Trabajo Final de Maestría

Maestría en Gestión Económica y Financiera de Riesgos

Escuela de Estudios de Posgrado - Facultad de Ciencias Económicas

Universidad de Buenos Aires

Modelo Estadístico de Aprobación Crediticia

Rodrigo Aparicio

Director de Tesis: Javier García Fronti

Co-director: Pablo Matías Herrera

Mayo de 2016

Índice

Índice de tablas

Índice de gráficos

Introducción

La Asignación de un Crédito Bancario forma parte del negocio habitual de las Entidades Financieras, proveyendo a los agentes económicos del financiamiento necesario para llevar a cabo proyectos de inversión, a cambio de una tasa de interés que sirva de rédito. Es muy importante que el prestatario cumpla con las obligaciones contraídas en el crédito, garantizando la solvencia del sistema financiero y permitiendo mantener la estabilidad del mismo.

La decisión de los agentes económicos de hacer frente a las obligaciones, depende de los flujos de fondos futuros del prestatario, de su patrimonio y del riesgo de impago o también conocido como “riesgo de crédito”.

Para las entidades financieras poder estimar la probabilidad de impago de los créditos, es muy importante dado que permite diseñar políticas de asignación crediticia que preserven la estabilidad del sistema, disminuyendo la cartera morosa y garantizando la solvencia de las entidades (White, 1975).

Para estimar la probabilidad de impago, que como hemos visto es un problema social con implicancias sobre la solvencia del sistema financiero, se propone utilizar metodología estadística de avanzada sobre datos recientes de carteras crediticias, para luego extrapolar resultados que permitan refinar las políticas futuras de asignación crediticia.

Los datos a utilizar corresponden al año 2015, que por las características Macroeconómicas sirven para hacer estimaciones sobre el riesgo de impago en los próximos años.

La justificación de la temática, es la siguiente. Los Modelos de Aprobación Crediticia combinan el conocimiento académico (principalmente estadístico asociado a los métodos de regresión avanzada), con las necesidades del negocio bancario vinculadas a la colocación de préstamos con baja probabilidad de mora.

El abordaje de estos temas presenta beneficios para las entidades financieras, la sociedad en su conjunto y forman parte de mis gustos personales y profesionales. Por tales motivos decidí profundizar en el análisis estadístico y financiero como marco general para el desarrollo del trabajo de tesis.

El objetivo general de la tesis es diseñar un modelo de aprobación crediticia que permita estimar la probabilidad de mora de potenciales tomadores de crédito. Para cumplimentarlo, se llevarán a cabo los siguientes objetivos específicos. En primer lugar, se va a analizar la información disponible para el desarrollo del modelo, contenido de las tablas, descripción de variables y obtención de la base final. En segundo lugar, se va a realizar el análisis univariados, separando las variables continuas de las categóricas para empezar a comprender el fenómeno. En tercer lugar, se va a realizar el análisis multivariado, profundizando la comprensión del problema mediante el estudio de correlaciones y calculando pesos relativos de variables. En este punto se Presenta el Modelo sobre la base de Desarrollo. En cuarto lugar, se va a confeccionar la validación del modelo, utilizando información distinta a la de desarrollo para dar cuenta que el modelo tiene un buen desempeño. En quinto lugar, se va a poner a prueba el Modelo utilizando información muy reciente, simulando como sería el funcionamiento del modelo en la realidad, también se conoce como validación por fuera de la ventana de Tiempo. Finalmente se va a desarrollar la estrategia de implementación del modelo.

La Hipótesis a contrastar es que la utilización de múltiples variables para predecir la probabilidad de que un cliente falle y no se recupere el crédito otorgado, mejora ampliamente los ratios de mora que representan un beneficio a las entidades financieras y la sociedad en su conjunto.

Para cumplimentar el objetivo establecido y verificar la hipótesis planteada, la tesis se estructura de la siguiente forma. En el capítulo uno, se establece un marco teórico del riesgo de crédito que permita comprender las diferentes ideas que se han planteado para resolver problemas de predicción de morosidad.

En el capítulo dos, de modelos predictivos, se evaluarán algunas de las técnicas de predicción estadística más utilizadas como las Redes Neuronales, Árboles de Decisión y Regresión Logística. Esta última cobra mayor relevancia dado que es la más utilizada por las Entidades Financieras del Mundo y sus conclusiones son muy intuitivas para que puedan ser interpretadas por los ejecutivos de cuentas que analizan solicitudes de crédito.

En el capítulo tres se realiza el relevamiento y el análisis de la información que se requiere para garantizar la correcta selección de la muestra que debe ser representativa del fenómeno a estimar, caso contrario los estimadores de máxima verosimilitud del modelo no serían predictores de la mora.

En el capítulo cuatro, se realiza El análisis univariado que es la primera aproximación que se utiliza para comprender el poder predictivo individual de cada variable para estimar la variable objetivo. Es muy importante resaltar que no es una condición suficiente como metodología de eliminación de variables no predictivas, sino que permite hacer una descripción inicial del comportamiento de cada variable individual. Es bueno recordar que si bien una variable individual que no posee poder predictivo para explicar el fenómeno, en combinación con otra puede ser relevante y susceptible de ser incorporada en la regresión.

En el capítulo cinco, se realiza el análisis multivariado que es la técnica más importante para decidir que variables incorporar al modelo. Esta técnica es complementaria del análisis univariado debido a que una variable predictiva también tiene que serlo en combinación con el resto de variables. Una de las cuestiones más importantes a la hora de encontrar el modelo de ajuste más adecuado para explicar la variabilidad de una característica cuantitativa es la correcta especificación del llamado modelo teórico. En otras palabras, debemos seleccionar de entre todas las variables candidatas a ser explicativas de la variable dependiente un subconjunto que resulte suficientemente explicativo, que podemos medirlo mediante el coeficiente de determinación o con los criterios de información, para obtener el mejor subconjunto de variables explicativas. El método que se utilizará es “Paso a Paso” o “Stepwise”.

En el capítulo seis, una vez obtenido el modelo predictivo, se debe validar con una muestra que no fue utilizada para su construcción a modo de control cruzado. Esta forma de validación se la conoce como validación fuera de la ventana dado que no se utiliza información de la muestra de desarrollo, y recrea condiciones reales de nuevas solicitudes crediticias. Este último paso permite aprobar o rechazar el modelo predictivo empíricamente. En el caso de un rechazo no hay que cometer el error de corregir alguna variable del modelo para que supere esta instancia, debido a que esa corrección va a mejorar la performance en validación pero luego al aplicarlo en la realidad va a perder poder predictivo. Por este motivo las instituciones financieras mantienen de forma separada profesionales de desarrollo y validación de modelos.

La tesis finaliza con las conclusiones.

Capítulo 1: Un Marco Teórico del Riesgo de Crédito

En el presente capítulo se expondrán las distintas visiones del Riesgo de Crédito dependiendo si la institución financiera es un Banco o una Compañía de Seguros. Ambas instituciones con objetivos distintos utilizan metodologías similares que vale la pena revisar para comprender mejor el fenómeno.

La problemática del riesgo de crédito

La incertidumbre es una problemática recurrente con la que debe vivir una institución financiera. Una amplia gama de fenómenos cuyo comportamiento es impredecible tienen impacto real en el desempeño de las entidades. En el caso de las compañías de seguros, éstas tienen que realizar erogaciones por diversos siniestros asegurados. En el caso de entidades bancarias, éstas tienen que crear reservas preventivas y capital para hacer frente a pérdidas generadas en la calidad crediticia de sus clientes y por cambios en la coyuntura macroeconómica que afecten sus carteras (Elizondo & Altman, 2003).

El análisis de desvíos en los factores cuyo comportamiento es impredecible, puede ser realizado por medio de varias herramientas estadísticas, que en el caso de las compañías de seguros, han dado lugar a la teoría del riesgo. Una de las aplicaciones que realizan tradicionalmente los actuarios, es encontrar la distribución de probabilidad de pérdida, generada por los activos financieros adquiridos por un conjunto de individuos.

El desarrollo de la teoría de riesgos ha permitido a las compañías aseguradoras conocer mejor la exposición de sus carteras y estimar las pérdidas a las que se exponen. Hace muy poco tiempo que se han explotado estas herramientas en el negocio bancario debido a la similitud que existe con el riesgo más importante que las entidades enfrentan, el riesgo de crédito.

La problemática de la aprobación crediticia

Los Modelos Estadísticos para Aprobación Crediticia o “Credit Scoring” han sido extensamente estudiados y utilizados por un sin fin de instituciones financieras hace más de 50 años, mostrando resultados exitosos en todo el mundo. En términos generales, los modelos de “Credit Scoring” se pueden definir como un conjunto de métodos y técnicas estadísticas que se utilizan para predecir la probabilidad de que un cliente falle y, por ende, no se recupere el crédito otorgado por alguna institución financiera.

Durante muchos años el pronóstico y la administración del riesgo financiero estaba en manos de los analistas financieros, quienes cuantificaban riesgos solamente a través de un conjunto de reglas propias del negocio.

La historia de los modelos de “Credit Scoring” se remonta al año 1936 cuando Fisher (McLachlan, 2004) introduce la idea de discriminar diferentes grupos dentro de una población específica y desarrolla la Función Discriminante que lleva su nombre. Lugo Durand en 1941 amplía esta idea aplicada en un contexto financiero para discriminar la capacidad de impago (Durand & others, 1941).

En 1958 Bill Fair y Earl Isaac (McCorkell, 2002) comenzaron con el desarrollo de un sistema analítico que hoy en día se conoce con el nombre de FICO Score (Fair Isaac Corporation (empresa fundada en 1956) Score) una de las herramientas más usadas a nivel mundial en relación al análisis de riesgo de créditos.

En los años 60 con la creación de nuevos instrumentos financieros, las tarjetas de crédito, los modelos de “Credit Scoring” tomaron relevancia y mostraron su real importancia y utilidad (Jarrow & Turnbull, 1995).

Este tipo de modelos es superior como predictor que cualquier juicio experto cualitativo (Choquet & Meyer, 1963). Otro hito importante en este contexto fue el desarrollo del Z-Score, publicado en 1968, para predecir quiebras, ha sido

aplicado en muchas empresas del sector financiero, y en ese momento ya señalaba la necesidad y ventajas de pasar de un análisis univariado de ratios e indicadores, a otro multivariado que contempla a todos de manera integral (Altman & others, 2000).

Un punto clave de la implementación y posicionamiento de este tipo de modelos, se explica por el explosivo aumento de las solicitudes crediticias y de evaluaciones de riesgo financiero, en donde las antiguas reglas de operación quedan inoperantes ante estos nuevos requerimientos y los altos volúmenes de información (Caouette, Altman, & Narayanan, 1998). Además, si se considera que los procesos deben ser rápidos y eficientes, el grado de automatización de la toma de decisiones sobre el otorgamiento de un crédito hace fundamental la aplicación de este tipo de modelos.

En términos prácticos, los modelos de “Credit Scoring” permiten una reducción significativa en los tiempos de ejecución de los distintos procesos financieros para el otorgamiento de un crédito, permitiendo con esto una mayor automatización y reduciendo en forma drástica la necesidad de la intervención humana en la evaluación y estimación del riesgo crediticio. Los principales usuarios de este tipo de modelos son los bancos e instituciones financieras, así como las compañías de seguro. Entre las principales características que tienen en común están la posibilidad de gestionar y administrar el riesgo, en donde al manejar importantes sumas de capital, pequeñas reducciones en el riesgo de la cartera significan enormes incrementos en la rentabilidad del negocio. Los beneficios reportados por la aplicación de estos modelos no sólo afectan a los bancos e instituciones financieras, sino que directamente a todos los clientes del sector financiero, pues reduce la discriminación errónea de clientes que solicitan algún crédito y provee un análisis más objetivo y acabado de las solicitudes, siendo importante destacar la incorporación de variables, concentrando en un solo modelo múltiples factores que pueden afectar el riesgo de una solicitud.

¿Cuál es la situación actual de la Argentina?

La Comunicación “A” N° 5203 del Banco Central de la República Argentina aprueba las normas sobre “Lineamientos para la gestión de riesgos en las entidades financieras” y establece que, a partir del 2 de enero de 2012, las entidades financieras deberán contar con un proceso integral para la gestión de riesgos, que incluya la vigilancia por parte del Directorio y de la Alta Gerencia para identificar, evaluar, seguir, controlar y mitigar todos los riesgos significativos. Este proceso deberá ser proporcional a la dimensión e importancia económica de la entidad financiera de que se trate como así también a la naturaleza y complejidad de sus operaciones, teniendo en cuenta los lineamientos contenidos en esta disposición. El proceso integral para la gestión de riesgos deberá ser revisado periódicamente en función de los cambios que se produzcan en el perfil de riesgo de la entidad y en el mercado.

Por todo lo expuesto, el desarrollo de un modelo de “Credit Scoring” tiene muchas ventajas respecto a la metodología anterior y es contemporáneo a la Normativa Vigente en Argentina.

Capítulo 2: Modelos Predictivos

Ante los retos de este milenio, no sólo en las ciencias económicas, biológicas y en el área agrícola, sino también en otras ramas, se requiere una labor eficiente en la organización y desarrollo de la investigación científica y el conocimiento que ésta genera, a lo que puede contribuir, en gran parte, la aplicación consecuente de modelos estadísticos, con el apoyo de las nuevas tecnologías de la información y la comunicación. En la literatura científica y docente hay diversas publicaciones que refieren a cualidades de los modelos estadísticos de regresión, ya sean teóricas como prácticas, y se expresan sus posibilidades descriptivas, explicativas o predictivas en el contexto de una situación determinada. Para contribuir al realismo, precisión y generalidad en la aplicación de los modelos estadísticos de regresión y otros, se proponen algunos en particular que sirven para estimar probabilidades de morosidad que interesan al presente trabajo.

Distintos Tipos de Modelos Predictivos

En general, cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado siguiendo aproximaciones distintas. El número de técnicas es muy grande y sólo puede crecer en el futuro. También aquí, sin pretender ser exhaustivos, se describen algunas técnicas que sirven para estimar probabilidad de morosidad.

Redes neuronales: Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra (Haykin, Haykin, Haykin, & Haykin, 2009).

Arboles de decisión: Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los deudores, a partir de datos históricos almacenados.

Su principal ventaja es la facilidad de interpretación. Categorizados como aprendizaje basado en similaridades, los árboles de decisión son uno de los algoritmos más sencillos y fáciles de implementar y a su vez de los más poderosos (Linoff & Berry, 2011). Este algoritmo genera un árbol de decisión de forma recursiva al considerar el criterio de la mayor proporción de ganancia de información (gain ratio), es decir, elige el atributo que mejor clasifica a los datos en base a un test de hipótesis chi-cuadrado (López, 2007).

Regresión Logística: Es el principal método que se utiliza en la estimación de morosidad debido a que es un problema de discriminación entre dos poblaciones. Una forma de abordar el problema es definir una variable de clasificación, y que tome el valor cero cuando el elemento pertenece a la primera población, “No Morosos”, y uno cuando pertenece a la segunda, “Morosos” (Agresti & Kateri, 2011).

El modelo Logístico se aplica a una amplia gama de situaciones donde las variables explicativas no tienen una distribución conjunta normal multivariante (Peña, 2002). Por ejemplo, si algunas son categóricas, podemos introducirlas en el modelo Logístico mediante variables ficticias como se hace en el modelo de regresión estándar. Una ventaja adicional de este modelo es que si las variables son normales verifican el modelo Logístico.

Metodología y técnicas a utilizar

Para realizar la estimación, se utilizará una muestra compuesta por solicitudes de crédito aprobadas con toda la información de mora, datos del sistema financiero como tenencias de producto, límites de crédito en otras entidades, y todo lo referido al comportamiento financiero de un cliente. La muestra se dividirá en dos partes “Desarrollo” y “Validación”, la parte de “Desarrollo” sirve para construir la fórmula de asignación de probabilidades, “Validación” sirve para controlar que el modelo de “Credit Scoring” tenga poder predictivo[1].

[1] Las Fórmulas 1 y 2 se obtuvieron de (Hosmer Jr & Lemeshow, 2004).

El enfoque a elegido es el Cuantitativo, utilizando el Análisis Estadístico de Regresión Logística como Técnica Relevante. Esta metodología consiste en hallar el conjunto de parámetros tales que maximicen la probabilidad de observar las realizaciones obtenidas si el modelo subyacente real que determina la mora fuera el especificado. Inicialmente, el modelo construye una puntuación como combinación lineal de posibles variables explicativas (el subíndice "i" representa a cada una de ellas y el "0" al término independiente o constante):

$$Puntuación = \beta_0 + \sum_i \beta_i x_i \quad (1)$$

Luego, la probabilidad de impago (PD) de cada cliente se computa en base a la puntuación obtenida con la siguiente transformación:

$$PD = \frac{1}{1+e^{-Puntuación}} \quad (2)$$

Las regresiones se llevan a cabo en base a realizaciones binomiales, en donde se considera si el cliente entró o no entró en mora.

Un atractivo de los modelos logísticos, es la directa interpretación de los coeficientes que se aplican a las variables involucradas, ya que el signo del ponderador (+/-) da cuenta de la lógica económica de la variable.

Capítulo 3: Relevamiento y Análisis de Información

El análisis exploratorio de información es fundamental para que el ajuste del modelo sea apropiado. Este análisis permite identificar aquellas fuentes de información externas e internas como potenciales variables predictivas, y seleccionar el subconjunto de datos representativo del fenómeno de impago que se estudiará en el presente trabajo.

Fuentes de Información Internas de la Entidad Financiera

Río Platense S.A. es una Entidad Financiera ficticia que opera en un mercado similar al argentino. Se recibió la base de datos "Información de Clientes" que contiene variables Demográficas y Saldos en Cuenta que consta de la siguiente información:

- **Número de observaciones**: 188.185
- **Variables**: 3
- **Listado alfabético de variables y atributos**:

Tabla 1: Nombre y Formato de Variables "Información de Clientes"

	Variable	Tipo	Longitud
1	TRAMITE	Clave	9
2	CONT_EDAD	Continua	2
3	CONT_SDOS	Continua	4

Fuente: Elaboración propia.

En donde "TRAMITE" es el código que identifica al cliente de "Río Platense S.A.", "CONT_EDAD" es la edad del cliente expresada en años y "CONT_SDOS" es el Saldo Promedio Mensual del Pasivo del Cliente.

Fuentes de Información Externas a la Entidad Financiera

Se recibió la base de datos “Información de Bureau” que contiene variables Comportamentales del Sistema Financiero:

- **Número de observaciones**: 188.185
- **Variables**: 6
- **Listado alfabético de variables y atributos**:

Tabla 2: Nombre y Formato de Variables “Información de Bureau”

	Variable	Tipo	Longitud
1	TRAMITE	Clave	9
2	CONT_RS	Continua	3
3	CONT_ANTLABMES	Continua	3
4	BINA_BANCARIZADO	Binaria	1
5	BINA_GPOPC	Binaria	1
6	CONT_EXIG	Continua	6

Fuente: Elaboración propia.

En donde “TRAMITE” es el código que identifica al cliente de “Río Platense S.A.”, “CONT_RS” es un coeficiente comportamental del sistema financiero que oscila entre 0 y 999, a medida que se acerca a 999 tiene mejor comportamiento en el sistema financiero, “CONT_ANTLABMES” es la antigüedad laboral en meses, “BINA_BANCARIZADO” es una marca que indica si el cliente está bancarizado en otra entidad, “BINA_GPOPC” indica si al menos una de las entidades financieras a las que pertenece el cliente forma comparte similares políticas crediticias con “Río Platense S.A.” y “CONT_EXIG” es el monto de Deuda Exigible informado al Banco Central de la República Argentina (BCRA) de todas las otras entidades financieras en las que opera el cliente.

Construcción de la Variable Objetivo del Modelo (Target)

El Objetivo o Target del Modelo es la Variable que se busca predecir, el caso de “Río Platense S.A.” se busca predecir Morosidad mayor a 90 días, para eso se utilizó la base de datos “Información de Morosidad” que contiene dicha información.

- **Número de observaciones**: 188.185
- **Variables**: 2
- **Listado alfabético de variables y atributos:**

Tabla 3: Nombre y Formato de la Variable Objetivo "Mora"

	Variable	Tipo	Longitud
1	TRAMITE	Clave	9
2	MORA	Binaria	1

Fuente: Elaboración propia.

Información de Solicitudes Rechazados en "Río Platense S.A." y Aprobadas en Otras Entidades

Se recibió la base de datos "Información de Solicitudes Rechazadas" que contiene información de Clientes que Solicitaron Productos Crediticios en "Río Platense S.A." pero fueron Rechazadas y en otras Entidades fueron Aprobadas. Esta información es muy valiosa debido a que en general no se conoce que hubiera sucedido con dicha Solicitud si hubiera sido Aprobada. Las variables son las mismas que figuran en las tablas anteriores.

- **Número de observaciones**: 61.186
- **Variables**: 9
- **Listado alfabético de variables y atributos:**

Tabla 4: Nombre y Formato de Variables "Información de Solicitudes Rechazadas"

	Variable	Tipo	Longitud
1	TRAMITE	Clave	9
2	MORA	Binaria	1
3	CONT_EDAD	Continua	2
4	CONT_SDOS	Continua	4
5	CONT_RS	Continua	3
6	CONT_ANTLABMES	Continua	3
7	BINA_BANCARIZADO	Binaria	1
8	BINA_GPOPC	Binaria	1
9	CONT_EXIG	Continua	6

Fuente: Elaboración propia.

Unión de Tablas y Obtención del Dataset Definitivo

Como fue mencionado a lo largo de este documento, se trabajaron distintas bases de datos con el fin de obtener la mejor información posible para realizar una predicción, y para facilitar el análisis se unificó toda la información en una sola base de datos “Dataset”. Adicionalmente, se dividió aleatoriamente la información para Desarrollar el Modelo y para Validar el poder Predictivo.

- **Número de observaciones**: 249.371
- **Variables**: 10
- **Listado alfabético de variables y atributos:**

Tabla 5: Nombre y Formato de Variables “Dataset”

	Variable	Tipo	Longitud
1	TRAMITE	Clave	9
2	BASE	Binaria	1
3	MORA	Binaria	1
4	CONT_EDAD	Continua	2
5	CONT_SDOS	Continua	4
6	CONT_RS	Continua	3
7	CONT_ANTLABMES	Continua	3
8	BINA_BANCARIZADO	Binaria	1
9	BINA_GPOPC	Binaria	1
10	CONT_EXIG	Continua	6

Fuente: Elaboración propia.

Las variables del renglón 1 y del 3 al 10 fueron descriptas en los párrafos anteriores, la variable “BASE” indica si el cliente corresponde a la información para Desarrollo “0” o para Validación “1”.

Conclusiones del Relevamiento y Análisis de Información

Se realizó el armado de información, quitando registros duplicados de la base de datos. También se integró información de bureau para enriquecer datos externos a la entidad y se generó una marca para identificar la Base del Desarrollo y la Base para la Validación del Modelo Predictivo.

Capítulo 4: Análisis Univariado

Con el objetivo de seleccionar las variables candidatas a formar parte del modelo predictivo de admisión, se analizará univariadamente las variables incluidas en la Base de Datos “Dataset”.

El set de variables candidatas del modelo está conformado por:

Edad: Representada por el campo “CONT_EDAD”, corresponde a la edad del cliente al momento de la observación. Dicha variable, identificada como continua, fue categorizada de forma tal que permita mejorar el poder predictivo. El caso que mejor discrimina es la transformación binaria de la variable utilizando como corte a mayores o iguales a 30 años “BINA_EDAD30”.

Categoría 1: Hasta 31 años de edad inclusive.
Categoría 2: Desde 31 años hasta 35 años inclusive.
Categoría 3: Desde 35 años hasta 39 años inclusive.
Categoría 4: Desde 39 años hasta 42 años inclusive.
Categoría 5: Desde 42 años hasta 46 años inclusive.
Categoría 6: Desde 46 años hasta 50 años inclusive.
Categoría 7: Desde 50 años hasta 54 años inclusive.
Categoría 8: Desde 54 años hasta 58 años inclusive.
Categoría 9: Desde 58 años hasta 63 años inclusive.
Categoría 10: Desde 63 años de edad.

Antigüedad Laboral: Representada por el campo “CONT_ANTLABMES” y corresponde a la antigüedad laboral del cliente. Dicha variable, identificada como continua, fue categorizada en función a criterios de segmentación estadística, deciles, para los cuales fueron observadas distintas tasas de mora empírica.

Categoría 1: Sin Antigüedad.
Categoría 2: Con Antigüedad hasta 14 meses inclusive.
Categoría 3: Desde 14 meses hasta 32 meses inclusive.

Categoría 4: Desde 32 meses hasta 46 meses inclusive.
Categoría 5: Desde 46 meses hasta 59 meses inclusive.
Categoría 6: Desde 59 meses hasta 74 meses inclusive.
Categoría 7: Desde 74 meses hasta 100 meses inclusive.
Categoría 8: Desde 100 meses hasta 142 meses inclusive.
Categoría 9: Desde 142 meses hasta 177 meses inclusive.
Categoría 10: Desde 177 meses.

Score Comportamental: Representado por el campo "CONT_RS", corresponde al score de comportamiento financiero. Dicha variable, identificada como continua, fue categorizada en función a criterios de segmentación estadística, deciles, para los cuales fueron observadas distintas tasas de mora empírica.

Categoría 1: Hasta 562 puntos de score inclusive.
Categoría 2: Desde 562 puntos hasta 687 puntos inclusive.
Categoría 3: Desde 687 puntos hasta 782 puntos inclusive.
Categoría 4: Desde 782 puntos hasta 837 puntos inclusive.
Categoría 5: Desde 837 puntos hasta 869 puntos inclusive.
Categoría 6: Desde 869 puntos hasta 892 puntos inclusive.
Categoría 7: Desde 892 puntos hasta 911 puntos inclusive.
Categoría 8: Desde 911 puntos hasta 928 puntos inclusive.
Categoría 9: Desde 928 puntos hasta 945 puntos inclusive.
Categoría 10: Desde 945 puntos de score.

Saldos Promedios en Cuantas Vista: Representados por el campo "CONT_SDOS" y corresponde al promedio de los últimos tres meses de las cuentas vista del cliente. Dicha variable, identificada como continua, fue categorizada en función a criterios de segmentación estadística, deciles, para los cuales fueron observadas distintas tasas de mora empírica.

Categoría 1: Hasta 795 pesos promedio inclusive.
Categoría 2: Desde 795 pesos hasta 1302 pesos inclusive.
Categoría 3: Desde 1302 pesos hasta 1810 pesos inclusive.
Categoría 4: Desde 1810 pesos hasta 2305 pesos inclusive.

Categoría 5: Desde 2305 pesos hasta 2800 pesos inclusive.
Categoría 6: Desde 2800 pesos hasta 3301 pesos inclusive.
Categoría 7: Desde 3301 pesos hasta 3802 pesos inclusive.
Categoría 8: Desde 3802 pesos hasta 4303 pesos inclusive.
Categoría 9: Desde 4303 pesos hasta 4805 pesos inclusive.
Categoría 10: Desde 4805 pesos promedio.

Deuda Exigible en el Sistema Financiero: Representado por el campo "CONT_EXIG" y corresponde al monto de deuda neta informado al Banco Central de la República Argentina (BCRA) por el resto de las entidades. Dicha variable, identificada como continua, fue categorizada en función a criterios de segmentación estadística, deciles, para los cuales fueron observadas distintas tasas de mora empírica.

Categoría 1: Hasta 200 pesos de deuda exigible inclusive.
Categoría 2: Desde 200 pesos hasta 240 pesos inclusive.
Categoría 3: Desde 240 pesos hasta 386 pesos inclusive.
Categoría 4: Desde 386 pesos hasta 579 pesos inclusive.
Categoría 5: Desde 579 pesos hasta 843 pesos inclusive.
Categoría 6: Desde 843 pesos hasta 1237 pesos inclusive.
Categoría 7: Desde 1237 pesos hasta 1982 pesos inclusive.
Categoría 8: Desde 1982 pesos de deuda exigible.

Bancarizado: Representada por el campo "BINA_BANCARIZADO", indica si tiene productos en el resto de las entidades financieras. Dicha variable está identificada como binaria.

Categoría 1: Los Clientes No Bancarizados figuran en la tabla con el número "0".
Categoría 2: Los Clientes Bancarizados figuran en la tabla con el número "1".

Entidades Financieras Similares en Políticas Crediticias: Representados por el campo "BINA_GPOPC", indica si el cliente tiene productos crediticios en al menos una Entidad Financiera cuya Política Crediticia sea similar a "Río Platense S.A.". Dicha variable está identificada como binaria.

Categoría 1: Los Clientes que no pertenecen a Entidades Financieras con Políticas Crediticias Similares figuran en la tabla con el número “0”.
Categoría 2: Los Clientes que pertenecen a Entidades Financieras con Políticas Crediticias Similares figuran en la tabla con el número “1”.

Selección de Variables a partir del Análisis Univariado

Una vez seleccionada la base de datos de desarrollo de los modelos y detectadas las variables que la componen, se efectuó el análisis univariado, con el objetivo de conocer su contenido detallado, en particular la relación de cada variable con la variable dependiente “MORA”.

De esta manera se busca probar:

- El sentido económico de la tendencia de la morosidad de cada variable, que puede ser:
    - Creciente: Si un mayor valor de la variable implica una mayor morosidad.
    - Decreciente: Si un mayor valor de la variable implica una menor morosidad.
    - Indeterminada: Si no existe una relación entre la variable y la mora.

  En este sentido, se observó el signo del coeficiente (en el caso de variables continuas) o la relación de magnitudes entre los coeficientes de diferentes categorías (en el caso de variables categóricas).

- El poder discriminante univariante de cada variable:

  Hay varias medidas que se utilizan normalmente para este fin: la distancia de Kolmogorov-Smirnov (Lilliefors, 1967), la curva ROC (Hanley & McNeil, 1982) y el índice de poder, comúnmente conocido como PowerStat (Dickey & Fuller, 1981). Para simplificar la cantidad de medidas se utilizará KS (solo en los análisis univariados).

A continuación, se describe brevemente los aspectos metodológicos de la distancia KS:

Distancia Kolmogorov-Smirnov (KS): La distancia de KS es aquella que mide la diferencia entre dos distribuciones de probabilidad. Es utilizada habitualmente, entre otros fines, para medir la adecuación de una distribución teórica a una frecuencia acumulada empírica, como insumo para un test de bondad de ajuste.
En modelos de estimación de probabilidad de default, la distancia KS es definida como la mayor distancia entre las distribuciones acumuladas de defaults y no-defaults. Conceptualmente, un modelo que refleje adecuadamente la situación de morosidad de la cartera, debería acumular rápidamente los no-defaults y más lentamente los defaults, de modo tal que la diferencia, entre ambas distribuciones, sea grande. En general, cuanto mayor sea el valor de la distancia KS, el modelo tendrá un mayor poder discriminante entre defaults y no-defaults.

Resultados del Análisis Univariado

A continuación se presentarán resultados de pruebas univariadas realizadas sobre base de desarrollo que arrojó aleatoriamente 124.517 clientes (49,93%).

Se dividió el análisis entre variables continuas y categóricas. Para las primeras, fueron analizados los siguientes conceptos:

- Tasa de mora por decil de la variable analizada
- Signo del coeficiente obtenido en la regresión univariada.
- Distancia KS.

Para las variables categóricas, se analizaron los siguientes conceptos:

- Tasa de mora por categoría.
- Distancia KS.

Variables Continuas

Edad (CONT_EDAD):

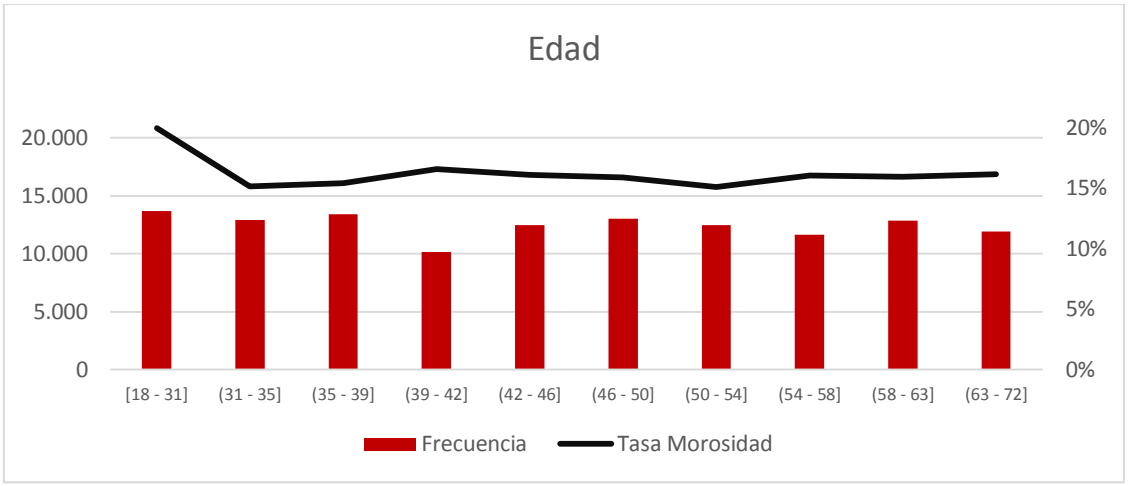
Tabla 6: Descriptiva de Variable "Edad"

CONT_EDAD	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
[18 - 31]	10.953	2.725	13.678	11%	20%	11%	13%	3,0%
(31 - 35]	10.974	1.959	12.933	10%	15%	21%	23%	2,2%
(35 - 39]	11.351	2.062	13.413	11%	15%	32%	33%	1,5%
(39 - 42]	8.469	1.678	10.147	8%	17%	40%	42%	1,7%
(42 - 46]	10.467	2.008	12.475	10%	16%	50%	52%	1,6%
(46 - 50]	10.962	2.071	13.033	10%	16%	61%	62%	1,3%
(50 - 54]	10.568	1.877	12.445	10%	15%	71%	71%	0,4%
(54 - 58]	9.755	1.860	11.615	9%	16%	80%	80%	0,3%
(58 - 63]	10.814	2.048	12.862	10%	16%	90%	90%	0,1%
(63 - 72]	9.992	1.924	11.916	10%	16%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS 3,0%***	

Fuente: Elaboración propia.

La Tabla 6 muestra información de la variable "Edad" abierta por Deciles. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

Gráfico 1: Frecuencias y Tasa de Morosidad de Variable "Edad"



Fuente: Elaboración propia.

- Signo del coeficiente: Negativo (posee sentido económico).
- Distancia KS: 3,0%.

Antigüedad Laboral (CONT_ANTLABMES):

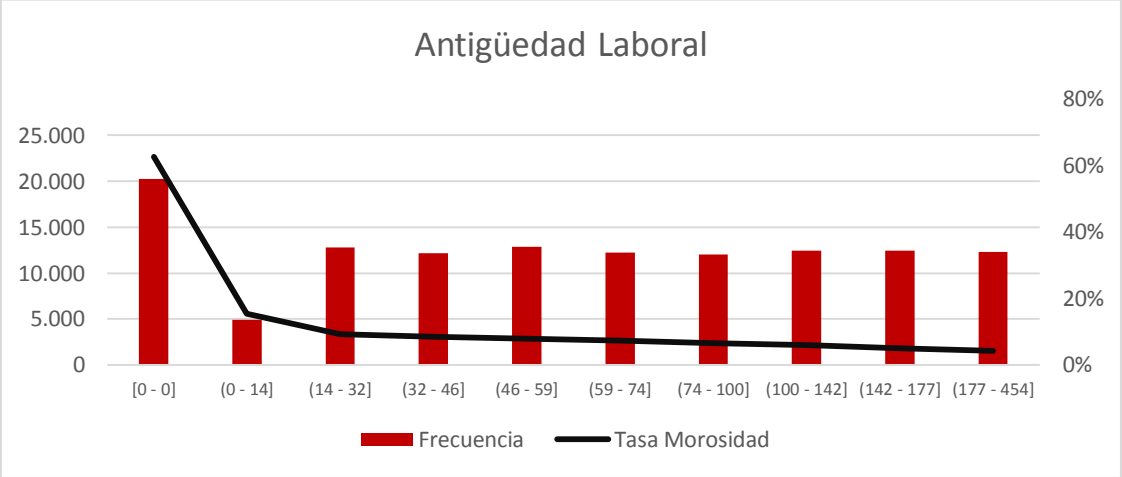
Tabla 7: Descriptiva de Variable "Antigüedad Laboral"

CONT_ANTLABMES	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
[0 - 0]	7.614	12.644	20.258	16%	62%	7%	63%	55,3%
(0 - 14]	4.149	754	4.903	4%	15%	11%	66%	55,0%
(14 - 32]	11.613	1.183	12.796	10%	9%	22%	72%	49,7%
(32 - 46]	11.118	1.035	12.153	10%	9%	33%	77%	44,2%
(46 - 59]	11.872	1.005	12.877	10%	8%	44%	82%	37,8%
(59 - 74]	11.370	886	12.256	10%	7%	55%	87%	31,3%
(74 - 100]	11.254	798	12.052	10%	7%	66%	91%	24,4%
(100 - 142]	11.707	755	12.462	10%	6%	77%	94%	16,9%
(142 - 177]	11.822	626	12.448	10%	5%	89%	97%	8,7%
(177 - 454]	11.786	526	12.312	10%	4%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS 55,3%***	

Fuente: Elaboración propia.

La Tabla 7 muestra información de la variable "Antigüedad Laboral" abierta por Deciles. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

Gráfico 2: Frecuencias y Tasa de Morosidad de Variable "Antigüedad Laboral"



Fuente: Elaboración propia.

- Signo del coeficiente: Negativo (posee sentido económico).
- Distancia KS: 55,3%.

Score Comportamental (CONT_RS):

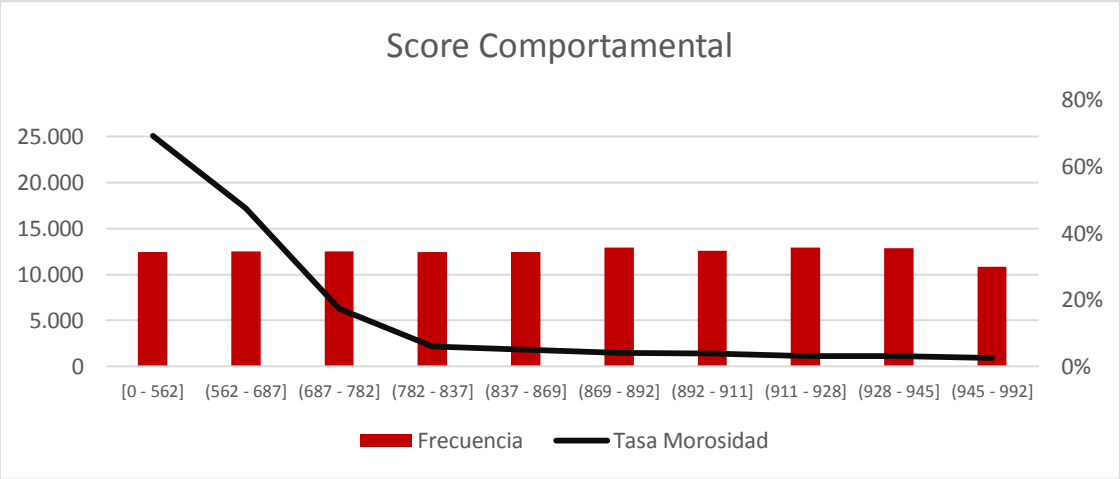
Tabla 8: Descriptiva de Variable "Score Comportamental"

CONT_RS	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
[0 - 562]	3.828	8.653	12.481	10%	69%	4%	43%	39,1%
(562 - 687]	6.560	5.954	12.514	10%	48%	10%	72%	62,3%
(687 - 782]	10.350	2.156	12.506	10%	17%	20%	83%	63,1%
(782 - 837]	11.684	748	12.432	10%	6%	31%	87%	55,6%
(837 - 869]	11.804	614	12.418	10%	5%	42%	90%	47,3%
(869 - 892]	12.386	529	12.915	10%	4%	54%	92%	38,0%
(892 - 911]	12.097	477	12.574	10%	4%	66%	95%	28,8%
(911 - 928]	12.536	402	12.938	10%	3%	78%	97%	18,7%
(928 - 945]	12.466	398	12.864	10%	3%	90%	99%	8,8%
(945 - 992]	10.594	281	10.875	9%	3%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS 63,1%***	

Fuente: Elaboración propia.

La Tabla 8 muestra información de la variable "Score Comportamental" abierta por Deciles. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

Gráfico 3: Frecuencias y Tasa de Morosidad de Variable "Score Comportamental"



Fuente: Elaboración propia.

- Signo del coeficiente: Negativo (posee sentido económico).
- Distancia KS: 63,1%.

Saldos Promedios en Cuantas Vista (CONT_SDOS):

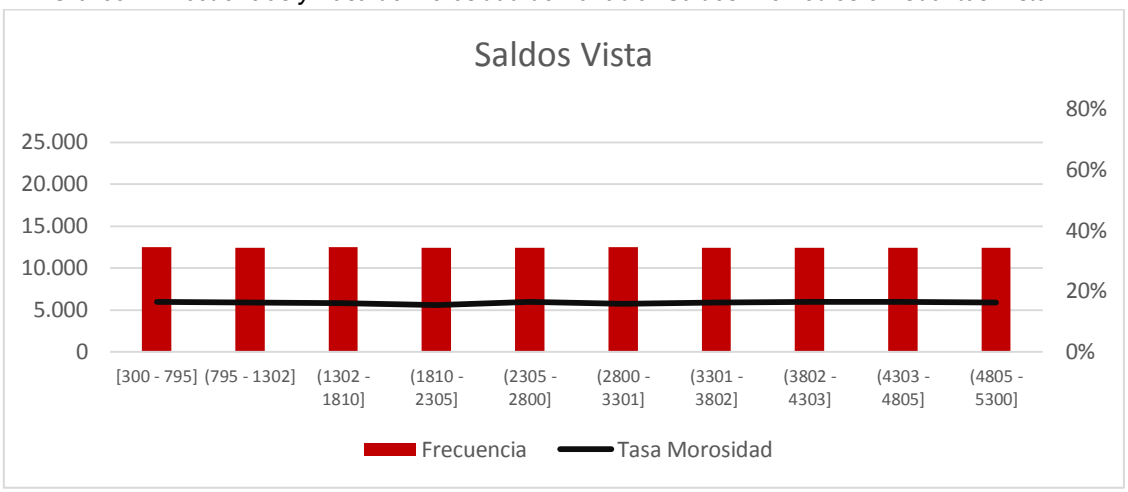
Tabla 9: Descriptiva de Variable "Saldos Promedios en Cuantas Vista"

CONT_SDOS	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
[300 - 795]	10.406	2.068	12.474	10%	17%	10%	10%	0,3%
(795 - 1302]	10.400	2.032	12.432	10%	16%	20%	20%	0,3%
(1302 - 1810]	10.467	2.010	12.477	10%	16%	30%	30%	0,2%
(1810 - 2305]	10.514	1.922	12.436	10%	15%	40%	40%	0,3%
(2305 - 2800]	10.407	2.043	12.450	10%	16%	50%	50%	0,2%
(2800 - 3301]	10.483	1.987	12.470	10%	16%	60%	60%	0,4%
(3301 - 3802]	10.414	2.024	12.438	10%	16%	70%	70%	0,4%
(3802 - 4303]	10.377	2.062	12.439	10%	17%	80%	80%	0,1%
(4303 - 4805]	10.413	2.044	12.457	10%	16%	90%	90%	0,0%
(4805 - 5300]	10.424	2.020	12.444	10%	16%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS***	***0,4%***

Fuente: Elaboración propia.

La Tabla 9 muestra información de la variable "Saldos Promedios en Cuantas Vista" abierta por Deciles. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

Gráfico 4: Frecuencias y Tasa de Morosidad de Variable "Saldos Promedios en Cuantas Vista"



Fuente: Elaboración propia.

- Signo del coeficiente: Pendiente muy cercana a cero (variable no predictiva).
- Distancia KS: 0,4%.

Deuda Exigible en el Sistema Financiero (CONT_EXIG):

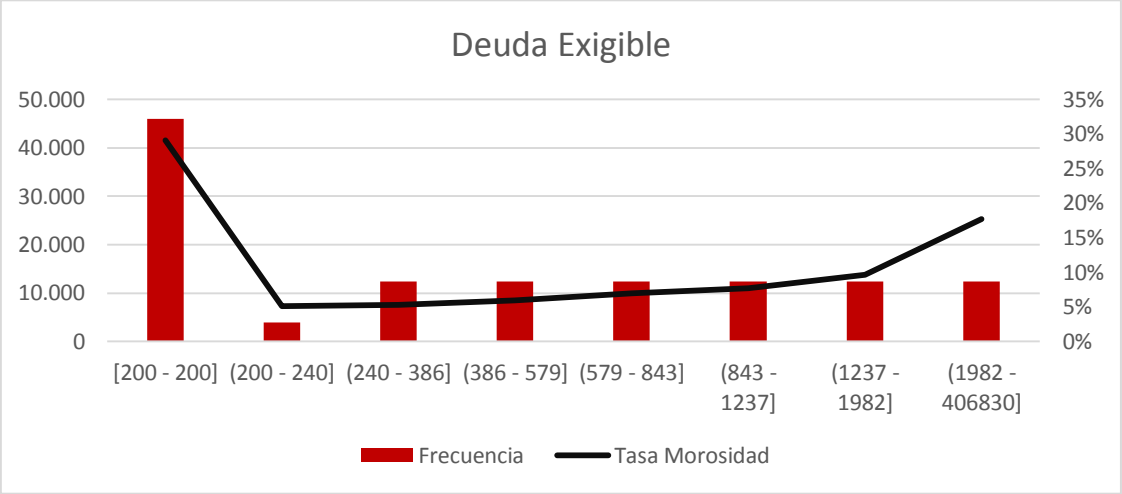
Tabla 10: Descriptiva de Variable "Deuda Exigible en el Sistema Financiero"

CONT_EXIG	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
[200 - 200]	32.575	13.376	45.951	37%	29%	31%	66%	34,9%
(200 - 240]	3.732	200	3.932	3%	5%	35%	67%	32,4%
(240 - 386]	11.773	660	12.433	10%	5%	46%	70%	24,3%
(386 - 579]	11.667	739	12.406	10%	6%	57%	74%	16,8%
(579 - 843]	11.591	874	12.465	10%	7%	68%	78%	10,0%
(843 - 1237]	11.489	963	12.452	10%	8%	79%	83%	3,8%
(1237 - 1982]	11.237	1.198	12.435	10%	10%	90%	89%	1,1%
(1982 - 406830]	10.241	2.202	12.443	10%	18%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS***	***34,9%***

Fuente: Elaboración propia.

La Tabla 10 muestra información de la variable "Deuda Exigible en el Sistema Financiero" abierta por Deciles. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

Gráfico 5: Frecuencias y Tasa de Morosidad de Variable "Deuda Exigible en el Sistema Financiero"



Fuente: Elaboración propia.

- Signo del coeficiente: Positiva (posee sentido económico).
- Distancia KS: 34,9%.

Variables Categóricas

Bancarizado (BINA_BANCARIZADO):

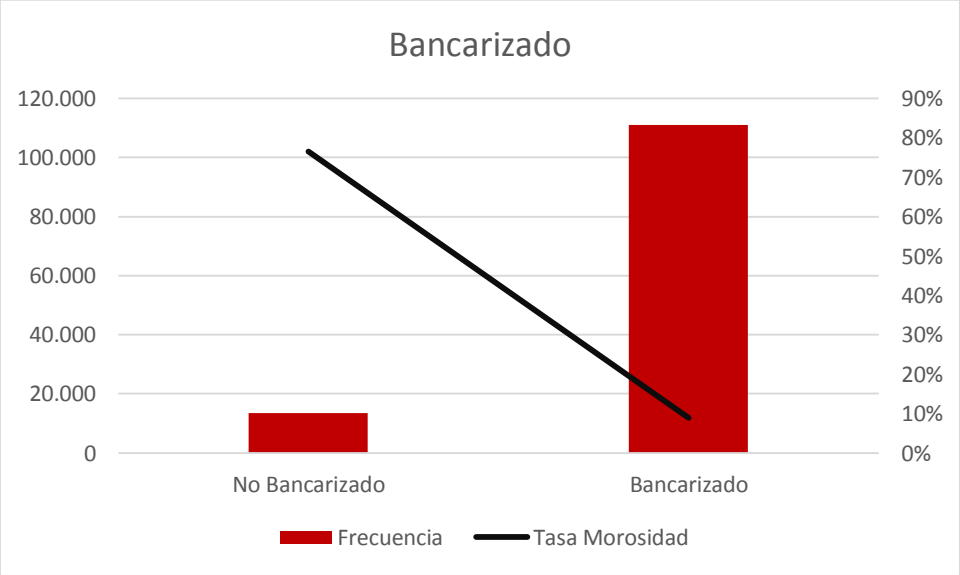
Tabla 11: Descriptiva de Variable "Bancarizado"

BINA_BANCARIZADO	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
No Bancarizado	3.142	10.278	13.420	11%	77%	3%	51%	47,8%
Bancarizado	101.163	9.934	111.097	89%	9%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS***	***47,8%***

Fuente: Elaboración propia.

La Tabla 11 muestra información de la variable "Bancarizado" por categoría. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

Gráfico 6: Frecuencias y Tasa de Morosidad de Variable "Bancarizado"



Fuente: Elaboración propia.

- Distancia KS: 47,8%.

Entidades Financieras Similares en Políticas Crediticias (BINA_GPOPC):

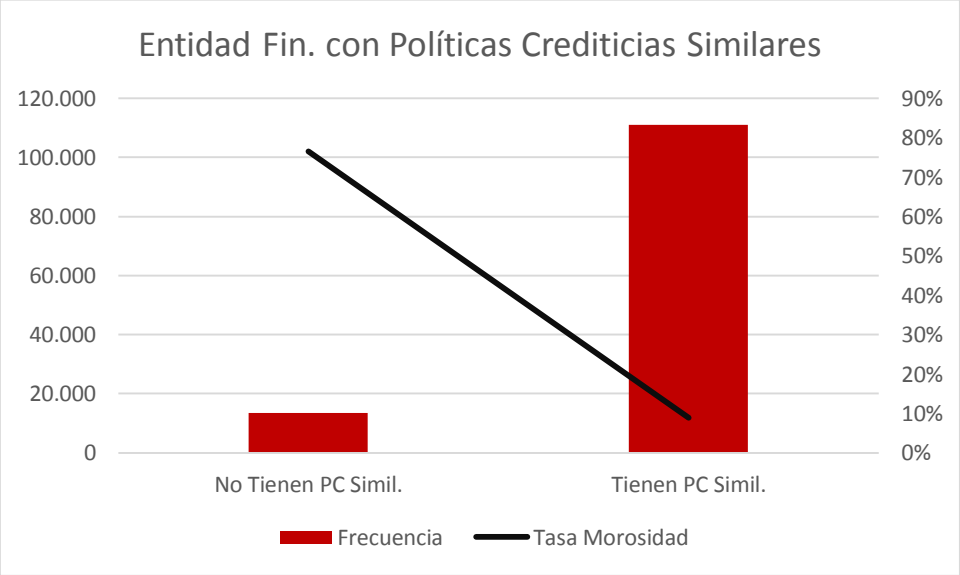
Tabla 12: Descriptiva de Variable "Entidades Financieras Similares en Políticas Crediticias"

BINA_GPOPC	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
No Tienen PC Simil.	3.142	10.278	13.420	11%	77%	3%	51%	47,8%
Tienen PC Simil.	101.163	9.934	111.097	89%	9%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS***	***47,8%***

Fuente: Elaboración propia.

La Tabla 12 muestra información de la variable "Entidades Financieras Similares en Políticas Crediticias" por categoría. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

Gráfico 7: Frecuencias y Tasa de Morosidad de Variable "Entidades Financieras Similares en Políticas Crediticias"



Fuente: Elaboración propia.

- Distancia KS: 47,8% (es igual que la variable Bancarizado, se elimina por estar repetida).

Edad Categorizada

Tabla 13: Descriptiva de Variable "Edad Categorizada"

Corte Edad con Mayor KS	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
Menores a 25	1.281	691	1.972	2%	35%	1%	3%	2,2%
26	1.010	297	1.307	1%	23%	2%	5%	2,7%
27	1.244	282	1.526	1%	18%	3%	6%	2,9%
28	1.518	337	1.855	1%	18%	5%	8%	3,1%
29	1.662	343	2.005	2%	17%	6%	10%	3,2%
30	2.012	375	2.387	2%	16%	8%	12%	3,1%
31	2.226	400	2.626	2%	15%	11%	13%	3,0%
32	2.482	450	2.932	2%	15%	13%	16%	2,8%
33	2.739	467	3.206	3%	15%	16%	18%	2,5%
Mayores a 34	88.131	16.570	104.701	84%	16%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS 3,2%***	

Fuente: Elaboración propia.

La Tabla 13 muestra información de la variable "Edad Categorizada" por categoría. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

La Edad, maximiza KS en 30 años (Distancia KS 3,2%), si bien es despreciable el poder discriminante, el negocio solicitó la incorporación al modelo.

Edad Binaria (BINA_EDAD30):

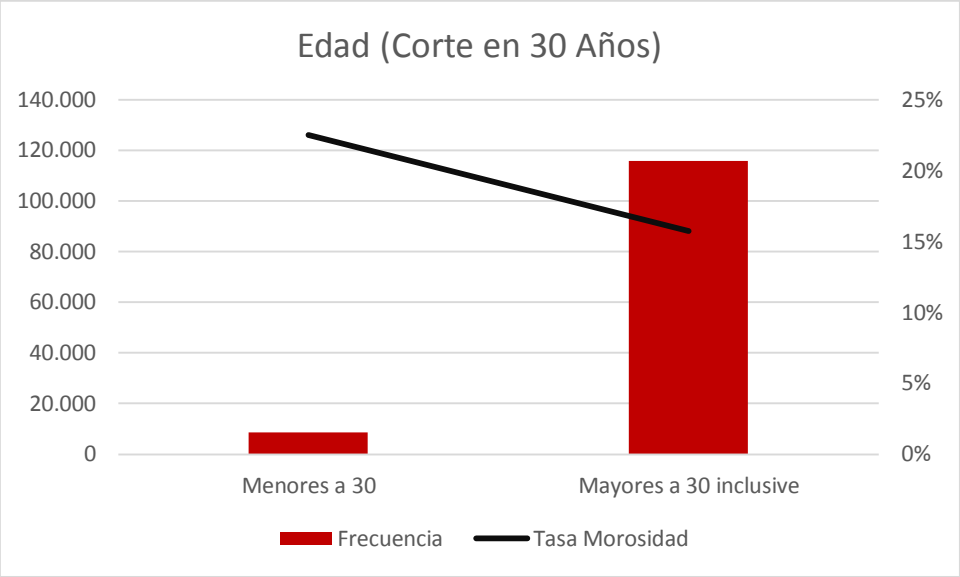
Tabla 14: Descriptiva de Variable "Edad Categorizada en 35 años"

BINA_EDAD30	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
Menores a 30	6.715	1.950	8.665	7%	23%	6%	10%	3,2%
Mayores a 30 inclusive	97.590	18.262	115.852	93%	16%	100%	100%	0,0%
Total	**104.305**	**20.212**	**124.517**				***Distancia KS 3,2%***	

Fuente: Elaboración propia.

La Tabla 14 muestra información de la variable "Edad Categorizada en 35 años" por categoría. Cantidad de Clientes "Morosos" y "No Morosos", Frecuencias Acumuladas y Distancia KS.

Gráfico 8: Frecuencias y Tasa de Morosidad de Variable "Edad Categorizada en 35 años"



Fuente: Elaboración propia.

Conclusiones del Análisis Univariado

El objetivo de este capítulo fue seleccionar las variables candidatas a formar parte del modelo predictivo de admisión. Para ello se analizaron univariadamente las variables incluidas en la Base de Datos "Dataset".

En base a este análisis, se decidió lo siguiente: eliminar la variable "Entidades Financieras Similares en Políticas Crediticia" debido a que se encuentra repetida; sugerir eliminar la variable Saldos Promedios en Cuantas Vista por no tener poder discriminante, se verificará en el Análisis Multivariado si combinada con el resto de variables mejora su performance. Debido a que el negocio solicitó la incorporación de la variable Edad al modelo, la mejor forma de ingresarla es en dos categorías, "Menores a 30 años" y "Mayores a 30 años inclusive". Igualmente, se verificará en el Análisis Multivariado que la inclusión no reduzca significativamente la predictibilidad del modelo de score.

Capítulo 5: Análisis Multivariado

Una vez identificadas las variables que presentan un buen ordenamiento de la tasa de morosidad, se procedió a la estimación del modelo completo. Para ello se recurrió a la estimación mediante la regresión logística. Esta metodología consiste en hallar el conjunto de parámetros tales que maximicen la probabilidad de observar las realizaciones obtenidas, si el modelo subyacente real que determina la mora fuera el especificado.

Inicialmente, el modelo construye una puntuación como combinación lineal de posibles variables explicativas (el subíndice "i" representa a cada una de ellas y el "0" al término independiente o constante):

$$Puntuación = \beta_0 + \sum_i \beta_i \, x_i \quad (1)$$

Luego, la probabilidad de impago (PD) de cada cliente se computa en base a la puntuación obtenida con la siguiente transformación:

$$PD = \frac{1}{1+e^{-Puntuación}} \quad (2)$$

Las regresiones se llevan a cabo en base a realizaciones binomiales, en donde se considera si el cliente entró o no entró en mora.

Un atractivo de los modelos logísticos, es la directa interpretación de los coeficientes que se aplican a las variables involucradas, ya que el signo del ponderador (+/-) da cuenta de la lógica económica de la variable.

Metodologías de Selección de Variables

Una de las cuestiones más importantes a la hora de encontrar el modelo de ajuste más adecuado para explicar la variabilidad de una característica cuantitativa es la correcta especificación del llamado modelo teórico. En otras palabras, debemos seleccionar de entre todas las variables candidatas a ser explicativas de la variable dependiente un subconjunto que resulte

suficientemente explicativo, que podemos medirlo mediante el coeficiente de determinación o con los criterios de información, para obtener el mejor subconjunto de variables explicativas.

En la práctica, la selección del subconjunto de variables explicativas de los modelos de regresión se deja en manos de procedimientos automáticos iterativos (Burden, Faires, & Palmas, 2002).

Los procedimientos más usuales son los siguientes:

Método “Regresivo” o “Backward”: se comienza por considerar incluidas en el modelo teórico a todas las variables disponibles y se van eliminando del modelo de una en una según su capacidad explicativa. En concreto, la primera variable que se elimina es aquella que presenta un menor coeficiente de correlación parcial con la variable dependiente (equivalente a un menor valor del estadístico t), y así sucesivamente hasta llegar a una situación en la que la eliminación de una variable más suponga un descenso importante en el coeficiente de determinación.

Método “Progresivo” o “Forward”: se comienza por un modelo que no contiene ninguna variable explicativa y se añade como primera de ellas a la que presente un mayor coeficiente de correlación, en valor absoluto, con la variable dependiente. En los pasos sucesivos se va incorporando al modelo aquella variable que presenta un mayor coeficiente de correlación parcial con la variable dependiente dadas las independientes ya incluidas en el modelo. El procedimiento se detiene cuando el incremento en el coeficiente de determinación debido a la inclusión de una nueva variable explicativa en el modelo ya no es importante.

Método “Paso a Paso” o “Stepwise”: es uno de los más empleados y consiste en una combinación de los dos anteriores (Siddiqi, 2012). En el primer paso se procede como en el método forward pero a diferencia de éste en el que cuando una variable entra en el modelo ya no vuelve a salir, en el procedimiento “paso

a paso" es posible que la inclusión de una nueva variable haga que otra que ya estaba en el modelo resulte redundante y sea expulsada.

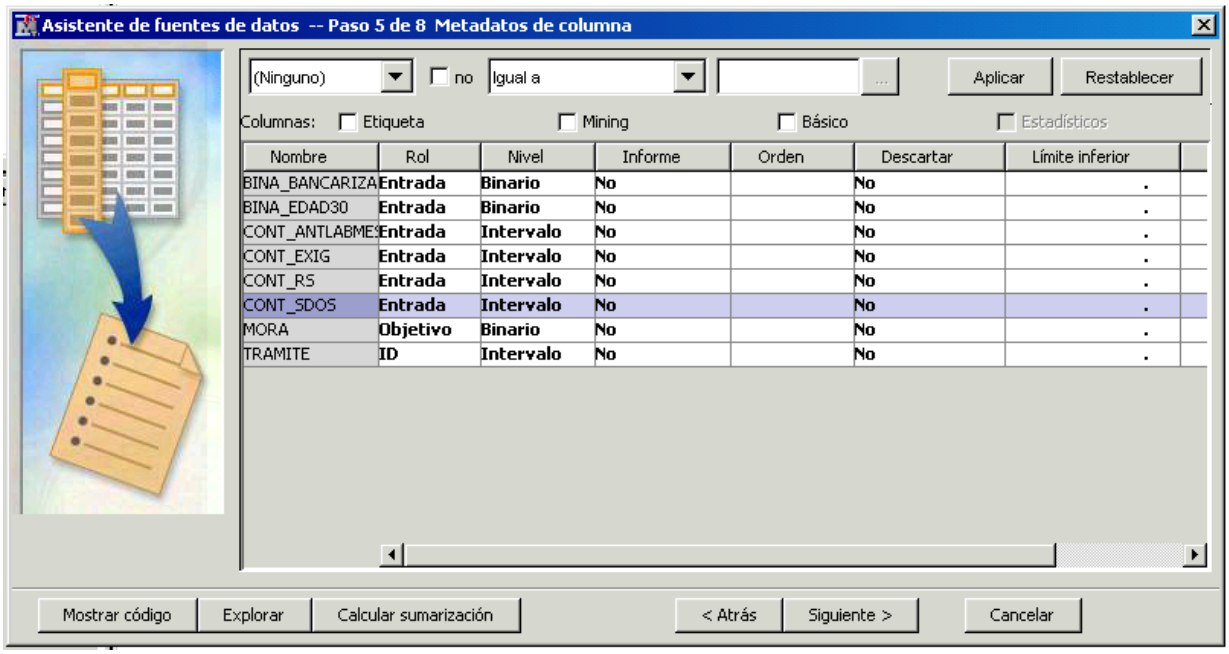
Este último método demora mayor tiempo de procesamiento para el Software, pero la posibilidad de quitar una variable que fue incluida erróneamente con un proceso multivariado ("progresivo"), o no volver a incluir una variable que fue descartada erróneamente con un proceso multivariado ("regresivo"), es el motivo por el cual se decide utilizar la metodología "paso a paso" para llevar adelante la selección de variables.

Resultados del Análisis Multivariado

A continuación se exponen los coeficientes del modelo (de todas las variables que superaron exitosamente el análisis univariado y "Saldos Vista"), que surgen del criterio de máxima verosimilitud, utilizando la regresión logística sobre la base de desarrollo, que arrojó aleatoriamente 124.517 clientes (49,93%).

En primer lugar, se definen las características de las variables que van a ser ingresadas al software estadístico:

Gráfico 9: Clasificación de Variables para ingresar a SAS Enterprise Miner



Fuete: Elaboración propia utilizando SAS Enterprise Miner.

Tabla 15: Salida de SAS Enterprise Miner luego de ejecutar una Regresión Logística

Tabla: Trazado de efecto

Variable ▲	Nivel	Coeficiente	T-valor	P valor	Variable de código de puntuación	Signo	Coeficiente absoluto	T-valor absoluto	Número TScore	Número de efecto
BINA_BANCARIZADO	0	0.796139	54.177...	0	_1_0	+	0.796139	54.17715	3	2
BINA_EDAD30	0	-0.20514	-10.613	2.59E-26	_2_0	-	0.20514	10.613	6	3
CONT_ANTLABMES		-0.00573	-25.8758	1.2E-147	CONT_ANT...	-	0.005734	25.87581	4	5
CONT_EXIG		.0001059	18.924...	7.18E-80	CONT_EXIG	+	.0001059	18.92443	5	6
CONT_RS		-0.0062	-95.7624	0	CONT_RS	-	0.006201	95.76241	1	4
CONT_SDOS		2.627E-6	0.3673...	0.71335	CONT_SD...	+	2.627E-6	0.367361	7	7
Intercept	1	3.455558	67.5818	0		+	3.455558	67.5818	2	1

Fuete: Elaboración propia utilizando SAS Enterprise Miner.

Se puede verificar que el P-Valor del Test de Hipótesis de Significatividad Individual de la variable Saldos Vista es 0,71335, muy lejos del 1% de tolerancia que se utiliza para la incorporación de la variable al modelo. El resto de las variables son Significativas Estadísticamente.

Se elimina la variable Saldo Vista en el Software y se vuelve a ejecutar la Regresión Logística.

Gráfico 10: Salida de SAS Enterprise Miner que muestra cómo se elimina una variable

Fuete: Elaboración propia utilizando SAS Enterprise Miner.

Tabla 16: Salida de SAS Enterprise Miner luego de ejecutar la segunda Regresión Logística

Tabla: Trazado de efectos

Variable ▲	Nivel	Coeficiente	T-valor	P valor	Signo	Coeficiente absoluto	T-valor absoluto	Número TScore	Número de efecto
BINA_BANCARIZADO	0	0.796115	54.17621	0	+	0.796115	54.17621	3	2
BINA_EDAD30	1	-0.20518	-10.6152	2.53E-26	-	0.20518	10.61524	6	3
CONT_ANTLABMES		-0.00573	-25.8755	1.3E-147	-	0.005734	25.87547	4	5
CONT_EXIG		.0001059	18.92491	7.11E-80	+	.0001059	18.92491	5	6
CONT_RS		-0.0062	-95.763	0	-	0.006201	95.76299	1	4
Intercept	1	3.462866	73.50566	0	+	3.462866	73.50566	2	1

Fuete: Elaboración propia utilizando SAS Enterprise Miner.

A continuación se exponen los coeficientes finales que surgen de la regresión logística, utilizando el criterio de máxima verosimilitud:

Tabla 17: Coeficientes Finales que surgen de la regresión logística definitiva en desarrollo

Variable	Categoría	Coeficiente	Prueba de Significatividad Individual **
BINA_BANCARIZADO	No Bancarizados	0.7961148805699232	< 0,0001
BINA_EDAD30	Mayores a 30 años inclusive	-0.20518028901302302	< 0,0001
CONT_ANTLABMES	No Corresponde *	-0.005733616687847308	< 0,0001
CONT_EXIG	No Corresponde *	0.000105903095364481	< 0,0001
CONT_RS	No Corresponde *	-0.006201012692318997	< 0,0001
Intercept o Constante		3.462866268240780	< 0,0001

Fuete: Elaboración propia.

Cabe destacar que las variables categóricas listadas (BINA_BANCARIZADO y BINA_EDAD30), incluyen un coeficiente adicional para la última categoría, el cual es igual a cero.

Dado que los p-valores de las pruebas de significatividad individual Chi-Cuadrado son inferiores a cualquier nivel de confianza que tolera “Río Platense S.A.”, se afirma que las variables incluidas en el modelo son significativas estadísticamente[2]. Se puede analizar la lógica económica de las variables del modelo, analizando el signo y el orden de los coeficientes de cada variable:

[2] * Las pruebas de significatividad individual consisten en la evaluación de la hipótesis de nulidad de los coeficientes evaluados individualmente, contra la hipótesis alternativa de coeficientes diferentes al nulo. Por lo tanto, si se obtiene un p-valor inferior al nivel de confianza seleccionado, se afirma que existe evidencia empírica para rechazar la hipótesis nula y por lo tanto, se asume que la variable es significativa.
** Al tratarse de una variable ingresada como continua, no está dividida en categorías y por lo tanto, el coeficiente estimado es único.

- Los coeficientes para “Edad” son decrecientes, lo que implica que a mayor categoría de la variable edad, corresponde un coeficiente más bajo. Por lo tanto, la relación entre edad y probabilidad será inversa, es decir, a mayor edad, menor será la probabilidad de morosidad.

- Los coeficientes de “Antigüedad Laboral” son decrecientes, lo que implica que a mayor antigüedad laboral, corresponde un coeficiente más bajo. Por lo tanto, la relación entre antigüedad y probabilidad será inversa, es decir, a mayor antigüedad, menor será la probabilidad de morosidad.

- Los coeficientes para “Score Comportamental” son decrecientes, lo que implica que a mayor score, corresponde un valor de coeficiente más bajo. Por lo tanto, la relación entre el score y probabilidad será inversa, es decir cuanto mayor sea el score, menor será la probabilidad de morosidad.

- Los coeficientes para “Deuda Exigible” son crecientes, lo que implica que a mayor exigible, corresponde un valor de coeficiente más alto. Por lo tanto, la relación entre el exigible y probabilidad será directa, es decir cuanto mayor sea el exigible, mayor será la probabilidad de morosidad.

- Los coeficientes para “Bancarizado” son decrecientes, lo que implica que a mayor categoría de la variable bancarizado, corresponde un coeficiente más bajo. Por lo tanto, la relación entre bancarizado y probabilidad será inversa, es decir, a mayor bancarización, menor será la probabilidad de morosidad.

De esta manera, puede concluirse que los resultados obtenidos, se comportan de acuerdo al sentido económico esperado. Asimismo, el valor de los criterios de información de Akaike y Schwartz son los que se exponen a continuación:

Tabla 18: Criterios de Información

AIC	65674,28
SC	65732,68

Fuete: Elaboración propia.

Cabe destacar que la importancia de dichos coeficientes, radica en indicar la conveniencia de utilizar un modelo u otro, a partir de la comparación de dichos coeficientes, en nuestro ejemplo sirvió para validar que si se eliminaba alguna de las variables que quedaron definitivas, se empeoraba el poder predictivo.

Tanto el AIC, criterio de información de Akaike (Akaike, 1974), como SC, criterio de información de Schwarz (Schwarz & others, 1978), son estadísticos que relacionan el cociente de la verosimilitud con el número de parámetros del modelo analizado. En general, se asume que cuanto menor sea este cociente mejor será el modelo. Es decir, en la comparación entre dos modelos alternativos, debería elegirse aquel cuyo valor de AIC o SC sea menor.

Las medidas de poder de este modelo (Distancia KS y ROC), se exponen a continuación:

Tabla 19: Medidas de Poder

KS	64,9%
ROC	88,8%

Fuete: Elaboración propia.

Curva ROC: La curva ROC proporciona una representación de la exactitud de la calificación dada por el modelo. También proporciona una medida de comparación entre distintos modelos.

Para representar la curva, se definen a continuación dos conceptos, fijando un punto de corte determinado:

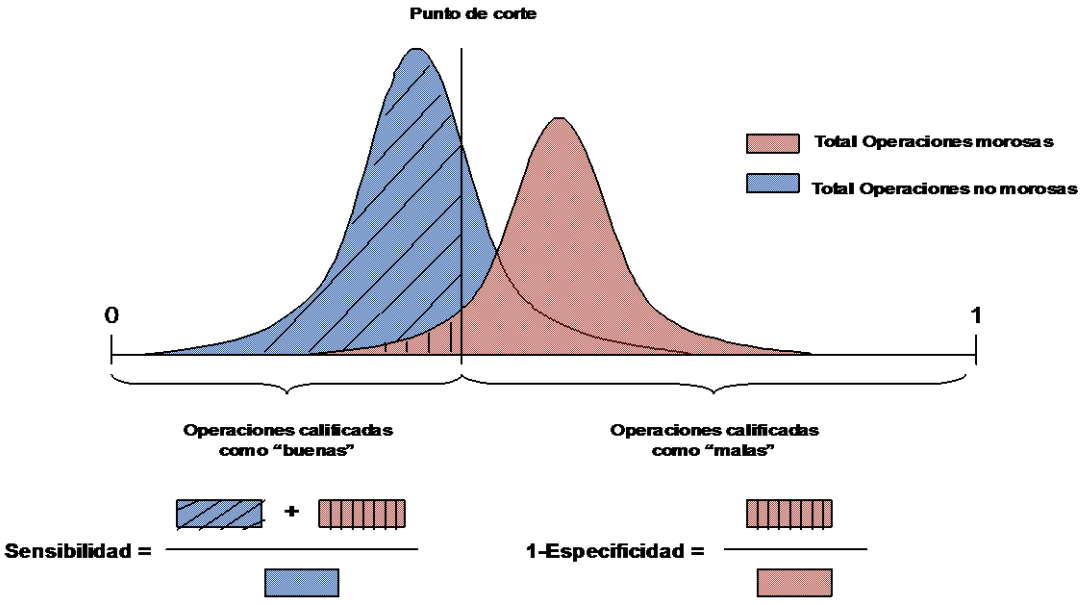
- Sensibilidad: Es la probabilidad de que el modelo clasifique correctamente a un cliente no moroso.

- Especificidad: Es la probabilidad de que el modelo clasifique correctamente a un cliente moroso.

Bajo la hipótesis de que la calificación otorgada por el modelo es un valor real que está entre 0 y 1 (caso de las regresiones logísticas), se puede establecer un punto de corte en la salida. Por ejemplo, el valor 0.5 es considerar "clientes malos" todos aquellos cuya puntuación otorgada por el modelo sea superior a 0.5, y "clientes buenos" aquellos cuya puntuación sea inferior a 0.5. Por lo tanto, la probabilidad de que el modelo califique correctamente a los clientes "buenos", fijado ese punto de corte, es el número de clientes "buenos" cuya puntuación es inferior a 0.5, entre el total de clientes "buenos".

La curva ROC se obtiene representando, para cada uno de los posibles puntos de corte, los pares de puntos (1 – Especificidad, Sensibilidad). Es decir, para cada valor del punto de corte (números reales entre cero y uno), se representa en el eje de abscisas el valor de "1 – Especificidad" (probabilidad de que el modelo no califique correctamente a un cliente moroso) y en el eje de ordenadas el valor de "Sensibilidad" (probabilidad de que el modelo califique correctamente a un cliente no moroso). Ambas distribuciones de 1-Especificidad y Sensibilidad, se representan gráficamente de la siguiente manera:

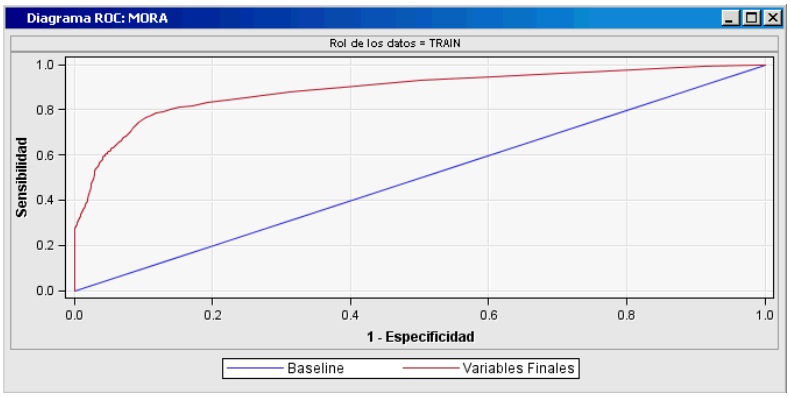
Gráfico 11: Discriminación de Distribuciones "Morosos" y "No Morosos"



Fuente: Elaboración propia en base a (Hair & Suárez, 1999).

La precisión del modelo se mide como el área que queda bajo la curva ROC. El modelo "perfecto" tendrá un área igual a 1, y el peor (aleatoriedad pura) un área igual a 0.5.

Gráfico 12: Curva ROC del Modelo Predictivo de desarrollo



Fuete: Elaboración propia utilizando SAS Enterprise Miner.

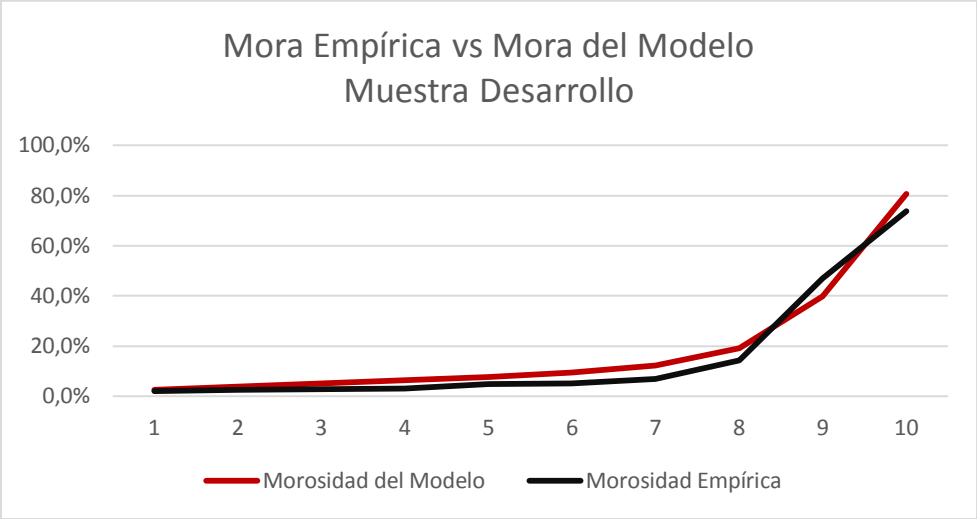
Por último, se efectuó un análisis por deciles de la variable probabilidad, comparando la probabilidad promedio de cada decil, con la tasa de mora observada. Los resultados obtenidos se exponen a continuación:

Tabla 20: Comparación de Morosidad del Modelo y Empírica

Decil	Morosidad del Modelo	Morosidad Empírica
1	2,5%	2,1%
2	3,8%	2,5%
3	5,2%	2,8%
4	6,5%	3,1%
5	7,8%	4,9%
6	9,5%	5,2%
7	12,3%	6,8%
8	19,2%	14,3%
9	39,9%	46,9%
10	80,6%	73,8%

Fuete: Elaboración propia.

Gráfico 13: Comparación de Morosidad del Modelo y Empírica



Fuete: Elaboración propia.

Puede observarse un ajuste razonable de la curva teórica del modelo, con relación a la curva empírica de tasas de mora.

Estudio de Correlaciones

Para verificar que no existen variables redundantes, es decir, variables con alto grado de correlación, se efectuó el estudio de los coeficientes de correlación de Pearson para las probabilidades calculados para cada variable.

Las mejores prácticas del negocio afirman que dos variables pueden tener los siguientes grados de correlación (en función a los coeficientes de correlación existentes entre ellas):

- Ausencia de correlación: si el coeficiente de Pearson entre ambas es inferior al 50%.
- Correlación Débil: si el coeficiente de Pearson entre ambas se ubica entre 50% y 70%.
- Correlación Fuerte: si el coeficiente de Pearson entre ambas es superior al 70%.

En general, se recomienda no incluir dos variables que presenten grados de correlación fuerte, excluyendo la de poder explicativo más bajo. Otro criterio para seleccionar la variable a excluir, consiste en calcular los valores de AIC y SC para los modelos con ambas variables, seleccionando aquel que presente criterios de información más bajos.

La matriz de correlaciones de Pearson entre las variables del modelo, fue obtenida a través de un procedimiento en el Software Estadístico SAS Enterprise Miner (DeLong, DeLong, & Clarke-Pearson, 1988).

Dicha matriz se expone a continuación:

Tabla 21: Matriz de Correlaciones de las variables explicativas del modelo

Correlaciones	Bancarizado	Edad	Score Comportamental	Anigüedad Laboral	Deuda Exigible	Políticas Crediticias
Bancarizado	1,00000	0,08490	0,33049	0,19066	0,09723	1,00000
Edad	0,08490	1,00000	0,09788	0,19333	0,03978	0,08490
Score Comportamental	0,33049	0,09788	1,00000	0,20621	0,01593	0,33049
Anigüedad Laboral	0,19066	0,19333	0,20621	1,00000	0,13145	0,19066
Deuda Exigible	0,09723	0,03978	0,01593	0,13145	1,00000	0,09723
Políticas Crediticias	1,00000	0,08490	0,33049	0,19066	0,09723	1,00000

Fuete: Elaboración propia.

Se puede corroborar que ningún par de variables posee un coeficiente de correlación superior al 50%, asumiendo entonces que entre ellas no existe correlación.

Conclusiones del Análisis Multivariado

Si bien el método de selección de variables “Paso a Paso” demora mayor tiempo de procesamiento para el Software, permite quitar una variable que fue incluida erróneamente con un proceso multivariado (“progresivo”), o no volver a incluir una variable que fue descartada erróneamente con un proceso

multivariado (“regresivo”). Este es el motivo por el cual se decide emplear la metodología “Paso a Paso” para llevar adelante la selección de variables.

Luego del estudio detallado de cada variable en la base de desarrollo, se estimaron los coeficientes para inferir la probabilidad de mora de una solicitud crediticia. Las medidas de poder son muy buenas, Distancia KS 64,9% y el índice que surge de la Curva ROC asciende a 88,8%, ambos valores son muy elevados y permiten concluir que el modelo tiene poder predictivo.

Por último, la comparación entre Morosidad del Modelo y Morosidad Empírica permite ver una diminuta diferencia entre ellas, evidenciando el buen poder predictivo que el modelo posee para inferir morosidad.

Capítulo 6: Validación del Modelo

A fin de determinar la validez del modelo en muestras diferentes, se estimarán las medidas de poder y el ajuste de la mora teórica a la empírica, sobre la población original del segmento. Esta población fue dividida aleatoriamente al inicio del desarrollo:

- Muestra de Validación: Aquellos registros de la Base de Datos "Dataset" cuyo valor del campo "BASE" es igual a "1".

A continuación se exponen los resultados obtenidos de estimar un modelo logístico en el cual la variable "PROBABILIDAD" sea la variable explicativa del modelo de default en las muestras mencionadas.

Muestra de Validación

Se obtuvo una tabla con las siguientes cifras de registros, morosos y tasas de morosidad:

Tabla 22: Descripción de datos de la muestra de validación

No Morosos	Morosos	Morosidad
104.948	19.906	15,9%

Fuete: Elaboración propia.

Para el modelo logístico recalculado, se efectuaron las pruebas de significatividad individual, y se calcularon los valores de los criterios de información de Akaike y Schwartz, así como también las medidas de poder:

Tabla 23: Salida de SAS Enterprise Miner luego de ejecutar la Regresión Logística en validación

Tabla: Trazado de efectos

Variable ▲	Nivel	Coeficiente	T-valor	P valor	Signo	Coeficiente absoluto	T-valor absoluto	Número TScore	Número de efecto
BINA_BANCARIZADO	0	0.810198	54.55017	0	+	0.810198	54.55017	3	2
BINA_EDAD30	1	-0.25051	-12.5275	5.28E-36	-	0.25051	12.52747	6	3
CONT_ANTLABMES		-0.00557	-24.9132	5.4E-137	-	0.005571	24.91316	4	5
CONT_EXIG		.0001007	18.11732	2.33E-73	+	.0001007	18.11732	5	6
CONT_RS		-0.00637	-97.0147	0	-	0.006374	97.01466	1	4
Intercept	1	3.52901	73.66675	0	+	3.52901	73.66675	2	1

Fuete: Elaboración propia utilizando SAS Enterprise Miner.

A continuación se exponen los coeficientes que surgen de la regresión logística, utilizando el criterio de máxima verosimilitud, para comparar con Desarrollo:

Tabla 24: Coeficientes Finales que surgen de la regresión logística definitiva en validación

Variable	Categoría	Coeficiente	Prueba de Significatividad Individual
BINA_BANCARIZADO	No Bancarizados	0.8101983311791262	< 0,0001
BINA_EDAD30	Mayores a 30 años inclusive	-0.2505096529638256	< 0,0001
CONT_ANTLABMES	No Corresponde	-0.005570976632219674	< 0,0001
CONT_EXIG	No Corresponde	0.000100701489870448	< 0,0001
CONT_RS	No Corresponde	-0.006374368065534111	< 0,0001
Intercept o Constante		3.5290103237290094	< 0,0001

Fuete: Elaboración propia.

Dado que los p-valores de las pruebas de significatividad individual Chi-Cuadrado son inferiores a cualquier nivel de confianza que tolera “Río Platense S.A.”, se afirma que las variables incluidas en la validación del modelo siguen siendo significativas estadísticamente.

Asimismo, el valor de los criterios de información de Akaike y Schwartz son los que se exponen a continuación:

Tabla 25: Criterios de Información

AIC	64069,39
SC	64127,80

Fuete: Elaboración propia.

Las medidas de poder de este modelo (Distancia KS y ROC), se exponen a continuación:

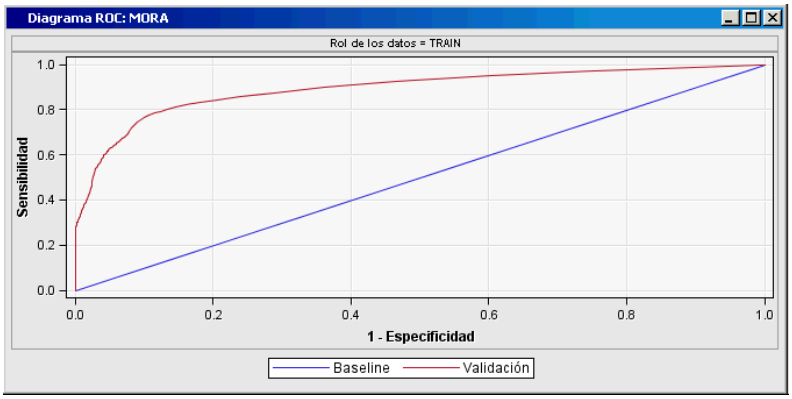
Tabla 26: Medidas de Poder

KS	66,5%
ROC	89,2%

Fuete: Elaboración propia.

La precisión del modelo se mide como el área que queda bajo la curva ROC. El modelo "perfecto" tendrá un área igual a 1, y el peor (aleatoriedad pura) un área igual a 0.5.

Gráfico 14: Curva ROC del Modelo Predictivo de validación



Fuete: Elaboración propia utilizando SAS Enterprise Miner.

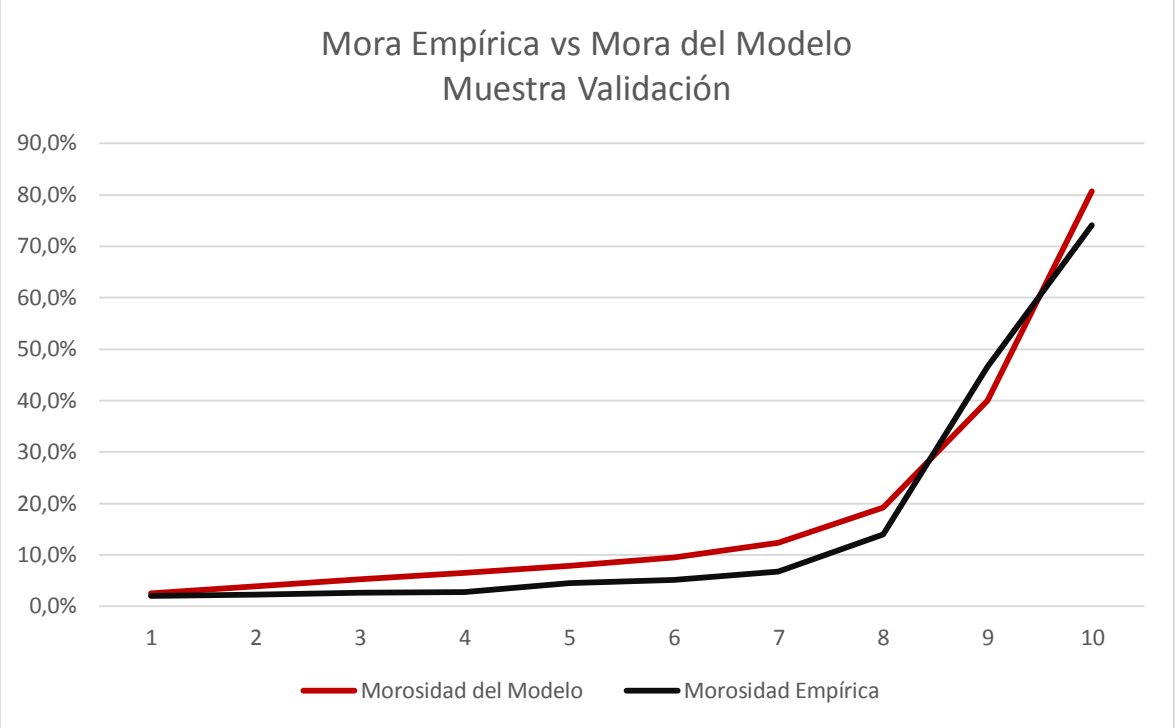
Se efectuó un análisis por deciles de la variable probabilidad, comparando la probabilidad promedio de cada decil, con la tasa de mora observada. Los resultados obtenidos se exponen a continuación:

Tabla 27: Comparación de Morosidad del Modelo y Empírica de validación

Decil	Morosidad del Modelo	Morosidad Empírica
1	2,5%	2,0%
2	3,8%	2,2%
3	5,2%	2,7%
4	6,5%	2,7%
5	7,8%	4,5%
6	9,5%	5,2%
7	12,3%	6,8%
8	19,2%	14,0%
9	40,0%	46,6%
10	80,7%	74,1%

Fuete: Elaboración propia.

Gráfico 15: Comparación de Morosidad del Modelo y Empírica de validación



Fuete: Elaboración propia utilizando SAS Enterprise Miner.

Al igual que en la muestra de desarrollo, puede observarse un ajuste razonable de la curva teórica del modelo, con relación a la curva empírica de tasas de mora.

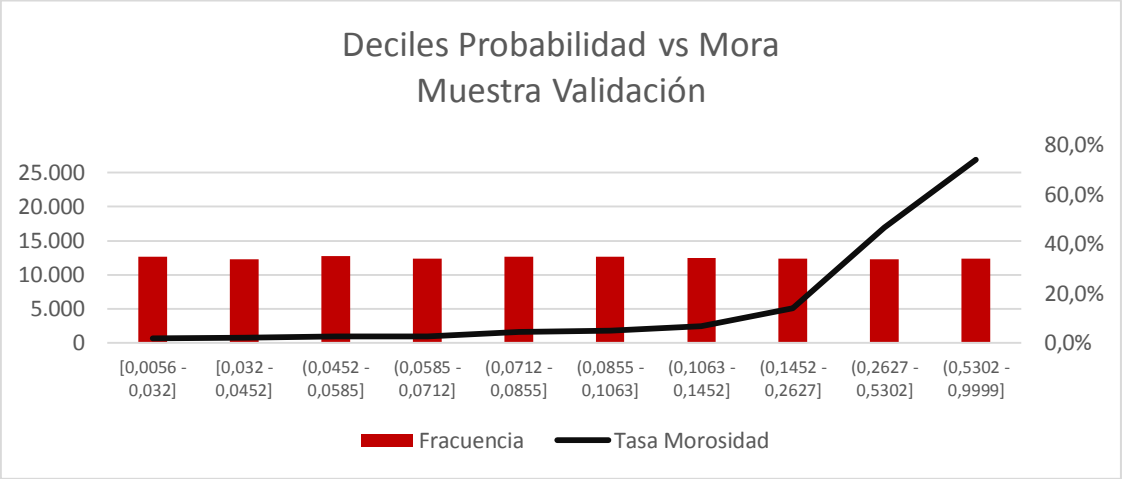
Por último, se efectuó un análisis detallado de las frecuencias acumuladas de Morosos y No Morosos de la variable probabilidad. Los resultados obtenidos se exponen a continuación:

Tabla 28: Descriptiva de Variable Score sobre la base de validación

PROBABILIDAD VALIDACION	No Morosos	Morosos	Total	Frecuencia	Morosidad	Frec. Acum. No Morosos	Frec. Acum. Morosos	Dif. KS
[0,0056 - 0,032]	12.387	257	12.644	10%	2,0%	12%	1%	10,6%
[0,032 - 0,0452]	12.044	276	12.320	10%	2,2%	23%	3%	20,8%
(0,0452 - 0,0585]	12.379	342	12.721	10%	2,7%	35%	4%	31,0%
(0,0585 - 0,0712]	12.047	334	12.381	10%	2,7%	47%	6%	40,9%
(0,0712 - 0,0855]	12.112	577	12.689	10%	4,5%	58%	9%	49,6%
(0,0855 - 0,1063]	11.982	651	12.633	10%	5,2%	70%	12%	57,9%
(0,1063 - 0,1452]	11.596	841	12.437	10%	6,8%	81%	16%	64,8%
(0,1452 - 0,2627]	10.633	1.730	12.363	10%	14,0%	91%	25%	66,5%
(0,2627 - 0,5302]	6.552	5.718	12.270	10%	46,6%	98%	53%	44,5%
(0,5302 - 0,9999]	3.216	9.180	12.396	10%	74,1%	101%	98%	2,1%
Total	**104.948**	**19.906**	**124.854**				***Distancia KS***	***66,5%***

Fuete: Elaboración propia.

Gráfico 16: Comparación de "Deciles de Score" vs "Tasa de Morosidad" en base Validación

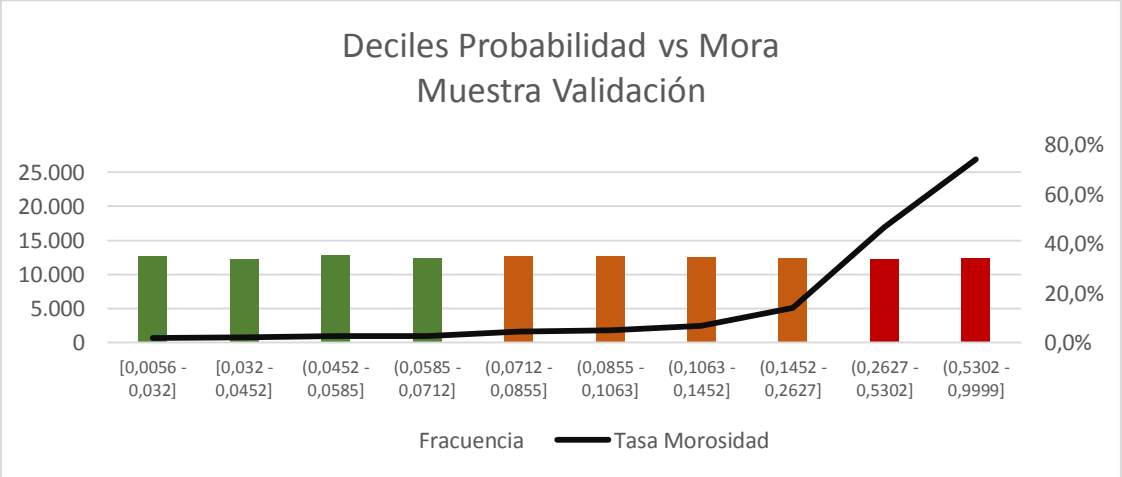


Fuete: Elaboración propia.

Estrategia de Implementación del Modelo

En base a la información provista por "Río Platense S.A.", se obtuvo un modelo de score con un alto nivel de predictibilidad. Esto se verificó luego de analizar "pruebas de significatividad individual", "criterios de información de Akaike y Schwartz", y "medidas de poder".

Gráfico 17: Comparación de "Deciles de Score" vs "Tasa de Morosidad" en base Validación



Fuete: Elaboración propia.

¿Cómo se aplica el modelo en el negocio?

Rojo: Como el modelo discrimina muy bien en los últimos dos deciles, se sugiere rechazar esas solicitudes crediticias.

Naranja: Para los deciles 5, 6, 7 y 8, dependerá de cuan agresiva quiera ser la Entidad Financiera para tomar la decisión de aprobar solicitudes.

Verde: En los primeros cuatro deciles, como la probabilidad de mora es casi nula, se recomienda aprobar esas solicitudes crediticias.

Conclusiones

En el desarrollo de la tesis, se realizó el armado de información, quitando registros duplicados de la base de datos. También se integró información de bureau para enriquecer datos externos a la entidad y se generó una marca para identificar la Base del Desarrollo y la Base para la Validación del Modelo Predictivo.

Luego del análisis univariado, se decidió eliminar la variable “Entidades Financieras Similares en Políticas Crediticia” debido a que se encuentra repetida.

Se sugiere eliminar la variable Saldos Promedios en Cuantas Vista por no tener poder discriminante, se verificará en el Análisis Multivariado si combinada con el resto de variables mejora su performance.

Debido a que el negocio solicitó la incorporación de la variable Edad al modelo, la mejor forma de ingresarla es en dos categorías, “Menores a 30 años” y “Mayores a 30 años inclusive”. Igualmente, se verificará en el Análisis Multivariado que la inclusión no reduzca significativamente la predictibilidad del modelo de score.

Si bien el método de selección de variables “Paso a Paso” demora mayor tiempo de procesamiento para el Software, permite quitar una variable que fue incluida erróneamente con un proceso multivariado (“progresivo”), o no volver a incluir una variable que fue descartada erróneamente con un proceso multivariado (“regresivo”). Este es el motivo por el cual se decide emplear la metodología “Paso a Paso” para llevar adelante la selección de variables.

Luego del estudio multivariado en la base de desarrollo, se estimaron los coeficientes para inferir la probabilidad de mora de una solicitud crediticia. Las medidas de poder son muy buenas, Distancia KS 64,9% y el índice que surge de la Curva ROC asciende a 88,8%, ambos valores son muy elevados y permiten concluir que el modelo tiene poder predictivo.

Por último, la comparación entre Morosidad del Modelo y Morosidad Empírica permite ver una diminuta diferencia entre ellas, evidenciando el muy buen poder predictivo que el modelo posee para inferir morosidad.

Referencias Bibliográficas

Agresti, A., & Kateri, M. (2011). *Categorical data analysis*. Springer. Retrieved from http://link.springer.com/10.1007/978-3-642-04898-2_161

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Altman, E. I., & others. (2000). Predicting financial distress of companies: revisiting the Z-score and ZETA models. *Stern School of Business, New York University*, 9–12.

Burden, R. L., Faires, J. D., & Palmas, O. (2002). *Análisis numérico*. Thomson Learning México. Retrieved from http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=BAC.xis&method=post&formato=2&cantidad=1&expresion=mfn=056893

Caouette, J. B., Altman, E. I., & Narayanan, P. (1998). *Managing credit risk: the next great financial challenge* (Vol. 2). John Wiley & Sons. Retrieved from https://books.google.com.ar/books?hl=es&lr=&id=FOJBUJOAN9AC&oi=fnd&pg=PR9&dq=Caouette,+J.+B.,+Altman,+E.+I.,+%26+Narayanan,+P.+(1998).+Managing+credit+risk:+the+next+great+financial+challenge+(Vol.+2).+John+Wiley+%26+Sons&ots=l5ZteOAMTX&sig=V_Se14gy3i8mpKAyZYmp4608MyY

Choquet, G., & Meyer, P.-A. (1963). Existence et unicité des représentations intégrales dans les convexes compacts quelconques. In *Annales de l'institut Fourier* (Vol. 13, pp. 139–154). Retrieved from http://archive.numdam.org/article/AIF_1963__13_1_139_0.pdf

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837–845.

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: Journal of the Econometric Society*, 1057–1072.

Durand, D., & others. (1941). Risk elements in consumer instalment financing. *NBER Books*. Retrieved from https://ideas.repec.org/b/nbr/nberbk/dura41-1.html

Elizondo, A., & Altman, E. I. (2003). *Medición integral del riesgo de crédito*. Editorial Limusa. Retrieved from https://books.google.com.ar/books?hl=es&lr=&id=UsK-

1Ajo44UC&oi=fnd&pg=PP1&dq=Medici%C3%B3n+integral+del+riesgo+de+cr%C3%A9dito&ots=Ul1aANHx7c&sig=AtdwVLS-6sL82LIEKjts-19SuNU

Hair, J. F., & Suárez, M. G. (1999). *Análisis multivariante* (Vol. 491). Prentice Hall Madrid. Retrieved from http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=AGRISUM.xis&method=post&formato=2&cantidad=1&expresion=mfn=000231

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Haykin, S. S., Haykin, S. S., Haykin, S. S., & Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3). Pearson Education Upper Saddle River. Retrieved from https://cise.ufl.edu/class/cap6615sp12/syllabus.pdf

Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons. Retrieved from https://books.google.com.ar/books?hl=es&lr=&id=Po0RLQ7USIMC&oi=fnd&pg=PR5&dq=+Hosmer,+D.+and+Lemeshow,+S.+(2000).+%E2%80%9CApplied+Logistic+Regression%E2%80%9D.+Wiley,+New+York.&ots=DoaXmg1jFQ&sig=uo10DbphJ_-ucEMd7kUMO8u-P90

Jarrow, R. A., & Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. *The Journal of Finance*, *50*(1), 53–85.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399–402.

Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons. Retrieved from https://books.google.com.ar/books?hl=es&lr=&id=AyQfVTDJypUC&oi=fnd&pg=PR37&dq=Linoff,+Gordon+S.+and+Berry,+Michael+J.+A.+(2011).+%E2%80%9CData+Mining+Techniques%E2%80%9D.+Wiley,+Indiana.&ots=KWLttsTRxI&sig=kiwN77CAvKpaeVHyE8vUGEE_pWg

López, C. P. (2007). *Minería de datos: técnicas y herramientas*. Editorial Paraninfo. Retrieved from https://books.google.com.ar/books?hl=es&lr=&id=wz-

D_8uPFCEC&oi=fnd&pg=PR4&dq=Miner%C3%ADa+de+datos:+t%C3%A9cnicas+y+herramientas&ots=ThZ0yn7w6H&sig=_O_gajJYb6mX7Fq2MSt5cTdusaU

McCorkell, P. L. (2002). The impact of credit scoring and automated underwriting on credit availability. In *The Impact of Public Policy on Consumer Credit* (pp. 209–227). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4615-1415-2_8

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition* (Vol. 544). John Wiley & Sons. Retrieved from https://books.google.com.ar/books?hl=es&lr=&id=O_qHDLaWpDUC&oi=fnd&pg=PR7&dq=Discriminant+Analysis+and+Statistical+Pattern+Recognition&ots=6FPiNPIS_S&sig=OsFO8M5K9blo8nLaZvHI6XwrFtM

Peña, D. (2002). *Análisis de datos multivariantes* (Vol. 24). McGraw-Hill Madrid. Retrieved from http://www.dpye.iimas.unam.mx/lety/archivos/cursoinegi/apoyos/ANAI%CC%80%C2%81LISIS%20DE%20DATOS%20MULTIVARIANTES(1).pdf

Schwarz, G., & others. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring* (Vol. 3). John Wiley & Sons. Retrieved from https://books.google.com.ar/books?hl=es&lr=&id=SEbCeN3-kEUC&oi=fnd&pg=PT7&dq=Siddiqi,+N.+(2012).+Credit+risk+scorecards:+developing+and+implementing+intelligent+credit+scoring+(Vol.+3).+John+Wiley+%26+Sons&ots=RtUP3PhLgQ&sig=dLMLvri7WLgVeHf0xlTdUy8NonA

White, R. W. (1975). *The probability of bankruptcy for American industrial firms*. Institute of Finance and Accounting.