UNIVERSIDAD TORCUATO DI TELLA

TESIS DE MAESTRIA EN ECONOMETRIA

CONSTRUCCIÓN DE UN MODELO DE CREDIT SCORING PARA POBLACIÓN NO BANCARIZADA DE LA CIUDAD AUTÓNOMA DE BUENOS AIRES UTILIZANDO EL CENSO NACIONAL DE PERSONAS, HOGARES Y VIVIENDAS COMO FUENTE COMPLEMENTARIA DE DATOS

Autor: Lic. Erica Anabela Grimberg

Directora: Ing. María del Rosario Bruera

2018

Contenido

Resumen

En la actualidad cada vez está más difundida la utilización de los modelos de Credit Scoring para estimar la probabilidad de que un individuo incurra en mora. La población no bancarizada es el segmento en el que existe mayor incertidumbre para predecir este evento. Este trabajo busca evaluar si el Censo Nacional de Personas, Viviendas y Hogares es útil como una fuente complementaria de información para integrar en los modelos de Credit Scoring de la población no bancarizada. La experiencia presentada se circunscribe a Capital Federal.

En el trabajo se describe en detalle la metodología utilizada tanto para el análisis de la fuente censal como para la construcción del modelo de Scoring final y su validación. Además, se incluye un análisis de puntos de corte y las especificaciones de implementación del modelo final.

Introducción

Todos los días, las distintas entidades que forman parte del sistema financiero se encuentran ante la necesidad de decidir sobre la aceptación de un nuevo cliente. La objetividad y la automatización de estas decisiones han tomado vital importancia en los últimos años. Los modelos de Credit Scoring desempeñan un papel fundamental en tal sentido. Estos modelos tienen el objetivo de evaluar a los solicitantes de productos de créditos y determinar su probabilidad de incurrir en default en una ventana de tiempo. Esta probabilidad se calcula a partir de modelos que surgen del análisis de datos y la búsqueda de patrones de comportamiento. Las estimaciones realizadas para predecir el default de individuos resultan más precisas cuando poseen información sobre comportamientos históricos de pagos. Sin embargo, existe el interés de poder predecir también el comportamiento en aquellas personas no bancarizadas para poder incluirlas en el circuito del crédito conociendo el riesgo que eso significa. Se entiende como población no bancarizada a todas aquellas personas físicas que no poseen productos de crédito informados ni en el Banco Central de la República Argentina ni en burós de crédito en los últimos 5 años.

El análisis sobre este segmento de la población, que posee escasa información en fuentes tradicionales, sostiene la necesidad de incorporar fuentes alternativas de datos para entender su comportamiento.

La propuesta de este trabajo consiste en poner a prueba una nueva fuente de datos para explicar la mora en la población no bancarizada de Capital Federal. Esta nueva fuente si bien es muy conocida, no es habitual encontrarla incorporada a los modelos de población no bancarizada. El Censo - si bien no posee una actualización frecuente - brinda variables estructurales por radio censal para todo el territorio de la República Argentina que permiten conocer características del entorno en el cuál vive el individuo. Se busca probar entonces, si esta información que es pública y de fácil acceso, logra contribuir en la predicción del default para el segmento no bancarizado.

En el Capítulo 1 se lleva a cabo una revisión de la historia de los modelos de Credit Scoring, los tipos y metodologías utilizadas y los desarrollos actuales encontrados sobre el segmento de interés.

En el Capítulo 2, se describe la nueva fuente de datos y se detalla la metodología utilizada para evaluar si esta nueva fuente resulta significativa para predecir la entrada en default, que será la variable objetivo.

En el Capítulo 3, se muestran los resultados y estimaciones obtenidos a partir la aplicación de la metodología descripta en el Capítulo 2

En el Capítulo 4 se presenta las pruebas de validación realizadas en una muestra distinta a la utilizada en el desarrollo.

En el Capítulo 5, se detallan los datos de la implementación, dado que este trabajo se inspira en necesidades reales del mercado actual y tiene una fuerte inspiración práctica.

Y finalmente en el Capítulo 6, se describen las conclusiones encontradas a partir del análisis.

Capítulo 1: Marco teórico y Antecedentes

La globalización, la evolución de la tecnología e internet y el desarrollo de distintos productos crediticios fueron algunos de los factores que fomentaron el enorme incremento de la industria financiera en las últimas décadas.

Dada la expansión de la industria de crédito, la necesidad de tomar decisiones objetivas, rápidas y consistentes adquirió cada vez más importancia. El interés por modelar el riesgo al momento de la admisión de un nuevo cliente y en el proceso de gestión de clientes existentes se convirtió en un objetivo fundamental. Credit Scoring es el uso de modelos estadísticos para transformar los datos relevantes en medidas numéricas que guían la toma de decisiones.

1.1 Historia de Modelos de Credit Scoring

Hasta la década del ’60 el juicio humano era el único factor que influía en la decisión de otorgar un crédito. Los responsables de otorgar préstamos usaban su experiencia, surgida de la observación del comportamiento de créditos de otros clientes, como base para juzgar a nuevos consumidores. El criterio para evaluar a un solicitante de crédito se basaba en las 5Cs: carácter (del solicitante), capacidad (para hacer frente al pago), capital (de respaldo), colateral (garantía de seguridad) y condiciones (factores externos). Este proceso, no sólo era lento sino que resultaba poco confiable a causa del error humano. Como evolución, los prestamistas estandarizaron la forma en que se tomaba la decisión mediante la utilización de un sistema de puntos que evaluaba las variables incluidas en el reporte de créditos. Esta nueva metodología ayudó a eliminar bastante el sesgo y la subjetividad que existía hasta entonces, sin embargo, todavía la puntuación estaba vinculada a medidas intuitivas de solvencia crediticia y no estaba verdaderamente basada en el comportamiento real del consumidor.

En la década del ‘60, el otorgamiento de crédito dio un gran salto cuando comenzaron a utilizarse modelos estadísticos que tenían en cuenta una gran cantidad de variables y combinación de variables. Estos modelos eran construidos con la información de pago de miles de consumidores reales. Tenían como objetivo determinar si las personas que solicitaban un crédito pagarían la deuda, cumplirían sus obligaciones y, en general, si actuarían del modo que la

entidad financiera consideraba aceptable. Esto causó un cambio drástico en la manera en que se otorgaban los créditos, se mutó de préstamos basados en relaciones personales a préstamos transaccionales.

El ingeniero Bill Fair y el matemático Earl Isaac, fundadores de la consultora Fair Isaac (FI, conocida como FICO) en San Francisco, fueron quienes introdujeron por primera vez el concepto de Credit Scoring a empresas otorgantes de crédito en 1958. Sin embargo, no fue sencillo competir con las costumbres arraigadas de estas entidades. American Investments fue la primera entidad en adoptar estos modelos y recién luego de observar el potencial de los mismos, otras entidades financieras comenzaron a adoptarlos para sus sistemas de decisión.

En 1963, FICO comenzó a dar servicio a Montgomery Ward[1], lo que ayudó a consolidar su posicionamiento a nivel mundial y desde entonces, comenzaron a dar servicio a muchas entidades otorgantes de crédito en los Estados Unidos.

Montgomery Ward contaba con un departamento de crédito en cada sucursal, pero gracias a la llegada de la computadora tuvo la posibilidad de centralizar la función de crédito.

Luego del suceso con Montgomery Ward, muchos otros retailers, como R.H. Macy, Gimbel's, Bloomingdale's, and J.C. Penney, avanzaron en el mismo sentido al observar este éxito.

Con el tiempo, otros rubros comenzaron a experimentar el uso de modelos estadísticos. A mediados de la de la década del '60, las compañías petroleras comenzaron a implementar los modelos de Credit Scoring debido a los problemas que experimentaban con sus operaciones de crédito, como robo de tarjetas, fraude, pérdidas de crédito. En este periodo, también se suman a la utilización de estos modelos, entidades emisores de tarjetas como Diners Club, American Express y Carte Blanche.

Muchas tarjetas eran emitidas sin cuota de ingreso, lo que incrementó el volumen y la competencia en el mercado, pero ocasionó grandes pérdidas. Esto último impulsó la necesidad de mejorar el proceso de toma de decisiones y comenzar con la implementación de Credit Scoring en este proceso. La mejora en los

[1] Montgomery Ward (MW) fue la primera megatienda a nivel internacional de compra por correo.

resultados no tardó en llegar y las tasas de mora se redujeron considerablemente.

Sin embargo, los solicitantes de crédito en un principio pensaban que “el crédito es un arte y no una ciencia”, por lo que no creían que una computadora pueda brindar mecanismos para tomar la decisión de otorgar un crédito. Además los desarrolladores de estos modelos, muchas veces eran incapaces de explicar por qué ciertas características eran favorables sobre otras. Sin embargo, en los años 1975 y 1976, la Reserva Federal declaró ilegal cualquier tipo de discriminación en el otorgamiento de créditos, a menos que la misma se base en modelos estadísticos. Esto favoreció al reconocimiento del uso de Credit Scoring. Por su parte, los resultados fueron demostrando la ventaja de utilizar estos modelos y los mismos fueron adquiriendo cada vez más aceptación y se fue masificando su utilización, incluso a otros productos, a medida que su costo y su velocidad mejoraron. A finales de los años ’70 y durante la década del ’80, estos modelos se comenzaron a aplicar en el proceso de otorgamiento de préstamos personales, descubiertos, financiamiento de automóviles y préstamos a pequeñas empresas. En sus inicios, mucho se hacía manualmente, recién en 1972, FICO realizó para Wells Fargo[2], la primera implementación totalmente automatizada de Credit Scoring.

La década del ’80, se caracterizó por la expansión y evolución de estos modelos. Por un lado, la utilización de esos modelos se hizo extensiva a otras áreas del ciclo de vida del crédito, como retención, comportamiento, cobranzas, entre otras. (Ver sección 1.2). Por otro lado, las técnicas utilizadas para el desarrollo de estos modelos se diversificaron. En sus comienzos estos modelos eran desarrollados usando técnicas estadísticas como análisis discriminante (DA) y probabilidad lineal (LPM), pero gracias al desarrollo de la tecnología y los softwares se pudieron experimentar otras técnicas como redes neuronales, sistemas expertos y regresiones logísticas, siendo esta última la más utilizada (Ver sección 1.3).

[2] Wells Fargo es un banco de Estados Unidos con operaciones en todo el mundo. Fue la primera entidad en fijar precios basados en el nivel de riesgo para pequeños empresas.

Aproximadamente en 1984, FICO desarrolló uno de los primeros *Score Credit Bureau*, para preseleccionar listas de contactos, utilizando datos de Buró. Así fue como FICO se convirtió en el pionero en desarrollar modelos matemáticos con información de Burós de Créditos[3], conocidos como FICO Scores.

El concepto ganó una amplia aceptación luego de que en 1987, MDS (Management Decision Systems)[4] desarrollara modelos de Score para predecir la quiebra para los tres principales burós (Equifax, Experian y TransUnion). Entre 1989 y 1991, FICO desarrolló scores para predecir el default.

Las mejoras en la tecnología que se dieron a partir de la década del '90, permitieron reducir los costos de almacenamiento de datos, lo que aumentó la posibilidad de que los prestadores puedan invertir en almacenamiento y minería de datos. También los prestadores comenzaron a comprar software para realizar desarrollos de modelos de Credit Scoring internos.

En este periodo, el Credit Scoring se expandió a otras nuevas áreas. Los efectos de la combinación de la mayor transparencia, tecnología y precios hizo posible llegar a los sectores con menores ingresos. La utilización de estos modelos en préstamos hipotecarios tuvo mayor resistencia, pero para fines de la década del '90, aproximadamente la mitad del mercado estadounidense utilizaba estos modelos para otorgar créditos hipotecarios. Así mismo, la aceptación de estos modelos generó un cambio sustancial en las aseguradoras, ya que permitió variar los términos y precios en función del riesgo. De manera similar ocurrió en el mercado de tarjetas de crédito un tiempo después.

En Argentina, ya para 2006 un estudio presentado por el Banco Central, mostró que estos modelos tenían una amplia difusión, sobre todo en carteras minoristas para la originación de productos.[5]

---

[3] Son llamadas Buró de Créditos las agencias de informe crediticio o centrales de riesgo que poseen información sobre el cumplimiento crediticio de entidades o individuos. En la actualidad, a nivel mundial hay 4 compañías que dominan esta industria. Dun & Bradstreet (D&B) es el principal jugador en información de crédito de empresas, mientras que Equifax, Experian y TransUnion dominan el mercado de individuos.

[4] MDS (Management Decision Systems) fundada en 1975/5 por John Coffman y Gary Chandler, fue la primera entidad en desarrollar score bureau de quiebra.

[5] Girault M. (2007). Modelos de Credit Scoring – Qué, Cómo, Cuándo y Para Qué-. Gerencia de Investigación y Planificación Normativa, Subgerencia General de Normas. Banco Central de la República Argentina (BCRA).

En la actualidad, los modelos de Credit Scoring son ampliamente utilizados en todo el mundo, para distintos productos, mercados y etapas del ciclo de vida del cliente. Los mismos pueden ser nutridos por diversas fuentes de información: propia de la entidad, provista por el cliente, proveniente de las agencias externas (Burós de Créditos) o de la Central de Deudores del Sistema Financiera.

1.2 Tipos de Modelos de Credit Scoring

En sus inicios los modelos de Credit Scoring se asociaban exclusivamente con el proceso de altas de nuevas solicitudes de crédito y con la decisión de "rechazar" o "aceptar" la nueva solicitud. En el siglo XXI, el concepto de Credit Scoring comenzó a utilizarse de una manera más amplia para referirse a todos aquellos modelos empleados para expandir o gestionar los créditos en general. Estos modelos reciben diferentes nombres dependiendo de: (i) la fuente de información; (ii) el objetivo o (iii) lo que está midiendo[6]. Siguiendo la división planteada por Raymond Anderson, los más comunes son:

Application Score: utilizado para originación de negocios. Combina información del cliente, transacciones anteriores y datos de Burós de Crédito. El desarrollo de modelos a medida es más habitual en entidades grandes y cuando los Burós de Crédito no cuentan con información suficiente para predecir la mora.

Behavioural Score: utilizado para gestión de cuentas, por lo general se centra en el comportamiento a nivel de cuenta. En su mayoría utilizan datos referentes al rendimiento del producto crediticio en cuestión. Con el tiempo, fue posible enriquecer estos modelos utilizando información demográfica, de Burós de Crédito y de otros productos. Estos modelos suelen ser a medida, desarrollados internamente en la entidad.

Collections Score: utilizado como parte del proceso de cobranzas, generalmente para impulsar acciones en los centros de llamadas. Suele combinar información propia del proceso de cobranzas (como promesa de pago, historial de contactos), con información de Buró de Crédito y otros productos.

[6] Anderson. R. (2007). The Credit Scoring Toolkit: theory and practice for retail Credit Risk Management and Decision Automation. Oxford University Press

Customer Score: a diferencia del Behavioural Score y Collections Score, que evalúan a nivel de cuenta, éste combina la información de todas las cuentas de un cliente. Suele ser útil para identificar el perfil de riesgo global de un cliente para mejorar la oferta y realizar gestiones de cuenta más apropiadas. Suele utilizarse para cross-selling.

Bureau Score: Es otorgado por un Buró de Crédito y sumariza los datos que figuran registrados sobre el individuo, con el objetivo de predecir la mora o la quiebra del individuo en alguna de sus cuentas. Algunas entidades lo usan como insumo para sus propios modelos y procesos, otras entidades, en general más pequeñas, lo utilizan como la única medida en el proceso de decisión. Suele ser muy valioso en la originación de nuevas cuentas, dado que brinda información sobre los solicitantes que no sería posible obtener de otro modo.

Los distintos scores buscan cubrir diferentes aspectos en el comportamiento del cliente, utilizando información proveniente de distintas fuentes: cliente, interna de la propia entidad y externa.

Muchas entidades otorgantes de crédito combinan sus scores internos con scores genéricos provenientes de los Burós de Crédito (Bureau Scores) en matrices de decisión o bien integran la información de los Burós de Crédito en sus scores internos, con el objetivo de modelar lo mejor posible los distintos aspectos en el comportamiento del cliente. Según describe Raymond Anderson (The Credit Scoring Toolkit, 2007) estos son las 4Rs: Risk (Riesgo), Response (Respuesta), Retention (Retención), Revenue (Ingresos).

En *Riesgo* es donde, tal vez, los modelos de Credit Scoring son más conocidos. Los scores de riesgo pueden dividirse en tres: de crédito, de fraude y de seguro. El riesgo de crédito busca predecir la morosidad e incluye a la mayoría de los scores mencionados anteriormente (Application, Behavioural, Customer, Collection, Customer, Bureau Score). Se pueden utilizar de manera aislada para tomar decisiones o bien combinado con score de retención, de respuesta y de ingresos. Por su parte, el score de riesgo de fraude busca identificar a quienes no tienen intención de pagar. Si bien el fraude ha ocasionado grandes pérdidas al sistema financiero, estos son modelos difíciles desarrollar por la baja cantidad de casos identificados y la difícil discriminación entre estos y los que

verdaderamente son incapaces de pagar. Por último, el riesgo de seguro, busca predecir el reclamo de seguro a corto plazo. Si bien no está relacionado con los datos de crédito se ha demostrado que existe una fuerte correlación entre los datos de crédito y los reclamos a las aseguradoras.

La *Respuesta* de un cliente o prospecto a una determinada acción o campaña que realiza una entidad no es un tema menor. Los Scores de Respuesta buscan lograr un mejor direccionamiento de las campañas de tal manera de limitar las mismas a aquellas personas que tienen mayor probabilidad de convertirse en clientes de la entidad o de aceptar un producto específico. Esto permite reducir costos de campaña y mejorar la tasa de aceptación.

La *Retención* de clientes suele ser otro punto clave en el negocio. Muchas veces los costos de adquisición son altos, por lo que se requiere que el cliente perdure para que sea rentable. Los modelos de Credit Scoring son utilizados para determinar si un cliente nuevo continuará siendo cliente. También son utilizados para predecir la inactividad de la cuenta o el cierre de la misma y diseñar estrategias para mantener activos a los clientes.

El *Ingreso* potencial que un cliente puede brindarle a una institución es un punto de interés para todas las entidades. Se utiliza Credit Scoring para identificar a aquellos individuos que hagan mayor uso de sus productos.

Siguiendo con la descripción de Raymond Anderson, los Scores de Riesgo, de Respuesta, de Retención y de Ingresos sirven para responder distintas necesidad del negocio y pueden ser utilizados en distintas etapas de la gestión de riesgo crediticio: marketing, procesamiento de nuevos negocios, gestión de cuentas y cobranzas.

En *marketing* una de las grandes preocupaciones o intereses que se persiguen es la adquisición de nuevas cuentas. Por esto los scores de riesgo y de respuesta suelen ser usados en conjunto, para lograr direccionar mejor la oferta, evitando costos por contactar a aquellos individuos que posiblemente no respondan favorablemente o que no superen las políticas de riesgo.

En el *procesamiento de nuevos negocios* es, quizás, la etapa donde los modelos de Credit Scoring tienen mayor popularidad. Se busca puntuar la solvencia

crediticia del postulante. En esta etapa no sólo se puede definir cuándo se acepta o se rechaza a un postulante sino también customizar la oferta en función a los scores y la información disponible.

En la etapa de *gestión de cuentas* se busca manejar las cuentas activas. Los behavioural scores (scores de comportamiento) se utilizan aquí para la administración de clientes existentes y se suelen actualizar a intervalos regulares de tiempo, por ejemplo una vez por mes.

En la etapa final de la gestión del crédito, se encuentra *cobranzas* y *recupero*. La particularidad de esta instancia es la urgencia de la tarea. Los clientes ya se encuentran en mora y se busca ordenar y valorizar la cartera para priorizar las gestiones cobros.

Se busca a través de esta descripción, entender la amplitud del concepto Credit Scoring y sus diferentes posibilidades de utilización para poder comprender mejor el alcance de este trabajo y visualizar de manera más clara su objetivo y a qué etapa de la gestión del crédito aplica. Dada esta descripción, se puede decir que este trabajo se centrará en el desarrollo de un modelo de Credit Scoring al cual se ha llamado Bureau Score, enfocado en medir el riesgo de crédito de un solicitante no bancarizado y pensado principalmente para ser utilizado en la etapa de procesamiento de nuevos negocios.

1.3 Técnicas de desarrollo de Credit Scoring

Existen diversas técnicas utilizadas para la construcción de los modelos de Credit Scoring. Estas se dividen en paramétricas, ya que requieren suposiciones sobre el comportamiento estadístico de los datos, y no paramétricas, que no requieren ningún supuesto. Dentro del primer conjunto, podemos encontrar regresión lineal, modelos probit y logit, análisis discriminante y dentro del segundo grupo, se encuentran algoritmos de árboles de decisión, redes neuronales, programación lineal, algoritmos genéticos, k-vecino más cercano, entre otros. Cada una de ellas tiene sus ventajas y desventajas dependiendo de las circunstancias y su aplicación dependerá del contexto del problema a resolver: estructura de los datos, las características usadas, la posibilidad de separar las poblaciones de interés usando esas características (población de buenos y malos) y el objetivo de clasificación (tasa de malos, clasificación

general, tasa de mala clasificación ponderada por el costo, tasa de malos dentro de los aceptados, alguna medida de rentabilidad, etc)[7]

Históricamente, el análisis discriminante y la regresión lineal eran las técnicas más ampliamente utilizadas para la construcción de modelos de Credit Scoring. Esto se debió a la sencillez de estas técnicas y a la disponibilidad de las mismas en diferentes software estadísticos. En la actualidad, la regresión logística, en general, es la técnica más utilizada para los modelos de Credit Scoring. Esto se debe a que es una técnica apropiada para estimar modelos cuya variables dependiente es binaria, las puntuaciones predichas suelen ser fácilmente traducibles en estimaciones de probabilidad, requiere menos supuestos y posee una mejor performance cuando las poblaciones de interés son muy desiguales.

A continuación se describen brevemente las técnicas más utilizadas para el desarrollo de estos modelos.

1.3.1 Técnicas Paramétricas

Las técnicas paramétricas realizan ciertos supuestos críticos sobre los datos subyacentes. Se describen a continuación: análisis discriminante, regresión lineal y modelos probit y logit.

Análisis Discriminante

El análisis Discriminante (DA) es una técnica estadística multivariada utilizada para clasificar una observación, dependiendo de sus características individuales, en uno de los grupos previamente definidos en los que se divide la población y que son excluyentes y exhaustivos entre sí. Busca identificar si existen diferencias significativas entre los grupos definidos y similitudes de los casos dentro de los grupos en función de un conjunto de variables independientes. De esta manera se establece una regla determinante que permita asociar a una observación con algunos de los grupos definidos con el menor error posible. Es utilizado en problemas donde la variable dependiente es categórica y permite establecer clasificaciones explicitas de dos o más grupos.

[7] Hand D.J., Henley W.E. (1995/6). Statistical Classification Methods in Consumer Credit Scoring: A review Journal of the Royal Statistical Society. Soc. A(1997) 160, Part 3, pp. 523-541.

Esta técnica considera simultáneamente todo el perfil completo de variables de una observación, identifica las características que definen a cada grupo y las interacciones entre ellas. Tiene la limitante de que sólo admite variables cuantitativas o métricas como independientes, requiere una serie de supuestos rígidos (linealidad de las relaciones, normalidad de las variables independientes, ausencia de multicolinealidad entre ellas, igualdad de matriz de covarianza entre grupos) y presenta la dificultad de que los resultados no son traducibles en probabilidades.

Al parecer la primera publicación del uso de esta técnica sobre modelos de scoring corresponde Durand (1941) donde mostraba que este método podría producir buenas predicciones del pago del crédito.

Regresión de probabilidad Lineal

Los modelos de probabilidad lineal (MPL) utilizan un enfoque de regresión por mínimos cuadrados ordinarios. Relacionan las variables independientes con la variable dependiente dicotómica de manera lineal, e intenta modelar a la variable dependiente tal como muestra la ecuación 1.3.1:

Ecuación 1.3.1: $Y_i = \beta' X_i + \varepsilon_i$, donde $E(\varepsilon_i / X_i) = 0$ y $E(\varepsilon_i) = 0$

Y_i es la variable dependiente que toma valores 0 y 1, β es un vector columna que agrupa a los p parámetros correspondientes a las variables explicativas que se encuentran en el vector columna X_i.

Si se define la probabilidad $P_i = \Pr(Y_i = 1/X_i)$, la esperanza de Y_i condicionado X_i es $E(Y_i/X_i) = P_i = \Pr(Y_i = 1/X_i) = \beta' X_i$. Es por esto que las estimaciones que se desprenden directamente de este modelo pueden ser interpretadas como la probabilidad de que ocurra el suceso definido como 1 en la variable dependiente. Una de las ventajas de este tipo de modelos, es la sencilla interpretación de los parámetros (igual al del Modelo Lineal General), ya que los mismos indican el efecto que una variación unitaria en cada una de las variables explicativas tiene sobre la probabilidad de que ocurra el suceso bajo estudio. Sin embargo, si bien las probabilidades son directamente estimadas por el modelo, esta técnica no tiene ninguna condición que restrinja las probabilidades al intervalo [0,1] por lo que la estimación de la variable dependiente podría tomar valores negativos o

mayores que 1, lo que carecería de sentido en el contexto de aplicación del método. Además, otro inconveniente que presentan es la violación de los supuestos de normalidad de los errores, homocedasticidad de la varianza. El primer supuesto no es válido dado que como ocurre con Y_i, ε_i puede tomar sólo dos valores, $\varepsilon_i = 1 - \beta' X_i$ cuando $Y_i = 1$ y $\varepsilon_i = -\beta' X_i$ cuadno $Y_i = 0$, por lo que no se podría decir que los errores se distribuyen normalmente. Por otra parte, podría demostrarse que las varianza de los errores, $V(\varepsilon_i) = P_i(1 - P_i)$, son heterocedásticos. La violación de estos supuestos provoca que los estimadores no sean eficientes, que las pruebas de hipótesis y los intervalos de confianza convencionales, no sean válidos. Además, en los modelos de esta característica, es común encontrar valores de R^2 subestimados, debido a que la suma de cuadrados de los residuos suele ser mayor a lo habitual.

Orgler (1970) fue el precursor del uso de esta técnica utilizándola en un modelo para préstamos comerciales. También recurrió a esta técnica un año después para construir un modelo de Credit Scoring para préstamos al consumo (Orgler, 1971), resaltando el buen poder predictivo de las variables sobre el comportamiento del cliente.

Modelo Probit y Logit

Considerando las dificultades mencionadas por el modelo de probabilidad lineal es posible transformar al modelo original para lograr acotar la probabilidad en el rango $[0,1]$. Si se desea obtener probabilidades entre 0 y 1, se requiere una función lineal monótona creciente que mapee la combinación lineal $\eta = \beta' X$ al intervalo $[0,1]$. Para esto, se puede adoptar un modelo para el cual los valores de probabilidad P_i estén circunscriptos al intervalo $[0,1]$: $P_i(\eta_i) = F(\beta' X)$. La idea consiste en utilizar una función de transformación $F(.)$ que tenga las siguientes características: diferenciable, monótona creciente y con rango $[0,1]$.[8]

El modelo no lineal sería $Y_i = F(\beta' X) + \varepsilon_i$ con $\varepsilon_i = E(Y_i / X_i) - F(\beta' X)$.

Existen varias funciones de distribución acumulada que cumplen estas características, sin embargo $F(.)$ suele tomar la forma de la función de

[8] Alamilla López N. E., Arauco Camargo S. (2009). Ensayos "Limitaciones del modelo lineal de probabilidad y alternativas de modelación microeconométrica"

distribución acumulada normal o logística. La primera da origen a los modelos Logit y la segunda a los modelos Probit.

Utilizando la distribución normal se tiene:

Ecuación 1.3.2: $P_i = F(\beta' X_i) = \phi(\beta' X_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta' X_i} e^{\frac{z^2}{2}} dz$ donde $z \sim N(0,1)$

Usando la distribución logística se tiene:

Ecuación 1.3.3: $P_i = F(\beta' X_i) = \frac{1}{1+e^{-\beta' X_i}} = \frac{e^{\beta' X_i}}{1+e^{\beta' X_i}}$

Estos modelos son estimados por máxima verosimilitud, el cual es un proceso iterativo y puede requerir varios intentos hasta alcanzar la convergencia. Esto requiere de un cálculo intensivo, lo cual fue posible gracias al desarrollo de la tecnología.

Generalmente, se prefiere el modelo Logit por la simplicidad de su función de distribución acumulada con respecto a la de la distribución normal, que involucra una integral y además, brinda una interpretación de los parámetros estimados más simple a través de la linealización del modelo.

En el Capítulo 2, se profundizará sobre los modelos Logit, ya que es la técnica elegida para modelar el problema tratado en este trabajo. Esto se debe a que: i) es una técnica apropiada para modelar variables dicotómicas, ii) los supuestos requeridos son menos pretenciosos comparados con otras técnicas, iii) las estimaciones arrojadas a partir de la utilización de estos modelos, son fácilmente traducibles a probabilidades iv) brinda estimaciones robustas de la probabilidad, dada la información disponible y v) es una técnica reconocida para el desarrollo de modelos de Credit Scoring.

Una de las primeras publicaciones sobre la regresión logística aplicada a modelos de Credit Scoring corresponde a Wiginton (1980), en donde también efectúa la comparación con análisis discriminante. Srinivasan y Kim (1987) también incluyó la regresión logística para comparar con otras técnicas. Srinivasan and Kim (1987) utilizaron la regresión logística para el proceso de evaluación de préstamos comerciales.

En la actualidad, se pueden encontrar una gran cantidad de trabajos que utilizan está técnica para modelos de Credit Scoring y comparan a la misma con otras técnicas.

1.3.2 Técnicas No Paramétricas

Se ha intentado utilizar técnicas no paramétricas para desarrollar modelos de Credit Scoring, ya que éstas no requieren prácticamente ningún supuesto para que su utilización sea válida. Algunas alternativas no paramétricas provienen del campo del aprendizaje automático como redes neuronales, algoritmos genéticos, k-vecino más cercano. Sin embargo, en algunos casos, estas técnicas son criticadas por falta de transparencia y un potencial sobreajuste. Otras técnicas no paramétricas son la: programación matemática y algoritmos de árboles de decisión.

Programación matemática

Los modelos de programación matemática se basan en optimizar un criterio objetivo como la proporción de solicitantes correctamente clasificados. Estos modelos permiten manejar una gran cantidad de variables, se adaptan a las necesidades de la entidad y no presentan inconvenientes cuando existe una relación determinística entre las variables. Proviene del campo de la investigación operativa para la asignación óptima de recursos. Incluye programación lineal, dinámica, entera, no lineal, entre otras.

En 1947 George Dantzig ideó el método simplex utilizado en el modelo de programación lineal mientras trabaja en la Fuerza Aérea de los Estados Unidos. El método resultó muy eficaz y fue rápidamente aplicado en otros ámbitos, también gracias al desarrollo de las computadoras.

Fueron Hand (1981), Showers y Chakrin (1981) y Kolesar y Showers (1985) quienes dieron el puntapié inicial para aplicar esta técnica en la actividad bancaria.

Árboles de decisión

Esta técnica consiste en una partición recursiva que busca la separación óptima de la muestra, de tal manera que las categorías de la variable respuesta ofrezcan

distintos perfiles de riesgo. Persigue el objetivo de crear grupos homogéneos y mutuamente excluyentes, minimizando la distancia entre los miembros de un grupo y maximizando la distancia entre los miembros de distintos grupos. Esta técnica ha ganado gran popularidad por la fácil interpretación de los resultados a través de la representación gráfica. Sin embargo, tiene la desventaja de que no es posible cuantificar la magnitud con la que cada variable contribuye en la predicción de la variable objetivo y dependiendo de la cantidad de datos que se tenga y de las aperturas que se realicen, se corre el riesgo de sobreajuste.

Esta metodología ha sido utilizada en diversas disciplinas como ciencia biológica, estadística e inteligencia artificial.

Friedman (1977) es de los primeros autores que utiliza la técnica de árboles de decisión para el desarrollo de modelos de Credit Scoring. Luego. Makowski (1985), Coffman (1986), Carter and Catlett (1987) utilizaron esta metodología para clasificar clientes en función del Score.

Redes Neuronales

Las redes neuronales son un algoritmo de inteligencia artificial inspirado en el funcionamiento de las redes neuronales de los organismos vivos. Busca a través de la auto-organización y aprendizaje automático (machine learning) captar la relación más adecuada entre las variables explicativas y la probabilidad de default, a la vez que determinar cuáles son las características más importantes que explican esta probabilidad. Este método permite procesar grandes volúmenes de datos, además es flexible ya que puede tratar relaciones no lineales y aprende de la información ya procesada para generar nuevos comportamientos. Sin embargo, el algoritmo que utiliza es muy complejo, costoso de implementar y difícil de interpretar, suele ser considerado como una “caja negra” y corren el riesgo de sobreajuste.

Rosenberg y Gleit (1994) describieron algunas aplicaciones de esta metodología a las decisiones de crédito corporativo y detección de fraude.

Algoritmos Genéticos

Los algoritmos genéticos están basado en la biología y los procesos de selección natural de Darwin. Esta metodología trata de encontrar un resultado óptimo a

través de un procedimiento sistemático. Tiene en cuenta como entrada a los individuos iniciales y selecciona quienes de ellos deben ser tenidos en cuenta para generar descendencia para la siguiente generación. De esta manera se itera hasta cumplir alguna de las condiciones de parada o hasta obtener un individuo que cumpla las restricciones iniciales.

Si bien los algoritmos genéticos se presentan como una herramienta dentro de Credit Scoring, rara vez se utilizan por su falta de trasparencia y por los grandes requerimientos de tecnología que implican.

Los principios básicos de los algoritmos genéticos fueron descriptos por Holland (1975). Fogerty y Ireson (1993) y Albright (1994) fueron uno de los primeros en describir la aplicación de esta técnica en los modelos de Credit Scoring.

K-vecinos más cercanos

Es una técnica no paramétrica muy simple. Se trata de clasificación supervisada y es utilizada para determinar la pertenencia a un grupo. Se parte de una base de entrenamiento que contiene todas las variables explicativas y también la clasificación deseada. Para un nuevo cliente se procede a calcular su distancia contra todos los casos analizados y sólo se consideran los k casos más cercanos al nuevo cliente. Finalmente este cliente se asignará al grupo al cual pertenece la mayoría de sus K vecinos mas cercanos.

En Credit Scoring está técnica no resulta muy práctica porque brinda sólo una clasificación y no un modelo de puntuación. Además, dependiendo de la cantidad de casos que se tengan en cuenta para el cálculo, puede tardar mucho tiempo dado que debe calcular la distancia de cada caso nuevo con todos los casos de la muestra y el proceso de asignación no es transparente.

Chatterjee y Barcun (1970) aplicó esta técnica en solicitantes de préstamos personales.

1.4 Modelos de Credit Scoring en población no bancarizada

Como se explicó anteriormente, los Burós de Crédito brindan a las entidades scores genéricos que se basan principalmente en el historial de pago de los individuos. Sin embargo, existe un amplio sector de la población que no cuenta

con productos de crédito. A este segmento se lo suele denominar POBLACION THIN, ya que los registros de información que poseen los Burós de Crédito sobre esta población son “delgados” por la escasez de datos. Los individuos pertenecientes a este segmento THIN, como dice Jennifer Tescher[9], se encuentran ante la dificultad de necesitar tener historial crediticio para obtener un crédito, pero a su vez necesitan tener crédito para construir una historia crediticia. De esta manera, es como este segmento, frecuentemente, queda relegado a productos y servicios financieros provenientes de entidades con menor formalidad que los bancos. Estas entidades suelen exigir tarifas más elevadas y muchas veces no son aportantes de los Burós de Crédito, por lo que no contribuyen en la construcción del historial de pago del individuo. Es posible encontrar sitios de internet, sobre todo de origen norteamericano, que brindan recomendaciones a este segmento de la población sobre cómo actuar para poder ir construyendo lentamente un historial crediticio, es decir, a que productos aspirar inicialmente.

En la actualidad, la *inclusión financiera* “se está convirtiendo en una prioridad para las autoridades, los organismos reguladores y las instituciones de desarrollo en todo el mundo”[10]. El Banco Mundial define *inclusión financiera* como el “acceso que tienen las personas y las empresas a una variedad de productos y servicios financieros útiles y asequibles que satisfacen sus necesidades —como pagos y transferencias, ahorro, seguros y crédito— y que son prestados de una manera responsable y sostenible.”

Más de 55 países, desde 2010, han firmado un compromiso para hacer frente a la inclusión financiera y muchos de ellos ya han elaborado estrategias nacionales para abordar este tema.

En Argentina, puntualmente, se llevaron a cabo algunas medidas que fomentaron la inclusión financiera. Entre estas medidas se puede mencionar la normativa del BCRA que impulsó la Cuenta Gratuita Universal, la bancarización de planes sociales, las políticas para localizar sucursales bancarias en zonas

[9] Jennifer Tescher es presidenta y CEO del Centro de Innovación en Servicios Financieros en Chicago.
[10] BancoMundial.org (2016). Inclusión financiera. Disponible: www.bancomundial.org/es/topic/financialinclusion/overview

con menor infraestructura bancaria y el Plan Nacional de Bancarización 2015-2019.

En este contexto, donde las entidades financieras incrementan sus nichos de mercado a personas carentes de información en Burós de Crédito, surge cada vez más el interés de buscar datos alternativas de información que permitan modelar de la mejor manera posible el riesgo de este segmento.

Se entiende como datos alternativos, todos aquellos que no provengan de datos de pago de bancos o de burós de crédito. Se pretende que los datos alternativos sirvan como fuente adicional de información con poder predictivo. Algunas de las fuentes alternativas de información pueden ser: datos de telefonía (TV, móvil, banda-ancha), servicios públicos (luz, gas, agua), registro de propiedades, alquiler, datos en línea, datos psicométricos.

En Estados Unidos, una serie de compañías desarrollan Score con datos alternativos para atacar este segmento. Entre estas compañías están: Cignifi, FactorTrust, First Access, eCredable, Happy Mango, Juvo, Neener Analytics, SharedLending, Trooly, TrustingSocial, ZestFinance[11]. Estas compañías se basan principalmente en información de pago y consumo de celulares, seguros, redes sociales y/o pagos a entidades no aportantes a los grandes Burós de Crédito. PRBC (Pay Rent, Built Credit) es otra empresa que desarrolla scores con información alternativa teniendo en cuenta información de pago de facturas (alquileres, pago de cuentas de celulares, internet y servicios públicos, préstamos estudiantiles, seguros). En este caso, el individuo debe crear una cuenta en el sitio web y cargar los datos de sus facturas para que luego PRBC verifique la información cargada y pueda calcular su score.

Así mismo, empresas vinculadas a scores con información tradicional de Burós de Crédito entienden que “la inclusión financiera para más personas es lo que el mercado está buscando” (Jim Wehmann, vicepresidente ejecutivo, Scores, FICO, 2016). Frente a estas necesidades del mercado actual, es que FICO, Equifax y LexisNexis Risk Solution se aliaron en U.S. para desarrollar FICO

[11] Mesropyan, E. (2017) Alternative Credit Scoring in the US: Innovators Applying Data Science to Unlock Financial Potential of ‘Thin-File’ Individuals. Medici. Disponible en: https://gomedici.com/alternative-credit-scoring-us-data-science-financial-potential-thin-file/

Score XD que utiliza información alternativa de datos de crédito para potenciar la predicción del comportamiento de pago en el segmento THIN. Este score combina información de pagos de teléfono fijo, celular e internet, datos de propiedades, registros públicos e información de Buró de Crédito, en caso de existir. Por su parte, TransUnion desarrolla un Score de crédito, llamado CreditVision Link, con información alternativa en el cual incorpora información sobre el comportamiento de pago de los individuos en entidades con menor formalidad que no aportan su información a los Burós de Crédito. También Experian lanzó al mercado "Extanded View" que incluye información de alquileres como parte del cálculo.

Esta práctica se fue extendiendo alrededor del mundo, lo cual permite transformar la manera de dar préstamos, lo que resulta muy importante sobre todo en países emergentes donde el porcentaje de no bancarizados es mayor.

En México hace unos años se publicó un trabajo empírico, *Metodología para un scoring de clientes sin referencias crediticias* (Espin-García, O. y Rodríguez-Caballero. C. (2013), en el cual se utilizan datos de un banco mexicano para analizar el segmento no bancarizado. En este se mostró que la información sociodemográfica que se puede recopilar en la solicitud resulta relevante para predecir el comportamiento de pago de este sector de la población.

EFL, empresa creada en 2006 como un proyecto de investigación en el Centro Internacional de Desarrollo de Harvard, desarrollo un score psicométrico. El mismo surge a partir de datos que se recopilan en un cuestionario que se le realiza al solicitante y "permite medir características como la confianza, autonomía, oportunismo, habilidades de razonamiento numérico y honestidad del individuo" (Klinger, 2015). EFL se convirtió en el principal proveedor mundial de este tipo de datos y actualmente brinda soporte a distintas empresas en África, Asia y Latinoamérica. Ofrece a las entidades financieras una evaluación digital para los solicitantes que culmina con el resultado del score psicométrico, que predice el comportamiento de pago a partir de la personalidad inferida mediante un cuestionario. EFL creó acuerdos con FICO, MasterCard, Equifax en algunos países para actuar como proveedor de este dato sobre todo para el segmento THIN.

De este modo, los datos alternativos trajeron al mercado una opción para llegar a la población THIN. Sin embargo, es importante tener en cuenta distintos factores que pueden condicionar su utilización.

Bailey Klinger cofundador de EFL, en su informe sobre *Scoring de Crédito Alternativo en Mercados Emergentes* (2015) propone dos indicadores para determinar si una fuente de datos alternativa debe ser considera: la disponibilidad y la capacidad de predicción de los mismos. En su informe, compara la eficiencia y disponibilidad de los datos en línea, datos móviles y datos psicométricos. Según Klinger, los datos en línea suelen tener baja cobertura y estar sesgados a un segmento de la población por lo que tienen una capacidad predictiva no muy alta, y además, pueden ser manipulados con facilidad y suelen ser difíciles de vincular a un individuo. Los datos móviles generalmente tienen una cobertura más alta y suelen ser más fáciles de vincular con los datos del individuo además de presentar mayor capacidad predictiva. En cuanto a los datos psicométricos, Klinger, plantea que su capacidad predictiva es muy buena, aunque depende de la calidad de las preguntas, y puede tener un gran potencial bien implementado.

Por otra parte, Schütte[12] propone tres dimensiones de análisis para tener en cuenta y evaluar la utilidad de cada fuente. Primero, sugiere que el analista debe considerar, si es que se trata de datos de pagos, si los mismos corresponden a productos o servicios asimilables a un crédito o a efectivo, es decir, si se recibe el beneficio antes de abonarlo. Se supone que mientras más similar es la transacción a un crédito estándar más útil será el dato, es el caso de los servicios públicos o las telecomunicaciones. La segunda dimensión que Schütte menciona es la cobertura que pueden tener esos datos y su relación con el costo, es decir, el volumen debe justificar el costo al cual se incurre por obtener esa información. Y por último, la viabilidad de concentración del dato, ya que podría provenir de muchos agentes distintos, como es el caso de pago de alquileres.

Es importante tener en cuenta también que para la utilización de fuentes alternativas de información se requiere revisar las regulaciones de cada país.

[12] Cheney, Julia S. (2008) Alternative Data and Its Use in Credit Scoring Thin and No-File Consumers. Federal Reserve of Philadelphia.

En Argentina, muchos de los datos alternativos mencionados no son tan simples de conseguir y aún no existen entidades que recopilen, regulen y administren ciertas fuentes de información. Como se mencionó anteriormente, muchas veces los costos por conseguir nuevas fuentes no son justificados por la baja cobertura, el bajo poder predictivo y la dificultad por vincular el dato a una persona.

Otro punto importante a tener en cuenta, es que estos datos, como los restantes que se utilizan para desarrollar el modelo, deben estar disponibles tanto al momento del desarrollo, como de la implementación del modelo. En el caso de los datos psicométricos, además de ser costosos por el tiempo que requieren para ser realizado el cuestionario (más de 15 minutos), tiene la desventaja de que no están disponibles para backtesting.

Analizando las posibles fuentes de información en función a las dimensiones que plantea Klinger, cobertura y predictibilidad, el Censo Nacional de Hogares, Viviendas y Personas se destaca por su disponibilidad y cobertura. Si bien la información Censal no se encuentra disponible a nivel de persona, sí se encuentra disponible a nivel de radio censal para todo el país. Los radios censales son pequeñas unidades territoriales cuyo tamaño depende de la densidad poblacional, pero pueden equivaler a una manzana en zonas muy pobladas. Por ejemplo, en Capital Federal, que es el objetivo de nuestro análisis, gran cantidad de radios censales equivalen a una manzana.

Con la premisa de que el entorno social de un individuo condiciona su comportamiento, se busca probar si las variables que describen el área en el que vive, contribuye a explicar su comportamiento de pago.

Capítulo 2: Descripción de datos y metodología

2.1 Descripción de nueva fuente de datos

El Censo Nacional de Población, Hogares y Viviendas de la República Argentina llevado a cabo en 2010 será evaluado como una fuente alternativa de información para el desarrollo del modelo de interés.

El Censo representa la mayor fuente de datos para conocer, cuantificar y analizar con el máximo nivel de desagregación la estructura demográfica, socioeconómica y la distribución espacial de la población.[13]

El Censo Nacional de Población, Hogares y Viviendas de la República Argentina es un censo de hecho, es decir, se obtienen datos de la persona en la vivienda donde haya pasado la noche anterior al día del relevamiento. Por este motivo, el operativo se lleva a cabo en un solo día para evitar duplicaciones de conteo. Se realiza cada 10 años permitiendo evaluar y medir los cambios ocurridos en una década como así también definir las bases para algunas de las políticas públicas de los siguientes 10 años. También constituye una gran fuente de información tanto para instituciones privadas como académicas.

Las unidades de análisis son tanto personas, como hogares y viviendas. En el Censo del 2010, al igual que el del año 1980 y 1991 se aplicó censo con muestra, es decir, a una muestra de viviendas particulares se les aplicó un cuestionario ampliado y a todas las restantes viviendas particulares, se les aplicó el cuestionario básico. Para viviendas colectivas, se aplicó un cuestionario específico.

El cuestionario básico, busca capturar variables sociodemográficas como: Sexo, Edad, Nivel Educativo, Características Básicas de la Vivienda, Condición de Actividad de las personas, etc. La información capturada a partir de este cuestionario permite construir los principales indicadores del país como Índice de Necesidades Básicas Insatisfechas (NBI), la Tasa de Desempleo o el Nivel Educativo de los Jefes de Hogar para cualquier agregación geográfica.

[13] Instituto Nacional de Estadísticas y Censo. (Argentina, 2010). *"Censo Nacional de Población, Hogares y Viviendas 2010 Censo del Bicentenario. Resultados definitivos Serie B Nº 2. Tomo 1"*.

El cuestionario ampliado, por su parte, es aplicado a una muestra probabilística. Éste, profundiza algunos aspectos tratados en el cuestionario básico, pero además indaga sobre otros aspectos de la población tales como Fecundidad, Pertenencia a pueblos originarios, Población Afrodescendiente, Previsión social y Cobertura de salud, entre otras.

La información relevada en el cuestionario ampliado se encuentra disponible en el REDATAM[14] a nivel de departamentos o partidos, mientras que la información relevada en el cuestionario básico se encuentra disponible a un mayor nivel de desagregación, a nivel de radio censal. Los radios censales, como se mencionó en la sección 1.4, son pequeñas unidades geográficas que en zonas muy pobladas pueden tener el tamaño de una manzana.

Es así que, utilizando esta fuente, es posible tener gran información sobre las características demográficas, habitacionales y sociales de cada uno de los radios que componen la Argentina y particularmente de los que componen la Capital Federal que es el objeto de estudio. La Capital Federal posee 3553 radios censales y para cada uno de estos se tiene la información agregada relevada en el cuestionario básico para los aspectos referidos a las personas, las viviendas y los hogares. De este modo, se cuenta con variables como: cantidad de hombres, cantidad de mujeres, cantidad de viviendas particulares, cantidad de ranchos, cantidad de viviendas con conexión satisfactoria a servicios públicos, cantidad de viviendas con calidad satisfactoria de materiales de construcción, cantidad de hogares con revestimiento en el interior o el cielorraso, etc, a nivel de radio censal. En el Anexo I se puede encontrar el listado completo de variables disponibles a nivel de radio censal.

Esta información tiene la ventaja de ser gratuita y como se dijo anteriormente se encuentra disponible para todo el país. Además se ofrece la cartografía de estos radios que permite ubicar exactamente cada código del radio en un lugar en el espacio. Esta cartografía junto al mapa de calles, permite ubicar cada radio de tal modo de saber entre que calles se circunscribe cada uno de ellos. Esto puntualmente es la clave para poder asociar a cada persona con las

[14] Para mayor información es posible visitar el sitio web del Instituto Nacional de Estadísticas y Censo de la República Argentina https://www.indec.gov.ar

características de su radio censal, ya que a partir del domicilio disponible en el Buró de Crédito es posible situar al individuo en el espacio y asociarlo con su radio de pertenencia y sus atributos. Profundizaremos sobre este punto en la sección 2.2.

2.1.1 Análisis en componentes principales

El análisis en componentes principales es una técnica que busca reducir la información contenida en los datos y facilitar su interpretación. Trata de explicar la relación entre las variables a partir de unas pocas variables no observables construidas como combinación lineal de las originales. En este trabajo el ACP se utiliza para definir un indicador que resuma la calidad de vida del radio censal con el fin de integrar este indicador al modelo de morosidad como variable explicativa.

Se toma como fuente la base de los 3553 radios censales de Capital Federal. Esta base contiene para cada radio las variables listadas en el Anexo I, donde cada una de ellas representa cantidad absoluta de hogares, viviendas o personas que cumplen determinada característica. A partir de esos datos se procede a relativizar cada una de las variables para que puedan ser comparables entre los distintos radios. Esto quiere decir que cada una de las variables es dividida por el total de las unidades de empadronamiento, correspondientes a la variable en cuestión, relevadas en ese radio. A modo de ejemplo, si originalmente la base posee como atributo el *total de hombres,* se construye la variable *porcentaje de hombres* como el cociente entre el *total de hombres* y el *total de personas* del radio; a partir del *total de jefes con universitario completo* se construye el *porcentaje de jefes con universitario completo* como el cociente entre el *total de jefes con universitario completo* y el *total de jefes* del radio; a partir del *total de viviendas tipo casa*, se crea el *porcentaje de casas* a partir del cociente entre el *total de viviendas tipo casa* y el *total de viviendas*; a partir del *total de hogares con computadora* se construye el *porcentaje de hogares con computadora* a partir del cociente entre el *total de hogar con computadora* y el *total de hogares*. Esta transformación se realiza con todas las variables para eliminar la influencia del total absoluto de viviendas, hogares y personas relevadas en el radio.

Se efectúa distintos análisis y trasformaciones a partir de las variables iniciales. Se realiza distintos análisis de correlaciones y pruebas de componentes principales con el fin de buscar un indicador que logre resumir en buena medida las características de un radio a partir de la información más relevante. Finalmente se seleccionan 14 variables que resultan de mayor interés para resumir las características de un radio censal. Estas variables involucran aspectos de la vivienda, del hogar y de las personas e incluyen también indicadores resúmenes elaborados por el INDEC. Las variables seleccionadas se listan en el Cuadro 2.1.1. La definición precisa de las mismas, tomada del diccionario que provee Instituto Nacional de Estadísticas y Censo de la República Argentina para las variables originales, se puede encontrar en el Anexo 2.

Cuadro 2.1.1: Listado de variables que caracterizan a los radios censales y fueron utilizadas para la construcción de componentes principales

Variables vinculadas al hogar	1	porc_compu_sí	Porcentaje de hogares que poseen computadora
	2	porc_hela_sí	Porcentaje de hogares que poseen heladera
	3	porc_baño_exclusivo	Porcentaje de hogares con baño exclusivo
	4	porc_combus_gas_red	Porcentaje de hogares con gas de red como combustible para cocinar
	5	porc_hacinam_más_3.00	Porcentaje de hogares con más de 3 habitaciones por cuarto
	6	porc_con_nbi	Porcentaje de hogares con al menos una Necesidad Básica Insatisfecha
variables vinculadas a las personas	7	porc_edujefes_alto	Porcentaje de jefes con nivel educativo alto
	8	porc_desocupado	Porcentaje de personas desocupadas
	9	porc_lee_ escribe_sí	Porcentaje de personas que saben leer y escribir
	10	porc_usa_compu_sí	Porcentaje de personas que usan la computadora
Variables vinculadas a las viviendas	11	porc_agua_red	Porcentaje de viviendas con agua de red
	12	porc_calidad_mat_1	Porcentaje de viviendas con Calidad de Material I
	13	porc_construc_satis	Porcentaje de viviendas con construcción satisfactoria
	14	porc_desag_red	Porcentaje de viviendas con desagote a red pública

En el cuadro 2.1.2 se muestran descriptivos básicos de las variables listadas.

Cuadro 2.1.2: Descriptivos básicos de las variables involucradas en los componentes principales para Capital Federal.

Variables	Cantidad de Casos	Media	Desviación estándar	Mínimo	p25	p50	p75	Máximo
porc_compu_sí	3.553	68,45	13,52	0,00	63,85	71,09	77,04	98,33
porc_hela_sí	3.553	97,29	5,75	0,00	97,52	99,27	99,73	100,00
porc_baño_exclusivo	3.553	94,69	8,87	0,00	93,89	98,18	99,57	100,00
porc_combus_gas_red	3.553	92,46	18,86	0,00	94,90	98,01	99,12	100,00
porc_hacinam_más_3.00	3.553	1,51	2,78	0,00	0,24	0,59	1,45	24,74
porc_con_nbi	3.553	6,15	9,59	0,00	0,60	1,88	7,59	88,89
porc_edujefes_alto	3.553	41,12	15,59	0,00	31,81	41,92	52,86	81,97
porc_desocupado	3.553	3,07	1,15	0,00	2,28	2,92	3,69	8,42
porc_lee_escribe_sí	3.553	96,50	2,22	0,00	95,96	96,74	97,45	100,00
porc_usa_compu_sí	3.553	74,65	9,04	0,00	71,08	76,18	80,61	97,54
porc_agua_red	3.553	99,56	2,07	0,00	99,62	100,00	100,00	100,00
porc_calidad_mat_1	3.553	86,39	15,08	0,00	84,96	90,16	93,88	100,00
porc_construc_satis	3.553	84,80	16,07	0,00	82,70	89,07	93,24	100,00
porc_desag_red	3.553	98,99	4,13	0,00	99,30	99,77	100,00	100,00

Estos valores dejan en evidencia una realidad de la Capital Federal. Permite observar, entre otras cosas, que la mitad de los radios de CABA poseen al menos el 2% de sus viviendas sin baño exclusivo. También se puede observar que existen radios que poseen la mayoría de sus hogares con condiciones básicas insatisfechas, ya que el máximo de *porc_con_nbi* supera el 80%. Este cuadro también deja en evidencia que mientras existen radios que poseen todas sus viviendas construidas de manera satisfactoria, un cuarto de los radios de CABA posee al menos el 18% de sus viviendas con construcción básica o insuficiente. Estas estadísticas varían mucho dependiendo de la geografía. En este trabajo el foco estará en Capital Federal que es el objetivo de estudio.

Para entender el sentido y fortaleza de la relación entre estas variables que describen al radio, se construye la matriz de correlación (Anexo III). A partir del análisis de la misma se puede observar la existencia de fuertes vínculos entre las distintas variables, incluso entre variables que originalmente caracterizaban a distintas unidades de análisis (personas, hogares o viviendas). Son muchas las relaciones que se podrían citar entre estas variables. A modo de ejemplo, si tomamos la variable porcentajes de hogares con computadora en el radio observamos que presenta una fuerte correlación positiva con el porcentajes de hogares con tenencia de heladera, de baño exclusivo, de gas de red como combustible para cocinar, con el porcentaje de jefes de hogar con nivel educativo

alto, con el porcentaje de personas que saben utilizar la computadora, con el porcentaje de viviendas con buena calidad de materiales y con construcción satisfactoria, mientras que presenta una fuerte correlación negativa con el porcentaje de hogares con hacinamiento y NBI.

Si bien a partir de la matriz de correlaciones se pueden observar fuertes relaciones entre las variables, se lleva a cabo el cálculo de dos medidas que permiten evaluar si efectivamente es apropiado utilizar componentes principales. Por un lado, el test de esfericidad de Bartlett nos permite decir que la matriz de correlación de estas variables es significativamente distinta a la identidad y por otro lado, el índice Kaiser-Meyer-Olkin nos indica que es "meritorio" un análisis de componentes principales. En el Anexo IV se puede encontrar una breve explicación de estas medidas con los resultados obtenidos.

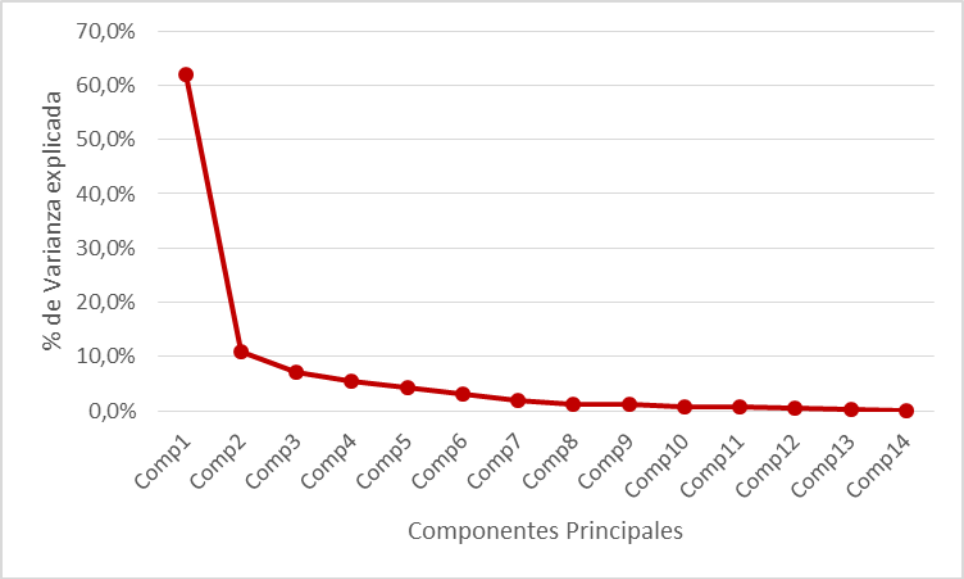
Asumiendo entonces que la correlación entre las variables es alta para aplicar una factorización, se avanza con el objetivo de resumir la información aportada por las variables originales a partir de un componente que surja como la combinación lineal de ellas y que pueda ser interpretado.

Si bien las variables que se utilizan están relativizadas y sus rangos de variación están limitados en el intervalo [0,100], los componentes principales fueron calculados a partir de la matriz de correlaciones para evitar que las variables con mayor variabilidad tengan mayor peso en el análisis.

A partir de la matriz de correlación de rango p (14) se calculan sus autovalores $\lambda_1, \dots, \lambda_p$ resolviendo $|R - \lambda I| = 0$ y sus vectores asociados $(R - \lambda_i I)a_i = 0$. Los componentes principales surgen de aplicar la transformación ortogonal A a las variables originales X para obtener unas nuevas variables Z no correlacionadas entre sí, $Z = XA$ donde $A'A = I$.

De este modo se obtienen 14 componentes ortogonales entre sí (no observables) que surgen como combinación lineal de las variables originales. Se obtiene un primer componente que logra explicar más de 60% de la variabilidad total de las variables y los restantes 13 componentes explican el resto de la variabilidad. En el siguiente gráfico se puede observar el porcentaje de varianza total explicada por cada componente (λ_i/p).

Gráfico 2.1.1: Porcentaje de varianza total explicada por cada componente



Los 3 primeros componentes son los que poseen autovalores asociados mayores a 1. (Ver Anexo V).

Conservando en una primera instancia estos 3 primeros componentes principales, se analiza la interpretación de los mismos para evaluar si es posible alcanzar el objetivo planteado inicialmente.

Para esto se calculan las correlaciones entre cada variable original estandarizadas y estos 3 primeros componentes. Dado que se parte de la matriz de correlaciones para el cálculo de componentes principales es posible obtener esta correlación con el siguiente cálculo $a_{ij} * \sqrt{\lambda_j}$. En el Anexo VI se encuentran los autovectores (a_i) de los primeros 3 componentes. En el cuadro 2.1.3 y el gráfico 2.1.2 se muestran estas correlaciones.

Tanto en el Gráfico 2.1.2 como en el Cuadro 2.1.3 es posible observar que el componente 1 guarda una fuerte correlación con todas las variables y en el sentido esperado. Es decir, en la mayoría de las variables involucradas en el análisis (11 de las 14 variables) se espera que cuanto mayor sea el porcentaje en el radio, mejor sean las condiciones de vida en el mismo. Tres de las variables involucradas tienen un sentido contrario a este, es decir cuanto mayor es el porcentaje de desocupados, de hogares con hacinamiento y de hogares con NBI, se espera que las condiciones del radio sean peores. Justamente estas últimas son las que se correlacionan fuertemente de manera negativa con el primer

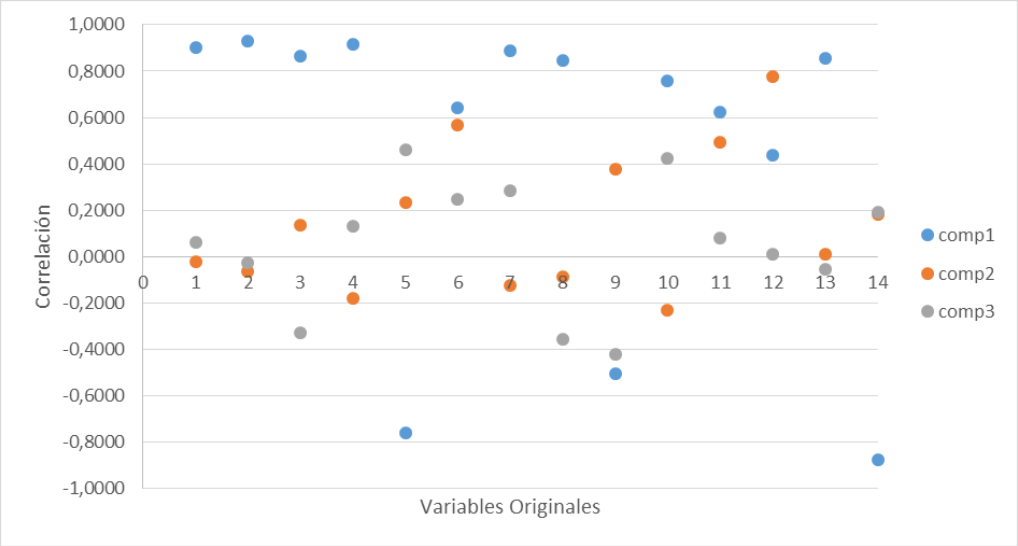
componente, mientas que todas las demás lo hacen de manera directa. Por este motivo, se podría decir que el primer componente resume en cierta medida la calidad de vida del radio y se podría esperar que cuanto mayor sea el valor de este indicador, mejores serán las condiciones de vida del mismo.

Considerando que con esta componente se obtiene el indicador que cumple en gran parte con nuestro objetivo inicial y además, observando que el componente 2 y 3 no poseen una interpretación clara y que en conjunto explican menos del 20% de la variabilidad total, se avanza principalmente con el análisis del primer componente.

Cuadro 2.1.3: Correlación de cada variable original con cada uno de los primeros 3 componentes principales

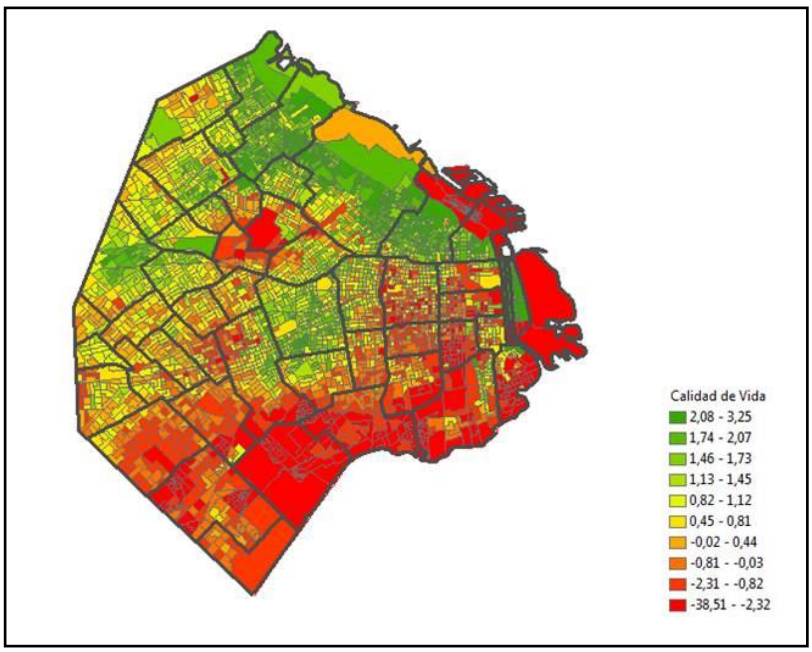
Variables	comp1	comp2	comp3
porc_compu_sí	0,9180	-0,1789	0,1316
porc_hela_sí	0,8644	0,1340	-0,3294
porc_baño_exclusivo	0,8468	-0,0882	-0,3546
porc_combus_gas_red	0,8561	0,0110	-0,0549
porc_hacinam_más_3.00	-0,8753	0,1809	0,1925
porc_con_nbi	-0,7586	0,2338	0,4599
porc_edujefes_alto	0,7598	-0,2322	0,4221
porc_desocupado	-0,5047	0,3795	-0,4206
porc_lee_escribir_sí	0,6430	0,5679	0,2489
porc_usa_compu_sí	0,8873	-0,1259	0,2823
porc_agua_red	0,4356	0,7756	0,0098
porc_calidad_mat_1	0,9000	-0,0227	0,0615
porc_construc_satis	0,9292	-0,0659	-0,0251
porc_desag_red	0,6220	0,4949	0,0782

Gráfico 2.1.2: Correlación de cada variable original con cada uno de los primeros 3 componentes principales



Se busca contrastar el nuevo indicador de calidad de vida con la realidad y así reducir la abstracción. Para lo cual se construye el mapa que muestra la distribución de este indicador en el espacio. En el gráfico 2.1.3 se muestra el componente 1 asociado a cada radio en el mapa de la Capital Federal.

Gráfico 2.1.3: Indicador de Calidad de Vida para Capital Federal a nivel de Radio Censal.



A partir del análisis gráfico, lo primero que se puede observar es una clara diferencia entre la zona sur de CABA y la zona Norte. Se destaca el corredor Norte con las mejores condiciones de vida, es decir, los radios ubicados, en Puerto Madero, Retiro, Recoleta, Palermo, Belgrano, Nuñez. Sin embargo, se observa también como el sector de la villa 31, ubicado en este área, se destaca con un bajo valor del indicador. En el centro de CABA, se ubica con altos valores Caballito y con bajos valores La Paternal y parte de Chacarita y hacia la izquierda se destaca con altos valores sobre todo Villa Devoto. El indicador elaborado pareciera estar acorde a la realidad conocida de la Capital Federal. (En el Anexo VII se encuentra el mapa con los barrios porteños)

2.2 Construcción de la muestra de desarrollo del modelo

Como se mencionó anteriormente, el trabajo aquí presentado busca desarrollar un “bureau score”, es decir, se trata de un modelo genérico que busca colaborar con las entidades del mercado en general, para discriminar en la población THIN los futuros buenos y malos clientes al momento de originar un producto de crédito.

El modelo genérico, busca contribuir con todas las entidades del mercado, por tal motivo considera información del mercado en general y no utiliza en su desarrollo una cartera específica de una entidad, ni tampoco la performance que un individuo haya tenido en una entidad o con un producto en particular. Siguiendo esta misma línea tampoco tiene en cuenta datos de la solicitud de los créditos, ya que es información propia y particular de cada entidad.

Diseño de la muestra

La base bajo estudio consiste en una muestra aleatoria de individuos que fueron consultados por alguna entidad del mercado en el Buró de Crédito en algún momento dentro del periodo que trascurre desde Agosto 2012 hasta Julio 2013. Se consideran los 12 meses distintos para evitar la estacionalidad. Además, para la extracción de la muestra se consideró únicamente el universo de individuos con domicilio en Capital Federal y que eran THIN al momento de la consulta. Se entiende que una persona es THIN, o no bancarizada, si no posee líneas de crédito informadas en BCRA o en Buró en los 5 años anteriores al momento de la consulta.

Cada una de las personas de la muestra fue enriquecida con la información disponible en el Buró de Crédito al momento de la consulta, con una historia de 5 años. Algunas de las variables disponibles en el Buró para esta población son: cantidad y tipo de consultas realizadas al Buró de Crédito en el pasado, antigüedad de las mismas, variables demográficas, información sobre tarjetas informada sin uso (tarjeta preembozada), condición de actividad, observaciones de morosidad en empresa de telecomunicación o retailers. En total, se dispone de aproximadamente 200 variables para la población analizada. Además, a partir del domicilio fue posible enriquecer esta base de datos con la información de la nueva fuente propuesta, el Censo Nacional de Hogares, Viviendas y Personas de 2010 y el indicador construido desarrollado en la sección 2.1.1, de esta manera, es posible adicionar alrededor de 150 variables, descriptas en la sección 2.1.

A todas las personas de la muestra se les calcula el objetivo de predicción (Good Bad).

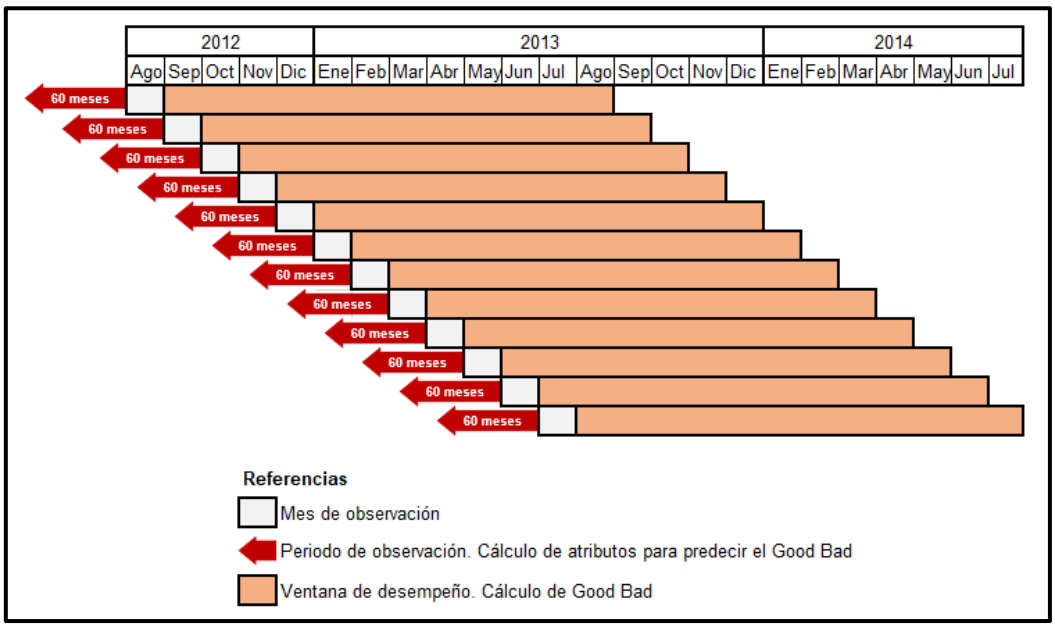
Definición del Objetivo de predicción (Good Bad)

La mayoría de las entidades del mercado aportan la información de sus líneas de crédito (tarjetas de crédito, préstamos y descubiertos en cuenta corriente) al Buró de Crédito, mientras que las entidades reguladas también lo hacen al BCRA. Esta información permite entre otras cosas conocer altas de productos de las personas y atrasos en los pagos. Mediante esta información fue posible construir la variable a predecir en el modelo. Para esto, cada una de las personas de la muestra fue observada durante el año siguiente a la consulta, para evaluar si dieron de alta un producto de crédito informado en Buró o en BCRA y si incurrieron o no en mora. Para el desarrollo del modelo, se consideran a aquellos individuos que tienen un comportamiento definido, o sea, que cumplen la condición de Malo o de Bueno. Se califica como Malo a aquel individuo que dentro de los 12 meses siguientes a la consulta registra en Buró o en BCRA un atraso mayor a 90 días reportado por alguna de las entidades aportantes. Por otra parte, se consideran como Buenos, a aquellos individuos que habiendo dado de alta un producto de crédito no registran atrasos en los 12 meses siguientes a la consulta.

La muestra final de 22.248 casos fue dividida de manera aleatoria en dos muestras con igual cantidad de casos: una muestra de entrenamiento, que se utilizará para desarrollar el modelo (metodología descripta en la Sección 2.3) y una muestra de validación, que será utilizada para evaluar la performance del modelo construido (Capitulo 4). No se incluyen en la muestra a aquellos individuos que no cumplen con la definición de Bueno o Malo. Es decir, se excluyen a aquellos individuos que en la ventana de desempeño tuvieron una mora mayor a 30 días de atraso pero menor a 90 y a aquellos individuos que no habiendo tenido atrasos, presentaron comportamiento por menos de 6 meses en la ventana de desempeño y no pueden ser considerados buenos. Tanto la muestra de entrenamiento como la de validación poseen similares características. La tasa de mora en ambas, bajo la definición presentada, es del 14%.

En el gráfico 2.2.1 se esquematiza la construcción de la muestra total.

Gráfico 2.2.1: Esquema de diseño de la muestra



2.3 Metodología de desarrollo del Modelo de Credit Scoring

Como se adelantó en la Sección 1.3.1, la metodología utilizada para llevar a cabo la estimación de este modelo de Scoring es el Modelo Logit, por las ventajas ya mencionadas.

Mediante la estimación de este modelo se intenta ordenar la población de interés en función de la probabilidad de incurrir en mora. Para cumplir este objetivo se considera la muestra de entrenamiento compuesta por 11.124 personas de Capital Federal que fueron consultadas al Buró de Créditos para dar de alta algún producto financiero entre Agosto 2012 y Julio 2013.

Teniendo en cuenta esta base se efectúa un análisis exhaustivo de las variables disponibles.

Análisis Univariado

El análisis de cada variable individualmente permite entender el dominio de cada una de ellas, la cantidad de valores perdidos, la variabilidad, la detección de valores extremos.

En términos generales, la muestra analizada posee similar cantidad de hombres que de mujeres y contiene una gran concentración de personas jóvenes, más del 50% posee edad menor a 35 años. Por otra parte, el 90% de las personas analizadas no tiene observaciones (ni mora en telco ni en retailers) registradas en Buró al momento de la consulta, pero el 94% había sido consultada alguna vez en los últimos 5 años, aunque sólo el 57% fue consultada alguna vez en los últimos 3 meses.

Gráfico 2.3.1: Distribución de la muestra según Sexo

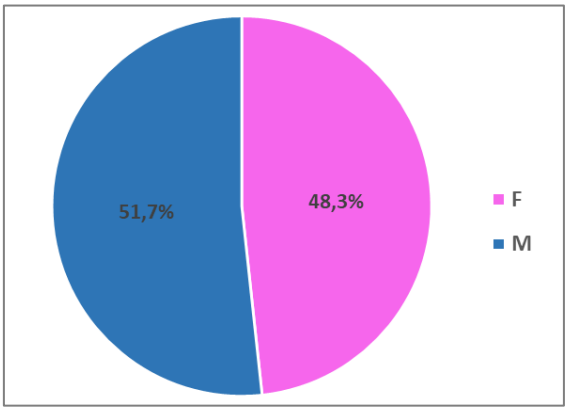


Gráfico 2.3.2: Distribución de la muestra según Rangos de Edad

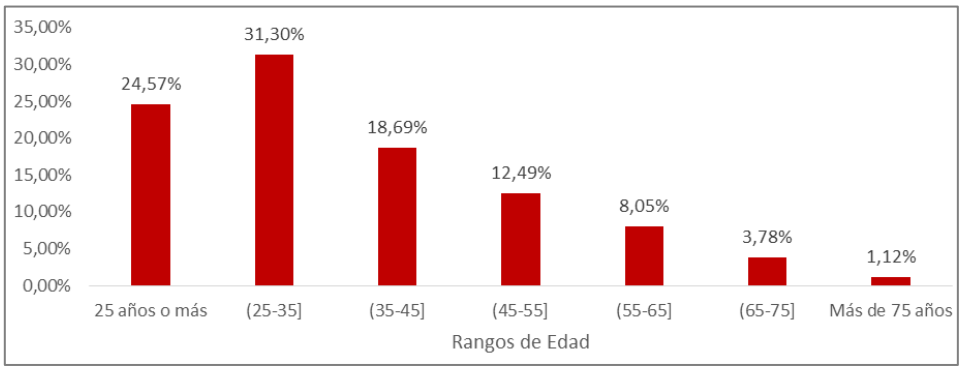


Gráfico 2.3.3: Distribución de la muestra según Cantidad de Consultas en los últimos 5 años

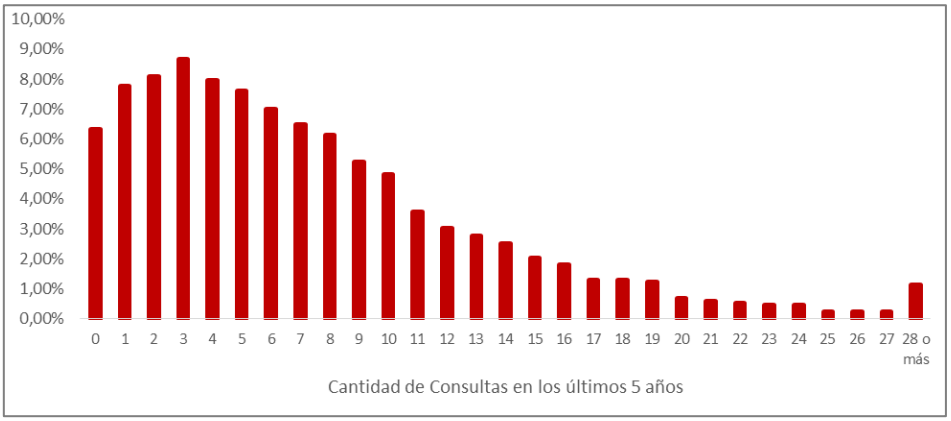
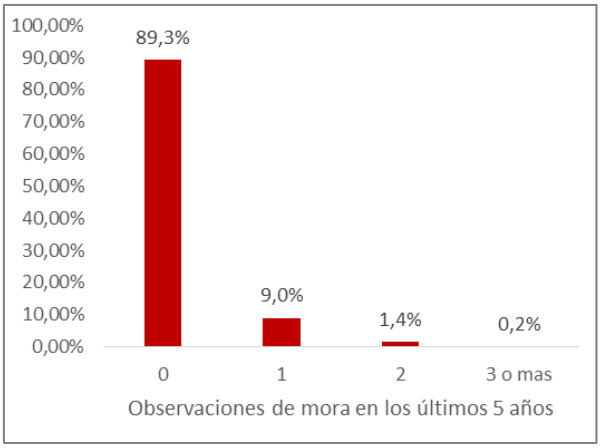


Gráfico 2.3.4: Distribución de la muestra según Cantidad de Observaciones en los últimos 5 años



Análisis de Poblamiento

En primer lugar, se realiza el análisis de poblamiento de cada una de las variables, para realizar una depuración inicial de la base. Es decir, para cada uno

de los atributos disponibles, se calcula el porcentaje de valores válidos (diferentes de cero o perdidos). De esta manera, es posible excluir a todos aquellos atributos que no aportan información por poseer el mismo valor para la mayoría de los registros. Alrededor de 160 atributos de Buró y del Censo fueron eliminados por tener más del 98% de sus valores con cero o valores perdidos.

Análisis Bivariado y cálculo del Information Value

El análisis bivariado consiste en el cruce de cada una de las variables disponibles con el Good Bad. En el caso de variables categóricas, se calcula la tasa de malos para cada una de las categorías de dicha variable. En el caso de variables continuas, en general se trabaja con deciles. Si existiesen valores perdidos, se conserva como una categoría aparte, a fin de analizar puntualmente este valor para efectuar la imputación del mismo.

El análisis bivariado ocupa un rol fundamental porque permite dar una idea más clara del sentido y la relación que se espera tenga cada variable con respecto al Good Bad. Además este análisis permite determinar la necesidad o conveniencia de realizar recodificaciones o transformaciones de la variable original. La técnica de árboles de decisiones es utilizada en ocasiones para definir la categorización óptima de variables independientes. En caso de crearse nuevos atributos, los mismos serán evaluados como nuevas variables.

A partir del análisis bivariado construido con todas las variables de la base y los nuevos atributos, es posible evaluar el poder predictivo de cada una de estas características. Para esto, Fair Isaac incorporó una medida que denominó Information Value, al cual denotamos como IV. Este indicador se calcula de acuerdo a la ecuación 2.3.1:

Ecuación 2.3.1: $IV = \sum_{i=1}^{n} \left[\left(\frac{G_i}{\sum G} - \frac{B_i}{\sum B} \right) \text{x}\, WOE_i \right]$

Donde $G =$ Buenos, $B =$ Malos, $\frac{G_i}{\sum G} =$ Porcentaje de Buenos, $\frac{B_i}{\sum B} =$ Porcentaje de Malos, $WOE_i = Ln\left[\left(\frac{G_i}{\sum G} \right) / \left(\frac{B_i}{\sum B} \right) \right]$ peso de la evidencia, $i =$índice entre las categorías de la variable evaluada, y $n =$ número total de categorías de la variable.

El IV suele ser difícil de interpretar porque no hay test estadísticos asociados a este indicador. Se toma como referencia que IV menor a 0,10 indica que una variable tiene un poder predictivo débil.[15] Sin embargo, hay que tener en cuenta que este indicador depende de las categorías definidas, tanto de las agrupaciones realizadas como de la cantidad de categorías, además de que evalúa a cada variable de manera independiente de las demás. Es por esto que aquí no se utiliza el IV para descartar las variables indiscriminadamente a la hora de estimar el modelo de Scoring, sino para tener una idea inicial sobre qué variables presentan mayor poder predictivo.

Entre las variables de Buró, las variables relacionadas con la cantidad de consultas cercanas al punto de observación y las observaciones de morosidad reportada por empresas de telecomunicación son las que poseen el IV más alto, rondando 0,40. Mientras que las variables relacionadas con la cantidad de tarjetas preembozadas son las que arrojaron el menor IV, rondando el 0,004.

En cuanto a las variables provenientes del Censo, el indicador de Calidad de Vida es el de mayor IV (0,15), mientras que el porcentaje de hogares cuyo combustible usado principalmente para cocinar es Gas a granel posee el IV más bajo (0,02).

Tratamiento de Valores Perdidos

Algunos de las variables involucradas en el análisis poseen valores perdidos para algunos de los registros de la muestra. Es importante entender el significado de la ausencia de valor para realizar la imputación.

En las variables provenientes del Censo, la existencia de valores perdidos se debe a que el domicilio del individuo registrado en Buró no pudo ser geocodificado y por lo tanto no puede ser vinculado a un radio censal.

En cuanto a las variables de Buró, el valor perdido en aquellas variables que representan cantidad, en general, suele significar ausencia y podría ser reemplazado por un 0, pero no siempre se da este caso.

[15] Anderson. R. (2007). The Credit Scoring Toolkit: theory and practice for retail Credit Risk Management and Decision Automation. Oxford University Press.

Para reemplazar los valores faltantes, se efectúa un análisis de los bivariados. De este modo es posible entender qué comportamiento tienen las distintas categorías de la variable para poder imputar el valor perdido por la categoría de la variable que muestre un comportamiento similar, es decir, similar tasa de malos o WOE. Como se dijo anteriormente, en el caso de que los valores faltantes superen el 98% de los casos, la variable es descartada.

Análisis de valores extremos

Las estimaciones que se efectúen con el modelo pueden verse afectadas por valores extremos. Por el tipo de variables involucradas, los valores extremos se pueden encontrar en los valores superiores de las variables, más que inferiores. Es necesario analizar el rango que tendrá cada una de las variables involucradas. Esto es importante dado que el score será calculado continuamente para nuevas personas, por lo que en la implementación del modelo será una condición importante definir el valor máximo y mínimo que puede tomar cada variable y así recodificar cada una de las variables involucradas antes de ser introducidas en la fórmula que brinda el score.

Para el desarrollo del modelo, es usual truncar las variables a partir de determinado valor, en este caso el percentil 99. Es decir, para las variables del Buró de Créditos, se le asignó el percentil 99 a todos aquellos individuos que poseen en estas variables un valor superior al mismo.

Estimación del Modelo de Scoring

El modelo consiste en estimar mediante variables explicativas (X_i) el comportamiento de la variable dependiente Bueno/Malo.

Si definimos el modelo logístico:

Ecuación 2.3.2: $y_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$

Se supone que ε_i sigue una distribución logística y y_i^* es una variable "latente" y lo que se observa es y_i, definida según la ecuación 2.3.3:

Ecuación 2.3.3: $y_i = \begin{cases} 1 \; si \; y_i^* > 0 \\ 0 \; si \; y_i^* \leq 0 \end{cases}$

Teniendo en cuenta estas definiciones se tiene:

Ecuación 2.3.4: $P_i = Prob(y_i = 1) = Prob(y_i^* > 0) = F(\beta' X_i) = \frac{1}{1+e^{-\beta' X_i}}$

El modelo a estimar intenta predecir la probabilidad de incurrir en mora según la función descripta.

A través de la linealización del modelo, se puede obtener una interpretación más clara de los parámetros estimados. (Ecuación 2.3.5)

Ecuación 2.3.5: $\ln\left(\frac{P_i}{1-P_i}\right) = ln\left(\frac{\frac{1}{1+e^{-\beta' X_i}}}{\frac{1}{1-\frac{1}{1+e^{-\beta' X_i}}}}\right) = ln\left(\frac{\frac{e^{\beta' X_i}}{1+e^{\beta' X_i}}}{\frac{1}{1+e^{\beta' X_i}}}\right) = \ln\left(e^{\beta' X_i}\right) = \beta' X_i$

La ecuación 2.3.5 es conocida como logit, que es el logaritmo natural de los odds $\left(\frac{P_i}{1-P_i}\right)$. El odds es el cociente entre la probabilidad de que el suceso ocurra (individuo incurra en default) y la probabilidad de que el suceso no ocurra y da cuenta de cuantas veces más probable es que el fenómeno suceda frente a que no suceda.

Este modelo es estimado mediante máxima verosimilitud por lo cual se utilizan software estadísticos capaces de resolver este proceso iterativo que requiere de cálculos intensivos. Más allá del componente tecnológico, se considera de gran importancia el conocimiento y experiencia del analista, ya que es quien diseña el experimento, determina los criterios para la selección del modelo, revisa manualmente los resultados y determina la coherencia de los mismos en base a su experiencia y conocimientos teóricos y prácticos, además de tener en cuenta las regulaciones legales vigentes de ese país sobre el uso de datos.

Como se mencionó al inicio, los supuestos de este tipo de modelos son menos estrictos que otros modelos paramétricos. Las exigencias de la metodología son:

1. Variables objetivo categórica. La variable dependiente binaria que indica si el individuo resulta bueno o malo, fue recodificada de tal manera de identificar como 1 a los individuos malos y con 0 a los que resultaron ser buenos pagadores.
2. Relación lineal entre las variables y el logaritmo de los odds, como se muestra en la ecuación 2.3.5.

3. Variables independientes no correlacionadas.
4. Independencia del término de error.
5. Uso de variables relevantes.

Para arribar al modelo final, inicialmente se tuvieron en cuenta todas las variables que poseían un poblamiento superior al 2%. Se utilizó la metodología *Forward Stepwise (Wald)*, que parte de un modelo sin variables y va incorporando una a una las variables que mejor explican los residuos. El modelo resultante es un modelo con un amplio set de variables que suele estar sobreajustado a la muestra de entrenamiento y con problemas de multicolinealidad. Sin embargo, este conjunto amplio de variables relevantes es utilizado para probar sucesivos modelos que surgen de incorporar de a una estas variables, respetando el orden en el cual entraron en este modelo, y eliminando aquellas que mostraban problemas de multicolinealidad. En cada uno de los modelos estimados se evalúa la performance que el modelo evidencia en las muestras de entrenamiento y validación. Es importante tener en cuenta que existe un momento en el cual la incorporación de una nueva variable no provoca mejora en la performance y redunda en un sobreajuste del modelo en la muestra de entrenamiento y hasta una baja de la performance en la muestra de validación.

Luego de identificar el conjunto de variables candidatas para el modelo final, se procede a estimar el modelo con estas variables y chequear el signo de los coeficientes, verificar que no haya multicolinealidad (a partir del Factor de Inflación de la Varianza), revisar la significatividad de todas las variables y evaluar la necesidad de incorporar alguna variable adicional que podría haber sido descartada en algún paso y de interés para el negocio. Para todas las variables involucradas se analiza minuciosamente el resultado del análisis bivariado para verificar que las variables que intervienen en el modelo ordenen adecuadamente la tasa de malos en todo el dominio de la variable.

Con el modelo final estimado, es habitual realizar una transformación de manera tal que en lugar de estimar la probabilidad de malo, el modelo brinde una puntuación que varíe entre 1 y 999, donde los puntajes más altos se relacionen a una menor probabilidad de incurrir en mora. Esta puntuación es conocida como Score.

El tratamiento de las variables realizado en la muestra de entrenamiento y la fórmula estimada del modelo final son aplicados a la muestra de validación para verificar la performance del modelo, con el objetivo final de lograr un ordenamiento de la tasa de mora de la población bajo estudio.

Capítulo 3: Presentación de Resultados

A continuación se presentan los resultados obtenidos a partir de la utilización de la metodología descripta en la muestra de entrenamiento.

3.1 Descripción de las variables del modelo

Las variables que intervienen en el modelo se muestran en el Cuadro 3.1.1.

Cada una de estas variables originales fue analizada en la muestra de entrenamiento para evaluar la necesidad de ser recodificada, transformada o bien si requiere imputación de valores perdidos. En esta sección se muestra el análisis bivariado de las variables tal como fueron introducidas en el modelo final.

Cuadro 3.1.1: Listado de variables que intervienen en el modelo por fuente

Fuente	Variable	Descripción
Buró	1.Edad	Edad
Buró	2.Sexo	Sexo
Buró	3.Cant_TC_Preembozadas_12m	Cantidad de Tarjetas de Crédito Preembozadas en los últimos 12 meses
Buró	4.Posee_empleador	Flag que indica si posee algún empleador informado
Buró	5.Obs_mora_60m	Cantidad de observaciones de morosidad (mora en telco o retail) en los últimos 60 meses
Buró	6.Meses_cons_TC2	Meses desde la última consulta realizada por una entidad emisora de TC, de segundo nivel
Buró	7.Cant_cons_3m	Cantidad de consultas totales en los últimos 3 meses
Buró	8.Cant_cons_no_Fi_12m	Cantidad de consultas de entidad no financieras en los últimos 12 meses
Buró	9.Cant_cons_TC1_24m	Cantidad de consultas de entidad emisores de TC de primer nivel, en los últimos 24 meses
Buró	10.Cant_cons_Fi1_60m	Cantidad de consultas de entidades financieras no bancarias de primer nivel, en los últimos 60 meses
Buró	11.Cant_cons_Fi2_60m	Cantidad de consultas de entidades financieras no bancarias de segundo nivel en los últimos 60 meses
Censo	12.porc_jefe_post_univ	Porcentaje de jefes del radio en el que vive en el individuo con estudios postuniversitarios
Censo	13.Ind_calidad_vida	Indicador de calidad de vida del radio en el que vive el individuo

3.1.1 Variables del Modelo provenientes de Buró

Las variables de Buró que intervienen en el modelo no poseen valores perdidos, a excepción de Meses_cons_TC2. El tratamiento realizado para cada variable se describe a continuación.

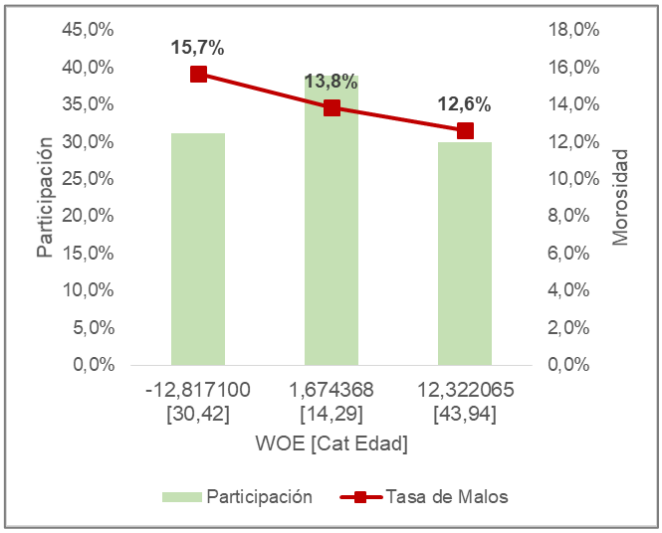
Variable 1: Edad

La edad fue recodificada según las categorías sugeridas por un árbol de decisión (Ver Anexo VIII). Como se observa en el Cuadro 3.1.2, la categoría con mayor mora son aquellos que se encuentran entre 30 y 42 años. Dada la particularidad de la relación de esta variable con el Good Bad, se recodifica esta variable de manera tal que cada categoría tome el valor del WOE correspondiente. El WOE ($WOE_i = Ln\left[\left(\frac{G_i}{\sum G}\right)/\left(\frac{B_i}{\sum B}\right)\right]$) es una transformación habitual que se suele utilizar para transformar variables cualitativas o agrupaciones de variables continuas o discretas. Permite crear una relación monótona con el Good Bad y mantiene una relación lineal con la función logística, lo cual resulta adecuado para este tipo de modelos.

Cuadro 3.1.2: Análisis Bivariado: Good Bad por Edad (en categorías)

CAT	MIN	MAX	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
1	18	29	38,9%	39,0%	38,3%	13,8%	1,674368	
2	30	42	31,1%	30,6%	34,7%	15,7%	-12,817100	0,009817
3	43	94	30,0%	30,5%	26,9%	12,6%	12,322065	

Gráfico 3.1.1: Participación y Tasa de Malos por Edad



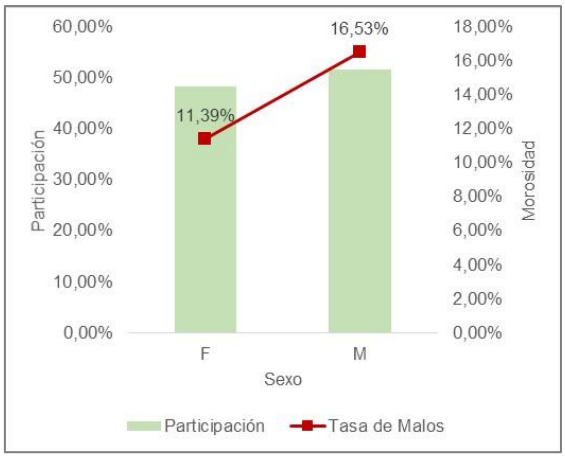
Variable 2: Sexo

El Sexo, fue recodificada de manera binaria, es decir toma el valor 1 si el individuo es mujer o 0 si es hombre. De esta manera, la categoría base es la masculina y se observa que la tasa de malos es mayor en este sexo.

Cuadro 3.1.3: Análisis Bivariado: Good Bad por Sexo

SEXO	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
F	1	48,29%	49,79%	39,16%	11,39%	24,0185392	0,0459439
M	0	51,71%	50,21%	60,84%	16,53%	-19,201983	

Gráfico 3.1.2: Participación y Tasa de Malos por Sexo



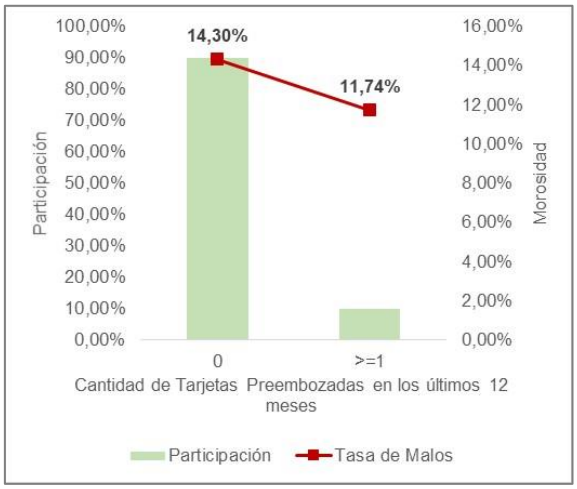
Variable 3: Cant_TC_Preembozadas_12m

Esta variable fue recodificada como una variable dicotómica. Un 1 indica que el individuo tiene informada a Buró por alguna entidad una tarjeta sin uso y un 0 caso contrario. La tasa de malos disminuye en aquellas personas que poseen TC preembozadas, es decir, que alguna entidad le emitió una TC.

Cuadro 3.1.4: Análisis Bivariado: Good Bad por Cantidad de Tarjetas Preembozadas en los últimos 12 meses

Cant_TC_ Preembozadas _12m	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0	0	90,12%	89,85%	91,75%	14,30%	-2,083787	0,0042994
>=1	1	9,88%	10,15%	8,25%	11,74%	20,63984	

Gráfico 3.1.3: Participación y Tasa de Malos por Cantidad de Tarjetas Preembozadas en los últimos 12 meses



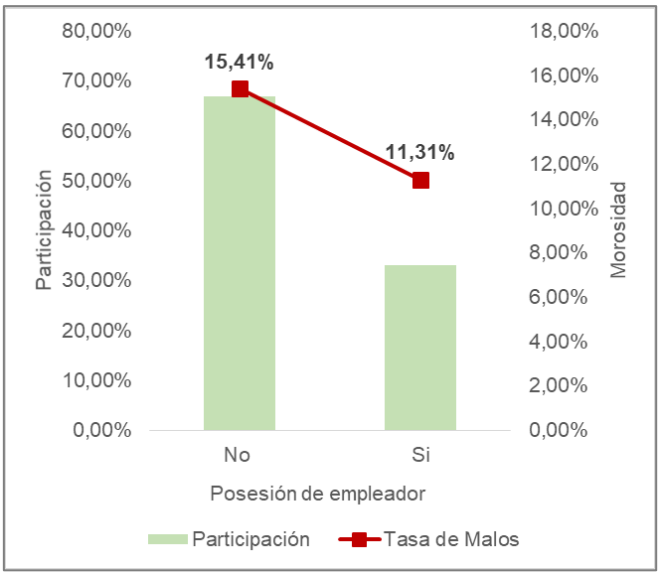
Variable 4: Posee_Empleador

Esta variable intenta resumir de algún modo si el individuo trabaja en relación de dependencia. Aquellas personas que poseen un empleador reciben un 1 en esta variable y tienen una menor tasa de malos.

Cuadro 3.1.5: Análisis Bivariado: Good Bad por Posesión de Empleador

Posee_ Empleador	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
No	0	66,93%	65,87%	73,38%	15,41%	-10,8003	0,0267931
Si	1	33,07%	34,13%	26,62%	11,31%	24,863176	

Gráfico 3.1.4: Participación y Tasa de Malos por Posesión de Empleador



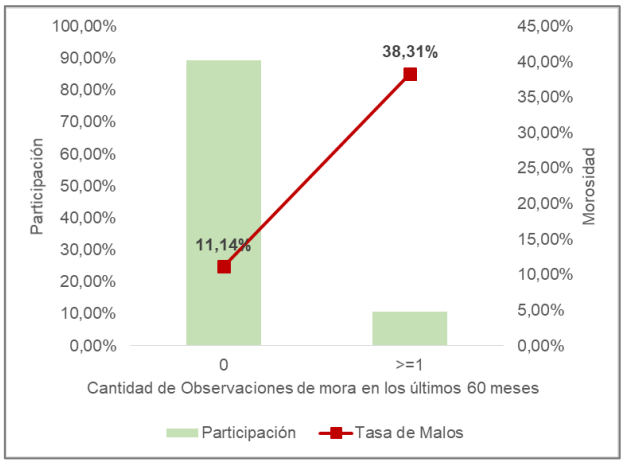
Variable 5: Obs_mora_60m

La variable fue recodificada como una variable dicotómica, donde 1 indica que el individuo posee alguna observación y se asocia con la mayor tasa de malos, mientas que 0 indica lo contrario.

Cuadro 3.1.6: Análisis Bivariado: Good Bad por Cantidad de observaciones en los últimos 60 meses

Obs_mora_60m	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0	0	89,28%	92,30%	70,76%	11,14%	26,57534	0,3447142
>=1	1	10,72%	7,70%	29,24%	38,31%	-133,4539	

Gráfico 3.1.5: Participación y Tasa de Malos por Cantidad de observaciones en los últimos 60 meses



Variable 6: Meses_cons_TC2

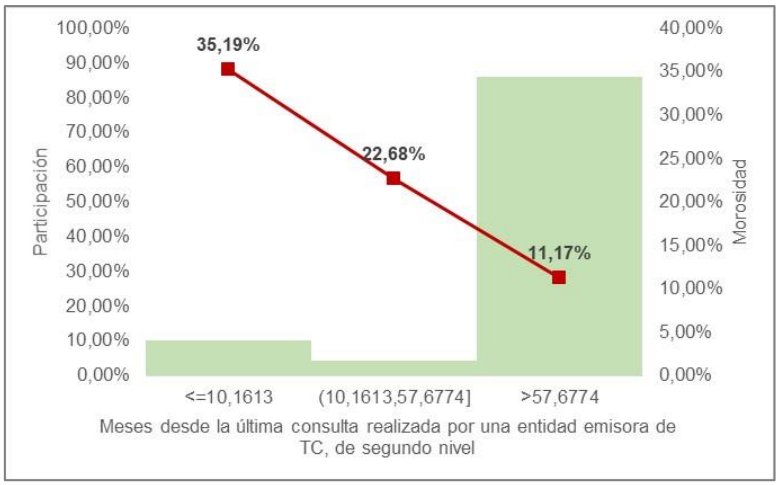
Esta variable se calcula como la diferencia entre la fecha de la consulta realizada por alguna entidad emisora de TC, de segundo nivel y el punto de observación. Se observa la mayor tasa de malos en aquellos que poseen consultas recientes de estas entidades, es decir, que estuvieron buscando recientemente TC en entidades de segundo nivel. En este caso la imputación de valores perdidos fue importante, dado que aquellos que nunca habían sido consultados por este tipo de entidades poseían un valor perdido en esta variable. A partir del análisis bivariado se decidió imputar a este valor la máxima antigüedad observada, dado que se observa una tasa de mora similar con este grupo. Es decir, si la consulta

es muy antigua, esto es similar a no haber sido consultado. En el Cuadro 3.1.7, se observa la variable ya imputada, tal como fue introducida en el modelo.

Cuadro 3.1.7: Análisis Bivariado: Good Bad por Meses desde la última consulta de una entidad emisora de TC, de segundo nivel

MIN	MAX	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0,0000	10,1613	9,99%	7,53%	25,02%	35,19%	-120,054162	
10,1614	57,6774	4,16%	3,74%	6,72%	22,68%	-58,4512636	0,28092827
57,6775	57,8065	85,85%	88,73%	68,27%	11,17%	26,2127916	

Gráfico 3.1.6: Participación y Tasa de Malos por meses desde la última consulta de una entidad emisora de TC, de segundo nivel



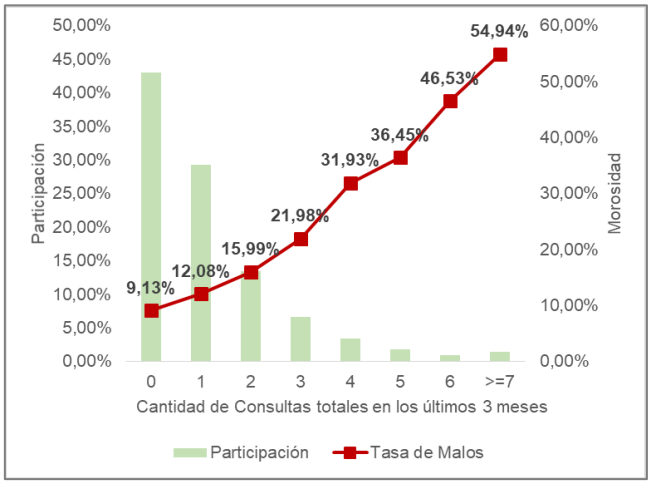
Variable 7: Cant_cons_3m

Esta variable se trunca de tal manera que el valor máximo este controlado, sea razonable y muestre un comportamiento ordenado en el bivariado. Se observa que a medida que la cantidad de consultas totales recientes aumenta, dando señales de personas buscadores de crédito, la tasa de malos aumenta.

Cuadro 3.1.8: Análisis Bivariado: Good Bad por Cantidad de consultas totales en los últimos 3 meses

Cant_cons_3m	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0	0	43,03%	45,50%	27,96%	9,13%	48,691266	0,3562456
1	1	29,25%	29,92%	25,14%	12,08%	17,403161	
2	2	13,49%	13,19%	15,36%	15,99%	-15,20639	
3	3	6,63%	6,01%	10,36%	21,98%	-54,43116	
4	4	3,41%	2,70%	7,74%	31,93%	-105,3916	
5	5	1,82%	1,35%	4,73%	36,45%	-125,5338	
6	6	0,91%	0,56%	3,01%	46,53%	-167,2249	
>=7	7	1,46%	0,76%	5,69%	54,94%	-200,9262	

Gráfico 3.1.7: Participación y Tasa de Malos por Cantidad de consultas totales en los últimos 3 meses



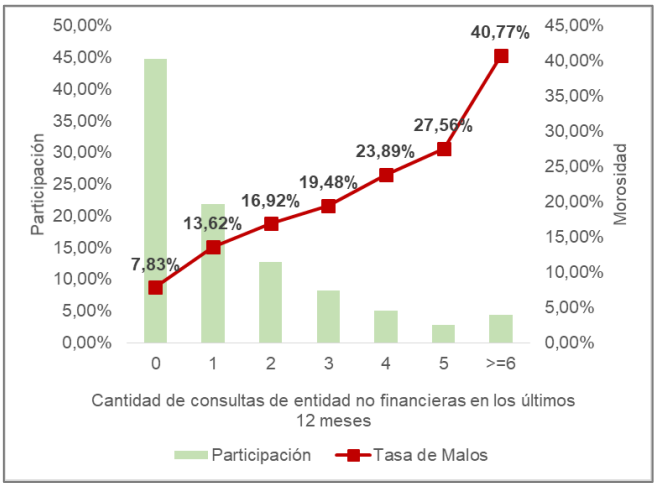
Variable 8: Cant_cons_no_Fi_12m

Esta variable recibe un tratamiento similar a la variable de consultas descripta anteriormente.

Cuadro 3.1.9: Análisis Bivariado: Good Bad por Cantidad de consultas de entidad no financieras en los últimos 12 meses

Cant_cons_no_Fi_12m	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0	0	44,75%	47,99%	24,95%	7,83%	65,396747	0,36644
1	1	21,91%	22,02%	21,24%	13,62%	3,5850505	
2	2	12,80%	12,37%	15,42%	16,92%	-22,00733	
3	3	8,26%	7,74%	11,45%	19,48%	-39,18209	
4	4	5,04%	4,47%	8,57%	23,89%	-65,21411	
5	5	2,80%	2,36%	5,50%	27,56%	-84,48976	
>=6	6	4,43%	3,05%	12,86%	40,77%	-143,7636	

Gráfico 3.1.8: Participación y Tasa de Malos por Cantidad de consultas de entidad no financieras en los últimos 12 meses



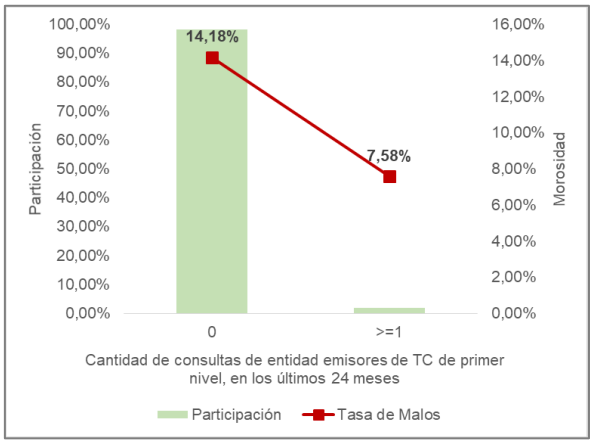
Variable 9: Cant_cons_TC1_24m

Esta variable fue recodificada como una variable dicotómica, donde 1 quiere decir que el individuo fue consultado por alguna entidad emisora de TC de primer nivel en los últimos 2 años y 0, significa lo contrario. Se efectúo esta agrupación por la escasa cantidad de casos en las restantes categorías. Se observa que el segmento de personas que tienen consultas de estas entidades de primer nivel, poseen una menor tasa de malos.

Cuadro 3.1.10: Análisis Bivariado: Good Bad por Cantidad de consultas de entidad emisores de TC de primer nivel, en los últimos 24 meses

Cant_cons_TC1_24m	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0	0	98,10%	97,96%	98,98%	14,18%	-1,03167	0,00711
>=1	1	1,90%	2,04%	1,02%	7,58%	68,9326	

Gráfico 3.1.9 Participación y Tasa de Malos por Cantidad de consultas de entidad emisores de TC de primer nivel, en los últimos 24 meses



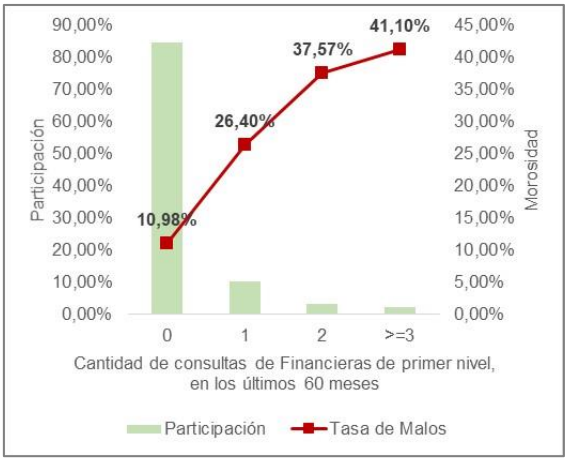
Variable 10: Cant_cons_Fi1_60m

Esta variable recibe el mismo tratamiento que Cant_cons_3m y Cant_cons_no_Fi_12m descriptas anteriormente.

Cuadro 3.1.11: Análisis Bivariado: Good Bad por Cantidad de consultas de Financieras de primer nivel, en los últimos 60 meses

Cant_cons_Fi1_60m	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0	0	84,41%	87,43%	65,96%	10,98%	28,17243	0,292838
1	1	10,28%	8,81%	19,32%	26,40%	-78,57323	
2	2	3,18%	2,31%	8,51%	37,57%	-130,3271	
>=3	3	2,12%	1,45%	6,21%	41,10%	-145,1322	

Gráfico 3.1.10: Participación y Tasa de Malos por Cantidad de consultas de Financieras de primer nivel, en los últimos 60 meses.



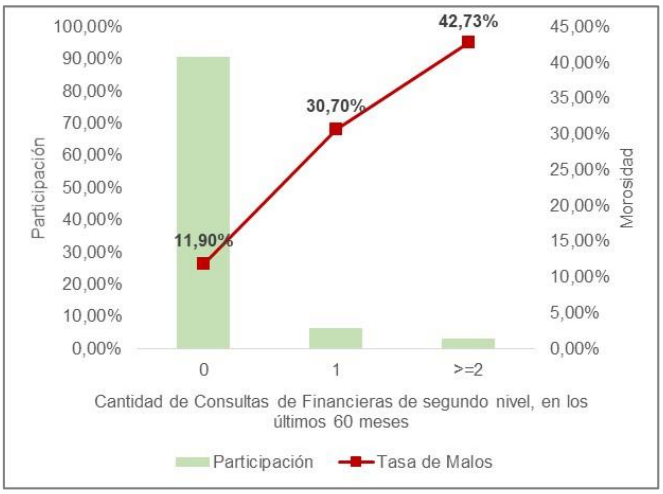
Variable 11: Cant_cons_Fi2_60m

Esta variable recibe el mismo tratamiento que Cant_cons_3m, Cant_cons_no_Fi_12m y Cant_cons_Fi1_60m descriptas anteriormente.

Cuadro 3.1.12: Análisis Bivariado: Good Bad por Cantidad de consultas de entidad Financieras de segundo nivel en los últimos 60 meses

Cant_cons_Fi2_60m	VALOR	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0	0	90,52%	92,79%	76,65%	11,90%	19,116148	
1	1	6,38%	5,15%	13,95%	30,70%	-99,71016	0,2301396
>=2	2	3,09%	2,06%	9,40%	42,73%	-151,83141	

Gráfico 3.1.11: Participación y Tasa de Malos por Cantidad de consultas de entidad Financieras de segundo nivel en los últimos 60 meses



3.1.2 Variables del Modelo provenientes del Censo

Las variables del modelo que corresponden a la nueva fuente de datos, no sufrieron grandes transformaciones con respecto a su versión original, sólo imputación de valores perdidos a las personas que no poseen datos correspondientes a la fuente censal. Esta ausencia de datos se debe a que en la base de datos no estaba guardada de manera precisa la dirección de modo que permita ubicar a la persona en un radio censal. Para el desarrollo y el testeo se imputaron los valores perdidos por el valor de la variable que se encuentre asociado a una tasa de mora similar. Sin embargo, al momento de implementarse, esto no sería necesario, dado que se espera que al utilizarse el

modelo, el individuo consultado brinde su dirección de manera voluntaria como requisito para obtener el producto de crédito deseado.

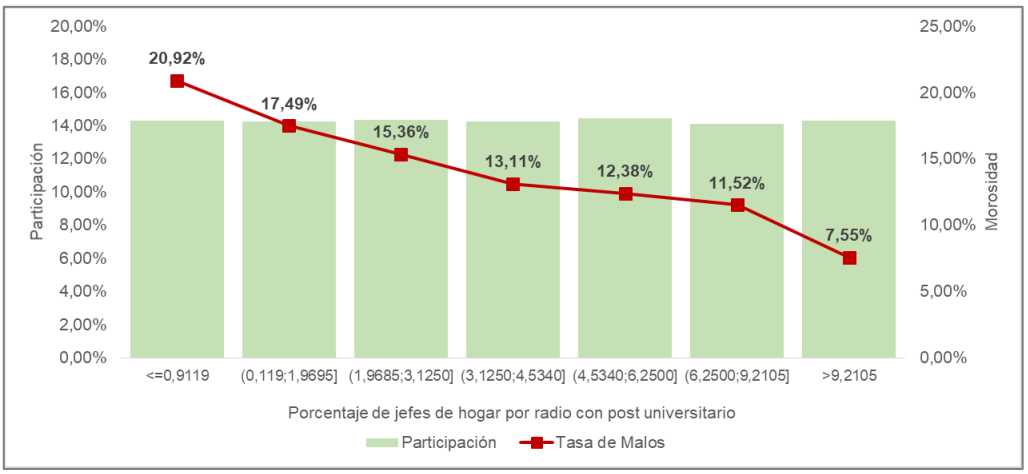
Como se puede observar en los análisis bivariados, éstas ordenan correctamente la mora, y por sí mismas, pueden crear grupos con morosidades muy distintas.

Variable 12: Porc_jefe_post_univ

Cuadro 3.1.13: Análisis Bivariado: Good Bad por Porcentaje de jefes del radio en el que vive en el individuo con estudios postuniversitarios

MIN	MAX	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
0,0000	0,9119	14,31%	13,17%	21,31%	20,92%	-48,11547	
0,9174	1,9685	14,24%	13,67%	17,72%	17,49%	-25,96131	
1,9737	3,1250	14,34%	14,12%	15,67%	15,36%	-10,44836	
3,1332	4,5340	14,26%	14,41%	13,31%	13,11%	7,976510	0,115139
4,5361	6,2500	14,45%	14,73%	12,73%	12,38%	14,55354	
6,2745	9,2105	14,12%	14,54%	11,58%	11,52%	22,74767	
9,2179	15,1117	14,28%	15,36%	7,68%	7,55%	69,37602	

Gráfico 3.1.12: Participación y Tasa de Malos por Porcentaje de jefes del radio en el que vive en el individuo con estudios postuniversitarios

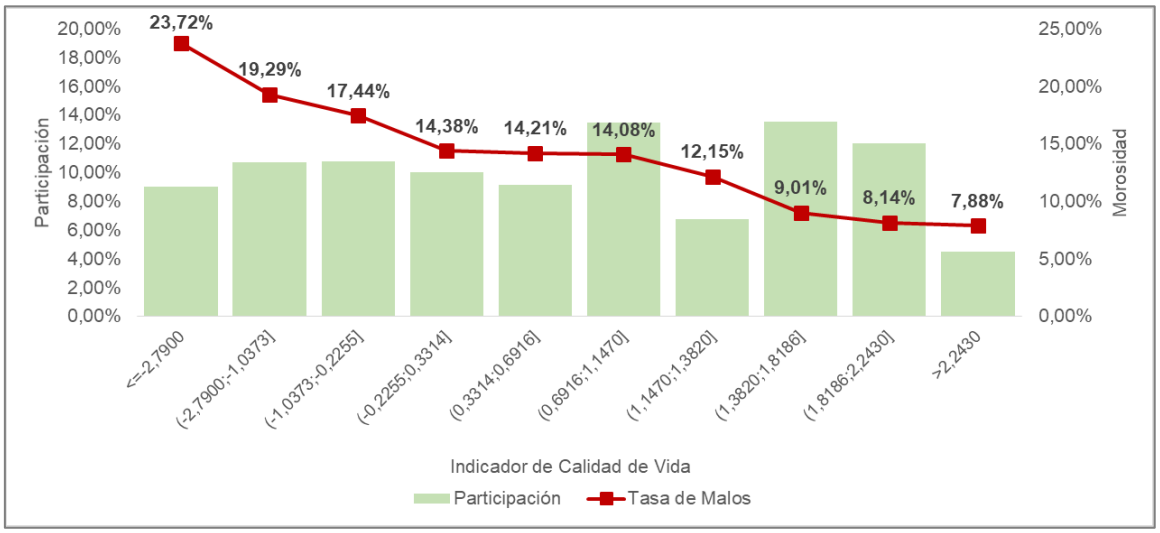


Variable 13: Ind_Calidad_vida

Cuadro 3.1.14: Análisis Bivariado: Good Bad por Indicador de calidad de vida del radio en el que vive el individuo

MIN	MAX	% TOTAL	% BUENOS	% MALOS	TASA DE MALOS	WOE	IV
-19,638	-2,7900	9,04%	8,02%	15,20%	23,72%	-63,8502	0,152915
-2,7670	-1,0373	10,73%	10,09%	14,68%	19,29%	-37,5354	
-1,0356	-0,2255	10,78%	10,36%	13,33%	17,44%	-25,1839	
-0,2227	0,3314	10,04%	10,01%	10,24%	14,38%	-2,22775	
0,3323	0,6916	9,14%	9,12%	9,21%	14,21%	-0,91023	
0,6922	1,1470	13,48%	13,48%	13,46%	14,08%	0,169481	
1,1479	1,3820	6,73%	6,88%	5,80%	12,15%	17,20276	
1,3834	1,8186	13,51%	14,32%	8,63%	9,01%	50,63063	
1,8200	2,2430	12,05%	12,89%	6,95%	8,14%	61,6983	
2,2457	2,5673	4,50%	4,82%	2,51%	7,88%	65,22588	

Gráfico 3.1.13: Participación y Tasa de Malos por Indicador de calidad de vida del radio en el que vive el individuo



3.2 Estimación de Modelo de Credit Scoring

La estimación del modelo presentado en la ecuación 2.3.4, tiene en cuenta todas las variables descriptas anteriormente, donde cada una de ellas interviene de manera significativa[16] en el modelo y con el signo esperado de acuerdo al bivariado. Además los factores de inflación de la varianza (FIV) se encuentran

[16] Se utiliza el test de Wald=$\hat{\beta}^2/\sigma_{\hat{\beta}}^2 \sim \chi_1^2$ y todas las variables son significativas el 99%.

en los valores aceptados (menor a 3), indicando ausencia de multicolinealidad (Ver Anexo IX). En el cuadro 3.2.1 se observan las estimaciones obtenidas.

Cuadro 3.2.1: Estimación de los Coeficientes del modelo

	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	Wald	gl	Sig.	$Exp(\hat{\beta})$	95% IC para $Exp(\hat{\beta})$	
							Inferior	Superior
Edad (WOE)	-0,008	0,003	8,197	1	0,004	0,9916	0,986	0,997
Sexo (F=1)	-0,364	0,061	35,656	1	0,000	0,695	0,617	0,783
Cant_TC_Preembozadas_12m	-0,306	0,107	8,147	1	0,004	0,736	0,597	0,909
Posee_empleador	-0,489	0,068	52,332	1	0,000	0,613	0,537	0,700
Obs_mora_60m	1,264	0,074	293,303	1	0,000	3,541	3,064	4,093
Meses_cons_TC2	-0,013	0,001	75,970	1	0,000	0,987	0,985	0,990
Cant_cons_3m	0,164	0,022	53,736	1	0,000	1,178	1,128	1,231
Cant_cons_no_Fi_12m	0,079	0,021	14,246	1	0,000	1,082	1,039	1,127
Cant_cons_TC1_24m	-0,826	0,283	8,529	1	0,003	0,438	0,252	0,762
Cant_cons_Fi1_60m	0,231	0,044	27,595	1	0,000	1,259	1,156	1,373
Cant_cons_Fi2_60m	0,368	0,059	38,427	1	0,000	1,445	1,286	1,624
porc_jefe_post_univ	-0,044	0,010	18,312	1	0,000	,957	0,938	0,977
Ind_calidad_vida	-0,044	0,009	22,419	1	0,000	0,957	0,940	0,975
Constante	-1,393	0,108	167,534	1	0,000	0,248		

Los coeficientes $\hat{\beta}_j$ permiten entender el sentido de la relación de cada variable con la variable dependiente y los test de hipótesis asociados a cada uno de ellos permiten testear si los mismos son significativamente distintos de 0. Si bien el coeficiente β_j no es directamente interpretable, su signo sí lo es. Un coeficiente con un signo positivo indica que si la variable incrementa, la probabilidad de que ocurra el evento aumenta y viceversa.

Se puede observar que las variables de la nueva fuente de información resultan ser significativas y con el signo esperado. Un incremento en el indicador de calidad de vida del radio o del porcentaje de jefes con educación post universitaria del radio, repercute en una disminución de la probabilidad de incurrir en mora.

Los Exp($\hat{\beta}_j$) suelen brindar una interpretación más clara de los resultados. Los Exp($\hat{\beta}_j$) son los odds ratio y representan cuánto mayor es la oportunidad de ser malo por el incremento unitario de la variable x_j, manteniendo el resto constante. Cuando el Exp($\hat{\beta}_j$) es mayor a 1 indica que un aumento de la variable dependiente incrementa los odds de incurrir en mora. Por el contrario un Exp($\hat{\beta}_j$)

menor a 1, indica que un aumento de la variable x_j repercute en una disminución de los odds de que el evento ocurra. Y cuando Exp($\hat{\beta}_j$) es igual a 1 quiere decir que cambios en la variable no modifican los odds, es decir, que la variable no contribuye a explicar la ocurrencia del evento. Se puede observar en el Cuadro 3.2.1, que todos los Exp($\hat{\beta}_j$) son significativamente distintos de 1, ya que ninguno de los intervalos de confianza contiene el 1, esto se relaciona con la significatividad de los $\hat{\beta}_j$.

De este modo, por ejemplo, si el porcentaje de jefes de hogar con estudios post universitario incrementase en una unidad, manteniendo el resto de las variables constante, los odds de ser malos disminuirían, resultando 0,957 de los odds calculados sin el incremento de esta unidad. En el Anexo X se encuentra la interpretación de Exp($\hat{\beta}_j$) de cada variable.

Los Exp($\hat{\beta}_j$) representan una medida estandarizada de asociación entre variables que puede servir para comparar el impacto que el incremento unitario de cada variable provoca en los odds de ser malos. Este concepto suele ser muy usado en epidemiología para evaluar el impacto que un factor de riesgo tiene sobre la oportunidad de ocurrencia del evento. Sin embargo, en los Modelos de Credit Scoring, las variables explicativas no suelen ser todas dicotómicas y esta medida no tiene en cuenta el rango de variación de cada variable. Por esto, sacar conclusiones sobre el impacto de esta variable en el modelo, observando sólo esta medida, puede ser confuso. Esto se debe a que en una variable dicotómica, el Exp($\hat{\beta}_j$) representará el cambio máximo que la variable x_j puede provocar en los odds, manteniendo todas las demás variables constantes, pero esto no es así en las variables con mayor rango. Para calcular el cambio máximo en los odds que puede provocar cada variable, ceteris paribus, se calculó Exp($\hat{\beta}_j * Rango_{x_j}$), es decir, la relación entre los odds cuando x_j es igual a su mínimo y cuando es igual a su máximo. A efectos comparativos, a aquellas variables para las cuales su valor es menor a 1, se calcula su inversa.

Gráfico 3.2.1: Máximo cambio en los odds que podría provocar cada variable del modelo

Las barras verdes, indican que Exp($\hat{\beta}_j * Rango_{x_j}$) es menor a uno por lo cual se calcula su inversa[17]. Se observa que cambios en las variables provenientes de la nueva fuente de información pueden llegar a duplicar y más la oportunidad de incurrir en mora. Por su parte, la variable que indica la existencia de observaciones de mora y la cantidad de consultas totales recientes parecieran ser la que mayor cambio provocan en la oportunidad de ser malo.

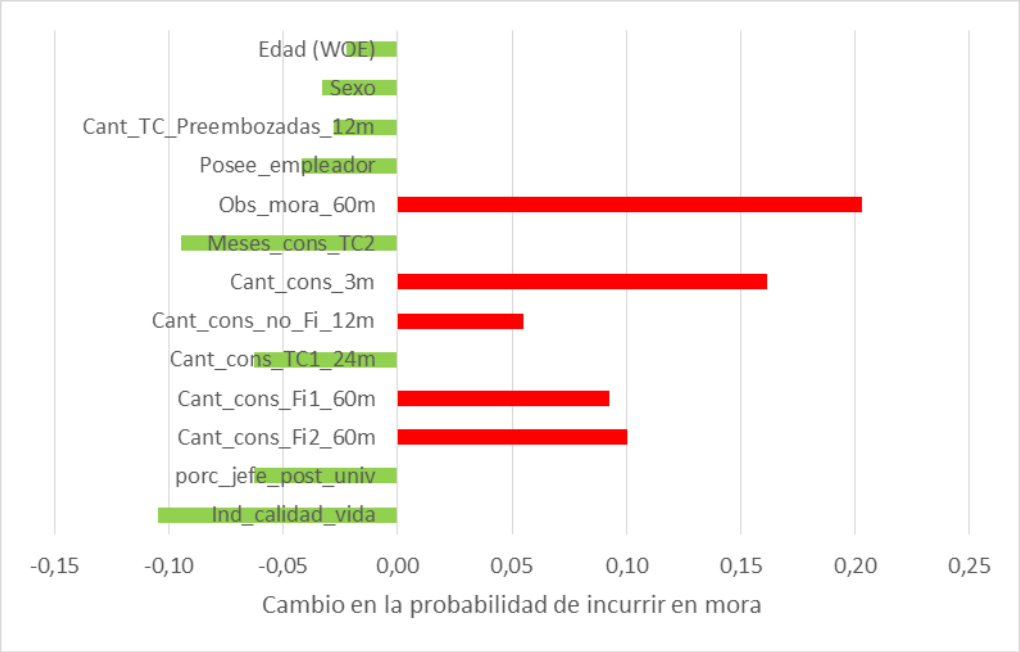
Si se busca evaluar el impacto que el incremento en cada variable puede provocar directamente en la probabilidad, se presenta el inconveniente que, a diferencia del modelo lineal, esto no depende solo del valor del coeficiente sino también del valor que toman las variables explicativas. Por este motivo, es común calcular los "efectos marginales medios", es decir, calcular el incremento en la probabilidad para valores puntuales de las variables explicativas, como podría ser la media o la mediana de las variables. Para este individuo "promedio"[18] también se podría calcular el impacto que tiene en la probabilidad

[17] Esto equivaldría a comparar el odds cuando la variable toma su valor máximo y luego su valor mínimo, dado que el coeficiente fue invertido.

[18] En este trabajo se considera a un individuo "promedio" a una mujer entre 18 y 29 años, sin TC preembozadas, ni datos sobre empleador, sin observaciones de mora en los últimos 60 meses, con una sólo consultas en los últimos 3 meses y una consulta en una entidad no financiera en el último año, sin cantidad de consultas en las demás variables involucradas, con domicilio ubicado en un radio que posee 4.6% de jefes de hogar con educación post universitaria y con indicador de calidad de vida igual a -0.11. Estos valores, dependiendo la variable que se trate, son la moda o el promedio, si el promedio no era un valor dentro de los valores posibles que puede tomar la variable, se considera el valor de la mediana.

de ser malo, el cambio máximo en cada variable, es decir, incrementar de su valor mínimo al máximo, considerando las demás variables constantes, en su valor “promedio” por ejemplo. En el Gráfico 3.2.2, se puede observar lo recién descripto.

Gráfico 3.2.2: Incremento en la probabilidad de incurrir en mora causado por la variación máxima de cada variable en un individuo “promedio”



Este gráfico muestra que las variables de la nueva fuente, en un individuo “promedio”, pueden disminuir la probabilidad de incurrir en mora, entre 0,07 y 0,10 puntos cada variable (siendo que la probabilidad varía entre 0 y 1). Esto indica que el contexto en el que vive el individuo se relaciona con su comportamiento como se esperaba.

En cuanto a las variables de Buró, se observa que Obs_mora_60m es la que provoca el mayor incremento en la probabilidad de incurrir en mora. Esto indica que un individuo “promedio” que posee antecedentes de mora en empresas de telecomunicaciones o retailers, presenta 0,20 puntos más de probabilidad de incurrir en mora en entidad aportantes a Buró o a BCRA que un individuo similar pero sin antecedentes de falta de pago.

Por su parte, las variables referidas a consultas, en general, suelen ser un buen indicio para identificar a aquellas personas que están buscando financiación en el mercado. El perfil de aquellas personas consultadas reiteradas veces y sobre

todo si esas consultas provienen de entidades no bancarias o de segunda línea, suele ser más riesgoso que aquellas consultadas por entidades bancarias o emisoras de tarjetas de crédito de primera línea. Así es que, el aumento de la cantidad de consultas recientes (cant_cons_3m) puede provocar un aumento de la probabilidad de incurrir en mora de hasta 0,16 puntos en un individuo "promedio". Del mismo modo, cada una de las variables que mide la cantidad de consultas provenientes de entidades financieras no bancarias (Cant_cons_Fi1_60m, Cant_cons_Fi2_60m) puede impactar en un aumento de la probabilidad de incurrir en mora de hasta 0,10 puntos en un individuo "promedio". Por su parte, si un individuo "promedio" presenta consultas antiguas de entidades emisoras de tarjetas de crédito de segundo nivel o no presenta (Meses_cons_TC2) posee una probabilidad de incurrir en mora de hasta 0,10 puntos menor que una persona de similares características pero en busca de financiación reciente de estas entidades (consultas de poca antigüedad). Por el contrario, tener consultas de una entidad emisora de TC de primera línea (Cant_cons_TC1_24m) resuelta favorable ya que, en un individuo "promedio", puede provocar una caída de la probabilidad de incurrir en mora de hasta 0,05 puntos.

Las variables que menor cambio provocan en la probabilidad de incurrir en mora en un individuo "promedio", al margen de edad y sexo, corresponde a can_tc_preembozadas_12m y posee_empleador. Poseer empleo informado o tener alguna TC preembozada se asocia con perfiles de menor riesgo. El empleo funciona como un indicador de que el individuo está inserto en el mercado laboral y la tenencica de TC preembozadas funcionan como indicio de que el individuo aplicó a políticas de crédito de alguna entidad del mercado. La disminución en la probabilidad de incurrir en mora que pueden provocar estas variables en un individuo "promedio" es inferior a 0,05 puntos.

Finalmente, luego de realizar todos los testeos correspondiente, se lleva a cabo la transformación del modelo, comentada en la sección 2.3. Esta transformación tiene en cuenta la siguiente fórmula:

Ecuación 3.2.1: $Score = round(1000 * (1 - p))$

Si $Score < 1$, se asigna el valor 1 y si $Score > 999$, se asigna el valor 999.

Capítulo 4: Validación del Modelo

Luego de la estimación del modelo y la interpretación de los coeficientes con la muestra de entrenamiento, es necesario evaluar su performance en una muestra distinta a la del desarrollo. Justamente para esto se utiliza la muestra de validación. A cada una de las 11.124 personas incluidas en dicha muestra se les calcula el score con la fórmula encontrada en la muestra de entrenamiento. Las tablas de performance, el análisis de la captura de malos, el cálculo del KS, la curva ROC, el Indice de Gini son algunas de las medidas de validación utilizadas.

Aunque resulta importante, quedan fuera del alcance de este trabajo, la validación con una muestra fuera de tiempo (OOT) y el análisis de estabilidad de las variables en el tiempo.

4.1 Medidas de Performance

Tablas de Performance

La tabla de performance es un cuadro muy usado en los modelos de Credit Scoring, permite entender la distribución del score, su capacidad para discriminar entre buenos y malos, para ordenar la tasa de mora y es un gran soporte para colaborar en las definiciones de riesgo. Asimismo, también sirve como una herramienta clave para determinar el punto de corte a la hora de la implementación.

Cuadro 4.1.1: Tabla de Performance Muestra de Validación

Bucket	Score		Total		Malos		Buenos		Tasa de Malos		KS
	Min	Max	Int(%)	Acum(%)	Int(%)	Acum(%)	Int(%)	Acum(%)	Int(%)	Acum(%)	
10	955	999	9,9%	9,9%	2,5%	2,5%	11,1%	11,1%	3,6%	3,6%	8,5%
9	943	954	10,4%	20,3%	3,3%	5,9%	11,6%	22,7%	4,6%	4,1%	16,8%
8	933	942	10,2%	30,5%	4,0%	9,9%	11,2%	33,9%	5,6%	4,6%	24,0%
7	923	932	9,2%	39,6%	4,5%	14,3%	9,9%	43,8%	7,0%	5,1%	29,5%
6	910	922	10,3%	49,9%	6,9%	21,2%	10,8%	54,6%	9,5%	6,1%	33,4%
5	892	909	10,2%	60,1%	7,5%	28,7%	10,7%	65,3%	10,4%	6,8%	36,6%
4	867	891	9,8%	69,9%	8,5%	37,1%	10,0%	75,4%	12,3%	7,6%	38,2%
3	813	866	10,1%	80,1%	13,8%	51,0%	9,5%	84,9%	19,4%	9,1%	33,9%
2	695	812	10,0%	90,0%	18,4%	69,4%	8,6%	93,4%	26,3%	11,0%	24,1%
1	1	694	10,0%	100,0%	30,6%	100,0%	6,6%	100,0%	43,7%	14,2%	0,0%

Esta tabla divide a la muestra de validación en 10 segmentos de prácticamente igual tamaño. En este cuadro se puede observar que existe una concentración de la distribución del score en valores altos. Uno de los datos más importante que muestra este cuadro es que el score logra un muy buen ordenamiento de la tasa de malos. Partiendo de una base que posee un 14,2% de mora, el score

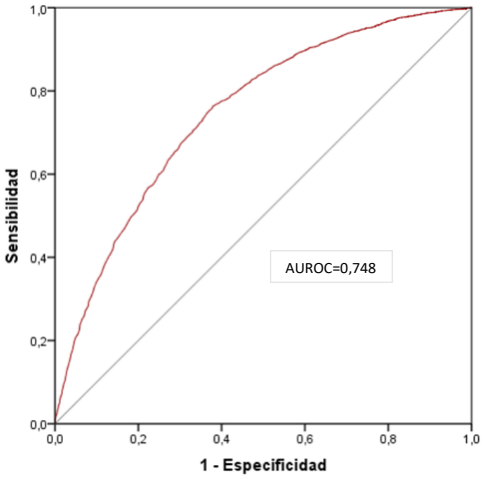
logra ordenar de tal manera que el mejor 10% de la población posee una morosidad menor al 4%, mientras que el peor 10% posee una mora superior al 40%, es decir, 10 veces mayor. Además permite establecer distintos escenarios fijando la tasa de aprobación o la proporción de aprobados tolerable (Ver Sección 4.2). También resulta interesante analizar la captura de malos y buenos en los peores y mejores rangos. Por ejemplo, se puede observar que en el mejor 30,5% se filtrarían el 9,9% de malos aproximadamente, mientras que más del 30% de los buenos estarían siendo capturados. Es posible ver así, la sensibilidad y especificidad del modelo que se muestra mejor en la curva ROC.

Curva ROC

La curva ROC (Características Operativas del Receptor) es ampliamente usada en distintos disciplinas como la medicina, la ingeniera, la psicología, además del Credit Scoring. Relaciona dos conceptos muy importantes: la sensibilidad, que es la habilidad de identificar los verdaderos positivos y la especificidad, que es la habilidad de identificar a los verdaderos negativos. En otras palabras, en el contexto de Credit Scoring es la capacidad que tiene el modelo de otorgar score altos a los buenos, para que puedan ser aceptados por las entidades y de marcar con score bajos a los malos para que sean rechazados por las mismas. Estas medidas varían de acuerdo al punto de corte que se elija. Por ejemplo, si se vuelve a la tabla de performance y se toma como punto de corte el score 867, se observa que el 75,4% de los buenos estarían siendo aceptados (Sensibilidad), y el 71,3% de los malos estarían siendo rechazados correctamente (Especificidad). La Curva ROC grafica la Sensibilidad, porcentaje de verdaderos positivos (buenos que estaría aceptando correctamente) por cada punto de corte, contra 1-Especificidad, porcentaje de falsos positivos (malos que estaría aceptando incorrectamente). En el gráfico 4.1.1 se muestra esta gráfica comparada con una línea recta que representa el azar. El estadístico que se utiliza relacionado con este gráfico es el AUROC[19], que mide el área bajo la curva y se espera que sea significativamente distinta de 0,5, que es el área de la recta relacionada con el azar.(Ver Anexo XI)

[19] El coeficiente de GINI está muy relacionado con el AUROC, GINI=2*AUROC-1.

Gráfico 4.1.1: Curva ROC

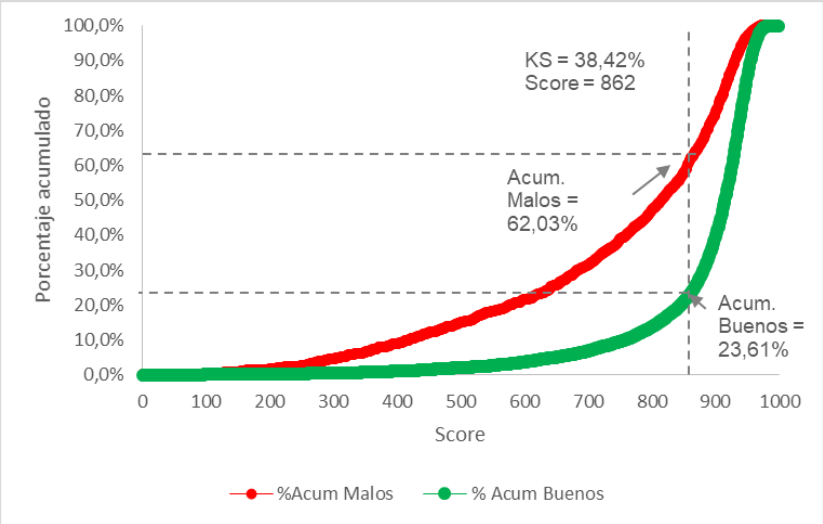


Estadístico Kolmogorov Smirnov

La tabla de performance muestra otro dato importante que es la separación entre la distribución de buenos y de malos, para cada intervalo, siendo el valor máximo el estadístico KS (Kolmogorov-Smirnov), 38,2 en este caso que se obtiene en el score 867. Si bien esta medida es muy reconocida en Credit Scoring, no se utiliza de manera aislada ya que muchas veces este corte de score no se adecúa a la realidad del negocio.

Es posible obtener el KS puntual, a partir de la distribución de malos y buenos calculada teniendo en cuenta todos los valores posibles del score, tal como se observa en el Gráfico 4.1.2.

Gráfico 4.1.2: Curva KS



Al igual que como se veía en la tabla de performance, se muestra la falta de casos en score muy bajos, sin embargo se observa la separación que el modelo logra entre ambas poblaciones. El KS del modelo es 38.42% y se obtiene en el score 862 donde las curvas logran su máxima separación, lo que resulta muy bueno sobre todo considerando que se trata de una población no bancarizada (segmento THIN).

Subsegmento sin datos

En la población analizada existen personas que no fueron consultadas nunca, que no tienen datos de empleador, ni de tarjetas preembozadas, ni observaciones de mora. Para este segmento el único dato que se tiene es la edad, el sexo y el domicilio. En este subsegmento, las variables de la fuente alternativa toman mayor importancia. Por este motivo, se busca evaluar si el score es capaz de ordenar la mora en esta subpoblación (1751 casos dentro de la muestra de validación cumplen las condiciones descriptas). Para esto se realiza una tabla de performance teniendo en cuenta esta subpoblación.

Gráfico 4.1.3: Tabla de performance – segmento sin datos –

Bucket	Score		Total		Malos		Buenos		Tasa de Malos		KS
	Min	Max	Int(%)	Acum(%)	Int(%)	Acum(%)	Int(%)	Acum(%)	Int(%)	Acum(%)	
5	945	999	20,2%	20,2%	8,0%	8,0%	21,2%	21,2%	3,1%	3,1%	13,2%
4	935	944	19,4%	39,5%	13,9%	21,9%	19,8%	41,0%	5,6%	4,3%	19,1%
3	924	934	20,6%	60,1%	20,4%	42,3%	20,6%	61,6%	7,8%	5,5%	19,3%
2	909	923	19,5%	79,6%	24,8%	67,2%	19,1%	80,7%	9,9%	6,6%	13,5%
1	1	908	20,4%	100,0%	32,8%	100,0%	19,3%	100,0%	12,6%	7,8%	0,0%

Desde luego la capacidad de discriminación del score se ve afectada porque es una población para la cual se tiene muy poca información. La mora de este segmento es de 7,8% y se observa una concentración de esta población en valores altos de score. Sin embargo, esta tabla de performance muestra que el score ordena correctamente la mora. El 20% de menor riesgo muestra una mora muy baja, 3%, mientras que el 20% más riesgoso cuadruplica esa tasa de mora, tiene un 12,6% de mora.

4.2 Análisis de Punto de Corte

En general las entidades definen un punto de corte o cut-off a partir del cual deciden rechazar al solicitante del crédito por considerarlo muy riesgoso o por

tener un riesgo mayor al que la entidad tolera. La tabla de performance es un soporte indispensable para definir este corte.

La definición del punto de corte está en concordancia con las políticas de la entidad. Entidades más conservadores se identificarán con cortes de score más altos, mientras que entidades dispuestas a asumir mayor riesgo se inclinan por cortes de score más bajos. La decisión del punto de corte es un trade off entre porcentaje de aceptados y tasa de Malos. Mientras mayor es la tasa de aceptados que se quiere, mayor será la mora que se deberá tolerar. Por el contrario mientras menor sea la tasa de malos que la entidad soporta, menor será el porcentaje de aceptados.

Es importante tener en cuenta que este es un score genérico, con lo cual, muchas entidades incorporan reglas duras en sus políticas de originación para customizar más la decisión y adecuarla a sus políticas. Otras entidades utilizan este score como insumo en scores propios. Y otro grupo de entidades, tal vez más nuevas o con menor experiencia o conforme con la performance del score genérico, únicamente utilizan este tipo de score. A veces, las entidades que utilizan sólo el score genérico como regla de decisión analizan el comportamiento del score, no en todo el mercado, sino en segmentos más similares a su mercado potencial y fijan el punto de corte teniendo en cuenta esa población de interés

Para identificar un punto de corte es necesario fijar el porcentaje de consultas que se desea aprobar o la tasa de malos que se está dispuesto a tolerar. Con estas definiciones la tabla de performance resulta una herramienta muy útil. Así es que, utilizando el cuadro 4.1.1, si se desea aprobar al 80% del total de consultados, se podría poner un punto de corte de 831 y se esperaría una tasa de malos de mercado del 9%. Si, por ejemplo, la entidad está dispuesta a tolerar un 6% de mora, el punto de corte debería ser 910 y se aprobarían el 50% de las consultas.

Capítulo 5: Especificaciones de implementación

Las especificaciones de implementación sirven para precisar los detalles del modelo, las recodificaciones de las variables y el detalle de la formula final para poder productivizar el modelo y disponibilizarlo para el cliente que quiera utilizarlo en una plataforma online.

Variables de entrada
DOCUMENTO (CUIT)
SEXO
DOMICILIO (O CPA)

Definiciones	
Edad	If age<=29 then Edad=1,674368038. If age>29 and age<=42 then Edad=-12,81710003. If age>42 then Edad=12,32206506.
Sexo	If Gender='F' then Sexo=1. Else 0.
Cant_TC_Preembozadas_12m	If Cant_TC_Preembozadas_12m>=1 then Cant_TC_Preembozadas_12m=1. Else 0.
Posee_empleador	If cuit_empleador is not null then posee_empleador=1. Else 0.
Obs_mora_60m	If cant_obs_mora_60m>=1 then obs_mora_60m=1. Else 0.
Meses_cons_TC2	Meses(Fecha_obs-Fecha_ult_cons_TC2). If Fecha_ult_cons_TC2 is null then Meses_cons_TC2=57,8065.
Cant_cons_3m	If Cant_cons_3m>=7 then Cant_cons_3m=7. Else Cant_cons_3m.
Cant_cons_no_Fi_12m	If Cant_cons_no_Fi_12m>=6 then Cant_cons_no_Fi_12m=6. Else Cant_cons_no_Fi_12m.
Cant_cons_TC1_24m	If Cant_cons_TC1_24m>=1 then Cant_cons_TC1_24m=1. Else 0.
Cant_cons_Fi1_60m	If Cant_cons_Fi1_60m>=3 then Cant_cons_Fi1_60m=3. Else Cant_cons_Fi1_60m.
Cant_cons_Fi2_60m	If Cant_cons_Fi2_60m>=2 then Cant_cons_Fi2_60m=2. Else Cant_cons_Fi2_60m.
porc_jefe_post_univ	Ver radio en donde se ubica la dirección
Ind_calidad_vida	Ver radio en donde se ubica la dirección

Variables de Salida	
Score	Z= Edad * (-0,00848500627874347) + Sexo * (-0,363606379647466) + Cant_TC_Preembozadas_12m * (-0,305851909058579) + Posee_empleador * (-0,488827141091382) + Obs_mora_60m * (1,26449414592627) + Meses_cons_TC2 * (-0,0127496403992877) + Cant_cons_3m * (0,163972323822083) + Cant_cons_no_Fi_12m * (0,0785993283096511) + Cant_cons_TC1_24m * (-0,825932070321406) + Cant_cons_Fi1_60m * (0,230603906478253) + Cant_cons_Fi2_60m * (0,368231175740056) + porc_jefe_post_univ * (-0,0436403487321436) + Ind_calidad_vida * (-0,0436158581873468) + (-1.39333038125742). Score=rnd(1000*(exp(-z)/(1+exp(-z)))) If score >999 then score=999. If score<1 then score=1.
Decisión	If score>=*cut-off* then decisión=Aceptar. If score<*cut-off* then decisión=Rechazar.

Capítulo 6: Conclusiones

En base a los resultados, se pudo observar que el Censo Nacional de Población, Hogares y Viviendas permite caracterizar a cada uno de los radios censales e identificar unidades geográficas diferentes entre sí en Capital Federal. Además, se observó que estas variables relacionadas con el contexto en el que vive el individuo muestran una relación con el comportamiento financiera. De este modo, los estudios presentados permiten concluir que esta nueva fuente de información contribuye en el modelo de score que busca explicar la probabilidad de default de los individuos no bancarizados. Extender este análisis a otras geografías constituye un nuevo desafío que vale la pena probar.

Es importante tener en cuenta que esta fuente de información no debe ser considerada como un reemplazo a las fuentes tradicionales sino como un complemento. Las variables analizadas del Censo tienen una buena capacidad predictiva y son capaces de ordenar la mora, sin embargo no constituyen la fuente principal, aunque desde luego aumentan su importancia en individuos que carecen de otro tipo de información.

Si bien la nueva fuente de información se actualiza cada 10 años, una persona podría asociarse a otras características en el transcurso de ese periodo si cambia su domicilio, ya que quedaría asociado a otro radio censal. Las características destacables que posee esta fuente de información son la accesibilidad gratuita y la cobertura nacional que junto con su capacidad predictiva la convierten en una fuente por demás interesante para ser tenida en cuenta en los modelos de riesgos que se desarrollen para la población no bancarizada.

Considerando el contexto mundial donde la inclusión financiera constituye uno de los principales ejes para un crecimiento inclusivo, la búsqueda para la detección de nuevas fuentes para el análisis del segmento no bancarizado debe ser un trabajo constante como así también el estudio por potenciar la capacidad de las fuentes ya estudiadas.

Anexo

Anexo I. Variables disponibles a nivel de radio censal

Variables de Vivienda en la base de radios censales	
Tipo de Vivienda Agrupado	Cantidad de viviendas Particulares
	Cantidad de viviendas Colectivas
Tipo de Vivienda Particular	Cantidad de viviendas tipo Casa
	Cantidad de viviendas particulares tipo Rancho
	Cantidad de viviendas particulares tipo Casilla
	Cantidad de viviendas particulares tipo Departamento
	Cantidad de viviendas particulares tipo Pieza en inquilinato
	Cantidad de viviendas particulares tipo Pieza en hotel familiar o pensión
	Cantidad de viviendas particulares tipo Local no construido para habitación
	Cantidad de viviendas particulares tipo Vivienda móvil
	Cantidad de viviendas particulares tipo Persona/s viviendo en la calle
Condición de ocupación	Cantidad de viviendas Con personas presentes
	Cantidad de viviendas Con todas las personas temporalmente ausentes
	Cantidad de viviendas En alquiler o venta
	Cantidad de viviendas En construcción
	Cantidad de viviendas Se usa como comercio, oficina o consultorio
	Cantidad de viviendas Se usa para vacaciones, fin de semana u otro uso temporal
	Cantidad de viviendas Por otra razón
Tipo de Vivienda Colectiva	Cantidad de viviendas colectiva tipo Hogar de ancianos
	Cantidad de viviendas colectiva tipo Hogar de menores
	Cantidad de viviendas colectiva tipo Colegio internado
	Cantidad de viviendas colectiva tipo Hospital
	Cantidad de viviendas colectiva tipo Prisión
	Cantidad de viviendas colectiva tipo Cuartel
	Cantidad de viviendas colectiva tipo Hogar de religiosos
	Cantidad de viviendas colectiva tipo Hotel turístico
	Cantidad de viviendas colectiva tipo Otros
	Cantidad de viviendas colectiva tipo Campamento/obrador
Calidad de los materiales	Cantidad de viviendas con materiales con Calidad 1
	Cantidad de viviendas con materiales con Calidad 2
	Cantidad de viviendas con materiales con Calidad 3
	Cantidad de viviendas con materiales con Calidad 4
Calidad de Conexiones a Servicios	Cantidad de viviendas con Calidad de Conexiones a Servicios Satisfactoria
	Cantidad de viviendas con Calidad de Conexiones a Servicios Básica
	Cantidad de viviendas con Calidad de Conexiones a Servicios Insuficiente
Calidad Constructiva de la Vivienda	Cantidad de viviendas con Calidad Constructiva Satisfactoria
	Cantidad de viviendas con Calidad Constructiva Básica
	Cantidad de viviendas con Calidad Constructiva Insuficiente
Cantidad de Hogares en la Vivienda	Cantidad de Viviendas con un hogar
	Cantidad de Viviendas con dos y más hogares

Variables de hogares en la base de radios censales	
Material predominante de los pisos	Cantidad de hogares cuyo Material predominante de los pisos es Cerámica, baldosa, mosaico, mármol
	Cantidad de hogares cuyo Material predominante de los pisos es Cemento o ladrillo fijo madera o alfombrado
	Cantidad de hogares cuyo Material predominante de los pisos es Tierra o ladrillo suelto
	Cantidad de hogares cuyo Material predominante de los pisos es Otro
Material predominante de la cubierta exterior del techo	Cantidad de hogares cuyo material predominante de la cubierta exterior del techo es Cubierta asfáltica o membrana
	Cantidad de hogares cuyo material predominante de la cubierta exterior del techo es Baldosa o losa (sin cubierta)
	Cantidad de hogares cuyo material predominante de la cubierta exterior del techo es Pizarra o teja
	Cantidad de hogares cuyo material predominante de la cubierta exterior del techo es Chapa de metal (sin cubierta)
	Cantidad de hogares cuyo material predominante de la cubierta exterior del techo es Chapa fibrocemento o plástico

	Cantidad de hogares cuyo material predominante de la cubierta exterior del techo es Chapa de cartón
	Cantidad de hogares cuyo material predominante de la cubierta exterior del techo es Caña, palma, tabla o paja con o sin barro
	Cantidad de hogares cuyo material predominante de la cubierta exterior del techo es Otro
revestimiento interior o cielorraso del techo	Cantidad de hogares con revestimiento interior o cielorraso del techo
	Cantidad de hogares sin revestimiento interior o cielorraso del techo
tenencia de agua	Cantidad de hogares con tenencia de agua Por cañería dentro de la vivienda
	Cantidad de hogares con tenencia de agua Fuera de la vivienda pero dentro del terreno
	Cantidad de hogares con tenencia de agua Fuera del terreno
Procedencia del agua para beber y cocinar por	Cantidad de hogares sin revestimiento interior o cielorraso del techo
	Cantidad de hogares sin revestimiento interior o cielorraso del techo
	Cantidad de hogares sin revestimiento interior o cielorraso del techo
	Cantidad de hogares sin revestimiento interior o cielorraso del techo
	Cantidad de hogares sin revestimiento interior o cielorraso del techo
	Cantidad de hogares sin revestimiento interior o cielorraso del techo
Tiene baño/letrina	Cantidad de hogares con Tiene baño/letrina
	Cantidad de hogares sin Tiene baño/letrina
tiene boton, cadena mochila para limpieza del inodoro	Cantidad de hogares con tiene botón, cadena mochila para limpieza del inodoro
	Cantidad de hogares sin tiene botón, cadena mochila para limpieza del inodoro
Desague del inodoro	Cantidad de hogares con Desagüe del inodoro por A red pública (cloaca)
	Cantidad de hogares con Desagüe del inodoro por A cámara séptica y pozo ciego
	Cantidad de hogares con Desagüe del inodoro por Sólo a pozo ciego
	Cantidad de hogares con Desagüe del inodoro por A hoyo, excavación en la tierra, etc.
Baño / letrina de uso exclusivo	Cantidad de hogares con Baño / letrina usado sólo por el hogar
	Cantidad de hogares con Baño / letrina compartido con otros hogares
Combustible usado principalmente para cocinar	Cantidad de hogar cuyo Combustible usado principalmente para cocinar es Gas de red
	Cantidad de hogar cuyo Combustible usado principalmente para cocinar es Gas a granel (zeppelin)
	Cantidad de hogar cuyo Combustible usado principalmente para cocinar es Gas en tubo
	Cantidad de hogar cuyo Combustible usado principalmente para cocinar es Gas en garrafa
	Cantidad de hogar cuyo Combustible usado principalmente para cocinar es Electricidad
	Cantidad de hogar cuyo Combustible usado principalmente para cocinar es Leña o carbón
	Cantidad de hogar cuyo Combustible usado principalmente para cocinar es Otro
Total de habitaciones o pieza s para dormir	Cantidad de hogares con 1 habitación o pieza para dormir
	Cantidad de hogares con 2 habitaciones o piezas para dormir
	.
	.
	Cantidad de hogares con 28 habitaciones o piezas para dormir
Total de habitaciones o piezas	Cantidad de hogares con 1 habitación o pieza
	Cantidad de hogares con 2 habitación o pieza
	.
	.
	Cantidad de hogares con 27 habitación o pieza
Heladera	Cantidad de hogares con heladera
	Cantidad de hogares sin heladera
Computadora	Cantidad de hogares con computadora
	Cantidad de hogares sin computadora
teléfono celular	Cantidad de hogares con teléfono celular
	Cantidad de hogares sin teléfono celular
teléfono línea	Cantidad de hogares con teléfono línea
	Cantidad de hogares sin teléfono línea
Tenencia de la vivienda y propiedad del terreno	Cantidad de hogares Propietario de la vivienda y del terreno
	Cantidad de hogares propietarios sólo de la vivienda
	Cantidad de hogares Inquilino
	Cantidad de hogares Ocupante por préstamo
	Cantidad de hogares Ocupante por trabajo
	Cantidad de hogares en Otra situación de tenencia

Total de personas en el hogar	Cantidad de hogares con 1 persona
	Cantidad de hogares con 2 personas
	Cantidad de hogares con 3 personas
	Cantidad de hogares con 4 personas
	Cantidad de hogares con 5 personas
	Cantidad de hogares con 6 personas
	Cantidad de hogares con 7 personas
	Cantidad de hogares con 8 y más personas
Hacinamiento	Cantidad de hogares con Hasta 0.50 personas por cuarto
	Cantidad de hogares con 0.51 - 0.99 personas por cuarto
	Cantidad de hogares con 1.00 - 1.49 personas por cuarto
	Cantidad de hogares con 1.50 - 1.99 personas por cuarto
	Cantidad de hogares con 2.00 - 3.00 personas por cuarto
	Cantidad de hogares con Más de 3.00 personas por cuarto
Al menos un indicador de NBI	Cantidad de hogares sin NBI
	Cantidad de hogares con NBI

Variables de Vivienda en la base de radios censales	
Tipo de Vivienda Agrupado	Cantidad de viviendas Particulares
	Cantidad de viviendas Colectivas
	Cantidad de viviendas Total
Tipo de Vivienda Particular	Cantidad de viviendas tipo Casa
	Cantidad de viviendas particulares tipo Rancho
	Cantidad de viviendas particulares tipo Casilla
	Cantidad de viviendas particulares tipo Departamento
	Cantidad de viviendas particulares tipo Pieza en inquilinato
	Cantidad de viviendas particulares tipo Pieza en hotel familiar o pensión
	Cantidad de viviendas particulares tipo Local no construido para habitación
	Cantidad de viviendas particulares tipo Vivienda móvil
	Cantidad de viviendas particulares tipo Persona/s viviendo en la calle
Condición de ocupación	Cantidad de viviendas Con personas presentes
	Cantidad de viviendas Con todas las personas temporalmente ausentes
	Cantidad de viviendas En alquiler o venta
	Cantidad de viviendas En construcción
	Cantidad de viviendas Se usa como comercio, oficina o consultorio
	Cantidad de viviendas Se usa para vacaciones, fin de semana u otro uso temporal
	Cantidad de viviendas Por otra razón
Tipo de Vivienda Colectiva	Cantidad de viviendas colectiva tipo Hogar de ancianos
	Cantidad de viviendas colectiva tipo Hogar de menores
	Cantidad de viviendas colectiva tipo Colegio internado
	Cantidad de viviendas colectiva tipo Hospital
	Cantidad de viviendas colectiva tipo Prisión
	Cantidad de viviendas colectiva tipo Cuartel
	Cantidad de viviendas colectiva tipo Hogar de religiosos
	Cantidad de viviendas colectiva tipo Hotel turístico
	Cantidad de viviendas colectiva tipo Otros
	Cantidad de viviendas colectiva tipo Campamento/obrador
Calidad de los materiales	Cantidad de viviendas con materiales con Calidad 1
	Cantidad de viviendas con materiales con Calidad 2
	Cantidad de viviendas con materiales con Calidad 3
	Cantidad de viviendas con materiales con Calidad 4
Calidad de Conexiones a Servicios	Cantidad de viviendas con Calidad de Conexiones a Servicios Satisfactoria
	Cantidad de viviendas con Calidad de Conexiones a Servicios Básica
	Cantidad de viviendas con Calidad de Conexiones a Servicios Insuficiente
Calidad Constructiva de la Vivienda	Cantidad de viviendas con Calidad Constructiva Satisfactoria
	Cantidad de viviendas con Calidad Constructiva Básica
	Cantidad de viviendas con Calidad Constructiva Insuficiente
Cantidad de Hogares en la Vivienda	Viviendas con un hogar
	Viviendas con dos y más hogares

Variables de Población en la base de radios censales	
Relación o parentesco con el jefe(a) del hogar	Cantidad de Jefe(a)
	Cantidad de Cónyuge o pareja
	Cantidad de Hijo(a) / Hijastro(a)
	Cantidad de Yerno / Nuera
	Cantidad de Nieto(a)
	Cantidad de Padre / Madre / Suegro(a)
	Cantidad de Otros familiares
	Cantidad de Otros no familiares
	Cantidad de Servicio doméstico y sus familiares
Sexo	Cantidad de varones
	Cantidad de mujeres
Edad	Cantidad de personas con 0 años
	Cantidad de personas con 1 años
	.
	Cantidad de personas con 110 años
En qué país nació	Cantidad de personas nacidas en Argentina
	Cantidad de personas nacidas en otro país
Sabe leer y escribir	Cantidad de personas que saben leer y escribir
	Cantidad de personas que no saben leer y escribir
Utiliza computadora	Cantidad de personas que utilizan computadora
	Cantidad de personas que no utilizan computadora
Edad en grandes grupos	Cantidad de personas entre 0 y 14 años
	Cantidad de personas entre 15 y 64 años
	Cantidad de personas entre 65 y más años
Edades quinquenales	Cantidad de personas entre 0-4 años
	Cantidad de personas entre 5-9 años
	Cantidad de personas entre 10-14 años
	Cantidad de personas entre 15-19 años
	Cantidad de personas entre 20-24 años
	Cantidad de personas entre 25-29 años
	Cantidad de personas entre 30-34 años
	Cantidad de personas entre 35-39 años
	Cantidad de personas entre 40-44 años
	Cantidad de personas entre 45-49 años
	Cantidad de personas entre 50-54 años
	Cantidad de personas entre 55-59 años
	Cantidad de personas entre 60-64 años
	Cantidad de personas entre 65-69 años
	Cantidad de personas entre 70-74 años
	Cantidad de personas entre 75-79 años
	Cantidad de personas entre 80-84 años
	Cantidad de personas entre 85-89 años
	Cantidad de personas entre 90-94 años
	Cantidad de personas entre 95 y más años
Condición de asistencia escolar	Cantidad de personas que asisten a un establecimiento educativo
	Cantidad de personas que asistieron a un establecimiento educativo
	Cantidad de personas que nunca asistieron a un establecimiento educativo
Nivel educativo que cursa o cursó	Cantidad de personas que cursa o curso nivel Inicial (jardín, preescolar)
	Cantidad de personas que cursa o curso nivel Primario
	Cantidad de personas que cursa o curso nivel EGB
	Cantidad de personas que cursa o curso nivel Secundario
	Cantidad de personas que cursa o curso nivel Polimodal
	Cantidad de personas que cursa o curso nivel Superior no universitario
	Cantidad de personas que cursa o curso nivel Universitario
	Cantidad de personas que cursa o curso nivel Post universitario
	Cantidad de personas que cursa o curso nivel Educación especial
Completó el nivel	Cantidad de personas que completaron el nivel cursado
	Cantidad de personas que no completaron el nivel cursado
Condición de actividad	Cantidad de personas ocupadas
	Cantidad de personas desocupadas
	Cantidad de personas económicamente inactivas
Jefes, nivel	Cantidad de jefes con máximo nivel alcanzado inicial (Jardín, preescolar)
	Cantidad de jefes con máximo nivel alcanzado Primario

	Cantidad de jefes con máximo nivel alcanzado EGB
	Cantidad de jefes con máximo nivel alcanzado Secundario
	Cantidad de jefes con máximo nivel alcanzado Polimodal
	Cantidad de jefes con máximo nivel alcanzado Superior no universitario
	Cantidad de jefes con máximo nivel alcanzado Universitario
	Cantidad de jefes con máximo nivel alcanzado Post universitario
	Cantidad de jefes con máximo nivel alcanzado Educación especial

Anexo II. Descripción de variables incluidas en el análisis de componentes principales.

La mayoría de las definiciones fueron tomadas del diccionario de la base de datos del Redatam que ofrece el INDEC[20], con excepción de la definición del indicador de Necesidades Básicas Insatisfechas (NBI) que fue tomado de la presentación de la distribución del Porcentaje de hogares y de población con NBI, según provincia. Total del país. Años 2001 y 2010[21]

Porcentaje de hogares que poseen computadora: indica la proporción de "hogares que cuentan con un aparato electrónico que se utiliza para el almacenaje, procesamiento de información" con respecto al total de hogares del radio.

Porcentaje de hogares que poseen heladera: es el cociente entre la cantidad de hogares que tienen "disponibilidad de un artefacto doméstico consistente en un receptáculo con paredes aislantes provisto de un motor que genera bajas temperaturas que permiten mantener frescos los alimentos y bebidas en su interior" y el total de hogares del radio.

Porcentaje de hogares con al menos una Necesidad Básica Insatisfecha: es la proporción de hogares que presentan al menos una necesidad básica insatisfecha con respecto al total de hogares del radio. Se considera que un hogar posee NBI si "presenta al menos una de las siguientes condiciones de privación:

[20] Base de datos REDATAM. Instituto Nacional de Estadística y Censos. República Argentina. Definiciones de la base de datos. Censo Nacional de Población, Hogares y Viviendas 2010 Censo del Bicentenario Serie Base de datos Censo 2010 Abril 2013. Disponible en https://www.indec.gov.ar/ftp/cuadros/poblacion/glosario_censo2010.pdf

[21] Definición de NBI tomada de la publicación del cuadro *Porcentaje de hogares y de población con Necesidades Básicas Insatisfechas (NBI), según provincia. Total del país. Años 2001 y 2010.* Disponible en https://www.indec.gob.ar/nivel4_default.asp?id_tema_1=4&id_tema_2=27&id_tema_3=66

- NBI 1. Vivienda: es el tipo de vivienda que habitan los hogares que moran en habitaciones de inquilinato, hotel o pensión, viviendas no destinadas a fines habitacionales, viviendas precarias.
- NBI 2. Condiciones sanitarias: incluye a los hogares que no poseen retrete.
- NBI 3. Hacinamiento: es la relación entre la cantidad total de miembros del hogar y la cantidad de habitaciones de uso exclusivo del hogar. Se considera hacinamiento crítico cuando en el hogar hay más de tres personas por cuarto.
- NBI 4. Asistencia escolar: hogares que tienen al menos un niño en edad escolar (6 a 12 años) que no asiste a la escuela.
- NBI 5. Capacidad de subsistencia: incluye a los hogares que tienen cuatro o más personas por miembro ocupado y que tienen un jefe que no ha completado el tercer grado de escolaridad primaria."

Porcentaje de hogares con baño exclusivo: es el cociente entre la cantidad de hogares en los cuales sus "miembros no comparten en forma habitual el baño con miembros de otro hogar" y el total de hogares del radio.

Porcentaje de hogares con gas de red como combustible para cocinar: es el cociente entre la cantidad de hogares cuyo principal combustible usado para cocinar es el gas de red y la cantidad de hogares totales.

Porcentaje de hogares con más de 3 habitantes por cuarto: es la relación entre la cantidad de hogares con más de 3 individuos por cuarto (sin contar baño/s y cocina/s) y la cantidad total de hogares.

Porcentaje de jefes con nivel educativo alto: surge del cociente entre la cantidad de jefes de hogar con nivel educativo alto y el total de jefes del radio. Se considera jefe de hogar a aquel que todos los mismos del hogar lo reconocen como tal. Se considera que el jefe tiene nivel educativo alto "si cursó o estaba cursando al momento del censo en la Argentina o en el exterior estudios de nivel Superior no universitario, Universitario o Post-universitario"

Porcentaje de personas que saben leer y escribir: es el cociente en la cantidad total de personas que "posee la capacidad de leer, escribir y comprender una frase sencilla sobre la vida cotidiana en cualquier idioma y el total de personas del radio".

Porcentaje de personas que usan la computadora: relación entre la cantidad de personas que “poseen la capacidad de manejar cualquier programa o software en una computadora. (Ej. Acceso a Internet, etc.)” y el total de personas del radio

Porcentaje de personas desocupadas: cociente entre la cantidad de personas desocupadas y el total de personas del radio. Según el Censo, se considera que “una persona está desocupada si durante las cuatro semanas anteriores al día del censo realizó acciones tendientes a establecer una relación laboral o iniciar una actividad empresaria”, no habiendo desarrollado ninguna actividad que genera bienes o servicios para el “mercado” la semana anterior a la fecha de referencia del censo.

Porcentaje de viviendas con agua de red: cociente entre la cantidad de viviendas que utilizan como principal fuente de abastecimiento para beber y cocinar el agua proveniente de red pública (agua corriente) y el total de hogares del radio.

Porcentaje de viviendas con Calidad de Material I: cociente entre la cantidad total de viviendas “presenta materiales resistentes y sólidos tanto en el piso como en techo; presenta cielorraso” y el total de viviendas del radio.

Porcentaje de viviendas con desagote a red pública: es el cociente entre la cantidad de viviendas que posee desagüe de inodoro a red pública sobre el total de viviendas. Se entiende “desagüe a red pública al sistema de cañerías interno que enlaza con una red de tuberías comunal de eliminación y tratamiento de las aguas servidas y materia sólida (líquidos cloacales).”

Porcentaje de viviendas con construcción satisfactoria: se refiere a la relación entre la cantidad de “viviendas que disponen de agua a red pública y desagüe cloacal” y el total de viviendas del radio.

Anexo III: Matriz de correlación de las variables de interés

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14
porc_compu_sí	1,000	,747	,752	,719	-,780	-,695	,841	-,493	,502	,951	,282	,778	,811	,476
porc_hela_sí	,747	1,000	,846	,712	-,739	-,773	,490	-,307	,551	,670	,490	,709	,761	,555
porc_baño_exclusivo	,752	,846	1,000	,664	-,770	-,853	,557	-,374	,468	,664	,343	,666	,738	,389
porc_combus_gas_red	,719	,712	,664	1,000	-,808	-,535	,518	-,376	,502	,692	,300	,888	,873	,563
porc_hacinam_más_3.00	-,780	-,739	-,770	-,808	1,000	,751	-,590	,412	-,439	-,724	-,212	-,796	-,829	-,450
porc_con_nbi	-,695	-,773	-,853	-,535	,751	1,000	-,510	,337	-,274	-,584	-,206	-,555	-,672	-,329
porc_edujefes_alto	,841	,490	,557	,518	-,590	-,510	1,000	-,496	,487	,877	,173	,632	,648	,354
porc_desocupado	-,493	-,307	-,374	-,376	,412	,337	-,496	1,000	-,212	-,483	-,047	-,446	-,457	-,174
porc_lee_escribe_sí	,502	,551	,468	,502	-,439	-,274	,487	-,212	1,000	,554	,676	,528	,515	,571
porc_usa_compu_sí	,951	,670	,664	,692	-,724	-,584	,877	-,483	,554	1,000	,300	,768	,784	,485
porc_agua_red	,282	,490	,343	,300	-,212	-,206	,173	-,047	,676	,300	1,000	,325	,324	,509
porc_calidad_mat_1	,778	,709	,666	,888	-,796	-,555	,632	-,446	,528	,768	,325	1,000	,977	,565
porc_construc_satis	,811	,761	,738	,873	-,829	-,672	,648	-,457	,515	,784	,324	,977	1,000	,549
porc_desag_red	,476	,555	,389	,563	-,450	-,329	,354	-,174	,571	,485	,509	,565	,549	1,000

Anexo IV. Test de Bartlett e índice KMO

El test de esfericidad de Bartlett (1950) compara la matriz de correlaciones mostrada en el cuadro 2.1.2 con la matriz identidad y plantea como hipótesis nula que la matriz de correlación entre las variables analizadas no es significativamente distinta a la matriz identidad. Bajo la hipótesis nula el determinante de la matriz de correlaciones R es igual a 1, $|R| = 1$, en cambio si las variables están fuertemente relacionadas $|R| \approx 0$. El test de Bartlett se basa en el determinante de esta matriz del siguiente modo:

$$-\left[n - 1 - \frac{2p + 5}{6}\right] \text{x } ln|R| \sim \chi^2_{(p^2-p)/2}$$

Donde p es el número de variables y n es la cantidad de observaciones.

En esta oportunidad se rechaza la hipótesis nula y la matriz de correlación resulta significativamente diferente a la identidad. Los resulados se presentan a continuación:

Bartlett test of sphericity

Chi-square = 65990.635
Degrees of freedom = 91
p-value = 0.000
H0: variables are not intercorrelated

Sin embargo, algunas referencias plantean que la particularidad de este test es que suele rechazar la hipótesis nula cuando n aumenta y recomiendan su uso cando $n/p < 5$, que no sucede en esta caso.

El índice de KMO, comparte el objetivo con el test de Bartlett ya que buscan evaluar si es conveniente factorizar las variables originales de una manera eficiente. La matriz de correlación es el punto de partida para tener una idea de la relación entre las variables, sin embargo, esta relación puede estar influenciada por una tercera, por esto se utiliza la matriz de correlación parcial para poder evaluar la relación entre las variables removiendo el efecto del resto. Cuando el KMO es cercano a 1 se dice que el análisis de componentes principales es adecuado, mientas que si el KMO es cercano a 0, el análisis de componentes principales no será relevante. La fórmula que calcula este índice se muestra a continuación:

$$\frac{\sum_i \sum_{i \neq j} \boldsymbol{r_{ij}^2}}{\sum_i \sum_{i \neq j} \boldsymbol{r_{ij}^2} + \sum_i \sum_{i \neq j} \boldsymbol{a_{ij}^2}}$$

Donde r_{ij} es la correlación simple y a_{ij} es la correlación parcial. Los resultados obtenidos se presentan a continuación como así también la escala que permite interpretar el índice.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy
KMO = 0.889

Interpretación de Índice de KMO:

1 >= KMO >= 0.9 muy bueno
0.9 >= KMO >= 0.8 meritorio
0.8 >= KMO >= 0.7 mediano
0.7 >= KMO >= 0.6 mediocre
0.6 >= KMO > 0.5 bajo
KMO <= 0.5 inaceptable

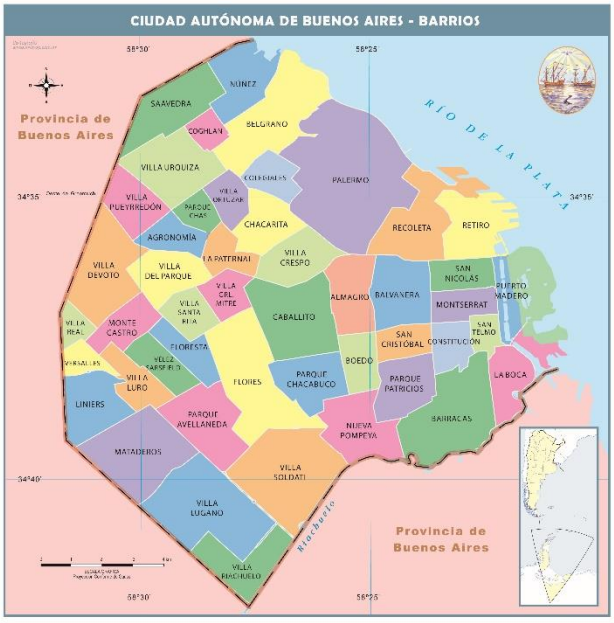
Anexo V. Autovalores de las componentes principales.

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	8,66458	7,1317	61,9%	61,9%
Comp2	1,53288	0,52235	11,0%	72,8%
Comp3	1,01053	0,244238	7,2%	80,1%
Comp4	0,766297	0,158197	5,5%	85,5%
Comp5	0,6081	0,174778	4,3%	89,9%
Comp6	0,433323	0,154723	3,1%	93,0%
Comp7	0,2786	0,0916743	2,0%	95,0%
Comp8	0,186925	0,00833931	1,3%	96,3%
Comp9	0,178586	0,0558102	1,3%	97,6%
Comp10	0,122776	0,0317908	0,9%	98,5%
Comp11	0,0909851	0,00960722	0,7%	99,1%
Comp12	0,0813779	0,0469971	0,6%	99,7%
Comp13	0,0343808	0,0237362	0,3%	99,9%
Comp14	0,0106446	.	0,1%	100,0%

Anexo VI. Autovectores de los 3 primeros componentes principales.

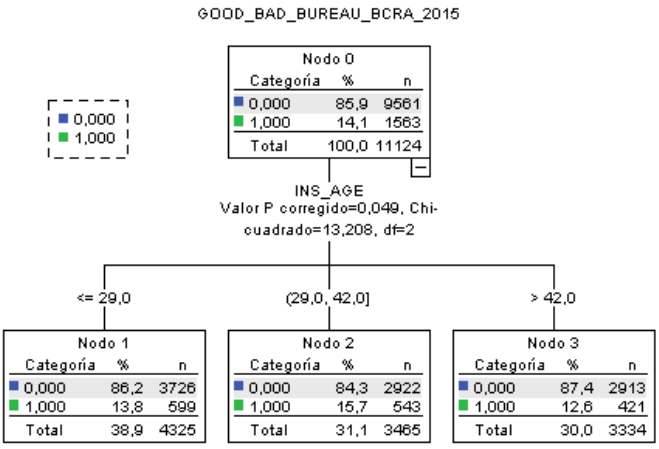
Variables	Comp1	Comp2	Comp3
porc_compu_sí	0,3119	-0,1445	0,1309
porc_hela_sí	0,2937	0,1083	-0,3277
porc_baño_exclusivo	0,2877	-0,0712	-0,3528
porc_combus_gas_red	0,2908	0,0088	-0,0546
porc_hacinam_más_3.00	-0,2974	0,1461	0,1915
porc_con_nbi	-0,2577	0,1888	0,4575
porc_edujefes_alto	0,2581	-0,1876	0,4199
porc_desocupado	-0,1715	0,3065	-0,4184
porc_lee_sí	0,2184	0,4587	0,2476
porc_usa_compu_sí	0,3015	-0,1017	0,2808
porc_agua_red	0,148	0,6264	0,0097
porc_calidad_mat_1	0,3058	-0,0183	0,0612
porc_construc_satis	0,3157	-0,0532	-0,025
porc_desag_red	0,2113	0,3997	0,0778

Anexo VII. Mapa de los Barrios Porteños.



Fuente: http://escuelasdeldistrito.blogspot.com.ar/2016/06/proyecto-mi-buenos-aires-querido.html

Anexo VIII. Árbol de decisión (muestra de entrenamiento) para explicar el Good Bad por la Edad



Anexo IX. Factor de inflación de la varianza

Variables	Estadísticas de colinealidad	
	Tolerancia	VIF
Edad (WOE)	0,984	1,016
Sexo	0,983	1,017
Cant_TC_Preembozadas_12m	0,989	1,011
Posee_empleador	0,984	1,016
Obs_mora_60m	0,950	1,052
Meses_cons_TC2	0,835	1,198
Cant_cons_3m	0,590	1,694
Cant_cons_no_Fi_12m	0,587	1,703
Cant_cons_TC1_24m	0,990	1,010
Cant_cons_Fi1_60m	0,772	1,296
Cant_cons_Fi2_60m	0,855	1,170
porc_jefe_post_univ	0,724	1,382
Ind_calidad_vida	0,722	1,386

Anexo X. Interpretación de los Exp($\hat{\beta}_j$) de cada variable

Exp($\hat{\beta}_1$) = El aumento en una unidad del WOE de edad, provoca una disminución de los odds de entrar en default, resultando 0.9916 de los odds calculados sin el incremento de esta unidad, manteniendo el resto de las variables constante. Dada la definición de WOE, la relación de estos con la variable dependiente definida de esta forma, siempre es negativa.

Exp($\hat{\beta}_2$) = Los odds de ser malo de los hombres es mayor que el de las mujeres. Ya que el odds de los hombres es 1,4385 (1/0.695) veces mayor que el odds de las mujeres, manteniendo el resto de las variables constante.

Exp($\hat{\beta}_3$) = Los odds de aquellos que no poseen TC preembozadas en los últimos 12 meses es mayor que de aquellos que poseen. Se observa que los odds de los que no poseen este tipo de TC es 1,3578 (1/0.736) veces mayor que los odds de aquellos que poseen.

Exp($\hat{\beta}_4$) = Los odds de los que no poseen empleador es 1,6304 (1/0,613) veces mayor que los odds de los que poseen.

Exp($\hat{\beta}_5$) = Los odds de aquellos que poseen observaciones de morosidad en los últimos 60 meses es 3.541 veces mayor que los odds de aquellos que no poseen observaciones de mora.

Exp($\hat{\beta}_6$) = Un incremento en una unidad de los meses desde la última consulta en TCs de segundo nivel, disminuye los odds. Los nuevos odds resultan 0.987 de los odds sin este incremento, manteniendo el resto de las variables constante.

Exp($\hat{\beta}_7$) = Un incremento en una unidad de la cantidad de consultas en los últimos 3 meses, incrementa los odds 1.178 veces.

Exp($\hat{\beta}_8$) = Un incremento en una unidad de la cantidad de consultas de entidades no financieras en los últimos 12 meses, incrementa los odds. Los nuevos odds son 1.082 veces mayor que previo al incremento, manteniendo el resto de las variables constante.

Exp($\hat{\beta}_9$) = Un incremento en una unidad de la cantidad de consultas de Entidades emisores de TC de primera línea en los últimos 24 meses, disminuye los odds a menos de la mitad. Los nuevos odds resultan 0.438 de los odds previo al incremento.

Exp($\hat{\beta}_{10}$) = Un incremento en una unidad de la cantidad de consultas en financieras de primera línea, aumenta los odds 1.259 veces.

Exp($\hat{\beta}_{11}$) = Un incremento en una unidad de la cantidad de consultas de Financieras de segunda línea en los últimos 60 meses, incrementa los odds. Los nuevos odds son 1.445 veces mayor que previo al incremento, manteniendo el resto de las variables constante.

Exp($\hat{\beta}_{13}$) = Un incremento en una unidad en el índice de calidad de vida, disminuye los odds. Se observa que los nuevos odds son 0. 957 de los odds previo al incremento, conservando el resto de las variables constante.

Anexo XI: Estadístico AUROC

Se observa que el intervalo de confianza, excluye al 0,5, con lo cual se puede decir que el área es significativamente distinta al área generada por el azar.

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
0,748	0,007	0,000	0,734	0,761

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

Bibliografía

Alamilla López N. E., Arauco Camargo S. (2009). *Ensayos "Limitaciones del modelo lineal de probabilidad y alternativas de modelación microeconométrica".*

Altman E. (2018). *Financial Ratios, Discriminant Analysis and the prediction of Corporate Bankruptcy.* Journal of Economics, Finance and Administrative Science. Vol. XXIII Nro 4.

Anderson. R. (2007). *The Credit Scoring Toolkit: theory and practice for retail Credit Risk Management and Decision Automation.* Oxford University Press.

Arena Diaz M. et al (2012) *Credit Scoring: Evaluación del Riesgo Crediticio de la cartera de microcreditos de una instutción Financiera en Uruguay.* Universidad de la República, Uruguay.

Avila G. et al. (2015). *Argentinean Banks & Basel III: 1Q15.* FixScr.

Balzarotti V., Castelpoggi F. (2009). *Modelos de puntuación crediticia: la falta de información y el uso de datos de una central de riesgos. BCRA.* Ensayos Económicos 56.

BancoMundial.org (2016). *Inclusión financiera.* Disponible: www.bancomundial.org/es/topic/financialinclusion/overview

Banda Ortiz H., Garza Morales R. (2014). *Aplicación teórica del método HoltWinters al problema de credit scoring de las instituciones de microfinanzas.* Mercados y Negocios. Volumen 15, N°2.

Cabo T. I. (2013) *Métodos de Bondad de Ajuste en Regresión Logística.* Master Oficial en Estadística Aplicada. Universidad de Granada.

Carballo, I. E. (2016) *Las claves para lograr una mejor inclusión financiera.* El EconomistaDiario. Disponible en: http://www.eleconomista.com.ar/2016-09-las-claves-para-lograr-una-mejor-inclusion-financiera/

Cardona Hernández P. (2004). *Aplicación de árboles de decisión en modelos de riesgo crediticio.* Revista Colombiana de Estadística Volumen 27 No 2. Págs. 139 a 151.

Carroll P., Rehmani S. (2017). *Alternative data and the unbanked.* Olyver Wyman. Financial Server

Cheney, Julia S. (2008) *Alternative Data and Its Use in Credit Scoring Thin and No-File Consumers.* Federal Reserve of Philadelphia.

Cicloderiesgo.com. (2015). *Revolución FINTECH La era de las Alianzas.* Colombia, N°20. Disponible en: http://cicloderiesgo.com/revista-ciclo-de-riesgo-edicion-20.pdf

Contón S. R. et al (2010). *Un Modelo de Credit scoring para instituciones de microfinanzas en el marco de Basilea II.* Journal of Economics, Finance and Administrative Science. Vol. 15 N° 28.

Cypher, S. (2016) *New Credit Score will help borrowers with Thin files.* Auto Credit Express. Disponible en: https://www.autocreditexpress.com/blog/new-credit-score-will-help-thin-files/

Dabós M. (2012) *Credit Scoring*. Univerdad de Belgrano.

Dratch, Dana. (2017) *6 ways to deal with no credit history*. Bankrate. Disponible en: https://www.bankrate.com/finance/credit-cards/6-ways-to-deal-with-limited-or-no-credit-1.aspx

Equifax-EFL. *EFL Score Psicométrico*. Disponible en: https://www.equifax.com/assets/honduras/eflCAMMX_folleto.pdf

Espin García, O. y Rodríguez Caballero. C. (2013). *Metodología para un scoring de clientes sin referencias crediticias.* Cuadernos de Economía, 32(59), 139-165.

FICO. (2016). *FICO, LexisNexis® Risk Solutions and Equifax® Partner to Expand Access to Credit with Debut of FICO® Score XD*. Washington.

FICO. *Expanding Credit Opportunities.* Disponible en: http://www.fico.com/independent/assets/FICO_Score_XD_Infographic_single_page_4156IN.pdf

G. Mera y M. Marcos. (2012) *Los censos de población como fuente de datos para trabajar a nivel microespacial (1980-2010)*. (pp. 137–161). Argentina.

Gestión. (2014). *Morosidad bancaria sube a 2,34% en marzo*. Perú. Disponible en: https://gestion.pe/economia/morosidad-bancaria-sube-234-marzo-2095542

Girault M. (2007). *Modelos de Credit Scoring – Qué, Cómo, Cuándo y Para Qué-*. Gerencia de Investigación y Planificación Normativa, Subgerencia General de Normas. Banco Central de la República Argentina (BCRA).

Hand D.J., Henley W.E. (1995/6). *Statistical Classification Methods in Consumer Credit Scoring: A review Journal of the Royal Statistical Society.* Soc. A(1997) 160, Part 3, pp. 523-541.

Hian Chye Koh et al. (2006) *A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques.* International Journal of Business and Information Volume 1 Number 1, pp 96-118.

Hosmer D. W. et al. (1997). *A comparison of goodness-of-fit tests for the logistic regression model.* Statistics in Medicine. Vol. 16, 965—980.

Iñiguez Salas C, Morales Arias M. G. (2009) *Selección de perfiles mediante regresión logística para muestras desproporcionadas, validación, monitoreo y aplicación en la proyección de provisiones.* Escuela politecnica Nacional. Ecuador.

Jones J. *Acquire more consumers without increasing your risk profile.* TransUnion. Disponible en: https://www.transunion.com/product/creditvision-link

Klecka W. R. (1980). *Discriminant Analysis. Sage University Paper.* Series/Number 07-019.

Klinger B. (2015). *Scoring de Crédito Alternativo en Mercados Emergentes.* EFL Global.

LosHornosLp. *La Influencia del Grupo en el Comportamiento del Individuo.* Disponible en: http://www.loshornoslp.com.ar/capacitacion/mi_libro/tema09.htm

MasterCard Social Newsroom. (2013) *MasterCard and EFL Partner to Drive Small Business Growth in Developing Economies.*

Medina Moral E. (2003) *Modelo de Elección Discreta.*

Mesropyan, E. (2017) *Alternative Credit Scoring in the US: Innovators Applying Data Science to Unlock Financial Potential of 'Thin-File' Individuals.* Medici. Disponible en: https://gomedici.com/alternative-credit-scoring-us-data-science-financial-potential-thin-file/

Pascual M. B. et al. (2015) *Un nuevo clasificador de préstamos bancarios a través de la minería de datos.*

Posada, A. (2014) *El Score Crediticio "Social": separando los datos del ruido.* TECHcetera.

PRBC - Payment Reporting Builds Credit. *What is alternative credit?* Disponible en: https: //www.prbc.com/home/AlternativeCredit

Rodríguez G. (2007) *Logit Models for Binary Data.* Lecture Notes on Generalized Linear Models. Princeton University

Rubio, J. (2014) *La Gestión del Riesgo de Crédito en las Instituciones de Microfinanzas.* Departamento de Economía Financiera y Contabilidad. Facultad de Ciencias Económicas y Empresariales. Universidad de Granada.

Sanchez Bilbao P. (2015) *Credit Scoring.* Universidad Cantabria. Santander.

Schreiner M. (2002). *Ventajas y Desventajas del Scoring Estadístico para las Microfinanzas.* Washington University en Saint Louis.

Siddiqi N. (2006) *Credit risk scorecards: developing and implementing intelligent credit scoring.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Spaulding, W. *Alternative Credit Bureaus and Alternative Credit Scores.* Thismatter.com.

Srinivasa V., Kim Y. (1987). *Credit Granting: A Comparative Analysis of Classification Procedures.* The Journal of Finance, Vol 42, No 3, Papers and Proceedings of the 45th Annual Meeting of the American Finance Association, New Orleans, Lousiana, December 28-30 1986, 665-681.

Steglich et al. (2017). *Efectos de Data Positiva en la Inclusión Financiera: El caso del Score+ de EFX para Uruguay*. Equifax D&A Analytics, Uruguay.

Thomas L. C. (2000) *A Survey of Credit and Behavioural Scoring: Forecasting financial risk of lending to consumers.* International Journal of Forecasting 16 (2000) 149–172.

TransUnion (2007) *The Importance of Credit Scoring for Economic Growth.*