Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

MAESTRÍA EN GESTIÓN ECONÓMICA Y FINANCIERA DE RIESGOS

TRABAJO FINAL DE MAESTRÍA

Datos alternativos en modelos de score para créditos a individuos en Argentina durante 2018

AUTOR: NATALI SOLANGE COHEN

DIRECTOR: MAURO EDGARDO SPERANZA

[ENERO 2021]

Dedicatoria

A la lucha por la igualdad de género.

Índice

Resumen Ejecutivo

El presente trabajo es un abordaje a la utilización de datos alternativos en la gestión de riesgo de crédito bancario. Este aspecto es sumamente relevante dado que una mejor gestión del riesgo de crédito trae aparejadas dos cuestiones principales: por un lado, menores pérdidas crediticias y, por otro lado, un ahorro en provisiones.

El objetivo principal de esta tesis es evaluar el impacto que posee la utilización de datos alternativos en la predicción acerca de la capacidad de repago ante la toma de deuda minorista en el sector bancario argentino. La hipótesis es que los datos alternativos mejoran la predicción acerca del riesgo crediticio de los solicitantes.

La principal conclusión de este trabajo es que la incorporación de datos no convencionales mejora la capacidad de predecir el comportamiento de pago ante tomas de deuda minoristas en tarjetas de crédito y paquetes bancarios.

Palabras claves: Riesgo de crédito, Modelos de *scoring*, Datos alternativos, Regresión logística, Indicadores de *performance*

"There´s a way to do it better – find it."
Thomas A. Edison

Introducción

Los modelos de score que actualmente utilizan los bancos comerciales para evaluar el otorgamiento de crédito se basan en la utilización de variables predictoras tradicionales, que no incorporan la nueva información que se puede desprender a partir del aprovechamiento de los grandes volúmenes de información que reciben los bancos. Por otro lado, las Fintech están ganando mercado otorgando créditos en base a modelos que incorporan información no tradicional. Esta desventaja en la que se encuentran los bancos es la que motivó el presente trabajo, el cual constituye un abordaje al análisis de incorporar variables alternativas a los modelos de score de individuos que utilizan los bancos con el objetivo de mejorar el poder predictivo de los mismos.

Esto es sumamente relevante dado que una mejor estimación del comportamiento futuro de las solicitudes de crédito que recibe un banco impacta, por un lado, en sus ganancias y, por otro lado, en las líneas de balance dado que disminuyen las pérdidas esperadas e inesperadas. Todo lo cual potencia el negocio bancario de dar préstamos a partir de los depósitos captados.

Desde sus inicios, la función principal de los bancos comerciales es el otorgamiento de créditos para cubrir las necesidades de los individuos y de las organizaciones. Para poder cumplir con su función, los bancos toman depósitos del público para luego otorgar créditos. Las regulaciones sobre los bancos nacen ante la necesidad de proteger dichos depósitos, dado que, si los bancos prestan dinero sin tener en cuenta el riesgo del deudor, los depósitos del público estarán desprotegidos.

Como dice Freixas y Rochet (2008),

> Banking operations may be varied and complex, but a simple operational definition of a bank is available: a bank is an institution whose current operations consist in granting loans and receiving deposits from the public. This is the definition regulators use when they decide whether a financial intermediary (…) has to submit to the prevailing prudential regulations for banks. This legal definition has the merit of insisting on the core activities of banks, namely, deposits and loans. (p. 1)

Los lineamientos para las regulaciones bancarias surgen para manejar los principales riesgos que enfrentan los bancos y son definidos por un grupo de reguladores que se

encuentran en el Bank of International Settlement (BIS) en Basilea. El primer acuerdo de Basilea surge principalmente para mitigar el riesgo de crédito y estableció capitales mínimos que deben mantener los bancos, los cuales dependen de su cartera crediticia.

Dentro de la gestión del riesgo de crédito, como dice Bessis (2015), la evaluación de la calidad crediticia de los deudores es un factor crítico. Para ello, existen sistemas de score, que son modelos estadísticos diseñados para distinguir entre buenos y malos pagadores. Dichos modelos son fundamentales para la gestión del riesgo de crédito.

Como dice Abdou y Pointon (2011),

> With the fast growth of the credit industry all over the world and portfolio management of huge loans, credit scoring is regarded as a one the most important techniques in banks, and has become a very critical tool during recent decades. Credit scoring models are widely used by financial institutions, especially banks, to assign credit to good applicants and to differentiate between good and bad credit. (p. 7)

Actualmente, en los modelos tradicionales de score para préstamos bancarios, se utilizan datos sociales, económicos e información histórica de comportamiento de pago dentro del sistema financiero para predecir capacidad de pago.

Como dice Mermelstein (2006),

> Dichos modelos mezclan entre el conjunto de variables predictoras de la morosidad algunas de tipo socio-demográfico relativas al deudor, junto con algunos ratios como el de loan-to-value, y además descansan en el supuesto de que el comportamiento de pagos pasado del deudor es buen predictor del comportamiento futuro (…), por lo que también suelen incorporar información de burós crediticios entre las variables independientes. (p. 40)

El problema de la metodología tradicional bancaria es que desaprovecha la nueva información que se puede obtener gracias al manejo de los grandes volúmenes de información que reciben las empresas (*Big Data*), la cual puede servir para explicar comportamiento de pago frente a un préstamo.

Por su parte, las *Fintech* están haciendo uso de nuevas variables que surgen a partir del procesamiento de *Big Data* para otorgar préstamos a los sectores que quedan marginados de recibir préstamos bancarios. A continuación, presentamos un fragmento de una

entrevista realizada a Marcos Galperín (CEO de MercadoLibre) donde se evidencia lo mencionado:

> —¿Y por qué ustedes sí les prestan?. —¿Qué vieron que los bancos no ven en esas pymes?
>
> —Nosotros vemos en **tiempo real** toda la historia de ventas que tuvieron en Mercado Libre o en el uso de Mercado Pago. Tenemos sistemas para trazar un **perfil muy preciso** de aquellos a los que les ofrecemos o nos piden crédito. No necesitamos papeles ni balances. Les prestamos a gente que no tiene un Veraz perfecto, incluso. Acá, si pasás los filtros, lo único que tenés que decir es cuánto querés, hacés doble click y te acreditamos la plata.
>
> —¿Y qué los lleva a ustedes a determinar que ese cliente es bueno?
>
> —Básicamente la **historia de ventas** que tienen en **Mercado Libre.** Vemos el volumen, el tiempo que llevan con nosotros y un montón de variables, la reputación que les dan sus contrapartes, los que les compran, cuánto tardan en entregar los productos… (Marcos Galperín 2017, iProfesional)

La hipótesis de la presente investigación es que la incorporación de datos alternativos mejora la estimación del riesgo de crédito. Para ello, vamos a realizar un trabajo de investigación empírica utilizando una base construida a partir de datos secundarios provistos por un banco anónimo, donde las unidades de análisis son las solicitudes de crédito de individuos en Argentina durante 2018. La cantidad y heterogeneidad de las solicitudes analizadas permite extrapolar los resultados al sistema bancario argentino.

El presente trabajo se estructura en tres capítulos principales. En el primero de ellos, se estudiará la utilización de grandes volúmenes de datos en la gestión del riesgo de crédito bancario. En primer lugar, se repasará la literatura acerca de los riesgos que enfrentan los bancos. En segundo lugar, se estudiará específicamente el riesgo de crédito y las regulaciones existentes para controlarlo. Finalmente, se pasará revista del uso de datos alternativos para la gestión del riego de crédito en los últimos tiempos.

En el segundo capítulo, se buscará construir un modelo de riesgo crediticio que permita evaluar la probabilidad de default de individuos dentro del sistema bancario argentino. Para ello, en primer lugar, se describirán los principales aspectos de un modelo de score. En segundo lugar, se repasarán las fórmulas utilizadas para construirlo y, finalmente, se estudiarán los principales indicadores de medición de la exactitud de los modelos de score.

En el tercer capítulo, se compararán modelos de riesgo crediticio con incorporación de variables alternativas evaluándolos en diferentes productos financieros. Para abordar este análisis, primero se realizará un estudio estadístico de las bases input utilizadas. Luego, se construirán un modelo de score con variables alternativas y un modelo de score restringido para cada tipo de producto; evaluando la *performance* de cada uno de ellos. Finalmente, se realizará la comparación de los resultados de *performance* para cada tipo de producto.

Planteamiento de los objetivos

El objetivo general de esta tesis es analizar el impacto que posee la incorporación de datos alternativos en la predicción del riesgo de crédito. Como objetivos específicos se considera describir el uso de los grandes volúmenes de datos en la gestión del riesgo de crédito bancario, construir un modelo de riesgo de crédito que permita evaluar la probabilidad de *default* de individuos dentro del sistema bancario argentino y comparar modelos de riesgo crediticio con incorporación de variables alternativas evaluándolos en diferentes productos financieros.

"This business is only for well-managed institutions. It's not meant to be done home alone without adult supervision"
Michael L. Brosnan

Capítulo 1: Utilización de los grandes volúmenes de datos en la gestión del riesgo de crédito bancario

En este capítulo se estudiará el aprovechamiento de los grandes volúmenes de datos en la gestión del riesgo de crédito bancario. En primer lugar, se estudiarán las características principales de los bancos y los riesgos que enfrentan. En segundo lugar, se estudiará específicamente el riesgo de crédito y las regulaciones existentes para administrarlo. Finalmente, se destacará la utilización reciente de datos alternativos para la gestión del riego de crédito.

1.1 Características de los bancos comerciales y los riesgos que enfrentan

Los bancos comerciales son entidades financieras, dado que se dedican en sus operaciones habituales a intermediar entre la oferta y demanda de recursos financieros; y, en Argentina, están regulados por el BCRA.

El BCRA también regula a otras entidades financieras: bancos de inversión (destinados principalmente a asistir a grandes empresas en la obtención de préstamos y de capital y en asesorarlas en procesos de restructuraciones, fusiones o adquisiciones), bancos hipotecarios, compañías financieras, sociedades de ahorro y préstamos para la vivienda u otros inmuebles (destinadas principalmente a otorgar créditos en base a un ahorro previo por parte del cliente) y cajas de crédito.[1]

1.1.1 Funciones principales de los bancos comerciales

Los bancos comerciales tienen como función principal tomar los depósitos de aquellos que poseen dinero y otorgar préstamos a los que necesitan fondeo. Tanto los que depositan como los que piden dinero pueden ser individuos, organizaciones o gobiernos[2]. La diferencia entre el interés que recibe el banco por los préstamos que realiza y el que paga

[1] Cabe notar que las empresas de seguros están excluidas del precedente listado, y esto se debe a que si bien son entidades financieras (dado que administran los ahorros del asegurado tal que puedan afrontar un eventual suceso futuro negativo), están reguladas específicamente por la SSN.

[2] La banca minorista es la que atiende a los individuos o a pequeñas empresas y la banca mayorista es la que presta servicios a grandes organizaciones o gobiernos.

por los depósitos que recibe se denomina *spread* y es su principal ganancia en la mayoría de los países (International Monetary Fund [IMF], 2012).

Los bancos comerciales toman depósitos de corto plazo y realizan préstamos a más largo plazo, lo cual es viable, en parte, gracias a que la mayoría de los depositantes no requieren del dinero a corto plazo y mantienen los depósitos en el banco. Asimismo, los bancos complementan el fondeo acudiendo a los mercados financieros: emitiendo obligaciones negociables o bonos, prestando títulos a cambio de dinero (operación denominada repo) o a través de la titulización (también conocida por el anglicismo securitización), que consiste en armar un paquete con préstamos que el banco posee en libros y venderlo en el mercado (IMF, 2012).

Asimismo, los bancos comerciales operan como un sistema de pagos doméstico e internacional dado que, a partir de la existencia de estos, se canalizan las transacciones monetarias entre las distintas partes: compradores y vendedores (aquí juegan un rol importante las tarjetas de crédito y débito), empleadores y empleados, acreedores y deudores, contribuyentes y gobierno, etc.

En un sistema de encaje fraccionario[3] los bancos comerciales tienen la función adicional de crear dinero. Según Rosignuolo (2017),

> La creación de dinero en un sistema de encaje fraccionario -tasa de efectivo mínimo o tasa de encaje inferior al 100%- es la resultante de una expansión primaria vía Banco Central (a través de los determinantes de la base: Sector Externo, Sector Gobierno y Sector Financiero) y de una creación secundaria a cargo de los bancos comerciales y del público, medida a través de los multiplicadores de la base monetaria. (p. 11)

Esta creación secundaria de dinero se debe a que los bancos comerciales no encajan el total de los depósitos, por lo tanto, parte del dinero depositado es vuelto a prestar (y así sucesivamente) generando una oferta monetaria superior a la generación inicial de dinero[4].

[3] El encaje es la porción de depósitos que un banco debe mantener en reservas líquidas y, por tanto, no se puede usar para inversiones ni préstamos. El encaje mínimo lo fija el banco central y, luego, los bancos comerciales pueden decidir encajar un porcentaje superior. Un sistema de encaje fraccionario es aquel en el cual en encaje regulatorio es menor al 100%.

[4] Como menciona Rosignulo (2017), la creación secundaria de dinero presenta filtraciones (el BCRA no puede regularla completamente) dado que depende, por un lado, de la decisión de encaje de los bancos comerciales y, por otro lado, de la decisión de cartera de los individuos entre los distintos tipos de depósito y el efectivo.

1.1.2 Riesgos financieros

El riesgo es la incertidumbre acerca de los eventos futuros y de su efecto negativo en los resultados de la entidad. La identificación, medición y control de los riesgos financieros se denomina administración financiera de riesgos (*Financial Risk Management*). Tanto los administradores de riesgos como los reguladores de los bancos tienen como objetivo aumentar la resiliencia de las entidades para enfrentar situaciones adversas (Bessis, 2015).

La clasificación de los distintos riesgos financieros se basa en la fuente de incertidumbre de estos. Jorion (2007) menciona la existencia de cuatro grandes grupos de riesgos: riesgo de crédito, riesgo de mercado, riesgo de liquidez y riesgo operacional. Sin embargo, la diversidad de riesgos a la que se enfrentan las instituciones financieras es más amplia. Esto se observa en la comunicación "A" 5398 del BCRA del 2013 donde detalla que las entidades financieras deben contar con un proceso integral de gestión de riesgos para evaluar y controlar los siguientes riesgos: crédito, liquidez, mercado, tasa de interés, operacional, titulización, concentración, reputacional y estratégico. Asimismo, con la resolución 30/2017 de la Unidad de Información Financiera (UIF) se agrega el requerimiento de incorporar al riesgo de lavado de activos y financiamiento del terrorismo dentro de la gestión integral de riesgos.

A continuación, pasaremos revista a los riesgos que enfrentan las entidades financieras exceptuando al riesgo de crédito, que será tratado en el apartado siguiente dado que constituye el núcleo del presente trabajo.

a) Riesgo de liquidez

El riesgo de liquidez se refiere a la capacidad de las entidades financieras de fondear los incrementos de los activos y cumplir con sus obligaciones a medida que éstas se hacen exigibles, sin incurrir en pérdidas significativas (BCRA, 2013). Según Jorion (2007), el riesgo de liquidez toma dos formas: riesgo de liquidez de activos y riesgo de liquidez de fondeo; el primero se refiere a la dificultad de poder realizar una transacción a los precios normales de mercado, lo cual puede ocurrir por variaciones de las condiciones de mercado o por las condiciones de mercado que poseen ciertos productos financieros; el segundo se refiere a la incapacidad de atender las obligaciones de pago, lo cual puede forzar la pronta liquidación de activos.

b) Riesgo de mercado

El riesgo de mercado es el riesgo de que la entidad sufra pérdidas debidas a fluctuaciones adversas del mercado que deprecien el valor de sus posiciones; las causas de las fluctuaciones incluyen a las tasas de interés, a los índices de acciones, al precio de *commodities* y a los tipos de cambio (Bessis, 2015). El riesgo de mercado se controla mediante límites en las exposiciones y en los nocionales con el indicador denominado valor en riesgo (VaR)[5] y con supervisiones específicas de los administradores de riesgo (*Risk Managers*).

c) Riesgo de Tasa de Interés

El riesgo de tasa de interés es la posibilidad que las tasas de fondeo crezcan por encima de las tasas que reciben las entidades financieras por sus colocaciones. Según Freixas y Rochet (2008), los bancos tienen una función de transformación de activos tal que transforman depósitos de corto plazo en préstamos de largo plazo y el riesgo de tasa de interés surge con la posibilidad de que las tasas de los depósitos crezcan por encima de las tasas contractuales de los préstamos que ofrecen.

d) Riesgo operacional

El riesgo operacional es el riesgo de pérdidas resultantes de inadecuados o fallidos procesos internos (fallas en el registro de las transacciones), de la actuación de las personas o sistemas (fraudes internos o externos) y de eventos externos; asimismo, incluye al riesgo legal, que es la exposición a penalidades y multas por acciones de supervisión o por acuerdos contractuales (Jorion, 2007).

e) Riesgo de titulización

Las titulizaciones son una fuente alternativa de financiación y de transferencia de riesgo a los inversores, pero generan nuevos riesgos para la entidad. Por un lado, la entidad tendrá riesgo de crédito por las posiciones compradas (por falta de pago de los activos subyacentes) y por las posiciones vendidas (por la posibilidad de cubrir el default de los activos subyacentes para evitar riesgo reputacional). Asimismo, las titulizaciones poseen riesgo de liquidez dado que si cae la liquidez del mercado se puede imposibilitar el lanzamiento de una titulización que está siendo estructurada. Por otro lado, tanto las titulizaciones que están siendo estructuradas como las posiciones compradas poseen

[5] El VaR es una medida estadística para medir el riesgo de una inversión. Indica la pérdida máxima que puede sufrir la entidad en un horizonte temporal determinado con un nivel de confianza dado.

riesgo de mercado debido a la variación adversa en los precios. También, las titulizaciones compradas poseen riesgo de concentración dependiendo de las características de los activos subyacentes. Finalmente, pueden surgir dificultades legales que impidan que una titulización en proceso de estructuración pueda ser vendida.

f) Riesgo de concentración

El riesgo de concentración surge de mantener posiciones que mantengan características similares: que sean con la misma contraparte o garante, que sean en la misma ubicación geográfica, que sean del mismo sector económico o que estén respaldadas por el mismo activo de garantía. Estas concentraciones pueden producir pérdidas para la entidad financiera producto de dificultades financieras de una contraparte específica, de riesgos comunes para un mismo sector geográfico o económico o por la caída de precio de la garantía.

g) Riesgo reputacional

Es aquel asociado a una percepción negativa sobre la entidad financiera por parte de las contrapartes, accionistas, inversores, tenedores de deuda, analistas de mercado que afecten adversamente la capacidad de la entidad para continuar con sus relaciones comerciales, para iniciar nuevos negocios o para acceder a fuentes de fondeo (BCRA, 2013). Cabe aclarar que la confianza del público y de los inversores en la entidad dependen de su reputación, por lo tanto, mantener el prestigio es importante para mantener el nivel de pasivos.

h) Riesgo estratégico

El riesgo estratégico es aquel asociado a variaciones desfavorables en los parámetros asociados a un proyecto que la entidad va a llevar a cabo o bien al riesgo de realizar un proyecto inadecuado.

i) Riesgo de lavado de dinero y financiamiento al terrorismo

Es el riesgo asociado a que la entidad financiera sea utilizada por terceros para realizar actos de lavado de dinero y/o financiamiento al terrorismo. Con la resolución 30/2017 la Unidad de Información Financiera (UIF) estableció que las entidades financieras deben incorporar en su proceso de autoevaluación de capital los riesgos asociados en cada una de sus líneas de negocio a que su estructura sea utilizada por terceros para los fines comentados y evaluar la efectividad de los controles realizados.

1.2 Sobre el riesgo de crédito y las principales regulaciones existentes

En este apartado, en primer lugar, repasaremos el concepto de riesgo de crédito para las entidades financieras y los factores utilizados para medirlo. Luego, estudiaremos las principales regulaciones que existen para administrarlo y controlarlo.

1.2.1 Riesgo de crédito

El riesgo de crédito nace a partir de las dificultades que presentan los deudores para el cumplimiento de pago de sus obligaciones contractuales, lo cual produce una incertidumbre sobre los flujos que esperan recibir las entidades financieras producto de los contratos pautados. Como dice Jorion (2007, p. 25), "*Credit risk is the risk of losses owing to the fact that counterparties may be unwilling or unable to fulfill their contractual obligations. Its effect is measured by the cost of replacing cash flows if the other party defaults*."

Como menciona Bessis (2015), el riesgo de crédito para la entidad también puede aumentar en aquellos casos que la calidad crediticia del deudor empeore, volviendo al crédito otorgado a dicha contraparte más riesgoso.

Las pérdidas potenciales que puede sufrir la entidad financiera son el principal, los intereses y los costos asociados a los esfuerzos de recuperación. Los factores para medir las pérdidas potenciales del riesgo de crédito son tres: la probabilidad de *default* (PD), la exposición al momento del *default* (EAD) y las pérdidas generadas a partir del *default* (LGD). Estos tres componentes de riesgo se utilizan para caracterizar el estado actual que posee un cliente o un producto frente al riesgo de crédito. De esta forma, si un deudor sufre un empeoramiento en su capacidad de pago, se va a ver reflejado en un aumento de la PD, por lo que migrará de estado.

A continuación, estudiaremos cada uno de los componentes que miden el riesgo de crédito.

I. Probabilidad de *default* (PD)

Como mencionamos previamente, la probabilidad de *default* es la posibilidad que un deudor no cumpla con sus obligaciones contractuales de pago de deuda.

Las características específicas a partir de las cuales se declara que una deuda está en *default* depende de la entidad que haga la evaluación. Las agencias de *rating* (entidades

cuya función principal es otorgar una valoración de riesgo de crédito a compañías o productos financieros) consideran que el *default* se produce al momento en que transcurre un día de atraso de al menos un dólar en el cumplimiento de las obligaciones pactadas (Bessis, 2015). Para los reguladores de las entidades financieras, el *default* de la cartera crediticia se define de forma distinta.

Los acuerdos de Basilea son recomendaciones sobre regulación bancaria emitidos por el Comité de Basilea de Supervisión Bancaria (CBSB) y tienen como objetivo conseguir una unidad normativa para los bancos de los distintos países. Según los acuerdos de Basilea II, el *default* ocurre si se da alguna o todas de las siguientes casuísticas: a) el banco considera improbable que el deudor pague completamente sus obligaciones sin mediar acciones por parte del banco, como la venta de garantías y b) el deudor está atraso en más de 90 días en cualquier obligación significativa con el banco (CBSB, 2006). Según los acuerdos de Basilea, los bancos deben utilizar esta definición de *default* para sus estimaciones internas de PD, LGD y EAD:

> A bank must use the reference definition of default for its internal estimations of PD and/or LGD and EAD. However, as detailed in paragraph 454, national supervisors will issue guidance on how the reference definition of default is to be interpreted in their jurisdictions. Supervisors will assess individual banks' application of the reference definition of default and its impact on capital requirements. (CBSB, 2006, p. 213)

Esta definición está alineada con el grado igual o superior a tres de la clasificación de deudores del BCRA, según la comunicación "A" 7156 (BCRA, 2020).

La probabilidad de *default*, dado un estado crediticio, depende de la situación macroeconómica que enfrenten los países. De esta forma, la probabilidad de *default* tenderá a aumentar en momentos desfavorables del ciclo económico y tenderá a disminuir en situaciones normales. Es por ello, que la PD puede calcularse *Point in Time* y será una probabilidad condicionada a la situación macroeconómica actual; o bien, puede calcularse *Through the Cycle* y será una PD de largo plazo no condicionada al ciclo (dado que surge del promedio de PDs en los distintos momentos del ciclo).

Los grandes bancos comerciales pueden calcular las PDs a partir de sus propias bases de datos con información histórica, dado que poseen grandes volúmenes de clientes para

realizar el cálculo; de esta forma, pueden contar los *defaults* observados y generar frecuencias de *default* para distintos momentos del tiempo (Bessis, 2015).

II. Exposición al momento del *default* (EAD)

La EAD es una estimación de la exposición (saldo de deuda) que tendrá el deudor con la entidad al momento del *default*, la cual es desconocida hasta que efectivamente se produce el *default*, dado que depende de factores inciertos.

En el caso de los préstamos que poseen un calendario de amortización contractual, el flujo de pagos efectivo puede diferir del calendario porque el cliente puede realizar pagos parciales o totales del principal. Asimismo, hay préstamos a tasa de interés variable que dependen de índices de mercado estocásticos, generando desconocimiento acerca del flujo exacto que percibirá la entidad financiera.

Por otro lado, el banco ofrece productos que poseen saldos fuera de balance, como las tarjetas de crédito o los adelantos de cuenta corriente. El saldo dentro de balance es el saldo efectivamente consumido por el cliente y el saldo fuera de balance es la diferencia entre el total de crédito otorgado por la entidad financiera y el consumo del cliente; ambos saldos son inciertos. Dado que se espera que los clientes aumenten el consumo de la línea de crédito en situaciones de dificultad financiera, es que se estima un *Credit Conversion Factor* (CCF) para estimar el saldo al momento del default de los productos con saldos fuera de balance (CBSB, 2006).

III. Pérdidas generadas a partir del *default* (LGD)

Cuando un deudor entra en *default* con una exposición determinada, la entidad financiera incurre en costos para recuperar el saldo adeudado. Estos costos son parte de las pérdidas generadas a partir del *default* y forman parte de la pérdida total. Asimismo, gracias a los esfuerzos de recuperación y las garantías que posea el crédito otorgado, se generan flujos positivos para la entidad financiera a lo largo de los meses posteriores al evento de *default*. La tasa de recupero es el porcentaje que representa el valor actual de los recobros menos los gastos sobre el total de la deuda al momento del *default*; la LGD es el complemento de la tasa de recupero.

La posibilidad de generar recobros a partir de las garantías que poseen los créditos va a depender de la liquidez de las garantías y de las condiciones de mercado. La incertidumbre acerca del valor al cual se venderán las garantías al momento de ejecutarlas

es incierto, por lo que el valor reconocido de las mismas por la entidad financiera es menor a su precio. Esta diferencia se denomina *haircut* y se utiliza como amortiguador de las fluctuaciones futuras (Bessis, 2015).

1.2.2 Aspectos regulatorios del riesgo de crédito

I. Regulaciones sobre el capital

Los riesgos financieros deben ser regulados y una de las formas de hacerlo es a través de la implementación de regulaciones sobre el capital. Estas regulaciones tienen el objetivo de aumentar la resiliencia de los bancos, tal que el capital sea la última línea de defensa para evitar el colapso en situaciones de stress. En 1988 se establece el primer acuerdo de Basilea que tiene como objetivo definir el capital regulatorio, que es el mínimo nivel de capital que deben mantener los bancos para hacer frente a pérdidas inesperadas. En este acuerdo se regula principalmente el riesgo de crédito.

> It should also be emphasised that capital adequacy as measured by the present framework, though important, is one of a number of factors to be taken into account when assessing the strength of banks. The framework in this document is mainly directed towards assessing capital in relation to credit risk (the risk of counterparty failure) but other risks, notably interest rate risk and the investment risk on securities, need to be taken into account by supervisors in assessing overall capital adequacy. (BIS, 1988, p. 2)

Para determinar el capital regulatorio, Basilea establece un ratio de solvencia (Cooke Ratio), que es un porcentaje de los activos que poseen los bancos ponderados por el riesgo de los mismos. Dichos ponderadores varían según el tipo de activo, pero no son calculados por cada entidad, sino que están fijados por el regulador. Con el acuerdo de Basilea II (publicado inicialmente en 2004 y con una versión definitiva hacia 2006), se introdujeron ponderadores de riesgo sensibles a los riesgos particulares de cada entidad y se incluyeron mitigadores de riesgo (como las garantías). En este sentido, para el cálculo del ponderador, se permite la utilización de estimaciones propias de los componentes del riesgo de crédito (PD, LGD y EAD), sujeto a modelos especificados por el regulador[6].

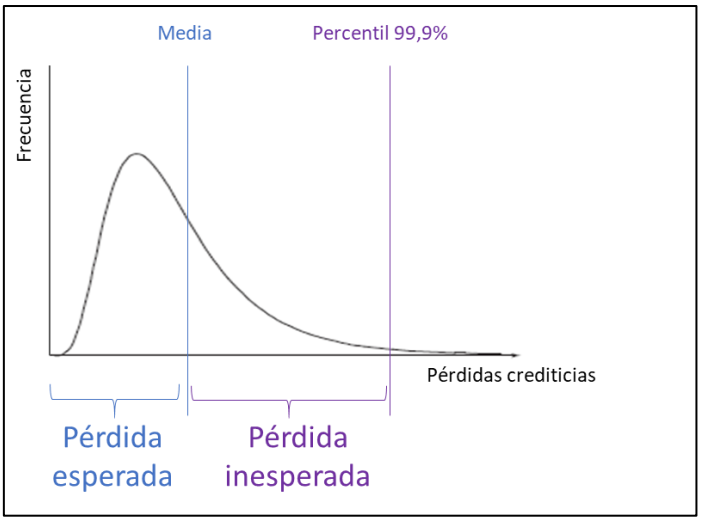
[6] Cabe aclarar que, según lo establecido por BIS (2006), solo los bancos que poseen sistemas internos de rating pueden realizar estimaciones propias de los factores de riesgo de crédito. El resto de las entidades debe seguir un enfoque estandarizado, utilizando ponderadores fijados por el supervisor.

En este segundo acuerdo, se estableció también que los bancos deben presentar un informe de autoevaluación de capital (IAC) en el que realicen su propia evaluación respecto al capital necesario para cubrir las pérdidas inesperadas teniendo en cuenta todos los riesgos que enfrentan; este capital se denomina capital económico. Para el cálculo del capital económico, Basilea permite el uso de modelos propios de valuación de riesgo y establece que debe ser comparado con el capital regulatorio a fin de aumentarlo en caso de ser necesario. Como se establece en el acuerdo de Basilea: "*The approach used must be based on the firm's internal economic capital approach, be well-documented and be subject to independent validation* (BIS, 2006, p. 262)."[7] Sin embargo, como menciona Hull (2015), los bancos intentan mantener el capital económico por debajo del capital regulatorio para no tener que aumentarlo.

El capital económico por riesgo de crédito es la máxima pérdida crediticia (pérdida inesperada), por sobre la pérdida esperada, que puede tener un banco para un horizonte de tiempo dado y con un determinado nivel de confianza. El horizonte temporal definido por el Comité de Basilea tanto para el cálculo del capital económico como para el cálculo del capital regulatorio es de 12 meses (BIS, 2006). Este cálculo se realiza a partir de la distribución de las pérdidas crediticias, la cual posee una asimetría hacia la derecha: mayor frecuencia en pérdidas de montos bajos y poca frecuencia en pérdidas de gran magnitud. De esta forma, para obtener la pérdida inesperada se obtiene el valor de pérdida para un cuantil determinado (que indica el nivel de confianza) de una distribución de pérdidas con un horizonte de tiempo dado y se resta la pérdida crediticia esperada. A continuación, se presenta un gráfico que expone lo comentado.

Gráfico 1.1 Distribución de pérdidas crediticias

[7] BIS (2006) establece que aquellos bancos que sigan el enfoque estandarizado (utilizando las estimaciones de PD, LGD y EAD establecidas por el regulador para el cálculo del capital regulatorio), deben utilizar sus estimaciones propias de parámetros en la evaluación del capital económico.



Fuente: Elaboración propia

Según Bessis (2015), el nivel de confianza para determinar el capital económico final es un parámetro clave para el banco dado que, si el capital económico está definido con una confianza del 99%, en promedio, en el 1% de los casos las pérdidas inesperadas serán superiores al capital, volviendo al banco insolvente. Por lo tanto, el complemento del nivel de confianza es la probabilidad de default del banco, la cual tiene una incidencia directa en el rating que le darán al banco las calificadoras de riesgo y, por ende, en su costo de financiamiento.

II. Regulaciones sobre las Provisiones

Si bien las regulaciones sobre el capital permiten administrar las pérdidas inesperadas, también existen regulaciones sobre las provisiones contables para administrar la pérdida crediticia esperada.

Recordando la definición de los componentes del riesgo de crédito realizada en el apartado 1.2.1, podemos comprender que la pérdida esperada de un préstamo es la probabilidad que entre en *default* (PD) multiplicada por el saldo de deuda (exposición) al momento de default (EAD) multiplicada por el porcentaje de la exposición que no será recuperado luego del default (LGD):

$$Pérdida\ esperada = PD * EAD * LGD$$

Para obtener la pérdida esperada total por riesgo de crédito, se suman las pérdidas esperadas de cada préstamo. Basilea establece que para cubrir las pérdidas esperadas los bancos deben realizar provisiones contables que estén alineadas a las Normas

Internacionales de Información Financiera (NIIF). Argentina, al ser un miembro del G-20 adopta las normativas de Basilea y son adaptadas localmente a través de los comunicados del BCRA.

En cuanto a las pérdidas esperadas, el regulador argentino (BCRA) estableció en la comunicación "A" 6430 que a partir del 01/01/2020 los bancos deben realizar sus provisiones de acuerdo con lo establecido en el punto 5.5 de las NIIF 9. El punto 5.5 de dichas normas incorpora el criterio de deterioro de valor para el cálculo de las pérdidas crediticias esperadas; de forma tal que, si un préstamo tuvo un incremento significativo de riesgo respecto al riesgo inicial con el cual se originó, el banco debe tener en cuenta no solo la pérdida esperada a 12 meses, sino que debe establecer la pérdida esperada durante toda la vida del activo (NIIF 9, 2014). Es decir, que para los activos que no hayan tenido un incremento significativo de riesgo, el horizonte temporal para evaluar la pérdida esperada es de 12 meses y para aquellos que hayan sufrido un incremento significativo de riesgo, el horizonte temporal es todo el período contractual.

Si bien cada entidad es la encargada de evaluar cuándo se produce un incremento significativo de riesgo, como criterio general, la norma establece que, si un activo posee más de 30 días de atraso, el mismo ha tenido un incremento significativo de riesgo. Cabe aclarar, que las entidades deben reconocer el valor temporal del dinero en el cálculo de la pérdida esperada.

1.3 Datos alternativos en la gestión del riesgo de crédito

La información tradicional utilizada en los modelos de crédito bancarios a individuos comprende datos socioeconómicos e información de bureau de créditos respecto a morosidad en el sistema financiero. Sin embargo, la evolución tecnológica puso un mayor volumen de información a disposición de las empresas.

En un informe de Emerj (consultora estadounidense de inteligencia artificial) de abril 2020 se refleja lo comentado previamente:

> While in the past lenders looked at only a few metrics like FICO score and income, companies have started looking at an individual's entire life and even their vast digital footprint to determine how likely they are to default. This is referred to as "alternative data" about potential borrowers. The idea is that extra

> data provides not just more insight into people with established FICO scores, but that it can be particularly useful for determining the creditworthiness of people without a traditional credit history. (Emerj, 2020)

Los modelos tradicionales de score crediticio que construyen los bureaus de crédito para el sistema financiero dependen principalmente de las siguientes variables: cantidad y tipo de cuentas de crédito que posee el individuo, longitud del historial crediticio, historial de pagos de los créditos tomados (en productos como: tarjetas de crédito, préstamos personales, préstamos prendarios, préstamos hipotecarios) y el porcentaje de crédito utilizado respecto al crédito otorgado (Equifax, *s.f.*)

Las variables alternativas que pueden ser utilizadas por los modelos de score comprenden, y no se limitan, a las siguientes:

- Información sobre pagos telefónicos (y de otros servicios)
- Información sobre el pago del alquiler
- Registros gubernamentales
- Hábitos de consumo
- Información de redes sociales
- Seguimiento en el historial de navegación de internet
- Seguimiento de la ubicación geográfica

Los principales bureaus de crédito a nivel internacional analizan datos alternativos en la construcción de los modelos de score.

Por ejemplo, el bureau de crédito FICO (creador de los scores con insignia FICO, los cuales son utilizados en la mayoría de las decisiones de crédito en Estados Unidos) incorporó información de pagos de alquileres lanzando en 2014 el modelo de score denominado FICO9. Asimismo, actualmente están analizando incorporar datos provenientes del uso de las cuentas a la vista y observan que les permite puntuar a 15 millones de consumidores estadounidenses que no tienen información suficiente para ser puntuados por su modelo de score tradicional (FICO, 2019).

También, las empresas FICO, Equifax y LexisNexis desarrollaron en conjunto un modelo de score denominado "FICO Score XD"[8] para la población estadounidense con poco o nulo historial crediticio en el sistema financiero que combina información tradicional con información de pagos telefónicos (fijos y celulares), de pagos de cable y de otros servicios, permitiendo valuar el riesgo crediticio de más del 70% de las solicitudes de crédito que quedaban sin ser puntuadas por el modelo de score tradicional (FICO, 2019).

Cabe aclarar, que no toda la información alternativa disponible resulta valiosa para ser incluida en los modelos de score. Según FICO (2015), la información debe contar con seis requisitos para que pueda formar parte de sus modelos de score: cumplir con los estándares regulatorios, tener suficiente profundidad histórica, cubrir la mayor cantidad de población posible, ser precisa, contar con poder predictivo y cumplir con el requisito de ortogonalidad (agregar valor por sobre la información contenida en el resto de las variables explicativas).

Por su parte, la empresa estadounidense Lenddo se dedica especialmente a la utilización de datos alternativos en modelos de score para aquellos individuos o pequeñas empresas que no cuentan con un historial crediticio suficiente para acceder al crédito. Su sistema se basa en generarle un score a cada individuo que desea ser calificado a partir de procesar información contenida en los *smartphones* de estos (como ser el historial de navegación, el uso de las redes sociales y la geolocalización).

Asimismo, la empresa ZestFinance (dedicada a generar scores crediticios a partir de modelos de decisión avanzados) recibió una inversión estratégica de Baidu (proveedor líder de búsquedas en internet en idioma chino) para utilizar datos de geolocalización, de historial de navegación y de historial de pagos en un modelo de score para el mercado chino (Businesswire, 2018).

Como menciona Equifax, la utilización de datos alternativos permite una mejor predicción del riesgo de crédito y una mayor inclusión financiera dado que disminuye la asimetría de información entre prestador y prestatario (Roadshow 2019). Los nuevos proveedores de crédito están creciendo a expensas de los jugadores tradicionales, en parte, gracias a la incorporación de datos alternativos en sus modelos de score (Forbes, 2019). Esta discrepancia en el uso de datos no tradicionales se refleja en un estudio realizado por

[8] Se lanzó una primera versión en 2016 denominada FICO Score XD y una versión mejorada en 2018 que cubre un mayor porcentaje de población.

Experian (uno de los principales bureaus de crédito de Estados Unidos) para el mercado estadounidense: "*Many fintech and other nonbank lenders routinely use an array of alternative data; banks and credit unions are incorporating limited forms of alternative data into their processes at a slower pace*" (Experian, 2018).

Mercadolibre lanzó en 2017 créditos a pequeñas y medianas empresas y en 2019 lanzó créditos a consumidores. Sus modelos de score se basan en la utilización de información no tradicional obtenida gracias a su plataforma comercial digital. Por ejemplo, para otorgar créditos a empresas procesan información no tradicional como ser datos sobre reputación y actividad, sobre el tiempo que tardan en responder las preguntas de los clientes, sobre el tiempo que tardan en enviar los pedidos y sobre el flujo de ventas (Infobae, 2017). Como dice Marcos de Santos, ejecutivo de Mercadolibre:

> Tenemos información financiera y también de comportamiento -qué tan rápido responden a problemas del cliente, por ejemplo-; son 2.000 variables que nos permiten entenderlas muy bien y nos hace incluir empresas que si las miraramos solo con los Bureau de crédito estarían afuera. (Infotechnology, 2019)

Asimismo, los créditos que ofrece Mercadolibre a los consumidores utilizan información del historial de consumo. Como dice Paula Arregui, ejecutiva de Mercadolibre: "Aprovechando toda la información que tenemos en nuestro sistema incluimos a mucha gente que los modelos de scoring tradicionales dejan afuera porque no los conoce" (Infotechnology, 2019).

En este capítulo, se hizo un repaso del uso de datos alternativos en la gestión de riesgo crediticio; en el capítulo siguiente, vamos a revisar qué es y cómo se construye un modelo de *score*.

Capítulo 2: Sobre la construcción de modelos de *credit scoring* para evaluar la probabilidad de *default* de individuos dentro del sistema bancario argentino

En este capítulo se intentará explicar cómo se construye un modelo de riesgo crediticio para evaluar la probabilidad de *default* de individuos dentro del sistema bancario argentino. En primer lugar, se describirán las principales características de un modelo de score. En segundo lugar, se repasará la técnica utilizada para construirlo y, finalmente, se estudiarán los principales indicadores de performance del modelo de *score*.

2.1 ¿Qué es un modelo de score?

Los modelos de *score* son utilizados para estimar la calidad crediticia de los deudores minoristas. Son modelos estadísticos que, en función de las características que posee cada individuo, le asignan un número con el objetivo de distinguir entre buenos y malos pagadores. Cuanto mejor gestionado esté el riesgo de crédito y, por ende, más precisa sea la estimación del riesgo de los créditos otorgados, menor será la morosidad de la cartera, impactando en las ganancias del banco.

Cabe aclarar que para estimar el riesgo de crédito de las empresas e instituciones se utilizan escalas de *rating*, las cuales ordenan a los deudores según su calidad crediticia. Estos *ratings* se construyen en función de datos cuantitativos, de datos cualitativos y del juicio experto del analista.

Los modelos de *score* son utilizados en distintos procesos de gestión dentro de los bancos, en mi opinión, los más relevantes son los siguientes:

- Admisión de solicitudes de crédito minoristas (*scores* de admisión)
- Seguimiento de la cartera minorista (*scores* de comportamiento)
- Planificaciones de producto / estrategias de *marketing*
- Evaluación por parte del sector de recuperaciones sobre qué clientes tienen mayor probabilidad de repago.
- Input para el cálculo de la probabilidad de *default*, que tiene impacto en el capital y las provisiones que deben mantener las entidades para cumplir con los estándares regulatorios. Asimismo, incide en variables de gestión como ser el cálculo del rendimiento ajustado al riesgo.

En este trabajo se analizarán los modelos de *score* de admisión, los cuales tienen incidencia en las nuevas solicitudes de crédito que reciben los bancos.

Es importante destacar que, según el informe de inclusión financiera del BCRA (2020), el 37,3% de los adultos contaban a septiembre 2019 con al menos un financiamiento otorgado por entidades financieras, cifra que se eleva al 48,8% considerando al sistema financiero ampliado (que incluye a los proveedores de crédito no considerados entidades financieras por el BCRA, como los créditos de las Fintech o las financiaciones de empresas comerciales). En dicho informe, se observa que el porcentaje de adultos financiados no tuvo variaciones significativas en los últimos tres años, pero sí varía por provincia[9].

A continuación, se presenta un ranking de los diez bancos con mayores préstamos otorgados a diciembre 2019 según información obtenida de la página web del BCRA.

Gráfico 2.1: Ranking de préstamos

Puesto	Entidad financiera	Préstamos otorgados (en miles de pesos)
1	NACION ARGENTINA	468.903.710
2	BCO GALICIA	302.307.504
3	SANTANDER RIO	266.431.073
4	BANCO PROVINCIA	240.919.209
5	MACRO SA	218.772.002
6	BANCO BBVA ARGENTINA	184.200.433
7	CIUDAD DE BS AS	113.477.608
8	HSBC BANK	107.099.679
9	ICBC	94.123.401
10	PATAGONIA SA	83.241.047

Fuente: Ranking de Préstamos (BCRA, diciembre 2019)

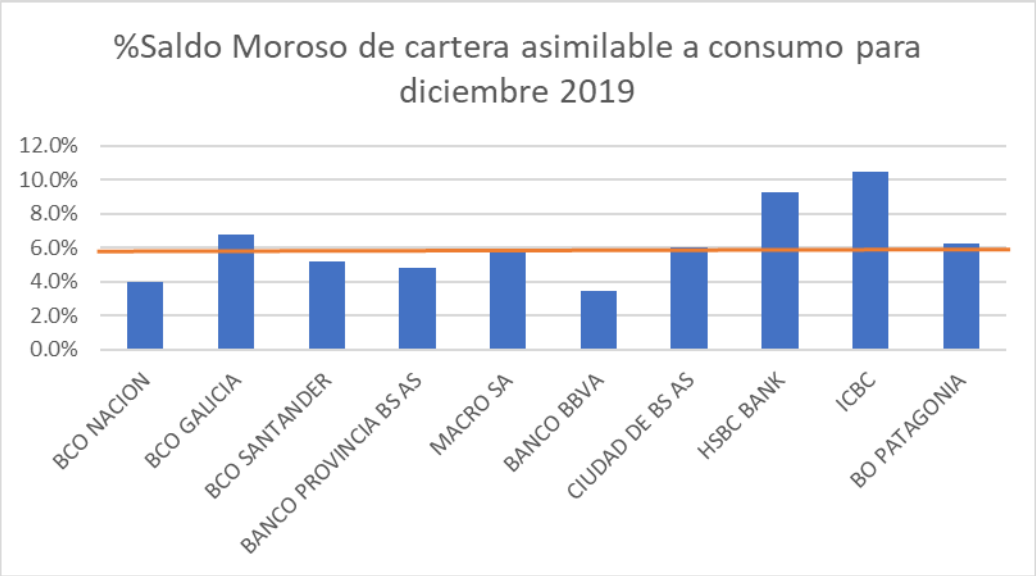
La morosidad de la cartera va a depender de los modelos de *score*, del comportamiento de la cartera y de las políticas crediticias (definidas en función del apetito al riesgo de cada entidad).

A continuación, presentamos un gráfico del porcentaje de saldo moroso sobre la cartera de consumo a diciembre 2019 para los principales bancos mencionados previamente y para todo el sistema financiero (se consideró como morosa a la situación BCRA mayor o igual a 3).

Gráfico 2.2: Proporción de saldos morosos de los principales bancos argentinos.

[9] Ver gráficos A.1 y A.2 del anexo.

Entidad financiera	%Saldo Moroso de cartera asimilable a consumo para diciembre 2019
BCO NACION	4.0%
BCO GALICIA	6.8%
BCO SANTANDER	5.2%
BANCO PROVINCIA BS AS	4.8%
MACRO SA	5.8%
BANCO BBVA	3.5%
CIUDAD DE BS AS	6.0%
HSBC BANK	9.3%
ICBC	10.5%
BO PATAGONIA	6.3%
Total sistema financiero	5.6%



Fuente: Central de deudores (BCRA, diciembre 2019)

2.1.1 Construcción de modelos de *score*

Los modelos de *score* tienen el objetivo de discriminar entre buenos y malos pagadores dentro de una población de individuos. La función de *score* es una combinación de las distintas variables explicativas y tiene como output el *score* estimado para cada cliente. Los coeficientes asociados a cada variable explicativa son el aporte que cada atributo le proporciona al *score* estimado.

En mi opinión, para construir un modelo de *score* hay que identificar, en primer lugar, si es un modelo de admisión o de comportamiento; en segundo lugar, se debe identificar el evento a explicar y la ventana temporal de observación y comportamiento; en tercer lugar, se debe definir la técnica estadística con la que se arma el modelo; luego, se debe contar con una base de datos y revisar la integridad de la información; a continuación, se deben seleccionar las variables explicativas y finalmente evaluar la performance del modelo con

la base utilizada para generarlo (*in the sample*) y con una base de testeo (*out of the sample*).

I. Admisión o comportamiento

Los modelos de admisión se utilizan para evaluar el riesgo crediticio de nuevas solicitudes de crédito (ya sea para nuevos clientes o para clientes existentes que solicitan nuevos créditos); en cambio, los modelos de comportamiento se utilizan para evaluar el desempeño que tendrán los créditos vigentes y utilizan información sobre el desempeño pasado en la entidad (Bessis, 2015). Un modelo de seguimiento puede servir para decidir a qué clientes se les varía el límite del descubierto de cuenta corriente o el límite de compra de las tarjetas de crédito.

II. Evento a explicar

Para evaluar el riesgo de crédito de solicitudes nuevas o de la cartera sana del banco (sin días de atraso) el evento que se intenta explicar suele ser el *default* del crédito otorgado con una ventana de desempeño de seis o doce meses. Para la cartera irregular, el evento a explicar puede ser la probabilidad de regularización o la probabilidad de pasar de un ciclo de atraso al siguiente en los próximos tres o seis meses (por ejemplo: de pasar de tener entre 1 y 30 días de atraso a tener entre 31 y 60 días de atraso). Los modelos de *score* sobre la cartera irregular suelen utilizarse para que la entidad pueda aplicar estrategias de cobro diferenciales segmentando a la población atrasada.[10]

III. Técnica estadística

Entre las técnicas utilizadas para construir los modelos de *scoring* se encuentra el análisis discriminante, los árboles de decisión, el aprendizaje automático a partir de patrones (inteligencia artificial), las redes neuronales y los modelos estadísticos *logit* y *probit*. En este trabajo no apuntamos a discutir las distintas técnicas, sino que utilizaremos el modelo *logit* porque la relación entre las variables explicativas y la explicada se obtiene fácilmente, porque los *scores* calculados se trasladan fácilmente a una probabilidad de *default* (cuanto mayor sea el *score*, mayor será la probabilidad de *default*) y porque, como

[10] Por ejemplo, habrá una población "distraída" que luego de tener unos pocos días de mora, regulariza el pago sistemáticamente; pero, otros grupos de población serán de mayor riesgo y el sector de recuperaciones de la entidad bancaria tendrá que poner más esfuerzos de recobro.

dicen Crouhy, Galai y Mark (2006), es la técnica comúnmente utilizada para construir modelos de score. En el apartado 2.2 detallaremos esta técnica.

IV. Base de datos

En esta instancia, se realiza un análisis univariado sobre las variables que forman parte de la base. El objetivo es conocer la distribución de las variables, la cantidad de valores faltantes y evaluar el poblamiento para cada valor posible de las variables (si una variable toma un mismo valor para todos los registros, no va a ser útil para discernir riesgo).

Para construir un modelo de score, se utiliza información pasada de atributos y desempeño crediticio de forma tal de generar un modelo que permita evaluar en el presente el riesgo que poseen los créditos. Por lo tanto, se debe contar con una base de datos con suficiente profundidad histórica para evaluar la performance pasada.

Como menciona Mermelstein (2006), la principal debilidad de los modelos de *scoring* es el sesgo de selectividad de la muestra ya que, generalmente, las fuentes de información suelen tener solamente solicitudes de crédito que terminaron transformándose en créditos otorgados. "En ese sentido, las muestras disponibles no reflejan a la totalidad del universo de solicitudes que se acercará a la oficina de evaluaciones de un prestamista (Mermelstein, 2006, p. 12)."

V. Selección de variables explicativas

Como dice Mermelstein (2006), las variables predictoras del modelo a implementar deben tener sentido económico y estadístico. Es decir, por un lado, deben tener una lógica económica *a priori* que fundamente incorporarlas en el modelo y, por otro lado, a partir del análisis de los datos, deben demostrar *a posteriori* la validez de incluirlas dentro del modelo.

Para analizar la significatividad estadística de cada variable se puede utilizar la prueba de significatividad individual, que indica si la variable es estadísticamente significativa (con un determinado nivel de confianza) para explicar el evento correspondiente (Wooldridge, 2010). Asimismo, se suelen realizar análisis bivariados para entender el poder discriminatorio de cada variable, que consiste en evaluar la tasa de malos por rango de score. Como medida estadística del análisis bivariado se suele utilizar el *information value*, que mide cuán buena es la variable para distinguir entre buenos y malos pagadores.

Respecto a las variables predictoras, como mencionamos previamente, los modelos de *score* tradicionales que utilizan los bancos se basan en información socioeconómica y en información provista por burós crediticios. Por ejemplo, en el modelo de *score* para individuos que construyen Crouhy, Galai y Mark (2006, p. 215), se utilizan las siguientes variables explicativas: años en el trabajo actual, casa propia o alquilada, nivel de bancarización del individuo (si posee caja de ahorro, cuenta corriente o tarjetas de crédito), ocupación, edad del solicitante y referencias crediticias.

VI. Performance del modelo

El objetivo del modelo es generar scores que discriminen la población que será morosa de la población que será sana. Los modelos se suelen construir de forma tal que, a mayor nivel de *score*, menor la morosidad esperada. Para evaluar la performance del modelo construido y también para comparar modelos entre sí, se pueden utilizar diversos indicadores estadísticos, como ser la prueba de Kolmogorov – Smirnov (KS), el coeficiente de Gini y el AUC (área bajo la curva ROC). Otra medida que suele utilizarse es la probabilidad de cometer el error de tipo 1 y la probabilidad de cometer el error de tipo 2. Los indicadores de performance serán explicados en el apartado 2.3.

2.2 Técnica estadística: Modelos *Logit*

Los modelos *logit*, al igual que los modelos *probit*, utilizan la técnica de regresión estadística multivariada (Bessis, 2015).[11]

La regresión estadística tiene el propósito de estimar relaciones económicas entre variables y probar teorías económicas (Wooldridge, 2010). Es decir, primero necesitamos contar con un modelo económico que plantee la relación entre variables independientes y la variable que se quiere explicar. Por ejemplo, si queremos explicar el salario en función de ciertas variables, podemos plantear el siguiente modelo económico que plantea Wooldridge (2010, p.4):

$$salario = f\ (educ, exper, capacitación) \qquad (1)$$

[11] El término multivariada significa que hay más de una variable explicativa en el modelo especificado.

Dónde la variable explicada *salario* representa el salario por hora que recibe un trabajador; *educ* son los años de escolaridad formal; *exper* refiere a los años de experiencia laboral y *capacitación* representa las semanas de capacitación laboral.

Al plantear el modelo econométrico a partir del modelo económico especificado se debe determinar la forma de la función *f()*. Si la forma funcional es lineal, entonces se trata de un modelo de regresión lineal y si la forma funcional es no lineal se trata de un modelo de regresión no lineal (como lo es el modelo *logit*). De esta forma, y siguiendo con el ejemplo de Wooldridge, se puede plantear el siguiente modelo de regresión lineal:

$$salario = \beta_0 + \beta_1 * educ + \beta_2 * exper + \beta_3 * capacitación + u \qquad (2)$$

Dónde $\beta_0, \beta_1, \beta_2$ son los parámetros del modelo econométrico que se van a estimar a través de algún método de optimización matemática[12] y u es el término de error, que contiene a otras variables no observables que afectan a la variable dependiente.

Las variables que se encuentran a la derecha de la ecuación (educación, experiencia laboral y capacitación) son los predictores de la variable que se intenta explicar (salario). El impacto de cada una de ellas en el salario va a estar determinado por los parámetros del modelo. De esta forma, para el modelo de regresión lineal planteado, si el parámetro estimado es positivo, ante un aumento de la variable independiente, se va a esperar un aumento del salario (y viceversa). Asimismo, si el parámetro estimado es cercano a cero, aportaría evidencia a favor de la falta de significatividad de la variable asociada para explicar la variable independiente.

El modelo de regresión lineal $y = \beta_0 + \beta X + u$[13], para una variable binaria que toma valor 1 en caso de pago y valor 0 en caso de no pago, tiene la ventaja que convierte la probabilidad que suceda el evento una función lineal de las variables explicativas. Esto sucede porque la variable *Y* toma dos valores: 1 y 0, por lo tanto, su esperanza es:

$$\mathrm{E(Y)} = 1 * \mathrm{P(Y = 1)} + 0 * \mathrm{P(Y = 0)} \qquad (3)$$

[12] Entre los más frecuentes, se encuentra el método de mínimos cuadrados y el método de máxima verosimilitud.

[13] Cabe aclarar que β y X son vectores, por lo cual el término βX representa la suma del producto entre cada parámetro y la variable explicativa correspondiente.

Despejando, obtenemos:

$$\mathrm{P(Y = 1) = E(Y)} \qquad (4)$$

Por definición de *Y*:

$$\mathrm{P(Y = 1) = E(Y) = E(}\beta_0 + \beta X + \mathrm{\mathcal{E})} \qquad (5)$$

Por propiedad de la esperanza y dado que la esperanza del error es cero, obtenemos lo comentado previamente:

$$\mathrm{P(Y = 1)} = \beta_0 + \beta E(X) \qquad (6)$$

Sin embargo, la desventaja del modelo lineal es que la variable dependiente puede tomar cualquier valor real, es decir, no está acotada entre 0 y 1.

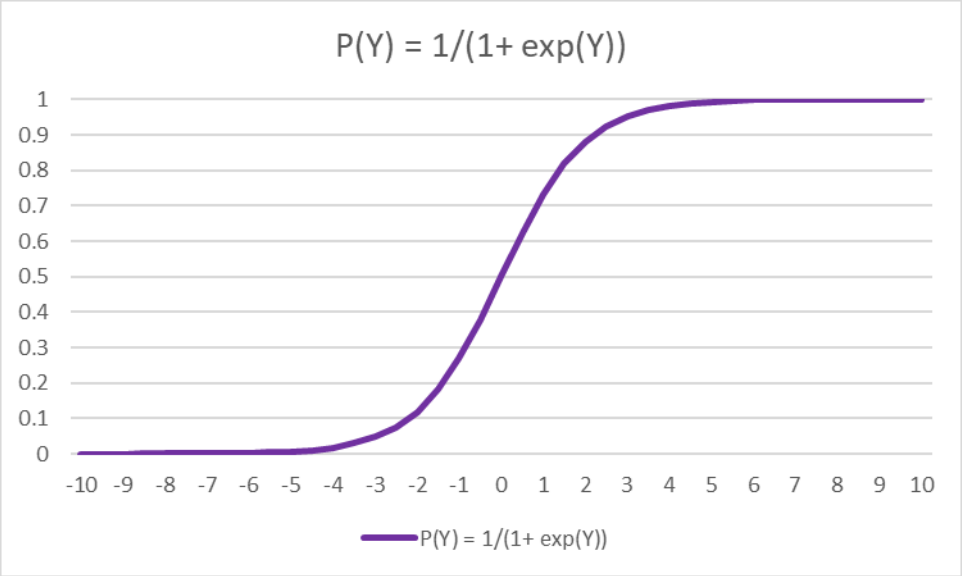
Los modelos de regresión no lineales *probit* y *logit* solucionan dicho problema transformando el valor *Y* que arroja el modelo lineal en un valor entre 0 y 1. Para hacer la transformación, utilizan funciones de distribución acumulada. El modelo *logit* utiliza la función de distribución acumulada logística, mientras que el modelo *probit* utiliza la función de distribución acumulada normal. Si bien ambas técnicas producen resultados similares (Bessis, 2015), como comentamos en el apartado anterior, nos vamos a detener en el modelo *logit*.

El modelo *logit* sigue expresando a la variable *Y* como función lineal de las variables independientes ($Y = \beta_0 + \beta' X + u$), pero en lugar de interpretar directamente a *Y* como valor de probabilidad, inserta la función *F(Y)* como argumento de la función de distribución acumulada logística. Por lo tanto, la probabilidad de *Y* queda expresada de la siguiente forma:

$$\mathrm{P(Y = 1) = 1/(1 + \exp\big(-(}\beta_0 + \beta' X + u\mathrm{)\big))} \qquad (7)$$

La función de distribución acumulada posee valores entre 0 y 1, por lo tanto, cada valor de la función *Y* (que dependerá de los valores que toman las variables independientes) tendrá asociado un valor entre 0 y 1, solucionando el problema que presentaba el modelo de regresión lineal.

Gráfico 2.3: Función logística



Fuente: Elaboración propia

Dado que la función de distribución es monótona creciente, cuanto menor sea *Y*, menor será la probabilidad acumulada, por lo tanto, menor será la probabilidad que pertenezca al grupo de individuos que cumplen con el pago de la deuda. Asimismo, cuanto mayor sea *Y*, mayor será la probabilidad acumulada, por lo tanto, mayor será la probabilidad que pertenezca al grupo de individuos que cumplen con el pago.

En el modelo *logit*, el método de optimización matemático utilizado para obtener estimadores de los parámetros del modelo *(β)* es el método de máxima verosimilitud. El método de máxima verosimilitud parte de los datos observados y busca maximizar la probabilidad que provengan de una función de forma conocida y parámetros desconocidos; los parámetros que maximicen dicha probabilidad serán los estimadores de máxima verosimilitud (Gourieroux y Jasiak, 2007).

Sean $(y_i, x_i), i = 1, \dots, n$ pares de observaciones independientes de default y características individuales y $f(Y, X, \beta)$ su función de distribución de forma conocida y parámetros desconocidos[14], el estimador de máxima verosimilitud está definido de la siguiente forma:

$$\hat{\beta} = \arg\max \sum_{i=1}^{n} \log f(y_i, x_i, \beta) \qquad (8)$$

[14] Es un modelo paramétrico ya que la forma funcional es conocida y los parámetros son desconocidos.

La función objetivo lleva el nombre de función de verosimilitud, si es diferenciable respecto al parámetro, el estimador de máxima verosimilitud debe satisfacer el siguiente sistema de condiciones de primer orden[15]:

$$\sum_{i=1}^{n} \frac{\partial \log f}{\partial \beta}(y_i, x_i, \hat{\beta}) = 0 \qquad (9)$$

Según Gourieroux y Jasiak (2007), las propiedades de los estimadores de máxima verosimilitud son las siguientes[16]:

1. Consistencia: el estimador tiende al verdadero valor del parámetro cuando el tamaño muestral tiende a infinito.
2. El estimador tiene una distribución asintótica normal multivariada.
3. El estimador es asintóticamente insesgado (su esperanza tiende al verdadero valor del parámetro cuando el tamaño muestral tiende a infinito).
4. El estimador es asintóticamente eficiente (su varianza es la mínima entre los estimadores asintóticamente insesgados cuando el tamaño muestral tiende a infinito).

En el caso del modelo *logit*, la función de densidad está dada por la distribución de Bernoulli[17]:

$$f(y_i, x_i, \beta) = \left[\frac{1}{1+\exp(-X'\beta)}\right]^{yi} \left[\frac{\exp(-X'\beta)}{1+\exp(-X'\beta)}\right]^{1-yi} \qquad (10)$$

El estimador de máxima verosimilitud del modelo *logit* debe satisfacer el siguiente sistema de condiciones de primer orden:

$$\sum_{i=1}^{n} \frac{\partial \log f}{\partial \beta}(y_i, x_i, \hat{\beta}) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i \left[y_i - \frac{1}{1+\exp\left(-X'\hat{\beta}\right)}\right] = 0 \qquad (11)$$

[15] Habrá tantas condiciones de primer orden como cantidad de parámetros.

[16] Deben cumplirse las condiciones de regularidad que, generalmente, se cumplen en las aplicaciones de riesgo de crédito.

[17] Recordar la función de probabilidad de Bernoulli: $f(x) = p^x(1-p)^{1-x}\ con\ x = \{1,0\}$.

Los residuos miden la diferencia entre el verdadero riesgo de cada individuo (y_i) y su riesgo estimado $(\frac{1}{1+\exp(-X'\hat{\beta})})$. Por lo tanto, estas condiciones de primer orden reflejan la ortogonalidad que debe existir entre las variables explicativas (x_i) y los residuos $\left(y_i - \frac{1}{1+\exp(-X'\hat{\beta})}\right)$. La ortogonalidad significa que los residuos y cada una de las variables explicativas no comparten información.

Con los parámetros estimados, la probabilidad que la variable Y sea igual a uno, es decir, la probabilidad que el individuo sea un buen pagador (si se define que 1 indica pago y 0 indica falta de pago) será la siguiente:

$$\mathrm{P(Y=1)} = 1/(1+\exp\left(-\left(X'\hat{\beta}\right)\right)) \qquad (12)$$

Esta probabilidad es el *score* que tendrá el individuo. Cuanto mayor sea el *score*, mayor será la probabilidad que el individuo pague la deuda tomada. El *score* es utilizado por el banco para generar un ranking de solicitudes de crédito y, mediante un *cut-off* elegido por el banco según su apetito al riesgo, separará a las solicitudes entre las aceptadas y las rechazadas.

2.3 Selección de variables explicativas y performance del modelo

La elección de las variables explicativas que se incorporarán al modelo econométrico se fundamenta, en primera instancia, en relaciones económicas. Luego, se puede refutar la incorporación de cada variable a través de análisis bivariados, análisis multivariados y de la performance general del modelo.

2.3.1 Análisis bivariados

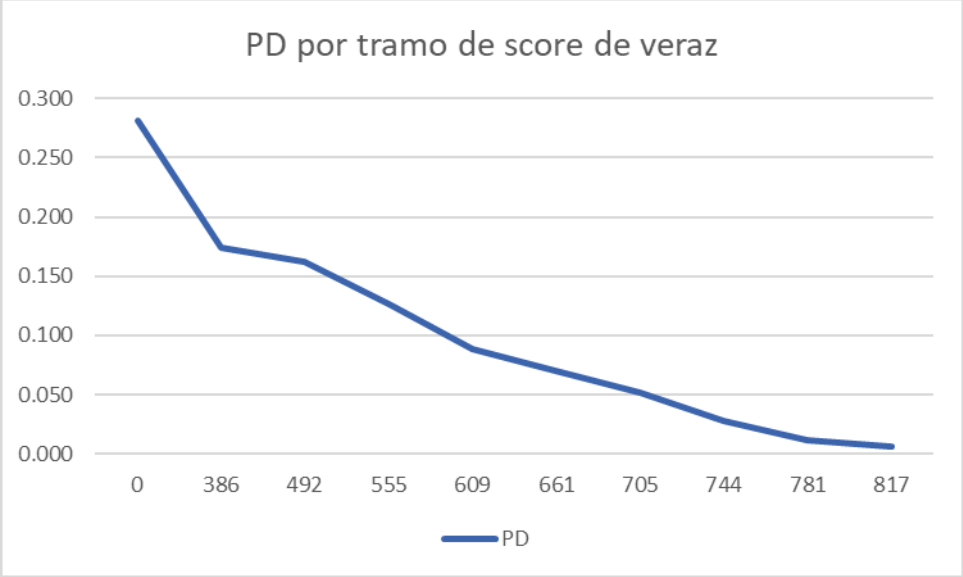
Para evaluar el poder de discriminación de las variables explicativas del modelo, se pueden realizar análisis bivariados. Uno de los análisis que se puede realizar consiste en separar la variable explicativa en rangos y evaluar la tasa de morosidad en cada uno de ellos. Es de esperar que una variable explicativa con buen poder de discriminación presente monotonicidad en la tasa de malos a medida que pasamos de un rango al siguiente. Por ejemplo, si se utiliza el *score* de veraz como variable explicativa para evaluar la probabilidad de pago de un préstamo, es de esperar que la tasa de malos por *bucket* de *score* sea menor a medida que aumenta el rango de *score* considerado. Cuanto

más difusa sea la tendencia observada en la tasa de malos a medida que nos movemos de un *bucket* al siguiente, menor poder de discriminación poseerá dicha variable.

A continuación, se presenta un cuadro y un gráfico de elaboración propia en los que se observa que la probabilidad de *default* a 12 meses disminuye a medida que aumenta el *score* de veraz[18].

Gráfico 2.4[19]**:** Análisis bivariado: PD por tramo de score veraz

Score veraz		Malos	Buenos	Total	PD
0	386	977	3815	4792	0.282
386	492	605	4160	4765	0.174
492	555	563	4304	4867	0.162
555	609	437	4289	4726	0.126
609	661	307	4469	4776	0.089
661	705	244	4526	4770	0.070
705	744	178	4570	4748	0.051
744	781	96	4703	4799	0.028
781	817	39	4767	4806	0.011
817	999	22	4701	4723	0.006



Fuente: Elaboración propia

Dado que la probabilidad de *default* disminuye a medida que aumenta el *score* de veraz, se puede deducir que el *score* de veraz es una buen predictor de la probabilidad de *default*.

[18] Los tramos de scores fueron separados en deciles.

[19] El gráfico fue realizado utilizando los datos del modelo de tarjetas, el cual será presentado en el capítulo siguiente.

Otro análisis bivariado que se puede realizar para observar la relación entre la variable explicativa y la variable objetivo es el cálculo del *weight of evidence* (WOE) y del *information value* (IV).

El WOE mide, para cada rango de valores de la variable explicativa, el desvío porcentual entre la distribución de buenos y la distribución de malos; de esta forma analiza el poder predictivo de cada rango de la variable explicativa en relación con la variable objetivo.

$$WOE = \ln\left(\frac{\%Buenos_i}{\%Malos_i}\right) \quad (13)$$

Para obtener el IV, en cada rango, se multiplica el WOE (que mide el desvío porcentual de las distribuciones) por la diferencia entre la distribución de buenos y malos (que mide la importancia entre las diferencias); luego, se realiza la sumatoria para todos los rangos. Dado que el IV analiza el poder predictivo total de la variable explicativa en relación con la variable objetivo, la medida puede utilizarse para comparar el poder predictivo con otras variables explicativas (Lin, 2013).

$$IV = \sum_i ((\%Buenos_i - \%Malos_i) * \ln\left(\frac{\%Buenos_i}{\%Malos_i}\right)) \quad (14)$$

La regla de oro para entender el poder predictivo de cada variable es la siguiente (Tibco, *s.f.*):

Information Value	Predictive Power
< 0.02	Useless
0.02 - 0.1	Weak
0.1 - 0.3	Medium
0.3 - 0.5	Strong
> 0.5	Suspiciously good; too good to be true

2.3.2 Análisis multivariado

El análisis multivariado es un procedimiento estadístico que consiste en analizar el aporte conjunto de varios factores para explicar un evento. Es decir, se analiza el poder explicativo de una variable teniendo en cuenta el aporte del resto de los regresores.

Para testear la significatividad de una variable dentro de un modelo *logit* se puede utilizar la prueba de Wald. La misma consiste en plantear como hipótesis nula que la variable explicativa no es significativa para explicar el evento (H_0: $\beta = 0$). Se rechaza H_0 cuando el estimador asociado a dicha variable explicativa ($\hat{\beta}$) es suficientemente distinto de cero.

Para realizar la prueba sobre *r* variables explicativas, se considera el siguiente estadístico de prueba[20]:

$$\xi w = \hat{\beta}'(\hat{V}\hat{\beta})^{-1}\hat{\beta} \qquad (15)$$

$\hat{V}\hat{\beta}$ es un estimador consistente de la matriz de varianzas y covarianzas de $\hat{\beta}$ y, bajo la hipótesis nula, ξw sigue una distribución asintótica chi-cuadrado con *r* grados de libertad (Gourieroux y Jasiak, 2007). La cantidad de grados de libertad es igual a la cantidad de variables que se incluyan en el análisis.

La conclusión de la prueba de Wald será:

$$\begin{cases} No\ rechazar\ H_0: \{\beta = 0\} & si\ \xi w\ <\ \chi_c^2(r) \\ Rechazar\ H_0: \{\beta = 0\} & si\ \xi w\ >\ \chi_c^2(r) \end{cases} \qquad (16)$$

Siendo *c* el nivel de confianza con el cual se realiza la prueba (en general, se realiza al 95% de confianza).

2.3.3 Performance del modelo

Los indicadores de *performance* de modelo que suelen utilizarse son el test Kolmogórov-Smirnov (KS), el coeficiente de Gini y el área debajo de la curva de la característica operativa del receptor (AUROC).

I. KS

La prueba de Kolmogórov-Smirnov (KS) es una prueba no paramétrica que determina la bondad de ajuste entre la distribución de buenos pagadores y la distribución de malos pagadores. Su valor indica la máxima diferencia que existe entre ambas distribuciones de

[20] Cuando se realiza una prueba sobre una sola variable explicativa, *r* es igual a uno. Si *r* es mayor a uno, se trata de una prueba para evaluar la significatividad de varias variables explicativas.

probabilidad acumuladas. Por lo tanto, es una medida del poder de discriminación del modelo.

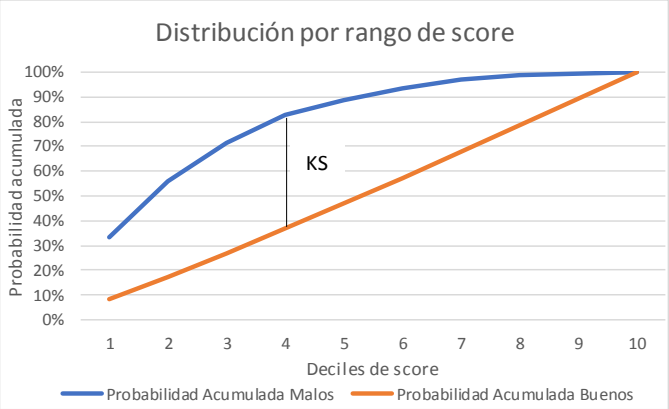
Para calcular el KS, se separa a la población en grupos de igual tamaño en función del *score* (*e.g.* deciles) y, para cada grupo, se obtiene la probabilidad acumulada de buenos pagadores y de malos pagadores. Luego, para cada grupo, se compara la diferencia absoluta entre ambas probabilidades acumuladas. El KS es la máxima diferencia absoluta encontrada.

$$KS = \max\{abs(prob\ acumulada\ Buenos - prob\ acumulada\ Malos)\}$$

A continuación, se presenta un gráfico de elaboración propia en el que se ilustra lo comentado.

Gráfico 2.5[21]**:** KS

Decil	Probabilidad Acumulada Malos	Probabilidad Acumulada Buenos	Diferencia absoluta
1	0.331	0.082	0.25
2	0.561	0.172	0.39
3	0.713	0.268	0.44
4	0.829	0.366	0.46
5	0.888	0.470	0.42
6	0.935	0.574	0.36
7	0.968	0.679	0.29
8	0.986	0.785	0.20
9	0.995	0.893	0.10
10	1.000	1.000	0.00
		KS	0.46



Fuente: Elaboración propia

II. Gini

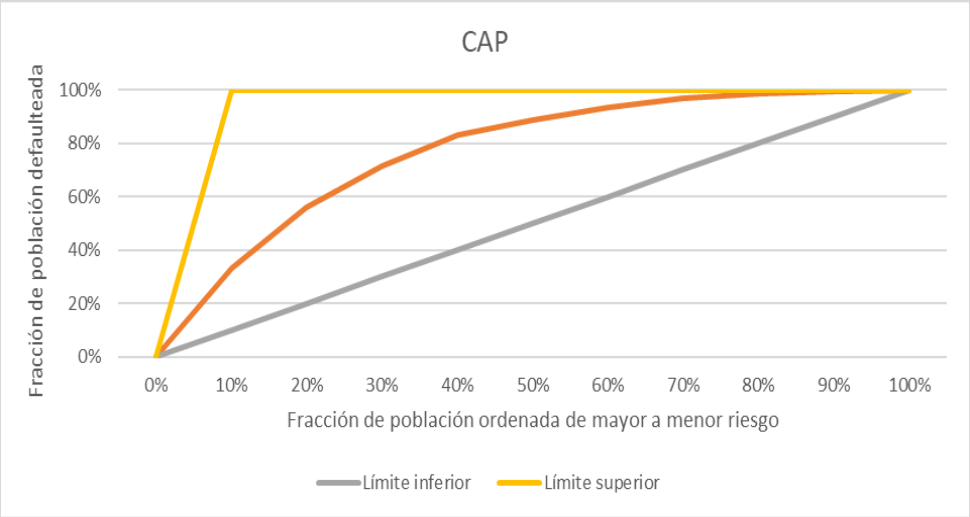
El Gini es un coeficiente que surge luego de los primeros trabajos de Lorenz (1905) para medir la desigualdad distributiva. Para visualizar la desigualdad, el autor presenta la Curva de Lorenz; la misma ordena la población por percentiles (de menor a mayor) en función de su ingreso relativo; el eje de las ordenadas mide el ingreso relativo acumulado de cada percentil. La igualdad perfecta ocurre en la línea de 45°donde cada percentil de la población posee el mismo ingreso.

[21] El gráfico fue realizado utilizando los datos del modelo de tarjetas, el cual será presentado en el capítulo siguiente.

Así como el coeficiente de Gini para medir desigualdad distributiva se mide a través de la Curva de Lorenz, el cálculo del Gini para medir el poder de discriminación de los modelos de *score* se realiza a partir del *cumulative accuracy profile* (CAP). El CAP grafica el porcentaje acumulado de la población en *default* para cada fracción de población ordenada de forma decreciente según su riesgo (de menor a mayor *score*). Si la tasa de *default* es de un 10%, un modelo de discriminación perfecta acumulará el 100% de la población en default en el primer decil de *score*. De esta forma, el límite superior del CAP se alcanza si el modelo discrimina perfectamente el riesgo. Por el contrario, si el *score* es independiente de la tasa de *default* (el modelo no discrimina riesgo), cada percentil de *score* tendrá la misma frecuencia relativa de default. De esta forma, el límite inferior del CAP está representado por la línea de 45°.

A continuación, se presenta un gráfico de elaboración propia en el cual se visualiza lo comentado.

Gráfico 2.6: CAP



Fuente: Elaboración propia

El coeficiente de Gini es el cociente entre dos áreas: el área desde el límite superior al límite inferior y el área desde el CAP hasta el límite inferior. Cuanto mejor sea la discriminación del modelo, mayor será el valor del Gini. El valor 0 indica que el modelo no discrimina el riesgo y el valor 1 indica discriminación perfecta; un coeficiente de Gini por encima de 0.6 se considera aceptable (Bessis, 2015).

El cálculo del coeficiente de Gini se lleva a cabo de diversas formas; una de las más extendidas es la fórmula de Brown (1994)[22]:

$$Gini = 1 - \sum_{i=0}^{k-1} (Y_{i+1} + Y_i)(X_{i+1} - X_i) \qquad (17)$$

Dónde *X* representa el porcentaje acumulado de *defaults*; *Y* representa el porcentaje acumulado de no defaults y *k* es la cantidad de intervalos.

En mi opinión, el Gini es una medida más completa que el KS ya que tiene en cuenta toda la distribución de probabilidad de buenos y malos pagadores y no solo la máxima diferencia entre ellas.

III. AUROC

El AUROC es una medida de performance que indica en qué porcentaje el modelo distingue correctamente entre buenos y malos pagadores. El AUROC es el área debajo de la curva ROC, la cual representa gráficamente los pares de puntos (sensibilidad;1-especificidad) que el modelo genera para cada punto de corte[23] (Fawcett, 2005).

La sensibilidad y la especificidad se pueden observar a partir de la matriz de confusión:

Matriz de confusión		
	no rechazado	rechazado
moroso	Falso positivo	Verdadero negativo
no moroso	Verdadero positivo	Falso negativo

La sensibilidad es la probabilidad de aceptar un crédito bueno (verdadero positivo/cantidad total de no morosos). El complemento de la sensibilidad es el error de tipo II: probabilidad de rechazar un crédito bueno (falso negativo/cantidad total de no morosos). La especificidad es la probabilidad de rechazar un crédito malo (verdadero

[22] La derivación matemática de la fórmula de Gini se puede ver en Derby (2003).

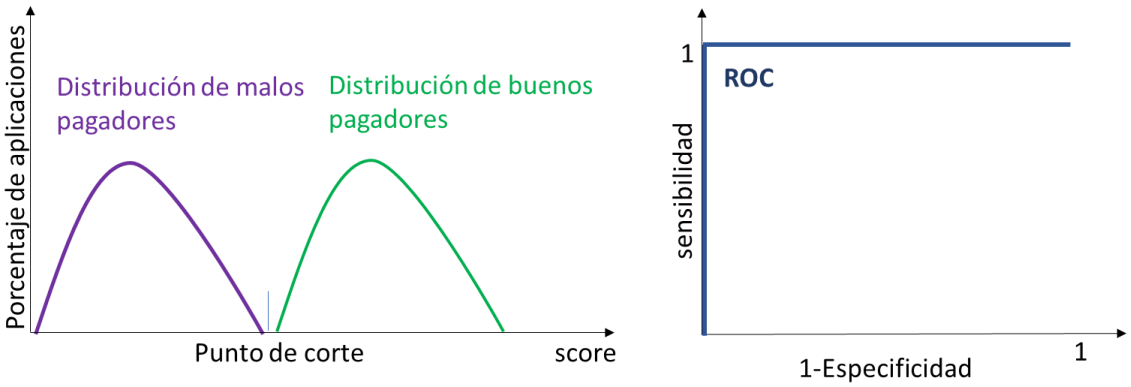
[23] El punto de corte es el score a partir del cual la entidad acepta una solicitud de crédito.

negativo/cantidad total de morosos). El complemento de la especificidad es el error de tipo I: probabilidad de aceptar un crédito malo (falso positivo/cantidad total de morosos).

El error de tipo I implica una pérdida de capital y de intereses para el banco, en cambio, el error de tipo II es un costo de oportunidad por no haber otorgado el crédito a un buen pagador (Bessis, 2015). El prestatario debe elegir un punto de corte que maximice su ganancia esperada, la cual depende de la ganancia por los créditos buenos otorgados, de la pérdida por los créditos malos otorgados y del costo de oportunidad de rechazar buenos pagadores[24].

Cuanto mayor sea el punto de corte que establezca el prestatario para aceptar el otorgamiento de un crédito, mayor será la probabilidad de cometer el error de tipo II y menor será la probabilidad de cometer el error de tipo I (y viceversa)[25]. Si el modelo de *score* permite separar completamente la distribución de buenos pagadores de la distribución de malos pagadores, el modelo de score es perfecto, ya que existirá un punto de corte para el cual la probabilidad de cometer ambos errores es cero. En este caso, el área debajo de la curva ROC es igual a 1. A continuación, presentamos un gráfico que ejemplifica lo comentado.

Gráfico 2.7: Curva ROC - modelo perfecto



Fuente: Elaboración propia

Aquí se observa que, aunque el modelo discrimine perfectamente entre buenos y malos pagadores, si se establece un punto de corte demasiado bajo, se estarán aceptando créditos

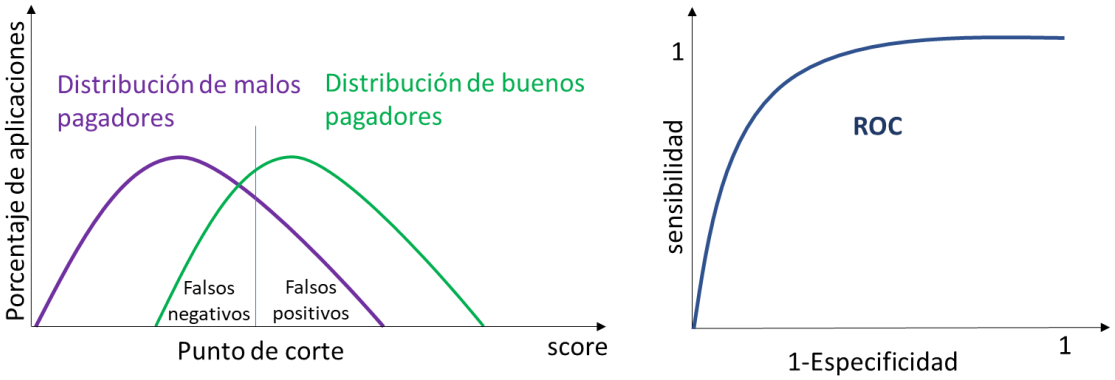
[24] Para más información sobre el cálculo del punto de corte óptimo, ver capítulo 2 de Gourieroux y Jasiak (2010).

[25] Esto ocurre en aquellos modelos de score que poseen un desempeño normal: la distribución de buenos pagadores se concentra en *scores* más altos y la distribución de malos pagadores se concentra en *scores* más bajos.

malos (1-especificidad > 0) y si se establece un punto de corte demasiado alto, se estarán dejando de aceptar algunos créditos buenos (sensibilidad < 1).

El error de tipo I y el error de tipo II no son cero en los modelos reales, por lo tanto, el área debajo de la curva ROC (AUROC) es menor a 1. A continuación, presentamos un gráfico que ejemplifica lo comentado.

Gráfico 2.8: Curva ROC – modelo normal



Fuente: Elaboración propia

En mi opinión, es una medida de performance más interesante que el Gini dado que no solo nos provee un valor para determinar cuán bueno es el modelo, sino que indica en qué medida el modelo distinguirá correctamente una solicitud de crédito.

Para calcular el AUROC, se puede utilizar la siguiente fórmula (Farris, 2010):

$$AUROC = \frac{Gini + 1}{2} \qquad (18)$$

Cuanto mayor es el AUROC, mayor es la probabilidad que el modelo clasifique correctamente una solicitud de crédito.

En el capítulo siguiente, se utilizarán bases de datos reales y se aplicará la metodología que hemos detallado en el presente capítulo.

Capítulo 3: Comparación de modelos de riesgo crediticio con incorporación de variables alternativas

El objetivo de este capítulo es comparar modelos de riesgo crediticio con incorporación de variables alternativas respecto a modelos de riesgo crediticio con información tradicional, realizando el contraste en diferentes productos financieros. Para ello, se segmentó el capítulo en tres partes. En primer lugar, se realizará un relevamiento de las bases input provistas por el banco anónimo con el objetivo de conocer las principales cualidades de la cartera, de definir la muestra que se utilizará para la construcción de los modelos y para determinar si es necesario realizar un ajuste sobre los datos input. En segundo lugar, se realizará la selección de las variables explicativas que incluiremos en cada modelo, se ejecutarán las regresiones logísticas correspondientes, se analizará la performance de los modelos y se realizarán validaciones *out of sample* de los modelos construidos. Finalmente, se compararán los resultados obtenidos bajo cada modelo para cada tipo de producto.

3.1 Fuentes de información

El trabajo de construcción de modelos de *score* tuvo una etapa inicial de análisis exploratorio de los datos input con el objetivo de conocer las características de la cartera y verificar la calidad del dato. Por lo tanto, en este apartado, se van a estudiar las distintas bases de información utilizadas en el análisis y las variables que las componen.

Las bases de datos utilizadas fueron provistas por un banco anónimo de Argentina. Se recibieron tres bases de datos con información de las características de nuevos préstamos otorgados a individuos entre enero y abril del 2018. Si bien el banco contaba con bases de datos para una ventana temporal más amplia, la información de las variables alternativas incluidas en las mismas solamente estaba disponible para el primer cuatrimestre del 2018 y, por ese motivo, se solicitaron las bases de información para la ventana temporal mencionada. La primera base agrupa solicitudes de tarjetas de crédito, la segunda base contiene solicitudes de paquetes bancarios[26] y la tercera base posee solicitudes de préstamos personales.

3.1.1 Descripción de variables

Base de tarjetas y paquetes

[26] El paquete bancario consiste en el otorgamiento de una cuenta corriente, de una tarjeta de crédito y de un préstamo pre acordado.

Las bases de tarjetas y paquetes poseen las mismas variables. A continuación, se presenta una tabla que las lista.

Tabla 3.1: Listado de variables de la base de tarjetas y paquetes

Variable	Tipo de variable	Característica
indicador_contrato	cuantitativa	Descriptiva
periodo	cuantitativa	Descriptiva
tipo_renta	cualitativa	Variables explicativas tradicionales
ingresoMensualOrdinario	cuantitativa	
marca_pyme	dummy	
marca_cliente_antig_hasta12m	dummy	
marca_antig_empleo_hasta24m	dummy	
cantidadTarjetaCredito	cuantitativa	
estadoCivil	cualitativa	
marcaPoseeAuto	dummy	
marcaPoseeCajaAhorro	dummy	
nivelEstudios	cualitativa	
tipoVivienda	cualitativa	
refCtaCorrienteYAhorro	cualitativa	
Grupo_edad	cualitativa	
cantidadConsultas	cuantitativa	
regulares_veraz	dummy	
scoreVeraz	cuantitativa	
INDICA_USA_HOME_BANKING	dummy	Variables explicativas alternativas
CANT_TRANSAC_HOME_BANKING	cuantitativa	
INDICA_USA_MOBILE	dummy	
CANTIDAD_TRANSACCIONES_MOBILE	cuantitativa	
no_default	dummy	Variable objetivo

Fuente: Elaboración propia

A continuación, se describen las variables siguiendo el orden en el que se presentan en la tabla anterior. Cabe aclarar que las variables cualitativas serán transformadas en variables *dummy* para poder ingresarlas dentro del modelo[27].

- Indicador del contrato: es el número de identificación con el cual el banco identifica a cada contrato.[28]
- Período: Es el período de otorgamiento del préstamo (toma valores de enero a abril del 2018).

[27] Si *m* es la cantidad de valores que puede tomar la variable cualitativa, se crean *m-1* variables *dummy*. Otra forma de ingresar variables cualitativas dentro del modelo *logit* es asignándole a cada valor posible de la variable cualitativa su *weight of evidence*..

[28] Para mantener el anonimato de los clientes del banco, se modificó la variable por números enteros aleatorios.

- Tipo de renta: indica el segmento de renta al cual pertenece el cliente que solicita el préstamo. Hay cuatro segmentos posibles: hasta $30.000 de ingresos, de $30.000 a $50.000, de $50.000 a $80.000 y desde $80.000.
- Ingreso mensual ordinario: es el ingreso mensual del cliente. Si bien, *a priori*, puede parecer multicolineal con la variable anterior, el banco otorga beneficios especiales para cada tipo de renta, por lo tanto, la información que presentan ambas variables es distinta.
- Marca pyme: indica si el préstamo solicitado es para el comercio del individuo o no.
- Marca antigüedad cliente hasta doce meses: indica con valor uno a los préstamos otorgados a clientes con antigüedad menor o igual a doce meses en la entidad bancaria. Es de esperar que los clientes antiguos tengan un mejor comportamiento; como menciona FICO (*s.f.*) en su página web, un mayor historial crediticio es señal de menor riesgo.
- Marca antigüedad empleo hasta 24 meses: indica con valor uno a los préstamos otorgados a clientes con antigüedad menor o igual a 24 meses en su empleo. Es esperable que los clientes con mayor antigüedad en su empleo tengan mejor desempeño crediticio.
- Cantidad de tarjetas de crédito: indica la cantidad de tarjetas de crédito que posee el cliente. Es de esperar que cuantas más tarjetas de crédito tenga el individuo, mejor será su probabilidad de pago; la página web de Equifax (*s.f.*) señala que la cantidad de productos crediticios que posee el individuo es un indicador relevante para los prestadores.
- Estado civil: indica con valor "S" a los clientes solteros, con valor "M" a los clientes casados, con valor "D" a los divorciados y con valor "W" a los viudos.
- Marca automóvil: identifica a los clientes que poseen automóvil al momento de la solicitud del préstamo.
- Marca caja de ahorro: indica si el cliente posee una caja de ahorro o no al momento de la solicitud.
- Nivel estudios: indica el grado de educación que posee el solicitante del préstamo al momento de la solicitud. Los valores que toma la variable son: A (sin educación), B (primario completo), C (secundario completo), D (terciario completo), E (universitario completo) y F (posgrado completo). La educación que

poseen los individuos puede ser una variable significativa para explicar comportamiento crediticio. En Gasparini y Cicowiez (2007) se realizan estudios que demuestran que los trabajadores con mayor nivel educativo poseen un ingreso mayor que los menos educados y que esta brecha se está incrementando con el paso del tiempo.

- Tipo de vivienda: identifica el tipo de vivienda del individuo al momento de la solicitud. Los valores que toma la variable son: H (propia), R (alquilada), P (vive con su familia) o M (ninguna de las anteriores).
- Referencia cuenta corriente y caja de ahorro: indica si el cliente posee cuenta corriente y/o caja de ahorro. Los valores que puede tomar la variable son: A (no posee ninguna), B (posee cuenta corriente solamente), C (posee caja de ahorro solamente) y D (posee caja de ahorro y cuenta corriente). Dado que existe en la base una variable que indica si el cliente posee caja de ahorro, va a existir multicolinealidad entre ambas variables y se excluirá esta última de los modelos.
- Grupo de edad: indica el segmento etario al cual pertenece el cliente. La variable puede tomar los siguientes valores: "Menor de 40 años", "Entre 40 y 60 años" y "Mayor a 60 años".
- Cantidad de consultas veraz: indica la cantidad de veces que se pidió un informe de veraz[29] del cliente.
- Regular veraz: indica con valor uno a los clientes bancarizados y sin antecedentes negativos.
- Score de veraz: es el score de bureau generado por la empresa Equifax[30]. El rango de valores que puede tomar la variable es de 1 a 999. Es de esperar que cuanto mayor sea el score veraz, mejor sea el desempeño crediticio del cliente.
- Indicador de uso de *home banking*: indica si el cliente utiliza o no la página web de la entidad bancaria. Es una variable alternativa ya que no pertenece al conjunto de variables que tradicionalmente utilizan los bancos para armar los *scores* (al igual que las tres variables que se mencionan a continuación).
- Cantidad de transacciones por *home banking*: indica la cantidad de transacciones que el cliente realizó mediante la página web del banco en el mes de la solicitud del nuevo préstamo.

[29] El informe de veraz contiene información del historial crediticio del cliente en el sistema financiero.

[30] En la sección 1.3 se presenta el detalle de las principales variables con las cuales se conforma el *score*.

- Indicador de uso de *mobile banking*: indica si el cliente utiliza o no la aplicación del banco en su celular.
- Cantidad de transacciones por *mobile banking*: indica la cantidad de transacciones que el cliente realizó mediante la aplicación móvil del banco en el mes de la solicitud del nuevo préstamo.
- Indicador de buen desempeño (variable objetivo): identifica con valor uno a los clientes que no tuvieron retraso de más de 90 días en los doce meses posteriores al otorgamiento del crédito y con valor cero al caso contrario.

Base de préstamos personales

A continuación, se listan las variables que contiene la base de préstamos personales.

Tabla 3.2: Listado de variables de la base de préstamos personales

Variable	Tipo de variable	Característica
identificador_contrato	cuantitativa	descriptiva
periodo	cuantitativa	descriptiva
marca_garantia	dummy	Variables explicativas tradicionales
pctFinanciacion	cuantitativa	
cuotaPrestamo	cuantitativa	
montoPrestamo	cuantitativa	
plazoPrestamo	cuantitativa	
relacionCuotaIngreso	cuantitativa	
tipo_renta	cualitativa	
ingresoMensualOrdinario	cuantitativa	
marca_pyme	dummy	
marca_cliente_antig_hasta12m	dummy	
marca_antig_empleo_hasta24m	dummy	
cantidadTarjetaCredito	cuantitativa	
estadoCivil	cualitativa	
marcaPoseeAuto	dummy	
marcaPoseeCajaAhorro	dummy	
nivelEstudios	cualitativa	
cantidadPersonasACargo	cuantitativa	
tipoVivienda	cualitativa	
refCtaCorrienteYAhorro	cualitativa	
grupo_edad	cualitativa	
regulares_veraz	dummy	
cantidadConsultas	cuantitativa	
scoreVeraz	cuantitativa	
INDICA_USA_HOME_BANKING	dummy	Variables explicativas alternativas
CANT_TRANSAC_HOME_BANKING	cuantitativa	
INDICA_USA_MOBILE	dummy	
CANTIDAD_TRANSACCIONES_MOBILE	cuantitativa	
no_default	dummy	Variable objetivo

Fuente: Elaboración propia

A continuación, se describen aquellas variables pertenecientes a la base de préstamos personales que no forman parte de las bases de tarjetas y paquetes.

- Marca garantía: indica si el préstamo otorgado por el banco está asegurado por una garantía.
- Porcentaje de financiación: es el llamado *loan to value*. Es el porcentaje que representa el préstamo solicitado respecto al bien total que se desea adquirir con dicho préstamo.
- Cuota del préstamo: es el valor de la primera cuota mensual que el cliente deberá pagar.
- Monto del préstamo: es el valor del préstamo solicitado.
- Plazo del préstamo: es la cantidad de meses que dura el contrato que realizó el cliente con la entidad bancaria.
- Relación cuota-ingreso: es el porcentaje que representa la primera cuota mensual respecto al ingreso mensual del cliente al momento de la solicitud.
- Cantidad de personas a cargo: es la cantidad de personas que dependen económicamente del cliente.

3.1.2 Análisis de datos

En este apartado comentaremos sobre las principales características de las bases input. Para mayor detalle, en el anexo se encuentra el análisis de estadística descriptiva para cada variable que forma parte de las bases de tarjetas, paquetes y préstamos personales.

Antes de proceder al detalle de las bases, es menester aclarar que, dado que los datos disponibles abarcan la ventana temporal de enero a abril 2018 (con observación de desempeño durante los doce meses posteriores), se seleccionará la ventana muestral de enero a marzo para el desarrollo de los modelos y la ventana de abril para la validación *out of sample* de los mismos.

Base de tarjetas

La base posee un total de 59.715 de tarjetas otorgadas entre enero y abril de 2018, el 7,3% (4.378 tarjetas) cae en *default* en alguno de los siguientes doce meses. Se encontraron valores faltantes (*missings*) solamente en tres variables y con un porcentaje relativo bajo: 2,3% en la variable *ingresoMensualOrdinario*, 0,02% en la variable *CantidadConsultas* y 2,35% en la variable *scoreVeraz*. El tratamiento realizado sobre los valores faltantes

consistió en reemplazar dichos valores por los valores medios encontrados para cada variable en la base de tarjetas.

Respecto a la variable *marca_pyme*, se encontró que solamente el 0,1% posee valor uno; dada su baja volumetría, posiblemente no sea una variable discriminante en el modelo de tarjetas.

Se observa que la renta media de los clientes pertenecientes a la base de tarjetas es de $31.043, su score veraz promedio es de 634 y la cantidad de tarjetas de crédito promedio que poseen los clientes al momento de la solicitud es de 1.21. La mayor parte de la población es soltera y menor a 40 años. El 49,1% tiene el secundario completo y el 48,6% tiene estudios superiores. El 84% de los solicitantes posee una antigüedad mayor a un año en el banco y el 61,6% tiene una antigüedad en el empleo menor o igual a dos años.

En relación con las variables alternativas, la proporción de clientes que usa *home banking* es mayor a la proporción que utiliza *mobile banking* (68,6% vs 50,7%); los clientes realizan un promedio mensual de 31,33 transacciones a través de la página web y 38,45 mediante la aplicación móvil.

Base de paquetes

La base posee un total de 54.144 de paquetes otorgados entre enero y abril de 2018, el 12,9% (6.976 paquetes) cae en *default* en los siguientes doce meses.

Se observan valores faltantes en las mismas tres variables que para el caso de tarjetas y con un porcentaje relativo también bajo; se aplicó el mismo tratamiento que para la base de tarjetas (reemplazar los valores faltantes por lo valores medios de cada variable en la base de paquetes). Respecto a la variable *marca_pyme*, el porcentaje de registros con valor igual a uno no es insignificante como en el caso de tarjetas, por lo cual, puede ser una variable útil para discernir riesgo.

El ingreso mensual promedio en la base de paquetes es de $29.902, el score veraz promedio es de 622 y la cantidad de tarjetas promedio es de 1.15 (valores por debajo de aquellos encontrados en la base de tarjetas). Si bien la mayor parte de la población es menor a 40 años, el porcentaje de población mayor a 40 años es más alto que en el caso de tarjetas. Asimismo, hay un mayor porcentaje de población con educación superior (50,8%).

Respecto a la variable de antigüedad en el banco, se observa que la mayor parte tiene una antigüedad inferior al año; lo cual es contrario a lo que sucede en la base de tarjetas. Esta casuística se considera razonable dado que es lógico que los clientes más nuevos soliciten paquetes y los clientes con mayor antigüedad quieran añadir una tarjeta a sus productos existentes.

En relación con las variables alternativas, el uso de la página web y de la aplicación móvil es menor que el encontrado en la base de tarjetas: el 57,5% utiliza *home banking* (con un promedio mensual de transacciones de 16) y el 30.1% utiliza la aplicación móvil (con un promedio mensual de transacciones de 18,32%).

Base de préstamos personales

La base posee un total de 12.020 de préstamos personales otorgados entre enero y abril de 2018, el 8,1% (968 paquetes) cae en *default* en los siguientes doce meses.

Se encontraron valores faltantes en las siguientes variables: *ingresoMensualOrdinario, CantidadConsultas, scoreVeraz, pctFinanciacion* y *cuotaPrestamo* con un porcentaje promedio de faltantes de 3.03%. Se reemplazan dichos valores por los valores medios encontrados en cada variable. La variable *marca_pyme* posee un porcentaje de valores iguales a uno superior al resto de las bases.

El ingreso mensual promedio de la base de préstamos personales es significativamente superior al encontrado en las otras bases; lo mismo sucede con la cantidad media de tarjetas y con el score veraz medio. Asimismo, la edad media de la población es mayor que la encontrada en las otras bases.

La mayor parte de la población posee una antigüedad en el banco mayor al año, una antigüedad en el empleo mayor a los dos años y presenta una garantía para respaldar el préstamo.

Los resultados encontrados se consideran razonables, dado que muestran que el banco prefiere otorgar préstamos personales a clientes con mayores ingresos, con antigüedad en el banco y en su trabajo, con un score veraz alto y con garantía que respalde los mismos.

Respecto a las variables alternativas: el 57% utiliza *home banking* (con un promedio mensual de transacciones de 29) y el 35% utiliza la aplicación móvil (con un promedio mensual de transacciones de 28%).

En el apartado siguiente, se va a construir un modelo de score para cada tipo de producto y se analizará su performance. Para ello, vamos a partir de las variables iniciales que recién comentamos y se seleccionarán aquellas que expliquen el comportamiento de la variable objetivo.

3.2 Construcción de modelos de score

El propósito de este apartado es la construcción de modelos de score para cada producto (tarjetas, paquetes y préstamos personales). Para ello, se seleccionarán las variables explicativas que formarán parte del modelo no restringido (modelo que incluye las variables alternativas); asimismo, se elegirán las variables explicativas que formarán parte del modelo restringido (el cual excluye las variables alternativas) y se elaborarán los indicadores de performance.

Para seleccionar los predictores que formen parte de los modelos de score vamos a utilizar una metodología de selección iterativa. La misma consiste en descartar las variables explicativas que no superan la prueba de significatividad individual considerando un nivel de confianza del 95%, de forma tal que el modelo final quede explicado solamente por las variables estadísticamente significativas. Como comentamos previamente, se seleccionaron los datos de enero a marzo 2018 para el desarrollo de los modelos de *score* y seleccionaron los datos de abril 2018 para la validación *out of the sample* de estos.

Para evaluar el desempeño de los modelos, se utilizarán los indicadores de performance descriptos en el capítulo anterior. Los mismos son: la probabilidad de error de tipo 1, la probabilidad de error de tipo 2[31], el KS, el Gini y el AUROC.

A continuación, presentaremos la construcción de los modelos e indicadores de performance para cada una de las bases. Cabe aclarar que los códigos SAS y la documentación técnica del proceso se encuentra en la parte B del anexo.

3.2.1 Modelo de tarjetas

El objetivo de este apartado es construir un modelo completo, que tenga en cuenta todas las variables explicativas y, a la vez, construir un modelo restringido, el cual no contemple las variables alternativas.

[31] El punto de corte que se definió en todos los modelos para determinar el error de tipo 1 y el error de tipo 2 fue la cantidad de clientes no morosos respecto al total de la muestra. En las entidades bancarias el punto de corte depende del apetito al riesgo de cada sector.

- Modelo de tarjetas completo

En primer lugar, se seleccionaron todas las variables que forman parte de la base (a excepción de las variables descriptivas), incluyendo las variables alternativas y se realizó la prueba de significatividad individual para todas ellas. A continuación, se presenta el resultado que arrojó la primera prueba de significatividad individual.

Tabla 3.3: Test de significatividad individual – Base Tarjetas

Effect	Pr > ChiSq
ingresoMensualOrdina	<.0001
marca_pyme	0.3885
marca_cliente_antig_	<.0001
marca_antig_empleo_h	<.0001
cantidadTarjetaCredi	0.0005
marcaPoseeAuto	0.1677
cantidadConsultas	<.0001
regulares_veraz	<.0001
scoreVeraz	<.0001
INDICA_USA_HOME_BANK	<.0001
CANT_TRANSAC_HOME_BA	<.0001
INDICA_USA_MOBILE	<.0001
CANTIDAD_TRANSACCION	<.0001
tipo_renta	0.0161
estadoCivil	0.0096
nivelEstudios	0.0001
tipoVivienda	<.0001
refCtaCorrienteYAhor	<.0001
Grupo_edad	0.0178

Fuente: Salida de SAS Studio

La variable de marca_pyme, tal como anticipamos, no muestra poder de discriminación, por lo tanto, será eliminada. Lo mismo sucede con la variable que indica si el cliente posee automóvil.

En segundo lugar, se seleccionaron solamente las variables explicativas cuyo p-valor no superaba el 5% y se volvió a realizar la prueba de significatividad individual para cada una de ellas. Dado que en esta segunda oportunidad todas las variables superaron dicha prueba, se realizó la regresión logística utilizando estos regresores.

En tercer lugar, se ejecutó la regresión logística en SAS Studio con las variables explicativas mencionadas. A continuación, se presentan los estimadores puntuales obtenidos para cada variable.

Tabla 3.4: Variables explicativas del modelo final de tarjetas y sus estimadores

Parameter		Estimate
Intercept		1.211
ingresoMensualOrdina		0.00002
marca_cliente_antig_		-0.4297
marca_antig_empleo_h		-0.1572
cantidadTarjetaCredi		0.0931
cantidadConsultas		-0.1715
regulares_veraz		0.2477
scoreVeraz		0.00348
INDICA_USA_HOME_BANK		0.1971
CANT_TRANSAC_HOME_BA		0.00375
INDICA_USA_MOBILE		0.3872
CANTIDAD_TRANSACCION		-0.0018
tipo_renta	Desde 30.000 hasta 50.000	0.2804
tipo_renta	Desde 50.000 hasta 80.000	0.2529
tipo_renta	Desde 80.000	0.000946
tipo_renta	Hasta 30.000	0
estadoCivil	D	-0.2454
estadoCivil	M	-0.1095
estadoCivil	S	-0.4811
estadoCivil	W	0
nivelEstudios	A	-0.4715
nivelEstudios	B	-0.7635
nivelEstudios	C	-0.4053
nivelEstudios	D	-0.2509
nivelEstudios	E	-0.3035
nivelEstudios	F	0
tipoVivienda	H	-0.5646
tipoVivienda	M	-0.3545
tipoVivienda	P	-0.0335
tipoVivienda	R	0
refCtaCorrienteYAhor	A	-0.7603
refCtaCorrienteYAhor	B	-0.6775
refCtaCorrienteYAhor	C	-0.4135
refCtaCorrienteYAhor	D	0
Grupo_edad	Desde 60	0.5947
Grupo_edad	Entre 40 y 60	0.0705
Grupo_edad	Menor a 40	0

Fuente: Salida de SAS Studio

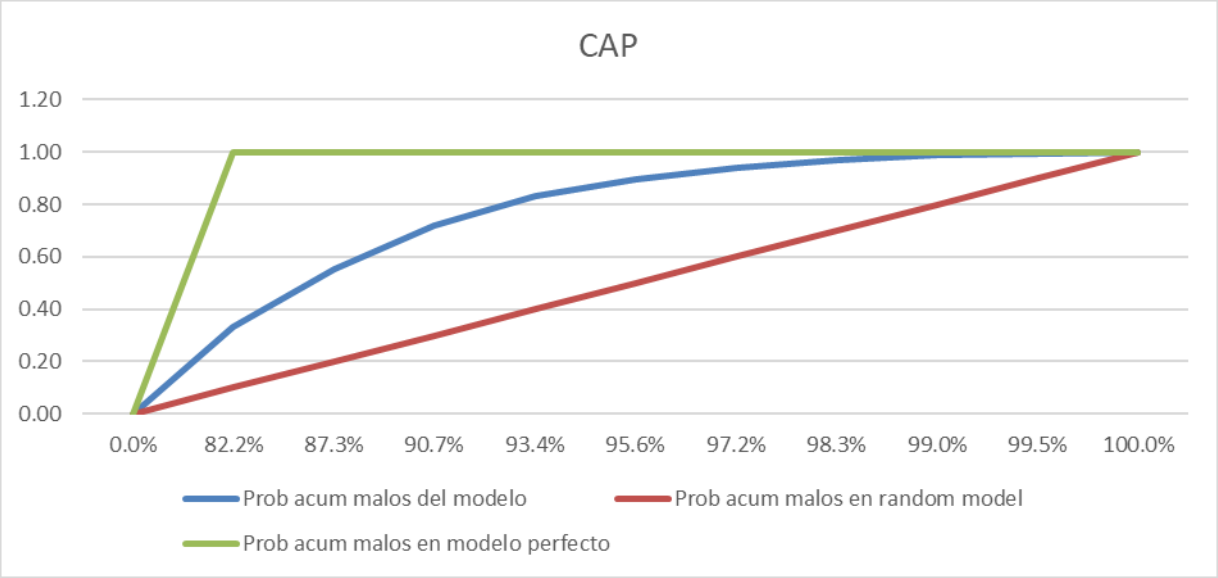
Se observa que los estimadores puntuales de las variables indicadoras de uso *de home banking* y de uso de *mobile banking* son mayores a cero, lo cual indica que el uso de dichas plataformas es señal de un buen comportamiento de pago del producto de tarjetas (recordemos que la variable objetivo toma valor uno si el cliente no entró en *default* en los siguientes doce meses del otorgamiento del crédito y cero en caso contrario).

Cabe aclarar que el estimador puntual igual a cero observado en ciertos valores de las variables cualitativas significa que dicho valor de la variable cualitativa quedó incorporado en el intercepto (recordar que por cada variable cualitativa con *k* posibles valores, se generaron *k-1* variables *dummy*).

En cuarto lugar, se realizaron las pruebas de performance sobre el modelo construido. A continuación, se presentan los resultados de las mismas junto con el CAP (*cumulative accuracy profile*), el cual grafica la probabilidad acumulada de malos por rango de *score*[32].

- Error de tipo 1: 33,9%
- Error de tipo 2: 19,3%
- KS: 46,4%
- Gini: 61,5%
- AUROC: 80,8%

Gráfico 3.1: CAP del modelo de tarjetas no restringido



Fuente: Elaboración propia

Finalmente, se realizó una validación *out of the sample* a partir de los datos de abril 2018. El proceso consistió en utilizar las variables explicativas y los estimadores puntuales del modelo de desarrollo y aplicarlos a las solicitudes de abril 2018 con el objetivo de evaluar la precisión del modelo en una población distinta a la utilizada para construirlo. Los indicadores de performance *out of the sample* muestran resultados similares a aquellos obtenidos para el desarrollo, lo cual le agrega robustez al modelo presentado[33].

[32] Los indicadores de performance fueron calculados de acuerdo a la metodología detallada en el capítulo anterior.

[33] Los indicadores de performance de la validación del modelo no restringido son los siguientes: error de tipo 1: 36,8% - error de tipo 2: 21,4%, KS: 42,1%, Gini 58%, AUROC: 79%.

- Modelo de tarjetas restringido

Para obtener el modelo final de tarjetas sin variables alternativas (modelo restringido), se realizó el mismo procedimiento que describimos anteriormente, pero, partiendo de todas las variables excepto las alternativas. Luego de los sucesivos filtros, se obtuvo un set de variables explicativas finales y se ejecutó la regresión logística. En el cuadro A.6 del anexo se presentan los estimadores puntuales de las variables explicativas que conforman el modelo restringido de tarjetas.

A continuación, se presentan los indicadores de performance del modelo final restringido:

- Error de tipo 1: 34,6%
- Error de tipo 2: 19,5%
- KS: 45,9%
- Gini: 60,1%
- AUROC: 80,0%

Al igual que para el modelo completo, se validó el modelo de tarjetas restringido con los datos del mes de abril 2018 y se obtuvieron indicadores de performance similares a aquellos obtenidos en el modelo construido[34].

De esta forma, concluimos con el trabajo sobre el producto de tarjetas y pasamos a detallar el procedimiento y los resultados obtenidos para el producto de paquetes.

3.2.2 Modelo de paquetes

En este apartado, se construirá un modelo de *score* para el producto paquetes partiendo de todas las variables explicativas y un modelo restringido, el cual no contempla estas últimas.

- Modelo de paquetes completo

Al igual que para el modelo de tarjetas, el primer paso consistió en seleccionar todas las variables que forman parte de la base (a excepción de las variables descriptivas). A cada variable se le realizó la prueba de significatividad individual con el objetivo de determinar

[34] Los indicadores de performance de la validación del modelo restringido son los siguientes: error de tipo 1: 36,9% - error de tipo 2: 19,5%, KS: 43,7%, Gini 56,8%, AUROC: 78,4%.

cuáles de ellas son posibles candidatas del modelo final. A continuación, se presenta el resultado que arrojó la primera prueba de significatividad individual.

Tabla 3.5: Test de significatividad individual – Base Paquetes

Effect	Pr > ChiSq
marca_pyme	<.0001
marca_cliente_antig_	<.0001
marca_antig_empleo_h	0.0387
cantidadTarjetaCredi	<.0001
ingresoMensualOrdina	0.363
marcaPoseeAuto	0.002
cantidadConsultas	<.0001
regulares_veraz	0.2444
scoreVeraz	<.0001
INDICA_USA_HOME_BANK	<.0001
CANT_TRANSAC_HOME_BA	<.0001
INDICA_USA_MOBILE	<.0001
CANTIDAD_TRANSACCION	<.0001
estadoCivil	0.0802
tipo_renta	<.0001
nivelEstudios	0.003
tipoVivienda	0.0069
refCtaCorrienteYAhor	<.0001
Grupo_edad	0.0003

Fuente: Salida de SAS Studio

Se observa que tres de las variables seleccionadas no superan dicha prueba, pero, a vez, todas las variables alternativas la superaron, por lo que son candidatas del modelo final.

Como segundo paso, se seleccionaron solamente las variables con p valor menor al 5% y se volvieron a realizar las pruebas de significatividad individuales. Todas las variables superaron las mismas, por lo que formarán parte del modelo final de paquetes.

Como tercer paso, se ejecutó la regresión logística para las variables seleccionadas; obteniendo los siguientes estimadores puntuales para cada una de ellas:

Tabla 3.6: Variables explicativas del modelo final de paquetes y sus estimadores

Parameter		Estimate
Intercept		0.681
marca_pyme		-0.4162
marca_cliente_antig_		-0.5383
marca_antig_empleo_h		0.0737
cantidadTarjetaCredi		0.0634
marcaPoseeAuto		-0.1156
cantidadConsultas		-0.1917
scoreVeraz		0.0035
INDICA_USA_HOME_BANK		0.2185
CANT_TRANSAC_HOME_BA		0.00505
INDICA_USA_MOBILE		0.439
CANTIDAD_TRANSACCION		-0.0024
tipo_renta	Entre 30.000 y 50.000	0.2382
tipo_renta	Entre 50.000 y 80.000	0.282
tipo_renta	Mayor a 80.000	0.5377
tipo_renta	Menor a 30.000	0
nivelEstudios	A	-0.0282
nivelEstudios	B	-0.3061
nivelEstudios	C	0.0311
nivelEstudios	D	0.043
nivelEstudios	E	0.139
nivelEstudios	F	0
tipoVivienda	H	-0.4476
tipoVivienda	M	-0.3554
tipoVivienda	P	-0.2975
tipoVivienda	R	0
refCtaCorrienteYAhor	A	-0.4097
refCtaCorrienteYAhor	B	-0.4884
refCtaCorrienteYAhor	C	-0.2008
refCtaCorrienteYAhor	D	0
Grupo_edad	Desde 60	0.2629
Grupo_edad	Entre 40 y 60	-0.0196
Grupo_edad	Menor a 40	0

Fuente: Salida SAS Studio

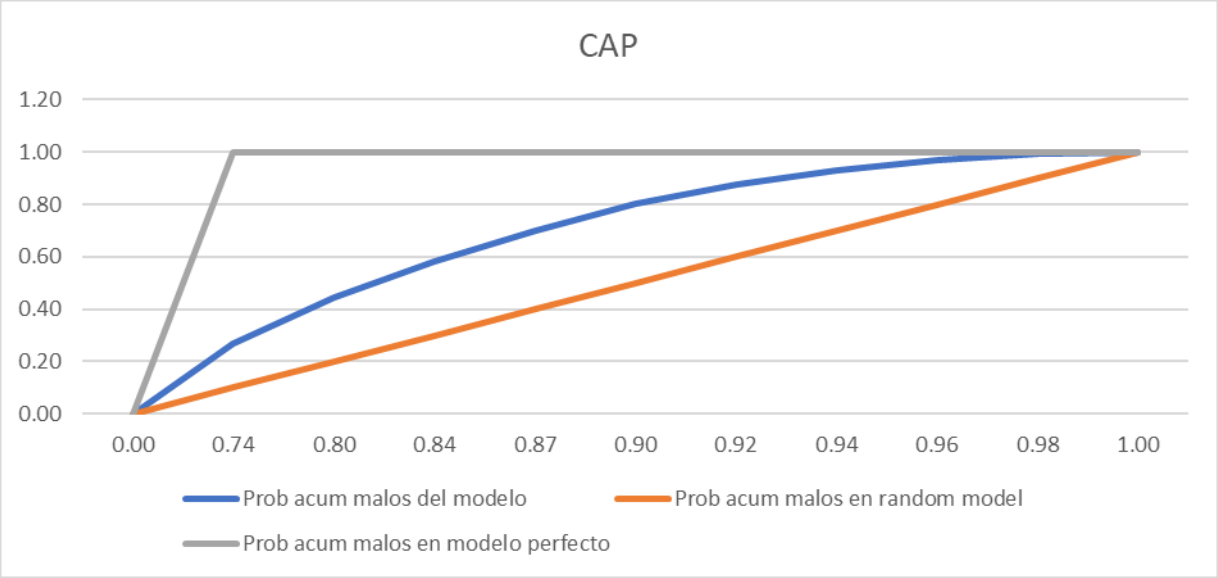
Al igual que para el modelo completo de tarjetas, se observa que las variables indicadoras de uso de *home banking* y de *mobile banking* poseen estimadores puntuales mayores a cero, indicando que, según el modelo, se espera un aumento de la probabilidad de pago si los solicitantes utilizan dichas plataformas. Asimismo, se observa que el grupo de mayores ingresos (renta superior a 80.000 ARS) posee un estimador puntual superior al resto de grupos de renta y que el score veraz posee un estimador puntual positivo; resultados coherentes con lo esperado para dichas variables.

En cuarto lugar, se realizaron las pruebas de performance sobre el modelo de paquetes y el CAP; los cuales se presentan a continuación:

- Error de tipo 1: 36,3%
- Error de tipo 2: 29,2%
- KS: 34,5%
- Gini: 49,3%

- AUROC: 74,6%

Gráfico 3.2: CAP del modelo de paquetes no restringido



Fuente: Elaboración propia

Finalmente, se realizó la validación *out of the sample* con los datos de abril 2018, observando indicadores de performance del modelo similares a los del desarrollo (el KS y la especificidad fueron incluso superiores en la validación)[35].

- Modelo de paquetes restringido

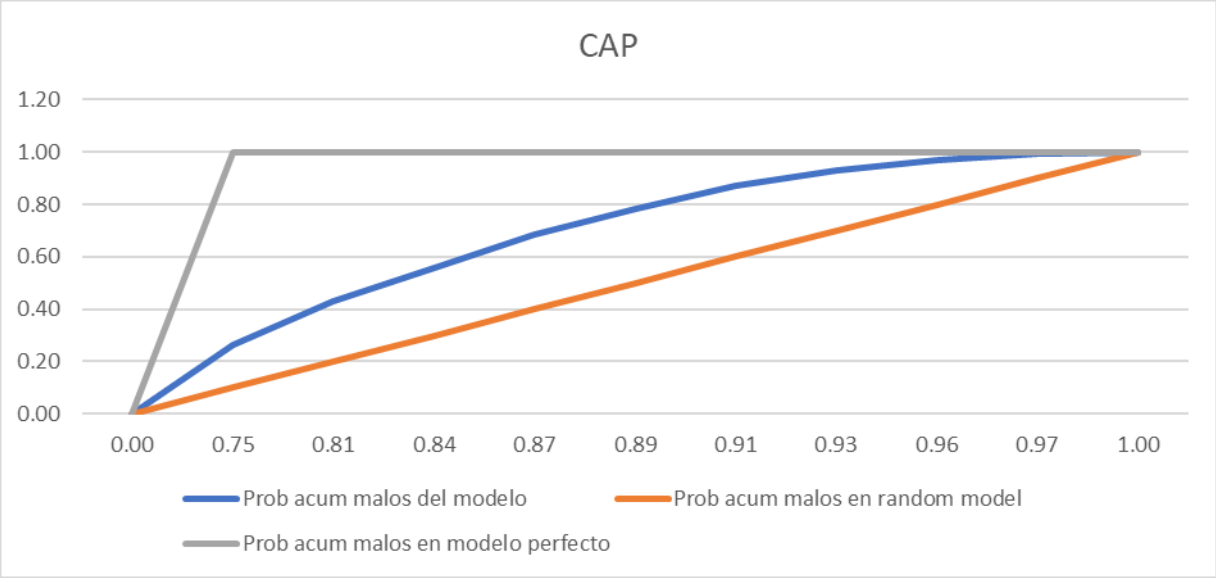
Para construir el modelo restringido, se siguió el mismo proceso descripto anteriormente: se realizaron sucesivas pruebas de significancia sobre las variables iniciales (sin considerar las alternativas) hasta conseguir el conjunto de variables que superaron la prueba. Luego, se realizó la regresión logística sobre el grupo de variables seleccionado (en el cuadro A.7 del anexo se presentan los estimadores puntuales de las variables seleccionadas) y se calcularon los indicadores de performance, los cuales presentamos a continuación junto con el CAP.

- Error de tipo 1: 37,3%
- Error de tipo 2: 29,8%
- KS: 32,7%
- Gini: 47,3%

[35] Los indicadores de performance de la validación del modelo completo fueron los siguientes: error de tipo 1: 35,9% - error de tipo 2: 27,5%, KS: 36,6%, Gini 48,7%, AUROC: 74,4%.

- AUROC: 73,7%

Gráfico 3.3: CAP del modelo de paquetes restringido



Fuente: Elaboración propia

Para validar el modelo, se utilizaron los datos de abril 2018 y se obtuvieron resultados superiores a los del desarrollo (mayores valores en KS, Gini y AUROC y menores valores en el error de tipo 1 y en el del tipo 2)[36].

3.2.3 Modelo de préstamos personales

En este apartado, se busca construir un modelo de *score* para el producto de préstamos personales partiendo de todas las variables explicativas y un modelo de score que utilice solamente las variables tradicionales.

- Modelo de préstamos personales completo

En primer lugar, se seleccionaron todas las variables explicativas de la base input y se realizaron las pruebas de significatividad individual. Solamente diez variables superaron dicha prueba. Luego, con las variables candidatas, se repitió la prueba y una de ellas no la superó. Finalmente, se seleccionaron las nueve variables candidatas y todas ellas superaron la prueba. Las variables que forman parte del modelo final son las siguientes: marca si el cliente posee garantía, antigüedad en el empleo y como cliente del banco,

[36] Los indicadores de performance de la validación *out of sample* para el modelo restringido de paquetes son los siguientes: error de tipo 1 – 36,3%, error de tipo 2 – 28,7%, KS – 34,9%, Gini – 47,6% y AUROC – 73,8%.

cantidad de tarjetas de crédito, cantidad de consultas al sistema veraz, *score* de veraz, cantidad de transacciones realizadas por *home banking*, marca si el cliente utiliza *mobile banking* y estado civil. Como se observa, dos de las variables alternativas forman parte del grupo reducido de variables explicativas del modelo final.

Como paso siguiente, se ejecutó la regresión logística con las variables seleccionadas y se obtuvieron los estimadores puntuales para cada una de ellas. Los mismos se presentan a continuación.

Tabla 3.7: Variables explicativas del modelo final de préstamos personales y sus estimadores

Parameter		Estimate
Intercept		-1.1526
marca_garantia		0.9571
marca_cliente_antig_		-0.3581
marca_antig_empleo_h		-0.2188
cantidadTarjetaCredi		0.1437
cantidadConsultas		-0.2571
scoreVeraz		0.00423
CANT_TRANSAC_HOME_BA		0.00369
INDICA_USA_MOBILE		0.2384
estadoCivil	D	0.3486
estadoCivil	M	0.6413
estadoCivil	S	0.2061
estadoCivil	W	0

Fuente: Salida SAS Studio

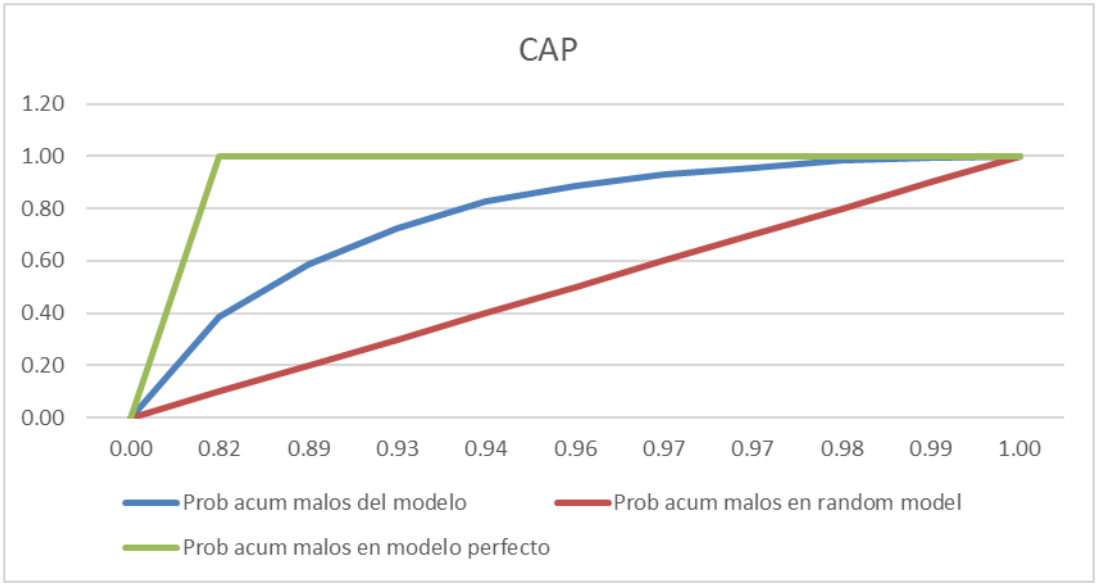
Se observa que ambas variables explicativas poseen un estimador puntual mayor a cero, indicando que a mayor uso de los sistemas de *mobile y home banking*, mejor comportamiento de pago posee el cliente. Asimismo, tal como se esperaba, las variables de garantía, score veraz y cantidad de tarjetas de crédito poseen valores mayores a cero (indicando un mejor comportamiento de pago cuanto más elevadas sean estas variables).

Como paso siguiente, se realizaron las pruebas de performance sobre el modelo de préstamos personales y se construyó el CAP; los mismos se presentan a continuación.

- Error de tipo 1: 7,5%
- Error de tipo 2: 8,7%
- KS: 46,3%
- Gini: 63,1%

- AUROC: 81,5%

Gráfico 3.4: CAP del modelo de préstamos personales no restringido



Fuente: Elaboración propia

Como paso final, se realizó la validación *out of sample* con los datos de abril 2018 obteniendo resultados aceptables en relación con aquellos obtenidos para el desarrollo[37].

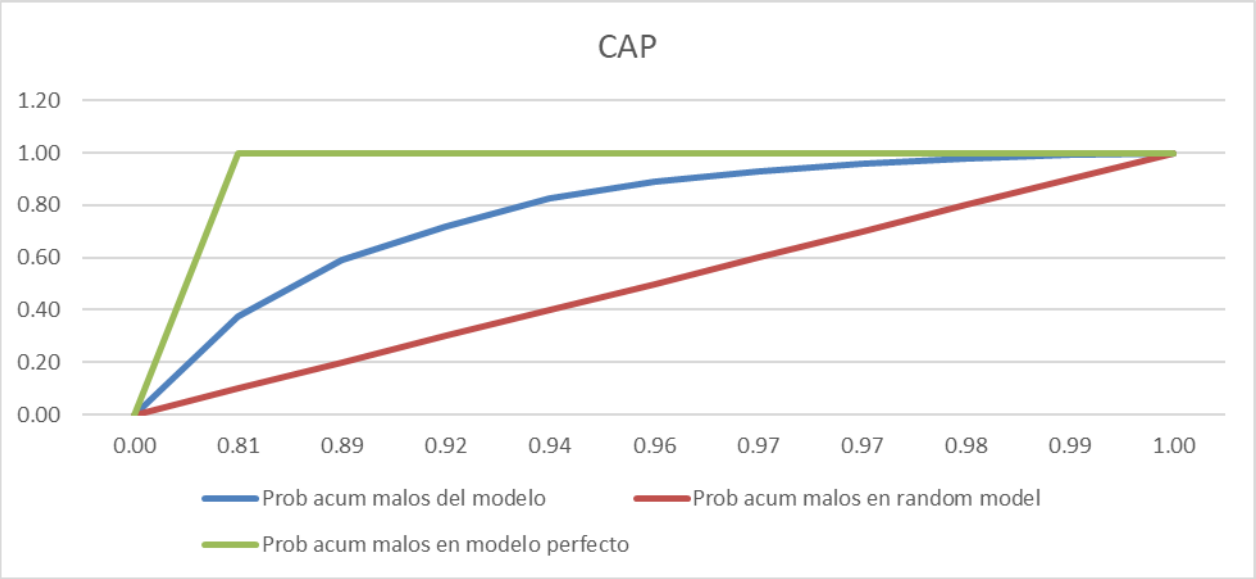
- Modelo de préstamos personales restringido

La construcción del modelo restringido se realizó con el mismo proceso iterativo descripto anteriormente. Para construir el modelo restringido, se siguió el mismo proceso descripto anteriormente. En el cuadro A.9 del anexo se presentan los estimadores puntuales de las variables predictoras que conforman el modelo final. Asimismo, sobre el modelo final, se calcularon los indicadores de performance, los cuales presentamos a continuación.

- Error de tipo 1: 7.4%
- Error de tipo 2: 8.9%
- KS: 46.0%
- Gini: 62.6%
- AUROC: 81.3%

[37] Los indicadores para la validación del modelo no restringido de préstamos personales son los siguientes: error de tipo 1 – 8,3%, error de tipo 2 – 11,9%, KS – 44,7%, Gini - 56,9% y AUROC – 78,5%.

Gráfico 3.5: CAP del modelo de préstamos personales restringido



Fuente: Elaboración propia

Para finalizar la construcción del modelo restringido, se realizó la validación del modelo con datos de abril 2018, obteniendo indicadores de performance cercanos a los del modelo restringido[38].

Aquí concluimos con el trabajo de construcción de modelos de *score* con datos alternativos y con datos tradicionales para los tres tipos de producto analizados; en el apartado siguiente, haremos la comparación de los resultados obtenidos bajo cada modelo.

3.3 Contraste de resultados

En este apartado realizaremos una evaluación de las discrepancias encontradas en los indicadores de performance de los modelos completos y de los modelos restringidos para cada tipo de producto.

Comparando los resultados de los modelos finales de tarjetas y paquetes, se puede observar una mejor *performance* del modelo no restringido, independientemente del indicador utilizado. A continuación, se presenta el resumen de los resultados encontrados para los modelos de tarjetas y paquetes.

[38] Los indicadores de performance de la validación del modelo restringido son los siguientes: error de tipo 1: 8,1% - error de tipo 2: 12,7%, KS: 43,2%, Gini 58,3%, AUROC: 79,1%.

Tabla 3.7: Comparación de resultados para el producto Tarjetas

Indicador de performance producto: Tarjetas	Modelo completo (A)	Modelo sin variables alternativas (B)
Error de tipo 1	33,9%	34,6%
Error de tipo 2	19,3%	19,5%
KS	46,4%	45,9%
Gini	61,5%	60,1%
AUROC	80,8%	80,0%

Fuente: Elaboración propia

Tabla 3.8: Comparación de resultados para el producto Paquetes

Indicador de performance producto: paquetes	Modelo completo (A)	Modelo sin variables alternativas (B)
Error de tipo 1	36,3%	37,3%
Error de tipo 2	29,2%	29,8%
KS	34,5%	32,7%
Gini	49,3%	47,3%
AUROC	74,6%	73,7%

Fuente: Elaboración propia

La elaboración de la matriz de confusión nos permitió obtener los errores de tipo 1 y de tipo 2 presentados en el cuadro anterior. Se observa que el modelo A, tanto para tarjetas como para paquetes, presenta un menor error de tipo 1 respecto al modelo B. Como detallamos en el capítulo 2, esto significa que el modelo A posee un menor porcentaje de aceptación de solicitudes que luego resultan morosas. Asimismo, el modelo A de tarjetas y paquetes presenta un menor error de tipo 2 que el modelo B, lo cual implica que el modelo A presenta una menor proporción de rechazo de solicitudes que tienen un buen comportamiento frente al pago de la deuda. Por lo tanto, con el modelo A se tendrán menores costos por riesgo de crédito dado que se aceptarán menos solicitudes morosas y se tendrá un menor costo de oportunidad al rechazar menos solicitudes no morosas.

Otro de los indicadores calculados fue el KS; se observa que el modelo A de tarjetas y paquetes presenta un KS más elevado que el modelo B, lo cual significa que las distribuciones por *score* de buenos pagadores y de malos pagadores del modelo A tienen un punto máximo de alejamiento mayor a las respectivas curvas del modelo B. Por lo tanto, bajo la medida de KS, el modelo A tiene un mayor poder de discriminación de riesgo que el modelo B.

Además de los indicadores mencionados, el contraste respecto al Gini y al AUROC, refleja el mejor resultado que posee el modelo A de tarjetas y paquetes. Esto nos indica que el modelo A clasifica mejor las solicitudes que el modelo B.

Los contrastes realizados parecen reflejar que los modelos de tarjetas y paquetes que incorporan datos alternativos presentan mejores resultados que los modelos restringidos que no incluyen variables alternativas.

En relación con el modelo de préstamos personales, al contrastar los resultados de los modelos finales, se puede observar una mejor *performance* del modelo no restringido para todos los indicadores desarrollados excepto por la probabilidad de error de tipo 2. A continuación, se presenta el resumen de los resultados encontrados para los modelos de préstamos personales.

Tabla 3.9: Comparación de resultados para el producto Préstamos Personales

Indicador de performance producto: préstamos personales	Modelo completo (A)	Modelo sin variables alternativas (B)
Error de tipo 1	7,5%	7,4%
Error de tipo 2	8,7%	8,9%
KS	46,3%	46,0%
Gini	63,1%	62,6%
AUROC	81,5%	81,3%

Fuente: Elaboración propia

A partir de la matriz de confusión, obtuvimos los valores de los errores de tipo 1 y de tipo 2. Respecto al primero de ellos, podemos observar que el modelo A presenta un valor levemente superior al valor del modelo B; esto quiere decir que el modelo B posee una menor proporción de aceptación de solicitudes morosas. Respecto al error de tipo 2, el modelo A presenta un menor valor en comparación al modelo B. Esto quiere decir que el modelo A posee una menor proporción de rechazo de solicitudes no morosas.

Respecto al resto de los indicadores de performance, se puede observar que el modelo A resulta más favorable en relación con el modelo B; por lo que el modelo A muestra una mayor exactitud para discernir riesgo.

Por lo tanto, si bien para la mayoría de los indicadores el modelo con información alternativa resulta preferible al modelo sin datos alternativos, los resultados en cuanto al

error de tipo 1 muestran resultados contrarios, indicando que el uso del modelo sin datos alternativos traería aparejado un menor costo de crédito por el rechazo de solicitudes morosas.

En el capítulo siguiente, se realizará una síntesis de los resultados obtenidos y se repasarán los objetivos propuestos y el abordaje dado a los mismos en el presente trabajo.

Conclusiones y futuras líneas de investigación

A lo largo del presente trabajo de investigación se ha abordado el objetivo principal planteado: analizar el impacto que posee la incorporación de datos alternativos en la predicción del riesgo de crédito. Atendiendo a los objetivos específicos, el primero de ellos consistió en estudiar el uso de los grandes volúmenes de información en la gestión de riesgo bancaria; este aspecto fue abordado en el capítulo 1. En el mismo, se describieron las principales funciones de un banco y los riesgos que enfrentan; luego, se pasó revista de las principales regulaciones existentes sobre el riesgo de crédito y, finalmente, se destacó la utilización de datos alternativos en la gestión del riesgo de crédito en los últimos tiempos.

El segundo objetivo específico planteado fue sentar las bases para la construcción de un modelo de *score* que ayude a evaluar la capacidad de repago de los clientes minoristas; el cual fue trabajado en el capítulo 2. En este capítulo, se explicó en qué consiste un modelo de *score*; luego, se hizo foco en el modelo de regresión logística y las fórmulas que lo componen; finalmente, se presentaron los principales indicadores de performance para los modelos de *scoring*.

El tercer y último objetivo trazado en la presente investigación fue la evaluación del aporte generado por datos alternativos en modelos de *score* para distintos productos bancarios; esto fue desarrollado en el capítulo 3. En dicho capítulo, primero realizamos un estudio sobre las bases input de cada tipo de producto bancario; luego, construimos un modelo de *score* para cada producto incorporando variables alternativas y construimos otro modelo de *score* restringido (el cual no utiliza datos alternativos). Finalmente, para cada producto, realizamos el cálculo y contraste de los indicadores de performance del modelo completo y del modelo restringido.

La hipótesis del trabajo fue que la incorporación de datos alternativos mejora la estimación del riesgo de crédito. En este sentido y gracias al trabajo realizado a lo largo de la presente investigación y a los resultados encontrados en el capítulo anterior, podemos corroborar parcialmente la hipótesis. En lo que respecta al producto de tarjetas y paquetes, podemos concluir que la incorporación de variables alternativas en la gestión del riesgo de crédito ayuda a mejorar la predicción del default. Esto se observa en los mejores indicadores de *performance* que poseen los modelos que incluyen variables alternativas en comparación con los modelos tradicionales para dichos productos

financieros. Por otro lado, en cuanto al producto de préstamos personales, el modelo con variables alternativas indica una mejor precisión en cuanto a la discriminación general de riesgo; pero, el modelo sin datos alternativos posee una leve diferencia a favor en cuanto al error que se comete aceptando créditos morosos. Es por este motivo que, si bien corroboramos la hipótesis del trabajo para los productos de tarjetas y paquetes, no podemos corroborarla para el producto de préstamos personales.

En conclusión, hemos estudiado el impacto de utilizar datos alternativos en distintos productos financieras y corroboramos el impacto positivo de éstos para los productos de tarjetas y paquetes. Asimismo, estudios futuros se requieren para poder profundizar en dos cuestiones principales. Por un lado, aunque en el presente trabajo se analizó el impacto de utilizar información alternativa asociada al uso de *home banking* y de *mobile banking* en la predicción del riesgo de crédito minorista, sería interesante que en futuros trabajos se puedan analizar otras variables que fueron introducidas en el capítulo 1, como ser la localización geográfica del individuo, las redes sociales o la información del historial de navegación en internet, entre otras.

Por otro lado, el presente trabajo se realizó con datos de aplicaciones a créditos de los primeros cuatro meses de abril 2018 (con observación de desempeño durante los doce meses siguiente; en este sentido, una futura línea de investigación podría ser realizar este mismo análisis en una ventana temporal más reciente. No obstante, hay que tener en cuenta que, durante la crisis de salud provocada por el Coronavirus en el año 2020, el gobierno otorgó moratorias en los vencimientos de créditos bancarios; por lo tanto, los datos de comportamiento de pago del 2020 difícilmente puedan ser utilizados para un análisis similar (al menos sin realizar un ajuste sobre los mismos), dado que el comportamiento frente a las obligaciones de deuda de los individuos se ve distorsionado al existir la opción legal de aplazar las mismas.

Bibliografía y Referencias bibliográficas

Abdou, H., & Pointon, J. (2011). *Credit scoring, statistical techniques and evaluation criteria: a review of the literature.* Greater Manchester, Inglaterra: John Wiley & Sons, Ltd.

Altman, E., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, 1721-1742.

Banco Central de la República Argentina. (13 de febrero de 2013). *Comunicación "A" 5398.* Obtenido de http://www.bcra.gov.ar/pdfs/comytexord/A5398.pdf

Banco Central de la República Argentina. (12 de enero de 2018). *Comunicación "A" 6430.* Obtenido de http://www.bcra.gov.ar/Pdfs/comytexord/A6430.pdf

Banco Central de la República Argentina. (9 de mayo de 2019). *Comunicación "A" 6693.* Obtenido de http://www.bcra.gob.ar/Pdfs/comytexord/A6693.pdf

Banco Central de la República Argentina. (1 de noviembre de 2020). *Comunicación "A" 7156.* Obtenido de https://www.bcra.gob.ar/Pdfs/Texord/t-cladeu.pdf

Banco Central de la República Argentina. (abril 2020). *Informe de Inclusión Financiera.*

Basel Committee on Banking Supervision. (1988). *International Convergence of Capital Measurement and Capital Standards.* Basel, Switzerland: Bank for International Settlements.

Basel Committee on Banking Supervision. (2006). *International Convergence of Capital Measurement and Capital Standards. A revised Framework.* Basel, Switzerland: Bank for International Settlements.

Basel Committee on Banking Supervision. (2006). *Sound credit risk assessment and valuation for loans.* Basel, Switzerland: Bank for International Settlements.

Bessis, J. (2015). *Risk Management in Banking.* West Sussex, Inglaterra: John Wiley & Sons, Ltd.

Brown, M. C. (1994). Using gini-style indices to evaluate the spatial patterns of health practitioners: Theoretical considerations and an application based on Alberta data. *Social Science & Medicine*, 1243-1256.

Business Wire. (18 de julio de 2016). *ZestFinance Receives Funding from Baidu to Fuel Development of Search-Based Underwriting Technology.* Obtenido de https://www.businesswire.com/news/home/20160717005040/en/ZestFinance-Receives-Funding-Baidu-Fuel-Development-Search-Based

Crouhy, M., Galai, D., & Mark, R. (2006). *The essentials of Risk Management.* Estados Unidos: McGraw-Hill.

Derby, N. (2003). *Mathematical Definition of the Gini Index.* Washington, Estados Unidos: University of Washington.

Emerj. (3 de abril de 2020). *Artificial Intelligence Applications for Lending and Loan Management.* Obtenido de https://emerj.com/ai-sector-overviews/artificial-intelligence-applications-lending-loan-management/

Equifax. (2 de mayo de 2018). *Credit Through the Ages: How Technology is Revolutionizing the Way We Assess Consumer Financial Behavior.* Obtenido de https://insight.equifax.com/credit-ages-technology/?intcmp=search

Equifax. (s.f.). *How Are Credit Scores Calculated?* Obtenido de https://www.equifax.com/personal/education/credit/score/how-is-credit-score-calculated/

Experian Information Solutions, Inc. (2018). *Alternative Data Across the Loan Life Cycle: How FinTech and Other Lenders Use It and Why.* Obtenido de https://www.experian.com/assets/consumer-information/reports/Experian_Aite_AltDataReport_Final_120418.pdf?elqTrackId=7714eff9f5204e7ca8517e8966438157&elqaid=3910&elqat=2

Farris, F. A. (2010). The Gini Index and Measures of Inequality. *The American Mathematical Monthly*, 851-864.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 861–874.

FICO. (8 de diciembre de 2015). *Not All Alternative Data Is Created Equal.* Obtenido de https://www.fico.com/blogs/not-all-alternative-data-created-equal

FICO. (29 de marzo de 2018). *FICO Continues to Expand Access to Credit with New FICO® Score XD 2.* Obtenido de https://www.fico.com/en/newsroom/fico-continues-expand-access-credit-new-fico-score-xd-2

FICO. (22 de mayo de 2019). *Leveraging Alternative Data to Extend Credit to More Borrowers.* Obtenido de https://www.fico.com/blogs/leveraging-alternative-data-extend-credit-more-borrowers

Forbes. (14 de agosto de 2019). *Alternative Data: The Great Equalizer To Lending Inequalities?* Obtenido de https://www.forbes.com/sites/forbestechcouncil/2019/08/14/alternative-data-the-great-equalizer-to-lending-inequalities/?sh=1e92db392449

Freixas, X., & Rochet, J.-C. (2008). *Microeconomics of Banking.* Cambridge, Estados Unidos: MIT Press.

Gasparini, L., & Cicowiez, M. (26-27 de Abril de 2007). *The socio-economic conditions in Argentina.* Buenos Aires, Argentina: Centro de Estudios Distributivos, Laborales y Sociales. Obtenido de researchgate: https://www.researchgate.net/profile/Leonardo_Gasparini2/publication/228458988_THE_SOCIO-ECONOMIC_CONDITIONS_IN_ARGENTINA/links/55dc55ae08ae9d6594945212/THE-SOCIO-ECONOMIC-CONDITIONS-IN-ARGENTINA.pdf

Gourieroux, C., & Jasiak, J. (2007). *The Econometrics of Individual Risk. Credit, Insurance, and Marketing.* New Jersey, Estados Unidos: Princeton University Press.

Hull, J. C. (2000). *Options, futures, and other derivatives.* Boston, Estados Unidos: Prentice Hall.

Hull, J. C. (2015). *Risk Management and Financial Institutions.* New Jersey, Estados Unidos: John Wiley & Sons, Inc.

International Monetary Fund. (marzo 2012). What Is a Bank? *Finance & Development*, 38 - 39.

iProfesional. (16 de julio de 2017). Obtenido de https://www.iprofesional.com/economia/252863-vinimos-a-ocupar-un-mercado-que-los-bancos-dejaron-libre

Jorion, P. (2007). *Value at Risk. The new benchmark for managing financial risk.* Estados Unidos: McGraw-Hill.

Lin, A. Z. (2013). *Variable Reduction in SAS by Using Weight of Evidence and Information Value*. Obtenido de https://support.sas.com/resources/papers/proceedings13/095-2013.pdf

Lorenz, M. (1905). Methods of Measuring the Concentration of Wealth. *American Statistical Association*, 209-219.

Mermelstein, D. (2006). *Defaults en carteras hipotecarias, macroeconomía y arreglos institucionales: más allá de los modelos de credit-scoring tradicionales.* Obtenido de https://www.researchgate.net/

Norma Internacional de Información Financiera 9. (24 de julio 2014). *Instrumentos Financieros.*

Road Show. (13 de febrero de 2019). *Scoring crediticio: cómo avanza Equifax en el uso de "data alternativa".* Obtenido de https://www.roadshow.com.ar/scoring-crediticio-como-avanza-equifax-en-el-uso-de-data-alternativa/

Rosignuolo, L. (2017). Principios de Economía Monetaria. Oferta y Demanda Monetaria, Banca Central y Política Monetaria. *Revista de Investigación en Modelos Financieros*, 1-38.

Unidad de Información Financiera. (2017). *Resolución 30-E/2017.* Ciudad de Buenos Aires, Argentina: Ministerio de Finanzas.

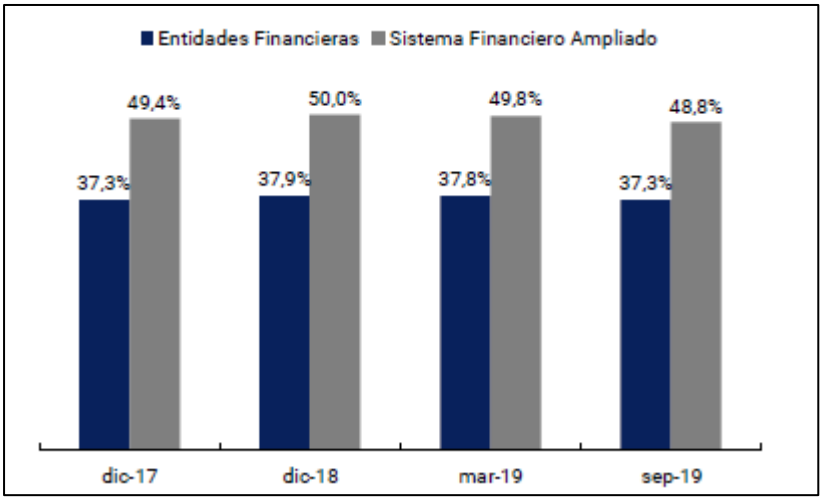
Wooldridge, J. M. (2010). *Introducción a la econometría. Un enfoque moderno.* México, D.F.: Cengage Learning Editores.

Anexos

1.1 Parte A: Material soporte referenciado en el cuerpo del trabajo

En esta sección del anexo se presentan los gráficos mencionados, pero no incluidos en los capítulos anteriores para no hacer extensa la lectura. Se introduce material tomado del informe de inclusión financiera del BCRA (gráficos A.1 y A.2 mencionados en el capítulo 1), se incluye el análisis de estadística descriptiva sobre las bases input (gráficos A.3 a A.5 mencionados en el capítulo 3) y se adjuntan las variables explicativas y los estimadores puntuales de los modelos restringidos (gráficos A.6 a A.8 mencionados en el capítulo 3).

Gráfico A.1: Porcentaje de adultos con al menos un financiamiento



Fuente: Informe de inclusión financiera (BCRA, abril 2020, p. 31)

Gráfico A.2: Porcentaje de adultos con al menos un financiamiento por provincia

Provincia	% de adultos con	
	Financiamiento de EEFF	Financiamiento de SFA
CABA	69,6%	79,8%
T. del Fuego	53,0%	62,8%
Neuquén	43,9%	54,3%
Santa Cruz	42,9%	53,3%
Córdoba	36,7%	50,7%
Chubut	40,2%	50,0%
Río Negro	38,5%	49,4%
La Pampa	42,5%	49,1%
Santa Fe	38,8%	48,8%
San Luis	33,4%	48,6%
Catamarca	30,0%	46,7%
Buenos Aires	36,0%	46,5%
Mendoza	32,2%	45,8%
Tucumán	31,3%	45,6%
La Rioja	28,2%	45,6%
Entre Ríos	32,7%	43,7%
S. del Estero	34,0%	43,3%
Salta	30,2%	42,4%
Chaco	27,2%	41,5%
San Juan	26,8%	41,1%
Misiones	29,4%	40,6%
Jujuy	28,5%	40,3%
Formosa	28,6%	39,3%
Corrientes	26,0%	39,2%
TOTAL NACIONAL	**37,3%**	**48,8%**

Fuente: Informe de inclusión financiera (BCRA, abril 2020, p. 32)

Gráfico A.3: Estadística descriptiva de la base de tarjetas

tipo_renta	Hasta 30.000	Desde 30.000 a 50.000	Desde 50.000 a 80.000	Desde 80.000	Missings
201801	9,323	3,856	1261	559	0
201802	9,620	3,784	1255	540	0
201803	10,515	4,673	1691	695	0
201804	7,752	2,721	948	522	0
frecuencia media	62.3%	25.2%	8.6%	3.9%	0.0%

marca_pyme	0	1	Missings
201801	14,980	19	0
201802	15,187	12	0
201803	17,551	23	0
201804	11,931	12	0
frecuencia media	99.9%	0.1%	0.0%

marca_cliente_antig_hasta12m	0	1	Missings
201801	12,510	2,489	0
201802	12,857	2,342	0
201803	14,842	2,732	0
201804	10,004	1,939	0
frecuencia media	84.1%	15.9%	0.0%

marca_antig_empleo_hasta24m	0	1	Missings
201801	9,523	5,476	0
201802	9,240	5,959	0
201803	10,944	6,630	0
201804	7,073	4,870	0
frecuencia media	61.6%	38.4%	0.0%

Estado Civil	D	M	S	W	Missings
201801	235	1,231	13489	44	0
201802	193	1,167	13808	31	0
201803	263	1,400	15861	50	0
201804	146	827	10938	32	0
frecuencia media	1.4%	7.7%	90.6%	0.3%	0.0%

marcaPoseeAuto	0	1	Missings
201801	11,805	3,194	0
201802	11,980	3,219	0
201803	13,698	3,876	0
201804	9,373	2,570	0
frecuencia media	78.5%	21.5%	0.0%

marcaPoseeCajaAhorro	0	1	Missings
201801	443	14,556	0
201802	679	14,520	0
201803	1,041	16,533	0
201804	1,118	10,825	0
frecuencia media	5.5%	94.5%	0.0%

Nivel estudios	A	B	C	D	E	F	Missings
201801	68	204	7836	2846	3857	188	0
201802	76	259	7406	3042	4218	198	0
201803	93	294	8573	3296	5096	222	0
201804	109	262	5526	2204	3668	174	0
frecuencia media	0.6%	1.7%	49.1%	19.1%	28.2%	1.3%	0.0%

TipoVivienda	H	M	P	R	Missings
201801	3,116	9,075	2693	115	0
201802	3,521	9,225	2357	96	0
201803	4,371	10,141	2938	124	0
201804	3,345	6,963	1579	56	0
frecuencia media	24.0%	59.3%	16.0%	0.7%	0.0%

refCtaCorrienteYAhorro	A	B	C	D	Missings
201801	382	61	7735	6821	0
201802	619	60	7973	6547	0
201803	968	73	8528	8005	0
201804	1,045	73	6060	4765	0
frecuencia media	5.0%	0.4%	50.7%	43.8%	0.0%

Grupo_edad	Menor a 40	Entre 40 y 60	Desde 60	Missings
201801	11,584	3,155	260	0
201802	11,926	3,009	264	0
201803	13,597	3,646	331	0
201804	9,295	2,450	198	0
frecuencia media	77.7%	20.5%	1.8%	0.0%

regulares_veraz	0	1	Missings
201801	3,681	11,318	0
201802	4,322	10,877	0
201803	4,884	12,690	0
201804	3,859	8,084	0
frecuencia media	28.0%	72.0%	0.0%

INDICA_USA_HOME_BANKING	0	1	Missings
201801	4,370	10,629	0
201802	5,213	9,986	0
201803	5,277	12,297	0
201804	3,913	8,030	0
frecuencia media	31.4%	68.6%	0.0%

INDICA_USA_MOBILE	0	1	Missings
201801	7,290	7,709	0
201802	7,706	7,493	0
201803	8,465	9,109	0
201804	5,996	5,947	0
frecuencia media	49.3%	50.7%	0.0%

no_default	0	1	Missings
201801	1,043	13,956	0
201802	1,160	14,039	0
201803	1,265	16,309	0
201804	910	11,033	0
frecuencia media	7.3%	92.7%	0.0%

ingresoMensualOrdinario	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	14,999	0	454,539,101.00	30,304.63	26,358.12	8,000.00	15,312.00	21,724.00	36,000.00	900,000.00
201802	14,952	247	452,154,644.00	30,240.41	27,520.02	8,000.00	15,000.00	21,000.00	36,000.00	900,000.00
201803	17,092	482	550,775,281.00	32,224.16	36,332.01	9,000.00	15,607.00	23,000.00	37,449.00	3,000,000.00
201804	11,276	667	354,112,052.00	31,404.05	31,367.57	9,000.00	15,000.00	20,300.00	36,000.00	1,240,000.00
promedio	14,580	349	452,895,269.50	31,043.31	30,394.43	8,500.00	15,229.75	21,506.00	36,362.25	1,510,000.00

cantidadTarjetaCredito	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	14,999	0	19,266.00	1.28	1.42	0.00	0.00	1.00	2.00	16.00
201802	15,199	0	18,414.00	1.21	1.42	0.00	0.00	1.00	2.00	16.00
201803	17,574	0	21,880.00	1.25	1.44	0.00	0.00	1.00	2.00	17.00
201804	11,943	0	13,347.00	1.12	1.38	0.00	0.00	1.00	2.00	16.00
promedio	14,929	0	18,226.75	1.21	1.41	0.00	0.00	1.00	2.00	16.25

cantidadConsultas	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	14,993	6	5,976.00	0.40	0.80	0.00	0.00	1.00	2.00	16.00
201802	15,197	2	5,713.00	0.38	0.78	0.00	0.00	1.00	2.00	16.00
201803	17,572	2	6,817.00	0.39	0.80	0.00	0.00	1.00	2.00	17.00
201804	11,943	0	4,175.00	0.35	0.74	0.00	0.00	1.00	2.00	16.00
promedio	14,926	2.5	5,670.25	0.38	0.78	0.00	0.00	1.00	2.00	16.25

scoreVeraz	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	14,997	2	9,484,822.00	632.45	163.69	15.00	532.00	663.00	36,000.00	903.00
201802	14,951	248	9,457,797.00	632.59	178.25	15.00	527.00	659.00	36,000.00	907.00
201803	17,090	484	10,871,245.00	636.12	190.72	13.00	523.25	662.00	37,449.00	904.00
201804	11,276	667	7,172,470.00	636.08	214.62	17.00	505.00	656.00	36,000.00	900.00
promedio	14,579	350.25	9,246,583.50	634.31	186.82	15.00	521.81	660.00	36,362.25	903.50

CANT_TRANSAC_HOME_BANKING	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	14,999	0	569,672.00	37.98	69.47	0.00	0.00	9.00	49.00	2,269.00
201802	15,199	0	484,381.00	31.87	59.71	0.00	0.00	5.00	41.00	1,441.00
201803	17,574	0	549,036.00	31.24	62.84	0.00	0.00	3.00	38.00	2,378.00
201804	11,943	0	289,544.00	24.24	50.56	0.00	0.00	1.00	27.00	724.00
promedio	14,929	0	473,158.25	31.33	60.64	0.00	0.00	4.50	38.75	1,703.00

CANTIDAD_TRANSACCIONES_MOBILE	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	14,999	0	599,914.00	40.00	70.77	0.00	0.00	3.00	56.00	948.00
201802	15,199	0	551,305.00	36.27	67.94	0.00	0.00	0.00	49.00	1,510.00
201803	17,574	0	746,023.00	42.45	75.70	0.00	0.00	3.00	58.00	1,521.00
201804	11,943	0	418,884.00	35.07	66.23	0.00	0.00	0.00	47.00	1,454.00
promedio	14,929	0	579,031.50	38.45	70.16	0.00	0.00	1.50	52.50	1,358.25

Fuente: Elaboración propia

Gráfico A.4: Estadística descriptiva de la base de paquetes

tipo_renta	Hasta 30.000	Desde 30.000 a 50.000	Desde 50.000 a 80.000	Desde 80.000	Missings
201801	9,372	2,296	1,080	577	0
201802	9,795	2,268	1,064	546	0
201803	11,323	2,892	1,336	766	0
201804	7,583	1,795	896	555	0
frecuencia media	70.3%	17.1%	8.1%	4.5%	0.0%

marca_pyme	0	1	Missings
201801	12,810	515	0
201802	13,068	605	0
201803	15,630	687	0
201804	10,264	565	0
frecuencia media	95.6%	4.4%	0.0%

marca_cliente_antig_hasta12m	0	1	Missings
201801	3,373	9,952	0
201802	3,192	10,481	0
201803	4,009	12,308	0
201804	2,496	8,333	0
frecuencia media	24.1%	75.9%	0.0%

marca_antig_empleo_hasta24m	0	1	Missings
201801	8,986	4,339	0
201802	8,775	4,898	0
201803	10,324	5,993	0
201804	6,627	4,202	0
frecuencia media	64.1%	35.9%	0.0%

Estado Civil	D	M	S	W	Missings
201801	397	1,511	11,204	213	0
201802	420	1,434	11,584	235	0
201803	476	1,662	13,907	272	0
201804	332	1,081	9,253	163	0
frecuencia media	3.0%	10.5%	84.9%	1.6%	0.0%

marcaPoseeAuto	0	1	Missings
201801	10,655	2,670	0
201802	10,866	2,807	0
201803	12,706	3,611	0
201804	8,553	2,276	0
frecuencia media	79.0%	21.0%	0.0%

marcaPoseeCajaAhorro	0	1	Missings
201801	5,193	8,132	0
201802	5,095	8,578	0
201803	6,355	9,962	0
201804	4,488	6,341	0
frecuencia media	39.0%	61.0%	0.0%

Nivel estudios	A	B	C	D	E	F	Missings
201801	95	214	6,563	2,264	4,074	115	0
201802	96	176	6,524	2,314	4,439	124	0
201803	94	191	7,461	2,744	5,705	122	0
201804	86	129	4,990	1,713	3,819	92	0
frecuencia media	0.7%	1.3%	47.2%	16.7%	33.3%	0.8%	0.0%

TipoVivienda	H	M	P	R	Missings
201801	3,802	7,134	2,260	129	0
201802	3,956	7,365	2,238	114	0
201803	5,014	8,977	2,202	124	0
201804	3,405	5,884	1,475	65	0
frecuencia media	29.9%	54.2%	15.1%	0.8%	0.0%

refCtaCorrienteYAhorro	A	B	C	D	Missings
201801	4,569	624	4,983	3,149	0
201802	4,507	588	5,605	2,973	0
201803	5,629	726	6,321	3,641	0
201804	4,006	482	3,950	2,391	0
frecuencia media	34.6%	4.5%	38.5%	22.4%	0.0%

Grupo_edad	Menor a 40	Entre 40 y 60	Desde 60	Missings
201801	8,528	3,669	1,128	0
201802	8,998	3,464	1,211	0
201803	10,824	4,049	1,444	0
201804	7,143	2,724	962	0
frecuencia media	65.6%	25.7%	8.8%	0.0%

regulares_veraz	0	1	Missings
201801	3,970	9,355	0
201802	4,674	8,999	0
201803	6,000	10,317	0
201804	4,247	6,582	0
frecuencia media	34.9%	65.1%	0.0%

INDICA_USA_HOME_BANKING	0	1	Missings
201801	5,168	8,157	0
201802	6,267	7,406	0
201803	6,726	9,591	0
201804	4,755	6,074	0
frecuencia media	42.3%	57.7%	0.0%

INDICA_USA_MOBILE	0	1	Missings
201801	9,236	4,089	0
201802	9,665	4,008	0
201803	11,290	5,027	0
201804	7,636	3,193	0
frecuencia media	69.9%	30.1%	0.0%

no_default	0	1	Missings
201801	1,676	11,649	0
201802	1,716	11,957	0
201803	2,127	14,190	0
201804	1,457	9,372	0
frecuencia media	12.9%	87.1%	0.0%

ingresoMensualOrdinario	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	12,974	351	374,674,216.00	28,878.85	38,756.33	6,456.00	14,500.00	19,909.00	33,000.00	2,100,000.00
201802	13,147	526	375,701,870.00	28,577.00	34,118.08	6,025.00	14,000.00	18,989.00	31,000.00	900,000.00
201803	15,356	961	473,466,960.00	30,832.70	51,336.64	4,480.00	14,000.00	19,190.00	35,000.00	3,069,265.00
201804	9,704	1125	303,909,094.00	31,317.92	39,495.39	5,250.00	13,441.00	19,000.00	35,000.00	1,702,828.00
promedio	12,795	740.75	381,938,035.00	29,901.62	40,926.61	5,552.75	13,985.25	19,272.00	33,500.00	1,943,023.25

cantidadTarjetaCredito	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	13,325	0	16,431.00	1.23	1.64	0.00	0.00	1.00	2.00	17.00
201802	13,673	0	15,619.00	1.14	1.63	0.00	0.00	1.00	2.00	17.00
201803	16,317	0	18,464.00	1.13	1.66	0.00	0.00	0.00	2.00	20.00
201804	10,829	0	11,842.00	1.09	1.66	0.00	0.00	0.00	2.00	19.00
promedio	13,536	0	15,589.00	1.15	1.65	0.00	0.00	0.50	2.00	18.25

cantidadConsultas	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	13,318	7	7,891.00	0.59	1.01	0.00	0.00	0.00	1.00	12.00
201802	13,668	5	7,412.00	0.54	0.96	0.00	0.00	0.00	1.00	11.00
201803	16,312	5	8,718.00	0.53	0.95	0.00	0.00	0.00	1.00	11.00
201804	10,828	1	5,514.00	0.51	0.92	0.00	0.00	0.00	1.00	11.00
promedio	13,532	4.5	7,383.75	0.54	0.96	0.00	0.00	0.00	1.00	11.25

scoreVeraz	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	13,324	1	8,365,188.00	627.83	154.76	19.00	534.00	650.00	750.00	901.00
201802	13,672	1	8,515,411.00	622.84	157.38	23.00	523.00	647.00	747.00	902.00
201803	16,314	3	10,105,677.00	619.45	156.94	27.00	513.00	636.00	746.00	903.00
201804	10,828	1	6,713,034.00	619.97	156.46	23.00	511.00	640.00	745.00	907.00
promedio	13,535	1.5	8,424,827.50	622.52	156.38	23.00	520.25	643.25	747.00	903.25

CANT_TRANSAC_HOME_BANKING	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	13,325	0	263,829.00	19.80	47.46	0.00	0.00	1.00	17.00	1,011.00
201802	13,673	0	228,614.00	16.72	52.31	0.00	0.00	0.00	12.00	3,363.00
201803	16,317	0	251,116.00	15.39	42.18	0.00	0.00	0.00	8.00	1,321.00
201804	10,829	0	129,727.00	11.98	35.56	0.00	0.00	0.00	4.00	942.00
promedio	13,536	0	218,321.50	15.97	44.38	0.00	0.00	0.25	10.25	1,659.25

CANTIDAD_TRANSACCIONES_MOBILE	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	13,325	0	256,825.00	19.27	51.97	0.00	0.00	0.00	11.00	956.00
201802	13,673	0	235,682.00	17.24	47.01	0.00	0.00	0.00	8.00	796.00
201803	16,317	0	318,459.00	19.52	52.42	0.00	0.00	0.00	10.00	1,144.00
201804	10,829	0	186,782.00	17.25	48.11	0.00	0.00	0.00	8.00	928.00
promedio	13,536	0	249,437.00	18.32	49.88	0.00	0.00	0.00	9.25	956.00

Fuente: Elaboración propia

Gráfico A.5: Estadística descriptiva de la base de préstamos personales

tipo_renta	Hasta 30.000	Desde 30.000 a 50.000	Desde 50.000 a 80.000	Desde 80.000	Missings
201801	617	766	960	656	0
201802	567	795	981	631	0
201803	670	959	1,178	836	0
201804	453	612	780	559	0
frecuencia media	19.2%	26.1%	32.4%	22.3%	0.0%

marca_garantia	0	1	Missings
201801	416	2583	0
201802	435	2539	0
201803	513	3130	0
201804	366	2038	0
frecuencia media	14.4%	85.6%	0.0%

marca_pyme	0	1	Missings
201801	2,315	684	0
201802	2,331	643	0
201803	2,754	889	0
201804	1,844	560	0
frecuencia media	76.9%	23.1%	0.0%

marca_cliente_antig_hasta12m	0	1	Missings
201801	2,431	568	0
201802	2,568	406	0
201803	3,167	476	0
201804	2,081	323	0
frecuencia media	85.2%	14.8%	0.0%

marca_antig_empleo_hasta24m	0	1	Missings
201801	2,569	430	0
201802	2,523	451	0
201803	3,086	557	0
201804	1,991	413	0
frecuencia media	84.6%	15.4%	0.0%

Estado Civil	D	M	S	W	Missings
201801	158	911	1,888	42	0
201802	148	858	1,920	48	0
201803	185	1043	2,357	58	0
201804	145	665	1,561	33	0
frecuencia media	5.3%	28.9%	64.3%	1.5%	0.0%

marcaPoseeAuto	0	1	Missings
201801	2,934	65	0
201802	2,893	81	0
201803	3,539	104	0
201804	2,318	86	0
frecuencia media	97.2%	2.8%	0.0%

marcaPoseeCajaAhorro	0	1	Missings
201801	1,068	1931	0
201802	1,066	1908	0
201803	1,317	2326	0
201804	856	1548	0
frecuencia media	35.8%	64.2%	0.0%

Nivel estudios	A	B	C	D	E	F	Missings
201801	3	25	1,823	471	671	6	0
201802	3	17	1,793	473	679	9	0
201803	6	19	2,252	576	784	6	0
201804	3	15	1,470	391	517	8	0
frecuencia media	0.1%	0.6%	61.0%	15.9%	22.1%	0.2%	0.0%

TipoVivienda	H	M	P	R	Missings
201801	1,967	511	499	22	0
201802	1,952	481	524	17	0
201803	2,413	558	650	22	0
201804	1,591	376	422	15	0
frecuencia media	65.9%	16.0%	17.4%	0.6%	0.0%

refCtaCorrienteYAhorro	A	B	C	D	Missings
201801	765	251	404	1,579	0
201802	825	207	374	1,568	0
201803	1,024	287	472	1,860	0
201804	688	166	343	1,207	0
frecuencia media	27.5%	7.6%	13.3%	51.7%	0.0%

Grupo_edad	Menor a 40	Entre 40 y 60	Desde 60	Missings
201801	1,627	1156	216	0
201802	1,667	1056	251	0
201803	1,993	1382	268	0
201804	1,323	876	205	0
frecuencia media	55.0%	37.2%	7.8%	0.0%

regulares_veraz	0	1	Missings
201801	247	2752	0
201802	251	2723	0
201803	363	3280	0
201804	263	2141	0
frecuencia media	9.4%	90.6%	0.0%

INDICA_USA_HOME_BANKING	0	1	Missings
201801	1,158	1841	0
201802	1,277	1697	0
201803	1,642	2001	0
201804	1,112	1292	0
frecuencia media	43.2%	56.8%	0.0%

INDICA_USA_MOBILE	0	1	Missings
201801	1,872	1127	0
201802	1,946	1028	0
201803	2,443	1200	0
201804	1,608	796	0
frecuencia media	65.5%	34.5%	0.0%

no_default	0	1	Missings
201801	219	2780	0
201802	214	2760	0
201803	314	3329	0
201804	221	2183	0
frecuencia media	8.1%	91.9%	0.0%

ingresoMensualOrdinario	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,993	6	187,383,822.77	62,607.29	97,722.86	3,500.00	35,000.00	50,000.00	75,000.00	3,000,000.00
201802	2,962	12	187,724,492.76	63,377.61	90,393.67	10,000.00	35,000.00	50,000.00	75,000.00	1,890,000.00
201803	3,625	18	232,860,020.17	64,237.25	86,680.14	5,300.00	36,000.00	51,000.00	76,840.73	2,000,000.00
201804	2,394	10	147,263,754.66	61,513.68	65,465.11	6,600.00	35,991.00	50,000.00	77,000.00	999,999.00
promedio	2,994	11.5	188,808,022.59	62,933.96	85,065.44	6,350.00	35,497.75	50,250.00	75,960.18	1,972,499.75

cantidadTarjetaCredito	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	6,715.00	2.24	1.95	0.00	1.00	2.00	3.00	15.00
201802	2,974	0	6,292.00	2.12	1.93	0.00	1.00	2.00	3.00	19.00
201803	3,643	0	7,498.00	2.06	2.02	0.00	1.00	2.00	3.00	19.00
201804	2,404	0	4,958.00	2.06	2.07	0.00	1.00	2.00	3.00	18.00
promedio	3,005	0	6,365.75	2.12	1.99	0.00	1.00	2.00	3.00	17.75

cantidadConsultas	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	2,285.00	0.76	1.13	0.00	0.00	0.00	1.00	10.00
201802	2,973	1	2,326.00	0.78	1.12	0.00	0.00	0.00	1.00	10.00
201803	3,642	1	2,692.00	0.74	1.12	0.00	0.00	0.00	1.00	9.00
201804	2,403	1	1,802.00	0.75	1.16	0.00	0.00	0.00	1.00	17.00
promedio	3,004	0.75	2,276.25	0.76	1.13	0.00	0.00	0.00	1.00	11.50

scoreVeraz	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	1,975,353.00	658.67	158.37	38.00	563.00	693.00	788.00	897.00
201802	2,973	1	1,941,362.00	653.00	158.43	35.00	554.00	688.00	781.00	898.00
201803	3,642	1	2,375,225.00	652.18	157.28	52.00	557.00	682.00	782.00	895.00
201804	2,403	1	1,537,389.00	639.78	163.39	43.00	535.75	669.50	774.00	895.00
promedio	3,004	0.75	1,957,332.25	650.91	159.37	42.00	552.44	683.13	781.25	896.25

CANT_TRANSAC_HOME_BANKING	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	110,222.00	36.75	66.76	0.00	0.00	5.00	47.50	758.00
201802	2,974	0	94,614.00	31.81	60.87	0.00	0.00	3.00	37.75	611.00
201803	3,643	0	95,439.00	26.20	54.94	0.00	0.00	0.00	27.50	783.00
201804	2,404	0	52,317.00	21.76	50.38	0.00	0.00	0.00	22.00	914.00
promedio	3,005	0	88,148.00	29.13	58.24	0.00	0.00	2.00	33.69	766.50

CANTIDAD_TRANSACCIONES_MOBILE	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	89,654.00	29.89	67.40	0.00	0.00	0.00	32.00	1,040.00
201802	2,974	0	82,159.00	27.63	63.83	0.00	0.00	0.00	25.00	961.00
201803	3,643	0	97,364.00	26.73	62.36	0.00	0.00	0.00	19.00	618.00
201804	2,404	0	64,530.00	26.84	65.54	0.00	0.00	0.00	21.00	1,050.00
promedio	3,005	0	83,426.75	27.77	64.78	0.00	0.00	0.00	24.25	917.25

pctFinanciacion	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,583	416	155,917.00	60.36	15.17	6.00	50.00	65.00	79.00	100.00
201802	2,539	435	151,617.00	59.72	14.98	2.00	50.00	64.00	79.00	100.00
201803	3,130	513	183,884.00	58.75	15.01	4.00	50.00	62.00	77.00	100.00
201804	2,038	366	118,321.00	58.06	15.35	6.00	50.00	63.00	78.00	100.00
promedio	2,573	432.5	152,434.75	59.22	15.13	4.50	50.00	63.50	78.25	100.00

cuotaPrestamo	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,984	15	24,430,205.00	8,187.07	9,892.86	420.00	4,277.00	6,325.00	9,466.00	381,405.00
201802	2,962	12	24,014,807.00	8,107.63	7,719.55	203.00	4,268.25	6,389.00	9,357.00	148,493.00
201803	3,637	6	30,940,394.00	8,507.12	8,052.86	307.00	4,392.00	6,463.00	9,835.00	93,229.00
201804	2,395	9	21,446,655.00	8,954.76	8,191.07	371.00	4,507.50	6,662.00	10,368.50	88,077.00
promedio	2,995	10.5	25,208,015.25	8,439.14	8,464.09	325.25	4,361.19	6,459.75	9,756.63	177,801.00

montoPrestamo	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	716,503,527.00	238,914.15	201,396.09	3,600.00	102,250.00	200,000.00	300,000.00	1,500,000.00
201802	2,974	0	701,815,982.00	235,983.85	205,066.14	4,000.00	100,000.00	200,000.00	300,000.00	1,800,000.00
201803	3,643	0	875,524,312.00	240,330.58	220,226.33	5,000.00	100,000.00	199,000.00	300,000.00	1,800,000.00
201804	2,404	0	584,208,381.00	243,015.13	220,152.47	5,000.00	100,000.00	199,100.00	300,000.00	1,800,000.00
promedio	3,005	0	719,513,050.50	239,560.93	211,710.26	4,400.00	100,562.50	199,525.00	300,000.00	1,725,000.00

plazoPrestamo	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	133,767.00	44.60	12.55	3.00	36.00	48.00	48.00	72.00
201802	2,974	0	131,884.00	44.35	12.51	4.00	36.00	48.00	48.00	72.00
201803	3,643	0	157,784.00	43.31	12.66	4.00	36.00	48.00	48.00	72.00
201804	2,404	0	101,433.00	42.19	13.46	6.00	36.00	48.00	48.00	72.00
promedio	3,005	0	131,217.00	43.61	12.79	4.25	36.00	48.00	48.00	72.00

relacionCuotaIngreso	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	2,112,027.42	704.24	17,255.63	0.00	9.37	13.27	19.66	528,250.00
201802	2,974	0	5,534,556.49	1,860.98	32,953.88	0.00	9.44	13.32	19.30	1,058,450.00
201803	3,643	0	5,769,522.44	1,583.73	26,921.96	0.00	9.46	13.58	19.76	777,600.00
201804	2,404	0	3,455,046.09	1,437.21	24,202.27	0.00	9.82	14.55	20.66	712,950.00
promedio	3,005	0	4,217,788.11	1,396.54	25,333.44	0.00	9.52	13.68	19.85	769,312.50

cantidadPersonasACargo	cantidad	missings	suma	promedio	desvío	mínimo	Q1	Q2	Q3	máximo
201801	2,999	0	339.00	0.11	0.46	0.00	0.00	0.00	0.00	5.00
201802	2,974	0	416.00	0.14	0.70	0.00	0.00	0.00	0.00	20.00
201803	3,643	0	566.00	0.16	1.54	0.00	0.00	0.00	0.00	85.00
201804	2,404	0	253.00	0.11	0.49	0.00	0.00	0.00	0.00	9.00
promedio	3,005	0	393.50	0.13	0.80	0.00	0.00	0.00	0.00	29.75

Fuente: Elaboración propia

Gráfico A.6: Variables explicativas del modelo restringido de tarjetas y sus estimadores

Parameter		Estimate
Intercept		1.7304
ingresoMensualOrdina		0.000026
marca_cliente_antig_		-0.3686
marca_antig_empleo_h		-0.1563
cantidadTarjetaCredi		0.1206
cantidadConsultas		-0.1522
regulares_veraz		0.2577
scoreVeraz		0.00355
tipo_renta	Desde 30.000 hasta 50.000	0.2441
tipo_renta	Desde 50.000 hasta 80.000	0.0924
tipo_renta	Desde 80.000	-0.2451
tipo_renta	Hasta 30.000	0
estadoCivil	D	-0.3681
estadoCivil	M	-0.2742
estadoCivil	S	-0.666
estadoCivil	W	0
nivelEstudios	A	-0.5387
nivelEstudios	B	-0.846
nivelEstudios	C	-0.4657
nivelEstudios	D	-0.3159
nivelEstudios	E	-0.3414
nivelEstudios	F	0
tipoVivienda	H	-0.7186
tipoVivienda	M	-0.4987
tipoVivienda	P	-0.1248
tipoVivienda	R	0
refCtaCorrienteYAhor	A	-0.9111
refCtaCorrienteYAhor	B	-0.8731
refCtaCorrienteYAhor	C	-0.4559
refCtaCorrienteYAhor	D	0

Fuente: Salida Sas Studio

Gráfico A.7: Variables explicativas del modelo restringido de paquetes y sus estimadores

Parameter		Estimate
Intercept		1.1753
marca_pyme		-0.4405
marca_cliente_antig_		-0.6269
marca_antig_empleo_h		0.0839
cantidadTarjetaCredi		0.074
marcaPoseeAuto		-0.1189
cantidadConsultas		-0.171
scoreVeraz		0.00356
tipo_renta	Entre 30.000 y 50.000	0.2847
tipo_renta	Entre 50.000 y 80.000	0.3214
tipo_renta	Mayor a 80.000	0.5924
tipo_renta	Menor a 30.000	0
nivelEstudios	A	-0.0943
nivelEstudios	B	-0.4152
nivelEstudios	C	-0.0239
nivelEstudios	D	-0.00824
nivelEstudios	E	0.1019
nivelEstudios	F	0
tipoVivienda	H	-0.5127
tipoVivienda	M	-0.4225
tipoVivienda	P	-0.3403
tipoVivienda	R	0
refCtaCorrienteYAhor	A	-0.599
refCtaCorrienteYAhor	B	-0.6827
refCtaCorrienteYAhor	C	-0.2927
refCtaCorrienteYAhor	D	0
Grupo_edad	Desde 60	0.1651
Grupo_edad	Entre 40 y 60	-0.0759
Grupo_edad	Menor a 40	0

Fuente: Salida Sas Studio

Gráfico A.8: Variables explicativas del modelo restringido de préstamos personales y sus estimadores

Parameter		Estimate
Intercept		-1.6023
marca_garantia		0.9924
marca_cliente_antig_		-0.3076
cantidadTarjetaCredi		0.1708
regulares_veraz		0.3006
cantidadConsultas		-0.2536
scoreVeraz		0.00435
estadoCivil	D	0.4714
estadoCivil	M	0.772
estadoCivil	S	0.3667
estadoCivil	W	0

Fuente: Salida Sas Studio

1.2 Parte B: Códigos SAS y documentación técnica

En esta sección del anexo, se presentarán los códigos SAS utilizados para generar la regresión logística, la salida de estos y la obtención de los indicadores de performance presentados en el capítulo 3. La sección se dividirá en tres partes, una por cada tipo de producto.

1.2.1 Modelo de tarjetas

Se importa la tabla input de tarjetas y se ejecuta la regresión logística con todas las variables explicativas como posibles predictoras.

Gráfico B.1: Código de SAS Studio para la primera iteración del modelo completo de tarjetas

```
/*Importo la base de tarjetas solamente*/

proc import datafile="I:/temp/Natali/Trabajo/base_card_trabajo.csv"
    dbms=csv
    out=tp.base_tarjetas
    replace
    ;
run;

/*1era corrida: Todas las variables explicativas*/
proc logistic data = tp.base_tarjetas (where=(periodo in (201801:201803))) covout;
    class  tipo_renta estadoCivil nivelestudios tipovivienda refctacorrienteyahorro Grupo_edad/ param = glm;
    model no_default (event="1") =  ingresoMensualOrdinario marca_pyme
    marca_cliente_antig_hasta12m marca_antig_empleo_hasta24m cantidadTarjetaCredito
    marcaPoseeAuto cantidadConsultas regulares_veraz scoreVeraz INDICA_USA_HOME_BANKING
    CANT_TRANSAC_HOME_BANKING INDICA_USA_MOBILE CANTIDAD_TRANSACCIONES_MOBILE
    tipo_renta estadoCivil nivelestudios    tipovivienda refctacorrienteyahorro Grupo_edad
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.2: Salida de SAS Studio para la primera iteración del modelo completo de tarjetas

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
ingresoMensualOrdina	1	28.5424	<.0001
marca_pyme	1	0.7437	0.3885
marca_cliente_antig_	1	75.1996	<.0001
marca_antig_empleo_h	1	15.3487	<.0001
cantidadTarjetaCredi	1	12.0333	0.0005
marcaPoseeAuto	1	1.9037	0.1677
cantidadConsultas	1	68.8070	<.0001
regulares_veraz	1	26.7111	<.0001
scoreVeraz	1	787.3156	<.0001
INDICA_USA_HOME_BANK	1	21.8652	<.0001
CANT_TRANSAC_HOME_BA	1	33.9547	<.0001
INDICA_USA_MOBILE	1	64.9751	<.0001
CANTIDAD_TRANSACCION	1	40.4635	<.0001
tipo_renta	3	10.3120	0.0161
estadoCivil	3	11.4378	0.0096
nivelEstudios	5	24.9351	0.0001
tipoVivienda	3	46.9158	<.0001
refCtaCorrienteYAhor	3	74.6017	<.0001
Grupo_edad	2	8.0531	0.0178

Fuente: Salida de SAS Studio

Como se observa en el gráfico presentado, dos de las variables explicativas presentan un p valor mayor al 5%; por lo tanto, se realiza una segunda regresión logística sin considerar dichas variables. A continuación, se presenta el código SAS.

Gráfico B.3: Código de SAS Studio para el modelo final de tarjetas no restringido

```
/*Sin variables con p valor >5%: Final*/
proc logistic data = tp.base_tarjetas (where=(periodo in (201801:201803))) covout;
    class  tipo_renta estadoCivil nivelestudios tipovivienda refctacorrienteyahorro Grupo_edad/ param = glm;
    model no_default (event="1") =  ingresoMensualOrdinario
    marca_cliente_antig_hasta12m marca_antig_empleo_hasta24m cantidadTarjetaCredito
    cantidadConsultas regulares_veraz scoreVeraz INDICA_USA_HOME_BANKING
    CANT_TRANSAC_HOME_BANKING INDICA_USA_MOBILE CANTIDAD_TRANSACCIONES_MOBILE
    tipo_renta estadoCivil nivelestudios    tipovivienda refctacorrienteyahorro Grupo_edad
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.4: Salida de SAS Studio para el modelo final de tarjetas no restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
ingresoMensualOrdina	1	28.5707	<.0001
marca_cliente_antig_	1	76.0843	<.0001
marca_antig_empleo_h	1	15.0950	0.0001
cantidadTarjetaCredi	1	12.3653	0.0004
cantidadConsultas	1	68.8671	<.0001
regulares_veraz	1	26.6945	<.0001
scoreVeraz	1	792.0610	<.0001
INDICA_USA_HOME_BANK	1	22.0810	<.0001
CANT_TRANSAC_HOME_BA	1	34.3089	<.0001
INDICA_USA_MOBILE	1	64.8946	<.0001
CANTIDAD_TRANSACCION	1	40.4822	<.0001
tipo_renta	3	10.2308	0.0167
estadoCivil	3	11.4707	0.0094
nivelEstudios	5	23.5620	0.0003
tipoVivienda	3	52.3987	<.0001
refCtaCorrienteYAhor	3	73.3822	<.0001
Grupo_edad	2	8.1098	0.0173

Fuente: Salida de SAS Studio

Se observa que todos los p valores son inferiores al 5%, por lo que este modelo no requiere una iteración adicional.

Gráfico B.5: Matriz de confusión y cálculo del KS, Gini y AUROC del modelo final de tarjetas no restringido

Punto de corte	0.92740517	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	33.9%	19.3%
moroso	670	2798		
no moroso	29273	15031	Sensibilidad	Especificidad
			66.1%	80.7%

KS, Gini, Auroc Modelo final

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.8222409	1152	3626	4778	0.332	0.082	0.100	0.2503363			
0.822240877	0.8726341	765	4012	4777	0.553	0.172	0.200	0.3803684	0.221	0.254	0.0560831
0.872634067	0.9067159	586	4191	4777	0.722	0.267	0.300	0.4547454	0.169	0.439	0.0742463
0.906715942	0.9340997	375	4402	4777	0.830	0.366	0.400	0.4635179	0.108	0.633	0.0684852
0.934099688	0.9560142	223	4554	4777	0.894	0.469	0.500	0.4250303	0.064	0.836	0.0537245
0.956014162	0.9719049	167	4610	4777	0.942	0.573	0.600	0.3691311	0.048	1.042	0.0501936
0.971904933	0.983259	98	4679	4777	0.971	0.679	0.700	0.2917782	0.028	1.252	0.0353797
0.983259037	0.9901883	59	4718	4777	0.988	0.785	0.800	0.2022994	0.017	1.464	0.0249085
0.990188343	0.9945911	26	4751	4777	0.995	0.893	0.900	0.1025601	0.007	1.678	0.012579
0.994591052	1	17	4761	4778	1.000	1.000	1.000	1.11E-16	0.005	1.893	0.0092771
								Aceptable			Bueno
							KS	46.4%		Gini	61.5%
										AUROC	80.8%

Fuente: Elaboración propia

A continuación, se presentan los cálculos realizados para la validación *out of the sample* (OOS) del modelo final de tarjetas no restringido.

Gráfico B.6: Matriz de confusión y cálculo del KS, Gini y AUROC de la validación OOS del modelo final de tarjetas no restringido

Punto de corte	0.92380474	El punto de corte se definió a partir de la proporción de morosos en la muestra		
Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	36.8%	21.4%
moroso	195	715		
no moroso	6976	4057	Sensibilidad	Especificidad
			63.2%	78.6%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.821012152	274	921	1195	0.301	0.083	0.100	0.21762206			
0.821012152	0.866298055	201	993	1194	0.522	0.173	0.200	0.34849846	0.221	0.257	0.0567563
0.866298055	0.897134917	152	1042	1194	0.689	0.268	0.300	0.42108749	0.167	0.441	0.07372886
0.897134917	0.923949512	88	1106	1194	0.786	0.368	0.400	0.41754606	0.097	0.636	0.06151217
0.923949512	0.947388274	89	1106	1195	0.884	0.468	0.500	0.41510354	0.098	0.837	0.08181948
0.947388274	0.965485047	46	1148	1194	0.934	0.572	0.600	0.36160151	0.051	1.041	0.05261578
0.965485047	0.979654734	30	1164	1194	0.967	0.678	0.700	0.28906687	0.033	1.250	0.04122298
0.979654734	0.988608016	18	1176	1194	0.987	0.785	0.800	0.20225776	0.020	1.463	0.028929
0.988608016	0.994092973	8	1186	1194	0.996	0.892	0.900	0.10355328	0.009	1.677	0.0147394
0.994092973	1	4	1191	1195	1.000	1.000	1.000	3.3307E-16	0.004	1.892	0.00831671
								Aceptable			Bueno
							KS	42.1%		Gini	58.0%
										AUROC	79.0%

Fuente: Elaboración propia

A continuación, se presentarán los códigos y documentación técnica involucrados en la generación del modelo final de tarjetas restringido (sin variables alternativas).

Gráfico B.7: Código de SAS Studio para la primera iteración del modelo de tarjetas restringido

```
/*Completo sin variables alternativas*/
proc logistic data = tp.base_tarjetas (where=(periodo in (201801:201803))) covout;
    class  tipo_renta estadoCivil nivelestudios tipovivienda refctacorrienteyahorro Grupo_edad/ param = glm;
    model no_default (event="1") =  ingresoMensualOrdinario marca_pyme
    marca_cliente_antig_hasta12m marca_antig_empleo_hasta24m cantidadTarjetaCredito
    marcaPoseeAuto cantidadConsultas regulares_veraz scoreVeraz
    tipo_renta estadoCivil nivelestudios    tipovivienda refctacorrienteyahorro Grupo_edad
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.8: Salida de SAS Studio para la primera iteración del modelo de tarjetas restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
ingresoMensualOrdina	1	49.0753	<.0001
marca_pyme	1	1.0256	0.3112
marca_cliente_antig_	1	57.3218	<.0001
marca_antig_empleo_h	1	14.7899	0.0001
cantidadTarjetaCredi	1	20.2730	<.0001
marcaPoseeAuto	1	2.7240	0.0988
cantidadConsultas	1	57.3719	<.0001
regulares_veraz	1	28.7491	<.0001
scoreVeraz	1	829.3742	<.0001
tipo_renta	3	10.3039	0.0162
estadoCivil	3	12.3586	0.0063
nivelEstudios	5	29.2945	<.0001
tipoVivienda	3	59.2338	<.0001
refCtaCorrienteYAhor	3	106.7871	<.0001
Grupo_edad	2	4.2051	0.1221

Fuente: Salida de SAS Studio

Como se observa, tres de las variables poseen un p valor superior al 5%, por lo que se realiza una segunda iteración sin las mismas.

Gráfico B.9: Código de SAS Studio para el modelo final de tarjetas restringido

```
/*Sin alternativas y sin p >5%*/
proc logistic data = tp.base_tarjetas (where=(periodo in (201801:201803))) covout;
    class  tipo_renta estadoCivil nivelestudios tipovivienda refctacorrienteyahorro / param = glm;
    model no_default (event="1") =  ingresoMensualOrdinario
    marca_cliente_antig_hasta12m marca_antig_empleo_hasta24m cantidadTarjetaCredito
    cantidadConsultas regulares_veraz scoreVeraz
    tipo_renta estadoCivil nivelestudios    tipovivienda refctacorrienteyahorro
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.10: Salida de SAS Studio para el modelo final de tarjetas restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
ingresoMensualOrdina	1	49.0126	<.0001
marca_cliente_antig_	1	58.4844	<.0001
marca_antig_empleo_h	1	15.1625	<.0001
cantidadTarjetaCredi	1	20.9129	<.0001
cantidadConsultas	1	57.4061	<.0001
regulares_veraz	1	29.1773	<.0001
scoreVeraz	1	841.8156	<.0001
tipo_renta	3	10.2370	0.0167
estadoCivil	3	14.5410	0.0023
nivelEstudios	5	27.3411	<.0001
tipoVivienda	3	66.3373	<.0001
refCtaCorrienteYAhor	3	104.6709	<.0001

Fuente: Salida de SAS Studio

Se observa que todos los p valores son inferiores al 5%, por lo que este modelo no requiere una iteración adicional.

Gráfico B.11: Matriz de confusión y cálculo del KS, Gini y AUROC del modelo final de tarjetas restringido

Punto de corte	0.92740517	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	34.6%	19.5%
moroso	675	2793		
no moroso	28991	15313	Sensibilidad	Especificidad
			65.4%	80.5%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.8267623	1092	3686	4778	0.315	0.083	0.100	0.23168099			
0.826762302	0.87198782	758	4019	4777	0.533	0.174	0.200	0.35953661	0.219	0.257	0.05619647
0.871987815	0.90543529	588	4189	4777	0.703	0.268	0.300	0.4345355	0.170	0.442	0.07500483
0.905435286	0.93240921	425	4353	4778	0.826	0.367	0.400	0.45883154	0.123	0.635	0.07784065
0.932409207	0.95438637	248	4528	4776	0.897	0.469	0.500	0.42813954	0.072	0.836	0.0597571
0.954386369	0.9710028	147	4630	4777	0.939	0.573	0.600	0.36602185	0.042	1.042	0.04418239
0.971002798	0.98252538	100	4677	4777	0.968	0.679	0.700	0.28929082	0.029	1.252	0.03611347
0.982525383	0.98955942	61	4716	4777	0.986	0.785	0.800	0.20043384	0.018	1.464	0.02575839
0.989559423	0.99379512	38	4739	4777	0.997	0.892	0.900	0.10442565	0.011	1.678	0.01838463
0.99379512	1	11	4767	4778	1.000	1.000	1.000	2.2204E-16	0.003	1.892	0.00600243
								Aceptable			Bueno
							KS	45.9%		Gini	60.1%
										AUROC	80.0%

Fuente: Elaboración propia

A continuación, se presentan los cálculos realizados para la validación *out of the sample* (OOS) del modelo final de tarjetas restringido.

Gráfico B.12: Matriz de confusión y cálculo del KS, Gini y AUROC de la validación OOS del modelo final de tarjetas restringido

Punto de cor	0.92380474	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	36.9%	19.5%
moroso	177	733		
no moroso	6957	4076	Sensibilidad	Especificidad
			63.1%	80.5%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−pa Mi-1	Bi=paBi+pa Bi-1	Zi=Ai*Bi
0	0.82214422	249	946	1195	0.274	0.086	125.426	0.1878836			
0.82214422	0.86664637	245	1133	1378	0.543	0.188	270.060	0.35442245	0.269	0.274	0.07381701
0.86664637	0.89340181	116	894	1010	0.670	0.269	376.069	0.40086534	0.127	0.458	0.05836955
0.89340181	0.92321196	121	1073	1194	0.803	0.367	501.390	0.43657868	0.133	0.636	0.08459128
0.92321196	0.94708089	69	1126	1195	0.879	0.469	626.817	0.41034539	0.076	0.835	0.06335061
0.94708089	0.9651994	45	1149	1194	0.929	0.573	752.138	0.35565382	0.049	1.042	0.0515123
0.9651994	0.97965448	37	1157	1194	0.969	0.678	877.459	0.29144594	0.041	1.251	0.05085274
0.97965448	0.98818407	16	1178	1194	0.987	0.785	1002.781	0.20225776	0.018	1.462	0.02571148
0.98818407	0.9934461	6	1188	1194	0.993	0.892	1128.102	0.1011742	0.007	1.677	0.01105574
0.9934461	1	6	1189	1195	1.000	1.000	1253.528	0	0.007	1.892	0.01247626
								Aceptable			Bueno
							KS	43.7%		Gini	56.8%
										AUROC	78.4%

Fuente: Elaboración propia

1.2.2 Modelo de paquetes

Se importa la tabla input de paquetes y se ejecuta la regresión logística con todas las variables explicativas como posibles predictoras.

Gráfico B.13: Código de SAS Studio para la primera iteración del modelo completo de paquetes no restringido

```
/*Importo la base de paquetes solamente*/

proc import datafile="I:/temp/Natali/Trabajo/base_pack_trabajo_2.csv"
    dbms=csv
    out=tp.base_paquetes
    replace
    ;
run;

/*Completo*/
proc logistic data = tp.base_paquetes (where=(periodo in (201801:201803)))  covout;
    class estadoCivil  tipo_renta nivelestudios tipovivienda refctacorrienteyahorro Grupo_edad/ param = glm;
    model no_default (event="1") = marca_pyme marca_cliente_antig_hasta12m
    marca_antig_empleo_hasta24m cantidadtarjetacredito ingresomensualordinario marcaposeeauto
    cantidadConsultas regulares_veraz scoreVeraz INDICA_USA_HOME_BANKING CANT_TRANSAC_HOME_BANKING
    INDICA_USA_MOBILE CANTIDAD_TRANSACCIONES_MOBILE
    estadoCivil  tipo_renta nivelestudios   tipovivienda refctacorrienteyahorro Grupo_edad
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.14: Salida de SAS Studio para la primera iteración del modelo completo de paquetes no restringido

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_pyme	1	33.1622	<.0001
marca_cliente_antig_	1	77.9726	<.0001
marca_antig_empleo_h	1	4.2762	0.0387
cantidadTarjetaCredi	1	15.6354	<.0001
ingresoMensualOrdina	1	0.8276	0.3630
marcaPoseeAuto	1	9.5194	0.0020
cantidadConsultas	1	176.7036	<.0001
regulares_veraz	1	1.3550	0.2444
scoreVeraz	1	1105.2311	<.0001
INDICA_USA_HOME_BANK	1	42.6159	<.0001
CANT_TRANSAC_HOME_BA	1	44.6064	<.0001
INDICA_USA_MOBILE	1	87.9164	<.0001
CANTIDAD_TRANSACCION	1	41.9725	<.0001
estadoCivil	3	6.7525	0.0802
tipo_renta	3	23.1404	<.0001
nivelEstudios	5	17.9211	0.0030
tipoVivienda	3	12.1528	0.0069
refCtaCorrienteYAhor	3	67.1072	<.0001
Grupo_edad	2	16.4564	0.0003

Fuente: Salida de SAS Studio

Como se observa, tres de las variables explicativas presentan un p valor mayor al 5%; por lo tanto, se realiza una segunda regresión logística sin considerar dichas variables. A continuación, se presenta el código SAS.

Gráfico B.15: Código de SAS Studio para el modelo final de paquetes no restringido

```
proc logistic data = tp.base_paquetes (where=(periodo in (201801:201803)))  covout;
    class tipo_renta nivelestudios  tipovivienda refctacorrienteyahorro Grupo_edad/ param = glm;
    model no_default (event="1") = marca_pyme marca_cliente_antig_hasta12m
    marca_antig_empleo_hasta24m cantidadtarjetacredito  marcaposeeauto
    cantidadConsultas  scoreVeraz INDICA_USA_HOME_BANKING CANT_TRANSAC_HOME_BANKING
    INDICA_USA_MOBILE CANTIDAD_TRANSACCIONES_MOBILE
    tipo_renta nivelestudios tipovivienda refctacorrienteyahorro Grupo_edad
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.16: Salida de SAS Studio para el modelo final de paquetes no restringido

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_pyme	1	33.5558	<.0001
marca_cliente_antig_	1	76.2603	<.0001
marca_antig_empleo_h	1	4.9475	0.0261
cantidadTarjetaCredi	1	16.0308	<.0001
marcaPoseeAuto	1	9.6212	0.0019
cantidadConsultas	1	181.3162	<.0001
scoreVeraz	1	1141.0367	<.0001
INDICA_USA_HOME_BANK	1	41.8641	<.0001
CANT_TRANSAC_HOME_BA	1	44.2811	<.0001
INDICA_USA_MOBILE	1	87.6084	<.0001
CANTIDAD_TRANSACCION	1	42.4458	<.0001
tipo_renta	3	41.4420	<.0001
nivelEstudios	5	18.6042	0.0023
tipoVivienda	3	12.0709	0.0071
refCtaCorrienteYAhor	3	69.9954	<.0001
Grupo_edad	2	15.0440	0.0005

Fuente: Salida de SAS Studio

Todas las variables poseen un p valor inferior al 5%, por lo que no se requiere una iteración adicional.

Gráfico B.17: Matriz de confusión y cálculo del KS, Gini y AUROC del modelo final de paquetes no restringido

Punto de corte	87.3%	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	36.3%	29.2%
moroso	1613	3906		
no moroso	24065	13731	Sensibilidad	Especificidad
			63.7%	70.8%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.74176836	1482	2850	4332	0.269	0.075	0.100	0.1931221			
0.74176836	0.804333137	971	3362	4333	0.444	0.164	0.200	0.28010856	0.176	0.240	0.04218296
0.804333137	0.842156257	754	3576	4330	0.581	0.259	0.300	0.32211433	0.137	0.423	0.05783425
0.842156257	0.870750406	663	3668	4331	0.701	0.356	0.400	0.34519748	0.120	0.615	0.07387852
0.870750406	0.895172472	547	3787	4334	0.800	0.456	0.500	0.34411385	0.099	0.812	0.08050175
0.895172472	0.917017742	424	3907	4331	0.877	0.560	0.600	0.31756863	0.077	1.016	0.078039
0.917017742	0.937153391	286	4044	4330	0.929	0.667	0.700	0.26239417	0.052	1.226	0.06354089
0.937153391	0.957701317	221	4110	4331	0.969	0.775	0.800	0.19369599	0.040	1.442	0.05773865
0.957701317	0.97673238	126	4206	4332	0.992	0.887	0.900	0.10524459	0.023	1.662	0.03794205
0.97673238	1	45	4286	4331	1.000	1.000	1.000	1.1102E-16	0.008	1.887	0.01538269
								Aceptable			Bueno
							KS	34.5%		Gini	49.3%
										AUROC	74.6%

Fuente: Elaboración propia

A continuación, se presentan los cálculos realizados para la validación *out of the sample* (OOS) del modelo final de paquetes no restringido.

Gráfico B.18: Matriz de confusión y cálculo del KS, Gini y AUROC de la validación OOS del modelo final de tarjetas no restringido

Punto de corte	86.5%	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	35.9%	27.5%
moroso	400	1057		
no moroso	6009	3363	Sensibilidad	Especificidad
			64%	73%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.74289688	363	720	1083	0.249	0.077	0.100	0.17231749			
0.742896881	0.79750937	259	825	1084	0.427	0.165	0.200	0.26205185	0.178	0.242	0.04296117
0.797509375	0.83348357	215	867	1082	0.574	0.257	0.300	0.31710573	0.148	0.422	0.06230353
0.833483568	0.86346236	207	876	1083	0.717	0.351	0.400	0.36570857	0.142	0.608	0.08640788
0.863462362	0.88871677	131	952	1083	0.806	0.452	0.500	0.35404017	0.090	0.803	0.07222026
0.888716768	0.91268671	97	985	1082	0.873	0.558	0.600	0.31551503	0.067	1.010	0.06723579
0.912686711	0.93412062	73	1010	1083	0.923	0.665	0.700	0.25785016	0.050	1.223	0.06126545
0.934120615	0.95565809	59	1024	1083	0.964	0.775	0.800	0.1890827	0.040	1.440	0.05830434
0.955658093	0.97523483	42	1041	1083	0.992	0.886	0.900	0.10683351	0.029	1.660	0.0478563
0.975234828	1	11	1072	1083	1.000	1.000	1.000	1.1102E-16	0.008	1.886	0.01423595
								Aceptable			Bueno
							KS	36.6%		Gini	48.7%
										AUROC	74.4%

Fuente: Elaboración propia

A continuación, se presentarán los códigos y documentación técnica involucrados en la generación del modelo final de paquetes restringido (sin considerar datos alternativos).

Gráfico B.19: Código de SAS Studio para la primera iteración del modelo de paquetes restringido

```
proc logistic data = tp.base_paquetes (where=(periodo in (201801:201803)))  covout;
    class estadoCivil  tipo_renta nivelestudios tipovivienda refctacorrienteyahorro Grupo_edad/ param = glm;
    model no_default (event="1") = marca_pyme marca_cliente_antig_hasta12m
    marca_antig_empleo_hasta24m cantidadtarjetacredito ingresomensualordinario marcaposeeauto
    cantidadConsultas regulares_veraz scoreVeraz
    estadoCivil  tipo_renta nivelestudios   tipovivienda refctacorrienteyahorro Grupo_edad
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.20: Salida de SAS Studio para la primera iteración del modelo de paquetes restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_pyme	1	37.6124	<.0001
marca_cliente_antig_	1	106.5468	<.0001
marca_antig_empleo_h	1	6.3179	0.0120
cantidadTarjetaCredi	1	19.3777	<.0001
ingresoMensualOrdina	1	0.8481	0.3571
marcaPoseeAuto	1	10.3458	0.0013
cantidadConsultas	1	146.7031	<.0001
regulares_veraz	1	0.2327	0.6295
scoreVeraz	1	1169.7565	<.0001
estadoCivil	3	5.5845	0.1337
tipo_renta	3	31.7074	<.0001
nivelEstudios	5	24.8572	0.0001
tipoVivienda	3	14.4929	0.0023
refCtaCorrienteYAhor	3	158.4638	<.0001
Grupo_edad	2	13.1099	0.0014

Fuente: Salida de SAS Studio

Como se observa, tres de las variables explicativas presentan un p valor mayor al 5%; por lo que se ejecuta una segunda regresión logística sin considerar dichas variables. A continuación, se presenta el código SAS.

Gráfico B.21: Código de SAS Studio para el modelo final de paquetes restringido

```
proc logistic data = tp.base_paquetes (where=(periodo in (201801:201803)))  covout;
    class tipo_renta nivelestudios  tipovivienda refctacorrienteyahorro Grupo_edad/ param = glm;
    model no_default (event="1") = marca_pyme marca_cliente_antig_hasta12m
    marca_antig_empleo_hasta24m cantidadtarjetacredito  marcaposeeauto
    cantidadConsultas  scoreVeraz
    tipo_renta nivelestudios tipovivienda refctacorrienteyahorro Grupo_edad
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.22: Salida de SAS Studio para el modelo final de paquetes restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_pyme	1	37.8112	<.0001
marca_cliente_antig_	1	105.3698	<.0001
marca_antig_empleo_h	1	6.4790	0.0109
cantidadTarjetaCredi	1	21.5857	<.0001
marcaPoseeAuto	1	10.2836	0.0013
cantidadConsultas	1	149.4380	<.0001
scoreVeraz	1	1201.0773	<.0001
tipo_renta	3	55.3037	<.0001
nivelEstudios	5	25.2803	0.0001
tipoVivienda	3	14.4232	0.0024
refCtaCorrienteYAhor	3	161.2269	<.0001
Grupo_edad	2	11.1368	0.0038

Fuente: Salida de SAS Studio

Dado que todos los p valores son inferiores al 5%, no se requiere una iteración adicional.

Gráfico B.23: Matriz de confusión y cálculo del KS, Gini y AUROC del modelo final de paquetes restringido

Punto de corte	87.3%	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	37.3%	29.8%
moroso	1642	3877		
no moroso	23690	14106	Sensibilidad	Especificidad
			63%	70%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−pa Mi-1	Bi=paBi+pa Bi-1	Zi=Ai*Bi
0	0.749419	1446	2886	4332	0.262	0.076	0.100	0.1856467			
0.749418996	0.80818196	917	3432	4349	0.428	0.167	0.200	0.26099673	0.166	0.244	0.04046129
0.808181957	0.84223759	720	3594	4314	0.559	0.262	0.300	0.29636572	0.130	0.429	0.05602022
0.842237593	0.8686976	697	3634	4331	0.685	0.358	0.400	0.32650897	0.126	0.621	0.07838221
0.868697598	0.8925944	553	3779	4332	0.785	0.458	0.500	0.32672416	0.100	0.817	0.08184075
0.892594404	0.91366935	468	3863	4331	0.870	0.561	0.600	0.30931555	0.085	1.019	0.08640661
0.913669346	0.93377615	329	4002	4331	0.930	0.666	0.700	0.26304357	0.060	1.227	0.07314787
0.933776146	0.9551414	212	4120	4332	0.968	0.775	0.800	0.19245009	0.038	1.442	0.05538933
0.955141397	0.97447698	133	4198	4331	0.992	0.887	0.900	0.1054787	0.024	1.662	0.04005249
0.974476977	1	44	4288	4332	1.000	1.000	1.000	0	0.008	1.887	0.01504043
								Aceptable			Bueno
							KS	32.7%		Gini	47.3%
										AUROC	73.7%

Fuente: Elaboración propia

A continuación, se presentan los cálculos realizados para la validación *out of the sample* (OOS) del modelo final de paquetes restringido.

Gráfico B.24: Matriz de confusión y cálculo del KS, Gini y AUROC de la validación OOS del modelo final de tarjetas restringido

Punto de corte	86.5%	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	36.3%	28.7%
moroso	418	1039		
no moroso	5970	3402	Sensibilidad	Especificidad
			64%	71%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi–paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.75157462	355	728	1083	0.244	0.078	0.100	0.16597315			
0.751574623	0.80406104	255	828	1083	0.419	0.166	0.200	0.25264204	0.175	0.244	0.0426525
0.804061039	0.83555555	214	869	1083	0.566	0.259	0.300	0.30679618	0.147	0.425	0.06238988
0.835555549	0.86298939	199	884	1083	0.702	0.353	0.400	0.34905468	0.137	0.612	0.08356394
0.862989386	0.88848315	145	938	1083	0.802	0.453	0.500	0.34848888	0.100	0.806	0.08023579
0.888483152	0.91137787	93	989	1082	0.865	0.559	0.600	0.30679156	0.064	1.012	0.06458577
0.91137787	0.93205467	89	994	1083	0.927	0.665	0.700	0.26181538	0.061	1.223	0.0747326
0.932054666	0.95364648	49	1034	1083	0.960	0.775	0.800	0.18511749	0.034	1.440	0.04842225
0.953646476	0.97375479	47	1036	1083	0.992	0.886	0.900	0.10683351	0.032	1.661	0.05357069
0.973754789	1	11	1072	1083	1.000	1.000	1.000	2.2204E-16	0.008	1.886	0.01423595
								Aceptable			Bueno
							KS	34.9%		Gini	47.6%
										AUROC	73.8%

Fuente: Elaboración propia

1.2.3 Modelo de préstamos personales

Se importa la tabla input de préstamos personales y se ejecuta la regresión logística con todas las variables explicativas como posibles predictoras.

Gráfico B.25: Código de SAS Studio para la primera iteración del modelo completo de préstamos personales no restringido

```
/*Importo la base de préstamos*/
proc import datafile="I:/temp/Natali/Trabajo/base_prestamos_trabajo.csv"
    dbms=csv
    out=tp.base_prestamos_trabajo_2
    replace
    ;
run;
/*Completo*/
proc logistic data = tp.base_prestamos_trabajo_2 (where=(periodo in (201801:201803)))  covout;
    class tipo_renta grupo_edad estadocivil nivelestudios  tipovivienda refctacorrienteyahorro/ param = glm;
    model no_default (event="1") = Marca_garantia pctFinanciacion cuotaPrestamo montoPrestamo plazoPrestamo
    relacionCuotaIngreso  ingresomensualordinario Marca_Pyme marca_cliente_antig_hasta12m marca_antig_empleo_hasta24
    cantidadTarjetaCredito marcaPoseeAuto cantidadPersonasACargo  regulares_veraz cantidadConsultas scoreVeraz
    INDICA_USA_HOME_BANKING CANT_TRANSAC_HOME_BANKING INDICA_USA_MOBILE CANTIDAD_TRANSACCIONES_MOBILE
    tipo_renta grupo_edad estadocivil nivelestudios tipovivienda refctacorrienteyahorro
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.26: Salida de SAS Studio para la primera iteración del modelo completo de préstamos personales no restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_garantia	1	31.8574	<.0001
pctFinanciacion	1	2.3457	0.1256
cuotaPrestamo	1	0.0460	0.8301
montoPrestamo	1	0.2605	0.6098
plazoPrestamo	1	0.1970	0.6572
relacionCuotaIngreso	1	0.0644	0.7996
ingresomensualordina	1	0.0003	0.9852
marca_pyme	1	2.2214	0.1361
marca_cliente_antig_	1	9.0145	0.0027
marca_antig_empleo_h	1	4.7391	0.0295
cantidadTarjetaCredi	1	20.5813	<.0001
marcaPoseeAuto	1	3.1067	0.0780
cantidadPersonasACar	1	0.1555	0.6933
regulares_veraz	1	4.6792	0.0305
cantidadConsultas	1	65.3828	<.0001
scoreVeraz	1	283.1817	<.0001
INDICA_USA_HOME_BANK	1	0.2093	0.6473
CANT_TRANSAC_HOME_BA	1	11.6230	0.0007
INDICA_USA_MOBILE	1	6.3896	0.0115
CANTIDAD_TRANSACCION	1	0.0986	0.7535
tipo_renta	3	0.9157	0.8216
grupo_edad	2	0.5444	0.7617
estadoCivil	3	14.5995	0.0022
nivelEstudios	5	6.8019	0.2358
tipoVivienda	3	1.3593	0.7151
refCtaCorrienteYAhor	3	5.2717	0.1529

Fuente: Salida de SAS Studio

Como se observa, solamente diez de las variables explicativas presentan un p valor inferior al 5%; por lo que se ejecuta una segunda regresión logística considerando solamente dichas variables. A continuación, se presenta el código SAS.

Gráfico B.27: Código de SAS Studio para la segunda iteración del modelo completo de préstamos personales no restringido

```
proc logistic data = tp.base_prestamos_trabajo_2 (where=(periodo in (201801:201803))) covout;
    class estadocivil/ param = glm;
    model no_default (event="1") = marca_garantia marca_cliente_antig_hasta12m marca_antig_empleo_hasta24m
    cantidadTarjetaCredito regulares_veraz cantidadConsultas scoreVeraz
    CANT_TRANSAC_HOME_BANKING INDICA_USA_MOBILE estadocivil
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.28: Salida de SAS Studio para la segunda iteración del modelo completo de préstamos personales no restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_garantia	1	104.7426	<.0001
marca_cliente_antig_	1	12.7704	0.0004
marca_antig_empleo_h	1	3.9945	0.0457
cantidadTarjetaCredi	1	19.4946	<.0001
regulares_veraz	1	3.6238	0.0570
cantidadConsultas	1	66.5713	<.0001
scoreVeraz	1	299.2552	<.0001
CANT_TRANSAC_HOME_BA	1	12.2429	0.0005
INDICA_USA_MOBILE	1	5.8814	0.0153
estadoCivil	3	13.8223	0.0032

Fuente: Salida de SAS Studio

Hay una de las variables cuyo p valor es superior al 5%, por lo que se ejecuta una iteración adicional sin considerar la misma.

Gráfico B.29: Código de SAS Studio para el modelo final de préstamos personales no restringido

```
proc logistic data = tp.base_prestamos_trabajo_2 (where=(periodo in (201801:201803))) covout;
    class estadocivil/ param = glm;
    model no_default (event="1") = marca_garantia marca_cliente_antig_hasta12m marca_antig_empleo_hasta24m
    cantidadTarjetaCredito cantidadConsultas scoreVeraz
    CANT_TRANSAC_HOME_BANKING INDICA_USA_MOBILE estadocivil
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.30: Salida de SAS Studio para el modelo final de préstamos personales no restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_garantia	1	103.3829	<.0001
marca_cliente_antig_	1	12.0492	0.0005
marca_antig_empleo_h	1	4.7469	0.0294
cantidadTarjetaCredi	1	25.7220	<.0001
cantidadConsultas	1	65.3297	<.0001
scoreVeraz	1	298.2798	<.0001
CANT_TRANSAC_HOME_BA	1	12.5359	0.0004
INDICA_USA_MOBILE	1	6.3445	0.0118
estadoCivil	3	13.8676	0.0031

Fuente: Salida de SAS Studio

Todas las variables presentan un p valor inferior al 5%, por lo que no se requiere una iteración adicional.

Gráfico B.31: Matriz de confusión y cálculo del KS, Gini y AUROC del modelo final de préstamos personales no restringido

Punto de corte	0.92231697	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	7.5%	8.7%
moroso	212	2238		
no moroso	6631	535	Sensibilidad	Especificidad
			93%	91%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.81551475	287	675	962	0.384	0.076	0.100	0.30809569			
0.815514751	0.89163967	152	810	962	0.588	0.167	0.200	0.42024693	0.203	0.244	0.04955667
0.891639667	0.92517044	104	857	961	0.727	0.264	0.300	0.46284178	0.139	0.432	0.06007538
0.925170441	0.94460304	75	887	962	0.827	0.364	0.400	0.46323211	0.100	0.628	0.06306656
0.944603037	0.95802763	44	917	961	0.886	0.467	0.500	0.41874055	0.059	0.832	0.04898007
0.958027627	0.9676642	32	930	962	0.929	0.572	0.600	0.35671894	0.043	1.040	0.04454304
0.967664195	0.97474594	20	941	961	0.956	0.678	0.700	0.2773928	0.027	1.251	0.03348758
0.974745942	0.98078798	20	942	962	0.983	0.785	0.800	0.19795392	0.027	1.463	0.03917198
0.980787977	0.98664619	9	952	961	0.995	0.892	0.900	0.10266193	0.012	1.677	0.02020032
0.986646191	1	4	958	962	1.000	1.000	1.000	1.1102E-16	0.005	1.892	0.0101311
								Aceptable			Bueno
							KS	46.3%		Gini	63.1%
										AUROC	81.5%

Fuente: Elaboración propia

A continuación, se presentan los cálculos realizados para la validación *out of the sample* (OOS) del modelo final de préstamos personales no restringido.

Gráfico B.32: Matriz de confusión y cálculo del KS, Gini y AUROC del modelo final de préstamos personales no restringido

Punto de corte	0.90806988	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	8.3%	11.9%
moroso	70	518		
no moroso	1665	151	Sensibilidad	Especificidad
			92%	88%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.79151591	79	162	241	0.357	0.074	0.100	0.28325626			
0.791515907	0.87701135	51	189	240	0.588	0.161	0.200	0.42744739	0.231	0.235	0.05423024
0.877011352	0.91522419	26	214	240	0.706	0.259	0.300	0.44706421	0.118	0.420	0.04936542
0.915224192	0.93839593	11	230	241	0.756	0.364	0.400	0.39147837	0.050	0.623	0.03100884
0.938395929	0.95293848	11	229	240	0.805	0.469	0.500	0.33635062	0.050	0.833	0.04147433
0.952938476	0.96335122	15	225	240	0.873	0.572	0.600	0.30115475	0.068	1.041	0.07067156
0.963351223	0.97212838	9	232	241	0.914	0.678	0.700	0.23560296	0.041	1.251	0.0509283
0.972128378	0.97871162	13	227	240	0.973	0.782	0.800	0.19044115	0.059	1.461	0.0859314
0.978711619	0.98512775	4	236	240	0.991	0.891	0.900	0.10043259	0.018	1.673	0.03027922
0.985127746	1	2	239	241	1.000	1.000	1.000	2.2204E-16	0.009	1.891	0.01710876
								Aceptable			Bueno
							KS	44.7%		Gini	56.9%
										AUROC	78.5%

Fuente: Elaboración propia

A continuación, se detallan los códigos y documentación técnica involucrados en la generación del modelo final de préstamos personales restringido.

Gráfico B.33: Código de SAS Studio para la primera iteración del modelo de préstamos personales restringido

```
proc logistic data = tp.base_prestamos_trabajo_2 (where=(periodo in (201801:201803)))  covout;
    class tipo_renta grupo_edad estadocivil nivelestudios   tipovivienda refctacorrienteyahorro/ param = glm;
    model no_default (event="1") = Marca_garantia pctFinanciacion cuotaPrestamo montoPrestamo plazoPrestamo
    relacionCuotaIngreso  ingresomensualordinario Marca_Pyme marca_cliente_antig_hasta12m marca_antig_empleo_hasta24
    cantidadTarjetaCredito marcaPoseeAuto cantidadPersonasACargo  regulares_veraz cantidadConsultas scoreVeraz
    tipo_renta grupo_edad estadocivil nivelestudios tipovivienda refctacorrienteyahorro
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.34: Salida de SAS Studio para la primera iteración del modelo de préstamos personales restringido

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_garantia	1	30.6895	<.0001
pctFinanciacion	1	1.4424	0.2297
cuotaPrestamo	1	0.0143	0.9047
montoPrestamo	1	0.4601	0.4976
plazoPrestamo	1	0.2516	0.6160
relacionCuotaIngreso	1	0.0887	0.7659
ingresomensualordina	1	0.0002	0.9881
marca_pyme	1	2.1670	0.1410
marca_cliente_antig_	1	9.0899	0.0026
marca_antig_empleo_h	1	3.9107	0.0480
cantidadTarjetaCredi	1	26.8314	<.0001
marcaPoseeAuto	1	3.1689	0.0751
cantidadPersonasACar	1	0.1443	0.7041
regulares_veraz	1	5.2748	0.0216
cantidadConsultas	1	62.7220	<.0001
scoreVeraz	1	290.9780	<.0001
tipo_renta	3	1.4013	0.7052
grupo_edad	2	1.6161	0.4457
estadoCivil	3	14.0393	0.0029
nivelEstudios	5	8.8957	0.1133
tipoVivienda	3	1.8131	0.6121
refCtaCorrienteYAhor	3	0.0689	0.9953

Fuente: Salida de SAS Studio

Como se observa, solo ocho de las variables explicativas presentan un p valor por debajo del 5%; por lo que se ejecuta una segunda regresión logística considerando solamente dichas variables. A continuación, se presenta el código SAS.

Gráfico B.35: Código de SAS Studio para la segunda iteración del modelo de préstamos personales restringido

```
proc logistic data = tp.base_prestamos_trabajo_2 (where=(periodo in (201801:201803))) covout;
    class estadocivil/ param = glm;
    model no_default (event="1") = marca_garantia marca_cliente_antig_hasta12m marca_antig_empleo_hasta24m
    cantidadTarjetaCredito regulares_veraz cantidadConsultas scoreVeraz
    estadocivil
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.36: Salida de SAS Studio para la segunda iteración del modelo de préstamos personales restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_garantia	1	104.9407	<.0001
marca_cliente_antig_	1	8.9165	0.0028
marca_antig_empleo_h	1	3.7806	0.0518
cantidadTarjetaCredi	1	32.4132	<.0001
regulares_veraz	1	4.9067	0.0268
cantidadConsultas	1	63.8280	<.0001
scoreVeraz	1	307.0067	<.0001
estadoCivil	3	12.8341	0.0050

Fuente: Salida de SAS Studio

Hay una de las variables cuyo p valor es superior al 5%, por lo que se ejecuta otra iteración sin considerar la misma.

Gráfico B.37: Código de SAS Studio para el modelo final de préstamos personales restringido

```
proc logistic data = tp.base_prestamos_trabajo_2 (where=(periodo in (201801:201803))) covout;
    class estadocivil/ param = glm;
    model no_default (event="1") = marca_garantia marca_cliente_antig_hasta12m
    cantidadTarjetaCredito regulares_veraz cantidadConsultas scoreVeraz
    estadocivil
    / clparm=both expb;
run;
```

Fuente: Elaboración propia

Gráfico B.38: Salida de SAS Studio para el modelo final de préstamos personales restringido

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
marca_garantia	1	116.0676	<.0001
marca_cliente_antig_	1	9.1240	0.0025
cantidadTarjetaCredi	1	35.2922	<.0001
regulares_veraz	1	5.7323	0.0167
cantidadConsultas	1	64.6974	<.0001
scoreVeraz	1	321.2869	<.0001
estadoCivil	3	13.4998	0.0037

Fuente: Salida de SAS Studio

Todas las variables presentan un p valor inferior al 5%, por lo que no se requiere otra iteración.

Gráfico B.39: Matriz de confusión y cálculo del KS, Gini y AUROC del modelo final de préstamos personales restringido

Punto de corte	0.92231697	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	7.4%	8.9%
moroso	222	2279		
no moroso	6590	525	Sensibilidad	Especificidad
			92.6%	91.1%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−paMi-1	Bi=paBi+paBi-1	Zi=Ai*Bi
0	0.81448571	280	682	962	0.375	0.077	0.100	0.29793561			
0.814485715	0.88987354	160	802	962	0.589	0.167	0.200	0.42169837	0.214	0.244	0.05230981
0.889873538	0.92449858	95	866	961	0.716	0.265	0.300	0.45123026	0.127	0.432	0.05497693
0.924498576	0.94324059	81	881	962	0.825	0.364	0.400	0.46032923	0.108	0.629	0.06823415
0.943240594	0.95738707	48	914	962	0.889	0.467	0.500	0.42153067	0.064	0.832	0.05344005
0.957387068	0.96777946	29	933	962	0.928	0.573	0.600	0.35515475	0.039	1.040	0.04037151
0.967779457	0.97485979	25	935	960	0.961	0.678	0.700	0.28319857	0.033	1.251	0.04185193
0.974859786	0.98044068	13	949	962	0.979	0.785	0.800	0.19359959	0.017	1.463	0.02545983
0.980440675	0.98587767	11	950	961	0.993	0.892	0.900	0.10121049	0.015	1.677	0.02469592
0.985877669	1	5	957	962	1.000	1.000	1.000	0	0.007	1.892	0.01266463
								Aceptable			Bueno
							KS	46.0%		Gini	62.6%
										AUROC	81.3%

Fuente: Elaboración propia

Finalmente, se presentan los cálculos realizados para la validación *out of the sample* (OOS) del modelo final de préstamos personales restringido.

Gráfico B.40: Matriz de confusión y cálculo del KS, Gini y AUROC del modelo final de préstamos personales restringido

Punto de corte	0.90806988	El punto de corte se definió a partir de la proporción de morosos en la muestra

Matriz de contingencia Modelo 1			Prob Error T1	Prob Error T2
	no rech	rech	8.1%	12.7%
moroso	72	494		
no moroso	1689	149	Sensibilidad	Especificidad
			92%	87%

KS, Gini, Auroc

PD		Malos	Buenos	Total	prob acumulada Malos	prob acumulada Buenos	paT	KS	Ai=paMi−pa Mi-1	Bi=paBi+pa Bi-1	Zi=Ai*Bi
0	0.79449538	84	157	241	0.380	0.072	0.100	0.30817112			
0.794495376	0.8795441	44	196	240	0.579	0.162	0.200	0.41748144	0.199	0.234	0.04651327
0.8795441	0.91815695	25	215	240	0.692	0.260	0.300	0.4321153	0.113	0.422	0.04772585
0.918156948	0.9399619	13	228	241	0.751	0.365	0.400	0.3864954	0.059	0.625	0.0367546
0.939961901	0.95381985	16	224	240	0.824	0.467	0.500	0.3562825	0.072	0.832	0.0602268
0.953819855	0.96493372	13	227	240	0.882	0.571	0.600	0.31112069	0.059	1.038	0.06108701
0.964933718	0.97302515	15	226	241	0.950	0.675	0.700	0.27546674	0.068	1.246	0.08456958
0.973025151	0.97903453	5	235	240	0.973	0.782	0.800	0.19044115	0.023	1.457	0.03296763
0.979034526	0.98536743	4	236	240	0.991	0.891	0.900	0.10043259	0.018	1.673	0.03027922
0.985367428	1	2	239	241	1.000	1.000	1.000	0	0.009	1.891	0.01710876
								Aceptable			Bueno
							KS	43.2%		Gini	58.3%
										AUROC	79.1%

Fuente: Elaboración propia