ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

ESTIMACIÓN DE ASIGNACIÓN DE CAPITAL SEMILLA PARA EMPRENDIMIENTOS DEL DISTRITO METROPOLITANO DE QUITO USANDO MODELO SCORING EN EL MARCO DEL PROGRAMA FONDOS DE LA CIUDAD DE QUITO (FONQUITO)

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

Autor: GERMÁN VICENTE HERRERA LOZADA

german.herrera@epn.edu.ec

Director: MENTHOR OSWALDO URVINA MAYORGA

menthor.urvina@epn.edu.ec

Quito, agosto, 2022

Ecuador

DECLARACIÓN

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

_____

Germán Vicente Herrera Lozada

Certifico que el presente trabajo de integración curricular fue desarrollado por GERMÁN VICENTE HERRERA LOZADA, bajo mi supervisión.

_____

Menthor Oswaldo Urvina Mayorga

Director

DECLARACIÓN DE AUTORÍA

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

GERMÁN VICENTE HERRERA LOZADA

MENTHOR OSWALDO URVINA MAYORGA

RESUMEN

Los emprendimientos que parten de una adecuada distribución financiera en sus operacionesde gestación contribuyen a encaminar un proyecto de emprendimiento sostenible, sin embargo, la posibilidad del mismo depende de factores como la propuesta de valor, experiencia, edad, capital, segmento de clientes, mercado, entre otros que determinan el éxito del producto o servicio ofertado. Existen programas de inversiones cuya finalidad es el incentivar económicamente a emprendimientos mediante la adjudicación de capital semilla, el cual permite insertar inversiones considerando los riesgos que el mismo está sometido en el mercado, cuya intención es la correctay adecuada adjudicación del capital en base a la información obtenida de cada factor. El presente estudio tiene como objetivo la construcción de un modelo tipo scoring con base a la informaciónde los factores que determinan el posicionamiento de un emprendimiento en el mercado local, mediante la asignación adecuada de capital semilla, que permita a este segmento la evolución; en ventas, capital humano y en impacto generado.

Palabras clave: Modelo scoring, capital semilla, emprendimiento, adjudicación.

ABSTRACT

The enterprises that start from an adequate financial distribution in their gestation operations contribute to start a sustainable entrepreneurship project, however, the possibility of it depends on factors such as the value proposition, experience, age, capital, customer segment, market, among others that determine the success of the product or service offered. There are investment programs whose purpose is to provide financial incentives to start-ups through the allocation of seed capital, which allows investments to be included considering the risks that the same is subjected to in the market, whose intention is the correct and appropriate allocation of capital based on the information obtained from each factor. The objective of this study is to construct a scoring model based on information on the factors that determine the positioning of an enterprise in the local market, through the appropriate allocation of seed capital, that allows this segment the evolution; in sales, human capital and in generated impact.

Keywords: Scoring model, seed capital, entrepreneurship, adjudication.

ÍNDICE

Capítulo I Introducción

Descripción del componente desarrollado

El emprendimiento es aquél reciente o nuevo proyecto empresarial que tiene un potencial de crecimiento gracias a una ventaja competitiva, tecnológica o no, para convertirse en al menos una pequeña empresa a quien la oferta y productos financieros desatienden sus necesidades, generando una falla que radica en la existencia de información asimétrica respecto de las verdaderas bondades y riesgos de los proyectos nuevos de estas empresas. Los factores de este segmento de empresarialidad buscan la materialización plena del potencial de crecimiento de los emprendimientos como el acceso a redes y servicios de apoyo empresarial de alto valor añadido que suplan y corrijan las carencias del emprendedor. Esta necesidad permite integrar el concepto de capital semilla, la cual es aquella inversión que contrarresta las debilidades de los emprendimientos al ser también un capital que aporta no únicamente en financiación sino también gran experiencia en áreas como estrategia, administración o comercial.

Los emprendimientos parten con un capital inicial, que bajo una adecuada distribución en sus operaciones de gestación, contribuyen a encaminar un proyecto de emprendimiento sostenible en el tiempo, sin embargo, la posibilidad de este depende de factores como; el componente diferenciador, experiencia, estatus de la etapa de desarrollo, proyecciones del uso del dinero,

estatus tributario con entidades públicas, entre otras, que determinan una adecuada ejecución del producto o servicio ofertado.

Existen programas de inversiones cuya finalidad es el incentivar económicamente a dichos emprendimientos mediante la adjudicación de capital semilla el cual permite insertar inversiones al segmento, considerando los riesgos que el mismo está sometido en el mercado, cuya intención es la correcta y adecuada adjudicación del capital en base a la información obtenida de cada factor. El presente estudio usa la información generada en estos programas para la implementación de un modelo de tipo scoring o de calificación, mismo que será desarrollado bajo el uso de software libre estadístico RStudio, cuyo fin permita determinar la estimación adecuada de adjudicación de capitales a posibles beneficiarios.

Definiciones

Capital semilla: - Es la inversión de recursos en la fase inicial de un proyecto, desde su concepción hasta el desarrollo de un proyecto innovador.

Desarrollo inicial: Son aquellos emprendimientos que cuentan con ventas y requieren recursos para cubrir una necesidad o aprovechar una oportunidad en el mercado, su finalidad es generar utilidad, empleo y desarrollo, así como, implementar nuevos productos y/o servicios novedosos o significativamente mejorados frente a los existentes en el mercado.

Emprendedor: Son personas naturales o jurídicas que persiguen un beneficio, trabajando individual o colectivamente. Pueden ser definidos como individuos que innovan, identifican y crean oportunidades, desarrollan un proyecto y organizan los recursos necesarios para aprovecharlo

Emprendimiento con potencial en convertirse en una MI PYME: Es un proyecto orientado

al desarrollo de un nuevo o significativamente mejorado bien o servicio cuyo factor fundamental es el uso del conocimiento que se genera a partir de procesos de investigación, desarrollo experimental y tecnológico o procesos creativos con base científica, cuyo fin último es su introducción en el mercado.

Equipo multidisciplinario: Es aquel formado por un grupo de personas con diferentes formaciones académicas, especializaciones técnicas y experiencias profesionales, que trabajan en conjunto ya sea de forma habitual o durante un tiempo determinado para llevar a cabo un proyecto.

Idea con prototipo: Se refiere a los proyectos formulados para el desarrollo de la primera versión del producto que se pretende lanzar al mercado, el cual permite validar que cumple con las características y funciones diseñadas con el objetivo de desarrollar nuevos o sustancialmente mejorados productos y/o servicios a los ya existentes en el mercado.

Innovación: Es el proceso creativo mediante el cual se genera un nuevo producto, diseño, proceso, servicio, método u organización, o añade valor a los existentes.

MiPyme: Son pequeñas empresas formadas por diferentes estructuras ya sean familiares, amigos o socios quienes deben aportar con capital para su desarrollo, se enfocan en un sector exclusivo en el mercado económico tienen pequeñas cantidades de trabajadores, capital e infraestructura que la diferencia de las grandes empresas que necesitan de financiamiento para empezar a operar.

Persona natural: Es aquel individuo que tiene la capacidad de ejercer derechos y contraer obligaciones, esta se la adquiere con la mayoría de edad.

Persona jurídica: Es una persona ficticia, capaz de ejercer derechos y contraer obligaciones civiles, y de ser representada judicial y extrajudicialmente.

Objetivo General

Implementar un modelo matemático de tipo scoring para determinar, bajo variables significativas, la asignación de capital semilla a beneficiarios del Distrito Metropolitano de Quito (DMQ) en marco del programa Fondos de la Ciudad de Quito (FonQuito).

Objetivos Específicos

- Levantamiento del flujo de información del programa FonQuito.
- Tratamiento de la data mediante proceso ETL (Extract Transform and Load).
- Delimitación de variables significativas que determinan la asignación de capital semilla a beneficiarios.
- Implementación de modelo (logit) para la información estructurada.
- Testeo y ajustes de modelo para validación.

Alcance

El presente estudio recolecta información del proceso de la primera convocatoria del programa FonQuito, ejecutada por la Corporación de Promoción Económica (Conquito), mismo que inicio en marzo de 2021. El programa en mención tiene información de candidatos del DMQ cuyos proyectos de emprendimiento tienen enfoque innovador o componente diferenciador. El modelo propuesto se desarrolla en función de las variables generadas durante el programa y son las que determinan la diferenciación de producto, potencialidad en el mercado objetivo, experiencia del proyecto de emprendimiento y grado de madurez o desarrollo de este.

Marco Teórico

La reciente pandemia ha generado oportunidades que pueden ser aprovechadas por los emprendedores, el gobierno local y también el ecosistema empresarial. De acuerdo con los expertos de organizaciones de ayuda, se han generado transiciones importantes “Dentro de las principales se encuentran: 1) masificación del uso de servicios digitales; 2) implementación de estrategias en las organizaciones que incluyan nuevas tecnologías de forma continua; 3) desarrollo de nuevos sectores como salud, teletrabajo, educación en línea, y todos los procesos de apoyo relacionados; 4) desarrollo de nuevos modelos de negocios que integren la sostenibilidad y la transformación digital como pilares críticos en los mismos; y 5) adopción de nuevas políticas públicas para acelerar

el emprendimiento, entre otros."[1]

Al considerar los objetivos planteados para el desarrollo sostenible en su agenda para el 2030 por el Programa de las Naciones Unidas para el Desarrollo (PNUD), comprenden acciones a ejecutarse con la finalidad de aportar con su cumplimiento, en tal sentido la Corporación de Promoción Económica Conquito ha creado el programa denominado Fondos de la Ciudad de Quito (FonQuito). El programa orienta sus objetivos como; reducir la pobreza en sus aristas para todos, incentivar el crecimiento económico sostenible e inclusivo, el empleo productivo y pleno decente, edificar infraestructuras resilientes, fomentar la industrialización sostenible e inclusiva y suscitar la innovación.

Este programa busca brindar a los emprendedores, actores de la economía popular y solidaria (EPS), micro, pequeñas y medianas empresas (Mipymes), los instrumentos necesarios para la puesta en marcha de su proyecto empresarial, consolidando sus empresas facilitando el acceso a fondos revolventes o subvencionables, para dotarles de instrucción y asesoramiento en gestión empresarial, diseño de modelo de negocios, herramientas digitales, mejoramiento de equipamiento y habilidades directivas.

El valor total del proyecto fue $1 045.000,00 y los montos financiados a proyectos beneficiarios para la primera convocatoria del programa fueron de $3.000. La convocatoria fue ejecutada a inicios del año 2021, etapa que se definió un formulario inicial con el cual los interesados o

[1] Los ecosistemas de emprendimiento de América Latina y el Caribe frente al COVID-19. Impactos, necesidades y recomendaciones: https://prodem.ungs.edu.ar/publicaciones_prodem/los-ecosistemas-de-emprendimiento-de-america-latina-y-el-carina-frente-al-covid-19-impactos-necesidades-y-recomendaciones/

candidatos a ser beneficiarios debieron cumplir con las bases delimitadas en la convocatoria tales como; propuesta con componente diferenciador, proyectos con un máximo de cinco años de actividad, gestación y ejecución en el DMQ, constar en lista blanca de entidades públicas como el Instituto Ecuatoriano de Seguridad Social (IESS), el Servicio de Rentas Internas (SRI) y el Servicio Nacional de Contratación Pública (SERCOP), el proyecto postulante con distinta etapa de desarrollo del mismo fue beneficiario de capital semilla por anteriores programas gestados por la corporación, equipo multiciplinario con al menos dos personas integrantes. La información recolectada posteriormente fue validada a través de medios oficiales y bajo metodologías de la corporación.

La intermediación financiera forma parte de las actividades principales que una organización sin fines de lucro como la Corporación Conquito puede ejecutar, impulsando económicamente al emprendimiento mediante la transferencia de recursos públicos y/o privados, provenientes de aquellos organismos con presupuestos asignados para este fin, a aquellos proyectos beneficiarios que requieren dichos recursos para invertirlos.

La ejecución y cumplimiento de este ejercicio conduce a la corporación hacia una exposición de un grupo de riesgos que desencadenan la deficiente distribución de estos recursos económicos hacia los beneficiarios. Por lo cual es imprescindible para la entidad integrar metodologías y políticas que direccionen a una apropiada gestión y mitigación de los riesgos como también el incorporar herramientas eficientes para la identificación, monitoreo, medición, mitigación, control y divulgación de estos.

En base a lo expuesto surge la metodología llamada scoring o calificación de candidatos (scoring), misma que tiene perspectivas distintas dependiendo de la conceptualización del autor, sin embargo como lo define D. J. Hand y Henley, es un método estadístico formalizado el cual

califica o clasifica a candidatos buenos y malos. A partir de los años 70 estos métodos han tenido gran crecimiento e importancia debido que éstos remplazaban a los métodos tradicionales para la decisión de concesiones de financiamiento a solicitantes de crédito, dichas decisiones se basaban en el juicio humano y experiencias de decisiones históricas.

Para Thomas, Edelman, y Crook el scoring es un método de calificación para el financiamiento cuya decisión se fundamenta en modelos de decisión y técnicas que permiten la concesión de créditos para agentes deficitarios. Por otro lado Shanmugapriya lo define como el proceso analítico del comportamiento histórico de los candidatos con la intención de discriminar aquellos en estatus de bancarrota y aquellos que no lo están.

Ya que existen algunas conceptualizaciones para un scoring, en el presente trabajo se considera como la calificación de financiamiento del candidato en base a la información que se conozca del mismo y su determinación parte de un conglomerado de metodologías que permitirán evidenciar la posibilidad que un candidato cumpla o no en su postulación a beneficiario de capital semilla.

Estos modelos de calificación o scoring, poseen ciertas funciones dentro del tiempo de vida del financiamiento comprende el otorgamiento, seguimiento, cobranza y recuperación, a pesar de ello, el presente trabajo se orienta en los métodos que sirven de apoyo para la toma de decisiones en el proceso otorgamiento de capital semilla, dichos métodos permiten cuantificar riesgos potenciales de candidatos con el fin de determinar quiénes podrán ser beneficiarios para adjudicación del capital, determinando variables significativas que permitan identificar a los postulantes cuyo perfil de riesgo se acondicione a los criterios establecidos en el programa.

Se cuentan con diversos métodos de modelamiento scoring, no obstante, se plantean modelos tradicionales como el de regresión logística, regularmente usados por instituciones financieras debido al poder de predicción y simplicidad de interpretación.

Estos métodos tradicionales tienen su ventaja al ser relativamente sencillos, fáciles de implementar e interpretar, características ausentes en los modelos de redes neuronales debido a la estructura compleja que limita la obtención de resultados óptimos. Por lo cual para todas las posibles situaciones es dificultoso la determinación de un "mejor" modelo en general.

El autor Zhang manifiesta que el estadístico Kolmogorov Smirnov (K-S) es un método mayormente usado para evaluar la influencia de clasificación de los modelos. Expone también que para investigadores o analistas, ningún método de evaluación representa una solución íntegra, sino mas bien se requiere la comprensión del problema y de los datos a estudio, para la correcta decisión de metodología y proceso de evaluación del modelo.

Esencia de los modelos

La finalidad de los modelos de regresión es la estimación de la posibilidad de incumplimiento de un candidato a beneficiario de capital semilla; lo cual es imprescindible una variable que identifique al candidato como bueno o malo, dicha variable es representada por la letra Y, llamada como variable dependiente en el modelo.

Esta variable dependiente es dicotómica que toma los valores:

$$Y = \begin{cases} 1: Si\ solicitante\ es\ definido\ como\ un\ buen\ candidato \\ 0: Si\ solicitante\ es\ definido\ como\ un\ mal\ candidato \end{cases} \quad (1)$$

El concepto de candidato bueno y malo detallado en ecuación (1), para la variable dependiente Y, es desarrollado con base a información demográfica y parámetros de evaluación generados en la corporación.

Estos modelos estiman la probabilidad de la variable dependiente Y pudiendo la misma tomar

el valor de 0 o 1, con base a un conjunto de variables independientes que son representadas mediante la letra X, dichas variables son de tipo cualitativas o cuantitativas. Estas variables se generan a partir de fuentes de información internas y externas del solicitante, como puede ser; información del proyecto, estado de desarrollo, demográfico, etc. En su mayoría dependerán de las características del emprendimiento solicitante que se esté considerando.

Regresión logística - logit

Los modelos logit pertenecen al conjunto de modelos de regresión con respuesta binaria, por otro lado, las variables independientes pueden ser cuantitativas o cuantitativas, o una combinación de ambas.

El modelo se fundamenta en la función de distribución logística y su estructura se presenta como:

$$P(Y = 1 \mid X) = F(z) = \frac{\exp(z)}{1 + \exp(z)}, \quad si, \quad -\infty < z < \infty \tag{2}$$

Conocido como $z = X^T\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$

En el cual:

• Y: variable dependiente binaria que puede tomar valores como; 0 (candidato malo) y 1 (candidato bueno).

• X: conjunto de *n* variables independientes $(x_1, x_2, \dots, x_n)$ vinculadas con la información del solicitante con el fin de explicar o predecir el valor que tomará Y.

• $F(z)$: como la función de probabilidad que depende del vector $\beta = (\beta_0, \beta_1, \dots, \beta_n)$, que vincula las variables independientes X con la variable dependiente Y. Dicha función sostiene un rango entre $[0, 1]$, llamada también función de distribución logística.

La finalidad del modelo consiste en hallar los coeficientes del vector β que mejor se ajusten a la ecuación (2).

Estimación de parámetros del modelo logit

En la estimación de coeficientes de β, pueden obtenerse con base al método de máxima verosimilitud (MV).

Sea un conjunto de k individuos, tal que, catalogarlos como buenos o malos candidatos está definido mediante la variable Y_i , con $i = 1, 2, \dots, k$.

Dado que cada candidato Y_i es una variable aleatoria tipo Bernoulli, pues toma los valores 0 o 1, se puede formular la probabilidad de que un evento suceda o no, como:

$$P(Y_i = 1) = P_i$$

$$P(Y_i = 0) = 1 - P_i$$

entonces la función de probabilidad:

$$f_i(Y_i) = P_i^{Y_i} \times (1 - P_i)^{1-Y_i}, \qquad donde\ i = 1, 2, 3, \dots, k \tag{3}$$

Por lo cual la ecuación (3) denota la probabilidad que $Y_i = 1\ o\ 0$ y dado que cada candidato es independiente entonces la probabilidad conjunta al observar k valores de la variable Y, se enuncia como:

$$f(Y_1, Y_2, \cdots, Y_k) = \prod_{i=1}^{k} f_i(Y_i) = \prod_{i=1}^{k} P_i^{Y_i} \times (1 - P_i)^{1-Y_i} \tag{4}$$

Dicha probabilidad conjunta es llamada función de verosimilitud (FV). Y considerando el logaritmo de la ecuación (4):

$$\ln\big(f(Y_1, Y_2, \cdots, Y_k)\big) = \sum_{i=1}^{k} [Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)]$$

$$= \sum_{i=1}^{k} [Y_i \ln P_i - Y_i \ln(1 - Y_i) + \ln(1 - P_i)]$$

$$= \sum_{i=1}^{k} \left[Y_i \ln\left(\frac{P_i}{1 - P_i}\right)\right] + \sum_{i=1}^{k} \ln(1 - P_i) \quad (5)$$

Como lo expresado en ecuación (2), la probabilidad de que un solicitante sea buen o mal candidato se representa como:

$$P_i = \frac{\exp({X_i}^T \beta)}{1 - \exp({X_i}^T \beta)} \quad (6)$$

la ecuación (6) se puede expresar como:

$$1 - P_i = \frac{1}{1 - \exp({X_i}^T \beta)} \quad (7)$$

y de igual manera,

$$\ln\left(\frac{P_i}{1 - P_i}\right) = {X_i}^T \beta \quad (8)$$

Tomando las ecuaciones (7) y (8) y remplazadas en ecuación (5) se expresa el logaritmo de la FV:

$$\ln\big(f(Y_1, Y_2, \cdots, Y_k)\big) = \sum_{i=1}^{k} Y_i\big({X_i}^T \beta\big) - \sum_{i=1}^{k} \ln\big(1 + \exp\big({X_i}^T \beta\big)\big) \quad (9)$$

La ecuación (9) depende de los coeficientes β, ya que Y_i y X_i son conocidas.

El tratamiento de máxima verosimilitud busca maximizar la ecuación (9), puesto que busca la máxima capacidad predictiva. Por lo tanto se deriva parcialmente respecto a cada incógnita, es decir, respecto a cada β_j , con $j = 1, \dots, n$, encontrando n ecuaciones no lineales, las cuales

deberán solventarse mediante procedimientos numéricos.

Estimados los valores respectivos de β se validan que maximicen la FV, considerando la condición de maximización de segundo orden. Por consiguiente se consiguen los coeficientes para así estimar la probabilidad de incumplimiento de un candidato, partiendo de la ecuación (2).

Conceptos del proceso de modelización

En la implementación de los modelos scoring basados en regresión logística es importante conceptualizar algunos términos:

Test Dickey – Fuller aumentada

El test de Dickey Fuller aumentada (ADF) faculta validar la estacionariedad de una serie temporal. Si dicha serie no es estacionaria, ésta muestra al menos una raíz unitaria, por tanto, se indaga contrastar la hipótesis.

- Ho: *la serie temporal contiene raíz unitaria*.
- Ha: *la serie temporal es estacionaria*.

El criterio para rechazar la hipótesis nula será si el valor absoluto (estadístico de la prueba ADF) es mayor a los valores críticos de la prueba (considerando al1%, 5%, o 10%).

Valor de Información (IV)

Es aquel valor cuantitativo que permite medir la facultad de predicción de una variable independiente. El valor del IV funciona con variables categóricas. Se calcula mediante la siguiente ecuación:

$$IV = \sum_{i=1}^{n} \left(\frac{b_i}{b} - \frac{m_i}{m} \right) \times \ln \left(\frac{\frac{y_i}{y}}{\frac{m_i}{m}} \right) \quad (10)$$

Donde el número de categorías es el valor n, la variable independiente y_i y m_i el número de buenos y malos candidatos en la categoría i. Finalmente, y y m representan respectivamente la cantidad total de buenos y malos candidatos en el periodo de modelo.

Análogamente se comprende que mientras más pequeño sea el valor de información, menos valor predictivo tendrá la variable independiente categorizada. A pesar de ello las variables cuyo VI sean mayores al 0.5 deberían ser evaluadas pues puede darse el caso de sobre estimación, mayores al 0,3 se consideran con un nivel de predicción fuerte, entre 0,1 y 0,3 como predictores de nivel medio, entre 0,02 y 0,1 como predicción débil y menores a 0,02 como no predictores.

La guía propuesta para los valores aceptados del valor de información de variables es:

RANGO VALOR	**NIVEL DE PREDICCIÓN**
$IV < 0.02$	No predictivo
$0.02 \leq IV < 0.1$	Débil
$0.1 \leq IV < 0.3$	Medio
$IV \geq 0.3$	Fuerte

Backward Stepwise y criterio de información de Akaike (AIC)

La técnica conocida como de pasos en reversa, consiste en integrar al modelo todas las variables y posterior ir excluyendo una tras otra acorde mediante algún criterio de evaluación. En consecuencia el AIC se usa como un criterio de selección de variables, mismas que son consideradas aquellas que logren alcanzar el menor valor de AIC.

El AIC castiga modelos con demasiados parámetros e investiga establecer la significancia de incluir nuevos parámetros en el modelo. Se define como:

Sea un grupo de m modelos, tal que

$$M_j^m \supset M_{j-1}^{m-1} \supset \cdots \supset M_{j-1}^1 \tag{11}$$

Con, $j > i,\ i > 0,$ subíndices que denotan el número de variables en cada modelo. Se indaga en el seleccionar aquel valor j que minimice la ecuación:

$$\text{AIC} = -2\log L + 2\text{k} \tag{12}$$

donde, L es función de máxima verosimilitud que se formula como:

$$L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2}\sum_{t=1}^{n}\frac{e_t^2}{\sigma^2} \tag{13}$$

Siendo, k la cantidad de parámetros en el modelo, n la cantidad de registros para la construcción del modelo y σ^2 el promedio de residuales e_t elevado al cuadrado. Como se denota el AIC castiga modelos conforme la desviación de datos reales, considerando al mejor modelo aquel cuyo valor sea el más bajo entre todos los modelos comparados, siendo este particular el modelo que mejor se acopla a la realidad de los datos.

Estadístico Kolmogorov – Smirnov (KS)

Este estadístico consiste en medir diferencias de funciones de distribución de buenos y malos candidatos para cada puntaje o score. El valor de este estadístico se formula como:

$$KS = \ Max_i \ \left(P_b(i) \ - \ P_g \ (i)\right) \tag{14}$$

Dado:

i: como valor score en el rango $L \ - \ H$, con, $L \ \leq \ i \ \leq \ H$.

$P_g \ (i)$, $P_b(i)$: representan la proporción en la población de buenos candidatos (good) y malos candidatos (bad) con score menor o igual a i.

$$P_g(i) = \frac{N_g(i)}{N_g} = \sum_{j=L}^{i} P_g(j)$$

$$P_b(i) = \frac{N_b(i)}{N_b} = \sum_{j=L}^{i} P_b(j) \tag{15}$$

donde:

N_g: cantidad total de buenos candidatos en la población.

N_b: cantidad total de malos candidatos en la población.

$N_g(i)$, $N_b(i)$: cantidad de buenos y malos candidatos en la población con scores menor o igual a i.

$$N_g(i) = \sum_{j=L}^{i} n_g(j) \qquad , y$$

$$N_b(i) = \sum_{j=L}^{i} n_b(j) \tag{16}$$

$P_g(i)$, $P_b(i)$: como la proporción de buenos (good) candidatos con score i y malos (bad)

candidatos con score i en la población.

$$P_g(i) = \frac{n_g(i)}{N_g} P_b(i) = \frac{n_b(i)}{N_b} \tag{17}$$

Después de todo, $n_g(i)\ y\ n_b(i)$ son la cantidad de buenos y malos candidatos con score i, en la población.

Este estadístico mide la facultad de discriminación de un modelo, adoptando valores entre 0 y 1. Tomando en consideración que aquel modelo con estadístico KS menor al 20% puede ser cuestionado y mayor al 70% probablemente sea considerado muy bueno para ser cierto. La empresa crediticia líder en gestión de la información, TransUnion, cuantifica la certidumbre de calificaciones de riesgos acorde con los estándares internacionales que se detallan en el cuadro:

RANGO VALOR	**NIVEL DE PREDICCIÓN**
$KS < 0.20$	Modelo no funcional
$0.20 \leq KS \leq 0.40$	Satisfactoria
$0.41 \leq KS \leq 0.50$	Bueno
$0.51 \leq KS \leq 0.60$	Muy bueno
$KS > 0.60$	Extraordinario

Coeficiente GINI

Este estadístico es usado para cuantificar la distinción entre buenos y malos candidatos en el modelo ya que toma valores entre 0 y 1, considerando que, si el coeficiente Gini es igual a 1, quiere decir que el modelo agrupa eficientemente a buenos y malos candidatos.

Este estadístico está dado por:

$$Gini = 1 - \sum_{i=L}^{H'} \left(P_b(i+1) - P_b(i)\right)\left(P_g(i+1) + P_g(i)\right) \quad (18)$$

Donde:

i: valor de score, en el rango $L - H$, con, $L \leq i \leq H$.

$P_g(i)$, $P_b(i)$ son la proporción de buenos y malos candidatos en la población respectivamente con score menor o igual a i.

Se debe tomar en cuenta que en modelos scoring, un coeficiente de Gini inferior al 35% es sospechoso y mayor o igual al 50% se considera satisfactorio.

Matriz de confusión; error, sensibilidad y especificidad

El método para valorar la predicción que tiene un modelo, consiste en el cálculo de porcentaje de candidatos que este discrimina de forma adecuada. Este valor porcentual de clasificación proviene de la matriz de confusión, misma que es elaborada bajo el siguiente algoritmo:

- Elección del punto de corte para los valores de calificación obtenidos.

- Agrupar a los candidatos con un valor de calificación por debajo del punto de corte elegido, estos serán considerados como malos candidatos esperados. Por otro lado, los candidatos valorados sobre el punto de corte son considerados como buenos candidatos esperados.

- Elaborar una tabla cruzada entre el grupo de discriminación real de bueno o malo candidato y la tabla anterior de bueno o malo esperado.

- Calcular las diferentes ratios como error, valor de sensibilidad y especificidad que pueden ser obtenidos del modelo.

Las ocurrencias correctamente clasificadas son denominadas como verdaderos positivos (buenos) y verdaderos negativos (malos). Por otro lado, al no ser bien clasificados se tienen falsos

positivos (malos clasificados como buenos) y a los falsos negativos (buenos clasificados como malos).

Adicional se determina la sensibilidad y especificidad como aquella característica del modelo para catalogar adecuadamente al candidato bueno y malo. Se presenta la matriz de confusión en su forma general:

Matriz de Confusión		**Real**	
		0: Malo	1: Bueno
Estimado	0: Malo	A	B
	1 : Bueno	C	D

Se presenta en la tabla los indicadores de eficiencia, mismos que serán usados para comparar los dos modelos elaborados con base a la matriz de confusión.

Nombre de Indicador	**Definición**	**Ecuación**
Tasa de aciertos	Cociente del número de predicciones correctas entre el total.	$\frac{A+D}{A+B+C+D}$
Error	Cociente del número de predicciones incorrectas entre el total.	$\frac{B+C}{A+B+C+D}$
Sensibilidad	Cociente del número de buenos clasificados correctamente entre el total de buenos.	$\frac{D}{B+D}$
Especificidad	Cociente del número de malos clasificados correctamente entre el total de malos.	$\frac{A}{A+C}$

Una manera de encontrar el punto de corte que permita igualar la predicción correcta de buenos y malos es mediante el uso del Índice de Youden:

$$Índice\ de\ Youden\ (YI) = max(Sensibilidad + Especificidad - 1) \quad (19)$$

El cual en una curva ROC, misma que será definida en el siguiente apartado, el índice de Youden es la distancia vertical máxima entre la curva y la diagonal. Siendo el punto de corte óptimo, aquel en el cual se alcanza el valor de YI.

Capítulo II Metodología

Para alcanzar los objetivos de este trabajo, se presenta una metodología de desarrollo, misma que consta de 3 fases. La primera fase recopila la información obtenida de la primera convocatoria del programa Fonquito en la cual se evidencia variables de tipo categóricas y numéricas codificadas en el aplicativo Excel de Office. La segunda fase usa los datos importados desde Excel para ser tratados bajo método de extracción, transformación y carga, proceso que es desarrollado en el programa RStudio versión 4.2.1. Por último la tercera fase consta del uso de la data tratada para la búsqueda de variables significativas para la creación del modelo el cual tendrá pruebas para validar los resultados con datos reales del programa.

Fase de desarrollo

El programa Fondos de la Ciudad de Quito inició con la etapa de recepción de información y postulaciones de interesados desde mediados de abril hasta mediados del mes de junio de 2021. En esta etapa del programa se recolectó información de 2181 proyectos postulantes interesados en iniciar el proceso para la adjudicación de capital semilla.

Es oportuno para un mejor entendimiento de los datos recolectados el desarrollar una breve descripción de las variables en el marco de la primera convocatoria del programa Fonquito. Para la primera convocatoria se ha permitido registrar proyectos cuya personería constituida fueron para personas naturales, personas jurídicas y asociaciones de la Economía popular y Solidaria (EPS).

Para los proyectos de personería jurídica y EPS deben estar regularizados por el organismo del sector público correspondiente en base al tipo de actividad productiva o servicio en el que están

constituidas y bajo mismo lineamiento del proyecto con que postula.

Para los proyectos de personería natural así como requisitos comunes para todos los que apliquen a la convocatoria están; que los miembros del proyecto deben ser mayores de edad con residencia permanente en Ecuador y radicados o constituidos en el DMQ, acreditar experiencia o desarrollo del proyecto mismo que está limitado hasta 60 meses desde su gestación, estar en lista blanca con los organismos públicos como; el Servicio de Rentas Internas (SRI), Instituto Ecuatoriano de Seguridad Social (IESS) y el Servicio Nacional de Contratación Pública (SERCOP).

Los proyectos postulantes deben situarse en una de las dos etapas; como idea con prototipo o desarrollo inicial. Así mismo deben contar con componente de mejoras significativas respecto al de competencias en el mercado en el cual estará categorizado acorde a ocho tipos de categorías y en base a productos o servicios que el proyecto postulante ofrece.

Descripción de variables de la primera convocatoria del programa Fonquito

Para la recolección de información se ha creado un formulario en línea mediante 49 preguntas, las cuales fueron implementadas en las plataformas gratuitas de recolección de información como; Gust y KoBoToolBox.

En el siguiente cuadro se describen 37 de estas variables puesto que las restantes 12 son variables de tipo texto que contienen información personal del representante y descripción del proyecto. Estas variables fueron valoradas y sintetizadas objetivamente por el equipo técnico de la corporación.

Detalle del nombre de variables con su respectiva descripción y categorización:

N°	Variable	Descripción	Tipo variable
1	"tipo_postulante"	Tipo de personería; Natural, Jurídica o EPS del proyecto	Categórica
2	"rango_edad_actual"	Descripción del rango de edad del representante del proyecto	Categórica
3	"genero_identifica"	Género al cual se identifica el representante del proyecto; Femenino, Masculino, LGBTIQ+, Prefiero no decirlo.	Categórica
4	"grupo_poblacional_identifica_representante"	Grupo poblacional con que se identifica el representante; Afroecuatoriano, Blanco, Indígena, Mestizo, Montubio, Otro.	Categórica
5	"representante_pertenece_algún_pueblo_nacionalidad_indígena"	El representante afirma o no pertenecer a algún pueblo o nacionalidad indígena.	Categórica
6	"representante_posee_carnet_CONADIS"	El representante afirma o no poseer carnet del CONADIS.	Categórica
7	"Nacionalidad_representante"	El representante se identifica con nacionalidad; Ecuatoriana o Extranjera.	Categórica
8	"tiempo_se_encuentra_desarrollando_proyecto"	Especifica el tiempo que el proyecto se encuentra desarrollando; Menos de 1 año, Entre 1 y 5 años, Más de 5 años.	Categórica
9	"etapa_actual_proyecto"	Etapa en que se encuentra el proyecto; Desarrollo Inicial o Idea con prototipo.	Categórica
10	"categoria_alinea_proyecto"	Identificación del proyecto con las 8 posibles categorías clasificadas.	Categórica

11	"sub_categoria_alinea_proyecto"	Identificación del proyecto en las subcategorías de; Productos, Servicios, Productos y servicios.	Categórica
12	"recibido_recursos_económicos_programasdecapitalsemilla_gestionadospor_CONQUITO"	Proyecto afirma o no haber recibido capital semilla en programas anteriores ejecutados por CONQUITO.	Categórica
13	"cantidad_total_integrantes_proyecto_emprendimiento_incluido_representante"	Indica la cantidad numérica de integrantes que conforma el proyecto postulante.	Numérica
14	"cantidad_personas_genero_femenino"	Indica la cantidad numérica de personas de género femenino que conforma el proyecto postulante.	Numérica
15	"cantidad_personas_jovenes_18a29años"	Indica la cantidad numérica de jóvenes de 18 a 29 años que integrantes que conforma el proyecto postulante.	Numérica
16	"cantidad_personas_capacidades_diferentes"	Indica la cantidad numérica de personas con capacidades diferentes que conforma el proyecto postulante.	Numérica
17	"cantidad_adultos_mayores"	Indica la cantidad numérica de personas adultas mayores que conforma el proyecto postulante.	Numérica
18	"cantidad_personas_pertenecientes_pueblos_nacionalidades"	Indica la cantidad numérica de personas pertenecientes a pueblos o nacionalidades indígenas que conforma el proyecto postulante.	Numérica
19	"Breve_Descripcion_proyecto	Verificación del tipo de componente	Categórica

	_compotente_diferente"	diferenciador del proyecto; Por producto, Por proceso, Por modelo de gestión, No cumple.	
20	"uso_adecuado_de_los_3000"	Verificación si el proyecto hará uso adecuado o no del capital semilla de ser el caso que sea ganador del mismo.	Categórica
21	"Evidencia_congruente"	Representante justifica/sustenta con documentación o evidencia el estado del desarrollo del proyecto postulante. (Si, No)	Categórica
22	"Documento_de_identidad_Visa_cedula"	La representante adjunta o no un documento de identificación válido y actual. (Si, No)	Categórica
23	"representante_presenta_certificado_cumplimiento_obligaciones_tributarias_SRI"	Representante del proyecto adjunta o no certificado generado en el SRI. (Si, No)	Categórica
24	"representante_está_aldia_cumplimiento_obligaciones_tributarias_SRI"	Representante del proyecto adjunta certificado generado en el SRI mismo que se encuentra al día en obligaciones tributarias. (Si, No)	Categórica
25	"representante_presenta_certificado_denoser_contratista_incumplido_SERCOP"	Representante del proyecto adjunta o no certificado generado en la SERCOP. (Si, No)	Categórica
26	"representante_noconsta_contratista_incumplico_SERCOP"	Representante del proyecto adjunta certificado generado en la SERCOP mismo que se encuentra como no ser contratista incumplido con el estado. (Si, No)	Categórica

27	"representante_presenta_certificado_cumplimiento_obligaciones_patronales_IESS"	Representante del proyecto adjunta o no certificado generado en el IESS. (Si, No)	Categórica
28	"representante_notiene_obligaciones_patronales_pendientes_IESS"	Representante del proyecto adjunta certificado generado en el IESS mismo que se encuentra como no mantener obligaciones patronales. (Si, No)	Categórica
29	"utilizaralos_recursos_decapitalsemilla_adquirir_insumosoproductos_paraser_comercializados"	Verificación de que el proyecto planifica hacer uso del capital semilla acorde a los lineamientos del programa.	Categórica
30	"emprendimiento_sera_implementado_en_el_DMQ"	Verificación de que el proyecto se ejecutara e implementará en el DMQ.	Categórica
31	"proyecto_postulante_ha_recibido_capitalsemilla_convocatorias_anteriores"	Verificación en que el proyecto tenga histórico de haber recibido o no capital semilla en programas similares.	Categórica
32	"emprendimiento_aplica_alamisma_etapa_de_desarrollo_con_lacual_recibio_capital_semilla"	Verificación si el proyecto tras recibir capital semilla en otro proyecto difiere de esta postulación acorde a su etapa de desarrollo.	Categórica
33	"equipo_emprendedor_cuenta_conalmenos_dos_personas_en_sus_integrantes"	Verificación para conocer el número mínimo de integrantes requeridos en el proyecto acorde las bases del programa.	Categórica
34	"representante_del_proyecto_completa_la_totalidad_informacion_delformulario_postulacion"	Validación si el postulante completo o no con totalidad la información solicitada en el formulario de inscripción.	Categórica

35	"representante_del_proyecto_postula_mas_deunavez_en_la_convocatoria_con_el_mismo_emprendimiento"	Validación para conocer si el mismo proyecto ha sido postulado mas de una vez por el mismo representante.	Categórica
36	"documentacion_legible"	Verificación si la documentación adjunta es legible y correctamente emitida por las entidades públicas de control.	Categórica
37	"APROBACION_CAPITAL_SEMILLA"	Variable binaria que determina si el proyecto es adjudicado de capital semilla o no.	Categórica

Recolección de información

La información consolidada llega a un total de 1842 registros de proyectos postulantes, siendo para el caso 339 registros retirados debido a incumplimientos por causales de rechazo conforme lo establecido en las bases de la primera convocatoria. Estos registros, los cuales son parte de estudio, cuentan con información de 49 preguntas realizadas a los postulantes, mismas que recolectan información del proyecto, equipo emprendedor, representante del proyecto e información complementaria adjunta.

La data fue creada, diseñada y supervisada desde el aplicativo Excel de Office 365, la misma fue alimentada por las plataformas de recolección de información como; Gust (https://gust.com/) y KoboToolBox (https://www.kobotoolbox.org/). Mediante programación en la nube de Microsoft

Automated los registros de ambas plataformas fueron enviados a una sola matriz en la nube, posterior a ello tras el cierre de etapa de postulación, inició la intervención del equipo técnico de la corporación Conquito para desarrollo y validación de información.

Una vez terminada la intervención y supervisión de la información estructurada, se realizó un cierre de acceso a la información, aislando a la data consolidada en el área de emprendimiento bajo tutela de técnicos de la corporación. Esta información fue cargada al software estadístico Rstudio donde se da inicio al proceso de limpieza y tratamiento de datos.

Proceso ETL

El archivo Excel denominado como “fq21.xlsx” se cargó en RStudio versión 4.2.1 (2022-06-23 ucrt) con el uso de un computador portátil con procesador de 1.8GHz (Gigahertz) y opción de overlock , memoria aleatoria RAM de 20 GB(Gigabytes) y disco en estado sólido.

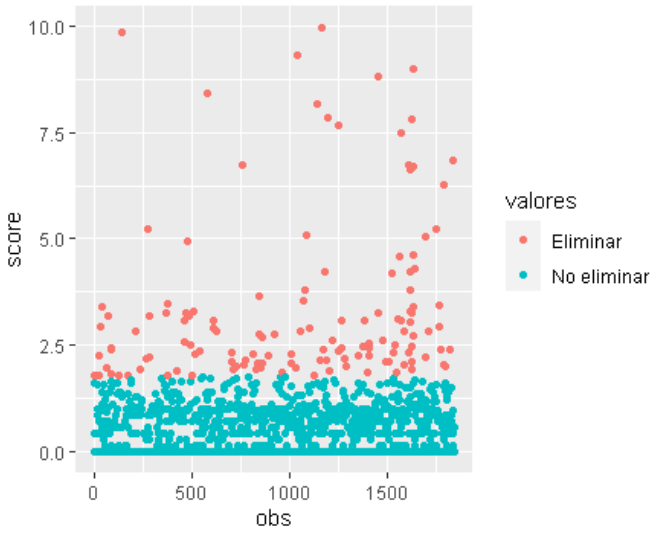
El primer paso consistió en la revisión y corrección de los nombres de niveles de las variables de tipo categóricas ya que tras la ejecución de la función ‘table()’ se evidenció variables en cuyos niveles existían diferencias de tipeo, tal como se muestra en el siguiente ejemplo:

```
> table(data$tipo_postulante)
Asociaciones EPS Persona jurídica Persona Jurídica  Persona natural  Persona Natural
            26               8              80              121             1607
```

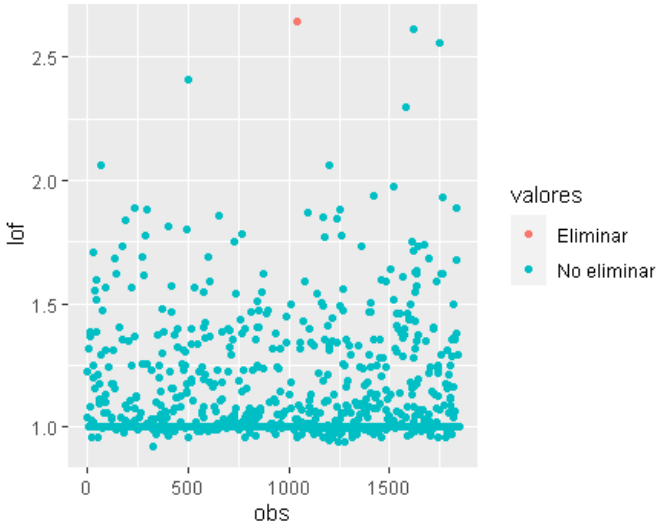
Posterior se evaluó la existencia de variables constantes mediante la función ‘constante()’ la

cual al ejecutar se evidenció no existen variables con valores constantes. De la misma manera, se valida si existen registros o campos con valores “NA” mediante la función ‘porcNA()’, esto debido a datos faltantes que pudieran existir, sin embargo, se evidenció que los valores porcentuales de campos vacíos son casi nulos con valores menores al 0.02% en cada variable.

Finalmente se realiza revisión de datos atípicos u outliers, mediante la separación de variables cuantitativas mismas que se ha tomado como cota el percentil 0,975, con el cual se ha encontrado 132 registros atípicos como se muestran:



En la gráfica se especifica los registros atípicos que se requieren eliminar o retirar de la data. Definiendo los límites para estos valores a eliminar se procede a reestructurar la data retirando estos 132 registros con lo que se obtiene el resultado gráfico siguiente:



Luego del proceso de retiro se ha concretado 1710 observaciones en la data del presente estudio.

Análisis de significancia de variables

Luego del proceso ETL de la información consolidada se procede a discriminar variables de tipo numéricas y de tipo categóricas. Las variables de tipo categóricas fueron procesadas por la función 'TestVI()' misma que nos muestra el nivel de predicción que estas representan en el modelo de estudio.

La función del valor de información resultante llamada como "dVI" en el script, determina las siguientes variables como candidatas para la predicción:

```
> dVI
                                                                                Variable          VI
1           emprendimiento_aplica_alamisma_etapa_de_desarrollo_con_lacual_recibio_capital_semilla 10.538452148
2  representante_del_proyecto_postula_mas_deunavez_en_la_convocatoria_con_el_mismo_emprendimiento 10.538452148
3                       proyecto_postulante_ha_recibido_capitalsemilla_convocatorias_anteriores 10.439883782
4          representante_del_proyecto_completa_la_totalidad_informacion_delformulario_postulacion 10.073383653
5                            equipo_emprendedor_cuenta_conalmenos_dos_personas_en_sus_integrantes  9.999399186
6                                                 emprendimiento_sera_implementado_en_el_DMQ  9.928491752
7       utilizaralos_recursos_decapitalsemilla_adquirir_insumosoproductos_paraser_comercializados  9.497273936
8                              representante_está_aldia_cumplimiento_obligaciones_tributarias_SRI  6.799190874
9                                                    Documento_de_identidad_Visa_cedula  6.791307050
10                  representante_presenta_certificado_cumplimiento_obligaciones_tributarias_SRI  6.705332094
11                                representante_notiene_obligaciones_patronales_pendientes_IESS  6.697582564
12                  representante_presenta_certificado_cumplimiento_obligaciones_patronales_IESS  6.643635536
13                                                        documentacion_legible  6.613039368
14                      representante_presenta_certificado_denoser_contratista_incumplido_SERCOP  6.597802838
15                                    representante_noconsta_contratista_incumplico_SERCOP  6.597802838
16                                         Breve_Descripcion_proyecto_compotente_diferente  2.520107844
17                                                          Evidencia_congruente  0.702088634
18                                                       uso_adecuado_de_los_3000  0.447224217
19                                                    categoria_alinea_proyecto  0.142404304
20                                                 sub_categoria_alinea_proyecto  0.136708635
21                                           tiempo_se_encuentra_desarrollando_proyecto  0.118140984
22                                                              rango_edad_actual  0.049491839
23                                                                tipo_postulante  0.042263952
24                                                       etapa_actual_proyecto  0.028101352
25                                           grupo_poblacional_identifica_representante  0.027151643
26                                                              genero_identifica  0.024102895
27                                 representante_pertenece_algún_pueblo_nacionalidad_indigena  0.010389402
28                                                    Nacionalidad_representante  0.009779508
29                                                 representante_posee_carnet_CONADIS  0.005249885
30               recibido_recursos_económicos_programasdecapitalsemilla_gestionadospor_CONQUITO  0.004942919
```

Como se detallan en la gráfica que antecede, existen 18 variables cuyo valor de información determina que tienen un nivel de predicción fuerte (aquellas mayores a 0.3).

Las variables de tipo numéricas fueron procesadas por la función ‘KS()’ la cual el valor resultante llamado “dKS” determina las siguientes variables para la predicción:

```
> dKS
                                                                Variable     KS
1 cantidad_total_integrantes_proyecto_emprendimiento_incluido_representante 0.0874
2                                                cantidad_adultos_mayores 0.0802
3                                          cantidad_personas_genero_femenino 0.0679
4                                     cantidad_personas_capacidades_diferentes 0.0329
5                                          cantidad_personas_jovenes_18a29años 0.0314
6                        cantidad_personas_pertenecientes_pueblos_nacionalidades 0.0177
>
```

Se observa que las variables cuyo valor estadístico KS no supera el mínimo estandarizado (20%), podrían no reflejar significancia en el modelo predictivo.

Debido los resultados de las funciones ‘KS()’ y ‘VI()’ implementadas en el script, surge la necesidad de usar una función más robusta que permita mostrar el valor de información para todas la variables candidatas a ser consideradas como predictivas.

Se considera la librería *Information* de RStudio en la que se hace uso de la función ‘create_infotables()’ ya que esta devuelve el valor de información para todas las variables de la

data. Para tal efecto la función nos devuelve los siguientes valores:

	Variable	IV
32	emprendimiento_aplica_alamisma_etapa_de_desarrollo_con_lacual_recibio_capital_semilla	3.776163478
35	representante_del_proyecto_postula_mas_deunavez_en_la_convocatoria_con_el_mismo_emprendimiento	3.776163478
31	proyecto_postulante_ha_recibido_capitalsemilla_convocatorias_anteriores	3.686361205
34	representante_del_proyecto_completa_la_totalidad_informacion_delformulario_postulacion	3.408059351
33	equipo_emprendedor_cuenta_conalmenos_dos_personas_en_sus_integrantes	3.342809567
30	emprendimiento_sera_implementado_en_el_DMQ	3.281135665
29	utilizaralos_recursos_decapitalsemilla_adquirir_insumosoproductos_paraser_comercializados	2.953264370
19	Breve_Descripcion_proyecto_compotente_diferente	2.508783011
22	Documento_de_identidad_Visa_cedula	1.367124937
24	representante_está_aldia_cumplimiento_obligaciones_tributarias_SRI	1.367124937
23	representante_presenta_certificado_cumplimiento_obligaciones_tributarias_SRI	1.326590878
28	representante_notiene_obligaciones_patronales_pendientes_IESS	1.319384341
27	representante_presenta_certificado_cumplimiento_obligaciones_patronales_IESS	1.298055079
36	documentacion_legible	1.287550330
25	representante_presenta_certificado_denoser_contratista_incumplido_SERCOP	1.280605144
26	representante_noconsta_contratista_incumplico_SERCOP	1.280605144
21	Evidencia_congruente	0.671339177
20	uso_adecuado_de_los_3000	0.424166617
8	tiempo_se_encuentra_desarrollando_proyecto	0.122658437
10	categoria_alinea_proyecto	0.106265470
11	sub_categoria_alinea_proyecto	0.104737710
15	cantidad_personas_jovenes_18a29años	0.094418963
2	rango_edad_actual	0.054419822
13	cantidad_total_integrantes_proyecto_emprendimiento_incluido_representante	0.046310313
17	cantidad_adultos_mayores	0.046204357
3	genero_identifica	0.028763713
14	cantidad_personas_genero_femenino	0.024279001
9	etapa_actual_proyecto	0.014223927
16	cantidad_personas_capacidades_diferentes	0.013133552
5	representante_pertenece_algún_pueblo_nacionalidad_indigena	0.011758438
4	grupo_poblacional_identifica_representante	0.008716680
7	Nacionalidad_representante	0.007700574
12	recibido_recursos_económicos_programasdecapitalsemilla_gestionadospor_CONQUITO	0.006295153
1	tipo_postulante	0.004491644
6	representante_posee_carnet_CONADIS	0.004016723
18	cantidad_personas_pertenecientes_pueblos_nacionalidades	0.003028230

El grafico anterior nos muestra el valor de información de las 36 variables que participan en la data, encontrando así variables no predictivas cuyo valor de información son menor al 0,02, así como aquellas que superan el 0,3 y 0,5 lo cual sugeriría existe un sobreajuste.

Formulación e implementación del modelo

Es necesario considerar un subconjunto de los datos para entrenar el modelo y otro subconjunto para testearlo. Así, también tomando en cuenta a dicha partición de manera que los datos sean balanceados en ambos subconjuntos tanto en el conjunto de datos de entrenamiento y el conjunto de testeo.

Observamos que la proporción de datos es desbalanceada debido a que de los 1.710 registros se

conoce que 1.604 fueron considerados como malos candidatos de capital semilla, mientras que los 106 restantes se consideran como buenos candidatos.

Ahora observemos la cantidad porcentual que representan los candidatos en la data:

```
> table(data$APROBACION_CAPITAL_SEMILLA) %>% prop.table()
       No        Si
0.94245385 0.05754615
```

Como se aprecia la cantidad porcentual de buenos candidatos es del 5,75% mientras que los candidatos malos alcanzan al 94,25%. Esto nos sugiere realizar un rebalanceo de los datos para los conjuntos de data de entrenamiento y testeo.

Posterior a ello se realizará pruebas de modelos con datos balanceados y no balanceados hasta determinar aquel que sea considerado como modelo idóneo bajo comparación de parámetros.

Modelo lineal generalizado

Para la construcción del modelo se realizó pruebas con datos balanceados y no balanceados, posterior se comparó los parámetros que determinan la efectividad entre los mismos. Estos parámetros son; el criterio de información de Akaike (AIC), el porcentaje de representación de los datos en el modelo (R^2), valor de precisión predictiva (ACC), valor de distinción entre buenos y malos candidatos (GINI) y el valor numérico que tiene el modelo para catalogar a candidatos buenos

y malos.

El primer modelo propuesto llamado *"model_1"* se realizó considerando el grupo de datos de entrenamiento balanceados e integrando todas las variables de la data.

Para tal efecto se tiene el siguiente resultado de la significancia de variables en el modelo:

	Pr(>|z|)
(Intercept)	<2e-16 ***
tipo_postulantePersona Jurídica	<2e-16 ***
tipo_postulantePersona Natural	<2e-16 ***
rango_edad_actual30 a 39 años	<2e-16 ***
rango_edad_actual40 a 49 años	<2e-16 ***
rango_edad_actual50 a 59 años	<2e-16 ***
rango_edad_actual60 a 64 años	<2e-16 ***
rango_edad_actualMás de 65 años	<2e-16 ***
genero_identificaLGBTIQ+	<2e-16 ***
genero_identificaMasculino	<2e-16 ***
genero_identificaPrefiero no decirlo	<2e-16 ***
grupo_poblacional_identifica_representanteBlanco	<2e-16 ***
grupo_poblacional_identifica_representanteIndígena	<2e-16 ***
grupo_poblacional_identifica_representanteMestizo	<2e-16 ***
grupo_poblacional_identifica_representanteMontubio	<2e-16 ***
grupo_poblacional_identifica_representanteOtro	<2e-16 ***
representante_pertenece_algún_pueblo_nacionalidad_indígenaSi	<2e-16 ***
representante_posee_carnet_CONADISSi	<2e-16 ***
Nacionalidad_representanteExtranjera	<2e-16 ***
tiempo_se_encuentra_desarrollando_proyectoMás de 5 años	<2e-16 ***
tiempo_se_encuentra_desarrollando_proyectoMenos de 1 año	<2e-16 ***
etapa_actual_proyectoIdea con prototipo	<2e-16 ***
categoria_alinea_proyectoAgricultura	<2e-16 ***
categoria_alinea_proyectoAlimentos frescos o procesados y bebidas	<2e-16 ***

categoria_alinea_proyectoCiencias de la vida y educación	<2e-16 ***
categoria_alinea_proyectoConfecciones textiles o calzado	<2e-16 ***
categoria_alinea_proyectoConstrucción	<2e-16 ***
categoria_alinea_proyectoSalud y bienestar	<2e-16 ***
categoria_alinea_proyectoTransporte y logística	<2e-16 ***
categoria_alinea_proyectoTurismo y cultura	<2e-16 ***
sub_categoria_alinea_proyectoProductos	<2e-16 ***
sub_categoria_alinea_proyectoProductos y servicios	<2e-16 ***

Los parámetros para este primer modelo con 31 variables fueron; el valor de penalización del modelo 904,96, el valor de representación predictiva del modelo con los datos (R^2) fue de 0,99, el valor de la precisión predictiva para discriminar entre buenos y malos candidatos (ACC) es 0,9908, la habilidad del modelo de catalogar correctamente al cliente bueno y malo son 1 y 0,9811 respectivamente y el estadístico usado para medir cuan bien el modelo scoring distingue entre los buenos y malos clientes (GINI) es de 0.9816.

El primero modelo parece ser el mejor candidato pues su representatividad de datos, el cual es cercano al 100%, así como la eficacia predictiva para discriminar y su valor GINI. Dichos parámetros sugieren que existe un sobre ajuste del modelo con lo cual equívocamente podría interpretarse como el mejor modelo propuesto.

El segundo modelo llamado *"model_2"* fue elaborado considerando todas las variables de la data de entrenamiento junto con datos no balanceados. Encontrando así el siguiente resultado de la significancia de variables en el modelo:

	Pr(>|z|)
(Intercept)	<2e-16 ***
tipo_postulantePersona Jurídica	<2e-16 ***
tipo_postulantePersona Natural	<2e-16 ***
rango_edad_actual30 a 39 años	<2e-16 ***
rango_edad_actual40 a 49 años	<2e-16 ***
rango_edad_actual50 a 59 años	<2e-16 ***
rango_edad_actual60 a 64 años	<2e-16 ***
rango_edad_actualMás de 65 años	<2e-16 ***
genero_identificaLGBTIQ+	<2e-16 ***
genero_identificaMasculino	<2e-16 ***
genero_identificaPrefiero no decirlo	<2e-16 ***
grupo_poblacional_identifica_representanteBlanco	<2e-16 ***
grupo_poblacional_identifica_representanteIndígena	<2e-16 ***
grupo_poblacional_identifica_representanteMestizo	<2e-16 ***
grupo_poblacional_identifica_representanteMontubio	<2e-16 ***
grupo_poblacional_identifica_representanteOtro	<2e-16 ***
representante_pertenece_algún_pueblo_nacionalidad_indígenaSi	<2e-16 ***
representante_posee_carnet_CONADISSi	<2e-16 ***
Nacionalidad_representanteExtranjera	<2e-16 ***
tiempo_se_encuentra_desarrollando_proyectoMás de 5 años	<2e-16 ***
tiempo_se_encuentra_desarrollando_proyectoMenos de 1 año	<2e-16 ***
etapa_actual_proyectoIdea con prototipo	<2e-16 ***
categoria_alinea_proyectoAgricultura	<2e-16 ***
categoria_alinea_proyectoAlimentos frescos o procesados y bebidas	<2e-16 ***
categoria_alinea_proyectoCiencias de la vida y educación	<2e-16 ***
categoria_alinea_proyectoConfecciones textiles o calzado	<2e-16 ***
categoria_alinea_proyectoConstrucción	<2e-16 ***
categoria_alinea_proyectoSalud y bienestar	<2e-16 ***
categoria_alinea_proyectoTransporte y logística	<2e-16 ***
categoria_alinea_proyectoTurismo y cultura	<2e-16 ***

sub_categoria_alinea_proyectoProductos <2e-16 ***

sub_categoria_alinea_proyectoProductos y servicios <2e-16 ***

sub_categoria_alinea_proyectoServicios <2e-16 ***

recibido_recursos_económicos_programasdecapitalsemilla_gestion<2e-16 ***

cantidad_total_integrantes_proyecto_emprendimiento_incluido_re<2e-16 ***

cantidad_personas_genero_femenino <2e-16 ***

cantidad_personas_jovenes_18a29años <2e-16 ***

cantidad_personas_capacidades_diferentes <2e-16 ***

cantidad_adultos_mayores <2e-16 ***

cantidad_personas_pertenecientes_pueblos_nacionalidades <2e-16 ***

Breve_Descripcion_proyecto_compotente_diferentePor Modelo de G<2e-16 ***

Breve_Descripcion_proyecto_compotente_diferentePor Proceso <2e-16 ***

Breve_Descripcion_proyecto_compotente_diferentePor Producto <2e-16 ***

uso_adecuado_de_los_3000Si <2e-16 ***

Evidencia_congruenteSi <2e-16 ***

Documento_de_identidad_Visa_cedulaSi <2e-16 ***

representante_presenta_certificado_cumplimiento_obligaciones_t<2e-16 ***

representante_está_aldia_cumplimiento_obligaciones_tributarias <2e-16 ***

representante_presenta_certificado_denoser_contratista_incumpl <2e-16 ***

representante_noconsta_contratista_incumplico_SERCOPSi <2e-16 ***

representante_presenta_certificado_cumplimiento_obligaciones_pa<2e-16 ***

representante_notiene_obligaciones_patronales_pendientes_IESSSi<2e-16 ***

utilizaralos_recursos_decapitalsemilla_adquirir_insumosoproduct<2e-16 ***

emprendimiento_sera_implementado_en_el_DMQSi <2e-16 ***

proyecto_postulante_ha_recibido_capitalsemilla_convocatorias_an<2e-16 ***

emprendimiento_aplica_alamisma_etapa_de_desarrollo_con_lacual_r<2e-16 ***

equipo_emprendedor_cuenta_conalmenos_dos_personas_en_sus_integr<2e-16 ***

representante_del_proyecto_completa_la_totalidad_informacion_de<2e-16 ***

representante_del_proyecto_postula_mas_deunavez_en_la_convocato<2e-16 ***

documentacion_legibleSi <2e-16 ***

Para este segundo modelo con 59 variables los parámetros fueron; el valor de penalización de 1.125,2, el valor de representación predictivo del modelo con los datos (R^2) de 0,98, el valor de la precisión predictiva para discriminar entre buenos y malos candidatos (ACC) es 0,9883, la habilidad del modelo de catalogar correctamente al candidato bueno y malo son 1 y 0,8353 respectivamente y el estadístico para medir cuan bien el modelo scoring discrimina entre los buenos y malos candidatos o GINI es de 0.9766.

En el modelo "model_2" se evidencia que el valor de penalización es mucho mayor que el "model_1" debido a la integración de una cantidad mayor de variables, la representatividad de los datos no varía mucho respecto al primer modelo, lo mismo se evidencia con la precisión predictiva para discriminar y difiere, tanto en la habilidad del modelo para catalogar correctamente al candidato malo, como en el GINI.

El siguiente cuadro se detalla la comparación paramétrica de los dos modelos:

Modelo \ Parámetros	**AIC**	R^2	**ACC**	**GINI**
"model_1"	904,96	*0,99*	0,9908	0.9816
"model_2"	1.125,2	*0,98*	0,9883	0.9766

Como los modelos previos se evaluaron en la data de entrenamiento se procedió a implementar tanto model_1 y model_2 en data testeo para verificar los cambios que puedan producirse.

Para el primer modelo se realizó la predicción usando la data test con datos balanceados, obteniendo así que el modelo presenta una precisión predictiva de 0,9142, una representación predictiva del modelo con los datos de 0,9142 y el estadístico usado para medir cuan bien el modelo scoring discrimina entre los buenos y malos candidatos de 0,8284.

En el segundo modelo se testeó la predicción usando la data test con datos sin balancear, encontrándose una precisión predictiva de 1, representación predictiva del modelo con los datos de

1 y estadístico GINI de 1. Lo anterior encontrado muestra que existe un sobre ajuste ocasionando que el modelo erróneamente sea considerado como uno totalmente perfecto.

En el siguiente cuadro se detalla los parámetros del modelo testeado en la data test:

Modelo \ Parámetros	R^2	**ACC**	**GINI**
"model_1"	*0,9142*	0,91	0.8284
"model_2"	*1*	1	1

Ventajas y desventajas Modelo 1:

Con lo expuesto en los cuadros comparativos se ha encontrado que el modelo model_1 es el que muestra mejores resultados acorde a los parámetros como; la representatividad de los datos en el modelo para ambas datas se mantiene por encima del 91%, la precisión predictiva del modelo mantiene un valor aceptable (superior a 0,9) y el valor para medir la distinción del modelo sobre los candidatos supera valores aceptables (mayor a 0,8). Lo cual el modelo en mención al estructurarse con datos balanceados muestra una mejor predicción que el modelo 2.

Por otro lado hay que considerar que este modelo integra todas las variables de la base de datos balanceada, obviando el valor de información que aporta independientemente cada una de las mencionadas y dejando a un lado el proceso de cálculo del valor de información (VI).

Ventajas y desventajas Modelo 2:

El modelo model_2 sigue por debajo los valores paramétricos estudiados del modelo 1, esto únicamente para datos sin balancear en la data de entrenamiento. Sin embargo al modelar con la data testeo se evidenció que se consiguen predicciones erróneas y que pueden mal interpretarse ya que las mismas son optimistas y sobre estiman la realidad.

De manera análoga en este modelo se ha considerado todas las variables de la data por lo cual su valor de penalización es mucho mayor que en el modelo 1, debido a la cantidad de variables

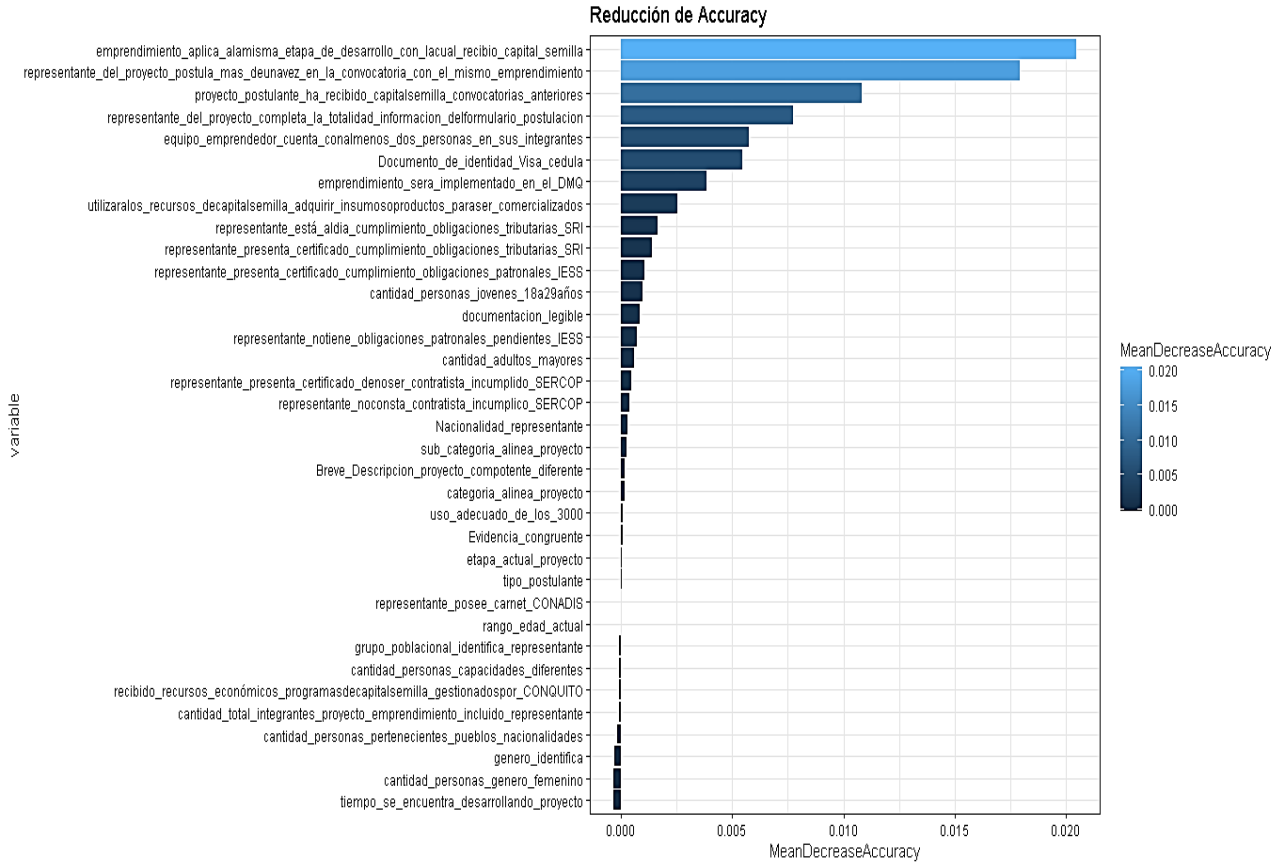
integradas en el mismo.

El tercer modelo llamado *"model_3"* es construido considerando los datos de entrenamiento sin balancear y las variables con mayor valor de información presentados en la sección 2.1.3, con lo que se evidencian parámetros como; el valor del AIC de penalización de 2124,4, el valor de representación predictivo del modelo con los datos (R^2) de 0,97, el valor de la precisión predictiva para discriminar entre buenos y malos candidatos (ACC) de 0,9766, la habilidad del modelo para catalogar correctamente al cliente bueno y malo son 0,9807 y 0,8909 respectivamente y el estadístico para medir cuan bien el modelo discrimina entre los buenos y malos clientes de 0.9532.

Bajo este último modelo y conocido el valor de los parámetros de los anteriores, podemos evidenciar que existe la apertura de ejecutar el criterio Backward Stepwise, con lo cual, se propone un nuevo modelo el cual se realizó considerando el valor de información (VI) y criterio de información de Akaike (AIC).

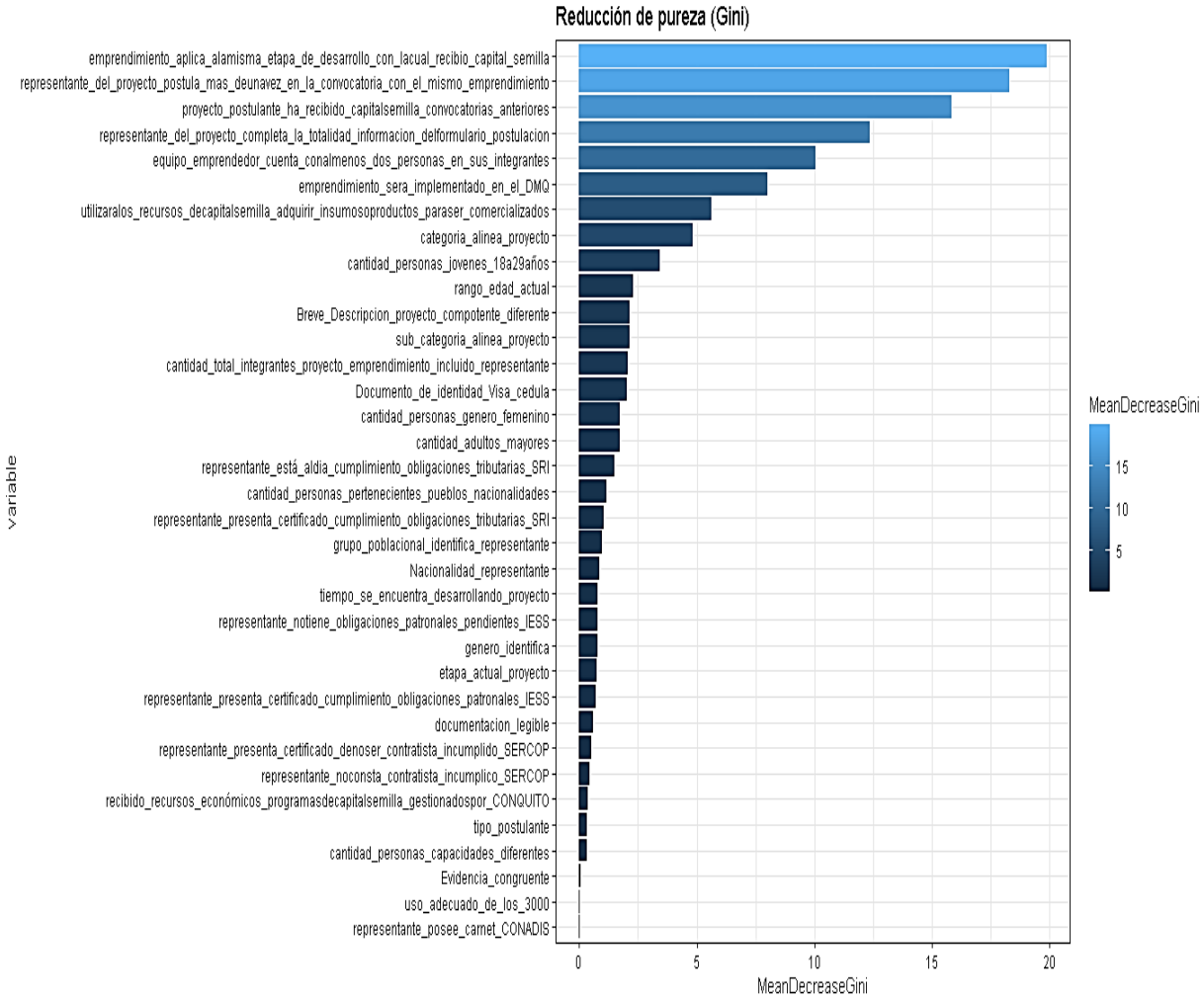
Modelo scoring tipo logit

El presente modelo fue testeado con base al modelo "model" y bajo simulaciones en las que se logra estabilizar una adecuada reducción en la precisión del modelo como se muestra en la siguiente gráfica donde se detallan las variables que generan un mejor accuracy:



El gráfico anterior muestra las variables que generan una mejor precisión al modelo bajo orden decreciente.

El gráfico de reducción de pureza del coeficiente GINI nos muestra de manera decreciente las variables que permite medir al modelo eficientemente en la distinción de buenos y malos candidatos.



Ambos análisis muestran que las variables listadas a continuación evidencian una influencia alta sobre las probabilidades de beneficiarios de capital semilla.

Variable:
tipo_postulante
rango_edad_actual
genero_identifica
grupo_poblacional_identifica_representante
representante_posee_carnet_CONADIS

Nacionalidad_representante
tiempo_se_encuentra_desarrollando_proyecto
etapa_actual_proyecto
categoria_alinea_proyecto
sub_categoria_alinea_proyecto
recibido_recursos_económicos_programasdecapitalsemilla_gestionadospor_CONQUITO
cantidad_total_integrantes_proyecto_emprendimiento_incluido_representante
cantidad_personas_genero_femenino
cantidad_personas_jovenes_18a29años
cantidad_personas_capacidades_diferentes
cantidad_adultos_mayores
cantidad_personas_pertenecientes_pueblos_nacionalidades
Breve_Descripcion_proyecto_compotente_diferente
uso_adecuado_de_los_3000
Evidencia_congruente
emprendimiento_aplica_alamisma_etapa_de_desarrollo_con_lacual_recibio_capital_semilla
representante_del_proyecto_completa_la_totalidad_informacion_delformulario_postulacion
documentacion_legible

El modelo propuesto bajo las variables listadas en la tabla anterior arroja los parámetros como el valor del AIC de penalización con 812,87, el valor de representación predictivo del modelo con los datos (R^2) es de 0,99, el valor de la precisión predictiva para discriminar entre buenos y malos candidatos (ACC) es 0,9916, la habilidad del modelo de discriminar correctamente al candidato bueno y malo son 0,9956 y 0,9296 respectivamente y el estadístico para medir cuan bien el modelo scoring discrimina entre los buenos y malos clientes es de 0.9832.

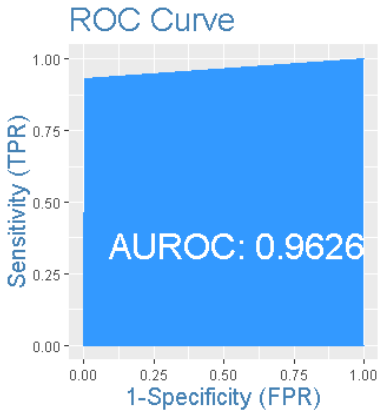
Este nuevo modelo propuesto fue puesto a prueba en la data de testeo para conocer su comportamiento al momento de evaluar sus predicciones, con lo que se ha obtenido que la precisión del modelo no vario mucho siendo un ACC de 0,978 y la sensibilidad empieza a verse un poco afectada en contraste con la especificidad, para el caso, se ha obtenido una sensibilidad de 0.9796 y especificad de 0.7917. El modelo mantiene un valor GINI de 0.9416.

En el siguiente cuadro se muestra la comparación de parámetros del modelo "model" aplicado en la data de entrenamiento y testeo:

Modelo \ Parámetros	**AIC**	R^2	**ACC**	**GINI**
model (entrenamiento)	812,87	*0,99*	0,9916	0.9832
model (testeo)	90	*0,97*	0,9708	0,9416

Dado lo mostrado en el cuadro anterior concluimos que el modelo llamado "model" es el mejor modelo pues la predicción de este tanto para la data de entrenamiento como de testeo muestran parámetros con pocas variaciones en representatividad de datos, en precisión y en distinción de buenos y malos candidatos.

Valor Auroc Modelo propuesto (model)

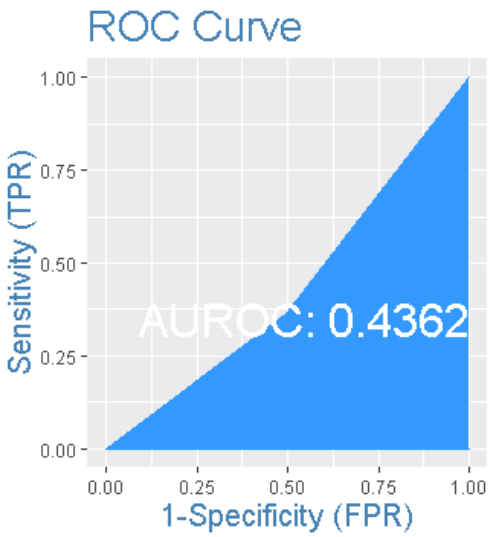


Elaboración: Propia

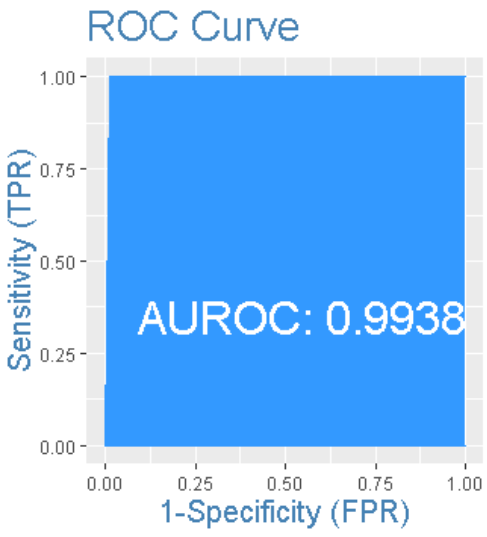
Como se observa el valor para medir la eficacia predictiva de un modelo, evaluando gráficamente la capacidad de este para discriminar entre buenos y malos candidatos es mediante el área bajo la curva de ROC (AUROC) el cual puede variar entre 0 y 1. Para el modelo construido se ha obtenido un valor de 0,9626, el cual bajo los criterios en los que; si el valor es igual a 1 las predicciones son perfectamente correctas, por otro lado si es 0 implica que las mismas son erróneas y un valor 0,5 indicaría que la clasificación se realiza de manera aleatoria.

Las curvas de los modelos *model_1* y *model_2* indican valores AUROC de 0.4362 y 0.9938 respectivamente, como se muestran en las gráficas siguientes:

Valor Auroc Modelo 1 (model_1)



Valor Auroc Modelo 2 (model_2)



Elaboración: Propia

Podría considerarse al *model2* como el modelo con la mejor clasificación de buenos y malos, sin embargo, recordemos que dicho modelo tiene un valor de penalización mucho mayor que la del modelo que se ha propuesto.

Dado que el modelo llamado “model” cumple con los mejores parámetros que los otros evaluados, se expone la siguiente regresión lineal de tipo *logit*:

$$Y_i = -2{,}971 - 0{,}146X_{1ai} - 0{,}526X_{1bi} + 0{,}144X_{2ai} + 0{,}175X_{2bi} - 0{,}662X_{2ci} + 0{,}515X_{2di} + 0{,}489X_{2ei} + 0{,}406X_{3ai} + 0{,}0467X_{3bi} + 0{,}382X_{3ci} - 1{,}014X_{4ai} - 0{,}594X_{4bi} - 0{,}404X_{4ci} - 0{,}619X_{4di} - 2{,}142X_{4ei} - 0{,}125X_{5i} - 0{,}216X_{6i} + 0{,}235X_{7ai} + 0{,}0581X_{7bi} - 0{,}0538X_{8i} + 0{,}508X_{9ai} + 0{,}658X_{9bi} + 1{,}591X_{9ci} + 0{,}746X_{9di} + 1{,}032X_{9ei} + 0{,}749X_{9fi} + 1{,}381X_{9gi} + 1{,}089X_{9hi} + 0{,}441X_{10ai} + 0{,}488X_{10bi} - 0{,}138X_{11i} - 0{,}046X_{12i} + 0{,}92X_{13i} - 0{,}146X_{14i} + 0{,}113X_{15i} - 0{,}099X_{16i} - 0{,}145X_{17i} - 0{,}901X_{18ai} - 0{,}496X_{18bi} - 0{,}672X_{18ci} + 0{,}286X_{19i} + 1{,}209X_{20i} + 2{,}886X_{21i} - 1{,}256X_{22i} + 0{,}633X_{23i}$$

Donde se conoce las variables:

X_{1ai}: será igual a 1 si tipo de postulante es persona jurídica y 0 caso contrario.

X_{1bi}: será igual a 1 si tipo de postulante es persona natural y 0 caso contrario.

X_{2ai}: será igual a 1 si rango de edad de postulante es de 30 a 39 años y 0 caso contrario.

X_{2bi}: será igual a 1 si rango de edad de postulante es de 40 a 49 años y 0 caso contrario.

X_{2ci}: será igual a 1 si rango de edad de postulante es de 50 a 59 años y 0 caso contrario.

X_{2di}: será igual a 1 si rango de edad de postulante es de 60 a 64 años y 0 caso contrario.

X_{2ei}: será igual a 1 si rango de edad de postulante es más de 65 años y 0 caso contrario.

X_{3ai}: será igual a 1 si género que se identifica postulante es LGBTIQ+ y 0 caso contrario.

X_{3bi}: será igual a 1 si género que se identifica postulante es masculino y 0 caso contrario.

X_{3ci}: será igual a 1 si postulante prefiere no especificar y 0 caso contrario.

X_{4ai}: será igual a 1 si grupo poblacional de postulante es blanco y 0 caso contrario.

X_{4bi}: será igual a 1 si grupo poblacional de postulante es indígena y 0 caso contrario.

X_{4ci}: será igual a 1 si grupo poblacional de postulante es mestizo y 0 caso contrario.

X_{4di}: será igual a 1 si grupo poblacional de postulante es montubio y 0 caso contrario.

X_{4ei}: será igual a 1 si grupo poblacional de postulante es otro y 0 caso contrario.

X_{5i}: será igual a 1 si postulante posee carnet CONADIS y 0 caso contrario.

X_{6i}: será igual a 1 si nacionalidad de postulante es extranjera y 0 caso contrario.

X_{7ai}: será igual a 1 si tiempo de desarrollo de proyecto es más de 5 años y 0 caso contrario.

X_{7bi}: será igual a 1 si tiempo de desarrollo de proyecto es menos de 1 año y 0 caso contrario.

X_{8i}: será igual a 1 si etapa del proyecto es idea con prototipo y 0 caso contrario.

X_{9ai}: será igual a 1 si categoría de proyecto es agricultura y 0 caso contrario.

X_{9bi}: será igual a 1 si categoría de proyecto es alimentos y 0 caso contrario.

X_{9ci}: será igual a 1 si categoría de proyecto es ciencias de la vida y 0 caso contrario.

X_{9di}: será igual a 1 si categoría de proyecto es textiles y 0 caso contrario.

X_{9ei}: será igual a 1 si categoría de proyecto es construcción y 0 caso contrario.

X_{9fi}: será igual a 1 si categoría de proyecto es salud y 0 caso contrario.

X_{9gi}: será igual a 1 si categoría de proyecto es transporte y 0 caso contrario.

X_{9hi}: será igual a 1 si categoría de proyecto es turismo y 0 caso contrario.

X_{10ai}: será igual a 1 si subcategoría de proyecto es productos y 0 caso contrario.

X_{10bi}: será igual a 1 si subcategoría de proyecto es productos&servicios y 0 caso contrario.

X_{11i}: será igual a 1 si ha recibido recursos económicos de programas anteriores gestionados por CONQUITO y 0 caso contrario.

X_{12i}: es la cantidad numérica del total de integrantes del proyecto de emprendimiento.

X_{13i}: es la cantidad numérica de personas de género femenino.

X_{14i}: es la cantidad numérica de la cantidad de personas jóvenes de 18 a 29 años.

X_{15i}: es la cantidad numérica de personas con capacidades diferentes.

X_{16i}: es la cantidad numérica de adultos mayores.

X_{17i}: es la cantidad numérica de personas pertenecientes a pueblos o nacionalidades.

X_{18ai}: será igual a 1 si el proyecto innovador es modelo de gestión y 0 caso contrario.

X_{18bi}: será igual a 1 si el proyecto innovador es por proceso y 0 caso contrario.

X_{18ci}: será igual a 1 si el proyecto innovador es por producto y 0 caso contrario.

X_{19i}: será igual a 1 si el proyecto expone un correcto uso adecuado de financiamiento y 0 caso contrario.

X_{20i}: será igual a 1 si el proyecto adjunta evidencia congruente y 0 caso contrario.

X_{21i}: será igual a 1 si el proyecto aplica a la misma etapa de desarrollo con la que recibió capital semilla en anteriores programas gestados por CONQUITO y 0 caso contrario.

X_{22i}: será igual a 1 si el proyecto completa con la totalidad de la información de postulación y 0

caso contrario.

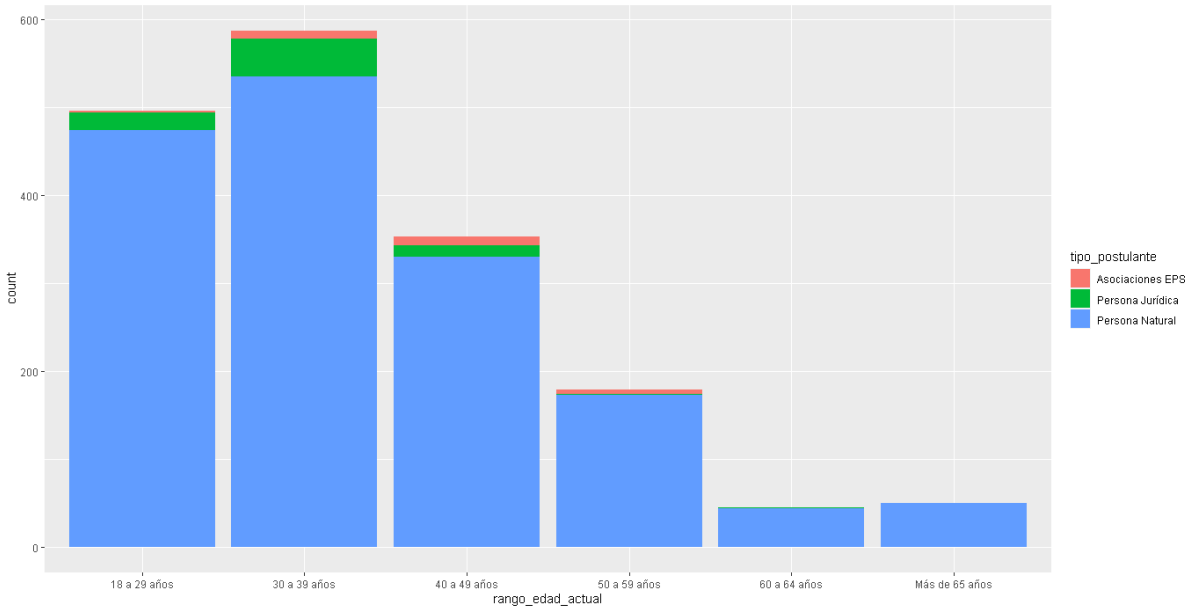
X_{23i}: será igual a 1 si el proyecto presenta documentación legible y 0 caso contrario.

Capítulo III Resultados, conclusiones y recomendaciones

Según el Índice de Actividad Emprendedora Temprana (TEA), el país mantiene un crecimiento del 29,62% en el año 2017 y del 36,2% en el año 2019 y que tras efectos de la pandemia se ha mantenido en aumento, sin embargo, al ser un país con alto porcentaje de TEA también mantiene un alto porcentaje de emprendimientos que claudican en el valle de la muerte y no superan una sobrevivencia de 48 meses.

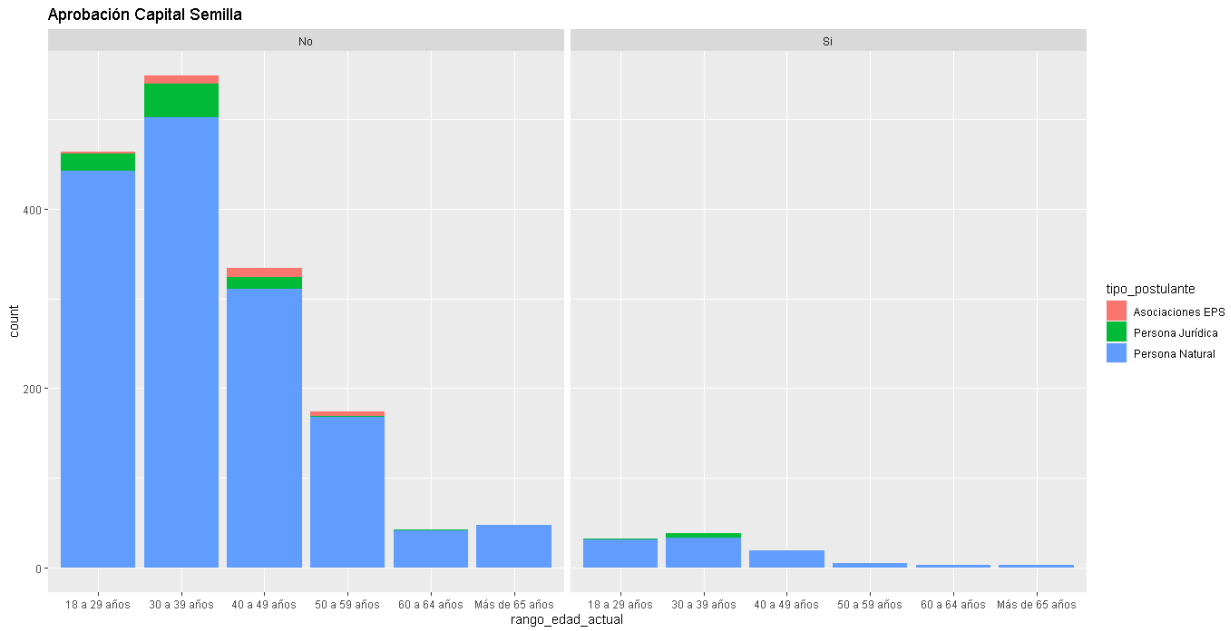
La ciudad de Quito integra más de un millón y medio de habitantes lo cual acorde al TEA comprende una gran cantidad de emprendedores que debido a la carencia de políticas gubernamentales para impulsar sus proyectos y también la falta de cultura financiera, provocan que muchos de los proyectos se mantengan en ideas o sin un prototipo.

El programa Fonquito, al ser el primer fondo de subvención en el D.M.Q en ejecutarse con finalidad de otorgar fondos no reembolsables bajo concepto de capital semilla, ha logrado generar 107 beneficiarios en su primera convocatoria, cada uno de estos con una adjudicación de capital semilla que benefició a proyectos en su gestación y desarrollo.



El gráfico anterior muestra que mediante el programa Fonquito se ha evidenciado que actualmente en la ciudad de Quito existen mayormente personas naturales jóvenes entre 18 a 29 años y adultos entre 30 a 49 años que conforma las candidaturas.

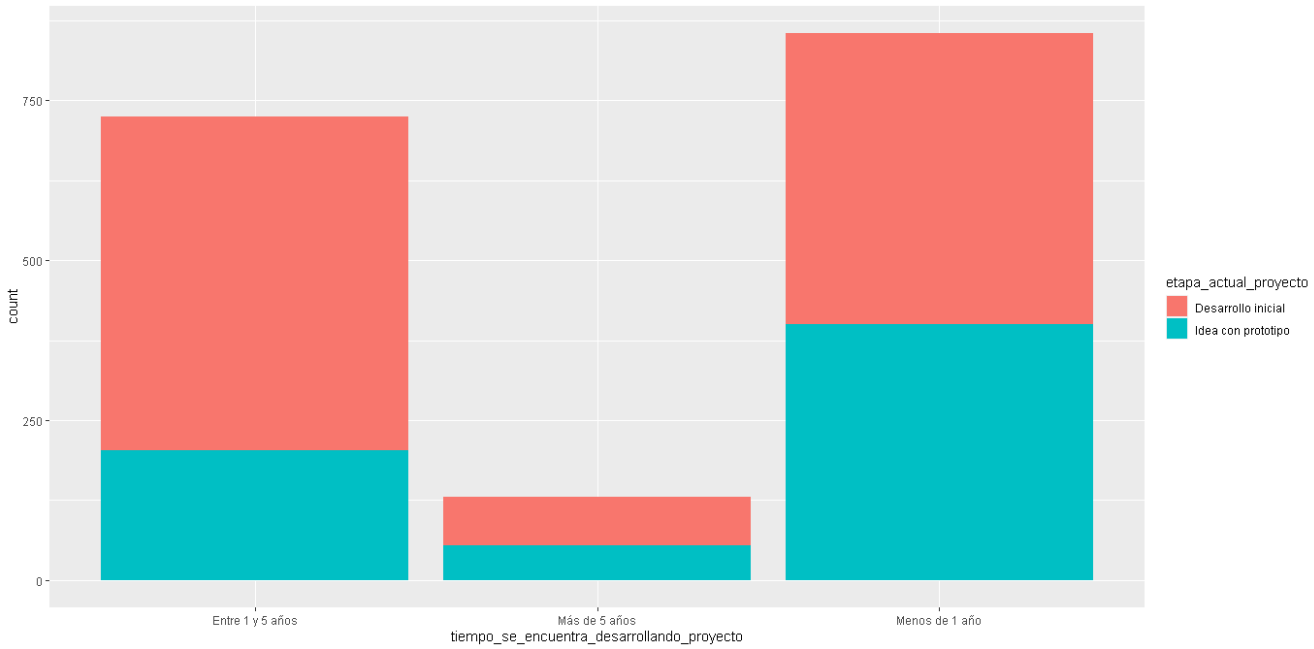
Los candidatos que aprobaron la adjudicación del capital semilla comprenden en su mayoría a personas naturales de entre 18 hasta 49 años:



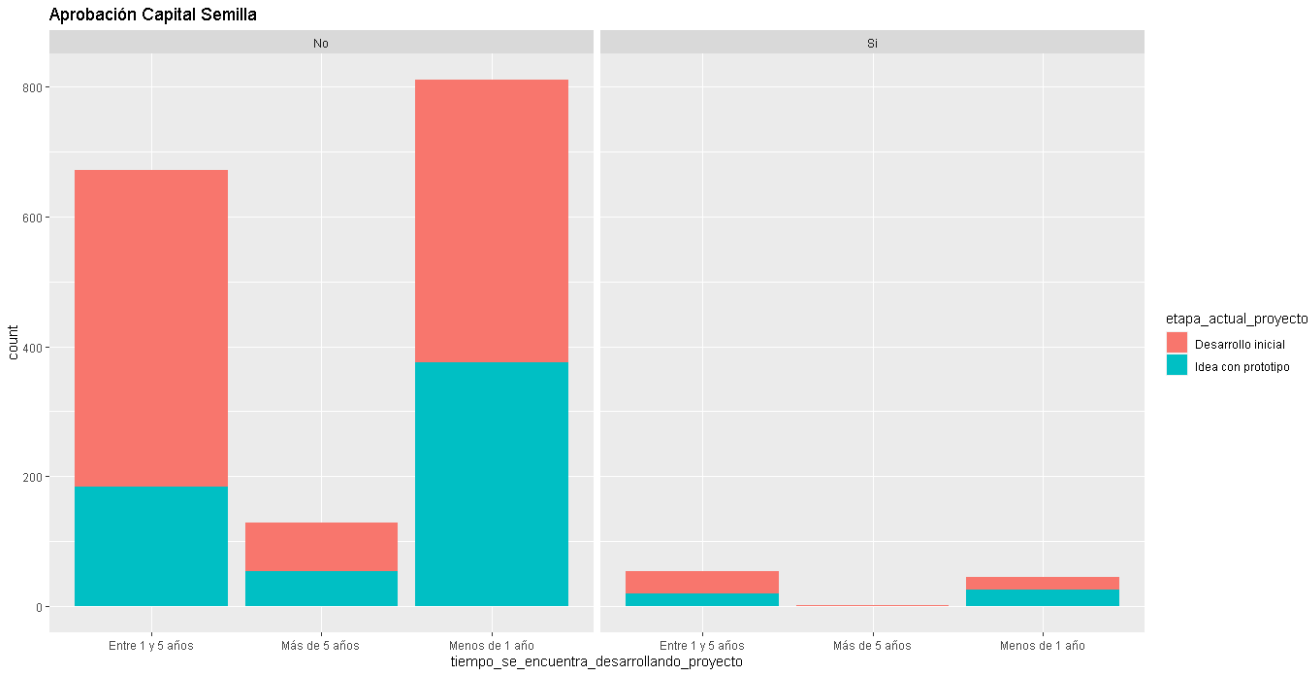
El modelo propuesto en este trabajo busca predecir a aquellos candidatos potenciales a la adjudicación de las convocatorias anuales que mantiene el programa Fonquito, para tal efecto se ha considerado el modelo cuyos parámetros mantienen un estándar aceptable para los mismos así como también se ha comparado dichos modelos con el fin de adecuar aquel con mejores predicciones proyectadas a la realidad.

Los postulantes que han iniciado un proyecto de emprendimiento cuyo tiempo de desarrollo sea entre 1 a 5 años, mayormente lo conforman proyectos con desarrollo inicial ya que estos están en etapa de realizar primeras ventas o inserción en mercado. También existe una cantidad balanceada de aquellos proyectos en etapas de desarrollo inicial e idea con prototipo cuyo tiempo de desarrollo es menor a un año, debido a que proyectos con idea prototipo se encuentran atravesando etapas de testeo para posterior integrarse a ser proyectos con desarrollo inicial.

Hay que enfatizar que una de las restricciones del programa Fonquito es la prohibición de adjudicación de capital semilla a proyectos cuyo tiempo de desarrollo supera los 5 años.

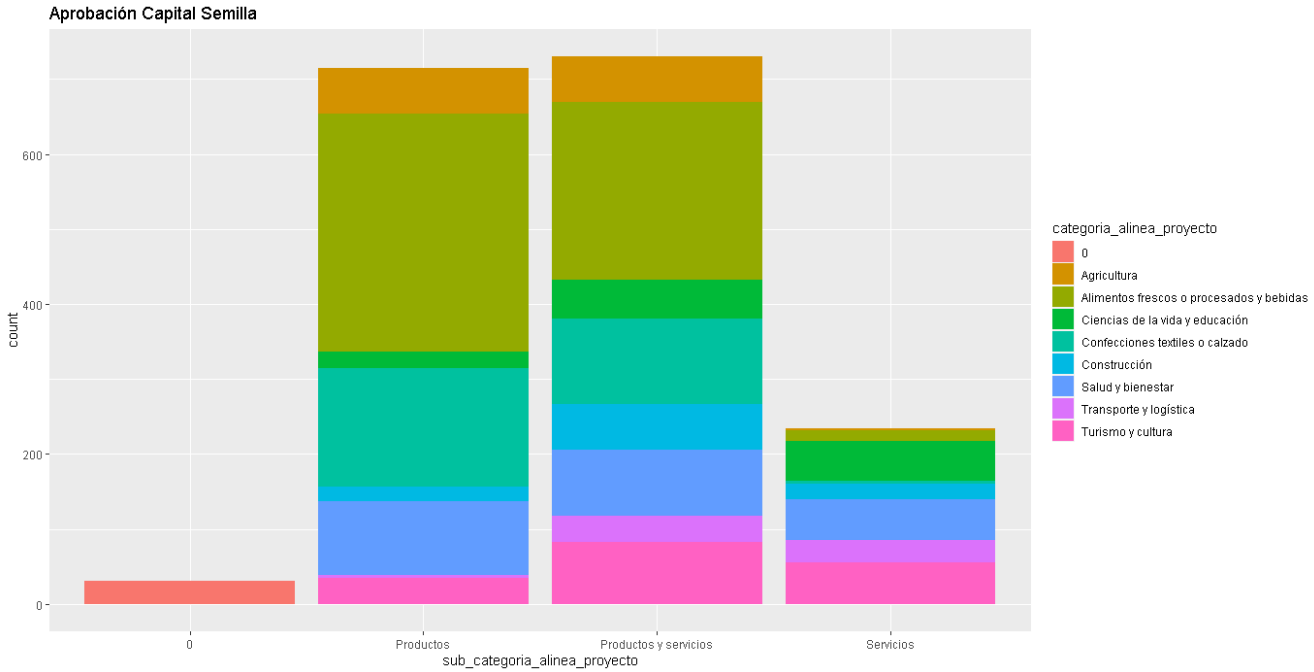


Se puede observar que los proyectos que fueron adjudicados de capital semilla fueron aquellos cuyo tiempo de desarrollo fueron desde los 0 hasta los 60 meses.

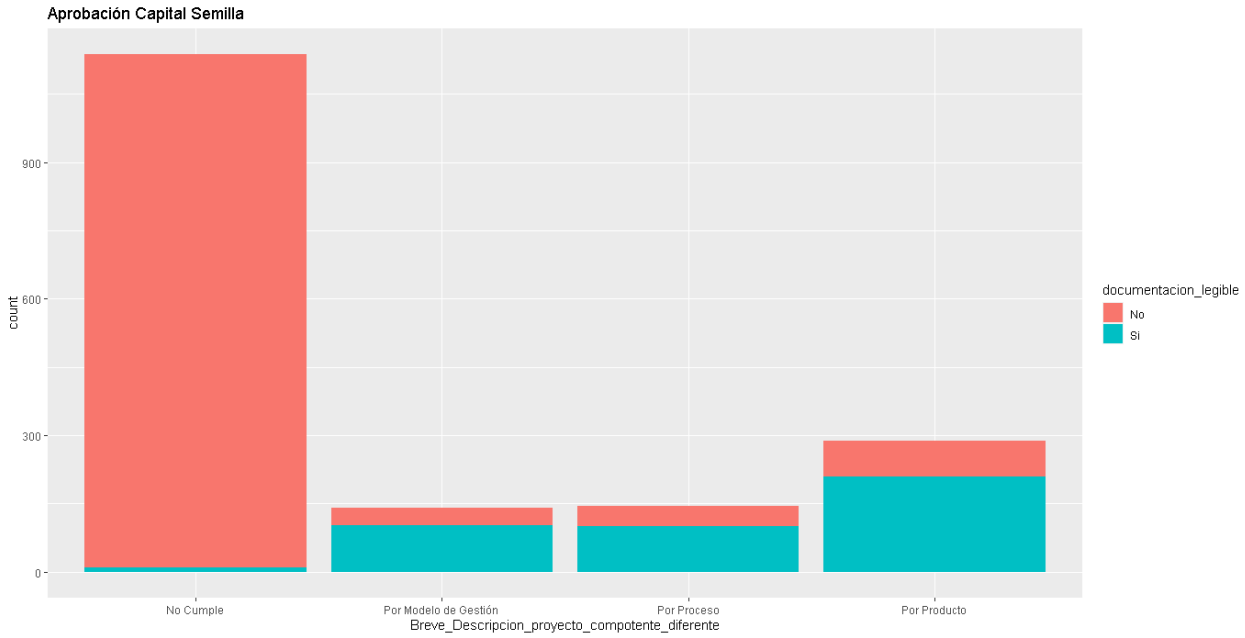


Los proyectos postulantes tienen 9 tipos de categorías que han sido clasificadas según el sector estratégico al que se orienta los productos o servicios o productos y servicios ofertados. La categoría “0”

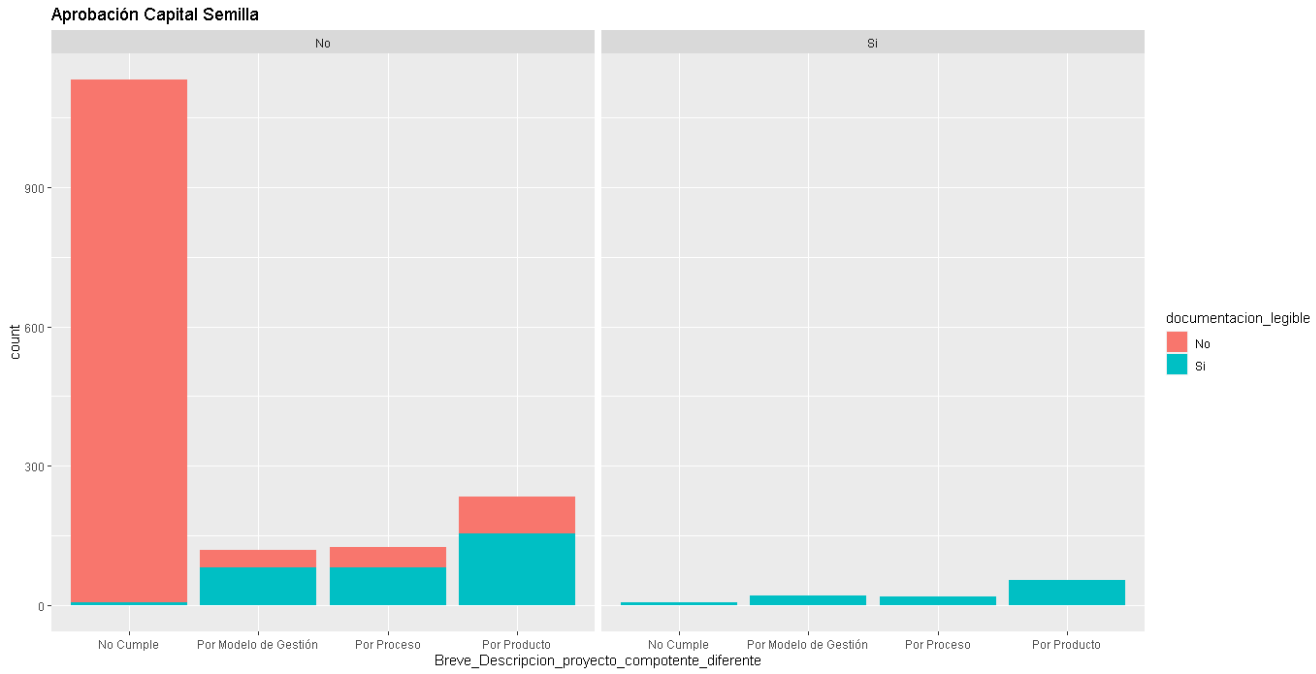
hace referencia a aquellos proyectos cuyos propietarios no definieron su categoría. Se evidencia que en la subcategoría de productos así como en productos y servicios predominan los alimentos frescos o procesados, por otro lado, en la subcategoría de servicios predominan aquellos de turismo y cultura.



Existen proyectos que tras validación de documentación se encontraron incompletos, no presentados o incluso con anomalías, lo cual permitió al equipo técnico valorar los proyectos acorde al componente diferenciador de los mismos, encontrando así proyectos que mayormente no están regularizados o no mantienen formalidad en su gestación. El componente diferenciador está clasificado por modelo de gestión, proceso o producto, siendo este último el que mayor caso concentra en la convocatoria.



Análogamente los proyectos que adjudicaron capital semilla son aquellos cuyo componente diferenciador fueron clasificados por su producto.



Lo anterior expuesto representa el objetivo de la primera convocatoria del Fonquito ya que busca brindar asistencia y seguimiento a la implementación de emprendimientos con potencial de convertirse en una MIPYME, que se encuentren en etapas de idea con prototipo o desarrollo inicial que cuenten con un componente diferenciador, con la finalidad de otorgar capital semilla que facilite su implementación e inserción en el aparato productivo en el D.M.Q.

El modelo planteado mantiene los objetivos del programa considerando los criterios generales, criterios de rechazo y excepciones que se estipulan en el mismo para garantizar el cumplimiento de las bases de la convocatoria y normativa legal vigente.

Para lograr esta etapa de adjudicación los beneficiarios se integraron a un proceso que duró entre siete y ocho meses, mismos que consistieron en procesos de migración de información para sistematizar, validar y desarrollar nuevos parámetros de decisión en base a criterios objetivos por parte de un equipo técnico designado por la corporación.

El proceso al ser realizado por personal de la corporación requirió de capital humano y tiempo invertido para el análisis de cada candidato postulante al programa, surgiendo así la necesidad de pronosticar en base a la información recolectada, aquellos proyectos o candidatos que tienen una alta probabilidad de ser considerados como candidatos tentativos a la adjudicación de capital semilla.

La necesidad conocida logra ser satisfecha cuando el tiempo de este proceso para la determinación de candidatos beneficiarios fue reducido haciendo uso de software estadístico que permite pronosticar desde la etapa de postulación de candidatos a aquellos proyectos que sean considerados por el modelo como propensos a ser beneficiados de dicho capital.

Para los tres modelos propuestos (*model1, model2* y *model*) se expone las siguientes matrices de confusión respectivas usando la data de entrenamiento:

Matriz *mode_l1*

```
> mc1
Confusion Matrix and Statistics

          Reference
Prediction  No  Si
        No 615  11
        Si   0 571

               Accuracy : 0.9908
                 95% CI : (0.9836, 0.9954)
    No Information Rate : 0.5138
    P-Value [Acc > NIR] : < 2.2e-16
```

Matriz *model_2*

```
> mc2
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1112   14
         1    0   71

               Accuracy : 0.9883
                 95% CI : (0.9805, 0.9936)
    No Information Rate : 0.929
    P-Value [Acc > NIR] : < 2.2e-16
```

Matriz *model*

```
> mc3
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1121    5
         1    5   66

               Accuracy : 0.9916
                 95% CI : (0.9847, 0.996)
    No Information Rate : 0.9407
    P-Value [Acc > NIR] : <2e-16
```

La matriz de confusión del primer modelo muestra que hay una cantidad de falsos positivos y nulos falsos negativos, esto genera una precisión y sensibilidad de 0,9908 y 1 respectivamente, mientras que el segundo modelo muestra una cantidad mayor de falsos positivos haciendo que la precisión del este modelo sea menor al primer modelo. Por otro lado el tercer modelo muestra una mejor precisión que los anteriores siendo de 0,9916.

El modelo propuesto “model” es testeado con la data de testeo llamada *test* cuya matriz de confusión se expone:

```
> mc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 479   5
         1  10  19

               Accuracy : 0.9708
                 95% CI : (0.9522, 0.9835)
    No Information Rate : 0.9532
    P-Value [Acc > NIR] : 0.03135
```

Para esta prueba se evidenció que la predicción responde a la realidad de la data de testeo pues la misma integra 513 observaciones las cuales 29 corresponden a candidatos beneficiarios de capital semilla. Adicional la precisión del modelo aplicado a esta data no ha variado y mantiene valores adecuados.

Resultados predicciones beneficiarios

El score estructurado para el modelo predictivo se aplicó a los datos de entrenamiento (train) que integra 1197 registros, y de igual forma se aplicó a los datos de testeo (test) y su tabla score respectiva se clasifica en rangos definidos previamente, para el caso en 2.

Cuenta de Var dependiente	Etiquetas de columna		
Etiquetas de fila	**0**	**1**	**Total general**
2	328	29	357
1	156		156
Total general	**484**	**29**	**513**

Cuenta de Var dependiente	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
10	3111	7074	10185
9	6364	3820	10184
8	7739	2446	10185
7	8210	1975	10185

La tabla anterior muestra que existen en efecto 29 beneficiarios correspondientes a la categoría dos en la data de testeo los cuales se proyectan a la realidad de la data. Adicional pronostica 484 malos candidatos pertenecientes a categorías 2 y 1.

Análogamente se realiza una tabla score para los datos de entrenamiento con el fin de validar la predicción estructurada, obteniendo la siguiente tabla:

Cuenta de Var dependiente	Etiquetas de columna		
Etiquetas de fila	**0**	**1**	**Total general**

2	328	71	399
1	798		798
Total general	**1126**	**71**	**1197**

Cuenta de Var dependiente	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
10	3111	7074	10185
9	6364	3820	10184
8	7739	2446	10185
7	8210	1975	10185

En efecto se aprecia que en la data entrenamiento existen 71 candidatos beneficiados de capital semilla correspondientes a la categoría 2 y 1126 malos candidatos pertenecientes a la categoría 1.

Se propone entonces dos casos al azar tomados de la base de datos del programa Fonquito el cual se va analizando un particular de la data para fines prácticos como se expone en los siguientes casos:

Caso 1: Candidato de personería jurídica natural, rango de edad de 18 a 29 años, género femenino, mestiza, no posee carnet CONADIS, ecuatoriana, proyecto de entre 1 y 5 años en etapa desarrollo inicial, categoría de agricultura que brinda productos y servicios, no ha recibido recursos económicos gestados por la corporación, el proyecto integra 3 personas, 1 de género femenino, 3 personas jóvenes de 18 a 29 años, componente innovador por modelo de gestión, cumple con la proyección del uso adecuado de capital semilla, adjunta evidencia congruente, el proyecto si aplica a la misma etapa de desarrollo con la que recibió capital semilla, cumple con la totalidad de la información del formulario y cumple con documentación legible.

Haciendo uso del modelo predictivo propuesto se tiene la ecuación:

$$Y_{caso1} = -2{,}971 - 0{,}146(0) - 0{,}526(1) + 0{,}144(0) + 0{,}175(0) - 0{,}662(0) + 0{,}515(0) +$$

$0{,}489(0) + 0{,}406(0) + 0{,}0467(0) + 0{,}382(0) - 1{,}014(0) - 0{,}594(0) - 0{,}404(1) - 0{,}619(0) - 2{,}142(0) - 0{,}125(0) - 0{,}216(0) + 0{,}235(0) + 0{,}0581(0) - 0{,}0538(0) + 0{,}508() + 0{,}658(1) + 1{,}591(0) + 0{,}746(0) + 1{,}032(0) + 0{,}749(0) + 1{,}381(0) + 1{,}089(0) + 0{,}441(0) + 0{,}488(1) - 0{,}138(0) - 0{,}046(3) + 0{,}92(1) - 0{,}146(3) + 0{,}113(0) - 0{,}099(0) - 0{,}145(0) - 0{,}901(1) - 0{,}496(0) - 0{,}672(0) + 0{,}286(1) + 1{,}209(1) + 2{,}886(1) - 1{,}256(1) + 0{,}633(1)$

El valor de Y es de 0,296 lo cual muy probablemente es mala candidata para ser beneficiaria de capital semilla.

Caso 2: Candidato de personería jurídica natural, rango de edad de 30 a 39 años, género femenino, mestiza, no posee carnet CONADIS, ecuatoriana, proyecto de entre 1 y 5 años en etapa desarrollo inicial, categoría de alimentos que brinda productos y servicios, no ha recibido recursos económicos gestados por la corporación, el proyecto integra 6 personas, 3 de género femenino, 1 personas jóvenes de 18 a 29 años, componente innovador por producto, cumple con la proyección del uso adecuado de capital semilla, adjunta evidencia congruente, el proyecto si aplica a la misma etapa de desarrollo con la que recibió capital semilla, cumple con la totalidad de la información del formulario y cumple con documentación legible.

$Y_{caso1} = -2{,}971 - 0{,}146(0) - 0{,}526(1) + 0{,}144(1) + 0{,}175(0) - 0{,}662(0) + 0{,}515(0) + 0{,}489(0) + 0{,}406(0) + 0{,}0467(0) + 0{,}382(0) - 1{,}014(0) - 0{,}594(0) - 0{,}404(1) - 0{,}619(0) - 2{,}142(0) - 0{,}125(0) - 0{,}216(0) + 0{,}235(0) + 0{,}0581(0) - 0{,}0538(0) + 0{,}508(1) + 0{,}658(0) + 1{,}591(0) + 0{,}746(0) + 1{,}032(0) + 0{,}749(0) + 1{,}381(0) + 1{,}089(0) + 0{,}441(0) + 0{,}488(1) - 0{,}138(0) - 0{,}046(6) + 0{,}92(3) - 0{,}146(1) + 0{,}113(0) - 0{,}099(0) - 0{,}145(0) - 0{,}901(1) - 0{,}496(0) - 0{,}672(0) + 0{,}286(1) + 1{,}209(1) + 2{,}886(1) - 1{,}256(1) + 0{,}633(1)$

En este caso el valor de Y es de 0.801 el cual tiene alta probabilidad de ser considerado un buen candidato para ser beneficiario de capital semilla mismo que adjudica un valor de $3000.

Conclusiones y recomendaciones

En este trabajo se expone un modelo matemático que permite simplificar las etapas que integra una convocatoria del programa Fonquito cuyo fin es la estimación de candidatos beneficiarios en base a las condiciones de la convocatoria desarrollada.

La aplicación de métodos matemáticos orientados a resolver problemas reales pueden influir para mejorar las políticas públicas al desarrollo del emprendimiento, impulsando la eficiencia y generando una propuesta de convocatorias continuas.

Se ha evidenciado que esta convocatoria enfoca a productos en categoría de alimentación y derivados, cuyos propietarios en su mayoría conforman personas jóvenes y adultos de hasta 49 años con proyectos en etapa de desarrollo inicial y con tiempo de gestación no mayor a 5 años.

Los ejercicios de aplicación del modelo fueron considerados tomando al azar dos proyectos aleatorios, uno de grupo de rechazados para adjudicación de capital semilla y uno del grupo de aprobados, mismos que posterior coincidieron con la realidad de la estimación.

El modelo propuesto fue testeado en las convocatorias actuales que se encuentran en ejecución para uso de la corporación y fines operativos del equipo técnico.

El programa Fonquito al ser una iniciativa impulsada por el municipio del Distrito Metropolitano de Quito puede ser considerado como parte de los programas que tienen intervención política por lo cual se recomienda encriptar la información recolectada en cada convocatoria y ejecutar el modelo para posterior hacer estudio de beneficiarios que puedan ser atípicos a los resultados pronosticados o de ser el caso realizar ajustes al modelo.

Referencias

1. Alonso Cánovas, D., y Tubau Sala, E. 2002. Inferencias bayesianas: una revisión teórica. Anuario de Psicología, pp 25-47.

2. Andreeva, G. (2005). European generic scoring models using survival analysis. Journal of the Operational research Society, 57(10), 1180-1187.

3. Baesens, B. (2003). Developing Intelligent Systems for Credit scoring using Machine Learning Techniques. (Tesis doctoral), Katholieke Universiteit Leuven, LIRIS, Louvain, Bel.

4. Bierman, H. y Hausman, W. H. (1970). The credit granting decision. Management Science, 16(8), 519-532.

5. Biggs, D., De Ville, B. y Suen, E. (1991). A Method of Choosing Multiway Partitions for Classification and Decision Trees. Journal of Applied Statistics, 18(1), 49-62.

6. Bouroche, J. y Tennenhaus, M. (1972). Some segmentation methods. Metra, 7, 407-418.

7. Boyle M., Crook J.N., Hamilton R. y Thomas L.C. (1992). Methods for credit scoring applied to slow payers in Credit scoring and Credit Control. Oxford: Oxford University Press.

8. Breiman, L., Friedman, J., Olshen, R. y Stone, C. (1984). Classification and Regression Trees.

Belmont: Wadsworth.

9. Cellard, I., Labbe, B. y Cox, G. (1967). Le programme Elisèe. Presentation et Applicaction. Metra, 3, 511-519.

10. Crook, J.N., Edelman, D.B. y Thomas, L.C. (2007). Recent development in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447-1465.

11. Dellaportas, P., Karlis, D. y Xekalaki, E. (1997). Bayesian analysis of finite poisson mixtures. Manuscript.

12. Desai V.S., Convay D.G., Crook J.N. y Overstreet G.A. (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms. IMA J. Mathematics applied in Business and Industry, 8, 323-346.

13. Gutiérrez Girault, M. A. (2007). Modelos de credit Scoring - Qué, cómo, cuándo y para qué. Munich: Munich Personal RePEc Archive. Recuperado el 11 de diciembre de 2016, de: https://mpra.ub.uni-muenchen.de/16377/

14. Herrera García, B. (2009). La supervisión de los bancos y el rol del Comité de Basilea para la supervisión bancaria. Contaduría y Administración(212), 41-48. Recuperado el 2 de mayo de 2017, de: http://www.ejournal.unam.mx/rca/212/RCA21204.pdf

15. Manjarrés V., F. A. (2012). Estructura y regulación del sistema financiero y bursátil. Saarbrücken: Editorial Académica Española.

16. Matich, D. J. (2001). Redes neuronales: conceptos básicos y aplicaciones (trabajo de curso, Informática aplicada a la Ingeniería de Procesos, orientación I, Universidad Tecnológica Nacional, Rosario). Recuperado el 11 de17. diciembre de 2016, de: https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadoral/monograias/matich-redesneuronales.pdf

18. Pérez Ramírez, F. O., y Támara Ayús, A. L. (2012). Análisis discriminante como seleccionador de variables influyentes en el cálculo de la probabilidad de incumplimiento. Revista Ciencias Estratégicas, 20(27), 307-322. Recuperado el 25 de abril de 2017, de: https://revistas.upb.edu.co/index.php/cienciasestrategicas/article/viewFile/1476/1437

19. Puertas Medina, R., y Martí Selva, M.L. (2013). Análisis del credit scoring. Revista de Administração de Empresas, 53(3), 303-315. Recuperado el 5 de noviembre de 2017, de: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S003475902013000300007&lng=en&tlng=en

20. Rayo, S. (2013). Gestión avanzada de riesgo de crédito. Seminario para gerentes de riesgos de las entidades de microfinanzas. Lima: Superintendencia de Bancos y Seguros y Banco Interamericano de Desarrollo. Recuperado