante conocer el tipo y la calidad de los datos ya que la calidad del modelo dependerá directamente de la información utilizada. En el caso del modelo de mejor hora de llamada (BTTC) se quiere predecir la probabilidad de que un cliente sea contactado en un horario, es por esta razón que es necesario que la información utilizada permita identificar las características de quienes son individuos contactados en los horarios y quienes no.

En este caso, para el desarrollo del modelo de regresión logística se cuenta con una gran cantidad de datos, es por esto que viene la necesidad de realizar una selección de las variables explicativas, para esto, se separan las variables numéricas continuas de las variables categóricas y se usan estadísticos e índices para medir el poder predictivo de cada una de las variables.

### 2.2.1. Medidas de Separación o Divergencia

Las medidas de separación o divergencia nos ayudan a conocer el poder predictivo de las variables numéricas continuas. Para modelos de respuesta binaria, como la regresión logística clásica, se usan pruebas como la de Kolmogorov-Smirnov y la de Anderson Darling para seleccionar las mejores variables explicativas de acuerdo al valor que tengan estos estadísticos, es decir, a mayor valor del estadístico mayor poder de predicción de la variable. En este caso, la variable respuesta tiene más de dos categorías, por tanto surge la necesidad de extender este concepto para el caso multinomial.

En el presente trabajo se usa la prueba de Kolmogorov-Smirnov, la que se detalla a continuación.

#### Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov (también conocida como prueba K-S) es una prueba no paramétrica que determina la bondad de ajuste de dos distribuciones continuas independientes, contrastando la hipótesis de que si estas son idénticas o no.

Utilizando el trabajo realizado por Leanna L. [5], se describe la generalización de el estadístico KS para más de dos muestras, al que se lo denotará por KSM (Kolmogorov-Smirnov Measure).

Para desarrollar el estadístico KSM, primero se describe el estadístico KS. En lugar de buscar diferencias en los promedios o medianas de la muestra, la prueba de KS busca la mayor diferencia en las distribuciones empíricas $\widehat{F}$ de dos muestras. Dadas una variable con dos respuestas posibles $y \in \{1, 2\}$ y una variable $x_s$ se define el

estadístico KS de la prueba de la siguiente manera

$$KS_s = \max_a |\widehat{F}_{s,1}(a) - \widehat{F}_{s,2}(a)| \tag{2.21}$$

$$= \max_a \left| \frac{\sum_{i=1}^{N_1} \mathbb{1}\{x_{s,1,i} \leq a\}}{N_1} - \frac{\sum_{i=1}^{N_2} \mathbb{1}\{x_{s,2,i} \leq a\}}{N_2} \right| \tag{2.22}$$

donde $\widehat{F}_{s,k}$ es la función de distribución empírica de $x_s$ cuando $y = k$, $N_k$ es el número de observaciones en la categoría $k$ y $x_{s,k,i}$ es la i-ésima muestra de $x_s$ cuando la variable respuesta es igual a k. Como la prueba de KS se formula a menudo como una prueba de hipótesis típica, las diferencias significativas entre distribuciones se declaran a menudo comparando un valor de p (valor de probabilidad) con una tasa de error de tipo I.

Si la variable de dependiente contiene más de dos categorías ($K > 2$) no se puede usar la prueba KS para identificar las variables explicativas con mayor poder predictivo. Sin embargo, hay dos características de la prueba KS que se pueden aprovechar para extenderlo a al nivel multinomial.

La primera es que el estadístico KS se ajusta a la definición de una métrica de distancia y ,la segunda, debido a que los valores de las funciones de distribución empírica están en el intervalo [0,1], los valores del estadístico KS también están en el intervalo [0,1]. Valores cercanos a uno indican una mayor diferencia en las distribuciones de las muestras.

Tomando en cuenta estas características se define KSM como la suma ponderada de los valores de KS de todas $\binom{k}{2}$ combinaciones por variable. Los pesos se toman proporcionales al tamaño total de la muestra, de modo que la medida KSM está dada por

$$KSM_s = \sum_{k=1}^{K} \sum_{k \neq k'} \frac{N_k + N_{k'}}{N(K-1)} KS_s(k, k') \tag{2.23}$$

donde $KS_s(k, k')$ es el valor del estadístico KS al comparar las distribuciones de una variable $x_s$ cuando la la variable respuesta es $k$ o $k'$ y $N$ es el tamaño total de la muestra. Como la suma de los pesos es 1, el estadístico $KSM$ está en el intervalo [0,1] y tiene la misma interpretación que el estadístico KS, valores cercanos a uno indican una mayor diferencia en las distribuciones de $x_s$ cuando la variable de respuesta tiene multiples categorías.

### 2.2.2. Medidas de Asociación

Las medidas de asociación son indicadores que miden el poder predictivo de las variables categóricas consideradas importantes para formar parte del modelo. En la práctica, las más utilizadas son la prueba de independencia Ji-cuadrado($\chi^2$) y el Valor de Información (IV), el segundo es usado cuando la variable respuesta es binaria, por tanto, en el presente trabajo se utilizará la prueba Ji-cuadrado($\chi^2$).

#### Prueba de Independencia Ji-cuadrado($\chi^2$)

El propósito en el presente trabajo, es utilizar el estadístico $\chi^2$, para estudiar la dependencia entre la variable dependiente politómica y las variables explicativas categóricas, sean estas binarias o politómicas. [6]

Se considera entonces, una variable cualitativa arbitraria $X$ con $X_1, ..., X_p$ categorías y la variable dependiente $Y$ con $Y_1, ..., Y_k$ categorías, su representación típica en tabla de contingencia es la siguiente

**Tabla 2.1:** Tabla de Contingencia

| **Y\X** | $X_1$ | $X_2$ | ... | $X_j$ | ... | $X_p$ | **Totales** |
|---|---|---|---|---|---|---|---|
| $Y_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1j}$ | ... | $n_{1p}$ | $n_{1.}$ |
| $Y_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2j}$ | ... | $n_{2p}$ | $n_{2.}$ |
| . | | | | | | | . |
| . | | | | | | | . |
| . | | | | | | | . |
| $Y_i$ | $n_{i1}$ | $n_{i2}$ | ... | $n_{ij}$ | ... | $n_{ip}$ | $n_{i.}$ |
| . | | | | | | | . |
| . | | | | | | | . |
| . | | | | | | | . |
| $Y_k$ | $n_{k1}$ | $n_{k2}$ | ... | $n_{kj}$ | ... | $n_{kp}$ | $n_{k.}$ |
| **Totales** | $n_{.1}$ | $n_{.2}$ | ... | $n_{.j}$ | ... | $n_{.p}$ | n |

Donde

$$n_{i.} = \sum_{j=1}^{p} n_{ij} \qquad \forall i = 1, ..., k \tag{2.24}$$

$$n_{.j} = \sum_{i=1}^{k} n_{ij} \qquad \forall j = 1, ..., p \tag{2.25}$$

y es inmediato que

$$n = n_{i.} = \sum_{j=1}^{p} n_{ij} = n_{.j} = \sum_{i=1}^{k} n_{ij} \tag{2.26}$$

La prueba Ji-cuadrado contrasta la hipótesis nula de independencia de las variables $X$ e $Y$ versus la hipótesis alternativa de existencia de asociación entre estas variables a un determinado nivel de significación $\alpha$ en base a la información recogida en la tabla de contingencia.

$$H_0\text{: } X \text{ e } Y \text{ son independientes}$$
$$H_1\text{: } X \text{ e } Y \text{ no son independientes}$$

Se define el valor $n'_{ij}$ como la frecuencia esperada que correspondería al par de categorías ($Y_i$,$X_j$) y está dado por

$$n'_{ij} = \frac{n_{i.}n_{.j}}{n} \quad \forall i = \{1, 2, ..., k\}; \quad j = \{1, 2, ..., p\} \tag{2.27}$$

Y el valor del estadístico asociado a la prueba puede ser calculado por

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{(n'_{ij} - n_{ij})^2}{n'_{ij}} \tag{2.28}$$

La medida contraria a la independencia es la asociación y se dice que dos variables están asociadas si aparecen juntas en mayor número de veces que el esperado si fuesen independientes, por tanto, si se rechaza la hipótesis nula entonces existe asociación y según sea la tendencia a coincidir o no se tendrán distintos grados de asociación.

En general, en tablas de $k \times p$ se utiliza el **coeficiente de contigencia de Pearson** definido por

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} \tag{2.29}$$

Este coeficiente varía entre $0 \leq C \leq \sqrt{\frac{q-1}{q}} < 1$ donde $q = \text{mín}\{k, p\}$ y valores cercanos a 0 indicarán independencia entre las variables y valores cercanos a 1 indicarán que existe relación entre las mismas

### 2.2.3. Árboles de Decisión

Para crear las variables dummy o agrupar las categorías de las variables cualitativas, se usa la técnica de árboles de decisión, este procedimiento crea un modelo de clasificación basado en árboles que clasifica casos en grupos o pronostica valores de una variable dependiente basada en los valores de las variables independientes. [7]

Este procedimiento puede ser usado para:

- **Segmentación:** Identifica las personas que pueden ser miembros de un grupo específico.
- **Estratificación:** Asigna los casos a una categoría de entre varias, por ejemplo, grupos de alto riesgo, bajo riesgo y riesgo intermedio.
- **Predicción:** Crea reglas y las utiliza para predecir eventos futuros, como la verosimilitud de que una persona cause mora en un crédito o el valor de reventa potencial de un vehículo o una casa.
- **Reducción de datos y clasificación de variables:** Selecciona un subconjunto útil de predictores a partir de un gran conjunto de variables para utilizarlo en la creación de un modelo paramétrico formal.
- **Identificación de interacción:** Identifica las relaciones que pertenecen sólo a subgrupos específicos y las especifica en un modelo paramétrico formal.
- **Fusión de categorías y discretización de variables continuas:** Vuelve a codificar las variables continuas y las categorías de los predictores del grupo, con una pérdida mínima de información.

El proceso iterativo de crecimiento que sigue la técnica de árboles es el siguiente: Primero se particiona la población en dos subconjuntos homogéneos, luego cada uno de estos subconjuntos es particionado nuevamente, el proceso se repeite recursivamente y termina cuando un subconjunto presenta una cantidad de individuos menor o igual al mínimo requerido, finalmente se establece el tipo de subconjunto dependiendo de su distribución en las categorías de la variable dependiente respecto a la distribución en el conjunto inicial.

Los métodos de crecimiento disponibles son [7]:

- **CHAID:** Detección automática de interacciones mediante ji-cuadrado (CHi-square Automatic Interaction Detection). En cada paso, CHAID elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. Las categorías de cada predictor se funden si no son significativamente distintas respecto a la variable dependiente.
- **CHAID exhaustivo:** Una modificación del CHAID que examina todas las divisiones posibles de cada predictor.
- **CRT:** Árboles de clasificación y regresión. CRT divide los datos en segmentos para que sean lo más homogéneos que sea posible respecto a la variable dependiente. Un nodo terminal en el que todos los casos toman el mismo valor en la variable dependiente es un nodo homogéneo y "puro".

- **QUEST:** Árbol estadístico rápido, insesgado y eficiente (Quick, Unbiased, Efficient Statistica Tree). Método rápido y que evita el sesgo que presentan otros métodos al favorecer los predictores con muchas categorías. Sólo puede especificarse QUEST si la variable dependiente es nominal.

## 2.3. Validación de la Especificación del Modelo de Regresión Logística Multinomial

De igual forma que en la regresión logística clásica, para estimar los coeficientes del modelo de regresión multinomial, se usa el método de máxima verosimilitud que consiste en encontrar los valores de los coeficientes que maximicen la probabilidad de obtener los valores de la variable dependiente en función de los datos proporcionados por la muestra.

Los cálculos para las estimaciones de los coeficientes de la regresión logística multinomial no son directos, por lo que es necesario usar métodos iterativos como el método de Newton-Rapson. Usando estos métodos se obtienen los coeficientes y sus errores estándar.

Dado que el modelo multinomial se puede ver como varios modelos logit clásicos, para validarlo, aparte de usar las estadísticas comunes, como pruebas sobre los coeficientes (estadístico de Wald), bondad de ajuste del modelo (estadístico chi-cuadrado), análisis de multicolinealidad y residuos se utilizan el estadístico KS y el área bajo la curva ROC (AUROC) para evaluar la calidad de discriminación.

### Multicolinealidad

Se define a la multicolinealidad como el problema de que una variable explicativa en el modelo de regresión sea una combinación lineal de las demás, es decir, que dos o más variables estén linealmente correlacionadas. Esto ocasiona un incremento exagerado en los errores estándar y en los coeficientes estimados, por tanto se estaría tratando con un problema mal especificado o mal condicionado [8].

Para estudiar el problema de multicolinealidad se utiliza el índice de condicionamiento (IC), definido por

$$IC = \sqrt{\frac{\lambda_{mx}}{\lambda_{min}}}$$

donde $\lambda_{mx}$ y $\lambda_{min}$ son los valores propios máximo y mínimo respectivamente, de la matriz de correlaciones de las variables explicativas.

Si $IC < 10$ no hay presencia de multicolinealidad; si $10 \leq IC \leq 15$ existe multico-

linealidad moderada y si $IC > 15$ hay multicolinelidad fuerte [8].

**Residuos de Devianza**

Se dispone de de una muestra de tanaño N con Q combinaciones diferentes de valores de las variables explicativas $X_1, ..., X_n$. Se denota a cada combinación de valores de las variables explicativas por $x_q = (x_{q0}, x_{q1}, ..., x_{qn})'$ con $x_{q0} = 1 \ \forall q = 1, ..., Q$. En cada una de estas combinaciones se tiene una muestra aleatoria de $d_q$ observaciones independientes de la variable de respuesta politómica $Y$, de entre las cuales se denota por $y_{j/q}$ al número de observaciones que caen en la categoría de respuesta $Y_j \ \forall j = 1, ..., k$. Así que se verifica que $\sum_{j=1}^{k} y_{j/q} = d_q$ y $\sum_{q=1}^{Q} d_q = N$. Se denota por $\widehat{m}_{j/q}$ la frecuencia esperada de respuesta $Y_j$ en la combinación $x_q$ de valores observados de las variables predictoras, estimada bajo el modelo y definida por $\widehat{m}_{j/q} = d_q \widehat{p}_{j/q}$

Se definen los residuos de la devianza o residuos estudentizados por

$$d_{j/q} = \left( 2 \left[ y_{j/q} \ln \left( \frac{y_{j/q}}{\widehat{m}_{j/q}} \right) \right] \right)^{\frac{1}{2}}$$

Con esta expresión se define el estadístico de chi-cuadrado de razón de verosimilitudes como

$$G^2 = \sum_{q=1}^{Q} \sum_{j=1}^{k} d_{j/q}^2$$

Para contrastar la significación estadística de los residuos se plantea el contraste

$$H_0 : d_{j/q} = 0$$

$$H_1 : d_{j/q} \neq 0$$

Bajo la hipótesis nula, el residuo $d_{j/q}$ tiene distribución asintóticamente normal con media 0 y varianza estimada $\widehat{\sigma}^2(d_{j/q}) < 1$. En este caso, se consideran significativos cuando el valor absoluto es mayor que 4, y se considera que la observación correspondiente es anormal. Se puede definir también los residuos de devianza ajustados o estandarizados como

$$d_{j/q}^s = \frac{d_q}{\widehat{\sigma}(d_{j/q})}$$

que tiene distribución normal estándar. Así que se rechaza la hipótesis nula con un nivel de significancia $\alpha$ cuando $|d_{j/q}^s| \geq z_{\alpha/2}$.

### KS

El estadístico KS (explicado a detalle en la sección anterior) permite medir la divergencia entre las distribuciones de las probabilidades estimadas mediante la regresión logística.

### Área bajo la curva ROC (Receiver Operating Characteristics)

El área bajo el ROC (AUROC: area under the curve ROC ) se ha convertido en un criterio de evaluación de desempeño estándar en problemas de reconocimiento de patrones de dos clases.

Utilizando el trabajo de Landgrebe y Duin [9] se extiende la medida AUC para el caso multinomial o multiclase y se la nombra VUS (volumen under the surface). Las covariables $x$ son clasificadas dentro de las categorías $K = Y_1, Y_2, ..., Y_k$ de la variable dependiente $Y$. Cada categoría tiene una distribución condicional $g(x|Y_j)$ y una probabilidad $p(Y_j)$. La asignación de las categorías se basa en la regla de Bayes, la cual asigna cada individuo su probabilidad más alta:

$$p(Y_j|x) = \frac{p(Y_j)g(x|Y_j)}{p(Y_1)g(x|Y_1) + p(Y_2)g(x|Y_2) + ... + p(Y_k)g(x|Y_k)} \tag{2.30}$$

luego para cada individuo se tomará

$$argmax_{j=1}^{k} p(Y_j|x) \tag{2.31}$$

En la práctica, desconocen las distribuciones condicionales de las categorías, éstas se estiman típicamente a partir de ejemplos representativos que se supone que se extraen aleatoriamente de la distribución verdadera, y se pueden usar en el mismo marco.

Entonces, usando las probabilidades estimadas el modelo de regresión multinomial, cada categoría tendrá una probabilidad de ocurrencia $p_j(x)$ y por las ecuaciones 2.30 y 2.31 a cada individuo le corresponderá $\text{máx}_{j=1}^{k} p_j(x)$.

Las clasificaciones se analizan a detalle por medio de la matriz de confusión (Tabla 2.2) de dimensión $k \times k$, donde los elementos de la diagonal representan las clasificaciones correctas en cada categoría y los elementos fuera de la diagonal los errores relacionados con cada categoría

El caso de dos categorías es muy conocido, con dos elementos fuera de las diagonales $r_{12}$ y $r_{21}$, popularmente conocidos como falsos negativos y falsos positivos respectivamente, y dos elementos diagonales $r_{11}$ y $r_{22}$ las verdaderas tasas positivas y verdaderas negativas. En este caso se obtiene un gráfico de las sensibilidad vs 1-espesificidad (Fi-

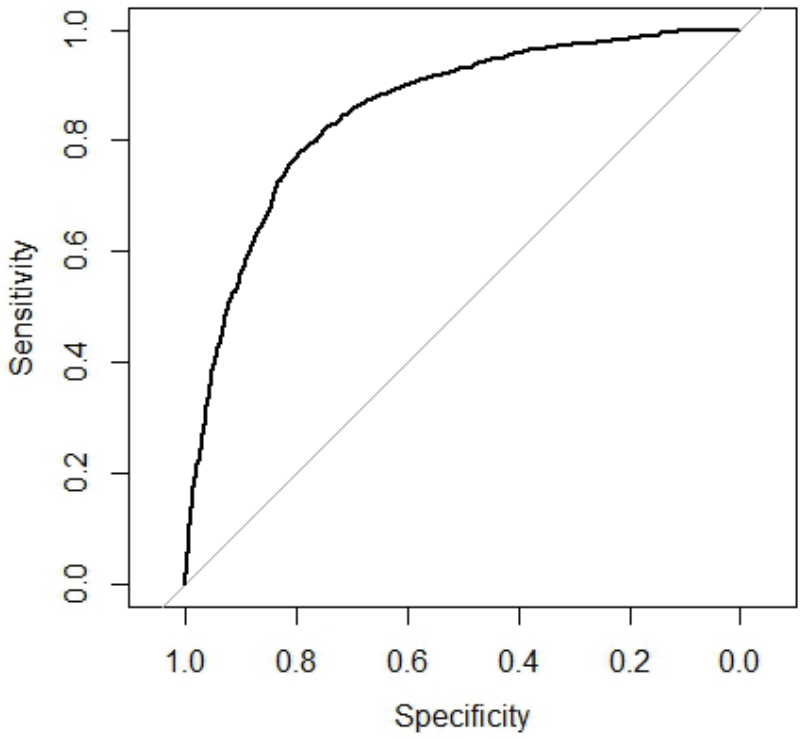


**Figura 2.1:** Curva Roc

gura 2.1), donde:

- sensibilidad: Probabilidad de ser un verdadero positivo.($r_{11}$)
- especificidad: Probabilidad de ser un verdadero negativo.($r_{22}$)

La clasificación perfecta resulta cuando el área bajo la curva es igual a 1, una clasificación pobre cuando el área es cercana a 0 y una clasificación aleatoria cuando el valor es 0,5 (puesto que es una porción del cuadrado unitario). Esta área se conoce como el área bajo el ROC (AUROC) y puede ser escrita como

$$AUC = \int r_{22} dr_{11} \tag{2.32}$$

La extensión del AUC al caso multinomial lleva al cálculo de el volumen bajo la superficie del hyperplano ROC. En este caso, se considera solamente las dimensiones ROC correspondientes a los elementos diagonales de la matriz de confusión. El VUS (Volumen Under the Surface) simplificado se puede escribir como:

$$VUS = \int ... \int \int r_{11} dr_{22} dr_{33} ... dr_{kk} \tag{2.33}$$

Esta medida permite evaluar la clasificación sobre todos los puntos responsables de las dimensiones ROC correspondientes a los elementos de la diagonal de la matriz

de confusión. Si sólo se consideran estos resultados, el VUS es similar al AUC en que si la clasificación es buena, resultará en un VUS alto y clasificaciones más pobres en un puntaje más bajo. Sin embargo, antes de que el VUS se aplique ciegamente, es importante caracterizar y comprender los límites de rendimiento entre clasificadores aleatorios y perfectos.

**Tabla 2.2:** Matriz de Confusión

| Real\Pronóstico | $Y_1$ | $Y_2$ | ... | $Y_k$ |
|---|---|---|---|---|
| $Y_1$ | $r_{11}$ | $r_{12}$ | ... | $r_{1k}$ |
| $Y_2$ | $r_{21}$ | $r_{22}$ | ... | $r_{2k}$ |
| . | . | . | ... | . |
| . | . | . | ... | . |
| . | . | . | ... | . |
| $Y_k$ | $r_{k1}$ | $r_{k2}$ | ... | $r_{kk}$ |

### Coeficiente de GINI

La medida del rendimiento del AUC está estrechamente relacionada con el coeficiente de Gini, que a veces se utiliza como medida alternativa. Esto se define más comúnmente como el doble del área entre la curva ROC y la diagonal (siendo esta área tomada como negativa en el raro caso de que la curva esté por debajo de la diagonal). La geometría elemental muestra que $Gini + 1 = 2 \times AUC$. En este trabajo trabajamos en términos de AUC, pero los resultados se aplican igualmente al coeficiente de Gini. [10]

### Tasa de clasificaciones correctas

Otra forma de validar el modelo es usando la tasa de clasificaciones correctas. Es decir, a partir del modelo ajustado, se clasifica cada observación en la categoría más probable, construyendo así una matriz de clasificación observados-predichos y se utiliza el porcentaje de clasificaciones correctas como una medida de la calidad de predicción. Se define como la proporción de individuos clasificados correctamente por el modelo y se calcula como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N. Un individuo es clasificado correctamente por el modelo cuando su valor observado de la variable respuesta Y $(Y_1, Y_2, \ldots, Y_k)$ coincide con su valor estimado por el modelo.

# Capítulo 3

# Desarrollo de un Modelo de Regresión Multinomial de mejor hora de llamada

Una de las características mas importantes de los portafolios de de crédito es la cantidad de operaciones que se generan, por tanto, el análisis de datos es muy importante para la toma de decisiones y para generar pronósticos, que por la homogeneidad de las operaciones resultan bastante confiables.

La información que se genera en cobranzas es abundante en diversos aspectos, sin embargo, lo más importante es que generalmente la información que es enviada por quien origina el portafolio (cedente) es validada previamente, ya que normalmente obedece los estándares de los organismos de control o pasan por un proceso de revisión de consistencia para realizar los balances contables.

Es por esto que toda la información relacionada con la variable producto es el insumo primario para generar la segmentación de la cartera y las estrategias. Por otro lado, se tiene un conjunto de datos que son de menor calidad, como las direcciones domiciliarias, números de teléfono, variables socio-demográficas y las que resultan de la captura de los sistemas de gestión cuando no se tienen bien definidos los árboles de gestión y las acciones a tomar en cada ramal; esta información es de menor calidad debido a que la mayoría de procesos crediticios y de cobranza son deficientes en la captura, validación e ingreso de datos en los sistemas, pues suelen ser actualizados frecuentemente sin guardar históricos y sin mantener consistencia para su validación.

El objetivo principal del modelo, es encontrar el horario del día donde la posibilidad de contactar a un cliente sea mayor, en otras palabras, se requiere separar aquellos individuos aquellos individuos que son posibles de contactar en un horario de los que no es posible contactarlos en ningún horario, por tanto, la información que resulta de la captura de los sistemas de gestión es la más importante para su desarrollo.

La metodología de desarrollo del modelo multinomial empieza con seleccionar la muestra, luego se define la variable dependiente para el modelo, se ajusta el modelo de regresión y finalmente se presentan los resultados y su validación.

## 3.1. Selección de la ventana de Muestreo

Se obtuvo información de Enero a Septiembre del 2016 del sistema principal de gestión, misma que fue minuciosamente analizada dentro del sistema de tal forma que sea consistente y válida para su uso.

Se seleccionó la población formada por todos los individuos a los que se los gestionó telefónicamente en los meses de Julio y Agosto del 2016, a los cuales se los denomina puntos de observación. Con el fin de tener suficiente información histórica se tomó en cuenta solo aquellos individuos que fueron gestionados en los últimos 6 meses antes de los puntos de observación.

**Tabla 3.1:** Distribución Población

| Pto Obs | Individuos | Porcentaje | Acumulado |
|---|---|---|---|
| **Julio 2016** | 3146 | 53 % | 53 % |
| **Agosto 2016** | 2845 | 47 % | 100 % |
| **Total** | 5991 | | |

A partir de esta información, se toman dos muestras aleatorias, una para modelamiento que consta del 60 % de los datos iniciales y otra de validación que consta del 40 % de los datos iniciales. A continuación se muestra la distribución de ambas muestras con respecto al punto de observación.

**Tabla 3.2:** Distribución Data Modelamiento

| Pto Obs | Individuos | Porcentaje | Acumulado |
|---|---|---|---|
| **Julio 2016** | 1888 | 53 % | 53 % |
| **Agosto 2016** | 1707 | 47 % | 100 % |
| **Total** | 3595 | | |

**Tabla 3.3:** Distribución Data Validación

| Pto Obs | Individuos | Porcentaje | Acumulado |
|---|---|---|---|
| **Julio 2016** | 1258 | 53 % | 53 % |
| **Agosto 2016** | 1138 | 47 % | 100 % |
| **Total** | 2396 | | |

Luego se definen las ventanas de tiempo de comportamiento y de desempeño. En la ventana de comportamiento se construyen las variables históricas que permitan pronosticar el valor de probabilidad de la variable dependiente y en la ventana de desempeño se define la variable dependiente, en otras palabras, este es el tiempo al que se hacen los pronósticos de probabilidad.

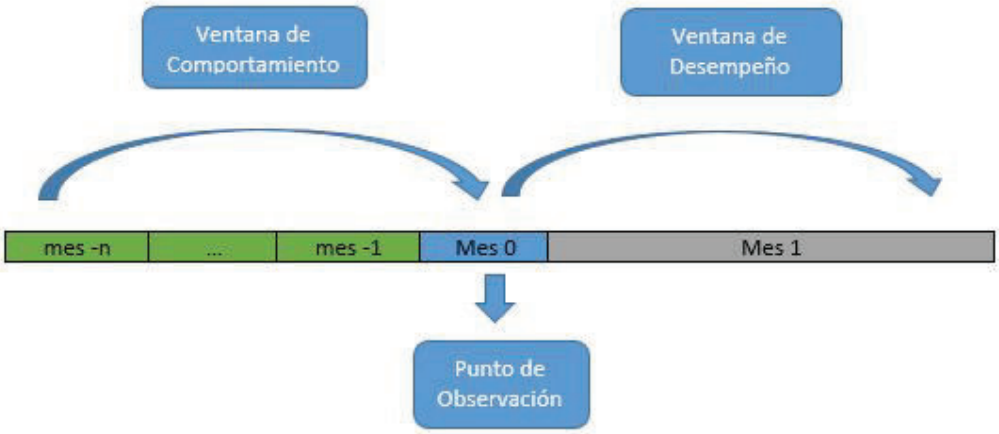


**Figura 3.1:** Ventanas de tiempo

En la figura 3.1 se definen las ventanas de tiempo de forma didáctica. Dado que la segmentación de los portafolios y la asignación de operaciones a ser gestionadas se realiza en períodos mensuales, se quiere pronosticar la probabilidad de contactar a un individuo en un horario determinado en el transcurso de un mes, por tanto la variable dependiente se define a un mes posterior al punto de observación.

## 3.2. Definición de la variable dependiente

La variable dependiente $Y$ es una variable cualitativa con más de dos categorías, cada categoría es un horario del día en el cual se puede contactar telefónicamente a un cliente. Para definir estas categorías se analizó la información de gestiones telefónicas de Enero 2016 a Septiembre 2016 tanto en conexiones efectivas y llamadas telefónicas realizadas en cada hora del día durante el mes.

Se observa un claro patrón de conexiones efectivas, donde en los horarios 7-9am y 13-16pm los porcentajes de conexión efectiva son más altos que en los horarios de 9-13pm y 16-21pm. A continuación se presenta el análisis en gráfico de barras correspondiente a los meses de Agosto 2016 y Septiembre 2016, de los meses restantes se pueden observar en el Anexo A.

AGOSTO:

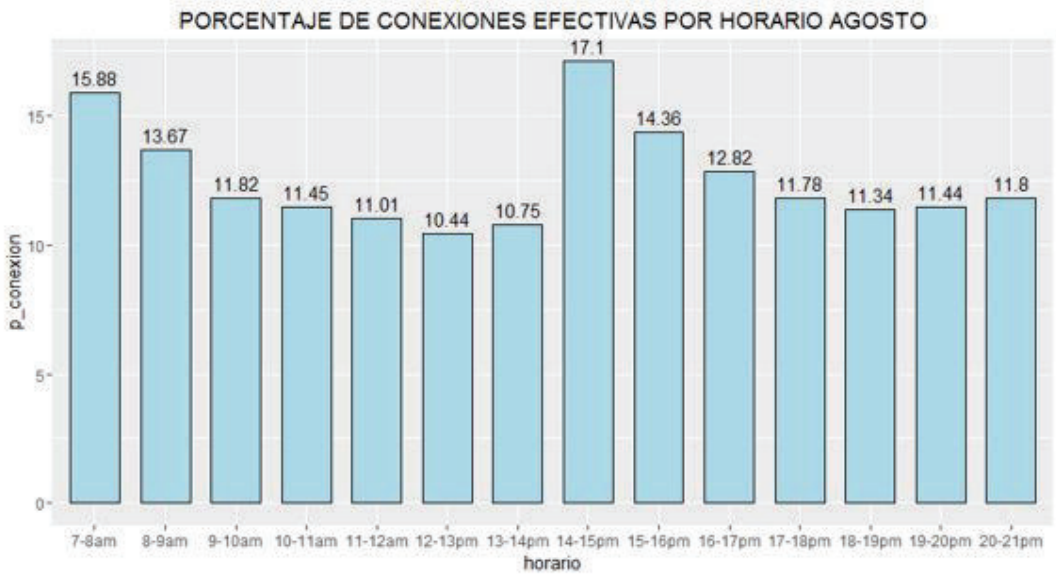


**Figura 3.2:** Conexión Agosto

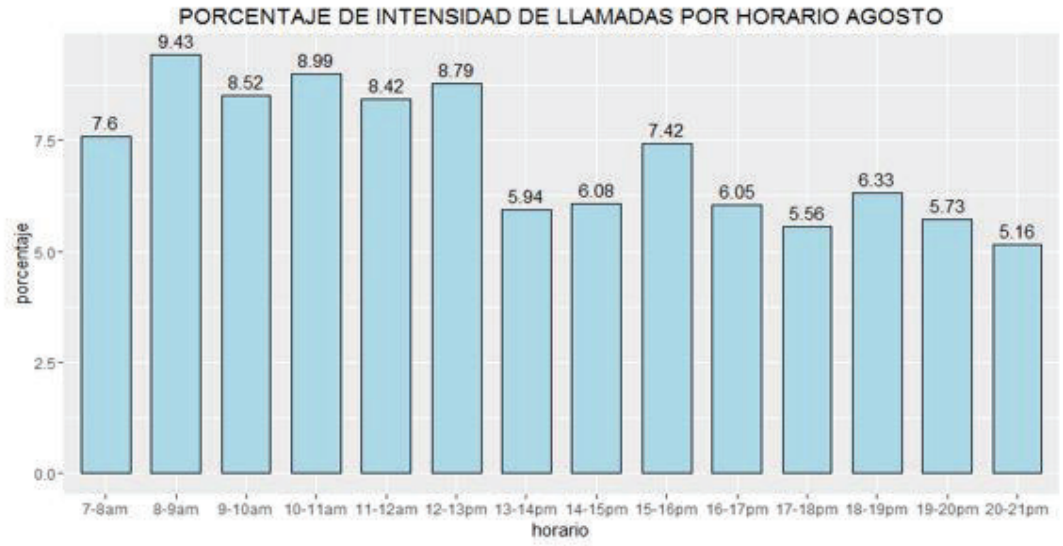


**Figura 3.3:** Llamadas Agosto

SEPTIEMBRE:

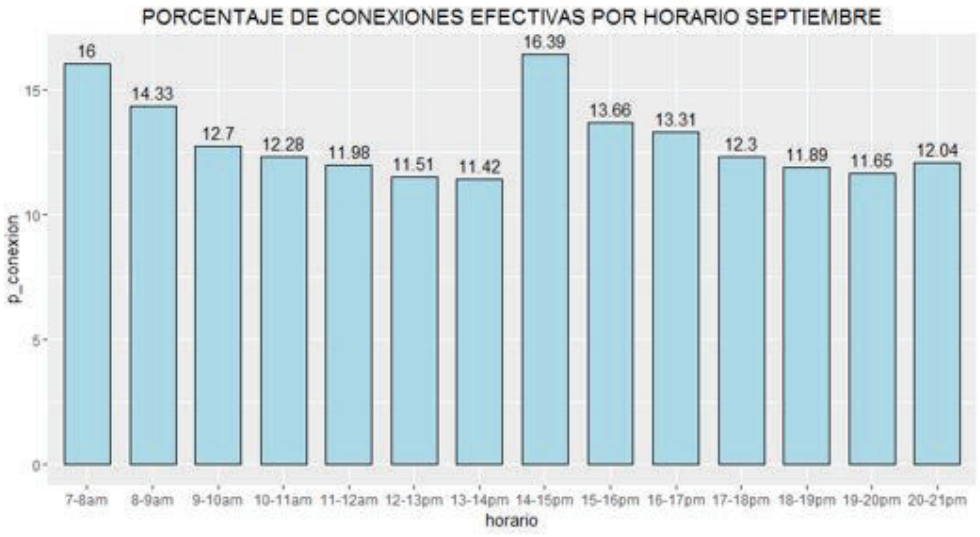


**Figura 3.4:** Conexión Septiembre

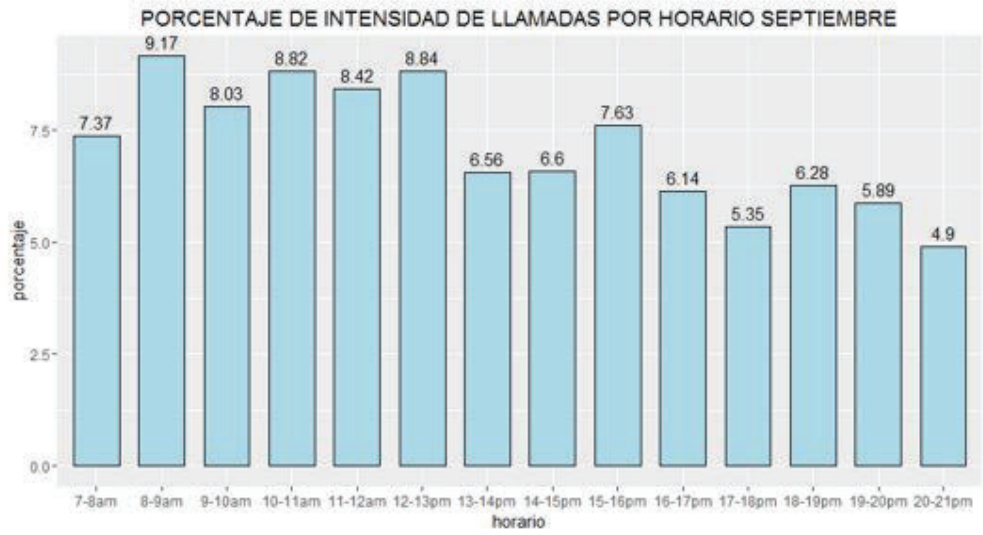


**Figura 3.5:** Llamadas Septiembre

En cuanto a llamadas telefónicas no se observa un patrón claro que se relacione con las conexiones efectivas. La intensidad de llamadas a cada hora es bastante variable, además se observa que una intensidad alta en un horario determinado no necesariamente se convierte en conexión efectiva alta.

A continuación se muestran matrices de correlación entre conexiones efectivas e intensidad de llamadas.

**Tabla 3.4:** Correlación Conexión Efectiva vs Llamadas

| Con/Llam | Ene | Feb | Mar | Abr | May | Jun | Jul | Ago | Sep |
|---|---|---|---|---|---|---|---|---|---|
| **Ene** | -0,123 | -0,005 | -0,042 | -0,073 | -0,191 | -0,130 | -0,037 | -0,037 | -0,010 |
| **Feb** | -0,056 | 0,043 | 0,010 | -0,021 | -0,129 | -0,091 | 0,011 | 0,016 | 0,046 |
| **Mar** | 0,014 | 0,123 | 0,083 | 0,047 | -0,064 | -0,026 | 0,064 | 0,069 | 0,099 |
| **Abr** | -0,126 | -0,012 | -0,059 | -0,097 | -0,202 | -0,150 | -0,079 | -0,083 | -0,042 |
| **May** | 0,014 | 0,113 | 0.069 | 0,035 | -0,071 | -0,024 | 0,049 | 0,052 | 0,091 |
| **Jun** | -0,072 | 0,050 | 0.000 | -0,047 | -0,136 | -0,096 | -0,053 | -0,055 | -0,010 |
| **Jul** | -0,145 | 0,009 | -0,038 | -0,081 | -0,195 | -0,132 | -0,098 | -0,102 | -0,076 |
| **Ago** | -0,078 | 0,086 | 0,039 | -0,002 | -0,125 | -0,066 | -0,035 | -0,042 | -0,009 |
| **Sep** | 0,036 | 0,165 | 0,122 | 0,091 | -0,027 | 0,010 | 0,094 | 0,093 | 0,111 |

En la matriz 3.4, valores positivos indican que si las llamadas aumentas las conexiones efectivas también lo harán, pero estos valores son bajos lo que corrobora el hecho descrito graficamente.

**Tabla 3.5:** Correlación Conexiones Efectivas

| Con | Ene | Feb | Mar | Abr | May | Jun | Jul | Ago | Sep |
|---|---|---|---|---|---|---|---|---|---|
| **Ene** | 1,000 | 0,964 | 0,970 | 0,981 | 0,968 | 0,943 | 0,963 | 0,965 | 0,963 |
| **Feb** | 0,964 | 1,000 | 0,979 | 0,969 | 0,961 | 0,958 | 0,907 | 0,910 | 0,915 |
| **Mar** | 0,970 | 0,979 | 1,000 | 0,980 | 0,985 | 0,974 | 0,941 | 0,949 | 0,956 |
| **Abr** | 0,981 | 0,969 | 0,980 | 1,000 | 0,986 | 0,982 | 0,959 | 0,952 | 0,939 |
| **May** | 0,968 | 0,961 | 0,985 | 0,986 | 1,000 | 0,969 | 0,938 | 0,939 | 0,951 |
| **Jun** | 0,943 | 0,958 | 0,974 | 0,982 | 0,969 | 1,000 | 0,946 | 0,939 | 0,918 |
| **Jul** | 0,963 | 0,907 | 0,941 | 0,959 | 0,938 | 0,946 | 1,000 | 0,990 | 0,957 |
| **Ago** | 0,965 | 0,910 | 0,949 | 0,952 | 0,939 | 0,939 | 0,990 | 1,000 | 0,977 |
| **Sep** | 0,963 | 0,915 | 0,956 | 0,939 | 0,951 | 0,918 | 0,957 | 0,977 | 1,000 |

La matriz 3.5 indica claramente que las conexiones efectivas en un mes están altamente correlacionadas con las conexiones de meses anteriores, es decir si la conexión efectiva en un mes es alta al mes siguiente será igual o mayor.

Dado el hecho de que el número de llamadas no está altamente correlacionado con las conexiones efectivas la variable dependiente se puede definir solo en base a las conexiones efectivas. Lógicamente, al aumentar en un gran volumen la intensidad a un mismo cliente la conexión efectiva con este será mayor, si su información para contactarlo es correcta y validada.

Con el objetivo de reducir las categorías de la variable dependiente correspondiente al horario de llamada, se agrupan horarios en base a el análisis descriptivo realizado anteriormente, tomando en cuenta resultados relativamente homogéneos en los valores, como porcentaje más altos y porcentajes bajos. Observando las gráficas mensuales, los

horarios donde existe mayor conexión son de 7am a 9am y de 13pm a 16pm mientras que 9am a 13pm y de 16pm a 21pm la conexión se reduce, por tanto se puede establecer los horarios de la siguiente manera:

- 1: 7-9 am
- 2: 9-13 pm
- 3: 13-16 pm
- 4: 16-21 pm

Una vez definidas las categorías de respuesta se procede a realizar la discriminación de los individuos que ingresarán a cada una de las categorías, en este caso a cada uno de los horarios de contacto o no contactarlo en ninguno, más específicamente ¿bajo que consideraciones se considera a un individuo contactado en cada horario?. Puesto que los análisis anteriores reflejan que un aumento de llamadas en un determinado horario no necesariamente implica un aumento de conexión, se tomará en cuenta solamente la conexión efectiva para definir la variable dependiente .

Al tener información de 9 meses de gestiones telefónicas, se tomarán los meses de Julio y Agosto como puntos de observación, y se tomará un mes hacia adelante como ventana de desempeño donde se definirá la variable dependiente, y los meses restantes serán tomados como ventana de comportamiento para la creación de las variables regresoras o explicativas independientes, como se muestra en la Figura 3.1.

Se definen las categorías de la variable dependiente $Y$ como sigue:
Sea $Y$ el contacto telefónico con las categorías de horario $Y_j = j$. Como se espera un aumento en la tasa de contactos efectivos cuando la tasa de conexión aumenta (definidas en la sección 1.1) realizando la menor cantidad de llamadas telefónicas posible, se define $pce_j$ el porcentaje de conexión efectiva en el horario $j$ como

$$pce_j = 100 * \frac{\#contactos\_efectivos_j}{\#llamadas\_totales_j}$$

Se etiqueta como contactado en el horario $j$ a todos los individuos cuyo valor máximo de porcentaje de conexión efectiva corresponde al horario $j$, si el valor máximo de porcentaje de conexión efectiva es 0 se los etiqueta como no contactado en ningún horario (NC) y la variable $Y$ toma el valor de 0 con $j = \{1, 2, 3, 4\}$.

$$Y = \begin{cases} j & \text{si } \max\limits_{j \in \{1,2,3,4\}} pce_j \neq 0 \\ 0 & \text{caso contrario} \end{cases}$$

Finalmente se presenta la distribución de la muestra de modelamiento de acuerdo a las categorías de la variable $Y$.

**Tabla 3.6:** Distribución Variable Dependiente Modelamiento

| **Y** | **Etiqueta** | **Individuos** | **Porcentaje** | **Acumulado** |
|---|---|---|---|---|
| 0 | NC | 2093 | 58 % | 58 % |
| 1 | 7-9 | 207 | 6 % | 64 % |
| 2 | 9-13 | 463 | 13 % | 77 % |
| 3 | 13-16 | 316 | 9 % | 86 % |
| 4 | 16-21 | 516 | 14 % | 100 % |
| **Total** | | 3595 | 100 % | |

En la tabla 3.6, se puede notar que las distribuciones de individuos contactados en el día y no contactados son bastante cercanas, de 42 % y 58 % respectivamente.

## 3.3. Selección de Variables Explicativas

En esta sección se seleccionará un subconjunto de variables explicativas considerando aquellas que representen una mayor divergencia entre las distribuciones de individuos en las categorías de la variable dependiente, para esto se usarán los estadísticos definidos en la sección 2.2.

La base de datos final consta de 139 variables explicativas, 5 variables categóricas y 134 variables numéricas. Variable como profesión, estado civil, nacionalidad no fueron tomadas en cuenta por escasez de población o por dudas acerca de su validación.

### 3.3.1. Filtrado de variables numéricas continuas

Para el filtrado de las variables numéricas, se utiliza el estadístico de KS definido a detalle en la sección 2.2.2. Para cada variable numérica continua $x_s$ dada se seleccionan los valores correspondientes a los individuos etiquetados con cada categoría de la variable dependiente, es decir, se crearán variables denotadas como $x_0$ a todos los individuos etiquetados como no contactados en ningún horario (NC) de la variable $x_s$, $x_1$ a todos los individuos etiquetados como contactados en el horario de 7-9am de la variable $x_s$, $x_2$ a todos los individuos etiquetados como contactados en el horario de 9-13pm de la variable $x_s$, $x_3$ a todos los individuos etiquetados como contactados en el horario de 13-16pm de la variable $x_s$ y $x_4$ a todos los individuos etiquetados como contactados en el horario de 16-21pm de la variable $x_s$, estas variables permitirán obtener una mejor partición de los del conjunto de individuos en la variable dependiente, es decir, permitirán determinar si la variable $x_s$ explica mucho mejor a la variable dependiente.

Para cada variable numérica continua se calcula el estadístico KSM.

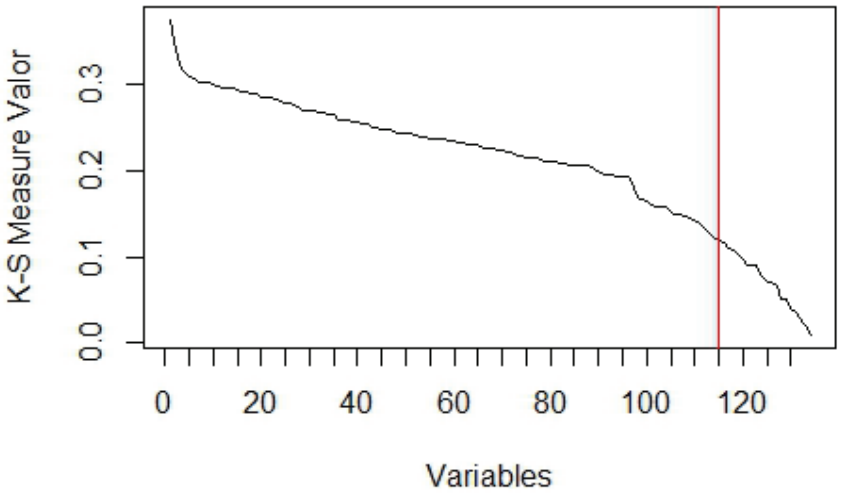


**Figura 3.6:** KSM por variable

En la figura 3.6 se puede ver que a partir de la variable 115 la medida del KSM tiende a 0, por tanto estas variables son las que se pueden descartar con seguridad.

De las variables restantes, se toman aquellas que tienen una correlación menor al 70 % ordenándolas de mayor a menor acuerdo a la medida de su KSM, así el conjunto de variables numéricas continuas candidatas para el modelo se reduce a 45.

### 3.3.2. Filtrado de variables categóricas

Para el filtrado de variables categóricas se utiliza el coeficiente de de contingencia de Pearson, definido a detalle en la sección 2.2.2, el cuál ayudará a establecer si la variable dependiente es explicada por la variable categórica mediante la prueba de independencia Ji-cuadrado, donde valores cercanos a 0 del coeficiente de contingencia indicarán independencia entre las variables y valores cercanos a 1 indicarán que existe relación entre las mismas.

En la figura 3.7 se presenta el gráfico del CCP por variable, en este caso se dispone de una cantidad baja de variables, y de acuerdo al gráfico se pueden descartar con seguridad las dos últimas variables y se utilizarán las tres restantes para construir variables dummy usando árboles de decisión.

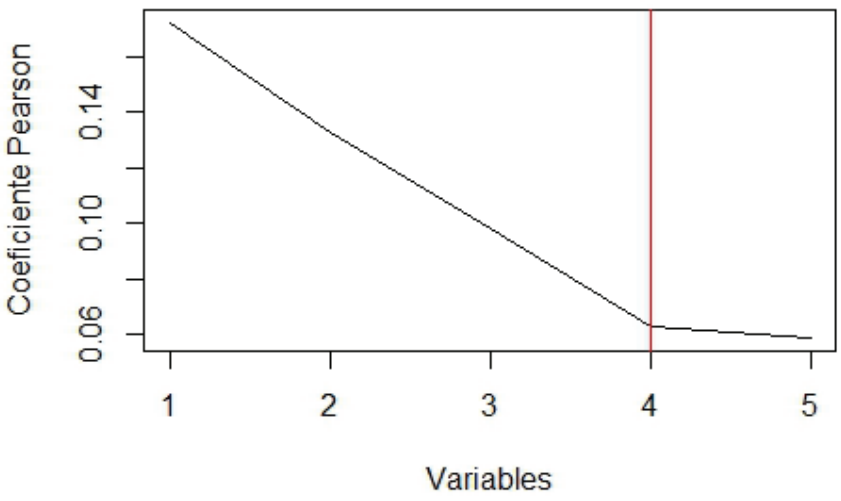


**Figura 3.7:** CCP por variable

Finalmente se obtiene una base de datos con las variables numéricas y categóricas con el mayor poder predictivo. Este filtrado de variables se realiza con el objetivo de descartar el mayor número de variables posible que no influyen en la discriminación de las categorías de la variable dependiente.

### 3.3.3. Descripción de variables explicativas

Luego de probar varios modelos de regresión con la categoría NC: no contactado en ningún horario, como categoría pivote y considerando los indicadores tales como la significancia (Estadístico de Wald), el estadístico KS y la medida KSM, el área bajo la curva ROC y el índice GINI se obtuvo un modelo constituido por 8 variables explicativas. A continuación se presentan el nombre de la variable con su respectivo coeficiente estimado, error estándar, significancia y los extremos del intervalo de confianza al 95 %, para cada categoría de horario.

**Tabla 3.7:** Variables Explicativas H1:7-9am

| 7-9am | Coeficiente | Std. Error | Significancia | Sup | Inf |
|---|---|---|---|---|---|
| **Intercepto** | -6,100 | 0,400 | 0,000 | -6,884 | -5,315 |
| **num_conex** | 0,483 | 0,045 | 0,000 | 0,394 | 0,571 |
| **p_conex** | 0,027 | 0,004 | 0,000 | 0,019 | 0,035 |
| **DIAGESTION** | 0,886 | 0,269 | 0,001 | 0,358 | 1,414 |
| **PRODUCTO** | 0,541 | 0,210 | 0,010 | 0,130 | 0,953 |
| **max_porc_conex_6M** | 0,008 | 0,003 | 0,002 | 0,003 | 0,014 |
| **num_conex_2M** | 0,096 | 0,025 | 0,000 | 0,047 | 0,146 |
| **min_num_conex_4M** | 0,419 | 0,121 | 0,001 | 0,182 | 0,657 |
| **rsaldo_inicial_2M_6M** | 0,800 | 0,270 | 0,003 | 0,272 | 1,328 |

**Tabla 3.8:** Variables Explicativas H2:9-13am

| 9-13pm | Coeficiente | Std. Error | Significancia | Sup | Inf |
|---|---|---|---|---|---|
| **Intercepto** | -3,976 | 0,261 | 0,000 | -4,488 | -3,465 |
| **num_conex** | 0,356 | 0,041 | 0,000 | 0,276 | 0,436 |
| **p_conex** | 0,026 | 0,003 | 0,000 | 0,019 | 0,032 |
| **DIAGESTION** | 0,185 | 0,141 | 0,189 | -0,091 | 0,461 |
| **PRODUCTO** | 0,548 | 0,148 | 0,000 | 0,257 | 0,839 |
| **max_porc_conex_6M** | 0,007 | 0,002 | 0,001 | 0,003 | 0,010 |
| **num_conex_2M** | 0,079 | 0,020 | 0,000 | 0,039 | 0,120 |
| **min_num_conex_4M** | 0,372 | 0,103 | 0,000 | 0,169 | 0,574 |
| **rsaldo_inicial_2M_6M** | 0,673 | 0,228 | 0,003 | 0,225 | 1,120 |

**Tabla 3.9:** Variables Explicativas H3:13-16pm

| 13-16pm | Coeficiente | Std. Error | Significancia | Sup | Inf |
|---|---|---|---|---|---|
| **Intercepto** | -5,126 | 0,343 | 0,000 | -5,798 | -4,453 |
| **num_conex** | 0,430 | 0,042 | 0,000 | 0,347 | 0,513 |
| **p_conex** | 0,028 | 0,004 | 0,000 | 0,021 | 0,035 |
| **DIAGESTION** | 0,732 | 0,194 | 0,000 | 0,352 | 1,113 |
| **PRODUCTO** | 0,820 | 0,182 | 0,000 | 0,462 | 1,177 |
| **max_porc_conex_6M** | 0,011 | 0,002 | 0,000 | 0,007 | 0,015 |
| **num_conex_2M** | 0,097 | 0,022 | 0,000 | 0,053 | 0,140 |
| **min_num_conex_4M** | 0,145 | 0,114 | 0,205 | -0,079 | 0,369 |
| **rsaldo_inicial_2M_6M** | 0,337 | 0,289 | 0,243 | -0,229 | 0,904 |

**Tabla 3.10:** Variables Explicativas H4:16-21pm

| 16-21pm | Coeficiente | Std. Error | Significancia | Sup | Inf |
|---|---|---|---|---|---|
| **Intercepto** | -4,131 | 0,266 | 0,000 | -4,653 | -3,608 |
| **num_conex** | 0,370 | 0,040 | 0,000 | 0,292 | 0,449 |
| **p_conex** | 0,026 | 0,003 | 0,000 | 0,020 | 0,033 |
| **DIAGESTION** | 0,265 | 0,139 | 0,056 | -0,007 | 0,538 |
| **PRODUCTO** | 0,874 | 0,156 | 0,000 | 0,568 | 1,180 |
| **max_porc_conex_6M** | 0,004 | 0,002 | 0,036 | 0,000 | 0,008 |
| **num_conex_2M** | 0,091 | 0,020 | 0,000 | 0,053 | 0,130 |
| **min_num_conex_4M** | 0,400 | 0,101 | 0,000 | 0,203 | 0,597 |
| **rsaldo_inicial_2M_6M** | 0,584 | 0,230 | 0,011 | 0,134 | 1,035 |

A continuación se describen a detalle las variables del modelo.

1. **num_conex:** Variable numérica. Número de conexiones efectivas al punto de observación.

2. **p_conex**: Variable numérica. Porcentaje de conexión efectiva al punto de observación.
$$p\_conex = \frac{conexiones\_efectivas}{total\_llamadas}$$

3. **PRODUCTO:** Variable dummy creada con árboles de decisión agrupada en TCF, OLLA DE ORO, MICROCREDITO.

$$PRODUCTO = \begin{cases} 1 & \text{si PRODUCTO=TCF} \\ 0 & \text{caso contrario} \end{cases}$$

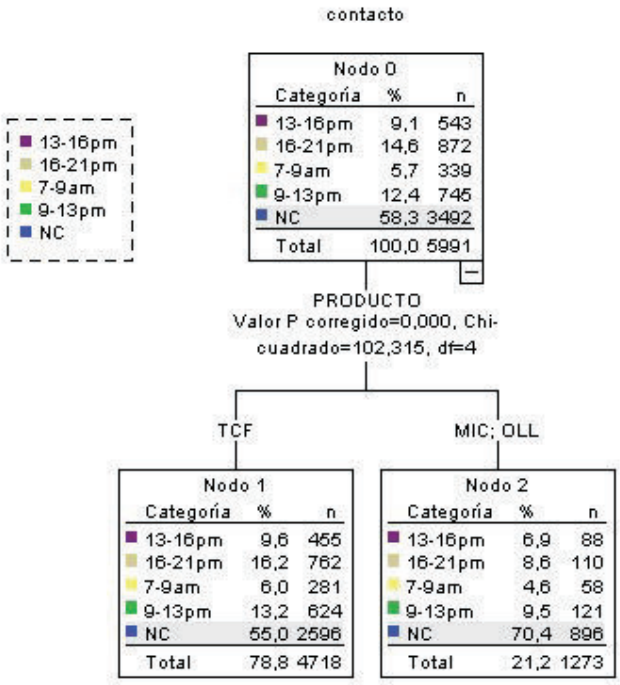


**Figura 3.8:** Variable PRODUCTO

Se asigna 1 al producto TCF, ya que según la clasificación que realiza el árbol de decisión, existe mayor posibilidad de contactar telefónicamente en este producto que en los otros dos.

4. **DIAGESTION:** Variable dummy creada con árboles de decisión, que contiene los días donde hubo mayor conexión efectiva al punto de observación, agrupada en

   - IS (Inicio de Semana): Lunes y Martes
   - MS (Mitad de Semana): Miércoles y Jueves
   - FD (Fin de Semana): Viernes y Sábado

$$DIAGESTION = \begin{cases} 0 & \text{si DIAGESTION=IS} \\ 1 & \text{caso contrario} \end{cases}$$

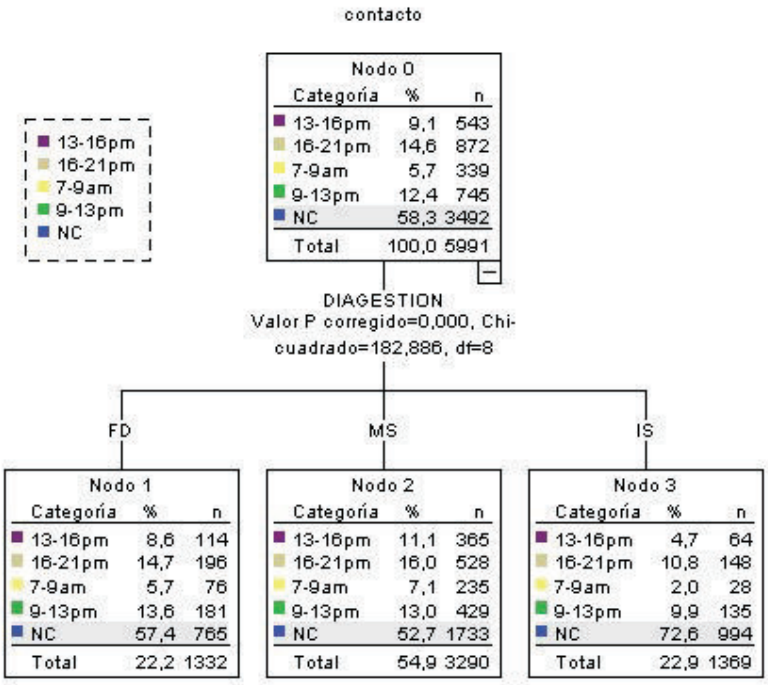


**Figura 3.9:** Variable DIAGESTION

Se asigna 0 al DIAGESTION IS, ya que según la clasificación del árbol de decisión, existe menor probabilidad de contacto efectivo en este grupo de días.

5. **max_porc_conex_6M:** Variable numérica. Máximo de los porcentajes de conexión efectiva en los últimos seis meses.

6. **num_conex_2M:** Variable numérica. Número de conexiones efectivas en los últimos dos meses.

7. **rsaldo_inicial_2M_6M:** Variable numérica. Razón entre el saldo inicial dos meses atrás antes del punto de observación y el saldo inicial seis meses atrás antes del punto de observación.

$$rsaldo\_inicial\_2M\_6M = \begin{cases} \frac{saldo\_ini\_2M}{saldo\_ini\_6M} & \text{si saldo\_ini\_6M}>0 \\ ln\_saldo\_ini\_2M & \text{caso contrario} \end{cases}$$

La transformación logarítmica se usa para obtener un modelo estable en el tiempo, puesto que si los valores de saldo inicial son altos, el impacto sobre la probabilidad

de contacto efectivo sería muy elevado, para que este efecto sea minimizado se acostumbra a usar esta transformación.

8. **min_num_conex_4M:** Variable numérica. Mínimo de conexiones efectivas en los últimos cuatro meses.

## 3.4. Interpretación de Parámetros

La interpretación de los coeficientes estimados se realiza en función de los coeficientes de ventajas, conocidos también como ODDS RATIO. Cuando se interpreta las odds ratios de cada variable, se asume que el resto de variables independientes se mantienen fijas. Se interpreta cada una de las variables independientes entre los distintos horarios de contacto tomando como referencia NC: no contactado en ningún horario. A continuación se presentan los coeficientes con estimados junto con los ODDS RATIO para cada categoría de horario.

**Tabla 3.11:** Odds Ratio H1: 7-9am

| 7-9am | Coeficiente | Odss |
|---|---|---|
| **Intercepto** | -6,100 | 0,002 |
| **num_conex** | 0,483 | 1,620 |
| **p_conex** | 0,027 | 1,027 |
| **DIAGESTION** | 0,886 | 2,425 |
| **PRODUCTO** | 0,541 | 1,718 |
| **max_porc_conex_6M** | 0,008 | 1,008 |
| **num_conex_2M** | 0,096 | 1,101 |
| **min_num_conex_4M** | 0,419 | 1,521 |
| **rsaldo_inicial_2M_6M** | 0,800 | 2,226 |

La tabla 3.11 se interpreta de la siguiente manera:

- La ventaja de contactar en el horario de 7-9am frente a no contactar en ningún horario es de 1,620 veces a medida que el número de conexiones efectivas aumenta en una unidad.
- La ventaja de contactar en el horario de 7-9am frente a no contactar en ningún horario es de 1,027 veces a medida que el porcentaje de conexión efectiva aumenta en una unidad.
- La ventaja de contactar en el horario de 7-9am entre semana(Miércoles o Jueves) o en fin de semana(Viernes o Sábado) frente a no contactar en ningún horario en inicio de semana (Lunes o Martes) es de 2,425 veces.

- La ventaja de contactar en el horario de 7-9am en el producto TCF frente a no contactar en ningún horario en otro producto (OLLA DE ORO o MICROCREDITO) es de 1,718 veces.
- La ventaja de contactar en el horario de 7-9am frente a no contactar en ningún horario es de 1,008 veces a medida que el máximo de los porcentajes de conexión efectiva en los últimos seis meses aumenta en una unidad.
- La ventaja de contactar en el horario de 7-9am frente a no contactar en ningún horario es de 1,101 veces a medida que el número de conexiones efectivas en los últimos dos meses aumenta en una unidad.
- La ventaja de contactar en el horario de 7-9am frente a no contactar en ningún horario es de 1,521 veces a medida que el mínimo de conexiones efectivas en los últimos cuatro meses aumenta en una unidad.
- La ventaja de contactar en el horario de 7-9am frente a no contactar en ningún horario es de 2,226 veces a medida que la razón entre el saldo inicial hace dos meses y el saldo inicial hace seis meses aumenta en una unidad.

**Tabla 3.12:** Odds Ratio H2: 9-13pm

| **9-13pm** | **Coeficiente** | **Odss** |
|---|---|---|
| **Intercepto** | -3,976 | 0,019 |
| **num_conex** | 0,356 | 1,428 |
| **p_conex** | 0,026 | 1,026 |
| **DIAGESTION** | 0,185 | 1,203 |
| **PRODUCTO** | 0,548 | 1,729 |
| **max_porc_conex_6M** | 0,007 | 1,007 |
| **num_conex_2M** | 0,079 | 1,083 |
| **min_num_conex_4M** | 0,372 | 1,450 |
| **rsaldo_inicial_2M_6M** | 0,673 | 1,959 |

La tabla 3.12 se interpreta de la siguiente manera:

- La ventaja de contactar en el horario de 9-13pm frente a no contactar en ningún horario es de 1,428 veces a medida que el número de conexiones efectivas aumenta en una unidad.
- La ventaja de contactar en el horario de 9-13pm frente a no contactar en ningún horario es de 1,026 veces a medida que el porcentaje de conexión efectiva aumenta en una unidad.

- La ventaja de contactar en el horario de 9-13pm entre semana(Miércoles o Jueves) o en fin de semana(Viernes o Sábado) frente a no contactar en ningún horario en inicio de semana (Lunes o Martes) es de 1,203 veces, sin embargo, esta variable no es significativa en esta categoría (tabla 3.8: Significancia 0,189), lo que indica que el contactar en el horario de 9-13pm frente a no contactar en ningún horario, no necesariamente está relacionado con que no haya conexión efectiva en inicio de semana.

- La ventaja de contactar en el horario de 9-13pm en el producto TCF frente a no contactar en ningún horario en otro producto (OLLA DE ORO o MICROCREDITO) es de 1,729 veces.

- La ventaja de contactar en el horario de 9-13pm frente a no contactar en ningún horario es de 1,007 veces a medida que el máximo de los porcentajes de conexión efectiva en los últimos seis meses aumenta en una unidad.

- La ventaja de contactar en el horario de 9-13pm frente a no contactar en ningún horario es de 1,083 veces a medida que el número de conexiones efectivas en los últimos dos meses aumenta en una unidad.

- La ventaja de contactar en el horario de 9-13pm frente a no contactar en ningún horario es de 1,450 veces a medida que el mínimo de conexiones efectivas en los últimos cuatro meses aumenta en una unidad.

- La ventaja de contactar en el horario de 9-13pm frente a no contactar en ningún horario es de 1,959 veces a medida que la razón entre el saldo inicial hace dos meses y el saldo inicial hace seis meses aumenta en una unidad.

**Tabla 3.13:** Odds Ratio H3: 13-16pm

| 13-16pm | Coeficiente | Odss |
|---|---|---|
| **Intercepto** | -5,126 | 0,006 |
| **num_conex** | 0,430 | 1,538 |
| **p_conex** | 0,028 | 1,028 |
| **DIAGESTION** | 0,732 | 2,080 |
| **PRODUCTO** | 0,820 | 2,269 |
| **max_porc_conex_6M** | 0,011 | 1,011 |
| **num_conex_2M** | 0,097 | 1,102 |
| **min_num_conex_4M** | 0,145 | 1,156 |
| **rsaldo_inicial_2M_6M** | 0,337 | 1,401 |

La tabla 3.13 se interpreta de la siguiente manera:

- La ventaja de contactar en el horario de 13-16pm frente a no contactar en ningún horario es de 1,538 veces a medida que el número de conexiones efectivas aumenta en una unidad.
- La ventaja de contactar en el horario de 13-16pm frente a no contactar en ningún horario es de 1,028 veces a medida que el porcentaje de conexión efectiva aumenta en una unidad.
- La ventaja de contactar en el horario de 13-16pm entre semana(Miércoles o Jueves) o en fin de semana(Viernes o Sábado) frente a no contactar en ningún horario en inicio de semana (Lunes o Martes) es de 2,080 veces.
- La ventaja de contactar en el horario de 13-16pm en el producto TCF frente a no contactar en ningún horario en otro producto (OLLA DE ORO o MICROCREDITO) es de 2,269 veces.
- La ventaja de contactar en el horario de 13-16pm frente a no contactar en ningún horario es de 1,011 veces a medida que el máximo de los porcentajes de conexión efectiva en los últimos 6 meses aumenta en una unidad.
- La ventaja de contactar en el horario de 13-16pm frente a no contactar en ningún horario es de 1,102 veces a medida que el número de conexiones efectivas en los últimos dos meses aumenta en una unidad.
- La ventaja de contactar en el horario de 13-16pm frente a no contactar en ningún horario es de 1,156 veces a medida que el mínimo de conexiones efectivas en los últimos 4 meses aumenta en una unidad, sin embargo, esta variable no es significativa en esta categoría (tabla 3.9: Significancia 0,205), lo que indica que el contactar en el horario de 13-16pm frente a no contactar en ningún horario, no necesariamente está relacionado con el mínimo de conexiones efectivas en los últimos cuatro meses.
- La ventaja de contactar en el horario de 13-16pm frente a no contactar en ningún horario es de 1,401 veces a medida que la razón entre el saldo inicial hace dos meses y el saldo inicial hace seis meses aumenta en una unidad, pero esta variable tampoco es significativa en esta categoría (tabla 3.9: Significancia 0,243), lo que indica que el contactar en el horario de 13-16pm frente a no contactar en ningún horario, no necesariamente está relacionado con la razón entre el saldo inicial hace dos meses y el saldo inicial hace seis meses.

**Tabla 3.14:** Odds Ratio H4: 16-21pm

| 16-21pm | Coeficiente | Odss |
|---|---|---|
| **Intercepto** | -4,131 | 0,016 |
| **num_conex** | 0,370 | 1,448 |
| **p_conex** | 0,026 | 1,027 |
| **DIAGESTION** | 0,265 | 1,304 |
| **PRODUCTO** | 0,874 | 2,396 |
| **max_porc_conex_6M** | 0,004 | 1,004 |
| **num_conex_2M** | 0,091 | 1,096 |
| **min_num_conex_4M** | 0,400 | 1,491 |
| **rsaldo_inicial_2M_6M** | 0,584 | 1,794 |

La tabla 3.14 se interpreta de la siguiente manera:

- La ventaja de contactar en el horario de 16.21pm frente a no contactar en ningún horario es de 1,448 veces a medida que el número de conexiones efectivas aumenta en una unidad.
- La ventaja de contactar en el horario de 16-21pm frente a no contactar en ningún horario es de 1,027 veces a medida que el porcentaje de conexión efectiva aumenta en una unidad.
- La ventaja de contactar en el horario de 16-21pm entre semana(Miércoles o Jueves) o en fin de semana(Viernes o Sábado) frente a no contactar en ningún horario en inicio de semana (Lunes o Martes) es de 1,304 veces.
- La ventaja de contactar en el horario de 16-21pm en el producto TCF frente a no contactar en ningún horario en otro producto (OLLA DE ORO o MICROCREDITO) es de 2,396 veces.
- La ventaja de contactar en el horario de 16-21pm frente a no contactar en ningún horario es de 1,004 veces a medida que el máximo de los porcentajes de conexión efectiva en los últimos seis meses aumenta en una unidad.
- La ventaja de contactar en el horario de 16-21pm frente a no contactar en ningún horario es de 1,096 veces a medida que el número de conexiones efectivas en los últimos dos meses aumenta en una unidad.
- La ventaja de contactar en el horario de 16-21pm frente a no contactar en ningún horario es de 1,491 veces a medida que el mínimo de conexiones efectivas en los últimos cuatro meses aumenta en una unidad.

- La ventaja de contactar en el horario de 16-21pm frente a no contactar en ningún horario es de 1,794 veces a medida que la razón entre el saldo inicial hace dos meses y el saldo inicial hace seis meses aumenta en una unidad.

Las variables obtenidas en el modelo final incluye las que se hubieran incluido en un análisis intuitivo, como el porcentaje de conexión efectiva y el número de conexiones efectivas ya que explicarían directamente si un individuo es contactable o no. Las variables máximo de los porcentajes de conexión efectiva en los últimos seis meses, el número de conexiones efectivas en los últimos dos meses y el mínimo número de conexiones efectivas en los últimos cuatro meses son las que mejor explican el comportamiento de cada individuo en conexiones efectivas en la ventana de tiempo definida, lo que de una manera intuitiva no hubiera sido posible detectar.

La variable razón entre el saldo inicial en los últimos dos meses sobre el saldo inicial en los últimos seis meses es una variable que explica la evolución de la deuda en la ventana de tiempo y explica que un individuo sea contactado o no, pues hay que recordar que la segmentación de la cartera para asignar a gestión se hace en función del saldo inicial.

En el tipo de producto, es mas fácil contactar a individuos en TCF pues es el producto que más volumen de individuos tiene con respecto a Olla de Oro y Microcrédito donde las gestiones telefónicas son con menor intensidad. El día al que se hace una llamada es bastante influyente en la hora de contacto, es por esto que esta variable es idónea para el modelar el contacto efectivo.

De manera general, se observa que los signos de los parámetros estimados por el modelo de cada categoría tienen sentido, lo que es una buena señal de que el modelo de estimaciones de probabilidad correctas, además, la interpretación de los parámetros van acorde con lo que se esperaría realmente.

## 3.5. Resultados y Validación del Modelo de BTTC

En esta sección se presentan los resultados obtenidos a través de los cuales se analizan la calidad de discriminación y predicción del modelo de regresión logística multinomial.

### 3.5.1. Multicolinelidad

Para medir el grado de multicolinealidad se utiliza el índice IC definido a detalle en la seccion 2.3, en este caso se tiene

$$IC = \sqrt{\frac{\lambda_{mx}}{\lambda_{min}}} = \sqrt{\frac{2{,}588709}{0{,}3585685}} = 2{,}686925$$

por tanto se puede concluir que el modelo no presenta problemas de multicolinealidad.

### 3.5.2. Residuos de Devianza

Para validar el modelo se utilizan los residuos de la devianza, considerando que los residuos que indican una falta de ajuste global son aquellos cuyo valor absoluto es mayor que 4, y se considera que la observacipon correspondiente es anormal.

**Tabla 3.15:** Estadísticas Residuos

| **Estadisticas/Horario** | **7-9am** | **9-13pm** | **13-16pm** | **16-21pm** | **NC** |
|---|---|---|---|---|---|
| **Min** | -0,416 | -0,351 | -0,373 | -0,428 | -0,907 |
| **1st Quartil** | -0,072 | -0,167 | -0,120 | -0,178 | -0,259 |
| **Median** | -0,026 | -0,081 | -0,049 | -0,084 | 0,118 |
| **Mean** | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| **3rd Quartil** | -0,013 | -0,052 | -0,025 | -0,049 | 0,223 |
| **Max** | 0,993 | 0,965 | 0,982 | 0,969 | 0,999 |

Como se puede observar en la tabla 3.15, entre los máximos y mínimos de los valores, todos los residuos en valor absoluto son menores de 1, por lo que no hay ninguna observación que se considere anormal.

### 3.5.3. Medidas de calidad de discriminación

#### KS y KSM

En esta etapa se analiza la máxima diferencia entre las distribuciones empíricas de la probabilidad de contacto efectivo estimada en cada categoría de horario, dada por

el estadístico KS y la medida KSM definidos detalladamente el la sección 2.2.1.

Para asegurarse de que no existen problemas de sobre ajuste del modelo a los datos de modelamiento, es decir, que el modelo presente buena discriminación únicamente para la información con la que fue desarrollado, es necesario evaluarlo con una base distinta a la modelamiento, en este caso se evalúa el modelo usando la muestra de validación descrita en la sección 3.1.

- **Contacto efectivo en el horario de 7-9am con respecto a no contactar en ningún horario.**
  Para este horario el modelo final resulta con un KS de 0.6361 y con la muestra de validación se obtiene un KS de 0.6268. En la figura 3.10 se observan la divergencia entre las distribuciones empíricas de probabilidad de contacto en el horario de 7-9am y la de probabilidad de no contactar en ningún horario, para la muestra de modelamiento (Izquierda) y para la muestra de validación (Derecha) respectivamente.

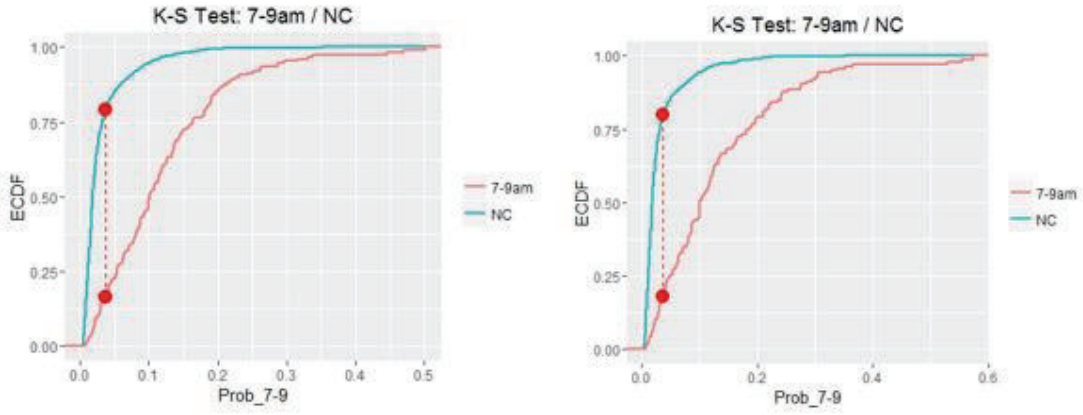


**Figura 3.10:** KS Modelamiento (Izq.) y Validación (Der.) 7-9am

- **Contacto efectivo en el horario de 9-13pm con respecto a no contactar en ningún horario.**
  Para este horario el modelo final resulta con un KS de 0,5589 y con la muestra de validación se obtiene un KS de 0,5989. En la figura 3.11 se observan la divergencia entre las distribuciones empíricas de probabilidad de contacto en el horario de 9-13pm y la de probabilidad de no contactar en ningún horario, para la muestra de modelamiento (Izquierda)y para la muestra de validación (Derecha) respectivamente.

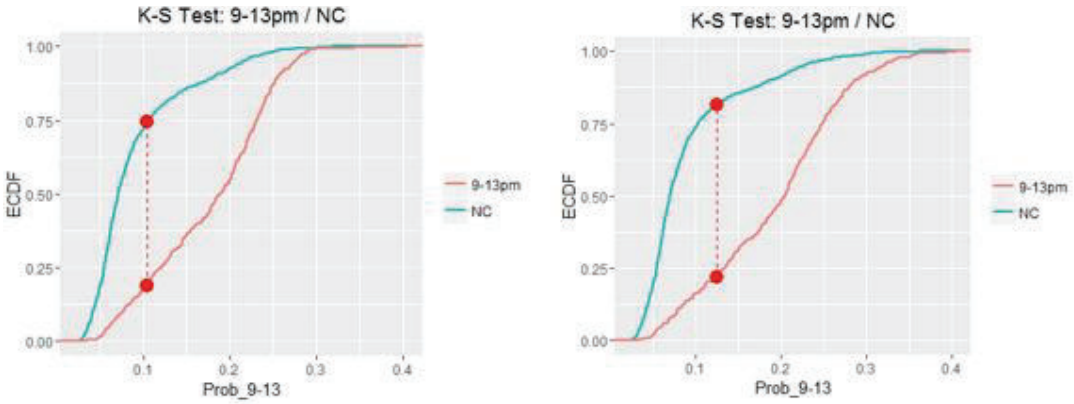


**Figura 3.11:** KS Modelamiento (Izq.) y Validación (Der.) 9-13pm

- **Contacto efectivo en el horario de 13-16pm con respecto a no contactar en ningún horario.**
  Para este horario el modelo final resulta con un KS de 0,5960 y con la muestra de validación se obtiene un KS de 0,5929. En la figura **??** se observan la divergencia entre las distribuciones empíricas de probabilidad de contacto en el horario de 13-16pm y la de probabilidad de no contactar en ningún horario, para la muestra de modelamiento (Izquierda) y para la muestra de validación (Derecha) respectivamente.

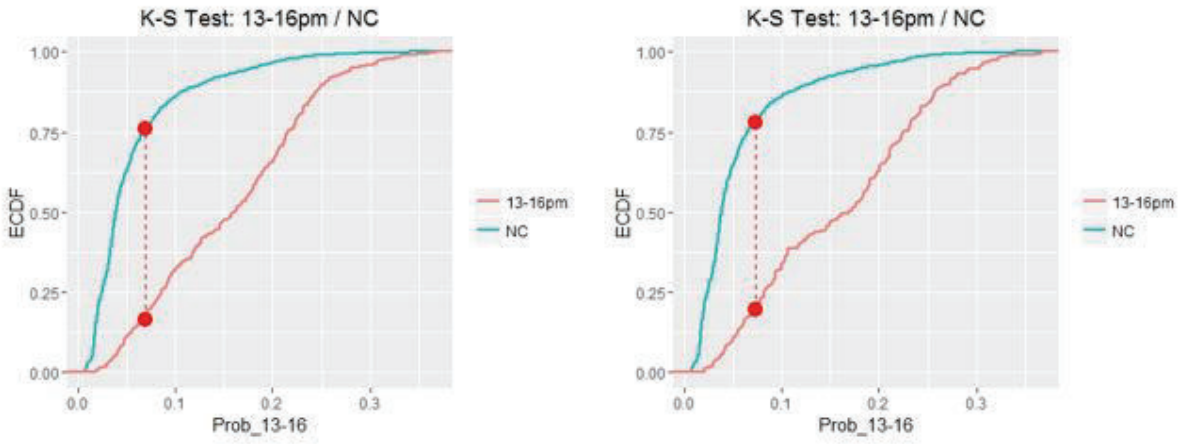


**Figura 3.12:** KS Modelamiento (Izq.) y Validación (Der.) 13-16pm

- **Contacto efectivo en el horario de 16-21pm con respecto a no contactar en ningún horario.**
  Para este horario el modelo final resulta con un KS de 0,5475 y con la muestra de validación se obtiene un KS de 0,5637. En la figura 3.13 se observan la divergencia entre las distribuciones empíricas de probabilidad de contacto en el horario de 16-21pm y la de probabilidad de no contactar en ningún horario, para la muestra de modelamiento y para la muestra de validación respectivamente.

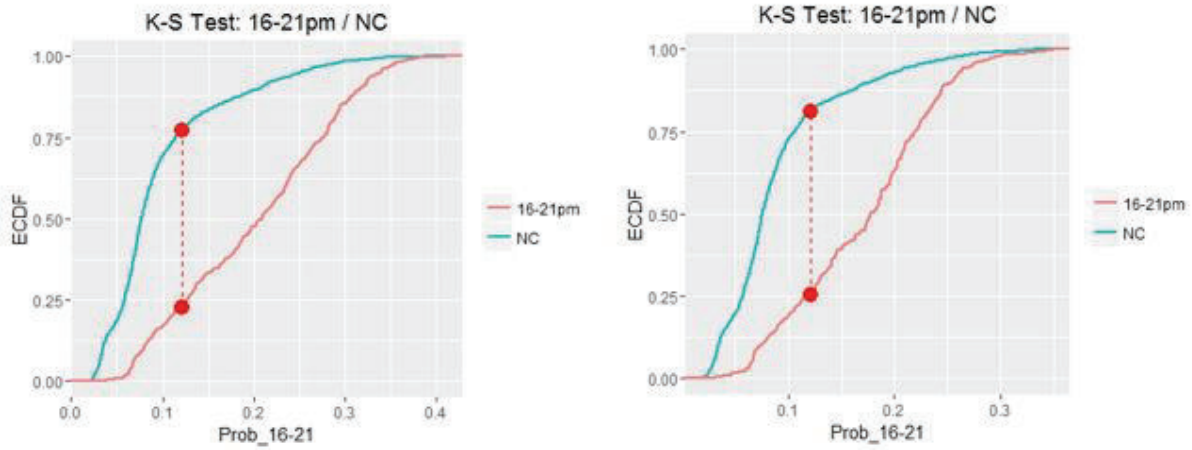


**Figura 3.13:** KS Modelamiento (Izq.) y Validación (Der.) 16-21pm

En cada categoría de horario los valores de KS son similares y superan el 50 % lo cual indica una alta divergencia entre las distribuciones y permite decir que el modelo no está sobreajustado.

Ahora, la medida KSM nos da una medida global de la calidad de discriminación del modelo, usando la definición de esta medida detallada en la sección 2.2.1 se obtiene un valor de 0,3988 con la muestra de modelamiento y un valor de 0,4056 con la muestra de validación.

### Área bajo la curva ROC

- **Contacto efectivo en el horario de 7-9am con respecto a no contactar en ningún horario.**
  Como resultado con la muestra de modelamiento se obtiene un índice AUROC de 0,882, lo cual quiere decir que existe una probabilidad de 0,882 de que la probabilidad estimada de contactar a un individuo en el horario de 7-9am sea mayor que la probabilidad estimada de un individuo de no contactarlo en ningún horario, elegidos aleatoriamente. Este valor permite decir que el modelo presenta un alto rendimiento de clasificación en esta categoría. A continuación se muestra la representación gráfica de de la curva ROC, con la probabilidad de contacto efectivo en el horario de 7-9am para la muestra de modelamiento y para la muestra de validación respectivamente.

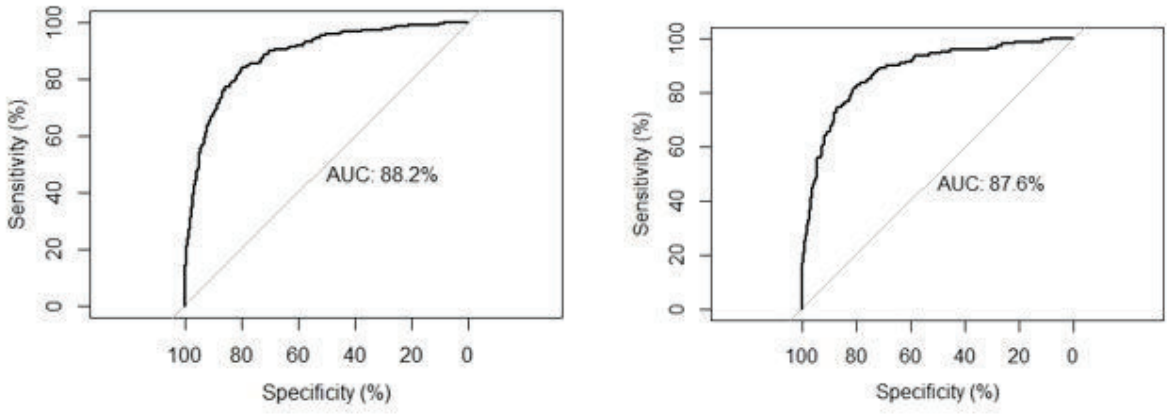


**Figura 3.14:** ROC Modelamiento (Izq.) y Validación (Der.) 7-9am

Con la muestra de validación se obtuvo un valor de 0,876, el cual es bastante cercano al obtenido con la muestra de modelamiento, como es de esperarse al no existir evidencias de sobreajuste con el estadístico KS.

- **Contacto efectivo en el horario de 9-13pm con respecto a no contactar en ningún horario.**

  Como resultado con la muestra de modelamiento se obtiene un índice AUROC de 0,836, lo cual quiere decir que existe una probabilidad de 0,836 de que la probabilidad estimada de contactar a un individuo de en el horario de 9-13pm sea mayor que la probabilidad estimada de un individuo de no contactarlo en ningún horario, elegidos aleatoriamente. Este valor permite decir que el modelo presenta un alto rendimiento de clasificación en esta categoría. A continuación se muestra la representación gráfica de de la curva ROC, con la probabilidad de contacto efectivo en el horario de 9-13pm para la muestra de modelamiento y para la muestra de validación respectivamente.

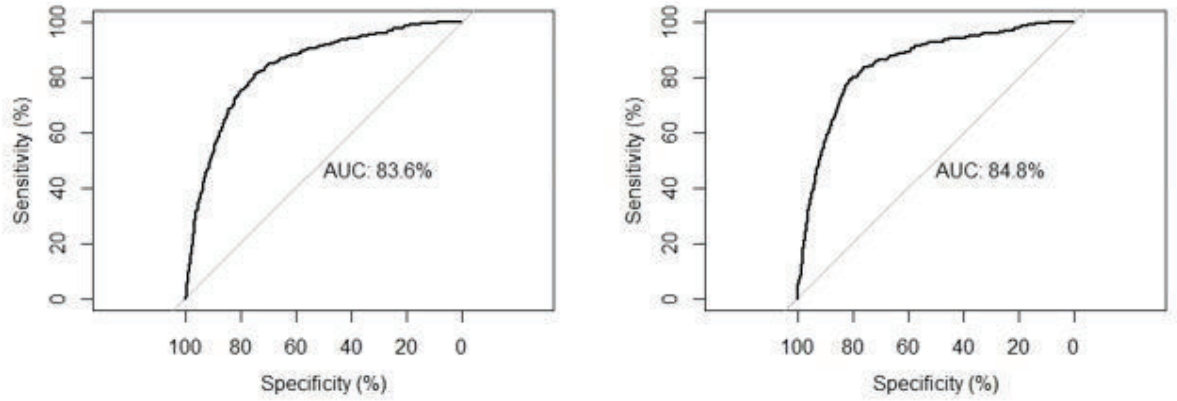


**Figura 3.15:** ROC Modelamiento (Izq.) y Validación (Der.) 9-13pm

Con la muestra de validación se obtuvo un valor de 0,848, el cual es bastante cercano al obtenido con la muestra de modelamiento, como es de esperarse al no existir evidencias de sobreajuste con el estadístico KS.

- **Contacto efectivo en el horario de 13-16pm con respecto a no contactar en ningún horario.**

  Como resultado con la muestra de modelamiento se obtiene un índice AUROC de 0,863, lo cual quiere decir que existe una probabilidad de 0,863 de que la probabilidad estimada de contactar a un individuo de en el horario de 13-16pm sea mayor que la probabilidad estimada de un individuo de no contactarlo en ningún horario, elegidos aleatoriamente. Este valor permite decir que el modelo presenta un alto rendimiento de clasificación en esta categoría. A continuación se muestra la representación gráfica de de la curva ROC, con la probabilidad de contacto efectivo en el horario de 13-16pm para la muestra de modelamiento y para la muestra de validación respectivamente.

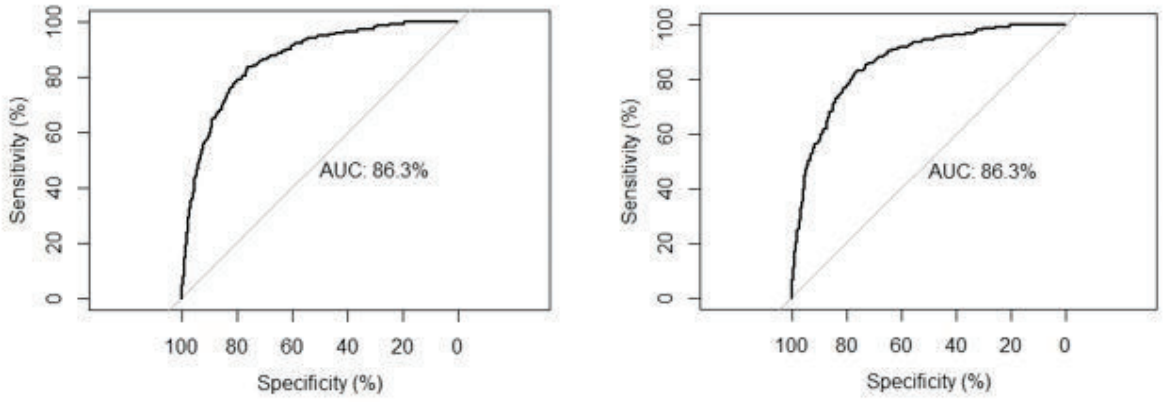


**Figura 3.16:** ROC Modelamiento (Izq.) y Validación (Der.) 13-16pm

Con la muestra de validación se obtuvo un valor de 0.8629, el cual es bastante cercano al obtenido con la muestra de modelamiento, como es de esperarse al no existir evidencias de sobreajuste con el estadístico KS.

- **Contacto efectivo en el horario de 16-21pm con respecto a no contactar en ningún horario.**

  Como resultado con la muestra de modelamiento se obtiene un índice AUROC de 0,831, lo cual quiere decir que existe una probabilidad de 0,831 de que la probabilidad estimada de contactar a un individuo de en el horario de 16-21pm sea mayor que la probabilidad estimada de un individuo de no contactarlo en ningún horario, elegidos aleatoriamente. Este valor permite decir que el modelo presenta un alto rendimiento de clasificación en esta categoría. A continuación se muestra la representación gráfica de de la curva ROC, con la probabilidad de contacto efectivo en el horario de 16-21pm para la muestra de modelamiento y para la muestra de validación respectivamente.

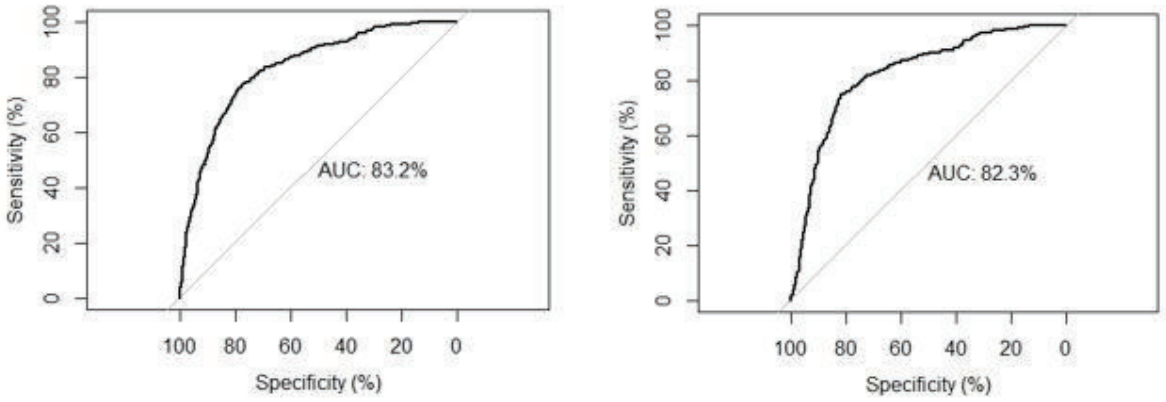


**Figura 3.17:** ROC Modelamiento (Izq.) y Validación (Der.) 16-21pm

Con la muestra de validación se obtuvo un valor de 0,823, el cual es bastante cercano al obtenido con la muestra de modelamiento, como es de esperarse al no existir evidencias de sobreajuste con el estadístico KS.

Finalmente, para obtener una medida global del rendimiento de clasificación del modelo, se calcula el VUS definido a detalle en la sección 2.3, tomando el máximo de las probabilidades estimadas de cada individuo obteniendose un valor de 0,6687 con la muestra de modelamiento y un valor de 0,66 con la muestra de validación. Por tanto se puede concluir que, en general, el modelo no tiene un alto rendimiento (esto en comparación con los valores de AUROC obtenidos para cada categoría que están por encima de 0.8).

**Coeficiente Gini**

El coeficiente Gini está relacionado con el indice AUROC mediante la igualdad

$$GINI = 2 * AUC - 1$$

Se muestran a continuación los valores de GINI para cada categoría.

- **Contacto efectivo en el horario de 7-9am con respecto a no contactar en ningún horario.**
  Se obtuvo un valor para el coeficiente de GINI de 0,765 para la muestra de modelamiento y un valor de 0,752 para la muestra de validación. Considerando los valores de GINI y AUROC se puede concluir que el modelo presenta un excelente rendimiento de clasificación para esta categoría.

- **Contacto efectivo en el horario de 9-13pm con respecto a no contactar en ningún horario.**
  Se obtuvo un valor para el coeficiente de GINI de 0,672 para la muestra de modelamiento y un valor de 0,697 para la muestra de validación. Considerando los valores de GINI y AUROC se puede concluir que el modelo presenta un excelente rendimiento de clasificación para esta categoría.

- **Contacto efectivo en el horario de 13-16pm con respecto a no contactar en ningún horario.**
  Se obtuvo un valor para el coeficiente de GINI de 0,727 para la muestra de modelamiento y un valor de 0,726 para la muestra de validación. Considerando los valores de GINI y AUROC se puede concluir que el modelo presenta un excelente rendimiento de clasificación para esta categoría.

- **Contacto efectivo en el horario de 16-21pm con respecto a no contactar en ningún horario.**
  Se obtuvo un valor para el coeficiente de GINI de 0,664 para la muestra de modelamiento y un valor de 0,647 para la muestra de validación. Considerando los valores de GINI y AUROC se puede concluir que el modelo presenta un excelente rendimiento de clasificación para esta categoría.

**Tablas de Clasificación**

Se muestran a continuación las tablas de clasificación para la variable dependiente $Y$ y la variable pronosticada $\widehat{Y}$

- **Contacto efectivo en el horario de 7-9am**
  Analizando los elementos de la diagonal de la tabla 3.16, se tiene que un 83,74 % y un 79,14 % de los individuos contactados en el horario de 7-9am y no contactados en ningún horario respectivamente son clasificados correctamente.

**Tabla 3.16:** Tabla de contingencia horario 7-9am Modelamiento

| $Y_1 \setminus \widehat{Y_1}$ | **NC** | **7-9am** |
|---|---|---|
| **NC** | 79,14 % | 20,86 % |
| **7-9am** | 16,26 % | 83,74 % |

Con la muestra de validación se obtiene la tabla 3.17, donde se tiene que un 81,88 % y un 79,71 % de los individuos contactados en el horario de 7-9am y no contactados en ningún horario respectivamente son clasificados correctamente. Los valores resultantes de la muestra de modelamiento y de la muestra de validación son bastante cercanos, de esto se puede concluir que el modelo presenta un excelente poder de clasificación en esta categoría.

**Tabla 3.17:** Tabla de contingencia horario 7-9am Validación

| $Y_1 \setminus \widehat{Y_1}$ | **NC** | **7-9am** |
|---|---|---|
| **NC** | 79,71 % | 20,29 % |
| **7-9am** | 18,12 % | 81,88 % |

- **Contacto efectivo en el horario de 9-13pm**
  Analizando los elementos de la diagonal de la tabla 3.18, se tiene que un 81,17 % y un 74,63 % de los individuos contactados en el horario de 9-13pm y no contactados en ningún horario respectivamente son clasificados correctamente.

**Tabla 3.18:** Tabla de contingencia horario 9-13pm Modelamiento

| $Y_2 \setminus \widehat{Y_2}$ | **NC** | **9-13pm** |
|---|---|---|
| **NC** | 74,63 % | 25,37 % |
| **9-13pm** | 18,83 % | 81,17 % |

Con la muestra de validación se obtiene la tabla 3.19, donde se tiene que un 78,10 % y un 81,59 % de los individuos contactados en el horario de 9-13am y no contactados en ningún horario respectivamente son clasificados correctamente.

Los valores resultantes de la muestra de modelamiento y de la muestra de validación son similares, de esto se puede concluir que el modelo presenta una buena calidad de clasificación en esta categoría.

**Tabla 3.19:** Tabla de contingencia horario 9-13pm Validación

| $Y_2 \setminus \widehat{Y_2}$ | **NC** | **9-13pm** |
|---|---|---|
| **NC** | 81,59 % | 18,41 % |
| **9-13pm** | 21,90 % | 78,10 % |

- **Contacto efectivo en el horario de 13-16pm**
  Analizando los elementos de la diagonal de la tabla 3.20, se tiene que un 83,53 % y un 75,97 % de los individuos contactados en el horario de 13-16pm y no contactados en ningún horario respectivamente son clasificados correctamente.

**Tabla 3.20:** Tabla de contingencia horario 13-16pm Modelamiento

| $Y_3 \setminus \widehat{Y_3}$ | **NC** | **13-16pm** |
|---|---|---|
| **NC** | 75,97 % | 24,03 % |
| **13-16pm** | 16,47 % | 83,53 % |

Con la muestra de validación se obtiene la tabla 3.21, donde se tiene que un 80,34 % y un 77,90 % de los individuos contactados en el horario de 13-16pm y no contactados en ningún horario respectivamente son clasificados correctamente. Los valores resultantes de la muestra de modelamiento y de la muestra de validación son bastante cercanos, de esto se puede concluir que el modelo presenta un excelente poder de clasificación en esta categoría.

**Tabla 3.21:** Tabla de contingencia horario 13-16pm Validación

| $Y_3 \setminus \widehat{Y_3}$ | **NC** | **13-16pm** |
|---|---|---|
| **NC** | 77,90 % | 22,10 % |
| **13-16pm** | 19,66 % | 80,34 % |

- **Contacto efectivo en el horario de 16-21pm Modelamiento**
  Analizando los elementos de la diagonal de la tabla 3.22, se tiene que un 77,48 % y un 77,27 % de los individuos contactados en el horario de 16-21pm y no contactados en ningún horario respectivamente son clasificados correctamente.

**Tabla 3.22:** Tabla de contingencia horario 16-21pm Modelamiento

| $Y_4 \setminus \widehat{Y}_4$ | **NC** | **16-21pm** |
|---|---|---|
| **NC** | 77,27 % | 22,73 % |
| **16-21pm** | 22,52 % | 77,48 % |

Con la muestra de validación se obtiene la tabla 3.23, donde se tiene que un 74,77 % y un 81,16 % de los individuos contactados en el horario de 13-16pm y no contactados en ningún horario respectivamente son clasificados correctamente. Los valores resultantes de la muestra de modelamiento y de la muestra de validación son bastante cercanos, de esto se puede concluir que el modelo presenta un excelente poder de clasificación en esta categoría.

**Tabla 3.23:** Tabla de contingencia horario 16-21pm Validación

| $Y_4 \setminus \widehat{Y}_4$ | **NC** | **16-21pm** |
|---|---|---|
| **NC** | 81,16 % | 18,84 % |
| **16-21pm** | 25,23 % | 74,77 % |

Finalmente, al tomar el máximo de las probabilidades estimadas de cada individuo, se presentan las tablas de contingencia para la muestra de modelamiento y para la muestra de validación.

**Tabla 3.24:** Tabla de contingencia multinomial modelamiento

| $Y \setminus \widehat{Y}$ | **7-9am** | **9-13pm** | **13-16pm** | **16-21pm** | **NC** |
|---|---|---|---|---|---|
| **7-9am** | **3,94** | 4,93 | 10,84 | 30,05 | 50,25 |
| **9-13pm** | 0,22 | **3,68** | 5,63 | 25,97 | 64,5 |
| **13-16pm** | 1,2 | 4,19 | **6,89** | 35,03 | 52,69 |
| **16-21pm** | 0,78 | 3,11 | 4,66 | **27,38** | 64,08 |
| **NC** | 0,1 | 0,62 | 0,67 | 3,99 | **94,62** |

Al comparar las tasas de clasificación correcta (elementos de la diagonal) con las tasa de clasificación errónea (elementos fuera de la diagonal) para cada categoría de la tabla 3.24, se observa que la clasificación correcta no es muy elevada excepto en la categoría NC.

Al comparar las tasas de clasificación correcta (elementos de la diagonal) con las tasa de clasificación errónea (elementos fuera de la diagonal) para cada categoría de la tabla 3.25, se observa que, al igual que en la tabla de modelamiento, la clasificación correcta no es elevada, excepto en la categoría NC.

Por tanto, no podemos concluir que al tomar las probabilidades mas altas para cada individuo, el modelo tiene un excelente poder de discriminación.

**Tabla 3.25:** Tabla de contingencia multinomial validación

| $Y \setminus \widehat{Y}$ | **7-9am** | **9-13pm** | **13-16pm** | **16-21pm** | **NC** |
|---|---|---|---|---|---|
| **7-9am** | **5,8** | 21,74 | 12,32 | 6,52 | 53,62 |
| **9-13pm** | 1,27 | **19,05** | 7,62 | 7,3 | 64,76 |
| **13-16pm** | 0,85 | 20,94 | **11,11** | 11,54 | 55,56 |
| **16-21pm** | 1,82 | 20,67 | 6,99 | **5,78** | 64,74 |
| **NC** | 0,22 | 2,1 | 0,8 | 1,81 | **95,07** |

Tomando en cuenta todas las medidas e indicadores expuestos anteriormente, resumidos en la tabla 3.26, se puede concluir que en términos de modelamiento, el mejor horario para tener contacto efectivo es el de 7-9am, seguido por el de 13-16pm.

**Tabla 3.26:** Resumen de resultados por categoría

| | **MODELAMIENTO** | | | **VALIDACION** | | |
|---|---|---|---|---|---|---|
| | **KS** | **AUROC** | **GINI** | **KS** | **AUROC** | **GINI** |
| **7-9am** | **0,636** | **0,882** | **0,752** | **0,6268** | **0,876** | **0,765** |
| **9-13pm** | 0,559 | 0,836 | 0,672 | 0,599 | 0,848 | 0,697 |
| **13-16pm** | **0,596** | **0,863** | **0,727** | **0,593** | **0,863** | **0,726** |
| **16-21pm** | 0,548 | 0,831 | 0,664 | 0,564 | 0,823 | 0,647 |

Por otro lado, al tomar la máxima probabilidad estimada para cada individuo, en medidas globales del modelo, los indicadores son relativamente bajos (tabla 3.27), por tanto no se puede concluir que el modelo multinomial da el mejor horario para contactar a un individuo, pero se pueden implementar mayores esfuerzos en los horarios de 7-9am y de 13-16pm para aumentar la contactibilidad.

**Tabla 3.27:** Resumen resultados globales

| **MODELAMIENTO** | | **VALIDACION** | |
|---|---|---|---|
| **KSM** | **VUS** | **KSM** | **VUS** |
| 0,3988 | 0,669 | 0,406 | 0,660 |

### 3.5.4. Tablas Performance

Las tablas performance, también conocida como tablas de desempeño o de rendimiento son una herramienta que permiten visualizar la calidad de discriminación de los modelos de regresión logística binomial. Para analizar el rendimiento de clasificación de un modelo, es usual realizar una partición en 10 intervalos de la probabilidad estimada y se analiza el número de individuos totales, el número de individuos contactados y no contactados y sus porcentajes en cada intervalo.

La tabla consta de los siguientes elementos:

- **Intervalo de Score:** (De-Hasta), Los intervalos son abiertos a la izquierda y cerrados a la derecha, resultantes de la partición en deciles de la probabilidad estimada. La probabilidad estimada se multiplica por 1000 para obtener un SCORE de cada individuo para ser contactado en cada horario.
- **Clientes:** Número de individuos en cada decil.
- **%Clien:** Porcentaje de individuos columna.
- **Acum Clien:** Porcentaje acumulado de individuos.
- **#NC:** Número de individuos no contactados en cada decil.
- **%NC:** Porcentaje de individuos no contactados columna.
- **Acum NC:** Porcentaje acumulado de individuos no contactados.
- **#Cont:** Número de individuos contactados en cada decil.
- **%Cont:** Porcentaje de individuos contactados columna.
- **Acum Cont:** Porcentaje acumulado de individuos contactados.
- **%NC D:** Porcentaje de individuos no contactados en cada decil.
- **%C D:** Porcentaje de individuos contactados en cada decil.
- **C:NC.-** Razón individuos contactados vs no contactados en cada decil.

Con respecto al campo **%Cont:**, se espera que los valores crezcan estrictamente a medida de la probabilidad estimada aumenta.

Con respecto al campo **%NC:**, es de esperar que los valores decrezcan a medida de la probabilidad estimada aumenta.

El campo **C:NC.-** se espera que crezca a medida que la probabilidad aumenta.

A continuación se presentan las tablas performance resultantes de la muestra de validación para cada horario.

- **Contacto efectivo en el horario de 7-9am**

**Tabla 3.28:** Tabla Performance 7-9am Validación

| Decil | De | Hasta | Clientes | %Clien | Acum_Clien | #NC | %NC | Acum_NC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 8 | 202 | 13.31 | 13.31 | 200 | 14,49 | 14,49 |
| 2 | 8 | 11 | 153 | 10,08 | 23,39 | 152 | 11,01 | 25,5 |
| 3 | 11 | 14 | 145 | 9,55 | 32,94 | 142 | 10,29 | 35,79 |
| 4 | 14 | 17 | 162 | 10,67 | 43,61 | 161 | 11,67 | 47,46 |
| 5 | 17 | 19,5 | 97 | 6,39 | 50 | 95 | 6,88 | 54,34 |
| 6 | 19,5 | 24 | 161 | 10,61 | 60,61 | 156 | 11,3 | 65,64 |
| 7 | 24 | 32 | 150 | 9,88 | 70,49 | 142 | 10,29 | 75,93 |
| 8 | 32 | 49,6 | 144 | 9,49 | 79,98 | 131 | 9,49 | 85,42 |
| 9 | 49,6 | 99 | 153 | 10,08 | 90,06 | 122 | 8,84 | 94,26 |
| 10 | 99 | 999 | 151 | 9,95 | 100,01 | 79 | 5,72 | 99,98 |

| Decil | De | Hasta | #Cont | %Cont | Acum_Cont | %NC_D | %C_D | C:NC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 8 | 2 | 1,45 | 1,45 | 0,99 | 0,01 | 0,01 |
| 2 | 8 | 11 | 1 | 0,72 | 2,17 | 0,99 | 0,01 | 0,01 |
| 3 | 11 | 14 | 3 | 2,17 | 4,34 | 0,98 | 0,02 | 0,02 |
| 4 | 14 | 17 | 1 | 0,72 | 5,06 | 0,99 | 0,01 | 0,01 |
| 5 | 17 | 19,5 | 2 | 1,45 | 6,51 | 0,98 | 0,02 | 0,02 |
| 6 | 19,5 | 24 | 5 | 3,62 | 10,13 | 0,97 | 0,03 | 0,03 |
| 7 | 24 | 32 | 8 | 5,8 | 15,93 | 0,95 | 0,05 | 0,06 |
| 8 | 32 | 49,6 | 13 | 9,42 | 25,35 | 0,91 | 0,09 | 0,1 |
| 9 | 49,6 | 99 | 31 | 22,46 | 47,81 | 0,8 | 0,2 | 0,25 |
| 10 | 99 | 999 | 72 | 52,17 | 99,98 | 0,52 | 0,48 | 0,91 |

En la tabla 3.28 se observa que el porcentaje de individuos contactados y la razón de individuos contactados vs los individuos no contactados aumentan a con la probabilidad de contacto efectivo y que el porcentaje de individuos no contactados tiende a decrecer a medida que la probabilidad aumenta como se esperaba.

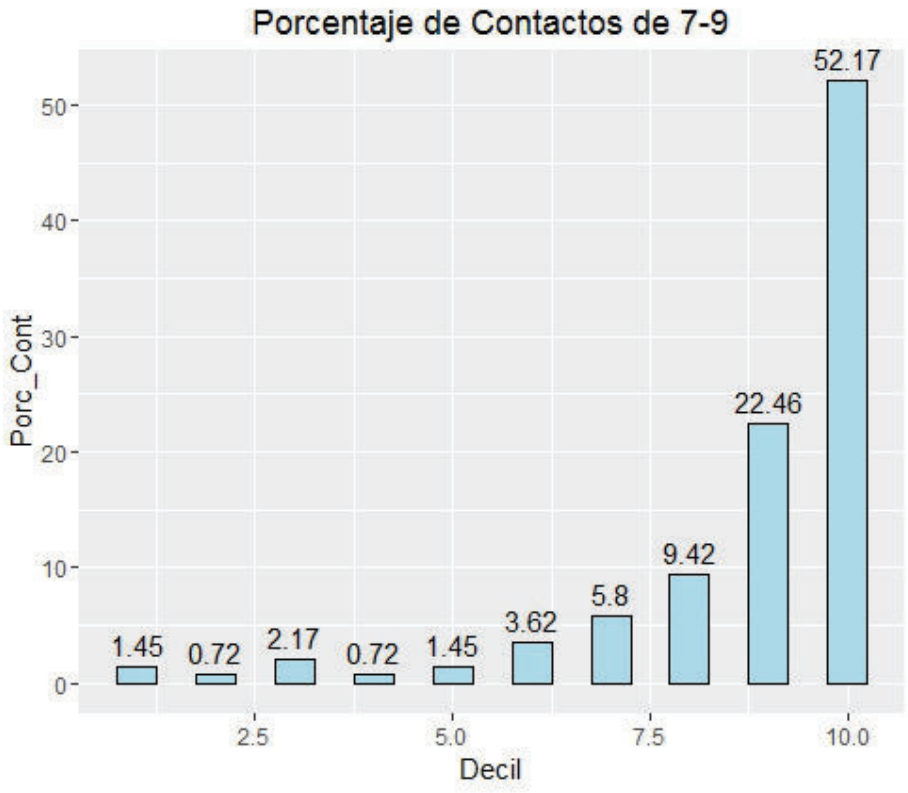


**Figura 3.18:** Porcentaje Contactados

En la figura 3.18 se observa detalladamente el crecimiento del porcentaje de individuos contactados en esta categoría.

- **Contacto efectivo en el horario de 9-13pm**

**Tabla 3.29:** Tabla Performance 9-13pm Validación

| Decil | De | Hasta | Clientes | %Clien | Acum_Clien | #NC | %NC | Acum_NC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 45 | 177 | 10,44 | 10,44 | 175 | 12,68 | 12,68 |
| 2 | 45 | 55 | 191 | 11,27 | 21,71 | 181 | 13,12 | 25,8 |
| 3 | 55 | 62 | 171 | 10,09 | 31,8 | 167 | 12,1 | 37,9 |
| 4 | 62 | 69 | 143 | 8,44 | 40,24 | 136 | 9,86 | 47,76 |
| 5 | 69 | 79 | 168 | 9,91 | 50,15 | 157 | 11,38 | 59,14 |
| 6 | 79 | 94 | 174 | 10,27 | 60,42 | 163 | 11,81 | 70,95 |
| 7 | 94 | 123 | 163 | 9,62 | 70,04 | 140 | 10,14 | 81,09 |
| 8 | 123 | 181,2 | 169 | 9,97 | 80,01 | 109 | 7,9 | 88,99 |
| 9 | 181,2 | 231,6 | 169 | 9,97 | 89,98 | 89 | 6,45 | 95,44 |
| 10 | 231,6 | 999 | 170 | 10,03 | 100,01 | 63 | 4,57 | 100,01 |

| Decil | De | Hasta | #Cont | %Cont | Acum_Cont | %NC_D | %C_D | C:NC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 45 | 2 | 0,63 | 0,63 | 0,99 | 0,01 | 0,01 |
| 2 | 45 | 55 | 10 | 3,17 | 3,8 | 0,95 | 0,05 | 0,06 |
| 3 | 55 | 62 | 4 | 1,27 | 5,07 | 0,98 | 0,02 | 0,02 |
| 4 | 62 | 69 | 7 | 2,22 | 7,29 | 0,95 | 0,05 | 0,05 |
| 5 | 69 | 79 | 11 | 3,49 | 10,78 | 0,93 | 0,07 | 0,07 |
| 6 | 79 | 94 | 11 | 3,49 | 14,27 | 0,94 | 0,06 | 0,07 |
| 7 | 94 | 123 | 23 | 7,3 | 21,57 | 0,86 | 0,14 | 0,16 |
| 8 | 123 | 181,2 | 60 | 19,05 | 40,62 | 0,64 | 0,36 | 0,55 |
| 9 | 181,2 | 231,6 | 80 | 25,4 | 66,02 | 0,53 | 0,47 | 0,9 |
| 10 | 231,6 | 999 | 107 | 33,97 | 99,99 | 0,37 | 0,63 | 1,7 |

En la tabla 3.29 se observa que el porcentaje de individuos contactados y la razón de individuos contactados vs los individuos no contactados aumentan a con la probabilidad de contacto efectivo y que el porcentaje de individuos no contactados tiende a decrecer a medida que la probabilidad aumenta como se esperaba.

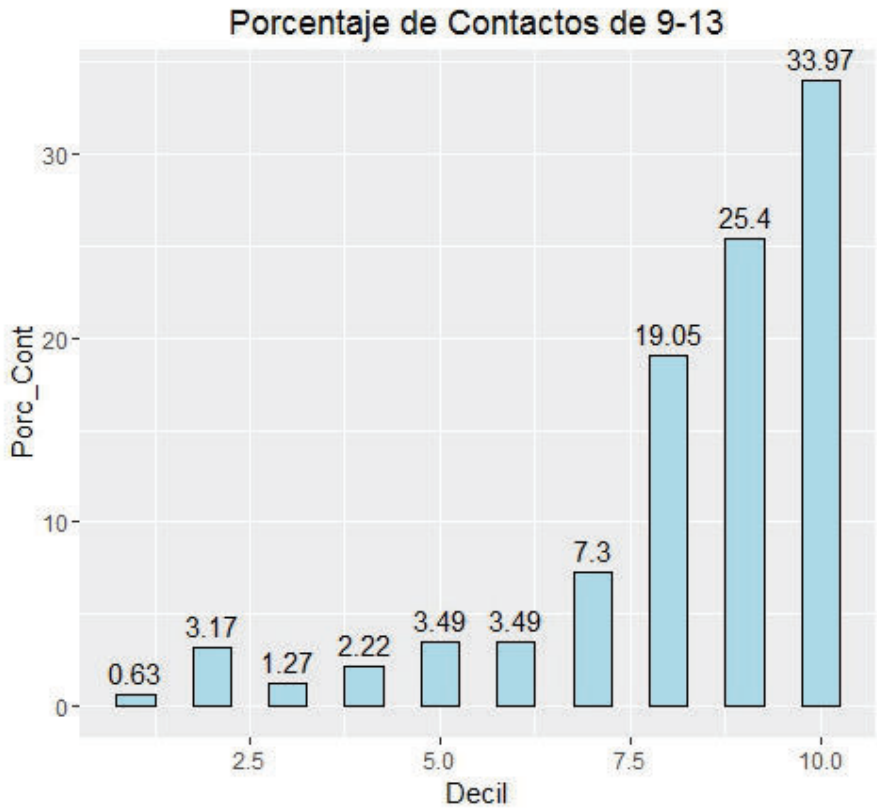


**Figura 3.19:** Porcentaje Contactados

En la figura 3.19 se observa detalladamente el crecimiento del porcentaje de individuos contactados en esta categoría.

- **Contacto efectivo en el horario de 13-16pm**

**Tabla 3.30:** Tabla Performance 13-16pm Validación

| Decil | De | Hasta | Clientes | %Clien | Acum_Clien | #NC | %NC | Acum_NC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 17 | 176 | 10,9 | 10,9 | 176 | 12,75 | 12,75 |
| 2 | 17 | 23 | 157 | 9,73 | 20,63 | 154 | 11,16 | 23,91 |
| 3 | 23 | 32 | 180 | 11,15 | 31,78 | 174 | 12,61 | 36,52 |
| 4 | 32 | 36 | 136 | 8,43 | 40,21 | 134 | 9,71 | 46,23 |
| 5 | 36 | 43 | 168 | 10,41 | 50,62 | 160 | 11,59 | 57,82 |
| 6 | 43 | 55 | 166 | 10,29 | 60,91 | 155 | 11,23 | 69,05 |
| 7 | 55 | 75 | 152 | 9,42 | 70,33 | 133 | 9,64 | 78,69 |
| 8 | 75 | 106 | 157 | 9,73 | 80,06 | 118 | 8,55 | 87,24 |
| 9 | 106 | 189 | 161 | 9,98 | 90,04 | 112 | 8,12 | 95,36 |
| 10 | 189 | 999 | 161 | 9,98 | 100,02 | 64 | 4,64 | 100 |

| Decil | De | Hasta | #Cont | %Cont | Acum_Cont | %NC_D | %C_D | C:NC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 17 | 0 | 0 | 0 | 1 | 0 | 0 |
| 2 | 17 | 23 | 3 | 1,28 | 1,28 | 0,98 | 0,02 | 0,02 |
| 3 | 23 | 32 | 6 | 2,56 | 3,84 | 0,97 | 0,03 | 0,03 |
| 4 | 32 | 36 | 2 | 0,85 | 4,69 | 0,99 | 0,01 | 0,01 |
| 5 | 36 | 43 | 8 | 3,42 | 8,11 | 0,95 | 0,05 | 0,05 |
| 6 | 43 | 55 | 11 | 4,7 | 12,81 | 0,93 | 0,07 | 0,07 |
| 7 | 55 | 75 | 19 | 8,12 | 20,93 | 0,88 | 0,12 | 0,14 |
| 8 | 75 | 106 | 39 | 16,67 | 37,6 | 0,75 | 0,25 | 0,33 |
| 9 | 106 | 189 | 49 | 20,94 | 58,54 | 0,7 | 0,3 | 0,44 |
| 10 | 189 | 999 | 97 | 41,45 | 99,99 | 0,4 | 0,6 | 1,52 |

En la tabla 3.30 se observa que el porcentaje de individuos contactados y la razón de individuos contactados vs los individuos no contactados aumentan a con la probabilidad de contacto efectivo y que el porcentaje de individuos no contactados tiende a decrecer a medida que la probabilidad aumenta como se esperaba.

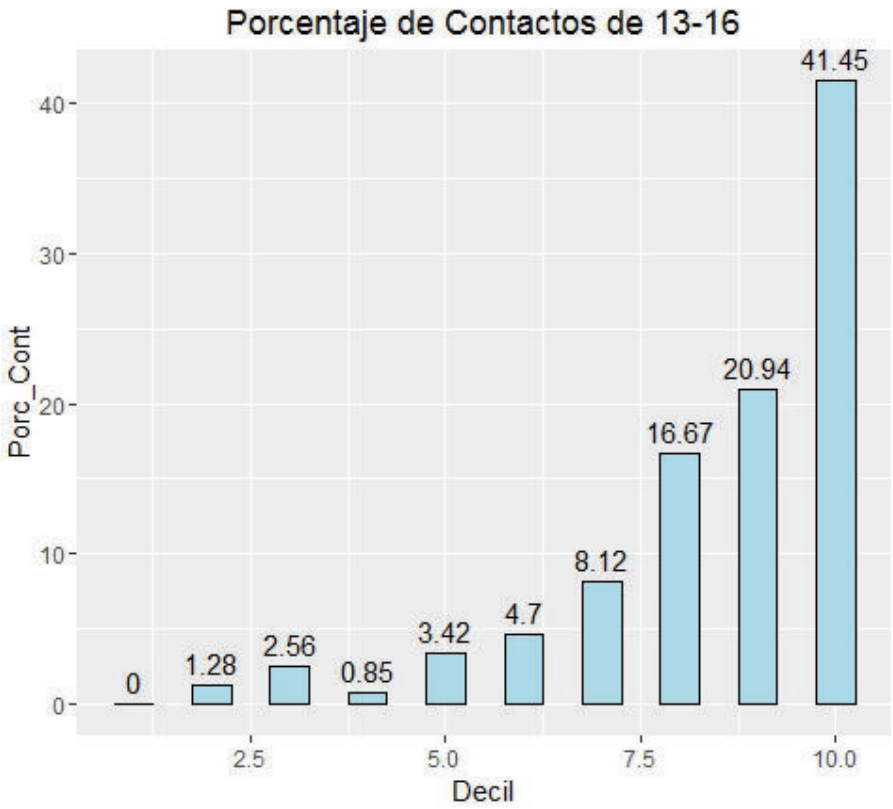


**Figura 3.20:** Porcentaje Contactados

En la figura 3.20 se observa detalladamente el crecimiento del porcentaje de individuos contactados en esta categoría.

- **Contacto efectivo en el horario de 16-21pm**

**Tabla 3.31:** Tabla Performance 16-21pm Validación

| Decil | De | Hasta | Clientes | %Clien | Acum_Clien | #NC | %NC | Acum_NC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 37 | 179 | 10,47 | 10,47 | 178 | 12,9 | 12,9 |
| 2 | 37 | 57 | 172 | 10,06 | 20,53 | 167 | 12,1 | 25 |
| 3 | 57 | 67 | 182 | 10,65 | 31,18 | 168 | 12,17 | 37,17 |
| 4 | 67 | 74 | 157 | 9,19 | 40,37 | 144 | 10,43 | 47,6 |
| 5 | 74 | 84 | 181 | 10,59 | 50,96 | 169 | 12,25 | 59,85 |
| 6 | 84 | 97 | 163 | 9,54 | 60,5 | 148 | 10,72 | 70,57 |
| 7 | 97 | 120 | 165 | 9,65 | 70,15 | 142 | 10,29 | 80,86 |
| 8 | 120 | 166 | 170 | 9,95 | 80,1 | 109 | 7,9 | 88,76 |
| 9 | 166 | 211 | 172 | 10,06 | 90,16 | 79 | 5,72 | 94,48 |
| 10 | 211 | 999 | 168 | 9,83 | 99,99 | 76 | 5,51 | 99,99 |

| Decil | De | Hasta | #Cont | %Cont | Acum_Cont | %NC_D | %C_D | C:NC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 37 | 1 | 0,3 | 0,3 | 0,99 | 0,01 | 0,01 |
| 2 | 37 | 57 | 5 | 1,52 | 1,82 | 0,97 | 0,03 | 0,03 |
| 3 | 57 | 67 | 14 | 4,26 | 6,08 | 0,92 | 0,08 | 0,08 |
| 4 | 67 | 74 | 13 | 3,95 | 10,03 | 0,92 | 0,08 | 0,09 |
| 5 | 74 | 84 | 12 | 3,65 | 13,68 | 0,93 | 0,07 | 0,07 |
| 6 | 84 | 97 | 15 | 4,56 | 18,24 | 0,91 | 0,09 | 0,1 |
| 7 | 97 | 120 | 23 | 6,99 | 25,23 | 0,86 | 0,14 | 0,16 |
| 8 | 120 | 166 | 61 | 18,54 | 43,77 | 0,64 | 0,36 | 0,56 |
| 9 | 166 | 211 | 93 | 28,27 | 72,04 | 0,46 | 0,54 | 1,18 |
| 10 | 211 | 999 | 92 | 27,96 | 100 | 0,45 | 0,55 | 1,21 |

En la tabla 3.31 se observa que el porcentaje de individuos contactados y la razón de individuos contactados vs los individuos no contactados aumentan a con la probabilidad de contacto efectivo y que el porcentaje de individuos no contactados tiende a decrecer a medida que la probabilidad aumenta como se esperaba.

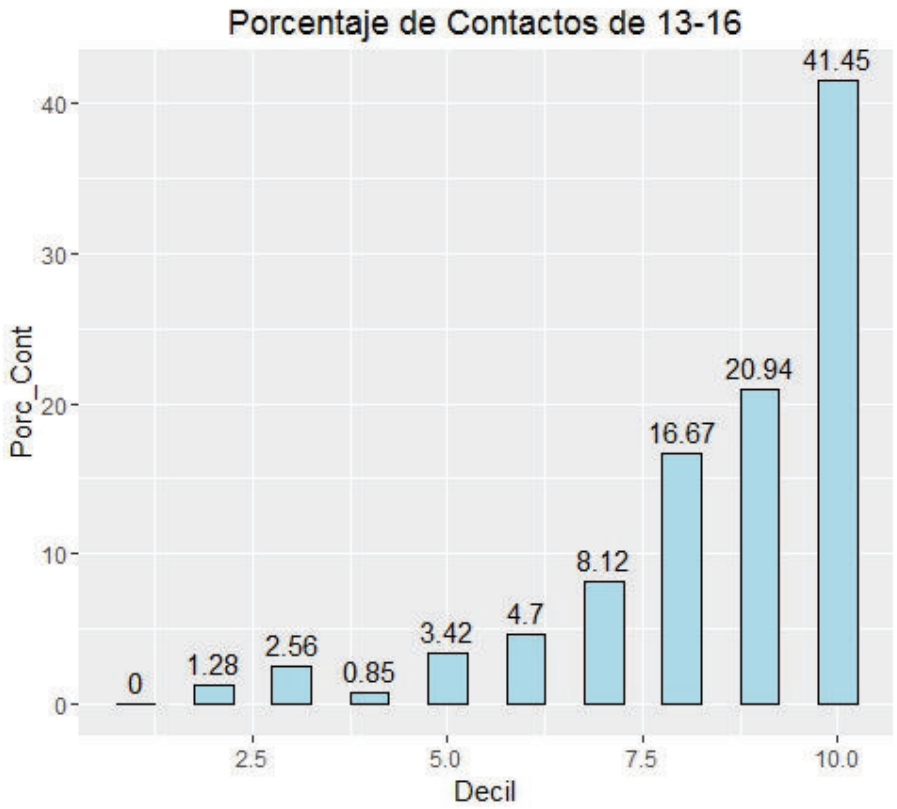


**Figura 3.21:** Porcentaje Contactados

En la figura 3.21 se observa detalladamente el crecimiento del porcentaje de individuos contactados en esta categoría.

# Capítulo 4

# Modelo de Contactabilidad General

Debido a la cantidad de individuos en la categoría NC de la regresión logística multinomial de la sección anterior, una alternativa para modelar este comportamiento sería mediante los llamados *modelos inflados en cero*. El exceso de ceros es común en la práctica, por lo que se han desarrollado modelos estadísticos que describen este fenómeno y, por lo tanto, permiten derivar conclusiones realistas y confiables a partir de las inferencias.

El inflamiento con ceros ocurre cuando hay una gran frecuencia de observaciones iguales a cero de tal manera que ninguna de las distribuciones discretas estándar proporciona un ajuste adecuado. Para modelar el exceso de ceros es fundamental entender la naturaleza del origen de los ceros. De acuerdo a lo anterior, los ceros se clasifican en dos tipos: ceros estructurales y ceros muestrales.

Para este caso de estudio, si un cliente no fue contactado en la ventana de tiempo 9 meses, sería un cero estructural, pero si solo aparece no contactado en la ventana de desempeño (3.1), sería un cero muestral. Así, un cero inevitable es un cero estructural, mientras que un cero que ocurre debido al mecanismo de muestreo es un cero muestral.

Una vez conocida la naturaleza de los ceros, se deben considerar los escenarios solo con ceros estructurales, solo con ceros muestrales o una combinación de ambos. A continuación se presentan los procedimientos mas comunes:

- **Modelos Lineales Generalizados:** Enfoque de modelación consiste en ajustar un modelo lineal generalizado en el cual la variable respuesta Y se modela con alguna distribución discreta que pertenece a la familia exponencial. Cuando se usa a la distribución Poisson, el modelo lineal generalizado es conocido como el modelo de regresión Poisson. Sin embargo, ante la presencia de exceso de ceros, el ajuste de un modelo de regresión Poisson generalmente será pobre. Por lo que se debe adecuar al modelo para que considere al exceso de ceros. Una exposición

amplia sobre la teoría y aplicaciones de los modelos de regresión Poisson sin exceso de ceros y en general de los modelos lineales generalizados.

- **Modelos en dos Partes:** Para analizar datos con exceso de ceros estructurales se han propuesto los modelos de dos partes, también conocidos como modelos condicionales. El enfoque consiste en primero modelar la presencia/ausencia (conteos diferentes de cero y ceros) con un modelo de regresión logística. Después se condiciona sobre los datos de conteo positivos y se modelan éstos con una distribución discreta con el cero truncado. Si se usa la distribución Poisson, se tiene el modelo Poisson de dos componentes. En estos modelos, todos los ceros se modelan en el componente presencia/ausencia.

  Un supuesto fundamental de este enfoque de análisis es que los ceros se originan a partir de un mecanismo simple que no afecta a las observaciones diferentes de cero. Una ventaja computacional de este enfoque es que es posible ajustar estos modelos en dos etapas. Primero se ajustan los ceros y los no ceros con el modelo de regresión logística y posteriormente se ajustan los conteos positivos usando la distribución Poisson con el cero truncado. De este modo, la log-verosimilitud del modelo es la suma de la log-verosimilitud de cada componente. Estos modelos son fáciles de ajustar e interpretar.

- **Modelos de Mezclas:** Los modelos de mezclas de distribuciones son combinaciones lineales convexas de distribuciones de probabilidad. Los ceros de estos modelos pueden ser una mezcla de ceros estructurales y muestrales. De esta clase de modelos, el llamado modelo de regresión Poisson Inflado con Ceros (PIC) es el más usado para datos de conteo con exceso de ceros. En estos modelos los ceros se dividen en dos grupos, uno tiene los ceros provenientes de la distribución que genera a la variable respuesta, el otro grupo tiene a los ceros extra. Los ceros del primer grupo se modelan con la distribución Poisson. Un cero en este grupo ocurre con probabilidad $1-p$ mientras que los ceros extra ocurren con probabilidad $p$.

  Otra distribución que se ha propuesto bajo este enfoque es la distribución binomial negativa. Esta distribución es particularmente adecuada para cuando además del exceso de ceros también se presenta sobredispersión. La recomendación es que si el exceso de ceros consiste de ceros muestrales, se debe usar un modelo de mezcla de distribuciones. Aunque la interpretación de estos modelos no es tan sencilla como la de los modelos de dos partes.

Hasta donde es de nuestro conocimiento, aún no se dispone de una discusión formal en la literatura acerca de cómo modelar datos con ceros estructurales y de muestreo.

Este es, al parecer, un problema abierto en modelación estadística. Es posible que un enfoque de inferencia Bayesiano, donde el modelo incorpore información acerca de los ceros muestrales como información a priori, pudiera dar respuestas de utilidad. [11]

Para el presente trabajo no fue necesario el uso de la metodología con modelos inflados en cero, dado que el porcentaje de no contactados no es elevado con respecto a los individuos contactados (en todos los horarios), por tanto, en esta sección se presenta el modelo resultante con dos categorías, equivalente a realizar una regresión logística binaria.

$$\sum_{j=1}^{k-1} p_j(x) = 1 - p_k(x)$$

El valor de $\sum_{j=1}^{k-1} p_j(x)$ es la probabilidad de tener contacto efectivo en cualquier horario y $p_k(x)$ es la probabilidad de no contactar en ningún horario, por tanto la variable dependiente sería

$$Y = \begin{cases} 1 & \text{si hubo contacto efectivo} \\ 0 & \text{caso contrario} \end{cases}$$

A continuación se presentan los resultados obtenidos para este caso, con la suma de las probabilidades estimadas para categoría de horario con el modelo multinomial.

- **KS**
  Para la muestra de modelamiento se obtiene un valor de KS (definido a detalle en la sección 2.2.1) de 0.56454 y se puede observar la separación de las distribuciones empíricas en la figura 4.1.

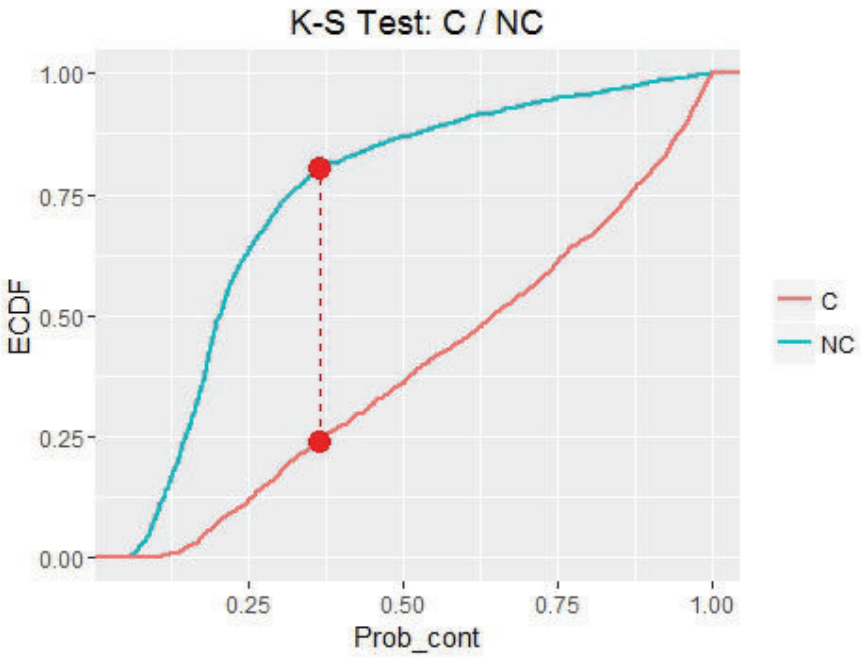


**Figura 4.1:** KS C/NC Modelamiento

Mientras que para la muestra de validación se obtiene un valor de KS de 0,58217 y se puede observar la separación de las distribuciones empíricas en la figura 4.2

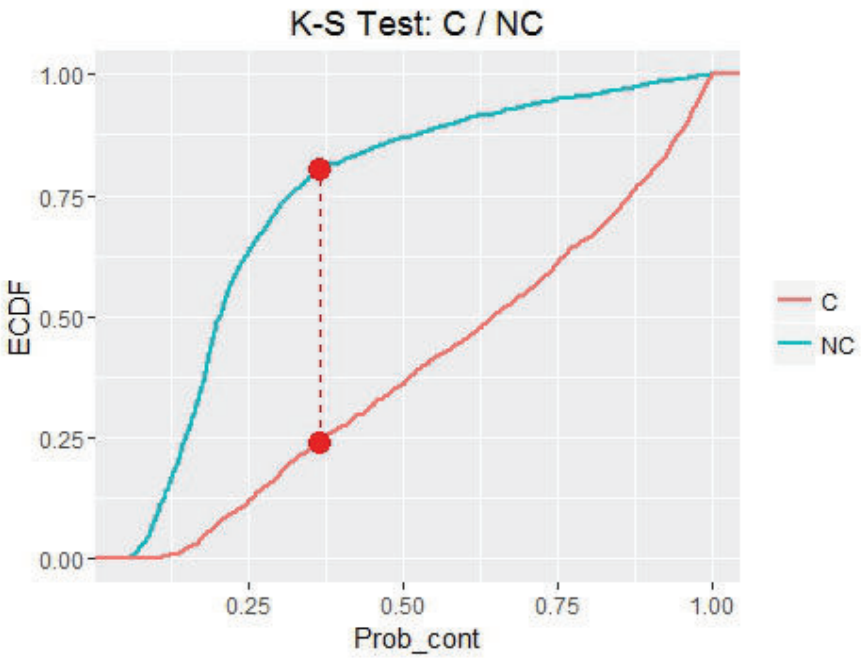


**Figura 4.2:** KS C/NC Validación

Al ser los valores de KS para la muestra de modelamiento y validación son muy similares y mayores a 0,5, por tanto se puede decir que el modelo es altamente discriminativo y no existe evidencia de sobreajuste.

- **Área bajo la curva ROC**

  En la figura 4.3 se tiene la representación gráfica de la curva ROC con la probabilidad de contacto efectivo estimada para la muestra de modelamiento y se obtiene un valor de índice AUROC de 0,8484 lo que quiere decir que existe una probabilidad de 0,8484 de que la probabilidad de contacto efectivo estimada de un individuo contactado se mayor que la probabilidad de contacto efectivo de un individuo no contactado elegidos aleatoriamente.

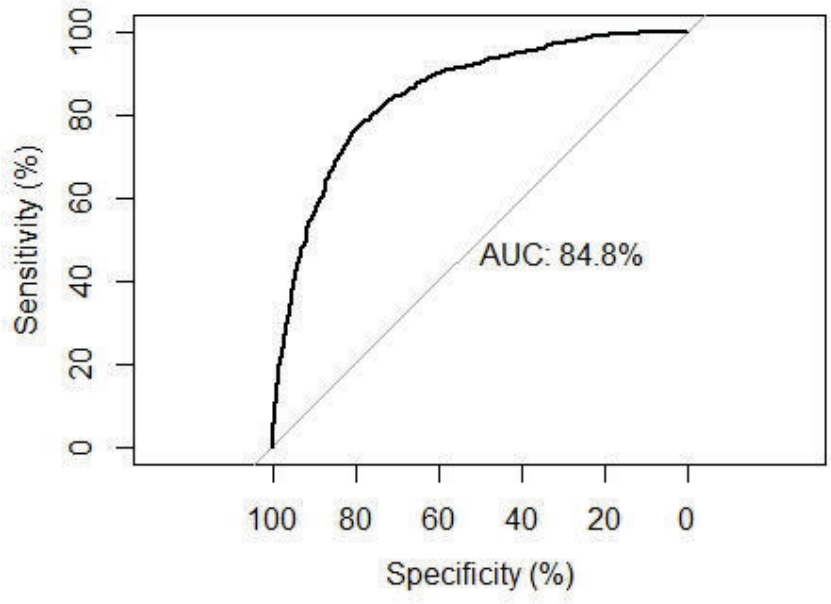


**Figura 4.3:** ROC C/NC Modelamiento

En la figura 4.4 se tiene la representación gráfica de la curva ROC con la probabilidad de contacto efectivo estimada para la muestra de validación y se obtiene un valor de índice AUROC (definido a detalle en la sección 2.3) de 0,8532. Se observa que la gráfica obtenida con la muestra de validación es muy similar a la obtenida con la muestra de modelamiento y el valor AUROC son muy similares, por tanto se puede decir que el modelo presenta un alto rendimiento de clasificación.

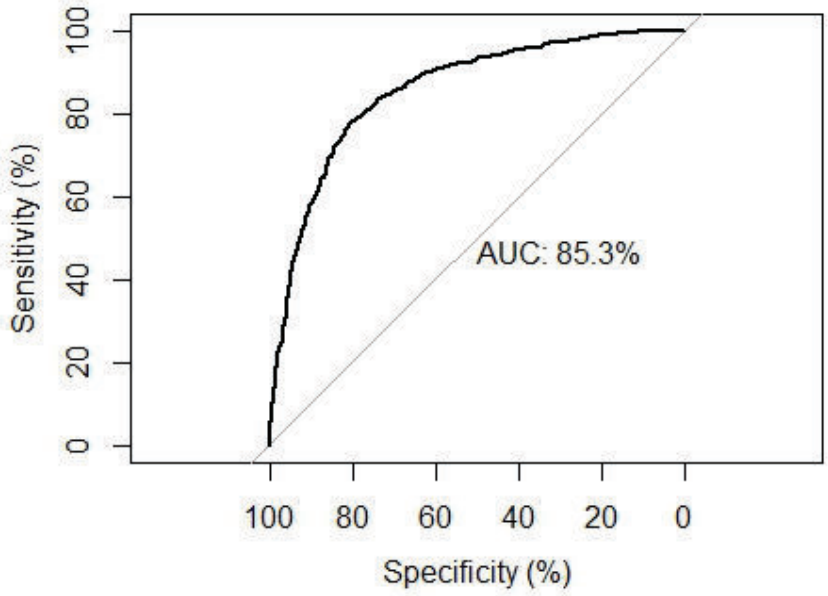


**Figura 4.4:** ROC C/NC Validación

- **GINI**
  Se obtuvo un valor del coeficiente GINI de 0,6968 para la muestra de modelamiento y un valor de 0,7064 para la muestra de validación, por tanto, considerando los valores de GINI y de AUROC se puede concluir que el modelo tiene un excelente rendimiento de clasificación.

- **Tablas de clasificación**
  Tabla de contingencia entre la variable dependiente $Y$ y la variable estimada $\widehat{Y}$ Analizando la diagonal de la tabla 4.1 se tiene un 76,09 % y un 80.30 % de individuos contactados y no contactados respectivamente son clasificados correctamente. Al ser estos valores elevados, se puede decir que el modelo tiene un alto poder de clasificación.

**Tabla 4.1:** Tabla de Clasificación Modelamiento

| $Y\widehat{Y}$ | **NC** | **C** |
|---|---|---|
| **NC** | 80,30 % | 19,70 % |
| **C** | 23,91 % | 76,09 % |

Para la muestra de validación se obtiene la tabla 4.2 donde se tiene que un 77,85 % y un 80,36 % de los individuos contactados y no contactados respectivamente fueron clasificados correctamente, estos valores son muy similares a los obtenidos

con la muestra de modelamiento, por tanto se puede concluir que el modelo presenta un excelente poder de discriminación.

**Tabla 4.2:** Tabla de Clasificación Validación

| $Y \setminus \widehat{Y}$ | **NC** | **C** |
|---|---|---|
| **NC** | 80,36 % | 19,64 % |
| **C** | 22,15 % | 77,85 % |

- **Tabla Performance**

  La tabla performance (definida a detalle en la sección 3.5.2) refleja el rendimiento del modelo binomial de contactabilidad.

**Tabla 4.3:** Tabla Performance C/NC Validación

| **Decil** | **De** | **Hasta** | **Clientes** | **%Clien** | **Acum_Clien** | **#NC** | **%NC** | **Acum_NC** |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 127 | 241 | 10,06 | 10,06 | 232 | 16,81 | 16,81 |
| 2 | 127 | 169 | 242 | 10,1 | 20,16 | 218 | 15,8 | 32,61 |
| 3 | 169 | 197 | 246 | 10,27 | 30,43 | 213 | 15,43 | 48,04 |
| 4 | 197 | 239 | 235 | 9,81 | 40,24 | 197 | 14,28 | 62,32 |
| 5 | 239 | 299 | 238 | 9,93 | 50,17 | 165 | 11,96 | 74,28 |
| 6 | 299 | 409 | 238 | 9,93 | 60,1 | 134 | 9,71 | 83,99 |
| 7 | 409 | 569,5 | 237 | 9,89 | 69,99 | 88 | 6,38 | 90,37 |
| 8 | 570 | 743 | 240 | 10,02 | 80,01 | 62 | 4,49 | 94,86 |
| 9 | 743 | 894 | 240 | 10,02 | 90,03 | 43 | 3,12 | 97,98 |
| 10 | 894 | 999 | 234 | 9,77 | 99,8 | 28 | 2,03 | 100,01 |

| **Decil** | **De** | **Hasta** | **#Cont** | **%Cont** | **Acum_Cont** | **%NC_D** | **%C_D** | **C:NC** |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 127 | 9 | 0,89 | 0,89 | 0,96 | 0,04 | 0:1 |
| 2 | 127 | 169 | 24 | 2,37 | 3,26 | 0,9 | 0,1 | 0:1 |
| 3 | 169 | 197 | 33 | 3,26 | 6,52 | 0,87 | 0,13 | 0:1 |
| 4 | 197 | 239 | 38 | 3,76 | 10,28 | 0,84 | 0,16 | 0:1 |
| 5 | 239 | 299 | 73 | 7,22 | 17,5 | 0,69 | 0,31 | 0:1 |
| 6 | 299 | 409 | 104 | 10,29 | 27,79 | 0,56 | 0,44 | 1:1 |
| 7 | 409 | 569,5 | 149 | 14,74 | 42,53 | 0,37 | 0,63 | 2:1 |
| 8 | 570 | 743 | 178 | 17,61 | 60,14 | 0,26 | 0,74 | 3:1 |
| 9 | 743 | 894 | 197 | 19,49 | 79,63 | 0,18 | 0,82 | 5:1 |
| 10 | 894 | 999 | 206 | 20,38 | 100,01 | 0,12 | 0,88 | 7:1 |

En la tabla 4.3 se observa que el porcentaje de individuos contactados y la razón de individuos contactados vs los individuos no contactados aumentan a con la probabilidad de contacto efectivo y que el porcentaje de individuos no contactados tiende a decrecer a medida que la probabilidad aumenta como se esperaba.

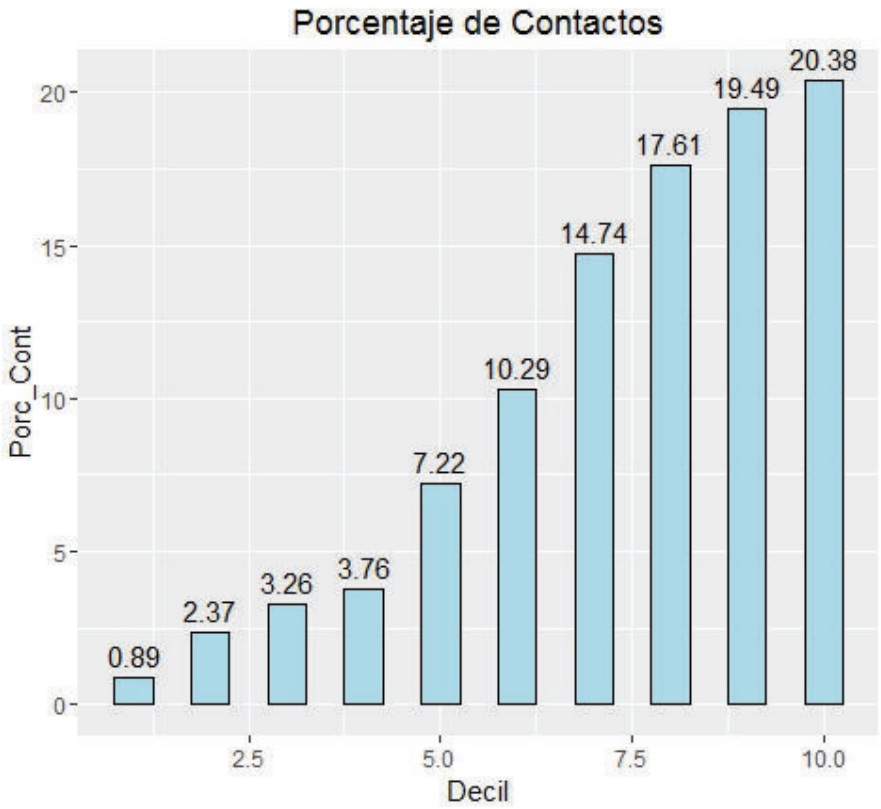


**Figura 4.5:** Porcentaje Contactados

En la figura 4.5 se observa detalladamente el crecimiento del porcentaje de individuos contactados por cada decil.

# Capítulo 5

# Conclusiones y Recomendaciones

- El mejor horario para contactar a un individuo puede ser modelado mediante herramientas matemáticas como la regresión logística multinomial basándose en información histórica de comportamiento y de gestión. Modelos de este tipo tendrán mejores características si la información con la que son alimentados es oportuna, completa y veraz.

- La regresión logística multinomial se caracteriza por realizar realizar k-1 regresiones logísticas binarias, donde k es el número de categorías de la variable dependiente, debido a esto, el modelo BTTC obtiene 4 regresiones logísticas con la categoría NC no contactar en ningún horario como categoría fija o pivote para cada regresión logística binaria. Para de obtener el menor número de variables que mejor expliquen el contacto efectivo en cada horario se analizó su significancia para cada horario de tal forma que en todos los horarios la mayor cantidad de variables sean significativas. De las tablas 3.7, 3.8, 3.9 y 3.10 y en base a este criterio, el horario 7-9am y el de 16-21pm son los mejores horarios, pues en estos horarios todas las variables son significativas.

- Tomando en cuenta las medidas de KS, AUROC y GINI de la tabla 3.26 el mejor horario, en términos de modelamiento es el horario de 7-9am seguido por el horario de 13-16pm, sin embargo el segundo tiene nos variables no significativas en el modelo (tabla 3.9). Por tanto el mejor modelo es en el horario de 7-9am.

- Al tomar la maxima probabilidad estimada de horario para cada individuo las medidas KSM y VUS son relativamente bajas (tabla 3.27) lo que indicaría que el modelo no es muy bueno para discriminar individuos contactados entre horarios de individuos no contactados en ningún horario, es decir que para establecer si a un individuo es mas probable contactarlo de 7-9am o de 9-13pm, se debe hacer un modelo entre estas dos categorías que las discrimine, por tanto si se tienen 4

categorías serían necesarios 6 modelos logísticos binarios. Variables que no fueron tomadas en cuenta por falta de confiabilidad y por exceso de datos perdidos como estado civil, profesión, sector de residencia y variables de alta importancia que no fue posible incluir por ausencia de información como números de teléfonos activos, si el teléfono es de casa, celular u oficina, indicador si el cliente realizó una llamada al call center, tiempos de llamadas podrían aumentar el poder de discriminación del modelo entre horarios.

- Para obtener mayor flexibilidad con las variables del modelo final, se recomienda tratar las k-1 regresiones logísticas de manera independiente, de este modo se pueden tratar las variables de manera más específica apoyándose en árboles de decisión donde se pueden encontrar las interacciones entre las variables que determinen el contacto efectivo en un horario determinado y así obtener variables diferentes en cada modelo logístico binomial. Esta metodología llevaría a construir otro tipo de modelo multinomial que podría dar mejores resultados de discriminación.

- Al sumar todas las probabilidades estimadas de las categorías de horario el resultado es la probabilidad estimada de contactar en cualquier horario, de esta forma el modelo multinomial se transforma un un modelo binomial que discrimina los individuos que probablemente serán contactados de los que probablemente no serán contactados. Este modelo resulta altamente discriminativo y de excelente calidad de discriminación por sus medidas de KS, AUC y GINI que son altas. El uso de este modelo ayuda a la optimización en la gestión telefónica dando prioridad a los individuos con mayor probabilidad de contacto efectivo (luego de realizada la segmentación de cartera) reduciendo costos y fuerza de trabajo.

- Dadas las probabilidades estimadas por el modelo para contactar a los individuos en los distintos horarios en el día, se deben implementar las estrategias de cobranza y las campañas en función de estas probabilidades para obtener resultados en la optimización del proceso de gestión telefónica.

- Se recomienda realizar continuamente una actualización de los números de teléfono de los individuos en los portafolios y para los casos en los que se tuviese varios números telefónicos para un mismo individuo, establecer el teléfono prioridad. Esto puede realizarse estableciendo un puntaje para cada teléfono, de tal forma que un puntaje alto indique mayor posibilidad de contactar a ese teléfono.

- Se recomienda la creación de un modelo de score comportamiento o de cobranza que permita separar los clientes que son regulares en sus pagos de los clientes que tienen problemas con cumplir sus obligaciones, así el modelo de BTTC se

centrará en la gestión de los clientes que tengan mayor probabilidad de entrar en default, así se logrará una optimización de todo el sistema de cobro.

# Anexos

# Anexo A

# Análisis Contacto efectivo y llamadas por hora

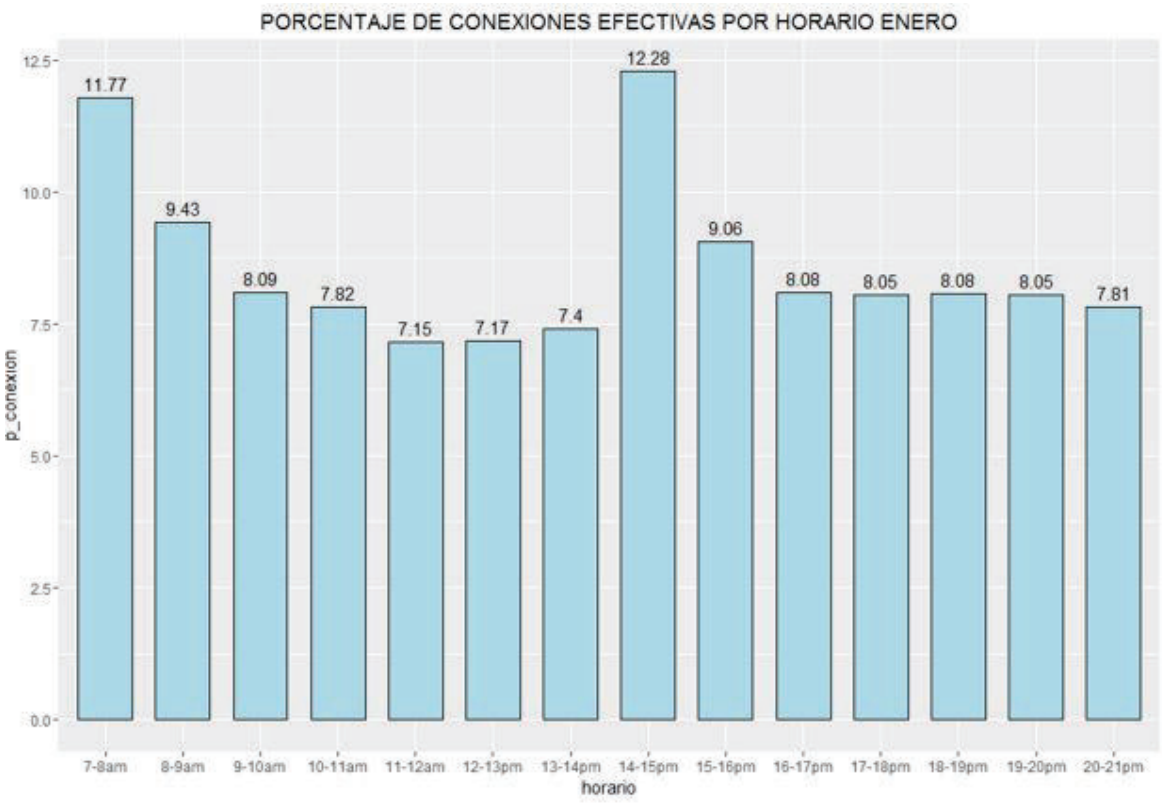


**Figura A.1:** Conexión Efectiva Enero

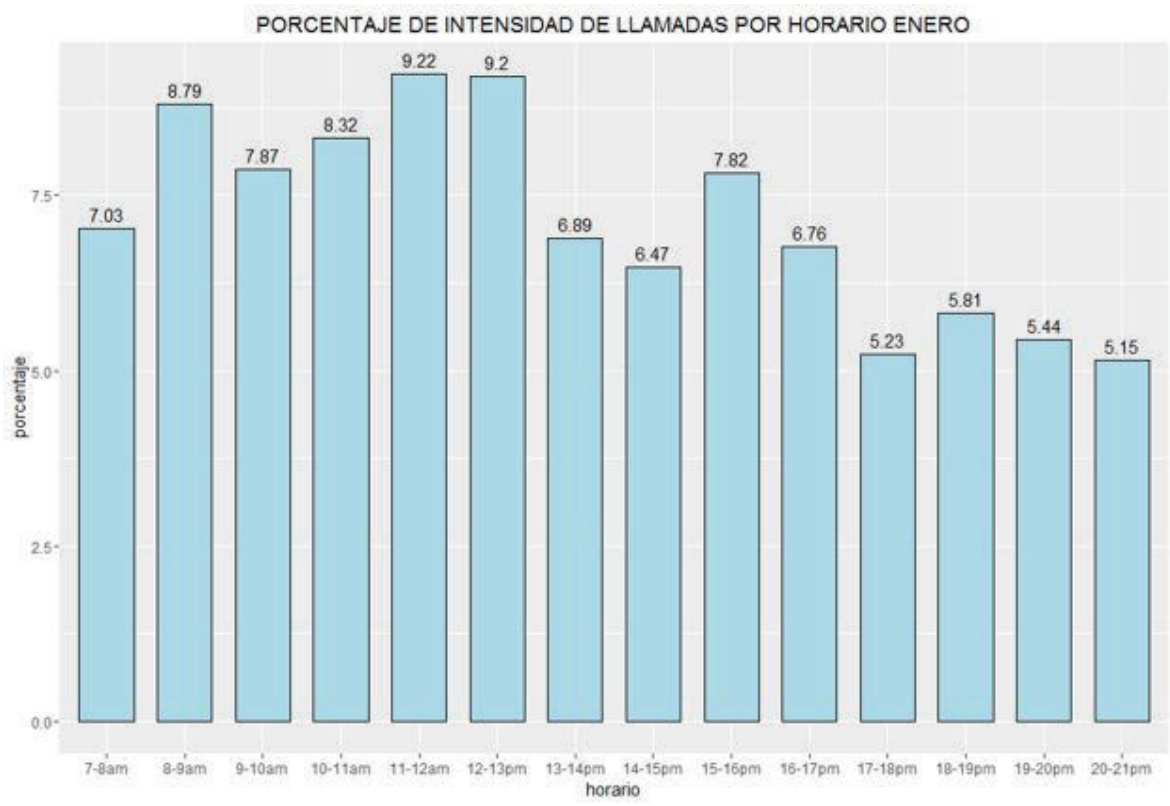


**Figura A.2:** Llamadas Realizadas Enero

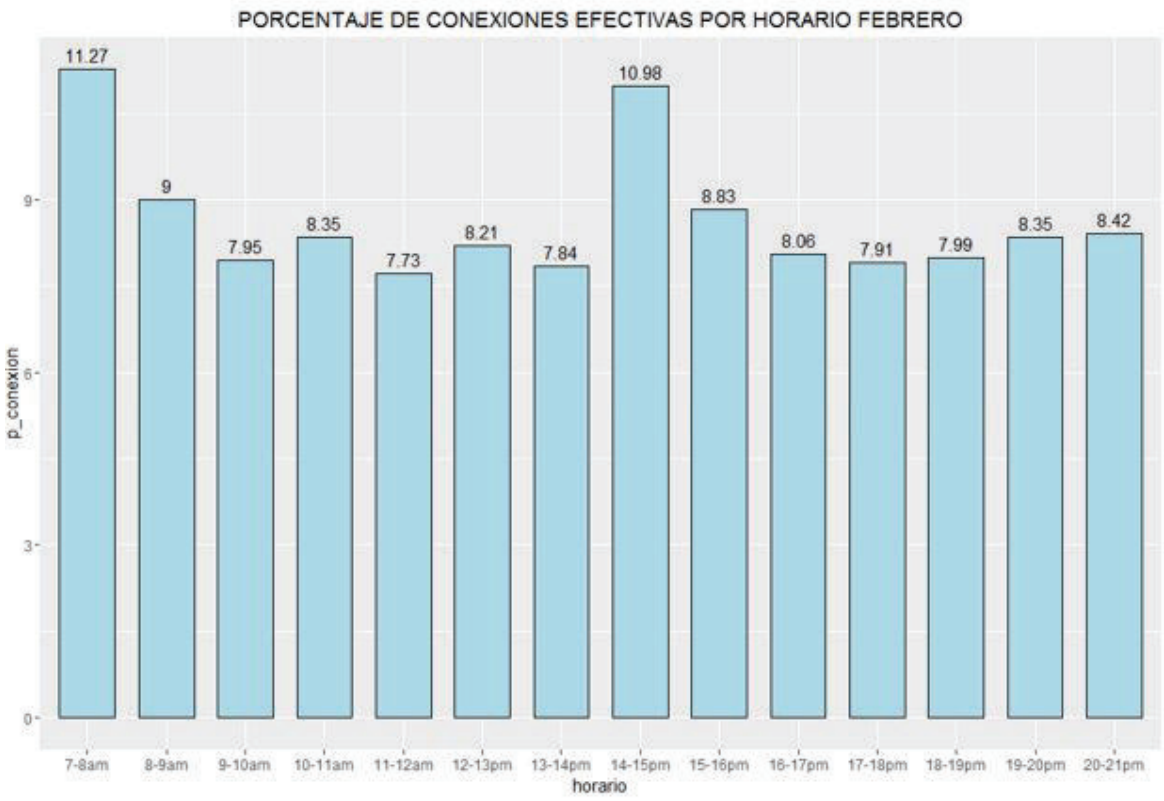


**Figura A.3:** Conexión Efectiva Febrero

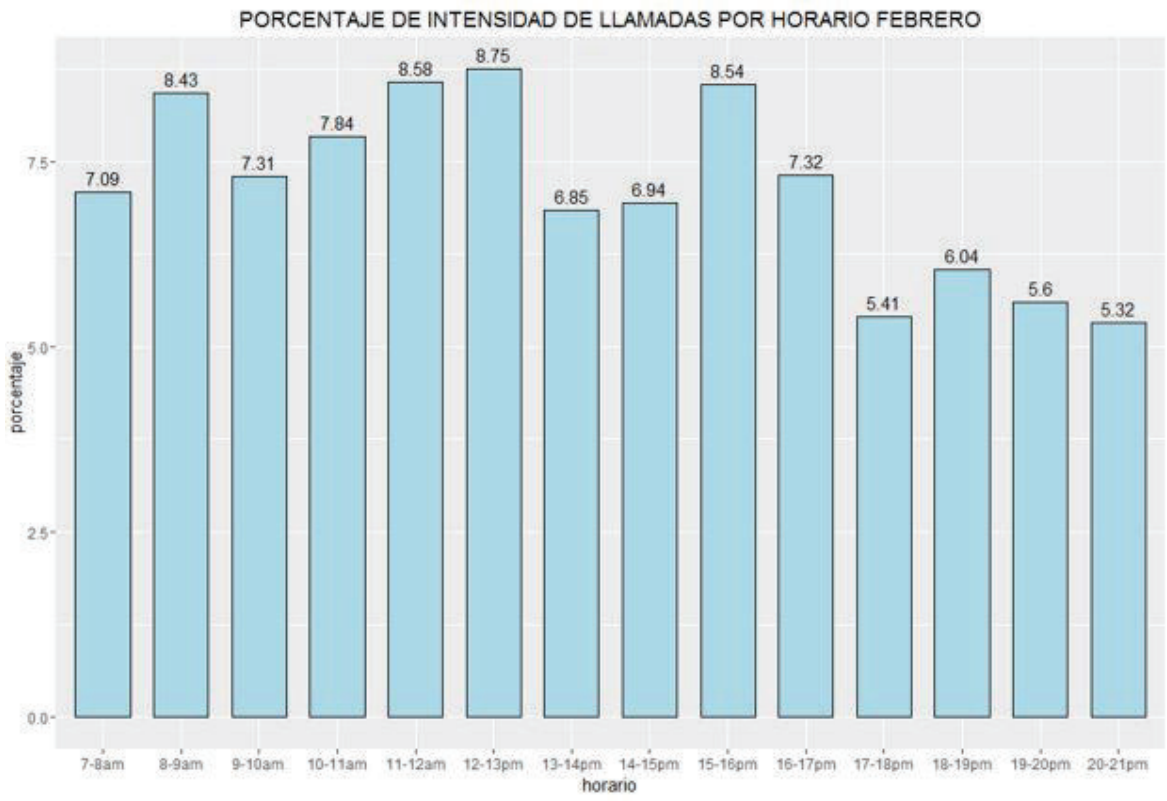


**Figura A.4:** Llamadas Realizadas Febrero

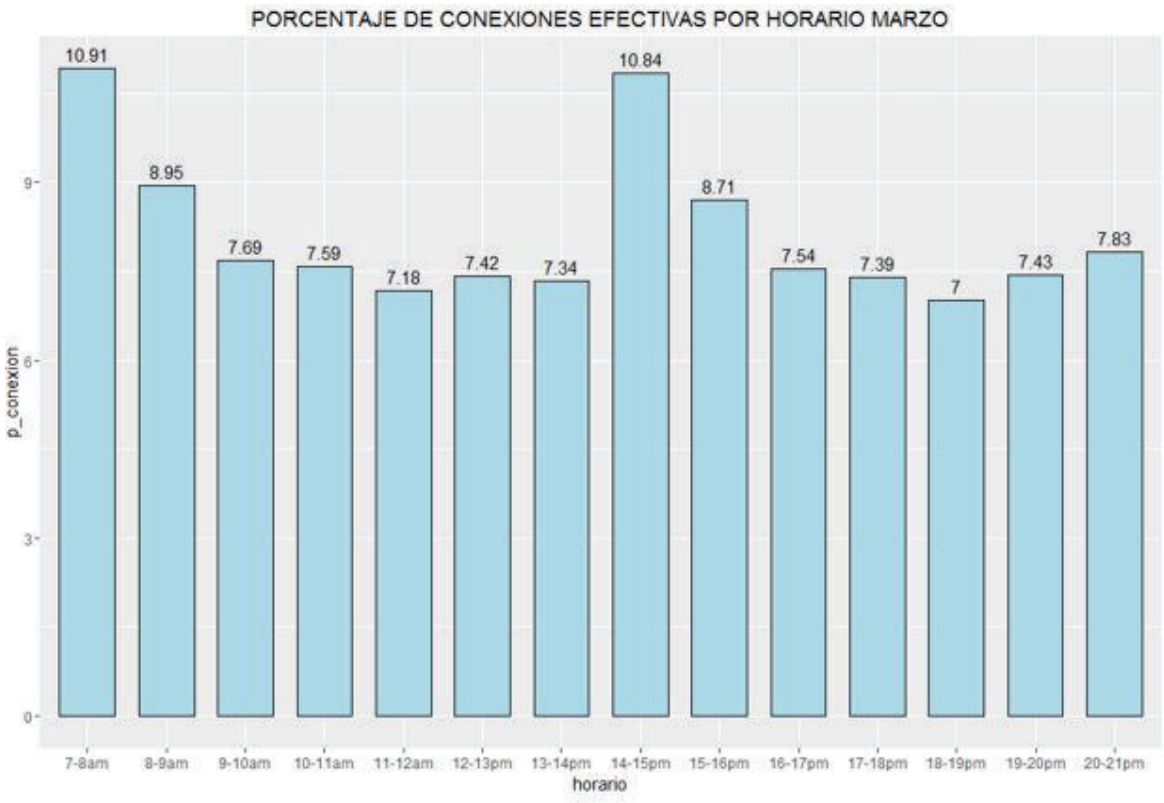


**Figura A.5:** Conexión Efectiva Marzo

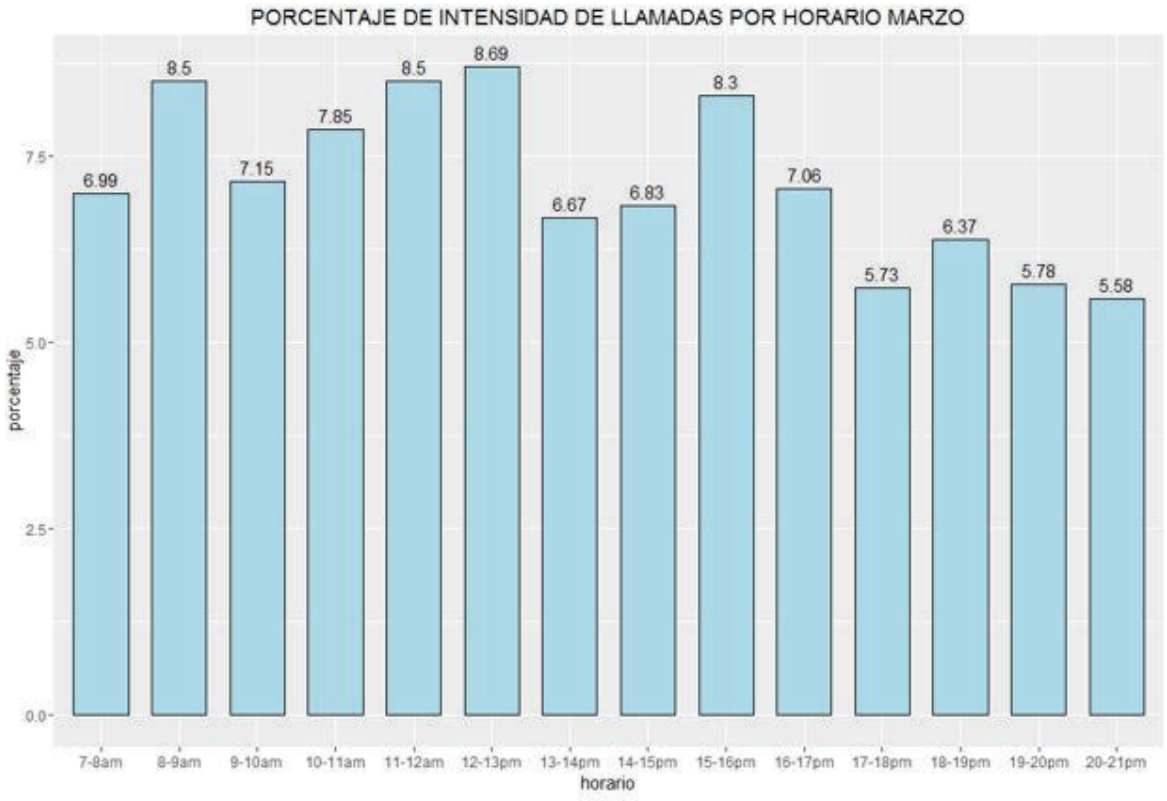


**Figura A.6:** Llamadas Realizadas Marzo

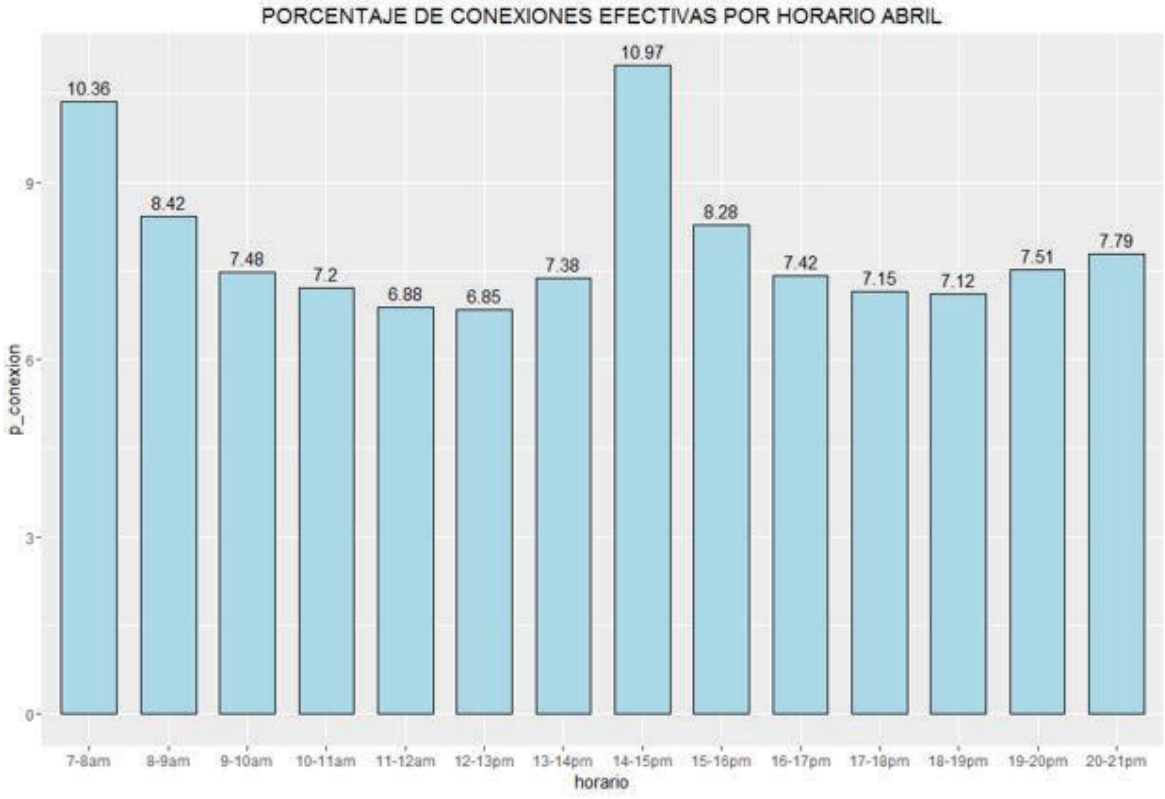


**Figura A.7:** Conexión Efectiva Abril

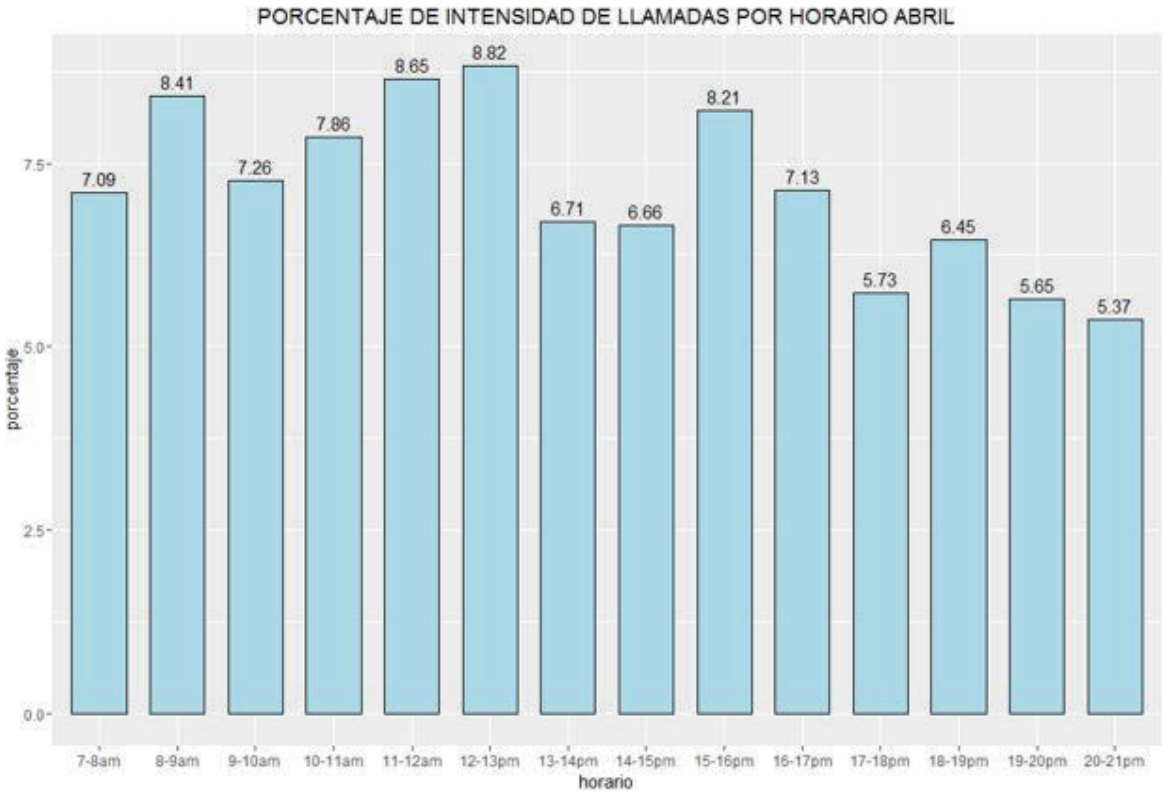


**Figura A.8:** Llamadas Realizadas Abril

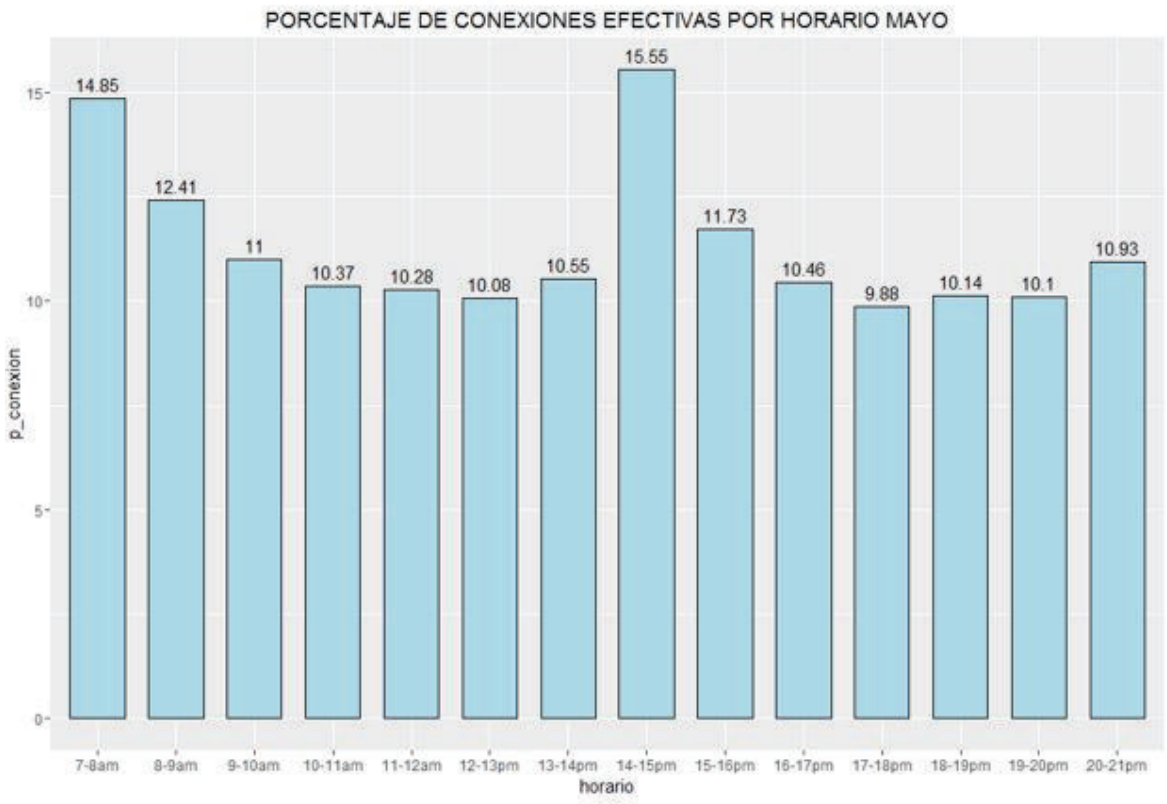


**Figura A.9:** Conexión Efectiva Mayo

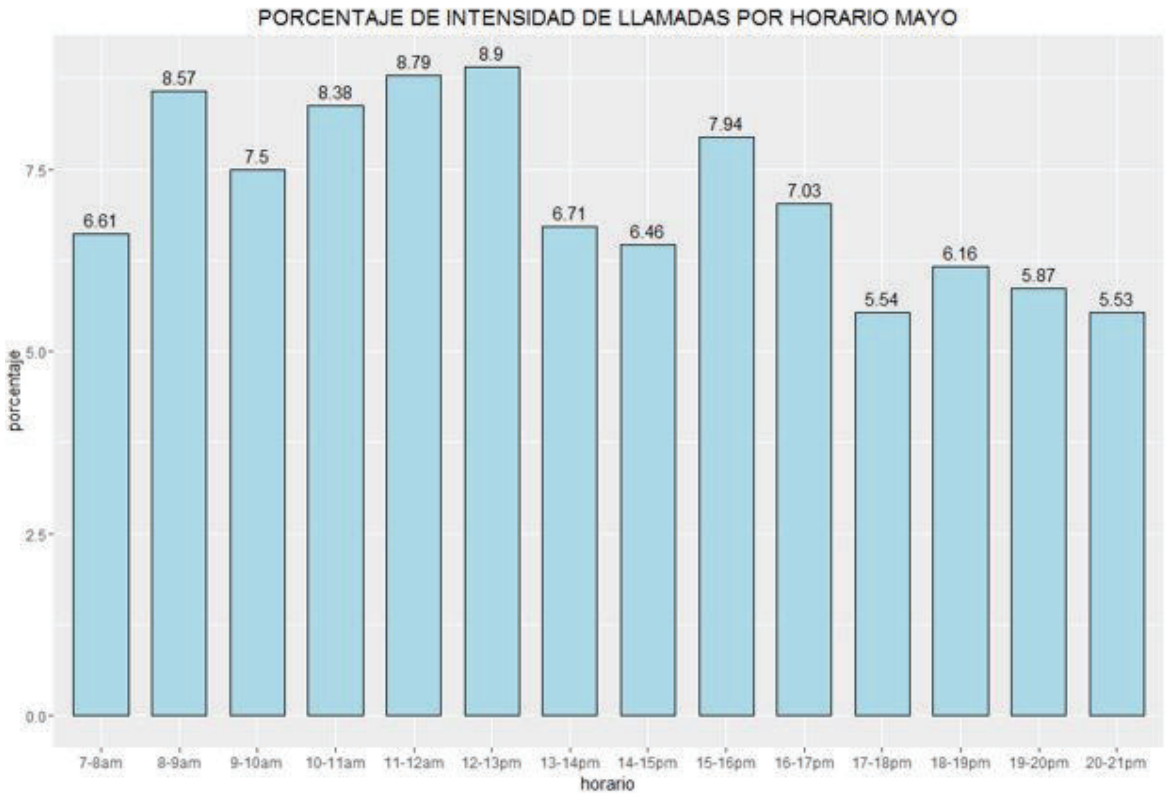


**Figura A.10:** Llamadas Realizadas Mayo

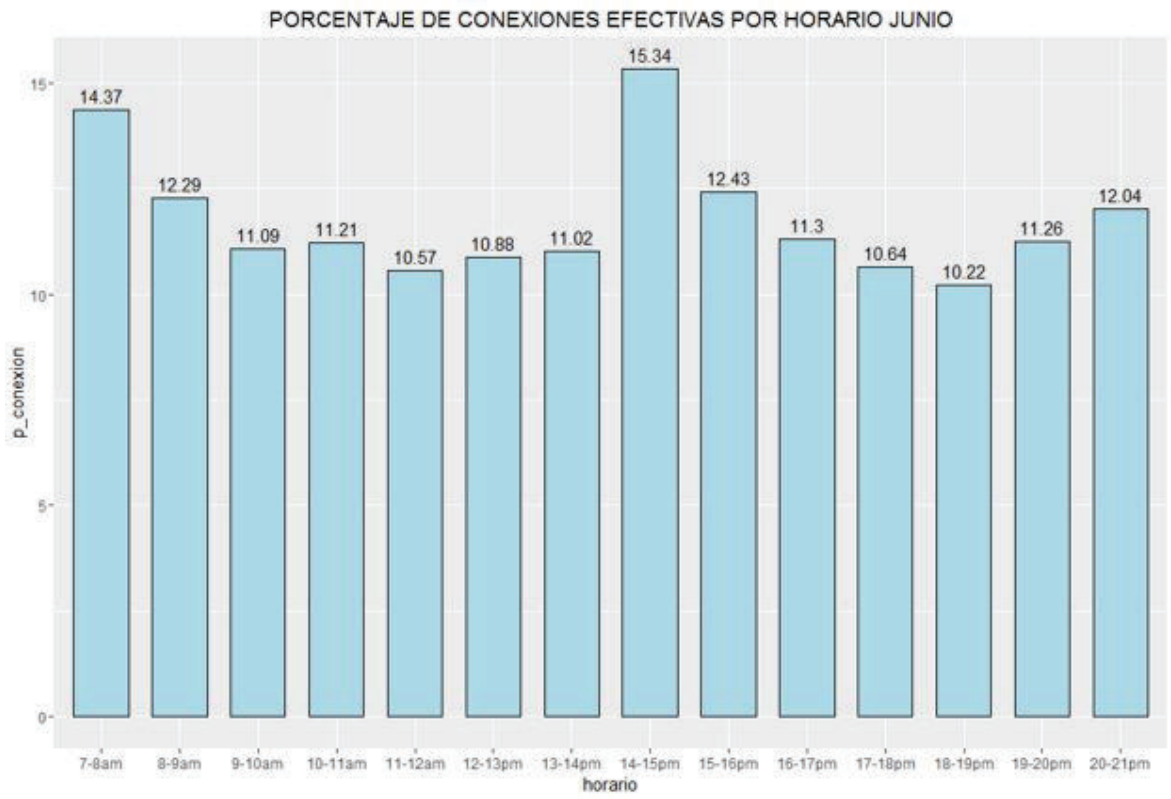


**Figura A.11:** Conexión Efectiva Junio

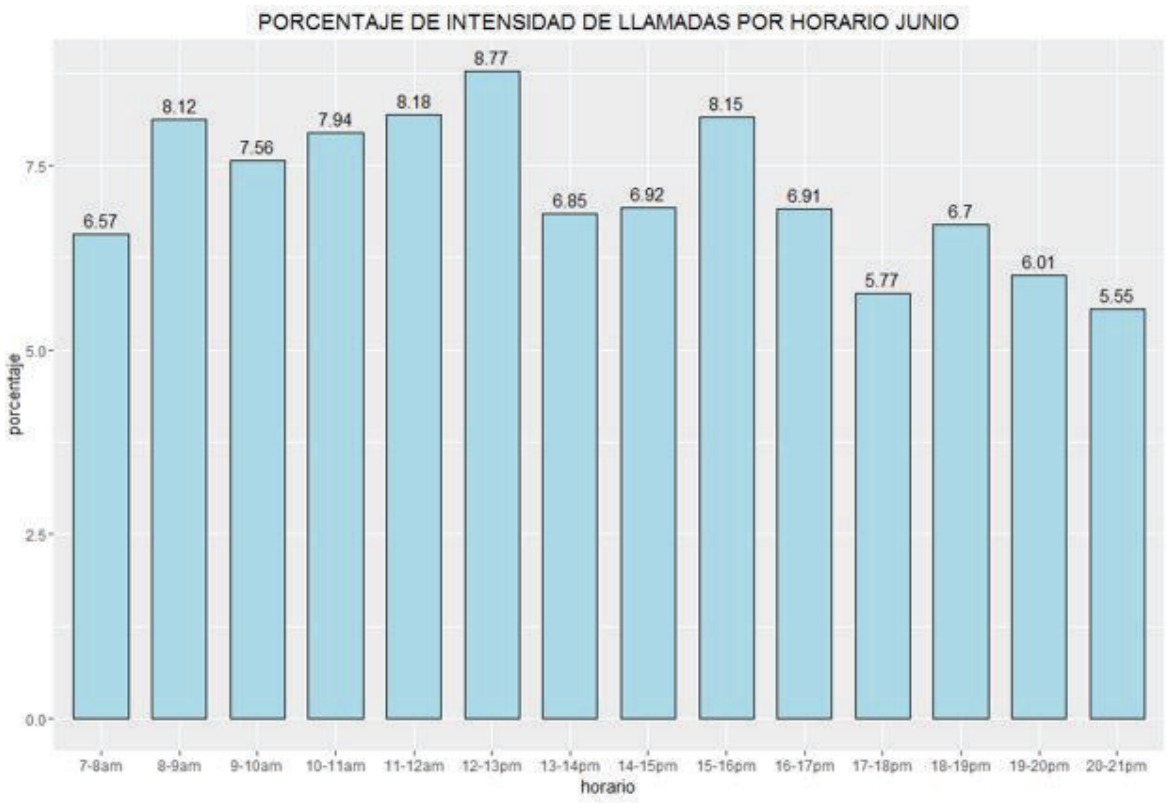


**Figura A.12:** Llamadas Realizadas Junio

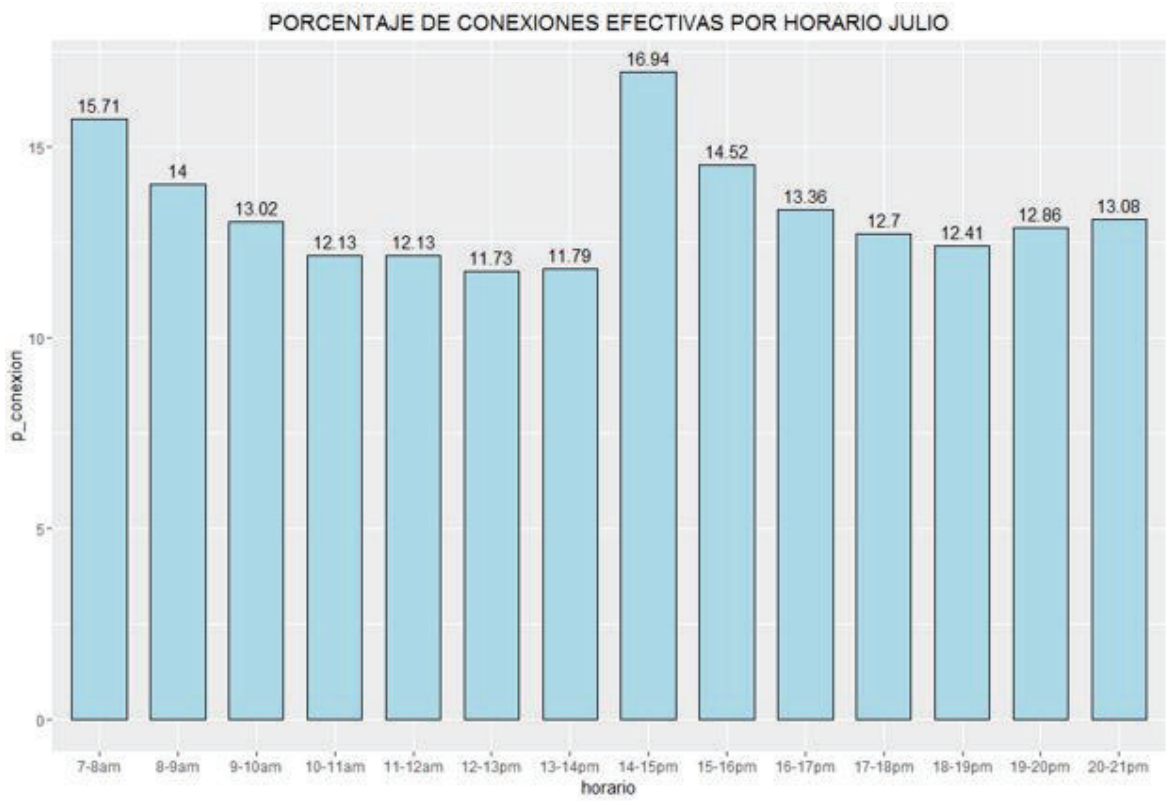


**Figura A.13:** Conexión Efectiva Julio

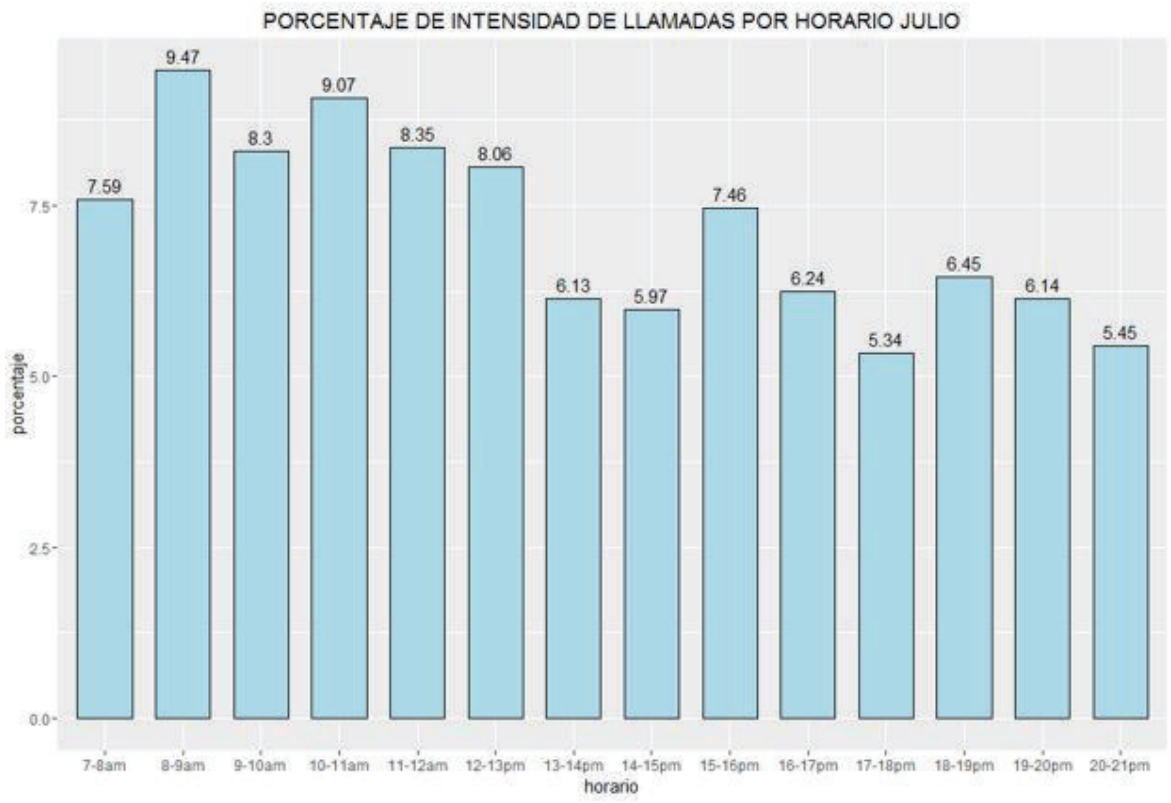


**Figura A.14:** Llamadas Realizadas Julio

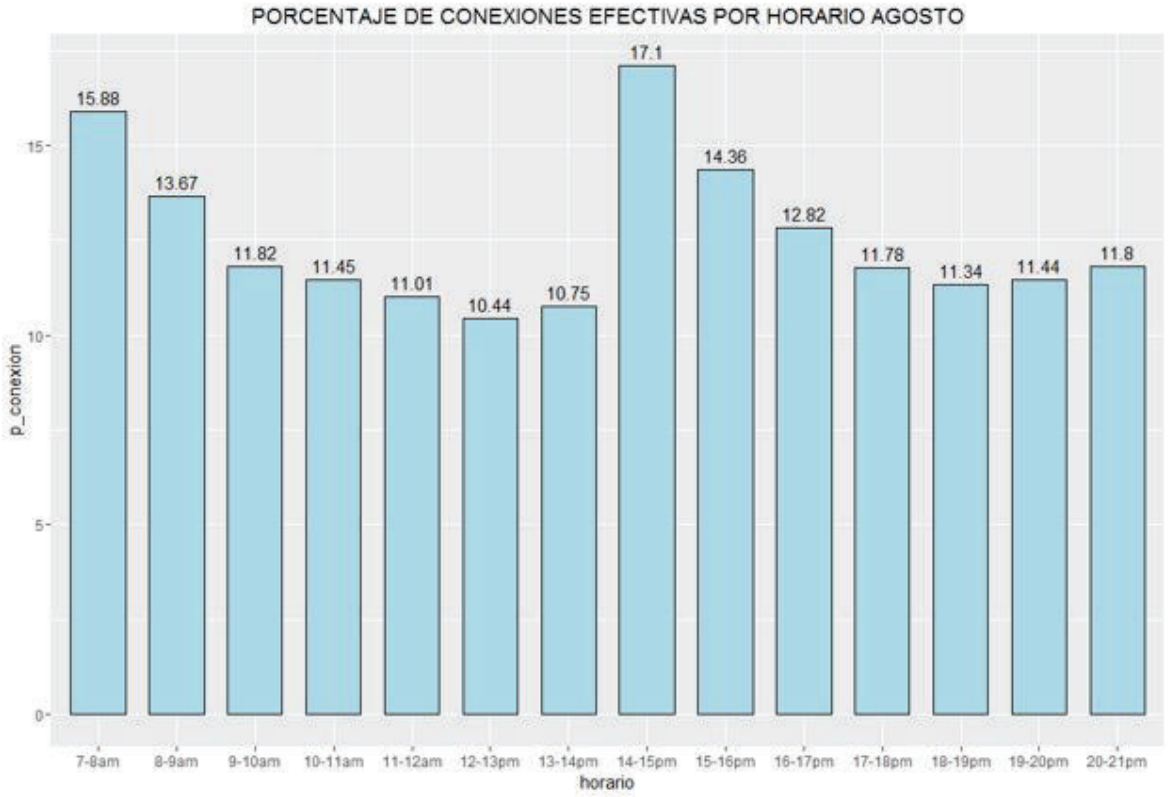


**Figura A.15:** Conexión Efectiva Agosto

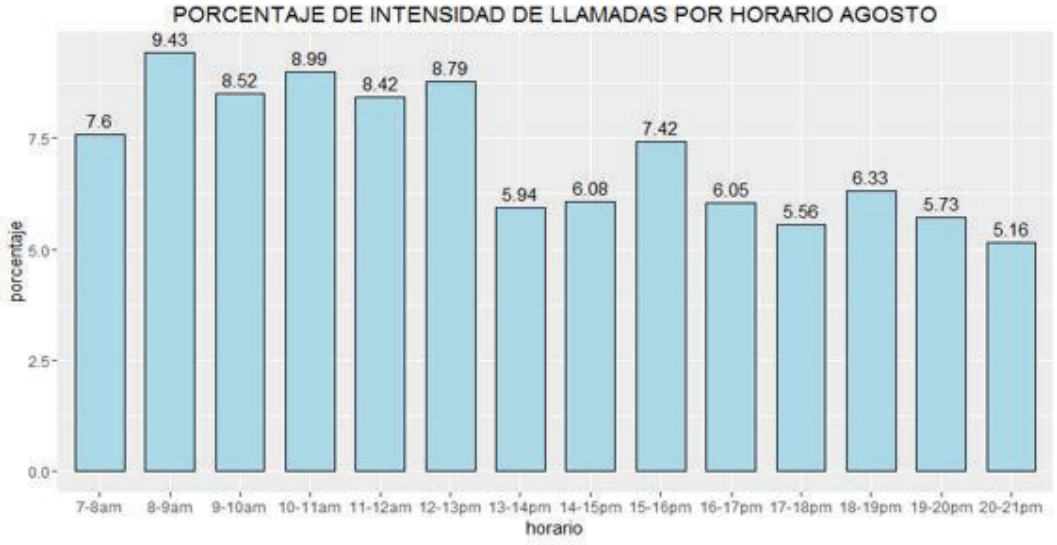


**Figura A.16:** Llamadas Realizadas Agosto

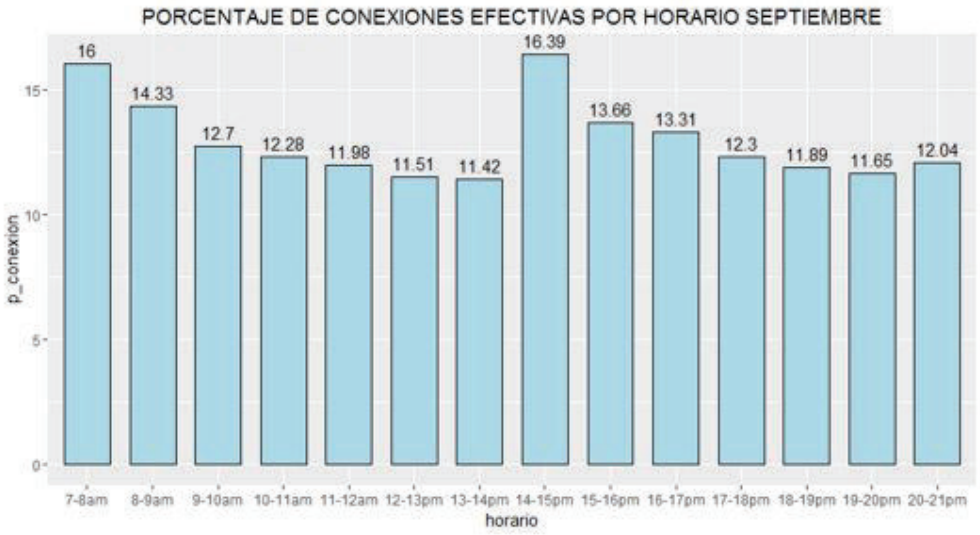


**Figura A.17:** Conexión Efectiva Septiembre

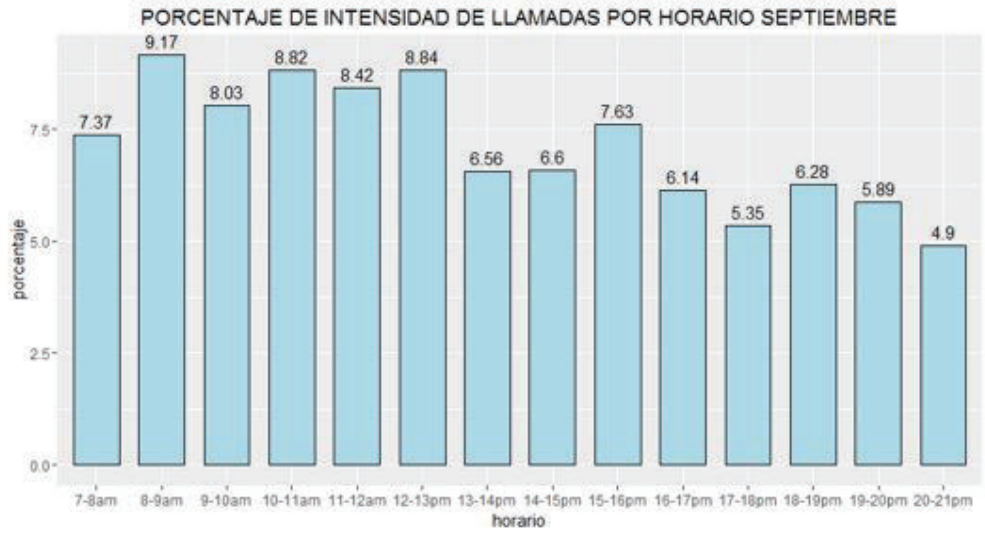


**Figura A.18:** Llamadas Realizadas Septiembre

# Anexo B

# Medidas de Asociación y Divergencia por variable

**Tabla B.1:** Coeficiente de Contingencia de Pearson

| No | Variable | Ji-Cuadrado |
|---|---|---|
| 1 | DIAGESTION | 0.1721 |
| 2 | PRODUCTO | 0.1324 |
| 3 | EsDependiente | 0.0983 |
| 4 | tipoDispositivo | 0.0627 |
| 5 | CIUDAD | 0.0589 |

**Tabla B.2:** KSM por variable

| No | Variable | KSM |
|---|---|---|
| 1 | rnum_conex_2M_6M | 0.375262077 |
| 2 | rsaldo_inicial_2M_4M | 0.335497107 |
| 3 | rnum_conex_2M_4M | 0.316031482 |
| 4 | rsaldo_inicial_2M_6M | 0.314919026 |
| 5 | refx_rec_4M_6M | 0.312047029 |
| 6 | ln_max_saldo_inicial_6M | 0.306029212 |
| 7 | ln_max_saldo_inicial_4M | 0.30486081 |
| 8 | porc_conex_2M | 0.303252057 |
| 9 | refx_rec_2M_4M | 0.300963414 |
| 10 | prom_porc_conex_4M | 0.299587655 |
| 11 | refxr_efr | 0.296230974 |
| 12 | ln_min_saldo_inicial_2M | 0.295923683 |
| 13 | ln_prom_saldoini_2M | 0.295846068 |
| 14 | rnum_conex_4M_6M | 0.294523467 |
| 15 | prom_porc_conex_2M | 0.292444093 |
| 16 | ln_max_efx_rec_2M | 0.291969608 |
| 17 | ln_max_saldo_inicial_2M | 0.289178964 |
| 18 | AtrasoPromSituacional | 0.288731777 |
| 19 | ln_prom_saldoini_6M | 0.287599169 |
| 20 | ln_prom_saldoini_4M | 0.286380319 |
| 21 | rsaldo_inicial_4M_6M | 0.286066335 |
| 22 | ln_prom_efx_rec_2M | 0.284231319 |
| 23 | max_porc_conex_2M | 0.283107835 |
| 24 | max_porc_conex_6M | 0.280374289 |
| 25 | SaldoInicial | 0.279435729 |
| 26 | ln_min_saldo_inicial_6M | 0.279335681 |
| 27 | ln_max_efx_rec_4M | 0.277563407 |
| 28 | prom_porc_conex_6M | 0.27223834 |
| 29 | min_días_mora_2M | 0.270019846 |
| 30 | ln_min_saldo_inicial_4M | 0.269472927 |
| 31 | EfectivoRecuperado | 0.268842355 |
| 32 | porc_conex_6M | 0.26778888 |
| 33 | SaldoDeudasSF | 0.267333667 |
| 34 | ln_max_efx_rec_6M | 0.265816974 |
| 35 | rdias_mora_2M_4M | 0.264603246 |
| 36 | AtrasoMaxSituacional | 0.258968111 |
| 37 | porc_conex_4M | 0.258914997 |
| 38 | refx_ef_2M | 0.258625344 |
| 39 | p_conex | 0.257197295 |
| 40 | rdias_mora_4M_6M | 0.254866702 |

| No | Variable | KSM |
|---|---|---|
| 41 | ln_prom_efx_rec_6M | 0.254406106 |
| 42 | refx_ef_4M | 0.252216856 |
| 43 | refr_sueldo | 0.249729571 |
| 44 | ln_min_efx_rec_4M | 0.249720826 |
| 45 | ci_4M | 0.24716889 |
| 46 | prom_ci_4M | 0.24716889 |
| 47 | prom_cp_4M | 0.24716889 |
| 48 | EfectivoXRecuperar | 0.244353448 |
| 49 | num_conex_2M | 0.242779653 |
| 50 | prom_conex_2M | 0.242779653 |
| 51 | min_porc_conex_2M | 0.242370956 |
| 52 | ln_prom_efx_rec_4M | 0.242293249 |
| 53 | max_días_mora_6M | 0.239495184 |
| 54 | prom_días_mora_6M | 0.239186231 |
| 55 | max_porc_conex_4M | 0.237966219 |
| 56 | ci_6M | 0.23617799 |
| 57 | prom_ci_6M | 0.23617799 |
| 58 | prom_cp_6M | 0.23617799 |
| 59 | DiasMora | 0.234725207 |
| 60 | max_num_conex_2M | 0.234310503 |
| 61 | Sueldo | 0.232632095 |
| 62 | rci_cp_4M | 0.230585291 |
| 63 | SaldoDeudasBS | 0.230368316 |
| 64 | max_días_mora_4M | 0.23023902 |
| 65 | ref_rec_2M_4M | 0.225389381 |
| 66 | prom_pagos_4M | 0.225264764 |
| 67 | prom_días_mora_4M | 0.225118526 |
| 68 | rci_cp | 0.223415514 |
| 69 | CuotasIniciales | 0.22308872 |
| 70 | refx_ef_6M | 0.222709974 |
| 71 | ln_max_ef_rec_4M | 0.222514471 |
| 72 | max_días_mora_2M | 0.221938876 |
| 73 | refx_rec_2M_6M | 0.221121858 |
| 74 | ln_prom_ef_rec_6M | 0.221071389 |
| 75 | ref_rec_2M_6M | 0.218053934 |
| 76 | ln_max_ef_rec_6M | 0.217503508 |
| 77 | max_ci_4M | 0.215314724 |
| 78 | CantidadDeudasSF | 0.212024296 |
| 79 | rdias_mora_2M_6M | 0.210851933 |
| 80 | rci_cp_6M | 0.210778507 |
| 81 | Edad | 0.210768134 |
| 82 | ln_max_ef_rec_2M | 0.208699434 |
| 83 | prom_días_mora_2M | 0.208289338 |
| 84 | ci_2M | 0.206533701 |
| 85 | prom_ci_2M | 0.206533701 |
| 86 | prom_cp_2M | 0.206533701 |

| No | Variable | KSM |
|---|---|---|
| 87 | num_conex_4M | 0.205587256 |
| 88 | prom_conex_4M | 0.205587256 |
| 89 | max_ci_6M | 0.200634787 |
| 90 | ref_rec_4M_6M | 0.200053137 |
| 91 | prom_pagos_6M | 0.198897511 |
| 92 | num_conex | 0.196041774 |
| 93 | ln_prom_ef_rec_4M | 0.195784204 |
| 94 | rci_cp_2M | 0.195488256 |
| 95 | max_ci_2M | 0.193878455 |
| 96 | ln_prom_ef_rec_2M | 0.193441207 |
| 97 | min_ci_2M | 0.1868967 |
| 98 | min_días_mora_6M | 0.169241755 |
| 99 | max_num_conex_4M | 0.166139914 |
| 100 | min_num_conex_2M | 0.163560795 |
| 101 | CargasFamiliares | 0.160899403 |
| 102 | ln_min_efx_rec_2M | 0.158115444 |
| 103 | num_conex_6M | 0.15766635 |
| 104 | prom_conex_6M | 0.15766635 |
| 105 | ln_min_ef_rec_2M | 0.150465425 |
| 106 | max_cp_6M | 0.149740114 |
| 107 | prom_pagos_2M | 0.149474038 |
| 108 | max_num_conex_6M | 0.147509262 |
| 109 | min_días_mora_4M | 0.142164485 |
| 110 | ln_min_efx_rec_6M | 0.14124194 |
| 111 | EsPagado | 0.13920622 |
| 112 | min_porc_conex_6M | 0.133367519 |
| 113 | max_pagos_2M | 0.127379716 |
| 114 | min_porc_conex_4M | 0.123062776 |
| 115 | CantidadDeudasBS | 0.116757954 |
| 116 | max_pagos_4M | 0.112394965 |
| 117 | max_cp_4M | 0.110122312 |
| 118 | min_pagos_4M | 0.104237892 |
| 119 | min_pagos_2M | 0.103135655 |
| 120 | ln_min_ef_rec_4M | 0.100428692 |
| 121 | min_ci_4M | 0.091160676 |
| 122 | min_num_conex_6M | 0.090951099 |
| 123 | max_pagos_6M | 0.090809981 |
| 124 | min_num_conex_4M | 0.077190016 |
| 125 | max_cp_2M | 0.069872945 |
| 126 | min_ci_6M | 0.069452512 |
| 127 | min_pagos_6M | 0.067205339 |
| 128 | min_cp_2M | 0.051018513 |
| 129 | CuotasPagadas | 0.050085712 |
| 130 | ln_min_ef_rec_6M | 0.040042148 |
| 131 | VivSaldoDeudas | 0.036398024 |
| 132 | VivCantidadDeudas | 0.029860539 |
| 133 | min_cp_4M | 0.021005814 |
| 134 | min_cp_6M | 0.009277446 |

# Anexo C

# Códigos implementados en R para el modelo de BTTC

## C.1. Test Kolmogorov-Smirnoff

```
###Test ks para dos muestras
KStest=function(x,y){

  if(length(unique(y))>2){
    print("Response is not bivariate")
  }

  SampA=x[y==unique(y)[1]]
  SampB=x[y==unique(y)[2]]

  stat=ks.test(SampA,SampB)$statistic

  return(ks.stat=stat)

}

###Test ks para mas de dos muestras
KSmeasure=function(X,y,cutoff=NA,plot=TRUE){

k=length(levels(as.factor(y)))
groups=levels(as.factor(y))
n=nrow(X)
```

```
p=ncol(X)

KS=rep(0,ncol(X))

for(i in 1:(k-1)){
for(j in (i+1):k){

temp=X[y==groups[i]|y==groups[j],]
tempy=y[y==groups[i]|y==groups[j]]

KStemp=apply(temp,2,KStest,y=tempy)

KS=KS+KStemp*(sum(y==groups[i]|y==groups[j]))/(n)
}}

if(plot==TRUE){

plot((KS[order(-KS)])[1:min(ncol(X),nrow(X))],type="l",xlab="Variables"
     ,ylab="K-S Measure Valor")
axis(side=1,at=seq(5,min(ncol(X),nrow(X)),by=5),labels=F)
abline(v=cutoff,col="red")

}
return(KS[order(-KS)])
}
```

## C.2. Coeficiente de contingencia de Pearson

```
jicuad<-function(var_dep,var)
{
  chisq<-chisq.test(var_dep,var)
  chisq<-as.numeric(chisq$statistic)
  ccp<-sqrt(chisq/(chisq+length(var)))
  ccp
}


test<-numeric()
for(i in 1:ncol(X))
```

```
{
  ind<-c(names(X)[i],round(jicuad(ml$ses,X[[i]]),4))
  test<-rbind(test,ind)
  test<-test[order(test[,2],decreasing=TRUE),]
  rownames(test)<-NULL
}
```

## C.3. Valor de Información (IV)

```
library(Information)
library(woe)
library(InformationValue)

IVtest=function(x,y){

  if(length(unique(y))>2){
    print("Variable respuesta no es bivariada")
  }
  dat<-data.frame(x,y)
  dat$y2<-ifelse(y==unique(y)[1],0,1)


  stat<-WOETable(X=dat$x, Y=dat$y2)
  vi<-sum(stat$IV)
  vi
}
#IVtest(x,y2)

VImeasure<-function(X,y,cutoff=NA,plot=TRUE){
k=length(levels(as.factor(y)))
groups=levels(as.factor(y))
n=nrow(X)
p=ncol(X)

VI=rep(0,ncol(X))

for(i in 1:(k-1)){
  for(j in (i+1):k){
```

```
    temp=X[y==groups[i]|y==groups[j],]
    tempy=y[y==groups[i]|y==groups[j]]

    VItemp=apply(temp,2,IVtest,y=tempy)

    VI=VI+VItemp*(sum(y==groups[i]|y==groups[j]))/(n)

  }}

if(plot==TRUE){

  plot((VI[order(-VI)])[1:min(ncol(X),nrow(X))],type="l",
  xlab="Nth-Largest VI Measure Value",ylab="VI Measure Value")
  axis(side=1,at=seq(5,min(ncol(X),nrow(X)),by=5),labels=F)
  abline(v=cutoff,col="red")
  }

return(VI[order(-VI)])
}
```

## C.4. Filtrado de Variables

```
library(dplyr, quietly = TRUE, warn.conflicts = FALSE)
library(ggplot2)
library(data.table)
#library(bit)
#library(bit64)
library(Information)
library(woe)
library(InformationValue)
options(scipen = 999)

BDD<-read.csv(file = "BDD_modelo.csv",sep = ",",dec=".", header = TRUE,
                stringsAsFactors=T)
BDD<-unique(BDD)
BDD$X<-NULL
colnames(BDD)
summary(BDD)
```

```
####VARIABLES CATEGORICAS###
BDD_cat<-BDD %>%select(NUMEROOPERACION, contacto ,PRODUCTO,CIUDAD,
                       tipoDispositivo ,DIAGESTION, EsDependiente)
#BDD_cat<-unique(BDD_cat)

write.csv(x = BDD_cat, file="BDD_cat.csv")

BDD_vi<-BDD_cat %>%select(-NUMEROOPERACION,-contacto)
contacto_vi<-BDD_cat$contacto

#Evaluacion IV#
VIM<-VImeasure(BDD_vi, contacto_vi, cutoff=NA, plot=TRUE)
names(VIM)
iv_value<-as.vector(VIM)
iv<-data.frame(Variable=names(VIM), iv_value)
View(iv)

#TABLAS DE CONTINGENCIA
t_contacto_etapa<-table(BDD_cat$contacto ,BDD_cat$Etapa)
prop.table(t_contacto_etapa)

t_contacto_ciudad<-table(BDD_cat$contacto ,BDD_cat$CIUDAD)
100*prop.table(t_contacto_ciudad ,2)

t_contacto_producto<-table(BDD_cat$contacto ,BDD_cat$PRODUCTO)
prop.table(t_contacto_producto)

t_contacto_disposit<-table(BDD_cat$contacto ,BDD_cat$tipoDispositivo)
100*prop.table(t_contacto_disposit ,2)

t_contacto_dia<-table(BDD_cat$contacto ,BDD_cat$DIAGESTION)
100*prop.table(t_contacto_dia ,2)

t_contacto_ecivil<-table(BDD_cat$contacto ,BDD_cat$EstadoCivil)
prop.table(t_contacto_ecivil ,2)

t_contacto_profesion<-table(BDD_cat$contacto ,BDD_cat$profesion)
round(prop.table(t_contacto_profesion ,2) ,2)
```

```
t_contacto_genero<-table(BDD_cat$contacto,BDD_cat$genero)
prop.table(t_contacto_genero,2)

t_contacto_depen<-table(BDD_cat$contacto,BDD_cat$EsDependiente)
prop.table(t_contacto_depen,2)

#Test chicudrado: CCP
chi<-numeric()
for(i in 1:ncol(BDD_vi))
{
  ind<-c(names(BDD_vi)[i],round(jicuad(contacto_vi,BDD_vi[[i]]),4))
  chi<-rbind(chi,ind)
  chi<-chi[order(chi[,2],decreasing=TRUE),]
  rownames(chi)<-NULL
}

plot((chi[,2])[1:min(ncol(BDD_vi),nrow(BDD_vi))],type="l",
     xlab="Variables",ylab="Coeficiente Pearson")

axis(side=1,at=seq(2,min(ncol(BDD_vi),nrow(BDD_vi)),by=5),labels=F)
abline(v=4,col="red")
View(chi)
write.csv(chi,"chisq.csv")

####VARIABLES NUMERICAS CONTINUAS###
BDD_num<-BDD %>%select(-CIUDAD,-tipoDispositivo,-PRODUCTO,
                       -DIAGESTION,-EsDependiente)
BDD_ks_num<-BDD_num %>%select(-NUMEROOPERACION,-contacto,-NUMEROCLIENTE)
BDD_ks_num<-unique(BDD_ks_num)
contacto_ks_num<-BDD_num$contacto

###MEDIDA DEL KS###
KSM<-KSmeasure(BDD_ks_num,contacto_ks_num,cutoff=115,plot=TRUE)
names(KSM)
ks_value<-as.vector(KSM)
ks<-data.frame(Variable=names(KSM),ks_value)
View(ks)
write.csv(x=ks,file = "ks_measure.csv")
#plot(cumsum(ks$ks_value),type="l")
```

```
####ELECCION DE VARIABLES SEGUN CORRELACION Y MAXIMO KS
#instalar paquet library(corrp)
mcor_num_gen<-cor(BDD_ks_num)#matriz de correlación
                              #de variables numéricas

mcor_ord_ks<-data.frame(mcor_num_gen[match(as.character(ks$Variable),
colnames(mcor_num_gen)), match(as.character(ks$Variable),
colnames(mcor_num_gen))])#matriz de correlación ordena por ks de mayor a
                         #menor por filas y columnas

fc_cor<-which(abs(mcor_ord_ks)>0.7 & (row(mcor_ord_ks) >
                                      col(mcor_ord_ks)),arr.ind = T)

#rownames(fc_cor)[which(fc_cor[,2]==j)]
if(nrow(fc_cor)>0){
var_elim<-numeric()
for(i in seq(1:nrow(fc_cor))){
  #i=4
  aux_var_elim<-c(fc_cor[i,1],fc_cor[i,2])
  if (!any(var_elim %in% aux_var_elim)){
    var_elim[i] <-max(fc_cor[i,])
    }
  }
  if(length(var_elim)>0){
    var_elim <- unique(var_elim[var_elim>0])
          }
}
var_elim<-var_elim[!is.na(var_elim)]
View(as.data.frame(var_elim))
v<-names(mcor_ord_ks)[var_elim]
#eleccion de las mejores variables
#numéricas según ks y correlación de
#la base de datos inicial
BDD_final<-unique(select(BDD,-one_of(v)))
write.table(BDD_final,"BDD_final.txt",dec = ",",sep = "\t")
summary(BDD_final)
colnames(BDD_final)
```

## C.5. Modelo Multinomial de BTTC

```
library(nnet)
library(reshape2)
#install.packages("leaps",dependencies = T)
#install.packages("pROC",dependencies = T)
library(leaps)
library(foreign)
library(ggplot2)
library(pROC)
#BDD_sub_num: base de datos consolidada con las mejores variables
#mayor ks y corralacion <0.7
str(BDD_final)
#BDD_final$CIUDAD<-as.factor(BDD_final$CIUDAD)
#BDD_final$DIAGESTION<-as.factor(BDD_final$DIAGESTION)

######Variables dummy y probabilidad en base a árboles de decición#####
BDD_final$PRODUCTO<-ifelse(BDD_final$PRODUCTO=="TCF",1,0)
BDD_final$CIUDAD<-ifelse(BDD_final$CIUDAD=="COSTA",0,1)
BDD_final$tipoDispositivo<-ifelse(BDD_final$tipoDispositivo=="CEL",0,1)
BDD_final$DIAGESTION<-ifelse(BDD_final$DIAGESTION=="IS",0,1)
BDD_final$EsDependiente<-ifelse(BDD_final$EsDependiente=="SI",1,0)

BDD_final$CargasFamiliares<-ifelse(is.na(BDD_final$CargasFamiliares),0,
                                   BDD_final$CargasFamiliares)
BDD_final$AtrasoMaxSituacional<-
                      ifelse(is.na(BDD_final$AtrasoMaxSituacional),
                                   0,BDD_final$AtrasoMaxSituacional)
BDD_final$AtrasoPromSituacional<-
                      ifelse(is.na(BDD_final$AtrasoPromSituacional),
                                   0,BDD_final$AtrasoPromSituacional)

######MODELO MULTINOM#####
BDD_final<-na.omit(BDD_final)
round(0.6*nrow(BDD_final))
round(0.4*nrow(BDD_final))

set.seed(10)
data_model<-sample_n(BDD_final,round(0.6*nrow(BDD_final)))
```

```
100*prop.table(table(data_model$contacto))

set.seed(10)
data_valid<-sample_n(BDD_final,round(0.4*nrow(BDD_final)))
100*prop.table(table(data_valid$contacto))

data_model$contacto<-relevel(data_model$contacto,ref="NC")
m<-multinom(contacto ~.,data=data_model %>%select(-NUMEROOPERACION,
                                                  -NUMEROCLIENTE),maxit=1000)
summary(m)

backwards = step(m, direction = "both",trace=0)
formula(backwards)

m2<-multinom(contacto ~ 1,data=data_model %>%select(-NUMEROOPERACION,
                                                    -NUMEROCLIENTE),maxit=1000)

#########MULTINOM######

mbest<-multinom(contacto ~  num_conex+p_conex+DIAGESTION+PRODUCTO+
                  max_porc_conex_6M+num_conex_2M+min_num_conex_4M+
                  rsaldo_inicial_2M_6M,data=data_model,maxit=1000)
#coeftest(mbest)
var_model<-
as.data.frame(c("num_conex","p_conex","DIAGESTION","PRODUCTO",
"SaldoDeudasSF","CantidadDeudasBS","VivSaldoDeudas",
"EsDependiente","AtrasoMaxSituacional","AtrasoPromSituacional",
"CargasFamiliares","Edad","refr_sueldo","max_porc_conex_6M",
"num_conex_2M","min_num_conex_4M","rnum_conex_2M_6M",
"ln_max_saldo_inicial_6M","rsaldo_inicial_2M_6M","max_cp_2M",
"refx_rec_4M_6M","refx_ef_2M","min_pagos_6M"))

colnames(var_model)<-c("Variable")
ks$Variable<-as.character(ks$Variable)
var_ks<-arrange(merge(ks,var_model),ks_value)
#Se retiran variables no significativas por orden de ks.

coef_mbest<-t(summary(mbest,Wald=TRUE)$coefficients)
write.csv(coef_mbest,"coeficientes_modelo.csv")
```

```
res_mbest<-t(summary(mbest,Wald=TRUE)$standard.errors)
write.csv(res_mbest,"residuos_modelo.csv")

###MULTICOLINEALIDAD###
mbest2<-multinom(contacto ~ 0 + num_conex + p_conex + DIAGESTION + PRODUC
                 max_porc_conex_6M + num_conex_2M + min_num_conex_4M +
                 rsaldo_inicial_2M_6M, data=data_model, maxit=1000)

#summary(mbest2)#Quitar el intercepto para multicolinealidad
mat_corr_model<-cor(data_model%%select(num_conex,p_conex,DIAGESTION,PROD
                                       max_porc_conex_6M,num_conex_2M,min_
                                       rsaldo_inicial_2M_6M))
vp<-eigen(mat_corr_model,symmetric = T,only.values = T)
val_prop<-vp$values
IC<-sqrt(max(val_prop)/min(val_prop))
IC

####RESIDUOS DE DEVIANZA####
residuos<-residuals(mbest)
summary(residuos)
shapiro.test(residuos[,1])
shapiro.test(residuos[,2])
shapiro.test(residuos[,3])
shapiro.test(residuos[,4])
#Prueba de wald
#Prueba de significancia, p-valores para los coeficientes
z<-summary(mbest)$coefficients/summary(mbest)$standard.errors
z

#prueba z para significancia de coeficientes ambas colas:
p<-(1-pnorm(abs(z),0,1))*2
p

#Directamente prueba significancia de wald para coeficientes:
library(AER)
coeftest(mbest)
write.csv(coeftest(mbest),"signif_model.csv")
```

```
#Pueba de verosimilitud, contraste condicional
anova(m2,mbest)

#Rcuadrado
m0<-multinom(contacto ~ 1,data=data_model%%select(-NUMEROOPERACION,
                                                  -NUMEROCLIENTE),maxit=1000)
dvf<-deviance(mbest)
dv0<-deviance(m0)
n<-12 #numero de variables predictoras
rmf<-1-((0.5*dvf+n+1)/(0.5*dv0+1))
rmf

######INTERVALOS DE CONFIANZA PARA LOS COEFICIENTES Y ODDS#######
ic<-confint(mbest,level=0.95)
exp(ic)
write.csv(ic,"inetervalos.csv")

odds<-t(exp(coef(mbest)))
write.csv(odds,"odds_modelo.csv")

######## KS MODELO####
head(pp<-fitted(mbest))
head(predict(mbest,type="probs"))
prob<-predict(mbest,type="probs")
colnames(prob)<-c("NC","H3","H4","H1","H2")

###KS DEL MODELO (MODELOS por categoria)
data_eval <- cbind(data_model, prob)
contacto <-ifelse(data_eval$contacto=="7-9am",1,
                  ifelse(data_eval$contacto=="9-13pm",2,
                  ifelse(data_eval$contacto=="13-16pm",3,
                  ifelse(data_eval$contacto=="16-21pm",4,0))))

contacto<-as.integer(contacto)

####CATEGORIA 1: 7-9#####
#contacto1<-ifelse(data_eval$contacto=="7-9am",0,1)
prob1 <- data_eval[["H1"]]
bivar1 <- data.frame(contacto,prob1)
```

```
# Funcion acumulada 1
Fn_1 <- ecdf(bivar1[bivar1[,1]==1,][,2])
# Funcion acumulada 0
Fn_NC1 <- ecdf(bivar1[bivar1[,1]==0,][,2])

minMax1<-seq(min(bivar1[bivar1[,1]==1,][,2],bivar1[bivar1[,1]==0,][,2]),
max(bivar1[bivar1[,1]==1,][,2],bivar1[bivar1[,1]==0,][,2]),
length.out = length(bivar1[bivar1[,1]==1,][,2]))
x01<-minMax1[which(abs(Fn_1(minMax1)-Fn_NC1(minMax1))==
max(abs(Fn_1(minMax1)-Fn_NC1(minMax1))))]
x01#punto de corte optimo
y01<-Fn_1(x01)
y11<-Fn_NC1(x01)

H1<-bivar1%%filter(contacto==1|contacto==0)
H1$grupo<-ifelse(H1$contacto==1,"7-9am","NC")

ggplot(H1, aes(x = prob1, group = contacto, color = grupo))+
  stat_ecdf(size=1) +
  xlab("Prob_7-9") +
  ylab("ECDF") +
  #geom_line(size=1) +
  geom_segment(aes(x = x01[1], y = y01[1], xend = x01[1], yend = y11[1]),
               linetype = "dashed", color = "red") +
  geom_point(aes(x = x01[1] , y= y01[1]), color="red", size=4) +
  geom_point(aes(x = x01[1] , y= y11[1]), color="red", size=4) +
  ggtitle("K-S Test: 7-9am / NC") +
  theme(legend.title=element_blank())

# Valor KS
ks1<-ks.test(bivar1[bivar1[,1]==0,][,2],bivar1[bivar1[,1]==1,][,2],
alternative = c("two.sided"))$statistic
n1<-length(bivar1[bivar1[,1]==0,][,2])
n<-nrow(data_model)
ksm1<-ks1*(n1+length(bivar1[bivar1[,1]==1,][,2]))/(4*n)

#Curva ROC
library(pROC)
```

```
plot.roc(H1$contacto,H1$prob1,print.auc=T,percent=T)
ROC1<-roc(H1$contacto,H1$prob1,direction=c("auto"),auc=TRUE,plot=F)
AUC1<-as.numeric(ROC1$auc)
GINI1<-2*AUC1-1
#multiclass.roc(contacto,data_eval$H1,percent=TRUE)


#####TABLA CLASIFICACION ENTRE C_NC Y LA PREDICCION P_C_NC Y PERFORMANCE
tablas<-function(prob,x0,contacto,cat,finq){
score1<-round(prob*1000)
corte<-x0
PC1_NC<-as.numeric(ifelse(prob>=corte,1,0))
SCORE<-cbind(contacto,PC1_NC,prob,score1)
colnames(SCORE)<-c("C_NC","PREDICCION_C1_NC","PROBABILIDAD","SCORE")

T_C<-table(SCORE[,1],SCORE[,2],dnn = c("C_NC","PC_NC"))

#Error de prediccion
print((T_C[1,2]+T_C[2,1])/sum(T_C))

#####PORCENTAJE RESPECTO AL TOTAL de C/NC por (fila)
T_C_p<-prop.table(T_C,1)
print(T_C_p)

T_C<-data.frame(T_C,round(T_C_p,2))
colnames(T_C)<-c("C_NC","PC_NC","FRECUENCIA","C_NC","PC_NC","PORC")
print(T_C)

###TABLA PERFORMANCE
Z<-data.frame(SCORE[,1],SCORE[,4])
Z<-Z[order(Z[,2],decreasing=TRUE),]

d <- quantile(Z[,2],prob=seq(0.1,1,finq))

razon_nc_c<-c(0)
A<-matrix(ncol=11,nrow=length(d))
A[1,1]<-1
A[length(d),2]<-999
```

```
for(i in 1:(nrow(A)-1))
{
  A[i,2]<-A[i+1,1]<-d[i]
}

TP<-data.frame(c(1:length(d)),A)
for (i in 1:(nrow(TP)))
{
  #num clientes en cada decil
  TP[i,4]<-nrow(subset(Z,subset=Z[,2]>TP[i,2] & Z[,2]<=TP[i,3]))
  TP[i,5]<-round(TP[i,4]/nrow(Z)*100,2)

  #no contactos en cada decil(MALOS)
  TP[i,6]<-nrow(subset(Z,subset=Z[,2]>TP[i,2] &
  Z[,2]<=TP[i,3] & Z[,1]==0))

  #contactos en cada decil (BUENOS)
  TP[i,8]<-nrow(subset(Z,subset=Z[,2]>TP[i,2] &
  Z[,2]<=TP[i,3] & Z[,1]==cat))

  #porc no_contactos decil
  TP[i,10]<-round((TP[i,6]/TP[i,4]),2)

  #porc contactos decil
  TP[i,11]<-round((TP[i,8]/TP[i,4]),2)

  #razón exito/fracaso
  if (TP[i,8]>0)
  {
    r<-round((TP[i,8]/TP[i,6]),0)
    razon_nc_c[i]<-round((TP[i,8]/TP[i,6]),2)
    TP[i,12]<-paste(c(r,1),collapse=":")
  }
  else
  {
    r<-round(TP[i,6],0)
    razon_nc_c[i]<-TP[i,6]
    TP[i,12]<-paste(c(r,0),collapse=":")
  }
```

```
}

for(j in 1:nrow(TP)){
#acum malos
TP[j,7]<-round(TP[j,6]/sum(TP[,6])*100,2)
#acum buenos
TP[j,9]<-round(TP[j,8]/sum(TP[,8])*100,2)
}

TP<-data.frame(TP[,1:5],cumsum(TP[,5]),TP[,6:7],cumsum(TP[,7]),TP[,8:9],
cumsum(TP[,9]),TP[,10:ncol(TP)],razon_nc_c)

colnames(TP)<-c("Decil","De","Hasta","No_Clientes","Porc_Clien",
"Porc_acum_Clien","No_NOCont","Porc_NOCont","Porc_acum_NOCont",
"No_Cont","Porc_Cont","Porc_acum_Cont","Porc_NOCont_Decil",
"Porc_Cont_Decil","ODDS","NOCONT/CONT")
TP
}
TP_7_9<-tablas(H1$prob1,x01,H1$contacto,cat=1,finq = 0.125)
write.csv(TP_7_9,"TP7_9.csv")

ggplot(TP_7_9,aes(x=Decil,y=Porc_Cont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_Cont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de Contactos de 7-9")

ggplot(TP_7_9,aes(x=Decil,y=Porc_NOCont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_NOCont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de No Contactos de 7-9")

####CATEGORIA 2: 9-13#####
prob2 <- data_eval[["H2"]]
bivar2 <- data.frame(contacto,prob2)

# Funcion acumulada / 9-13pm
Fn_2 <- ecdf(bivar2[bivar2[,1]==2,][,2])
# Funcion acumulada / NC de 9-13am
Fn_NC2 <- ecdf(bivar2[bivar2[,1]==0,][,2])
```

```
minMax2<-seq(min(bivar2[bivar2[,1]==2,][,2],bivar2[bivar2[,1]==0,][,2]),
max(bivar2[bivar2[,1]==2,][,2],bivar2[bivar2[,1]==0,][,2]),
length.out = length(bivar2[bivar2[,1]==2,][,2]))
x02<-minMax2[which(abs(Fn_2(minMax2)-Fn_NC2(minMax2))==
max(abs(Fn_2(minMax2)-Fn_NC2(minMax2))))]
x02#punto de corte optimo
y02<-Fn_2(x02)
y12<-Fn_NC2(x02)

H2<-bivar2 %>% filter(contacto==2|contacto==0)
H2$grupo<-ifelse(H2$contacto==2,"9-13pm","NC")

ggplot(H2, aes(x = prob2, group = contacto, color = grupo))+
  stat_ecdf(size=1) +
  xlab("Prob_9-13") +
  ylab("ECDF") +
  #geom_line(size=1) +
  geom_segment(aes(x = x02[1], y = y02[1], xend = x02[1], yend = y12[1]),
               linetype = "dashed", color = "red") +
  geom_point(aes(x = x02[1] , y= y02[1]), color="red", size=4) +
  geom_point(aes(x = x02[1] , y= y12[1]), color="red", size=4) +
  ggtitle("K-S Test: 9-13pm / NC") +
  theme(legend.title=element_blank())

# Valor KS
ks2<-ks.test(bivar2[bivar2[,1]==0,][,2],bivar2[bivar2[,1]==2,][,2],
alternative = c("two.sided"))$statistic
n2<-length(bivar2[bivar2[,1]==0,][,2])
ksm2<-ks2*(n2+length(bivar2[bivar2[,1]==2,][,2]))/(4*n)

#Curva ROC
library(pROC)
plot.roc(H2$contacto,H2$prob2,print.auc=T,percent=T)
ROC2<-roc(H2$contacto,H2$prob2,direction=c("auto"),auc=TRUE,plot=F)
AUC2<-as.numeric(ROC2$auc)
GINI2<-2*AUC2-1

#roc(contacto2,data_eval$H2)
```

```
#multiclass.roc(contacto, data_eval$H2)
#plot.roc(contacto, data_eval$H2)

###### TABLAS DE CLASIFICACION Y PERFORMANCE
TP_9_13<-tablas(H2$prob2,x02,H2$contacto,cat=2,finq = 0.125)
write.csv(TP_9_13,"TP_9_13.csv")

ggplot(TP_9_13,aes(x=Decil,y=Porc_Cont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_Cont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de Contactos de 9-13")

ggplot(TP_9_13,aes(x=Decil,y=Porc_NOCont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_NOCont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de No Contactos de 9-13")

####CATEGORIA 3: 13-16####
prob3 <- data_eval[["H3"]]
bivar3 <- data.frame(contacto,prob3)

# Funcion acumulada / 13-16pm
Fn_3 <- ecdf(bivar3[bivar3[,1]==3,][,2])
# Funcion acumulada / NC de 13-16pm
Fn_NC3 <- ecdf(bivar3[bivar3[,1]==0,][,2])

minMax3<-seq(min(bivar3[bivar3[,1]==3,][,2],bivar3[bivar3[,1]==0,][,2]),
max(bivar3[bivar3[,1]==3,][,2],bivar3[bivar3[,1]==0,][,2]),
length.out = length(bivar3[bivar3[,1]==3,][,2]))
x03<-minMax3[which(abs(Fn_3(minMax3)-Fn_NC3(minMax3))==
max(abs(Fn_3(minMax3)-Fn_NC3(minMax3))))]
x03#punto de corte optimo
y03<-Fn_3(x03)
y13<-Fn_NC3(x03)

H3<-bivar3 %%filter(contacto==3|contacto==0)
H3$grupo<-ifelse(H3$contacto==3,"13-16pm","NC")

ggplot(H3, aes(x = prob3, group = contacto, color = grupo))+
```

```
  stat_ecdf(size=1) +
  xlab("Prob_13-16") +
  ylab("ECDF") +
  #geom_line(size=1) +
  geom_segment(aes(x = x03[1], y = y03[1], xend = x03[1], yend = y13[1]),
               linetype = "dashed", color = "red") +
  geom_point(aes(x = x03[1] , y= y03[1]), color="red", size=4) +
  geom_point(aes(x = x03[1] , y= y13[1]), color="red", size=4) +
  ggtitle("K-S Test: 13-16pm / NC") +
  theme(legend.title=element_blank())

# Valor KS
ks3<-ks.test(bivar3[bivar3[,1]==0,][,2],bivar3[bivar3[,1]==3,][,2],
alternative = c("two.sided"))$statistic
n3<-length(bivar3[bivar3[,1]==0,][,2])
ksm3<-ks3*(n3+length(bivar3[bivar3[,1]==3,][,2]))/(4*n)

#CURVA ROC
plot.roc(H3$contacto,H3$prob3,print.auc=T,percent=T)
ROC3<-roc(H3$contacto,H3$prob3,direction=c("auto"),auc=TRUE,plot=F)
AUC3<-as.numeric(ROC3$auc)
GINI3<-2*AUC3-1

#multiclass.roc(contacto,data_eval$H2)
#plot.roc(contacto,data_eval$H2)

####TABLAS DE CLASIFICACION Y PERFORMANCE
TP_13_16<-tablas(H3$prob3,x03,H3$contacto,cat=3,finq=0.125)
write.csv(TP_13_16,"TP_13_16.csv")

ggplot(TP_13_16,aes(x=Decil,y=Porc_Cont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_Cont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de Contactos de 13-16")

ggplot(TP_13_16,aes(x=Decil,y=Porc_NOCont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_NOCont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de No Contactos de 13-16")
```

```
####CATEGORIA 4: 16-21 ####
prob4 <- data_eval[["H4"]]
bivar4 <- data.frame(contacto,prob4)

# Funcion acumulada / 13-16pm
Fn_4 <- ecdf(bivar4[bivar4[,1]==4,][,2])
# Funcion acumulada / NC de 9-13am
Fn_NC4 <- ecdf(bivar4[bivar4[,1]==0,][,2])

minMax4<-seq(min(bivar4[bivar4[,1]==4,][,2],bivar4[bivar4[,1]==0,][,2]),
max(bivar4[bivar4[,1]==4,][,2],bivar4[bivar4[,1]==0,][,2]),
length.out = length(bivar4[bivar4[,1]==4,][,2]))
x04<-minMax4[which(abs(Fn_4(minMax4)-Fn_NC4(minMax4))==
max(abs(Fn_4(minMax4)-Fn_NC4(minMax4))))]
x04#punto de corte optimo
y04<-Fn_4(x04)
y14<-Fn_NC4(x04)

H4<-bivar4%%filter(contacto==4|contacto==0)
H4$grupo<-ifelse(H4$contacto==4,"16-21pm","NC")

ggplot(H4, aes(x = prob4, group = contacto, color = grupo))+
  stat_ecdf(size=1) +
  xlab("Prob_16-21") +
  ylab("ECDF") +
  #geom_line(size=1) +
  geom_segment(aes(x = x04[1], y = y04[1], xend = x04[1], yend = y14[1]),
               linetype = "dashed", color = "red") +
  geom_point(aes(x = x04[1] , y= y04[1]), color="red", size=4) +
  geom_point(aes(x = x04[1] , y= y14[1]), color="red", size=4) +
  ggtitle("K-S Test: 16-21pm / NC") +
  theme(legend.title=element_blank())

# Valor KS
ks4<-ks.test(bivar4[bivar4[,1]==0,][,2],bivar4[bivar4[,1]==4,][,2],
alternative = c("two.sided"))$statistic
n4<-length(bivar4[bivar4[,1]==0,][,2])
ksm4<-ks4*(n4+length(bivar4[bivar4[,1]==4,][,2]))/(4*n)
```

```
#CURVA ROC
plot.roc(H4$contacto,H4$prob4,print.auc=T,percent=T)
ROC4<-roc(H4$contacto,H4$prob4,direction=c("auto"),auc=TRUE,plot=F)
AUC4<-as.numeric(ROC4$auc)
GINI4<-2*AUC4-1

#multiclass.roc(contacto,data_eval$H4)
#plot.roc(contacto,data_eval$H4)

####TABLAS DE CLASIFICACION Y PERFORMANCE
TP_16_21<-tablas(H4$prob4,x04,H4$contacto,cat=4,finq=0.125)
write.csv(TP_16_21,"TP_16_21.csv")

ggplot(TP_16_21,aes(x=Decil,y=Porc_Cont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_Cont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de Contactos de 16-21")

ggplot(TP_16_21,aes(x=Decil,y=Porc_NOCont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_NOCont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de No Contactos de 16-21")

#######KS MULTINOMIAL#####
ksm1+ksm2+ksm3+ksm4

####CONTACTABILIDAD GENERAL####
contacto_total<-ifelse(data_eval$contacto=="NC",0,1)
contacto_total<-as.integer(contacto_total)

prob_cont<-unlist(apply(prob[,2:5],1,sum))
bivar <- data.frame(contacto_total,prob_cont=prob_cont)

# Funcion acumulada /CON
Fn_c <- ecdf(bivar[bivar[,1]==1,][,2])
# Funcion acumulada / NC
Fn_NC <- ecdf(bivar[bivar[,1]==0,][,2])
```

```
minMax<-seq(min(bivar[bivar[,1]==1,][,2],bivar[bivar[,1]==0,][,2]),
max(bivar[bivar[,1]==1,][,2],bivar[bivar[,1]==0,][,2]),
length.out = length(bivar[bivar[,1]==1,][,2]))
x0<-minMax[which(abs(Fn_c(minMax)-Fn_NC(minMax))==
max(abs(Fn_c(minMax)-Fn_NC(minMax))))]
x0#punto de corte optimo
y0<-Fn_c(x0)
y1<-Fn_NC(x0)

bivar$grupo<-ifelse(bivar$contacto_total==1,"C","NC")

ggplot(bivar, aes(x = prob_cont, group = contacto_total, color = grupo))+
  stat_ecdf(size=1) +
  xlab("Prob_cont") +
  ylab("ECDF") +
  #geom_line(size=1) +
  geom_segment(aes(x = x0[1], y = y0[1], xend = x0[1], yend = y1[1]),
               linetype = "dashed", color = "red") +
  geom_point(aes(x = x0[1] , y= y0[1]), color="red", size=4) +
  geom_point(aes(x = x0[1] , y= y1[1]), color="red", size=4) +
  ggtitle("K-S Test: C / NC") +
  theme(legend.title=element_blank())

####TABLAS DE CLASIFICACION Y PERFORMANCE
TP_C<-tablas(prob_cont,x0,contacto_total,cat=1,finq = 0.1)
write.csv(TP_C,"TP_C.csv")

ggplot(TP_C,aes(x=Decil,y=Porc_Cont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_Cont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de Contactos")

ggplot(TP_C,aes(x=Decil,y=Porc_NOCont)) +
geom_bar(stat="identity",fill="lightblue",colour="black",width=0.5) +
geom_text(aes(label=round(Porc_NOCont,2)), vjust=-0.5, colour="black") +
ggtitle(label = "Porcentaje de No Contactos")

# Curva de buenos
#plot(Fn_c,do.points = FALSE,verticals=T,col='green',main='KS Test')
```

```
# Curva de malos
#lines(Fn_NC, lty=3,do.points = FALSE, verticals=T, col='red')
# Valor KS
ks.test(bivar[bivar[,1]==0,][,2], bivar[bivar[,1]==1,][,2],
alternative = c("two.sided"))
plot.roc(bivar$contacto_total, bivar$prob_cont, print.auc=T, percent=T)
ROC<-roc(bivar$contacto_total, bivar$prob_cont, direction=c("auto"),
auc=TRUE, plot=F)
AUC<-as.numeric(ROC$auc)
GINI<-2*AUC-1

plot(Fn_NC,do.points = FALSE, verticals=T, col='green',main='KS Test')
lines(Fn_1,lty=3,do.points = FALSE, verticals=T, col='black')
lines(Fn_2,lty=3,do.points = FALSE, verticals=T, col='yellow')
lines(Fn_3,lty=3,do.points = FALSE, verticals=T, col='blue')
lines(Fn_4,lty=3,do.points = FALSE, verticals=T, col='red')
```

# Anexo D

# Árboles de Decisión y algoritmos en SPSS

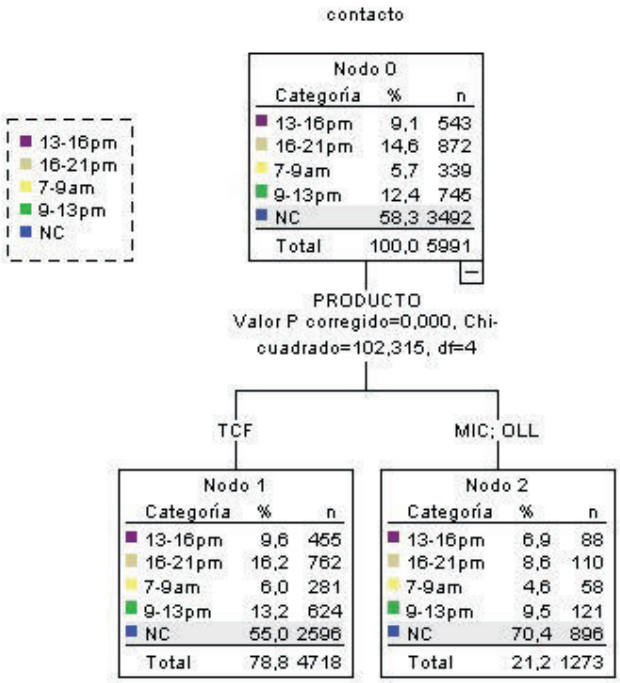


**Figura D.1:** Variable PRODUCTO

```
STRING pre_001 (A7).

/* Node 1 */.
DO IF (PRODUCTO NE "MIC"  AND  PRODUCTO NE "OLL").
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.550233.
END IF.
EXECUTE.

/* Node 2 */.
DO IF (PRODUCTO EQ "MIC" OR PRODUCTO EQ "OLL").
COMPUTE nod_001 = 2.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.703849.
END IF.
EXECUTE.
```

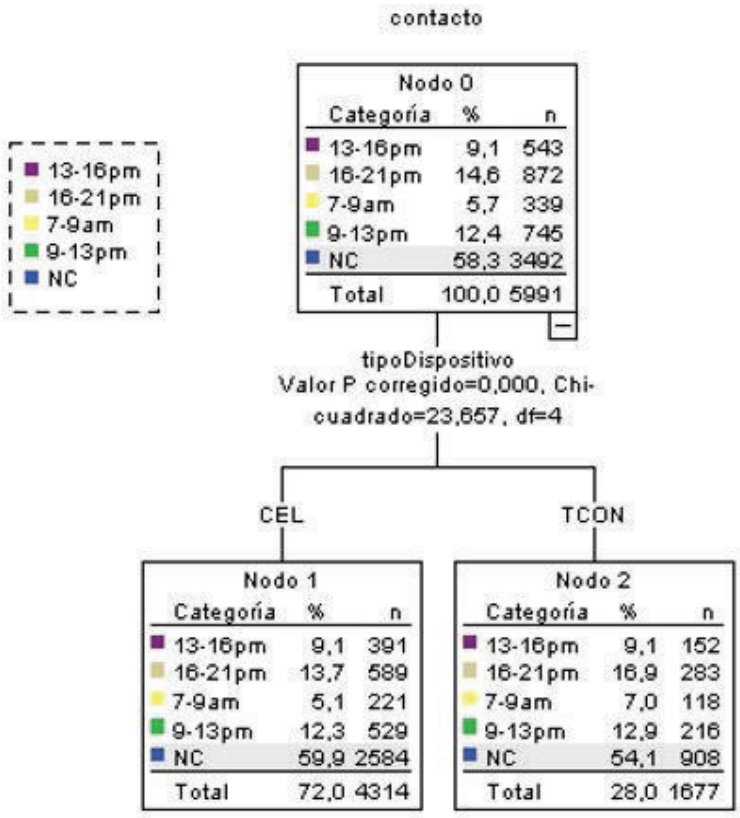


**Figura D.2:** Variable TipoDispositivo

```
STRING pre_001 (A7).
/* Node 1 */.
DO IF (tipoDispositivo NE "TCON").
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.598980.
END IF.
EXECUTE.

/* Node 2 */.
DO IF (tipoDispositivo EQ "TCON").
COMPUTE nod_001 = 2.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.541443.
END IF.
EXECUTE.
```

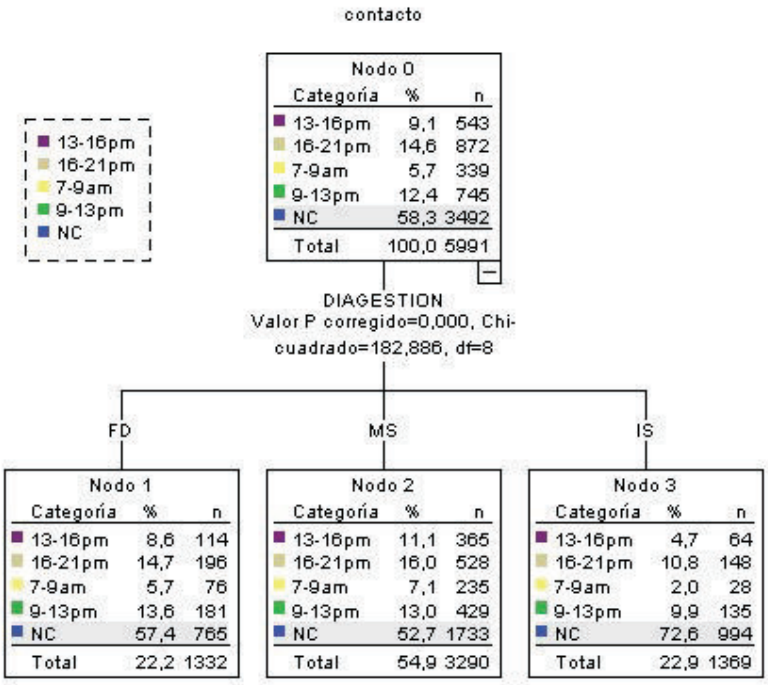


**Figura D.3:** Variable DIAGESTION

```
STRING pre_001 (A7).
/* Node 1 */.
DO IF (DIAGESTION EQ "FD").
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.574324.
END IF.
EXECUTE.

/* Node 2 */.
DO IF (DIAGESTION NE "FD" AND DIAGESTION NE "IS").
COMPUTE nod_001 = 2.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.526748.
END IF.
EXECUTE.
```

```
/* Node 3 */.
DO IF (DIAGESTION EQ "IS").
COMPUTE nod_001 = 3.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.726077.
END IF.
EXECUTE.
```

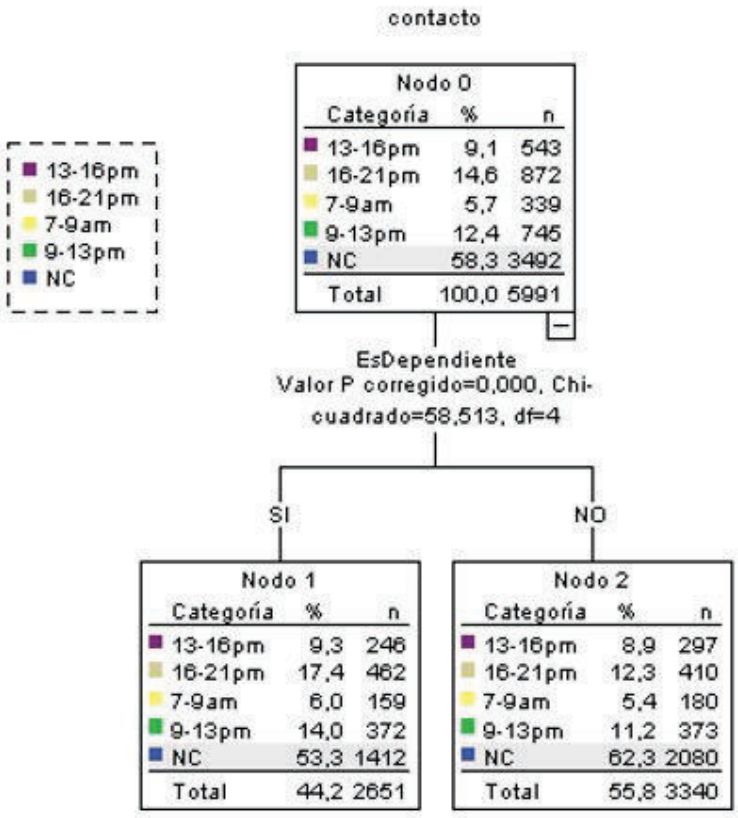


**Figura D.4:** Variable EsDependiente

```
STRING pre_001 (A7).
/* Node 1 */.
DO IF (EsDependiente EQ "SI").
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.532629.
END IF.
EXECUTE.
```

```
/* Node 2 */.
DO IF (EsDependiente NE "SI").
COMPUTE nod_001 = 2.
COMPUTE pre_001 = 'NC'.
COMPUTE prb_001 = 0.622754.
END IF.
EXECUTE.
```

# Referencias


[1] H. Bayrak, "Determining best time to reach customers in a multi-channel world ensuring right party contact and increasing interaction likelihood," *US 2013/0060587 A1*, vol. 1, 2013.

[2] V. P. Fernández, "Regresión logística multinomial," *Departamento de Estadística e Investigación Operativa. ETS de Ingenierías Agrarias.Valladolid.*, vol. 1, 2004.

[3] C. R. D. y Juan Cáceres Hernández, "Modelos de elección discreta y especificaciones ordenadas: uma reflexión metodológica," *ESTADÍSTICA ESPAÑOLA*, vol. 49, 2007.

[4] D. W. Hosmer and S.Lemeshow, *Applied Logistic Regression*. Wiley Interscience, 2000.

[5] S. C. Loftus, L. L. House, M. C. Hughey, J. B. Walke, M. H, and L. K. Belden, "Dimension reduction for multinomial models via a kolmogorov-smirnov measure (ksm)," 2015.

[6] A. Agresti, *AN INTRODUCTION TO CATEGORICAL ANALYSIS*. Wiley Interscience, 2007.

[7] I. SPSS, "Decision trees 20 manual," 2011.

[8] A. Castro, *Regresión Lineal*. Monografías de Matemática y Estadística, Quito.

[9] T. Landgrebe and R. P. Duin, "A simplified extension of the area under the roc to the multiclass domain," *Elect. Eng., Maths and Comp. Sc., Delft University of Technology, The Netherlands*, vol. 1, 2006.

[10] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Kluwer Academic Publishers*, vol. 45, 2001.

[11] M. de Lourdes Velasco Vázquez, "Un modelo de regresión poisson inflado con ceros para analizar datos de un experimento de fungicidas en jitomate," 2008.

[12] M. Ángeles Dueñas Rodríguez, "Modelos de respuesta discreta en r y aplicacion con datos reales," 2011.

[13] W. J. Conover, "Several k-sample kolmogorov-smirnov tests," *JSTOR*, vol. 36, 1965).

[14] Y. Croissant, "Estimation of multinomial logit models in r: The mlogit packages," 2015.

[15] A. Novales, *ECONOMETRIA*. Mc Graw Hill, 1993.

[16] D. Gujarati, *ECONOMETRIA*. Mc Graw Hill, 2009.

[17] J. Barreiro, "Modelo logit multinomial: Una aplicaciÓn regional al sector lÁcteo," *AEEADE*, vol. 4-1, 2004.

[18] C. Beltrán, "Aplicación del análisis de regresión logística multinomial en la clasificación de textos académicos: Biomet," *INFOSUR*, vol. 5, 2011.

[19] M. Arriaza, *Guía Práctica de Análisis de Datos*. IFAPA, 2006.

[20] "Mejores prácticas en estrategias de cobranzas," 2008.

[21] I. C. C. Tecnology, "Guía de outbound performance."