UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"CONSTRUCCIÓN DE UN MODELO DE SCORE PARA LA EVALUACIÓN DE CLIENTES POTENCIALES EN UNA ENTIDAD FINANCIERA"

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERA EN ESTADÍSTICA INFORMÁTICA

ALESSANDRA STEFANY CARRASCO REYES

LIMA – PERÚ

2024

CONSTRUCCIÓN DE UN MODELO DE SCORE PARA LA EVALUACIÓN DE CLIENTES POTENCIALES EN UNA ENTIDAD FINANCIERA

INFORME DE ORIGINALIDAD

7%	6%	1%	3%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	busquedas.elperuano.pe Fuente de Internet	2%
2	Submitted to Universidad ESAN -- Escuela de Administración de Negocios para Graduados Trabajo del estudiante	<1%
3	repositorio.espe.edu.ec Fuente de Internet	<1%
4	Submitted to Universidad Nacional Agraria La Molina Trabajo del estudiante	<1%
5	Submitted to Pontificia Universidad Catolica del Peru Trabajo del estudiante	<1%
6	qdoc.tips Fuente de Internet	<1%
7	Ceballes-Serrano, C.C., S. Garcia-Lopez, J.A. Jaramillo-Garzon, and G. Castellanos-Dominguez. "A strategy for classifying	<1%

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"CONSTRUCCIÓN DE UN MODELO DE SCORE PARA LA EVALUACIÓN DE CLIENTES POTENCIALES EN UNA ENTIDAD FINANCIERA"

PRESENTADO POR
ALESSANDRA STEFANY CARRASCO REYES

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE INGENIERA EN ESTADÍSTICA INFORMÁTICA

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

.....	.....
Dr. Cesar Higinio Menacho Chiok	Mg. Sc. Jesús Eduardo Gamboa Unsihuay
PRESIDENTE	ASESOR
.....	.....
Dr. Jaime Carlos Porras Cerrón	Mg. Sc. Rolando Jesus Salazar Vega
MIEMBRO	MIEMBRO

Lima - Perú
2024

DEDICATORIA

Dedicado a mis padres, que lucharon mucho por darme una vida llena de oportunidades y quienes me enseñaron que puedo lograr cada uno de mis sueños. A mi abuela, que siempre anheló con verme siendo profesional e independiente. A mi hermana, que es mi cómplice y apoyo incondicional en cada aspecto de mi vida. A Cody, por ser mi compañero en cada una de las noches de desvelo.

AGRADECIMIENTO

Agradecimiento especial al profesor Jesús Gamboa por el apoyo brindando, asesorándome durante la elaboración del presente trabajo.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El riesgo de crédito se encuentra dentro de las principales preocupaciones de las entidades financieras, por ello mantener una adecuada gestión es de vital importancia. Como parte de las estrategias definidas por las entidades se hace uso de análisis y modelos estadísticos para la admisión de clientes, generación de campañas, monitoreo de la calidad de cartera de clientes, manteniendo su apetito de riesgo. En el presente trabajo se describe la construcción de un modelo de score, específicamente, para clientes potenciales de una campaña crediticia de una entidad financiera, utilizando modelos *Random Forest* para la selección de variables y ajustando un modelo de Regresión Logística para la predicción de clientes morosos y no morosos. Dicho modelo logró una mejora de 14 puntos porcentuales en el indicador de *Kolgomorov Smirnov* en comparación con el modelo anterior de score.

Palabras clave: Riesgo de crédito, Score, *Credit Scoring*, Regresión Logística, Balanceo de Clases, *Random Forest*.

ABSTRACT

Credit risk is one of the main concerns of financial institutions, therefore maintaining an adequate management is very important. As part of the strategies defined by the entities, statistical analysis and models are used for the admission of clients, generation of financial campaigns, monitoring of the quality of the client portfolio, maintaining their risk appetite. This paper describes the construction of a model, specifically for potential customers of a credit campaign of a financial institution, using Random Forest models for the selection of variables and fitting a Logistic Regression model for the prediction of payment default. This model achieved an improvement of 14 percentage points in the Kolgomorov Smirnov compared to the previous scoring model.

Keywords: Credit Risk, Score, Credit Scoring, Logistic Regression, Random Forest.

I. INTRODUCCIÓN

En el presente trabajo se expone la problemática y la construcción de un modelo de score que permita perfilar a los clientes potenciales de una entidad financiera en Perú, bajo las directrices de la Resolución N° 0083-2022-CU-UNALM asociada al proceso de titulación bajo la modalidad de suficiencia profesional.

La entidad financiera de estudio es una empresa de capital extranjero, que inició sus operaciones en Perú hace menos de 10 años y tiene como único producto financiero las líneas de crédito revolventes; dichas líneas pueden ser usadas para compras en las tiendas de electrodomésticos pertenecientes al mismo grupo de dicha entidad o en productos de construcción en establecimientos asociados. Las tiendas de electrodomésticos, que representan al principal canal de venta, se encuentran ubicadas en las zonas de Lima Sur, Lima Norte y provincias a nivel nacional.

En el año 2018, la bachillera en Estadística Informática, quien suscribe el presente trabajo, se desempeñó como Analista de Riesgo de Crédito en una entidad financiera del Perú, supervisada por la Superintendencia de Banca y Seguros (SBS). Bajo dicha posición se realizaron diversos análisis, principalmente relacionados al seguimiento de la mora e identificación de segmentos de clientes que generaban menor y mayor riesgo. En relación con ello, y en comunicación con la Gerencia General, se identificó la necesidad de modificar la campaña crediticia "Pre Aprobado" y utilizar más información del sistema financiero, con el fin de poder perfilar a los clientes potenciales, es decir, de identificar subsegmentos con diferentes niveles de riesgo, los cuales deberán contar con tasas de interés diferenciadas.

En el presente documento se presentará en primera instancia la importancia de la gestión del riesgo para las empresas del sistema financiero, una visión preliminar de la entidad financiera y la problemática a la que se busca brindar una solución. Enseguida, se presentará el objetivo general y los objetivos específicos que se buscan alcanzar. Posteriormente se detallará la

metodología empleada y luego se expondrá los resultados, la discusión, para cerrar con las conclusiones y recomendaciones obtenidas.

1.1 Problemática

Las entidades financieras buscan mantener su rentabilidad y lograr la continuidad del negocio, principalmente mediante la óptima evolución de su cartera de clientes. Para lograrlo, las entidades buscan sumar esfuerzos en dos aristas: incrementar sus colocaciones y mantener el riesgo en niveles aceptables, los cuales son definidos de manera autónoma por cada compañía. Por ello, las empresas definen políticas y procedimientos que permitan calcular y monitorear continuamente los indicadores para cada uno de los riesgos a los que la entidad se ve expuesta.

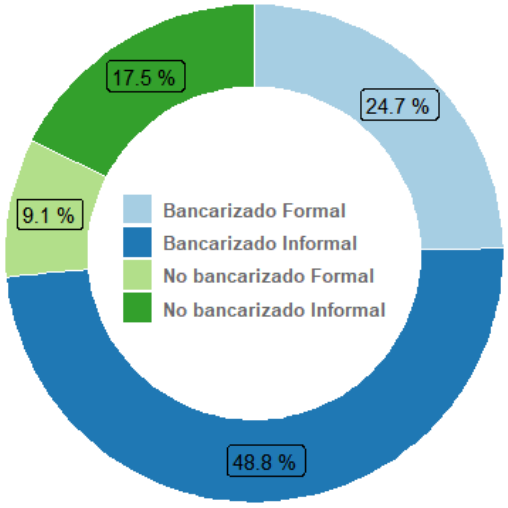
Dentro de los riesgos que causan mayor preocupación en las empresas del sistema financiero se encuentra el riesgo de crédito, el cual se define como la posibilidad de pérdida dado que los clientes incumplan con su compromiso de pago (Elizondo & Altman, 2004, p. 47). En cuanto a la gestión de riesgo crediticio, y en relación con la tendencia de utilizar los datos para generar conocimiento, las empresas utilizan información de fuentes internas y externas para analizar segmentos y/o variables de interés; asimismo, dichos datos permiten la elaboración de modelos estadísticos que permiten definir o modificar políticas y estrategias de la empresa, así como elaborar campañas que logren incrementar las colocaciones asociadas a cierto nivel de riesgo.

En particular, la entidad en estudio aplica una gestión de riesgos basándose principalmente en la gestión de riesgo de crédito, debido a que dicha empresa busca bancarizar a un perfil de clientes que no es el común de otras entidades del sistema financiero, incluyendo en los créditos a personas no bancarizadas previamente y/o con trabajos informales. Además de las características sociodemográficas de los clientes objetivos, se debe considerar la ubicación geográfica de las tiendas físicas, las cuales según el conocimiento de negocio presentan un riesgo adicional por estar asociadas a distritos con mayor índice de inseguridad y fraudes; por lo que se obtiene que la cuota de riesgos por créditos de la entidad es superior en comparación con otras entidades financieras. En relación con ello, en la figura 1 se puede observar que, respecto a la actividad económica, 1 de cada 3 clientes son formales y 2 de cada 3 son

informales, y en cuanto a la distribución de clientes según el nivel de bancarización se obtiene que aproximadamente el 70% son bancarizados y el 30% son no bancarizados previamente.

Figura 1

Participación de clientes por bancarización y tipo de actividad económica



Nota. El gráfico muestra la participación de la cartera de clientes del año 2017 según las principales variables definidas por la entidad: bancarización y tipo de trabajo.

En específico a los niveles de riesgo dentro de la entidad, y evaluando la cartera de clientes, la entidad ha definido subsegmentos, los cuales son asociados a planes financieros. Dentro de ellos, se encuentra el plan financiero asociado a la Campaña Pre Aprobado; dicha campaña es construida a partir de la base de datos proporcionada por la SBS, considerando como clientes potenciales a aquellas personas que presentan un buen historial en el sistema financiero, y en consecuencia, el riesgo asociado a esta campaña debería ser inferior en comparación con el resto de los planes financieros que cuenta la empresa.

Como parte de la gestión de riesgo crediticio, la entidad definió indicadores para monitorear la mora, siendo uno de los más importantes el correspondiente a mora temprana referente al primer incumplimiento de pago denominado First Payment Default (*FPD*). En tal seguimiento, se observó que el valor del FPD iba aumentando paulatinamente, dicha situación generó alerta debido a que la campaña Pre Aprobado debería presentar un performance adecuado, por ser el

segmento de menor riesgo, y esto conllevaría a incurrir en un incremento de las provisiones e impacto en la rentabilidad. Por ello, la empresa requería implementar un ajuste a dicha campaña, el cual debía ser soportado por un modelo de score que explote más información crediticia. Dicha propuesta permitió perfilar clientes dentro de ese segmento y posteriormente el área de Finanzas pudo evaluar y ajustar las tasas de interés para asegurar la rentabilidad esperada por la empresa.

Al respecto, se encuentran trabajos precedentes en Perú como el realizado por David y Chuquipul (2008), el cual estaba enfocado en construir un modelo de score que permita estimar el parámetro de Probabilidad de Default (PD) para el cálculo de la Pérdida Esperada; asimismo, se encuentra el estudio realizado por Rayo Cantón et al. (2010), el cual tiene la peculiaridad de desarrollar un modelo de scoring al sector de microfinanzas que presenta un comportamiento diferente al de los bancos. En línea con lo mencionado, se cuenta con el estudio desarrollado por Puertas y Marti (2013), basado en una entidad financiera de España, en el cual se precisa la importancia del uso del credit scoring en la evaluación de solicitudes de crédito con el fin de superar subjetividades de los analistas, así como también se muestra la relevancia de incluir en la metodología la división de la base de datos en muestra de entrenamiento y test, y la realización de pruebas de validación con el objetivo de asegurar un buen performance del modelo.

Es importante mencionar que, una particularidad que se presenta en el modelamiento para predecir el incumplimiento de pago es el desbalanceo de clases, por lo que es importante incluir dentro de la metodología un paso que le dé tratamiento a esta situación. En relación con ello, Namvar et al. (2018) compararon los tres tipos de remuestreo que permiten balancear las clases, considerando técnicas de submuestreo, sobre muestreo e híbridos y evaluaron cual presentaba mejor ajuste para un conjunto de datos de préstamos sociales; asimismo, Moscato et al. (2021) probaron técnicas de los tres tipos de remuestreo en el modelamiento de regresión logística, Random Forest y Perceptrón Multicapa, y comparó que combinación de remuestreo y modelo presentaba mejores indicadores de ajuste.

Finalmente, respecto a la amplia gama de modelos e información que pueden soportar el credit scoring, Barddal et al. (2020) compararon el uso de un modelo de regresión logística versus modelos machine learning e indicaron que es una buena vía combinar los modelos tradicionales con machine learning debido a que actualmente existe mayor información proveniente de diversas fuentes; también, Ashofteh y Bravo (2021) explicaron que es posible incluir datos no tradicionales, como información telefónica, cuando se busca puntuar poblaciones sin historia en el sistema financiera.

1.2 Objetivos

Objetivo General

Construir un modelo de regresión logística para los clientes potenciales de una campaña crediticia.

Objetivos Específicos

- **a.** Crear variables mediante la reingeniería de los campos insumo proporcionados en el Reporte Crediticio Consolidado.
- **b.** Balancear los datos en función a la variable objetivo "probabilidad de default temprano".
- **c.** Identificar las variables candidatas para el modelo de score.
- **d.** Estimar los parámetros del modelo de regresión logística.
- **e.** Validar el ajuste del modelo de regresión logística.

II. REVISIÓN DE LITERATURA

2.1 Regulación y supervisión del sistema financiero

La Superintendencia de Banca, Seguros y AFP (SBS) es el ente nacional encargado de regular y monitorear al sistema financiero y seguros, y responsable de la prevención del lavado y financiamiento del terrorismo (PLAFT). De acuerdo con la información proporcionada en su web oficial (https://www.sbs.gob.pe/), la SBS tiene como misión asegurar el buen funcionamiento de las empresas supervisadas, resguardando la estabilidad financiera y conducta de mercado, con la finalidad de velar por el bienestar de los ciudadanos.

Este organismo creó marcos normativos, con un enfoque basado en riesgos, en los que se incluye estatutos y directrices asociadas, principalmente, a riesgo de crédito, riesgo de mercado y riesgo operacional, tomando como referencia estándares internacionales como los Acuerdos de Basilea; asimismo, implementó normas que abarcan riesgos propios y relevantes de la realidad peruana. En ese marco, indica que las entidades deben contar con una adecuada gestión integral de riesgos, que permita identificar posibles eventos de pérdida, que se encuentre acorde al apetito definido autónomamente y que involucre a la totalidad de la empresa, líneas de negocio y procesos. (Superintendencia de Banca, Seguros y AFP, 2017, Resolución 272)

En esta línea, en la Resolución 272-2017 [SBS] se define las funciones principales de la Unidad de Riesgos para poder llevar a cabo la gestión integral de riesgos:

> a) Proponer las políticas, procedimientos y metodologías apropiadas para la gestión integral de riesgos en la empresa, incluyendo los roles y responsabilidades. b) Participar en el diseño y permanente mejora y adecuación de los manuales de gestión de riesgos. c) Velar por una adecuada gestión integral de riesgos, promoviendo el alineamiento de la toma de decisiones de la empresa con el sistema de apetito por el riesgo. d) Guiar la integración entre la gestión de riesgos, los planes de negocio y las actividades de gestión empresarial. e) Establecer un lenguaje común de gestión de riesgos basado en las definiciones de esta norma y de los demás reglamentos aplicables. f) Estimar las necesidades de patrimonio que permitan cubrir los riesgos que enfrenta la empresa y alertar a la gerencia y al comité de riesgos o directorio,

según sea el caso, sobre las posibles insuficiencias de patrimonio efectivo. g) Informar a la gerencia y al comité de riesgos o directorio, según sea el caso, los aspectos relevantes de la gestión de riesgos para una oportuna toma de decisiones. h) Informar al comité de riesgos o directorio, según sea el caso, acerca de los riesgos asociados al lanzamiento de nuevos productos, y a los cambios importantes en el ambiente de negocios, el ambiente operativo o informático, de forma previa a su lanzamiento o ejecución; así como de las medidas de tratamiento propuestas o implementadas. (p. 22)

Respecto al desarrollo y a la implementación de los modelos internos de riesgo de crédito, se indica que debe existir un área encargada de la elaboración de los sistemas de clasificación de la entidad utilizando data histórica del incumplimiento, así como la revisión de la existencia, aplicabilidad y coherencia del sistema de calificaciones en las distintas unidades de la empresa. También, señala la importancia de la validación interna de dichos modelos de riesgo, en donde se debe evaluar aspectos cuantitativos y cualitativos de la calidad de datos utilizados, el uso del sistema de calificaciones internas, parámetros estimados para la gestión del riesgo y la comparación, por lo menos anual, entre las tasas observadas del incumplimiento y las probabilidades de incumplimiento estimadas. (SBS, 2019, Resolución 14354)

Además de ser el organismo regulador y supervisor, proporciona a las entidades financieras supervisadas reportes e información de interés, entre ellas una base de datos denominada “Reporte Crediticio Consolidado”, o por sus siglas RCC, la cual contiene el historial crediticio de todas las personas que cuentan con algún tipo de deuda en el sistema financiero. Este reporte consolida la información brindada por cada ente financiero y contiene los campos: la identidad de las personas bancarizadas, la cantidad de empresas con las que se tiene algún compromiso crediticio, nombre de las empresas, los saldos para cada uno de los tipos de créditos y la calificación SBS (Normal, CPP, Deficiente, Dudoso y Pérdida), la cual depende de los días de atraso. Dicha calificación es utilizada por las empresas para la admisión de clientes, campañas crediticias, monitoreo de cartera, benchmarking, entre otros.

2.2 Gestión del riesgo de crédito

Las entidades financieras buscan lograr una adecuada gestión de los riesgos, los cuales se entienden como la posibilidad que se materialice un evento que genere pérdidas, debido a alguna vulnerabilidad latente de la entidad (González et al., 2018, p.54), para lo cual se debe evaluar los diferentes tipos de riesgos, tales como el riesgo de crédito, el riesgo operacional, el riesgo tecnológico, entre otros.

En particular, en la Resolución S.B.S N°3780-2011, emitida por la Superintendencia de Banca, Seguros y AFP, se explica el riesgo de crédito como la "posibilidad de pérdidas por la incapacidad o falta de voluntad de los deudores, contrapartes, o terceros obligados, para cumplir sus obligaciones contractuales". También se menciona que como parte de la gestión del riesgo de crédito se debe contar con políticas, indicadores, umbrales y modelos para identificar, monitorear y controlar el riesgo de crédito, con el fin de mantener a la empresa dentro del apetito de riesgo.

Al respecto, el credit scoring es un elemento clave para poder cuantificar el potencial riesgo de crédito, utilizando la relación entre las múltiples variables insumo y la variable correspondiente al default para la construcción de los scorecards; la implementación del credit scoring en las empresas permite superar los distintos juicios de expertos y reemplazarlos por fundamentos matemáticos (Baesens et al., 2016, p.95). En otras palabras, es un método que permite puntuar y ordenar a los clientes en términos del riesgo individual asociado. Al respecto, se busca construir un modelo, basándose en datos históricos, que identifique variables importantes que permitan clasificar entre clientes buenos y clientes malos, considerando que los puntajes altos estarán asociados a clientes con buen rendimiento, caso contrario el puntaje será bajo (Mester, 1997, p.4)

Respecto a la construcción de los scorecards, Siddiqui (2017) presenta la fórmula para convertir las probabilidades obtenidas en los modelos de score a una escala numérica, la cual se presenta a continuación:

$$Score = Offset + Factor * \ln(odds)$$

En relación con el objetivo mencionado, en investigaciones pasadas se encontró a David y Chuquipul (2008), quienes en su investigación para optar al grado de Magister en Finanzas, analizaron el uso del modelo de scoring enfocándose en las tarjetas de crédito. Dicha investigación, centrada en Lima Metropolitana, tenía como objetivo construir una herramienta que sirva de ayuda a los analistas de créditos. Se desarrolló un modelo de regresión logístico, el cual incluyó variables explicativas de tipo sociodemográficas e información de fuentes internas. En esta investigación se concluyó que el modelo construido representa al parámetro de Probabilidad de Default (PD) y que permitiría ser usado como un modelo interno para el cálculo de la Pérdida Esperada (PE), según el Acuerdo de Basilea II.

Asimismo, Cantón et al. (2010) presentaron un modelo de scoring orientado a instituciones de microfinanzas de Perú; asimismo, en la investigación se expone la utilidad del modelo desarrollado para la fijación de tasas de interés y su posible uso en la gestión de riesgo crediticio de acuerdo con el Acuerdo de Basilea II. Se presenta la necesidad de desarrollar un modelo de score para este tipo de entidades que, por el comportamiento de su público objetivo, difieren de la banca tradicional y del nivel de riesgo; en relación con ello, se indica que no existe una vasta literatura dedicada al scoring en instituciones de microfinanzas. Además, se explica que para la construcción de un modelo de score se puede hacer uso de técnicas como el análisis discriminante, modelos de probabilidad lineal, modelos logit, modelos de programación lineal, redes neuronales y árboles de clasificación. En dicha investigación se desarrolló un modelo logístico con nueve variables explicativas.

2.3 Tratamiento de valores perdidos

Es frecuente que existan valores perdidos en los datos a analizar y/o estimar, para el tratamiento de esta situación se utiliza distintos procedimientos, tales como los métodos tradicionales de imputación, imputación simple e imputación múltiple.

En cuanto a los métodos tradicionales de imputación, Galván y Medina (2007) explican el análisis con datos completos (listwise), donde se eliminan los registros que presenten valores faltantes en algunas de las variables, simplificando la base de datos a los registros que tienen información en cada uno los campos; el análisis con los datos disponibles (pairwise deletion),

el cual utiliza toda la información disponible de cada variable, pero considerando diferentes tamaños de muestra; y reponderación, donde las observaciones se ponderan utilizando algún modelo de probabilidad con información completa. Además, Galván y Medina (2007) indican que dentro de los procedimientos de imputación simple se encuentran la imputación por el método de medias no condicionadas, en donde se utiliza la media muestral de los valores existentes para reemplazar cada valor faltante, la imputación por medias condicionadas para datos agrupados, la imputación con variables ficticias, imputación mediante una distribución no conocida y la imputación por regresión. Álvarez y Muñoz (2009) afirman que dentro de las ventajas de la imputación simple se encuentra que son sencillas de aplicar a los datos sin que exista una pérdida relevante de eficiencia, en comparación con la imputación múltiple.

Por otro lado, Galván y Medina (2007) explican que la imputación múltiple usa métodos de simulación de Monte Carlo, donde se sustituye los valores faltantes a partir de un número de simulaciones (entre 3 a 10); asimismo, expone que la metodología consiste en que en cada simulación se analice los datos completos y luego se combinen los resultados para estimaciones robustas.

2.4 Normalización de variables

En la etapa de preprocesamiento, es decir, en la etapa previa al modelamiento se requiere evaluar el tipo y el rango de valores de cada una de las variables candidatas. Al respecto, Singh B. y Singh D. (2019) explican que la normalización es un paso indispensable en la construcción de modelos, el cual tiene como finalidad neutralizar el impacto de las variables numéricas con rangos de valores más amplios respecto a aquellas con rangos más pequeños en la discriminación de las clases y patrones; no obstante indica que pese a que la normalización asegura que las variables tengan una contribución numérica uniforme, esto no implica que tales variables sean igual de relevantes en el modelo de clasificación. A su vez, mencionan los tipos de normalización: los métodos basados en la media y desviación estándar, tales como el Z-score, Mean Centered, escala de Pareto, escala de estabilidad de variables; métodos basados en los valores mínimos y máximos, tales como normalización min-max, normalización max; normalización de la escala decimal; normalización de la mediana y de la desviación absoluta de la mediana; y normalización sigmoidal.

Con respecto a ello, Borkin et al. (2019) presentaron el método de normalización min-max, el cual realiza un escalamiento de los valores de las variables, convirtiéndolas en el rango de [0,1] o [-1,1] de acuerdo con la naturaleza de cada variable, y su fórmula se presenta a continuación:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

2.5 Balanceo de clases

En el proceso de modelamiento se asume que la participación de las clases que serán predichas presenta una proporción similar; no obstante, en la vida real se presentan situaciones donde existen clases con mucha más frecuencia. Tal situación la podemos observar en el análisis de riesgo de crédito, pues dentro de sus objetivos pretende clasificar a los clientes como "morosos" y "no morosos", y para ello se construyen modelos utilizando como variable dependiente el default, la cual presenta una distribución de clases desproporcionada, pues aquellos clientes "morosos" cuentan con una participación muy pequeña en comparación a los "no morosos"; caso contrario, la empresa no sería rentable. En la actualidad, y según el reporte de la SBS "Ratios de morosidad según días de incumplimiento" (https://www.sbs.gob.pe/), para mayo 2022 la morosidad promedio de más de 30 días de incumplimiento para los bancos es del 4.2%, para financieras es el 6.9% y para Edpymes 7.9%. Otros ejemplos de clases desbalanceadas se encuentran cuando las entidades buscan detectar los fraudes a los que se ven expuestos o cuando se pretende predecir la fuga de clientes de una organización.

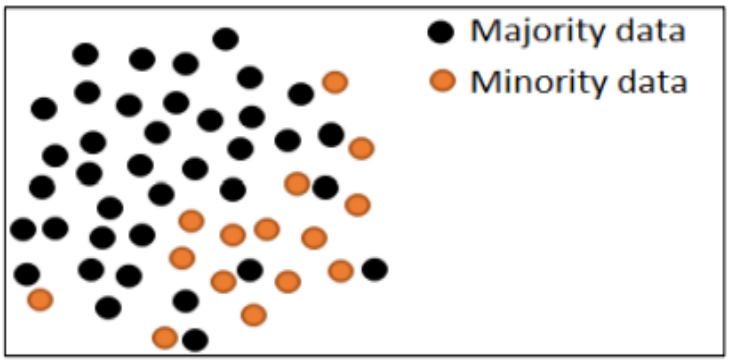
Por ello, se debe realizar un diseño inteligente para superar el sesgo que se genera hacia la clase mayoritaria. En este sentido, el balanceo de datos busca equilibrar los datos de la muestra de entrenamiento del modelo, mediante el sobremuestreo, que es la generación de nuevos objetos de los grupos minoritarios, o mediante el submuestreo, que es la eliminación de objetos de la clase mayoritaria (Krawczyk, 2016). En relación con ello, se sugiere realizar pruebas suficientes para evaluar la mejora de un método sobre el otro en el performance del modelo. Es importante considerar los aspectos positivos y negativos de cada técnica, teniendo en cuenta que en los métodos de sobremuestreo se podría crear datos que difieran de la distribución original de la variable; mientras que en el caso de los métodos de submuestreo es relevante tener en cuenta que al eliminar mucha información podría generar sobreentrenamiento en el set reducido. (Hoyos, 2019, p.10).

El balanceo de clases de la variable dependiente es un paso necesario para poder obtener estimaciones confiables de la probabilidad de incumplimiento de pago. Namvar et al. (2018) utilizaron en la construcción del modelo de score la comparación de técnicas de submuestreo (*RUS, IHT*), sobremuestreo (*ROS, ADASYN, SMOTE*) e híbridos (*SMOTE TOMEK, SMOTE ENN*), evaluándolos en los modelos de Regresión Logística, Análisis Discriminante Lineal y Random Forest; e indicaron que para evaluar datos desbalanceados no es adecuado usar la precisión como métrica de evaluación, dado que favorece a la clase mayoritaria, y en cambio recomienda el uso del ROC, sensibilidad y especificidad para evaluar los modelos. De igual manera, Moscato et al. (2021) contrastaron las técnicas de remuestreo anteriormente mencionadas y las evaluaron en los modelos de Regresión Logística, Random Forest y Perceptrón Multicapa; a la vez, se propuso considerar también el ratio de falsos positivos en la evaluación del performance de los modelos, debido a que dicha métrica impacta en la cantidad de créditos que pueden ser denegados indebidamente y por consecuencia en los ingresos de la entidad.

Por el lado del submuestreo, la técnica Tomek Link permite eliminar observaciones ruidosas y se basa en identificar los enlaces Tomek. Al respecto, Ai-jun y Peng (2020) indican que se forma un enlace Tomek cuando dos observaciones x e y pertenecen a distintas clases y son los vecinos más próximos, es decir, no existe una tercera observación z que presente menor distancia euclidiana que $dist(x, y)$; por consiguiente, una de las observaciones que conforman el enlace Tomek puede ser ruido o que el par se encuentran en el límite de decisión. Luego de detectar los enlaces Tomek se procede a eliminar la observación, de dicho enlace, que pertenece a la clase mayoritaria, logrando así disminuir la clase mayoritaria y separar adecuadamente las clases. En la figura 2 se puede observar un conjunto de datos con clases desbalanceadas, donde se visualiza que los límites entre la clase minoritaria y mayoritaria no son claros y en la figura 3 se visualiza el efecto de la aplicación del Tomek Link cuando las observaciones que se encuentran en el límite de decisión son suprimidas.

Figura 2

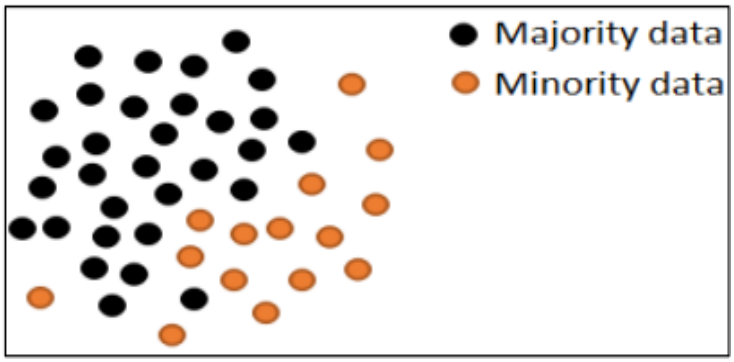
Conjunto de datos con clases desbalanceadas



Nota. Captado de Research on Unbalanced Data Processing Algorithm Base Tomeklinks-Smote, por Ai-jun y Peng, 2020.

Figura 3

Aplicación de la técnica Tomek Link



Nota. Captado de Research on Unbalanced Data Processing Algorithm Base Tomeklinks-Smote, por Ai-jun y Peng, 2020.

Respecto a SMOTE (Synthetic Minority Over-sampling Technique), Chawla et.al (2002), indicaron que es una técnica donde se realiza el sobremuestreo con la clase minoritaria introduciendo nuevos valores creados denominados "sintéticos" a lo largo del segmento que une los k-vecinos más cercanos de la dicha clase. Presentaron además el pseudo código para el algoritmo $SMOTE\ (T, N, k)$

Input: Número de muestras de la clase minoritaria (T), Porcentaje de sobremuestreo (N), número de vecinos cercanos (k).

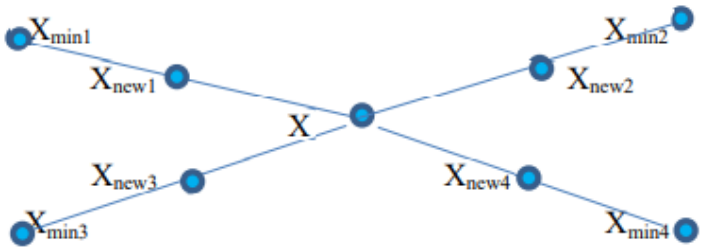
Output $(N/100)$*T

1. Si $N < 100$
2. Entonces se aleatoriza las T muestras de clases minoritaria.
3. $T = (N/100) * \mathrm{T}$
4. $N = 100$
5. Endif
6. $N = (int) = (N/100)$
7. $k = número\ de\ vecinos\ más\ cercanos$
8. $numattrs = Número\ de\ atributos$
9. $Sample[][]$: Matriz de las muestras originales de la clase minoritaria.
10. $newindex$: Recuenta el número de muestras sintéticas generadas, empezando desde 0.
11. $Synthetic[][]$: Matriz para las muestras sintéticas. (Se calcula los k vecinos más cercanos para cada muestra de clase minoritaria).
12. for $i \leftarrow 1$ to T:
13. Calcula los k vecinos más cercanos para i, y guarda los índices en la matriz $nnarray.Poblar\ (N, i, nnarray)$.
14. Endfor
15. While $N \neq 0$
16. Elegir un número aleatorio entre 1 y k, llamado nn. En este paso se elige uno de los k vecinos más cercanos a i.
17. for $attr \leftarrow 1$ to $numattrs$
18. Calcular $dif = Sample\left[nnarray[nn]\right][attr] - Sample[i][attr]$
19. Calcular $gap = número\ aleatorio\ entre\ 0\ y\ 1$.
20. $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
21. Endfor
22. $newindex++$
23. $N = N - 1$
24. Endwhile
25. Return. (Chawla et.al, 2002, p. 327)

En síntesis, y en relación con el pseudo código anterior, Cai et.al (2008) indicaron que la técnica del SMOTE crea nuevas observaciones al interpolar entre algunas observaciones minoritarias que se encuentran próximas, ampliando los límites de decisión de la clase minoritaria en el espacio de la clase mayoritaria, evitando el sobreajuste.

Figura 4

Sobremuestreo mediante SMOTE



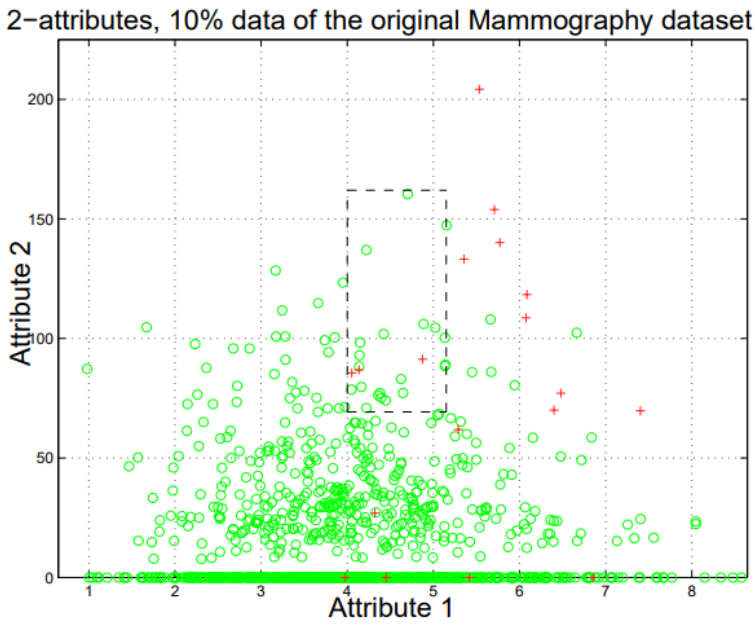
Nota. Se observa las observaciones sintéticas creadas entre las observaciones más cercanas X y X_{min-i}. Captado de Research on Unbalanced Data Processing Algorithm Base Tomeklinks-Smote, por Ai-jun y Peng, 2020.

Además, Chawla et.al (2002) ejemplifican el uso del SMOTE en árboles de decisión e indica que el sobremuestreo bajo esta técnica permite obtener regiones de aprendizaje de gran tamaño muestral y no específicas tal como se puede observar en la figura 5, es decir, evita crear regiones específicas para la clase minoritaria y de esa forma evitar el sobreajuste.

Por último, la técnica SMOTE-Tomek Link se centra en identificar los enlaces Tomek para luego poder borrar ambas observaciones que conforman dicho enlace, tanto la perteneciente a la clase minoritaria como el de la clase mayoritaria, y luego se realiza sobremuestreo, lo cual es ejemplificado en la figura 6.

Figura 5

Ejemplo de región de decisión generada posterior al sobremuestreo sintético



Nota. La clase minoritaria es representada mediante el (+) y el rectángulo puntuado representa la región de decisión para la base de datos *"Mammography"*. Adoptado de *SMOTE: técnica de sobremuestreo minoritario sintético. Revista de investigación de inteligencia artificial* (p.327), por Chawal et al., 2002.

Complementando, Ai-jun y Peng (2020) presentan el algoritmo de SMOTE- Tomek Link:

Algoritmo SMOTE-Tomek Link

Input: Conjunto de datos.

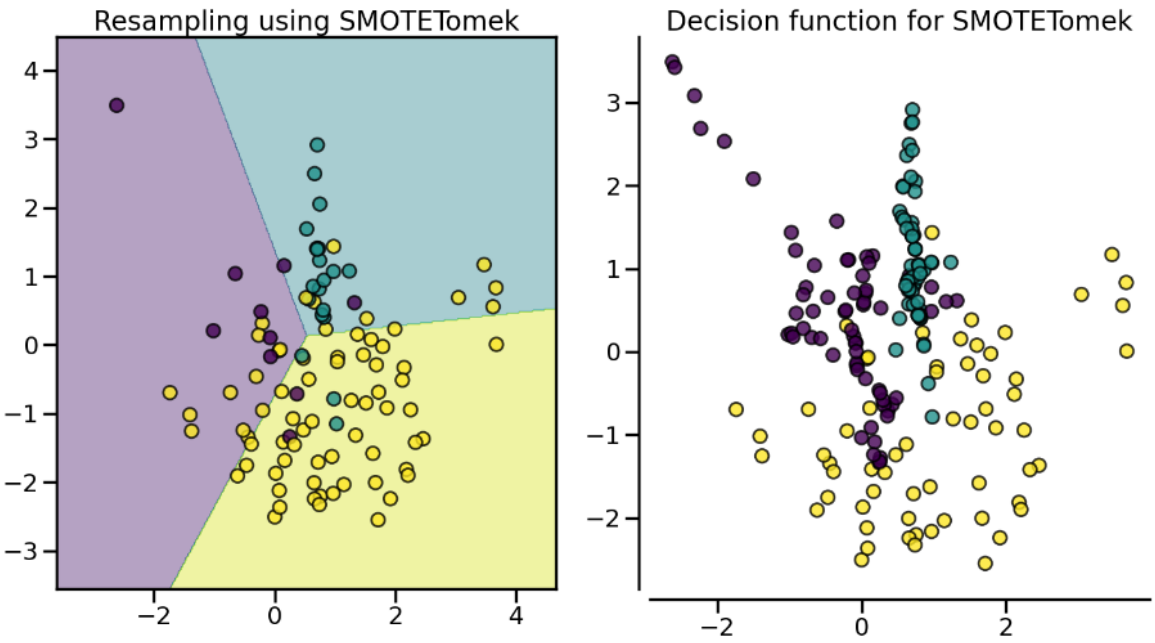
Output: Nuevo conjunto de datos ajustados.

1. Se toman dos observaciones x_i y x_j, y se calcula la distancia entre ellas.
2. Se toma una tercera observación, y se calculan las siguientes distancias $dist(x_i, x_k)$ y $dist(x_j, x_k)$.
3. Se verifica si hay enlaces Tomek, cuando se cumple que $dist(x_i, x_j) < dist(x_i, x_k)$ y $dist(x_i, x_j) < dist(x_j, x_k)$, siendo x_i y x_j son de distintas clases.

4. Se itera todo el conjunto de datos, a fin de identificar todos los enlaces Tomek y suprimirlos.
5. Después de la eliminación de los enlaces Tomek, se procede a realizar el sobremuestreo de la clase minoritaria, con el objetivo de obtener un nuevo conjunto de datos balanceado. (p.15)

Figura 6

Balanceo mediante SMOTE-Tomek



Nota. Se observa que primero se suprimen observaciones de la clase amarilla en la región lila y posteriormente se realiza el sobremuestreo de la clase morada. Capturado de Ilustración del Resampling [Imagen], por Áridas et al., 2017, Imbalanced learn (http://jmlr.org/papers/v18/16-365.html).

2.6 Selección de variables

Dinov (2018) explica que las técnicas de selección de variables se centran en métodos de filtro, métodos envolventes y métodos integrados. Los métodos de filtrado se enfocan en seleccionar los *features* (variables) con altos valores basados en X^2 o en la evaluación de las interacciones entre las variables candidatas y la variable respuesta, basándose en estadísticas como la

correlación, encontrándose dentro de esta categoría el *Information gain*. Los métodos envolventes realizan procesos iterativos para incorporar o eliminar variables, incluye la selección *forward, backward elimination y stepwise*. Los métodos integrados pueden utilizar varios clasificadores, predictores o procedimientos de *clustering*, estos métodos se realizan durante el proceso de construcción del modelo, es decir, la selección de las variables se va realizando en cada iteración del entrenamiento del modelo, como ejemplo de esta técnica se encuentra los árboles de decisión, *random forests*, selección de características usando *weighted-Support Vector Machine (SVM)*.

En relación con los métodos mencionados, un indicador de gran utilidad para evaluar la relevancia de aquellas variables seleccionadas es el *Information Value* (IV), Barddal et al. (2020) indicaron que tal métrica permite medir individualmente la relación de cada variable predictora y la variable dependiente. En esto sentido, considerando que la variable x_i es dividida en j particiones, se podría calcular el valor IV de x_i de acuerdo con:

$$IV(x_i) = \sum_j (\%\ de\ no\ default\ en\ j - \%\ de\ default\ en\ j) \times WOE(x_i, j),$$

siendo el $WOE(x_i, j) = \ln\left(\frac{\#\ eventos\ de\ no\ default\ in\ j}{\#\ de\ default\ en\ j}\right)$.

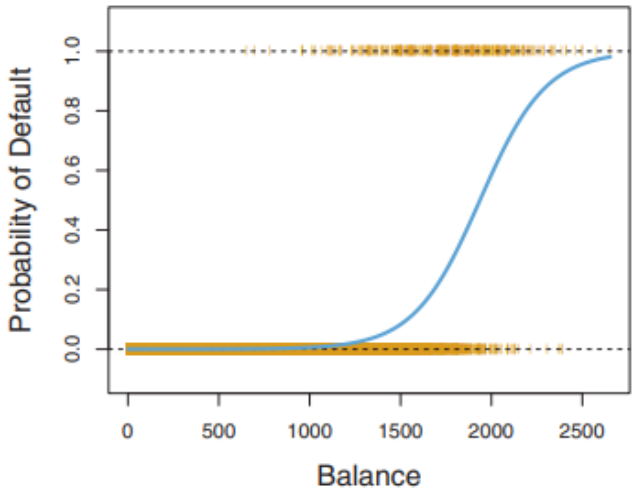
2.7 Regresión logística

El modelo de regresión logística estudia la relación entre variables explicativas y la variable dependiente dicotómica. El uso más común del modelo en mención en la industria bancaria es predecir la probabilidad de incumplimiento de pago (Miyamoto, 2014). Por consiguiente, dicha probabilidad asociada a cada cliente permite lograr un ordenamiento del riesgo, lo cual es conocido como credit scoring.

La regresión logística requiere que la variable dependiente sea dicotómica y utiliza para el cálculo de la probabilidad la función logística, la cual presenta una curva sigmoide dado que permite obtener como resultado probabilidades que se encuentran entre 0 y 1, como se muestra en la figura 7.

Figura 7

Ejemplo de ajuste de una curva logística



Nota. El gráfico de la curva logística muestra que independientemente del valor que tome la variable Balance, la Probabilidad del Default no toma valores menores a 0 ni mayores a 1. Adoptado de *An Introduction to Statistical Learning with Applications in R* (p.131), por T. Hastie et al., 2013.

En línea con lo mencionado, se plantea una variable aleatoria Y, la cual sigue una distribución Binomial, $Y \sim Bin(n, \pi)$. En vista de que la variable respuesta es dicotómica no es adecuado el uso de una regresión lineal, ya que este modelo ajustaría una recta o hiperplano y ello podría estimar valores diferentes a 0 y 1. Por tanto, se utiliza una función sigmoide que permita realizar estimaciones de probabilidad dentro del intervalo [0,1].

$$p(Y = 1) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} \quad (1)$$

Considerando la presencia de variables predictoras, Hastie et al. (2013) mostraron la función logística (2) y el odds ratio (3), el cual compara la probabilidad de que ocurra el evento estudiado con la probabilidad que no ocurra.

$$p(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 x_1}} \quad (2)$$

$$\frac{p(Y=1)}{1-p(Y=1)} = e^{\beta_0+\beta_1 x_1} \quad (3)$$

Del logaritmo de la ecuación (3) se obtiene el log odds, también llamado logit, el cual según se observa en la ecuación (4) es lineal en X. Asimismo, Hastie et al. (2013) mencionaron la relación entre el coeficiente estimado β_1 y $p(Y = 1)$ cuando X incrementa en una unidad, esto es, si β_1 es positivo entonces $p(Y = 1)$ incrementará, mientras que si β_1 es negativo existirá una disminución en $p(Y = 1)$.

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 x_1 \quad (4)$$

A través del método de máxima verosimilitud se realiza la estimación de los coeficientes, donde β_0 es el intercepto y β_1 mide el cambio promedio del logit al incrementar una unidad de x_1. Además, a medida de extrapolación, se presenta la ecuación de la regresión logística múltiple (5) y la ecuación para realizar predicciones en (6).

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (5)$$

$$p(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \quad (6)$$

2.8 Evaluación del modelo

Para la evaluación de modelos de clasificación se pueden utilizar diferentes métricas, de acuerdo con el objetivo de la construcción del modelo. En relación con ello, es de gran utilidad la matriz de confusión que permite comparar las clases observadas y predichas y, en consecuencia, construir métricas de evaluación del performance.

Tabla 1

Matriz de confusión

Clase predicha	Clase observada	
	Moroso	No moroso
Moroso	TP	FP
No Moroso	FN	TN

Nota. Se presenta de modo ilustrativo los resultados de un modelo de clasificación para predecir a los clientes como “Moroso” y “No moroso”.

La matriz de confusión permite evaluar el evento de interés; en específico en el credit scoring el evento de interés es detectar a los clientes “Morosos” y “No morosos”. En relación con ello, la figura 3 está compuesta por:

- True Positive (TP): Cantidad de observaciones que son “Moroso” y que fueron predichos correctamente como “Moroso”.
- True Negative (TN): Cantidad de observaciones que son “No moroso” y que fueron predichos correctamente como “No moroso”.
- False Negative (FN): Cantidad de observaciones que son “Moroso” y que fueron incorrectamente predichos como “No moroso”.
- False Positive (FP): Cantidad de observaciones que son “No moroso” y que fueron incorrectamente predichos como “Moroso”.

De la matriz presentada se construyen las principales métricas de evaluación tales como precisión (7), que representa la tasa de observaciones que fueron correctamente; sensibilidad (8), que mide la proporción de observaciones correctamente clasificadas respecto a la clase objetivo y la especificidad (9) que mide la proporción de observaciones correctamente clasificadas respecto a la clase de falla,

$$precisión = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

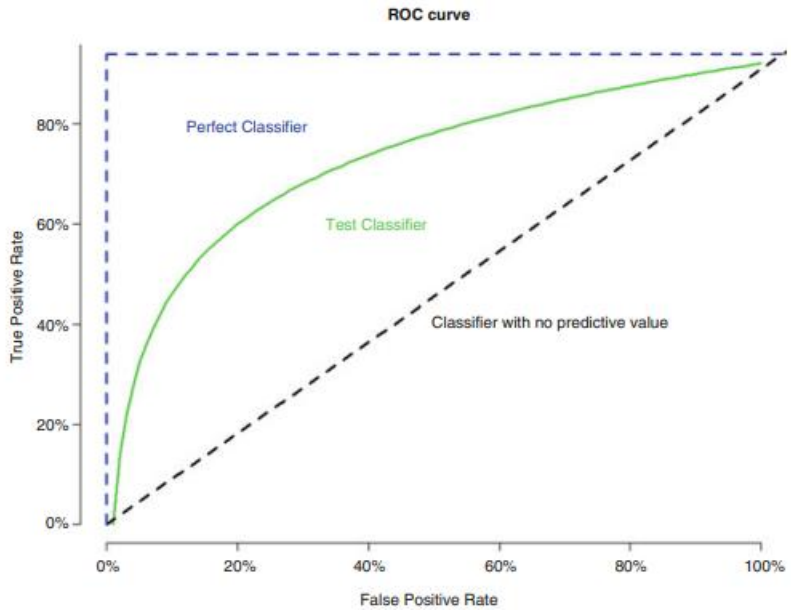
$$sensibilidad = \frac{TP}{TP+FN} \quad (8)$$

$$especificidad = \frac{TN}{TN+FP} \quad (9)$$

Adicionalmente, Dinov (2018) sugiere utilizar la curva Receiver Operating Characteristic (ROC) para evaluar clasificadores, el cual permite visualizar el tradeoff entre el ratio de falsos positivos y el ratio de verdaderos positivos. En la figura 8 se puede observar que cuando el clasificador es perfecto presenta 0% de falsos positivos y 100% de verdaderos positivos (línea punteada azul); no obstante, en la realidad el performance de los modelos de clasificación se asemeja a la curva verde.

Figura 8

Curva ROC



Nota. Tomado de *Data Science and Predictive Analytics* (p.488), por I. Dinov, 2018.

Asimismo, Dinov (2018) indica los valores estándar del área bajo la curva ROC para evaluar modelos de clasificación, donde los valores que se encuentran entre 0.5 y 0.6 representan a modelos que no discriminan, si los valores que se encuentran entre 0.6 y 0.7 significa que el modelo tiene una pobre discriminación, los valores entre 0.7 y 0.8 implican una discriminación aceptable, entre 0.8 y 0.9 significan una discriminación excelente y por encima del 0.9 presenta una discriminación perfecta.

III. DESARROLLO DEL TRABAJO

3.1 Delimitación del trabajo

Para la elaboración del presente trabajo se consideró a todos los clientes de la campaña "Preaprobado" del periodo 2017 y abarcó cada una de las sucursales de la entidad financiera, las cuales de acuerdo con el nicho de mercado definido por la organización se encuentran dispersas en Lima Norte (Carabayllo, Comas, Puente Piedra), Lima Sur (Lurín, San Juan de Miraflores, Villa María del Triunfo, Villa El Salvador), Lima Este (Ate, San Juan de Lurigancho, Huaycán), la Provincia Constitucional del Callao y la mayoría de los departamentos, a excepción de Apurímac, Huancavelica y Moquegua.

En relación con el tipo de investigación, el estudio es de naturaleza descriptivo debido a que dentro de los objetivos se pretendió identificar aquellas variables relevantes que influyeron en el comportamiento de pago de los clientes de la campaña en cuestión y es de tipo explicativo dado que, una vez identificado el conjunto de variables importantes, éste debe ser capaz de explicar y predecir a aquellos clientes de la campaña Pre Aprobado que caerán en incumplimiento de pago.

3.2 Fuentes de información

En primera instancia es importante comprender los campos que conforman el RCC, el cual es el insumo principal para el modelamiento y se constituye de dos bases de datos en su versión original (sin tratamiento). En la tabla 2 se puede observar los campos de la primera base, referida a información de los deudores en el sistema financiero (dícese así de toda persona que cuenta con algún tipo de préstamo), en la cual se tiene los campos: codigosbs, tipo de persona jurídica, tipo de documento de identidad, documento de identidad, calificación crediticia visto por persona y número de identidades financieras. La segunda base (tabla 3) contiene información de la deuda e incluye los siguientes campos: codigosbs, entidad financiera, tipo de crédito, días de mora, saldo, cuenta contable y calificación crediticia por entidad; esto implica que, un mismo deudor tendrá la cantidad de registros correspondientes a la combinación de número de entidades y cuentas contables asociadas.

Tabla 2

Tabla resumen del RCC

Variable	Definición	Valores	Tipo de variable
CODIGOSBS	Es el código que la SBS le asigna a cada deudor en el sistema financiero.	Códigos numéricos únicos para cada deudor.	Nominal
Tipo documento identidad	Tipo de identificación del deudor	1= DNI, 2= Carné de extranjería, 5= Pasaporte	Nominal
Documento de identidad	Número de identificación	-	Nominal
Tipo de persona	Indica si el deudor es persona natural u otro tipo.	1= Persona Natural, 2= Persona Jurídica, 3= Mancomunadas.	Nominal
Entidades	Número de entidades con las que el deudor cuenta un compromiso de pago.		Discreta
Clasificación Normal	Porcentaje en la que el deudor se encuentra en calificación "Normal", es decir, con morosidad no mayor a 8 días.	Valor entre 0 y 100.	Continua
Clasificación CPP	Porcentaje en la que el deudor se encuentra en calificación "CPP", es decir, con morosidad entre 9 y 30 días.	Valor entre 0 y 100.	Continua
Clasificación Deficiente	Porcentaje en la que el deudor se encuentra en calificación "Deficiente", es decir, con morosidad entre 31 y 60 días.	Valor entre 0 y 100.	Continua
Clasificación Dudoso	Porcentaje en la que el deudor se encuentra en calificación "Dudoso", es decir, con morosidad entre 61 y 120 días.	Valor entre 0 y 100.	Continua
Clasificación Pérdida	Porcentaje en la que el deudor se encuentra en calificación "Pérdida", es decir, con morosidad mayor a 120 días.	Valor entre 0 y 100.	Continua

Tabla 3

Tabla de saldos del RCC

Variable	Definición	Valores	Tipo de variable
CODIGOSBS	Es el código que la SBS le asigna a cada deudor en el sistema financiero.	Códigos numéricos únicos para cada deudor.	Nominal
Entidad	Código de la entidad asignado por la SBS.	Código numérico.	Nominal
Tipo de crédito	Dentro de cada entidad pueden existir varios créditos asociados a distintos tipos de crédito.	8= Medianas empresas, 9= Pequeñas empresas, 10= Microempresas, 11= Consumo revolvente, 12= Consumo no revolvente, 13= Hipotecario.	Nominal
Días de mora	Número de días de mora calculado desde la primera obligación de pago, según tipo de crédito y entidad.	-	Discreta

Saldo	Los saldos asociados a cada cuenta contable representado en soles.	-	Continua
Cuenta contable	Indica a qué categoría corresponde el saldo de deuda. Para cada entidad se presentan múltiples cuentas como por ejemplo colocaciones brutas, créditos castigados, entre otros.	Los primeros dígitos permiten identificar el tipo de cuenta contable, siendo las más comunes: 1401 = Créditos vigentes, 1403= Créditos reestructurados, 1404 = Refinanciados, 1405= Vencidos, 1406 = Cobranza judicial, 1408= Rendimientos devengados, 7205 = Líneas de crédito no utilizadas, 8103= Incobrables castigadas.	Nominal
Calificación	Calculado de acuerdo con los días de mora, según tipo de crédito y entidad.	0= Normal, 1= CPP, 2= Deficiente, 3= Dudoso, 4= Pérdida	Ordinal

Por otro lado, la variable objetivo o variable dependiente, que será utilizada en el estudio se define como el default temprano denominado First Payment Default (FPD). Esto es, un cliente tendrá la categoría de "FPD" cuando incumple el pago de su primera cuota y será "NOFPD" cuando no presenta atraso en su primera cuota.

En relación con ello, con el modelo de regresión logística se busca predecir la probabilidad de que los clientes potenciales de una campaña crediticia sean FPD.

3.3 Procedimientos

Considerando la experiencia profesional de la bachillera en Estadística Informática, cuyo cargo fue de Analista de Riesgo de Crédito, le permitió identificar aquellos segmentos que iban aumentando el riesgo para la entidad, comunicando al Comité de Riesgos los resultados mensuales de los diferentes tipos de mora, los cuales se realizaban utilizando indicadores y/o análisis como cosechas, migraciones (rollrates), alertas tempranas, cálculo del KS para evaluación de modelos, entre otros.

En particular, en el monitoreo de la mora temprana se detectó el incremento del FPD para la Campaña Pre Aprobado, el cual correspondía al mejor segmento, y por ello, la Gerencia General

comunicó la necesidad de realizar cambios en dicha campaña. Por esta razón, se hizo la propuesta de construir un modelo que permita predecir aquellos clientes potenciales que presentarían atraso en el cumplimiento de pago.

Para ello, se siguió la siguiente metodología (el flujo de procedimientos se puede observar en la figura 10):

a. Reingeniería de variables.

En esta etapa se buscó crear un nuevo set de variables en base a los campos proporcionados en el Reporte Crediticio Consolidado (RCC). Es importante mencionar que esta etapa fue de mucha importancia para poder tener mayor diversidad de variables candidatas y posteriormente identificar las más importantes; considerando que la base inicial proporcionada por la SBS contaba con pocos campos, lo que limitaba el proceso de aprendizaje del modelo. El proceso de reingeniería de variables se realizó utilizando el software Microsoft SQL Server, el cual es ideal para manejo de base de datos.

Para la explotación de la información fue necesario trabajar previamente las bases proporcionadas, a fin de obtener una sola tabla en la que se consolide la información a nivel cliente y, por ende, obtener información de su calificación final como deudor (Normal, CPP, Deficiente, Dudoso, Pérdida), monto de deuda, monto de líneas crédito utilizada, montos de líneas de no utilizados, entre otros.

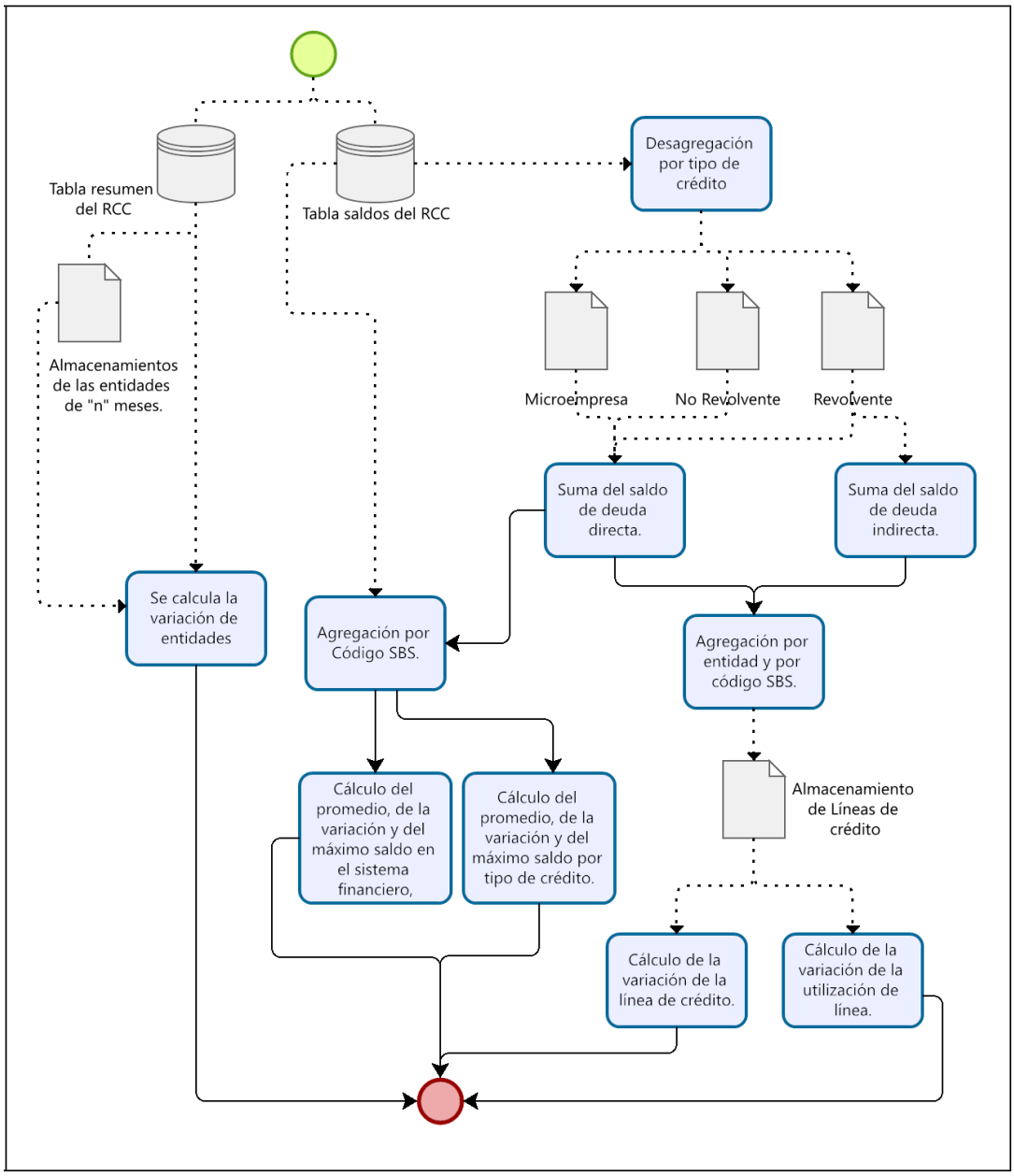
Mencionado lo anterior, como criterios para la construcción de la campaña Pre Aprobado se utilizaba información del RCC, limitándose a filtrar de la base de datos aquellas personas que contaban con un excelente historial crediticio considerando que hayan tenido 12 meses con calificación Normal (que hayan tenido un atraso de pago de máximo 8 días) y que cumplan con criterios de sobreendeudamiento de la entidad como un límite en el número de entidades.

Por ello, para cumplir el objetivo de construir un modelo que incorpore más información del sistema financiero, se consideró la construcción de variables que sirvan como candidatas en el

modelamiento. Para tal elaboración se requirió crear previamente tablas y variables intermedias (el flujo se puede visualizar en la figura 9), a fin de que puedan obtenerse los nuevos campos vistos por Código SBS.

Figura 9

Proceso de creación de variables



Las variables construidas se presentan a continuación.

Variación del número de entidades:

Funcionalidad: Tuvo como finalidad identificar si el deudor aumenta o disminuye la cantidad de entidades con las que tiene un compromiso de pago. Esta información es importante porque presenta una idea del nivel de endeudamiento de la persona.

Input y construcción: Se utilizó el campo Entidades de la tabla 1, el cual se encuentra a nivel de código SBS. Para ello, se tomó la cantidad de entidades de los últimos "n" meses, y posteriormente se calculó la variación en esos "n" meses. Por ejemplo, si se desea conocer cómo varió la cantidad de empresas en los últimos 3 meses, se compara el mes en el que se está evaluando con la cantidad de empresas de los dos meses preliminares, siguiendo la siguiente formula:

$$Variación\_Entidades\_n = (\frac{Entidades_{final}}{Entidades_{inical}} - 1) \times 100$$

Variables resultantes para ser candidatas:

- Variación en el número de entidades en los últimos 3 meses.
- Variación en el número de entidades en los últimos 6 meses.

Variación del saldo por tipo de crédito

Funcionalidad: Posibilitó identificar el incremento o decremento del endeudamiento de las personas en los meses próximos a la solicitud del crédito para cada tipo de crédito, siendo los créditos Microempresas dedicados a financiar actividades de producción, comercialización u otros servicios y los créditos Revolventes y No revolventes los asociados a atender gastos no empresariales.

Input y construcción: Se utilizó el campo Tipo de crédito de la tabla 2, para desagregar el saldo por créditos Microempresa, Revolvente y No revolvente. Luego, se utilizaron las cuentas contables de orden 14, las cuales reflejaban el saldo de créditos vigentes, restructurados, refinanciados y vencidos. Es importante mencionar que la tabla 2 contiene registros por cada

cuenta contable, por cada tipo de crédito y por cada entidad; por ello, fue necesario unificar los saldos a nivel de código SBS. Después se tomó el promedio de saldos para los últimos "n" meses y el promedio para los primeros "(12-n)" meses, con el fin de calcular su variación.

$$Variación\_Saldo\_Microempresa\_n = (\frac{PromedioMicroempresa_n}{PromedioMicroempresa_{12-n}} - 1) \times 100$$

$$Variación\_Saldo\_Revolvente\_n = (\frac{PromedioRevolvente_n}{PromedioRevolvente_{12-n}} - 1) \times 100$$

$$Variación\_Saldo\_NoRevolvente\_n = (\frac{PromedioNoRevolvente_n}{PromedioNoRevolvente_{12-n}} - 1) \times 100$$

En el caso del saldo total en el sistema financiero no se realizó la distinción por tipo de crédito, se limita el cálculo a tomar el saldo y agruparlo a nivel de código SBS.

Variables resultantes para ser candidatas:

- Variación del saldo en el tipo de crédito Microempresa en los últimos 3 meses.
- Variación del saldo en el tipo de crédito Microempresa en los últimos 6 meses.
- Variación del saldo en el tipo de crédito Revolvente en los últimos 3 meses.
- Variación del saldo en el tipo de crédito Revolvente en los últimos 6 meses.
- Variación del saldo en el tipo de crédito No Revolvente en los últimos 3 meses.
- Variación del saldo en el tipo de crédito No Revolvente en los últimos 6 meses.
- Variación del saldo total en el sistema financiero en los últimos 3 meses.
- Variación del saldo total en el sistema financiero en los últimos 6 meses.

<u>Promedio del saldo por tipo de crédito</u>

Funcionalidad: Permitió determinar el nivel de deuda actual por cada tipo de crédito. Con el objetivo de no tomar un solo valor al momento de la evaluación, se consideró el promedio.

Input y construcción: En primer lugar, se utilizó el campo Tipo de crédito de la tabla 2, para desagregar el saldo por créditos Microempresa, Revolvente y No revolvente. A la vez, como se quería utilizar la deuda actual, se utilizaron las cuentas contables de orden 14, las cuales

reflejaban el saldo de créditos vigentes, restructurados, refinanciados y vencidos. Posteriormente, se realizó la unificación del saldo existente para cada tipo de crédito y visto por código SBS. Finalmente, se calculó el promedio del saldo de los últimos "n" meses previos a la solicitud de crédito.

$$Prom\_Microempresa\_n = \frac{SaldoMicroempresa_1 + \cdots + SaldoMicroempresa_n}{n}$$

$$Prom\_Revolvente\_n = \frac{SaldoRevolvente_1 + \cdots + SaldoRevolvente_n}{n}$$

$$Prom\_NoRevolvente\_n = \frac{SaldoNoRevolvente_1 + \cdots + SaldoNoRevolvente_n}{n}$$

En el caso del saldo total en el sistema financiero no se realiza la distinción por tipo de crédito, se limita el cálculo a tomar el saldo y agruparlo a nivel de código SBS.

Variables resultantes para ser candidatas:

- Promedio de saldo del tipo de crédito Microempresa en los últimos 3 meses.
- Promedio de saldo del tipo de crédito Microempresa en los últimos 6 meses.
- Promedio de saldo del tipo de crédito Revolvente en los últimos 3 meses.
- Promedio de saldo del tipo de crédito Revolvente en los últimos 6 meses.
- Promedio de saldo del tipo de crédito No Revolvente en los últimos 3 meses.
- Promedio de saldo del tipo de crédito No Revolvente en los últimos 6 meses.
- Promedio del saldo total en el sistema financiero en los últimos 3 meses.
- Promedio del saldo total en el sistema financiero en los últimos 6 meses.

Saldo Máximo por tipo de crédito

Funcionalidad: Permitió identificar los picos altos de deuda en el sistema financiero.

Input y construcción: Análogamente a las variables precedentes, se calculó el saldo por tipo de crédito y a nivel de código SBS para cada uno de los últimos "n" meses. Posteriormente, se calcula el valor máximo de los últimos "n" meses.

$$Max\_Saldo\_Microempresa\_n = \max\left(SaldoMicroempresa_1, \dots, SaldoMicroempresa_n\right)$$

$$Max\_Saldo\_Revolvente\_n = \max\left(SaldoRev_1, \dots, SaldoRev_n\right)$$

$$Max\_Saldo\_NoRevolvente\_n = \max\left(SaldoNoRev_1, \dots, SaldoNoRev_n\right)$$

En el caso del saldo total en el sistema financiero no se realizó la distinción por tipo de crédito, se limita el cálculo a tomar el saldo y agruparlo a nivel de código SBS.

Variables resultantes para ser candidatas:

- Máximo de saldo en el tipo de crédito Microempresa en los últimos 3 meses.
- Máximo de saldo en el tipo de crédito Microempresa en los últimos 6 meses.
- Máximo de saldo en el tipo de crédito Revolvente en los últimos 3 meses.
- Máximo de saldo en el tipo de crédito Revolvente en los últimos 6 meses.
- Máximo de saldo en el tipo de crédito No Revolvente en los últimos 3 meses.
- Máximo de saldo en el tipo de crédito No Revolvente en los últimos 6 meses.
- Máximo de saldo total en el sistema financiero en los últimos 3 meses.
- Máximo de saldo total en el sistema financiero en los últimos 6 meses.

<u>Variación de la línea de crédito</u>

Funcionalidad: Evaluar el nivel crediticio, considerando que cada entidad del sistema financiero realiza evaluaciones autónomas de las solicitudes de crédito y de acuerdo con ello brindan las líneas crediticias. Las líneas de crédito corresponden únicamente a los créditos consumo revolvente.

Input y construcción: Se filtró los tipos de crédito con valor a 11, que corresponde a los créditos de consumo Revolvente. Después se sumó los montos de deuda directa con los montos de deuda indirecta para obtener el monto de la línea de crédito, utilizando los saldos asociados a las cuentas contables de orden 14 y orden 72 (Véase tabla 2). Tras ello, se tomó el promedio de las

líneas de crédito de los últimos “n” meses y se compara con el promedio de las líneas de crédito de los primeros “(12-n)” meses.

$$Variación\_Línea\_n = (\frac{PromedioLínea_n}{PromedioLínea_{12-n}} - 1) \times 100$$

Variables resultantes para ser candidatas:

- Variación de la línea del crédito Revolvente comparando el primer y segundo semestre.
- Variación de la línea del crédito Revolvente comparando los primeros 3 meses y los últimos 9 meses.

Variación de la utilización de la línea de crédito

Funcionalidad: Pretendió analizar la evolución de la utilización de las líneas de crédito.

Input y construcción: Se tomó como referencia el monto calculado de línea de crédito con los pasos precedentes, y finalmente la utilización de la línea de crédito fue calculada como el ratio entre el monto de deuda directa (asociado a las cuentas contables de orden 14) y la línea de crédito. Posteriormente, se tomó el promedio de la utilización líneas de crédito de los últimos “n” meses y se comparó con el promedio de la utilización de las líneas de crédito de los primeros “(12-n)” meses.

$$Variación\_utilización\_n = (\frac{PromedioUtilización_n}{PromedioUtilización_{12-n}} - 1) \times 100$$

Variables resultantes para ser candidatas:

- Variación de la utilización de la línea del crédito Revolvente comparando el primer y segundo semestre.
- Variación de la utilización de la línea del crédito Revolvente comparando los primeros 3 meses y los últimos 9 meses.

b. Acotamiento de valores.

Se verificó si las variables presentaban valores atípicos mediante gráficos de cajas. Luego se realizó el cálculo de los percentiles 90, 95 y 99, a fin de determinar qué valor es el adecuado para acotar las variables.

c. Escalamiento de variables

Se realizó el escalamiento de variables mediante el método de mínimo y máximo, con la finalidad de que la construcción del modelo no se vea afectada por las diferentes escalas de variables, ya que hay variables referidas a montos de deuda, las cuales podían variar hasta los 5 millones, y otras referidas a variaciones, las cuales podían variar de -1 a 8.

d. Tratamiento de valores perdidos.

Se determinó la cantidad de valores perdidos en cada una de las variables candidatas generadas en el paso anterior. La imputación de estos valores se realizó mediante el método de medias no condicionadas.

e. Balanceo de la variable dependiente.

La naturaleza de la variable dependiente, que corresponde al default temprano, es no balanceada; debido a que existe una gran diferencia entre la participación de las clases de "default" y "no default". Esta situación podría generar que el modelo estadístico realice mejores predicciones en la clase de "no default"; sin embargo, el interés central de la entidad financiera es poder determinar la probabilidad de que un cliente sea "default". De acuerdo con lo mencionado, en esta etapa se ejecutaron las técnicas de balanceo de datos de Random Undersampling, Random Oversampling, Tomek link y SMOTE, con la finalidad de que el modelo a ser construido aprenda y realice predicciones adecuadas en ambas clases.

f. Selección de variables importantes.

Del set de variables candidatas, construidas mediante reingeniería, se utilizó Random Forest y se seleccionó aquellas que presentaron mayor importancia relativa, tomando como referencia las diferentes técnicas de balanceo. Posteriormente, se revisó aquellas variables más relevantes obtenidas de las combinaciones de los 04 métodos de balanceo y bajo criterio de experto se realizó una depuración de aquellas variables que tienen un significado similar, con la finalidad de seleccionar las realmente más importantes, y a la vez, no incluir información redundante en el modelo de score.

g. Construcción del modelo de regresión logística

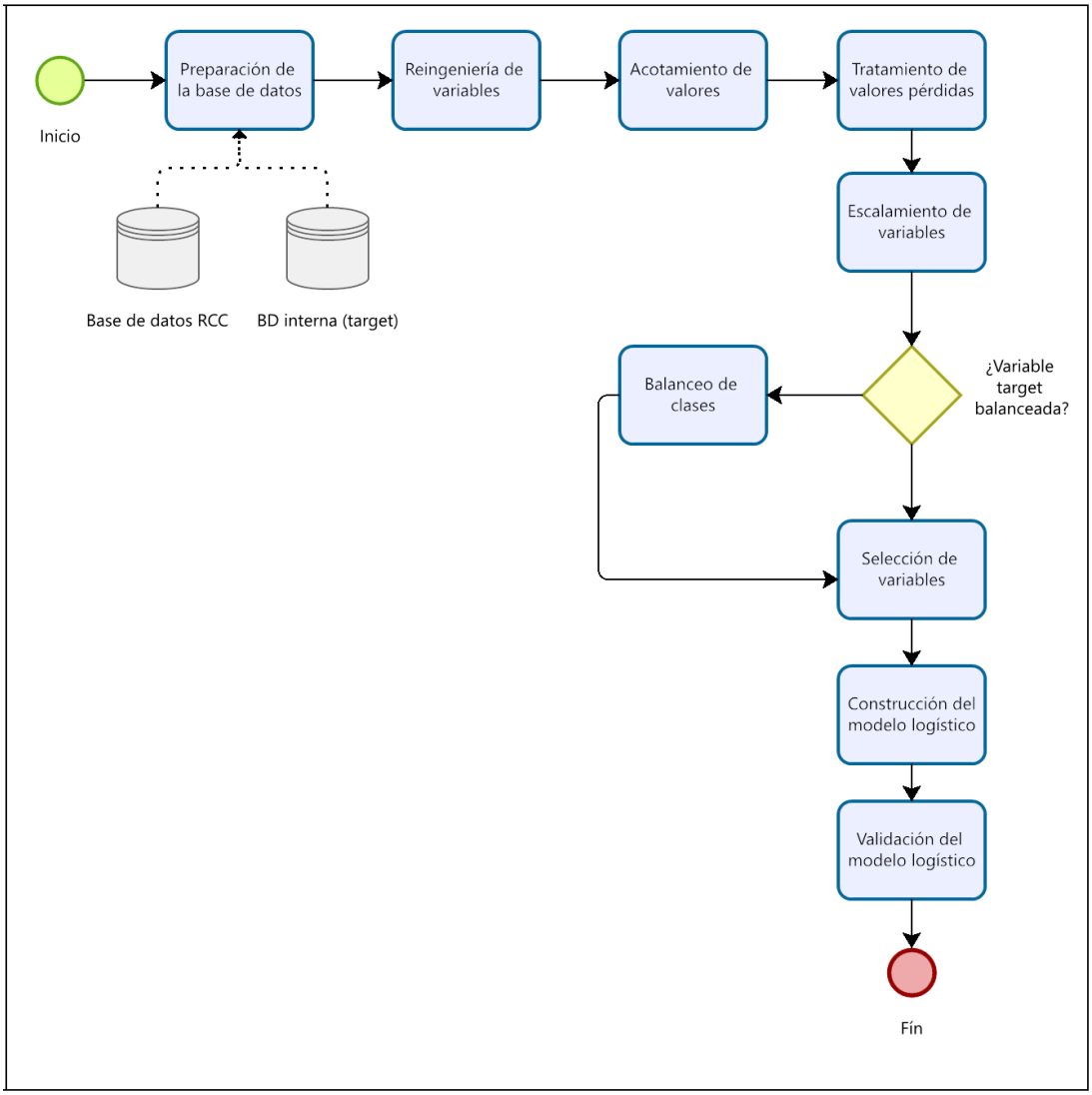
Se realizó la estimación de parámetros del modelo de regresión logística; asimismo, se evaluó que las variables finalistas aporten significativamente y que el ajuste del modelo sea adecuado.

h. Validación del modelo de regresión logística.

En la última etapa de la metodología se verificó que el modelo logístico presente un buen performance en una muestra diferente a la utilizada en la etapa de construcción. Por esta razón, se utilizó una muestra de validación, utilizando indicadores como la tasa de correcta clasificación, la curva ROC, entre otros.

Figura 10

Flujo de procedimientos



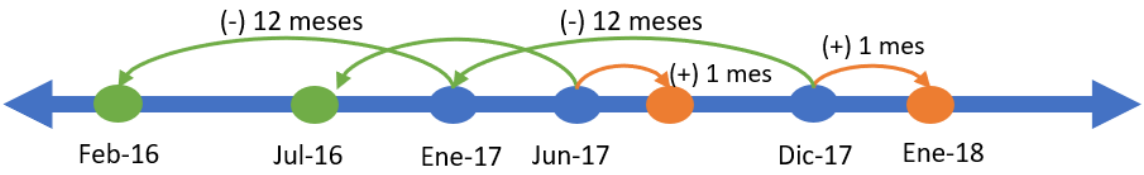
IV. RESULTADOS Y DISCUSIÓN

La base de datos utilizada corresponde a los clientes pertenecientes a la Campaña Pre Aprobado del año 2017, contando con un total de 38,062 de registros; evaluando el comportamiento de dichos clientes de acuerdo con la mora temprana, denominada FPD. Para ello, se tomó una ventana de tiempo de 1 año (enero a diciembre 2017) y el análisis de variables se realizó con la información del sistema financiero de 1 año antes de la adquisición del préstamo y, por último, para evaluar la mora temprana se evalúa si al siguiente mes después del préstamo (maduración del crédito = 1 mes), el cliente presentó atraso en sus pagos (FPD) o si pagó al día su primera cuota (NO FPD).

En la figura 11 se presenta el esquema de construcción de la base de modelamiento, donde se tiene que las cosechas de enero a diciembre del 2017 son representadas por los puntos azules, las flechas verdes indican los 12 meses que se debe retroceder para empezar a analizar su información financiera y las fechas naranjas indican el mes de evaluación que es el mes posterior a la cosecha; por ejemplo, para evaluar el comportamiento de la cosecha de junio del 2017, se retrocedió a julio del 2016 y se utilizó los 12 meses posteriores (julio 2016 a junio 2017) para analizar la información del sistema financiera encontrada en el RCC, y para evaluar si los clientes de dicha cosecha en estudio resultaron FPD o No FPD se verificó su morosidad en el mes de julio del 2017.

Figura 11

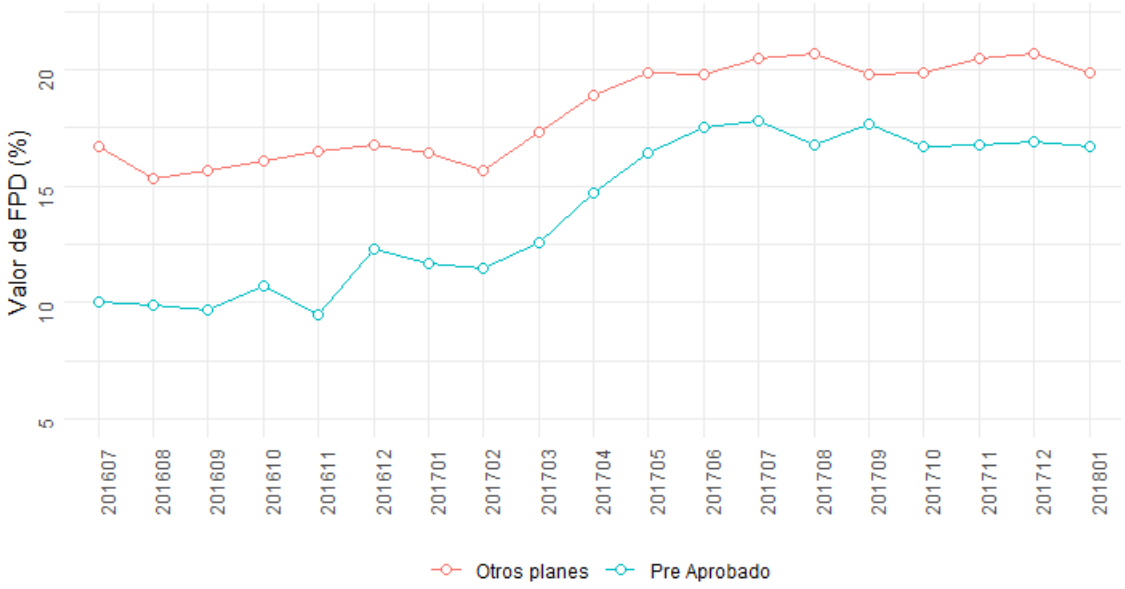
Ventanas de desempeño



La motivación para la construcción del modelo fue el incremento considerable del FPD de la campaña Pre Aprobado, el cual se acentuó a partir de abril del 2017 (tal como se muestra en la figura 12), acercándose al FPD del resto de planes financieros de la entidad. Esta situación es indeseable, debido a que la campaña utilizaba filtros más exigentes como contar con un buen comportamiento financiero en los últimos 12 meses, a diferencia del resto de planes que incluso son específicos para personas sin historial en el sistema financiero. De acuerdo con el presupuesto definido por la organización el FPD de la campaña en mención debía ser del 10%, y dado que este indicador sugiere cuáles serán los resultados de otros indicadores de morosidad, fue necesario realizar modificaciones y más controles en relación con la campaña, tanto a nivel de operatividad en la admisión de clientes como en la identificación de variables que intervienen en la predicción del default.

Figura 12

Evolución del FPD

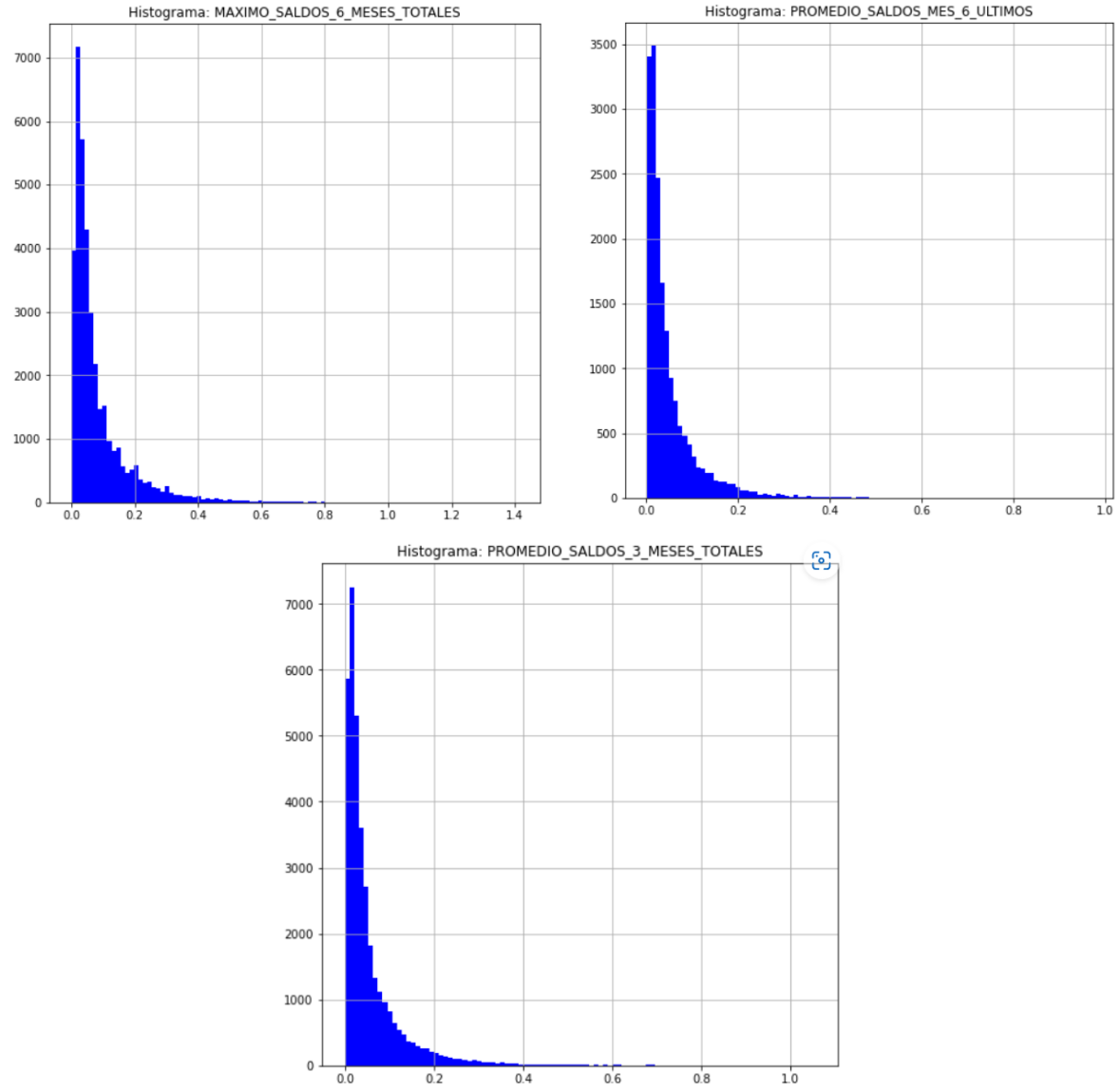


Posterior a la creación del conjunto de variables se realizó la revisión de gráficos univariados, con el fin de observar la asimetría y algunos picos en la distribución, así como estadísticos de tendencia central. En específico, se usó los histogramas para verificar que la distribución de cada variable no contradiga el sentido económico, es decir, si se esperaba por conocimiento de

la realidad que una variable tenga mayor concentración en valores pequeños, entonces el histograma debía mostrar asimetría positiva. A modo de ejemplo, en la figura 13 se presentan los histogramas de las variables “Saldo máximo en el sistema financiero en los últimos 6 meses” (MAXIMO_SALDOS_6_MESES_TOTALES), “Saldo promedio en crédito Microempresa en los últimos 6 meses” (PROMEDIO_SALDOS_MES_6_ULTIMOS), “Saldo promedio en el sistema financiero en los últimos 3 meses” (PROMEDIO_SALDOS 3_MESES_TOTALES), donde se corrobora que el comportamiento de tales variables va acorde a lo esperado.

Figura 13

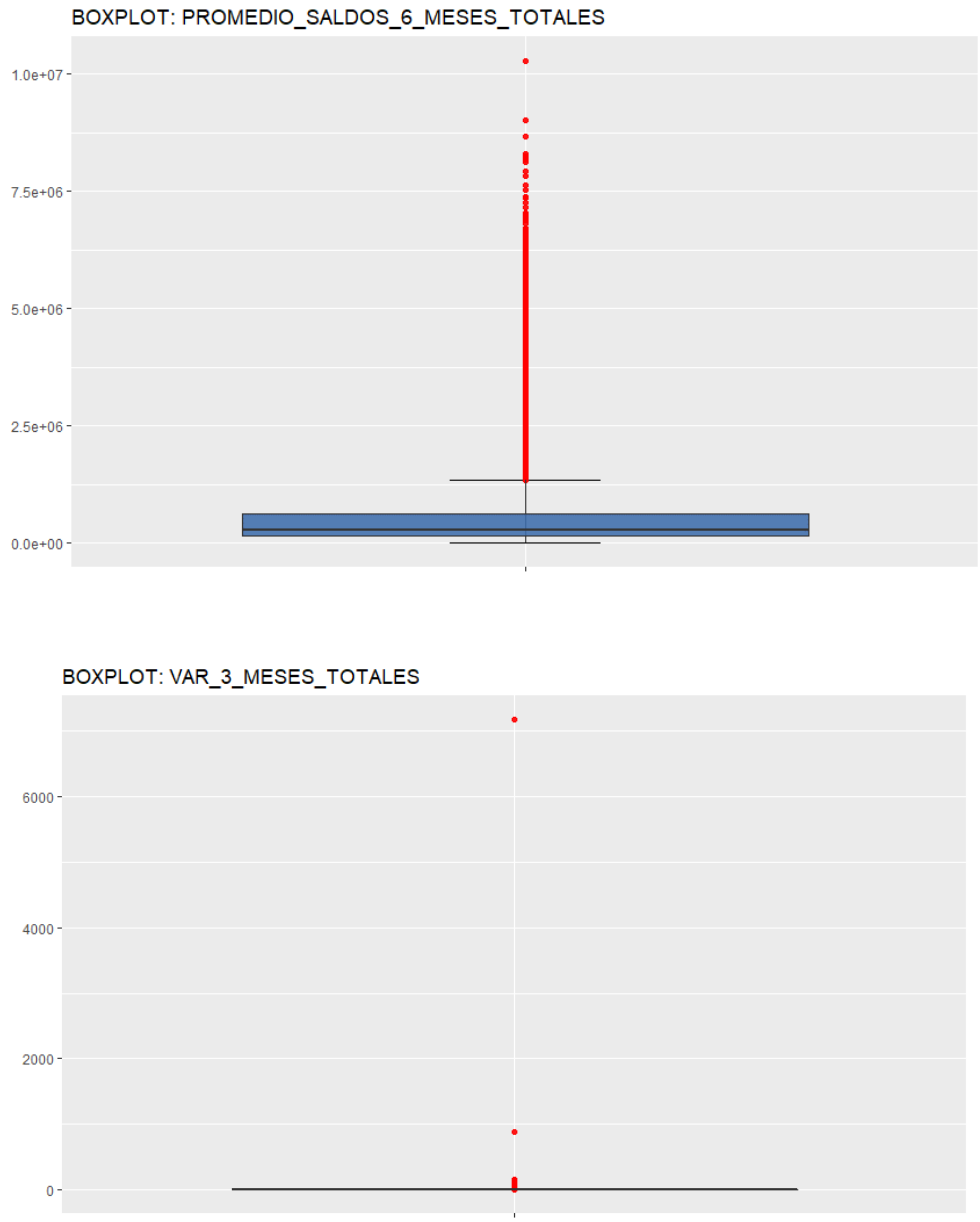
Histogramas



Adicionalmente, se revisó en gráficos de cajas si había presencia de datos atípicos en las variables candidatas, con el objetivo de acotarlas en caso sea necesario y que no afecten en la construcción del modelo. Como ilustración, en la figura 14 se observa que las variables contienen claramente valores no usuales; por ejemplo, la variable "Variación de la deuda en el sistema financiero en comparación con el último trimestre" (VAR_3_MESES_TOTALES) tiene una observación que toma el valor 1000 y otra con el valor de 7000, las cuáles se alejan fuertemente del resto de datos.

Figura 14

Boxplot



Complementariamente, se analizó los percentiles de cada una de las variables, a fin de poder determinar qué valor era el adecuado para utilizarse como cota superior. Al respecto, en la tabla 4 se observó que los valores atípicos se presentaban a partir del 1% de cola superior de los datos. Entonces, como parte del tratamiento de los datos se realizó el acotamiento de datos, considerando que todo valor que sea superior al percentil 99 de los datos, será reemplazado por dicho percentil. Es importante recalcar que no se realizó el acotamiento de valores inferiores, debido a que la naturaleza de las variables de deuda no permite valores negativos y cuando se realizó la revisión de valores mínimos, en todas las variables, se encontraban acorde a lo esperado.

Tabla 4

Estadísticos de variables candidatas

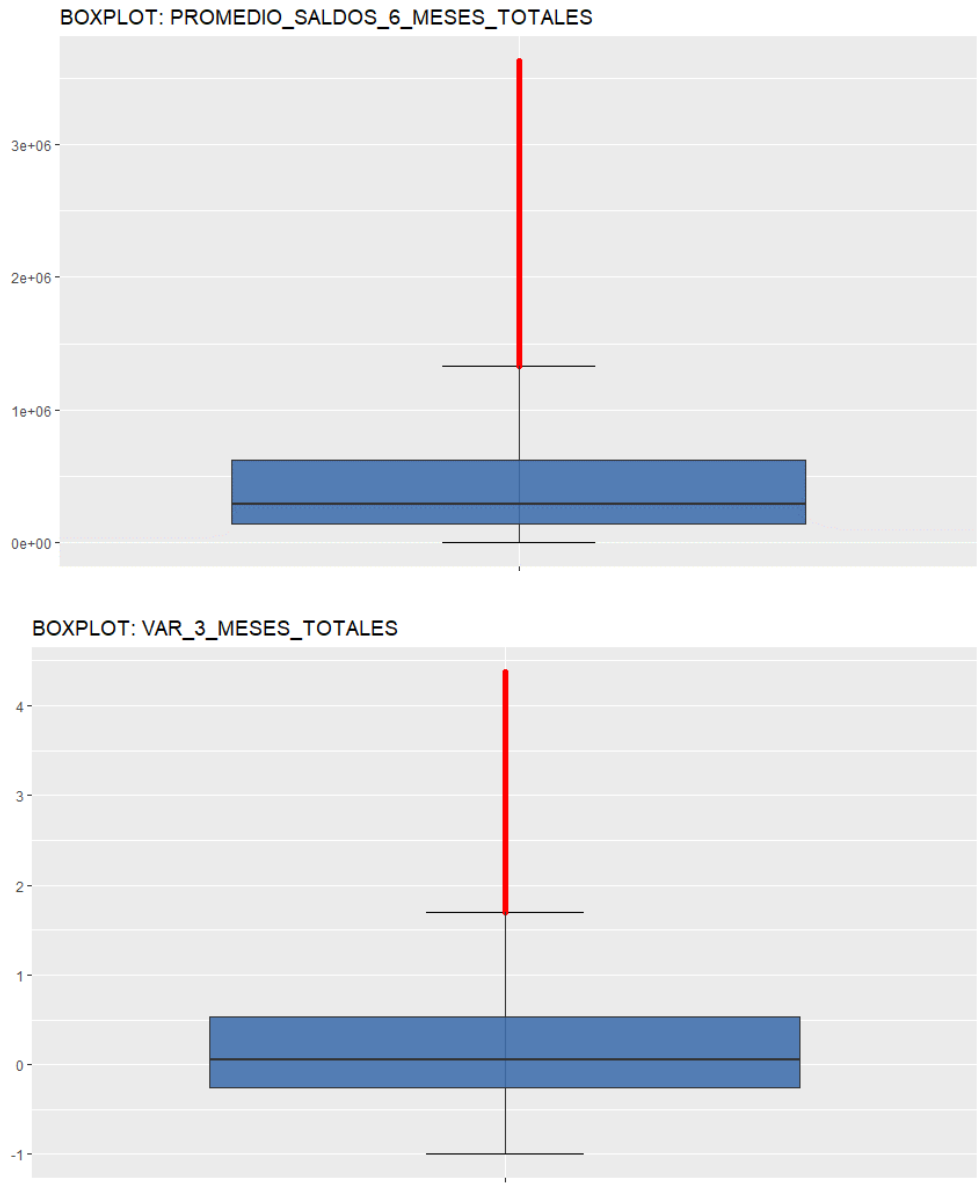
ESTADÍSTICOS	VAR_3_MESES_MES	VAR_6_MESES_MES	VAR_3_MESES_REV	VAR_6_MESES_REV	VAR_3_MESES_NREV	VAR_6_MESES_NREV	VAR_3_MESES_TOTALES	VAR_6_MESES_TOTALES
count	15590.000	14327.000	9048.000	7749.000	20807.000	19173.000	35761.000	34693.000
mean	0.274	0.416	1.908	9.824	0.209	0.347	0.538	1.789
std	0.809	1.302	84.068	742.077	1.360	2.643	38.294	229.103
min	-0.977	-0.967	-1.000	-0.999	-1.000	-0.957	-0.996	-0.997
25%	-0.243	-0.202	-0.229	-0.193	-0.321	-0.288	-0.253	-0.216
50%	0.090	0.133	0.049	0.077	-0.067	-0.044	0.048	0.088
75%	0.543	0.615	0.527	0.642	0.364	0.426	0.526	0.627
90%	1.180	1.353	1.484	1.709	1.053	1.251	1.213	1.438
95%	1.783	2.110	2.632	3.042	1.744	2.070	1.856	2.228
99%	4.098	5.111	8.118	9.474	4.743	5.761	4.375	5.596
max	4.099	32.507	7179.808	65204.000	55.205	220.599	7179.808	42519.500

Nota. En modo ilustrativo se presenta los estadísticos de las primeras variables.

En la figura 15 se presenta el gráfico de cajas de dos variables posterior al acotamiento realizado, en las que se puede observar el efecto de aplicar la limpieza de valores atípicos, comparándolo con la figura 14.

Figura 15

Boxplot post acotamiento



Asimismo, se revisó los datos faltantes en cada una las 30 variables (tabla 5) con la finalidad de que se realice el tratamiento adecuado. Era importante identificar qué variables contienen propiamente valores faltantes o si se debe a casos en los que verdaderamente el valor de la variable debía ser igual a 0. Como muestra, se observa que la variable "VAR_6_MESES_REV" presenta aparentemente una gran cantidad de missing; no obstante, esto se debe a que una persona bancarizada puede no tener todos los tipos de créditos y en este caso, la proporción de

observaciones sin datos corresponde a las personas que no presentan deuda del tipo Revolvente, por lo que se reemplazó el valor "null" por 0. Caso contrario, no es coherente que las variables referidas a deudas totales en el sistema financiero (tipo de crédito de Pequeña empresa, Mediana empresa, Microempresa, Revolvente, No Revolvente o Hipotecario) presenten valores igual a 0, dado que si una persona se encuentra en el RCC es porque debe tener algún compromiso de pago con alguna entidad.

Tabla 5

Identificación de datos faltantes

VARIABLES	N° MISSING	PORCENTAJE
VAR_6_MESES_REV	30313	0.796411
VARIACION_UTIL_REV_6_VS_6MESES	29294	0.769639
VARIACION_UTIL_REV_3_VS_9_MESES	29145	0.765724
VAR_3_MESES_REV	29014	0.762283
PROMEDIO_SALDOS_REV_3_ULTIMOS	27868	0.732174
PROMEDIO_SALDOS_REV_6_ULTIMOS	26836	0.70506
MAX_SALDOS_REV_3_ULTIMOS	25630	0.673375
MAX_SALDOS_REV_6_ULTIMOS	24945	0.655378
VAR_6_MESES_MES	23735	0.623588
VARIACION_LINEA_REV_6_VS_6MESES	23101	0.606931
VAR_3_MESES_MES	22472	0.590405
VARIACION_LINEA_REV_3_VS_9_MESES	21976	0.577374
PROMEDIO_SALDOS_MES_3_ULTIMOS	21551	0.566208
PROMEDIO_SALDOS_MES_6_ULTIMOS	20247	0.531948
MAX_SALDOS_MES_3_ULTIMOS	19781	0.519705
VAR_6_MESES_NREV	18889	0.496269
MAX_SALDOS_MES_6_ULTIMOS	18854	0.49535
VAR_3_MESES_NREV	17255	0.453339
PROMEDIO_SALDOS_NREV_3_ULTIMOS	16160	0.42457
PROMEDIO_SALDOS_NREV_6_ULTIMOS	13327	0.350139
MAX_SALDOS_NREV_3_ULTIMOS	12596	0.330934
MAX_SALDOS_NREV_6_ULTIMOS	10956	0.287846
VAR_6_MESES_TOTALES	3369	0.088513
VAR_3_MESES_TOTALES	2301	0.060454
PROMEDIO_SALDOS_3_MESES_TOTALES	2202	0.057853
PROMEDIO_SALDOS_6_MESES_TOTALES	929	0.024408
MAXIMO_SALDOS_3_MESES_TOTALES	807	0.021202
MAXIMO_SALDOS_6_MESES_TOTALES	571	0.015002
DIF_ENTIDADES_3_MESES	0	0
DIF_ENTIDADES_6_MESES	0	0

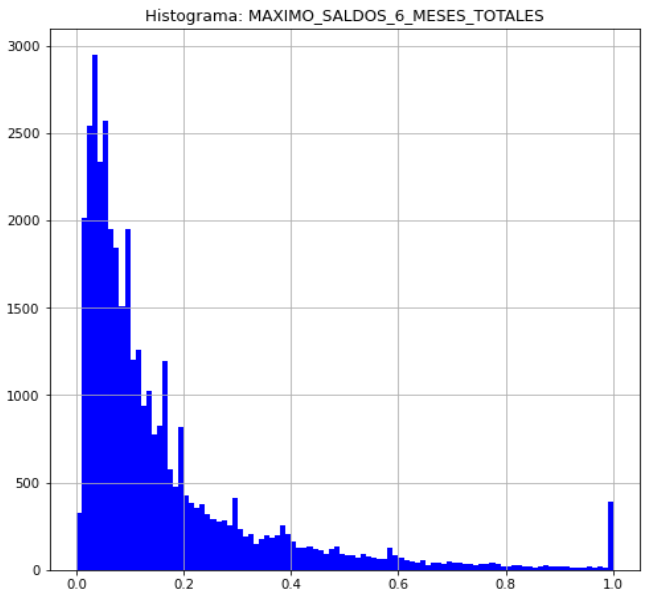
Se determinó aquellas únicas variables que se encontraban bajo la casuística de presentar valores faltantes, a las cuales se les realizó la imputación por la media (tal estadístico fue calculado posterior al acotamiento). Tales variables se presentan a continuación:

- “VAR_6_MESES_TOTALES”
- “VAR_3_MESES_TOTALES”
- “PROMEDIO_SALDOS_3_MESES_TOTALES”
- “PROMEDIO_SALDO_6MESES_TOTALES”
- “MAXIMO_SALDOS_3_MESES_TOTALES”
- “MAXIMO_SALDOS_6_MESES_TOTALES”

Respecto a los valores de las variables, existen algunas referidas al promedio de saldos, saldos máximos y variaciones que, aunque son variables continuas cuentan con diferentes escalas; así como variables referidas al número de entidades, que son variables discretas. Entonces, se procedió a utilizar el método de normalización mínimo – máximo, con el objetivo de que cada una de las 30 variables candidatas presenten la misma escala y se encuentran en el rango de [0,1], tal como se muestra en la figura 16.

Figura 16

Histograma



Nota. Se puede observar que posterior a la normalización, la distribución de la variable MAXIMO_SALDOS_6_MESES_TOTALES se encuentra en el rango de [0,1].

Posterior a ello, se realiza la división de la base de clientes en dos tipos de muestras: muestra de entrenamiento y test. Se consideró que el 70% de la base sea para entrenar el modelo y el

resto del 30% es para corroborar el performance del modelo, es decir, se tiene que 26,643 clientes serán usados el modelamiento y 11,419 para el contraste del modelo. Esta división se muestra en la tabla 6.

Tabla 6

Muestras de entrenamiento y test

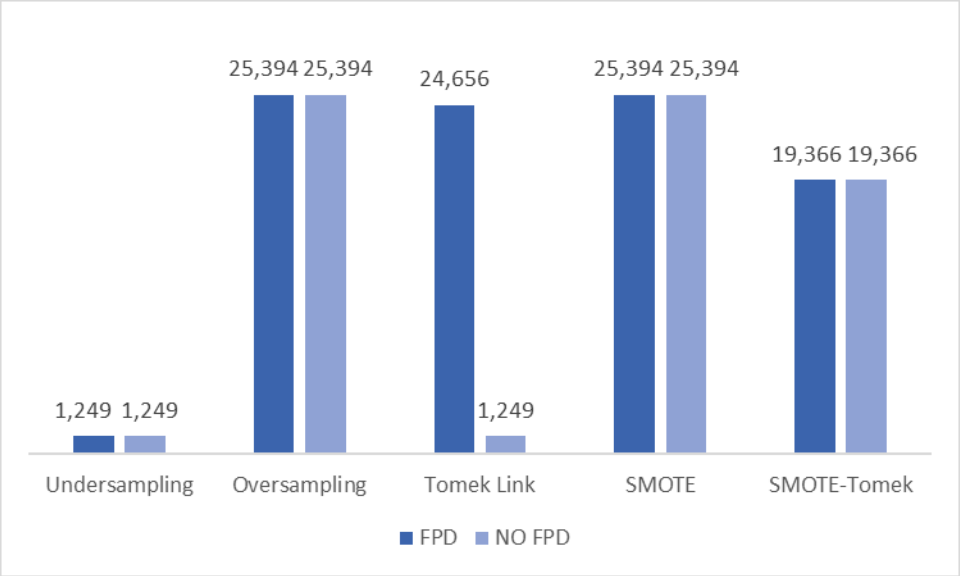
Respuesta	Entrenamiento	Test
NO FPD	25,394	10,883
FPD	1,249	536
Total	26,643	11,419

Previo al modelamiento, es importante realizar el balanceo de clases, para lograrlo se probaron 5 métodos de balanceo: undersampling, oversampling, Tomek link, SMOTE, SMOTE -Tomek. Al utilizar el método de submuestreo, la variable objetivo quedó balanceada con 1,249 registros, coincidiendo con el número de registros de la clase minoritaria (FPD); mientras que para el método de sobremuestreo (Oversampling, SMOTE), la variable objetivo se compuso por 25,394 registros correspondientes al total de la clase mayoritaria (NO FPD). En el caso de la técnica de SMOTE-Tomek se realizó el balanceo de clases, disminuyendo la cantidad de observaciones de la clase mayoritaria; mientras que con la técnica de Tomek Link, la proporción de desbalanceo no mejoró en gran medida.

Luego de contar con los 5 conjuntos de datos resultantes de los métodos de balanceo indicados, se procedió a probar qué variables de las 30 candidatas son las más adecuadas para clasificar si un cliente será FPD o NO FPD. Para lograr dicho objetivo se ajustó cada data a un modelo de Random Forest y como resultado se obtuvo la importancia asociada a cada variable, permitiendo identificar el orden de relevancia de cada una y facilitando la selección de variables. Por último, de las variables más relevantes de cada uno de los 5 casos se discutió cuáles serían las variables finalistas para probar en el modelo de regresión logística, basándose en aquellas que tuvieron mayor importancia relativa y de acuerdo con el visto bueno de la Gerencia de Riesgos.

Figura 17

Métodos de balanceo de clases

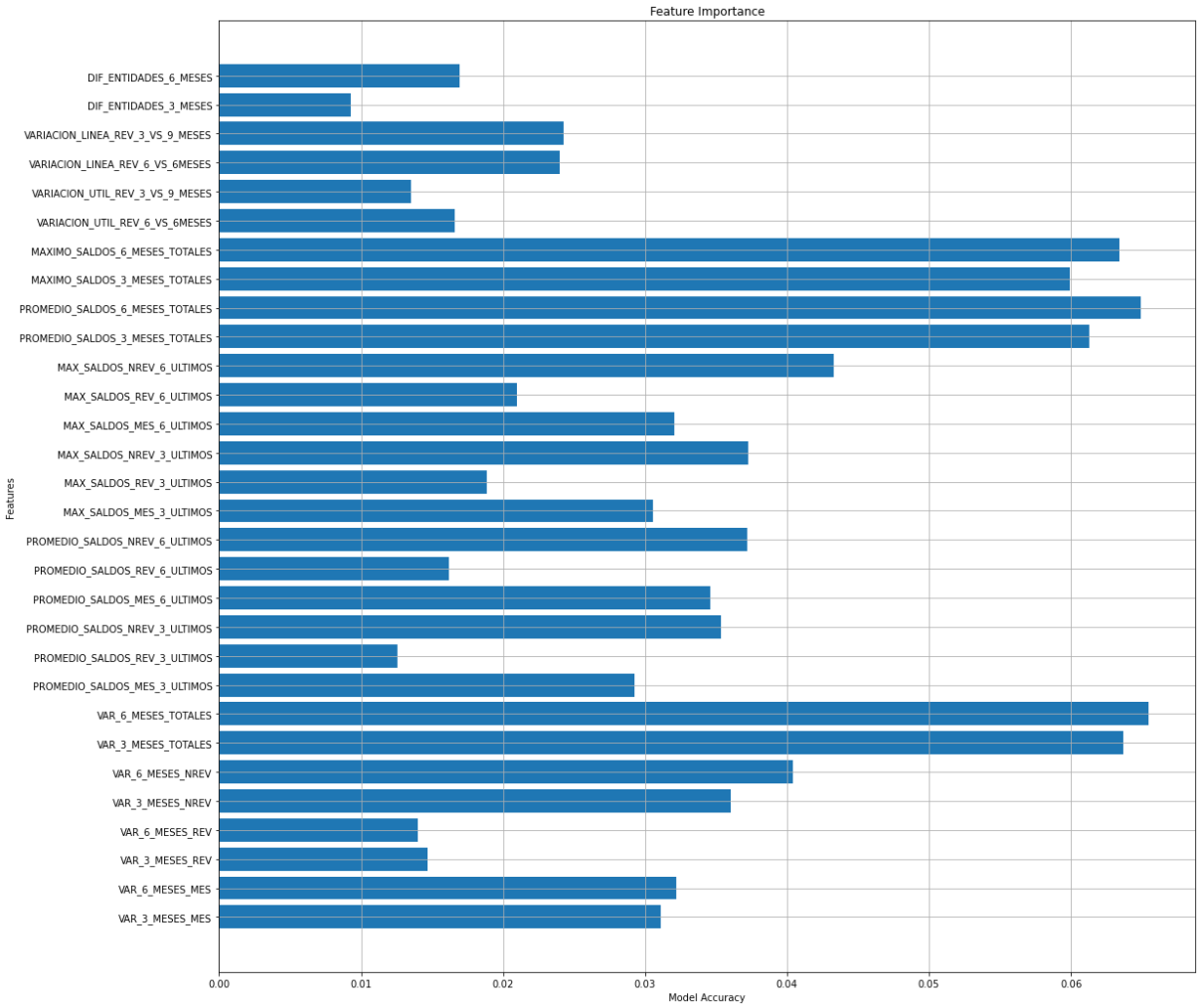


Nota. Los valores corresponden netamente al balanceo de la muestra de entrenamiento.

En primera instancia se ajustó el modelo de Random Forest a los datos balanceados por el método Undersampling. De tal ajuste se obtuvo que las variables con mayor importancia, como se muestra en la figura 18, fueron las de saldo máximo en el sistema financiero, para 3 y 6 meses (MAXIMO_SALDOS_6_MESES_TOTALES, MAXIMO_SALDOS_3_MESES_ TOTALES), el promedio de los saldos en el sistema financiero, para 3 y 6 meses (PROMEDIO_SALDOS_6_MESES_TOTALES, PROMEDIO_SALDOS_3_MESES_ TOTALES), la variación del saldo en el sistema financiero considerando el último trimestre y semestre (VAR_3_MESES_TOTALES, VAR_6_MESES_TOTALES), y continuando con el grado de relevancia se tiene a variables referentes al tipo de crédito No Revolvente (MAX_SALDOS_NREV_6_ULTIMOS, VAR_6_MESES_NREV). Es importante indicar que para la selección de variables finalistas se consideró no incluir información redundante, por ejemplo, no incluir variables que presenten el mismo concepto, pero con diferente umbral de cálculo.

Figura 18

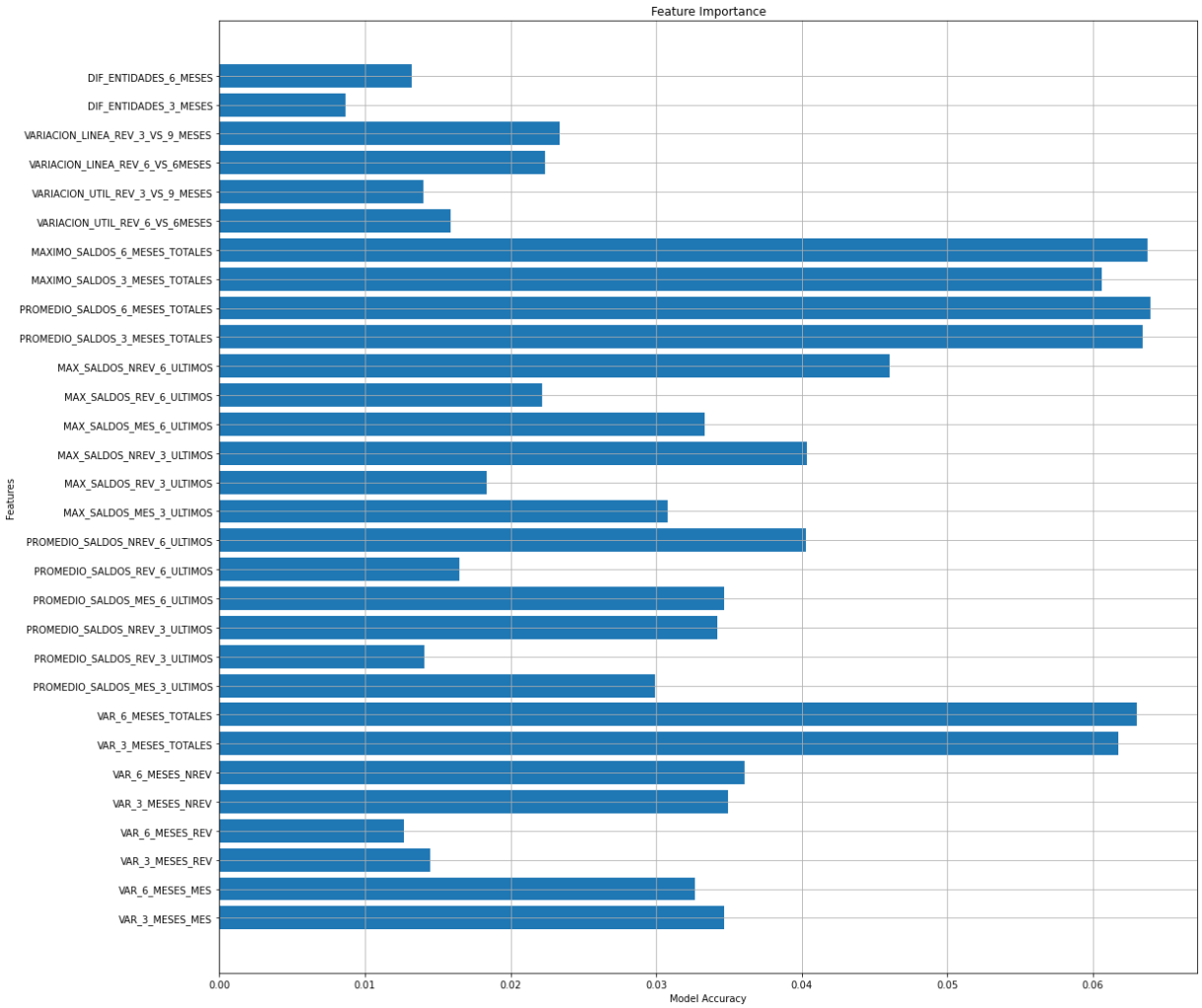
Importancia de variables en datos balanceados con Undersampling



En el caso del conjunto de datos balanceado mediante Oversampling, al realizar el modelado de Random Forest, se revisó la importancia de variables (figura 19) y se obtuvo que las variables de mayor relevancia son las referentes al saldo máximo en el sistema financiero en el último trimestre y semestre (MAXIMO_SALDOS_3_MESES_TOTALES, MAXIMO_SALDOS_6_MESES_ TOTALES), aquellas correspondientes a la variación del total de deuda (VAR_3_MESES_TOTALES, VAR_6_MESES_TOTALES) y respecto al máximo saldo de tipo No Revolvente en el último semestre (MAX_SALDOS_NREV_6_ULTIMOS). Es importante recalcar que, pese a que se tratan de dos métodos con enfoques opuestos, los métodos de Undersampling y Oversampling convergen a las mismas variables más importantes.

Figura 19

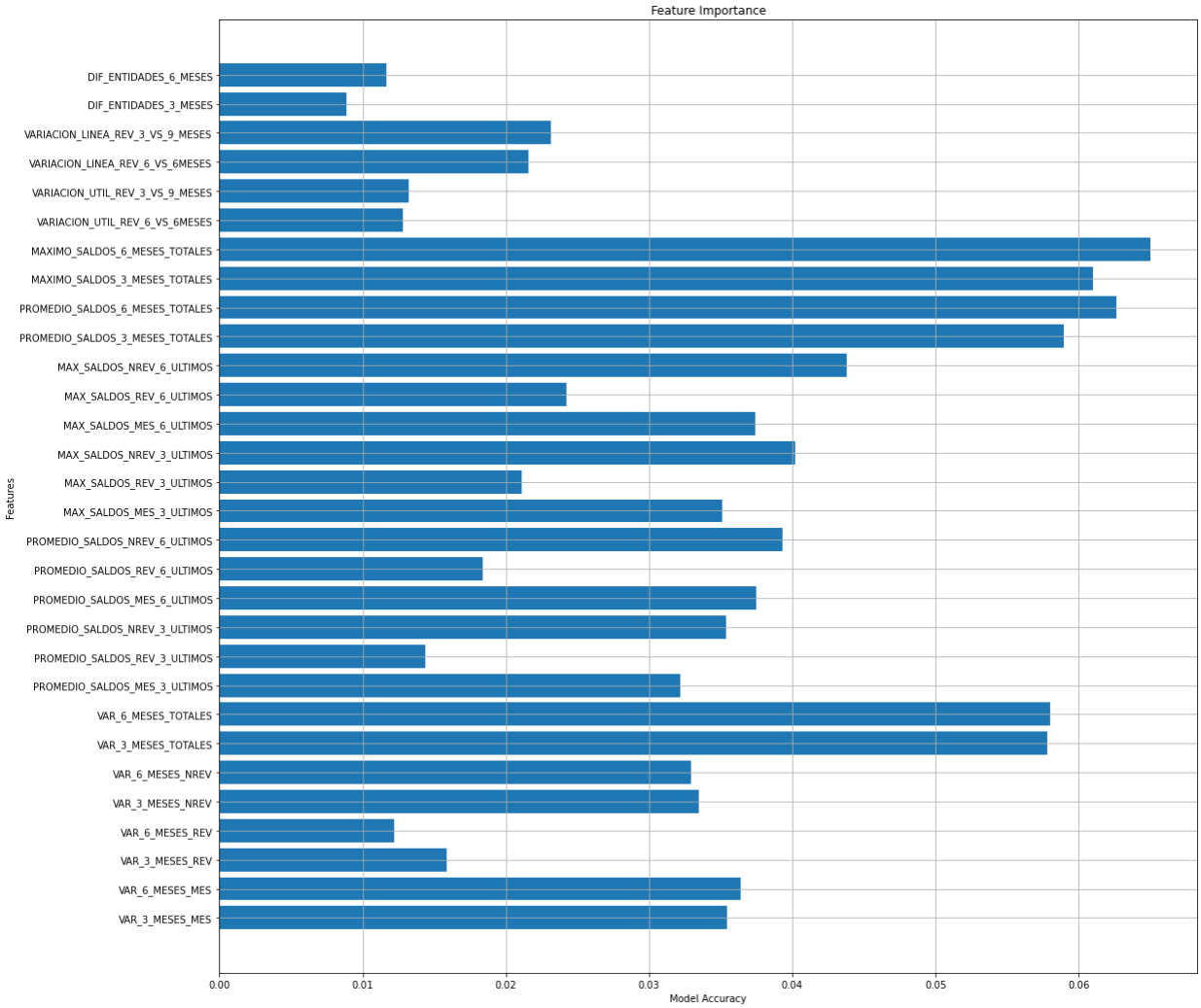
Importancia de variables en datos balanceados con Oversampling



De igual forma, en la figura 20 se observa que las variables más importantes de la base balanceada por el método Tomek-Link coinciden con lo identificado para las bases balanceadas mediante Undersampling y Oversampling.

Figura 20

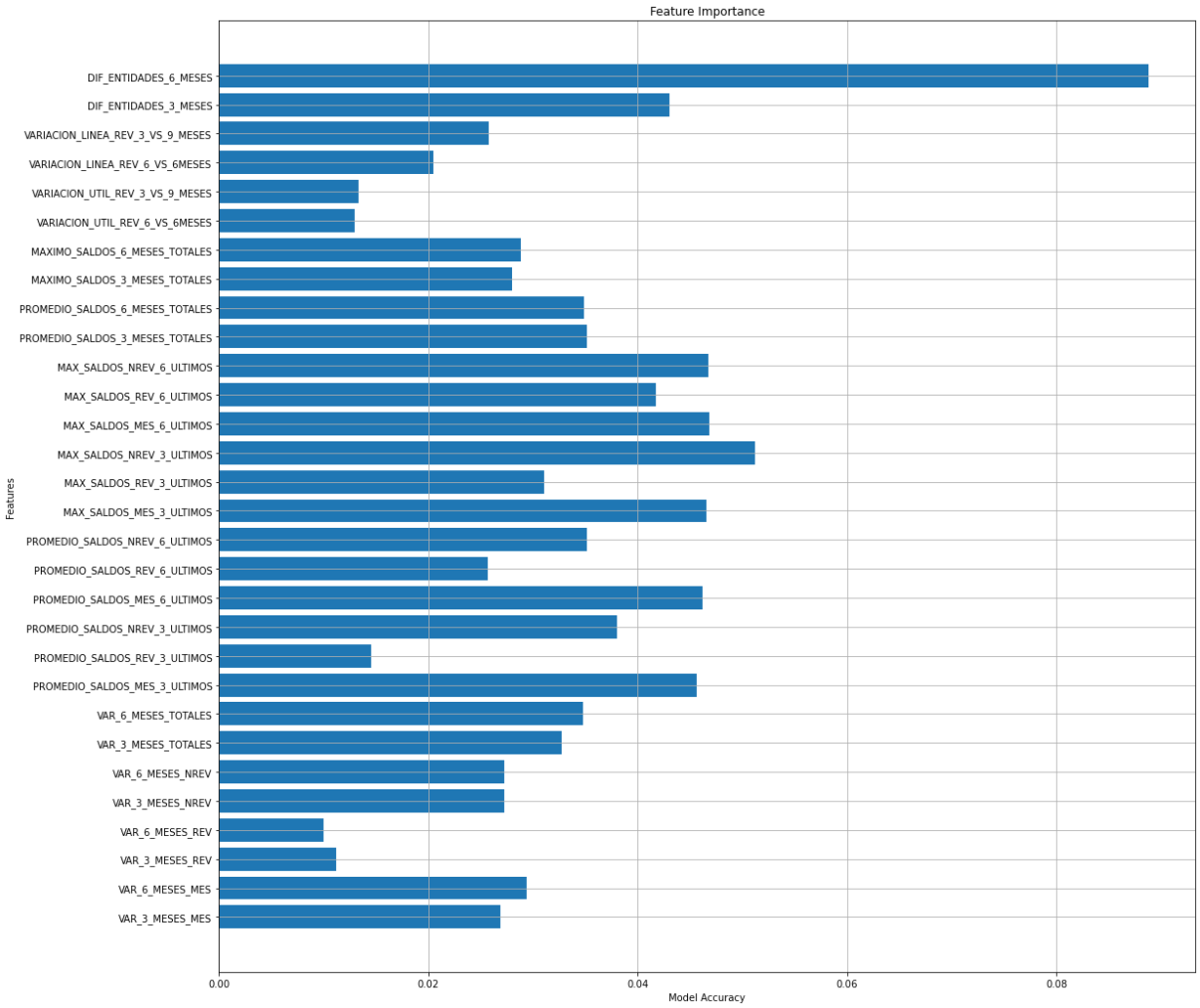
Importancia de variables en datos balanceados con Tomek Link



Caso contrario a lo obtenido previamente, con la base balanceada por el método de SMOTE se puede observar en la figura 21 que la variable más relevante es la diferencia de entidades en los últimos 6 meses previos al desembolso del crédito (DIF_ENTIDADES_6_MESES), seguida por la variable de saldo máximo del tipo de crédito No Revolvente en el último trimestre y semestre previo a la solicitud de crédito (MAX_SALDOS_NREV_3_ULTIMOS, MAX_SALDOS_NREV_6_ULTIMOS), y las variables del promedio de saldo en el tipo de crédito Microempresa en el último trimestre y semestre (PROMEDIO_SALDOS_MES_3_ULTIMOS, PROMEDIO_SALDOS_MES_6_ULTIMOS).

Figura 21

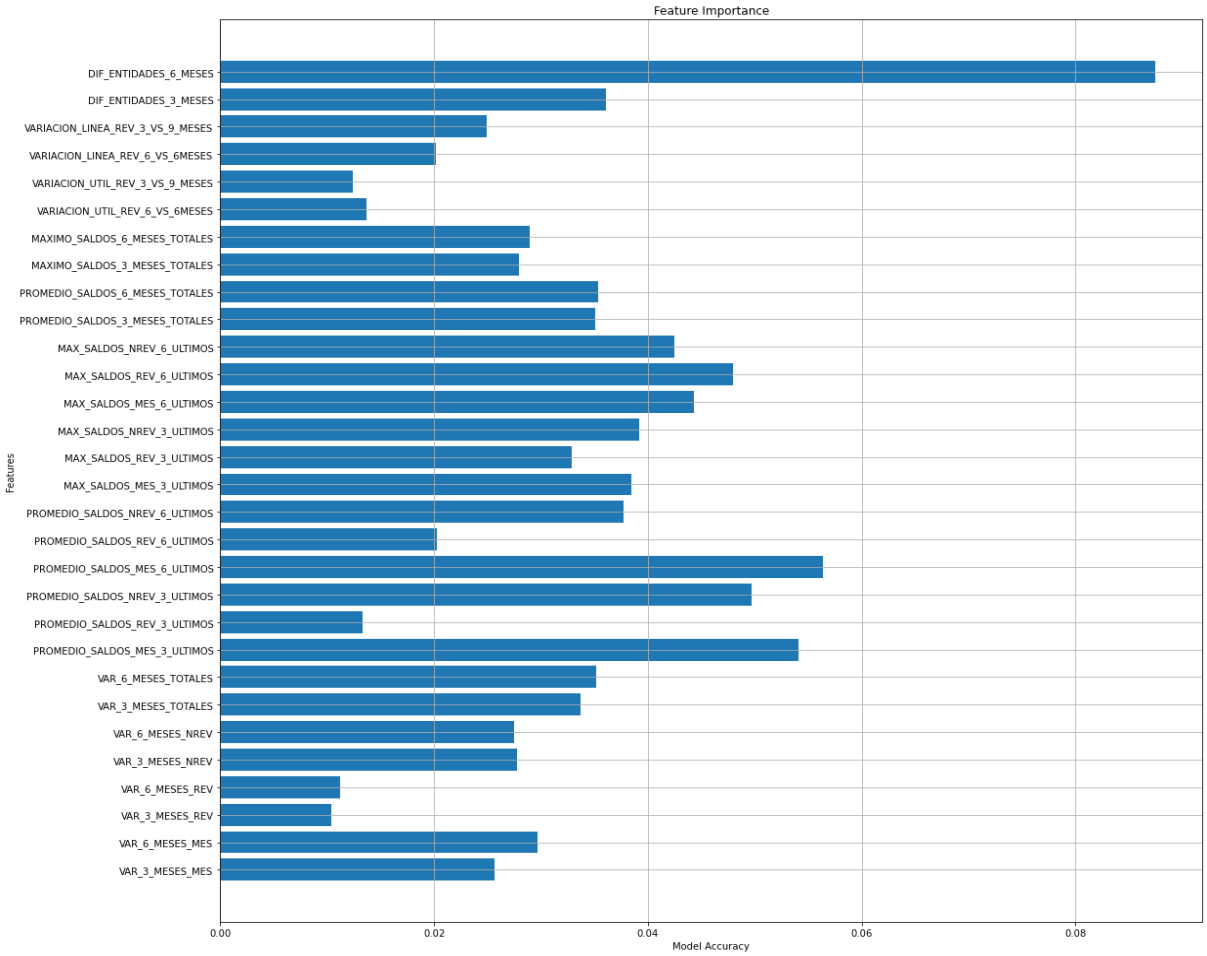
Importancia de variables en datos balanceados con SMOTE



Finalmente, se puede visualizar en la figura 22 que, similar a lo obtenido en SMOTE, el método de SMOTE-Tomek presenta una diferencia marcada en la importancia de las variables más representativas. Esto es, se identifica que la variable más relevante es la diferencia de entidades en el último semestre (DIF_ENTIDADES_6_MESES), seguido por el promedio de saldos del tipo de crédito Microempresa del último trimestre y semestre (PROMEDIO_SALDOS_MES_3_ULTIMOS, PROMEDIO_SALDOS_MES_6_ULTIMOS), luego por el promedio de saldos del tipo de crédito No revolvente del último trimestre (PROMEDIO_SALDOS_NREV_3_ULTIMOS) y finalmente el saldo máximo en el crédito Revolvente en los últimos 6 meses (MAX_SALDOS_REV_6_ULTIMOS).

Figura 22

Importancia de variables en datos balanceados con SMOTE- Tomek



Posterior a la identificación de las variables más importantes para cada una de las muestras tratadas por los métodos de Undersampling, Oversampling, Tomek Link, SMOTE, SMOTE-Tomek, se procedió a seleccionar las variables finalistas candidatas que serán probadas en el modelo de regresión logística. Es importante indicar que, si dos variables con el mismo concepto presentaban similar importancia relativa (por ejemplo, VAR_3_MESES_MES y VAR_6_MESES_MES) solo se eligió una de ellas para no incluir información relevante al modelo. A modo de resumen, en la tabla 7 se presenta las importancias relativas de las 5 técnicas probadas, donde las más relevantes se encuentran en formato negrita.

Tabla 7

Resumen de importancia de variables

	Undersampling	Oversampling	Tomek Link	Smote	Smote- Tomek limk
'VAR_3_MESES_MES'	0.0311	0.0347	0.0355	0.0268	0.0257
'VAR_6_MESES_MES'	0.0322	0.0327	0.0364	0.0294	0.0297
'VAR_3_MESES_REV'	0.0147	0.0144	0.0159	0.0112	0.0104
'VAR_6_MESES_REV'	0.0140	0.0126	0.0122	0.0100	0.0112
'VAR_3_MESES_NREV'	0.0360	0.0349	0.0335	0.0272	0.0278
'VAR_6_MESES_NREV'	**0.0404**	0.0360	0.0329	0.0273	0.0274
'VAR_3_MESES_TOTALES'	**0.0637**	**0.0617**	**0.0578**	0.0327	0.0337
'VAR_6_MESES_TOTALES'	**0.0655**	**0.0630**	**0.0580**	0.0347	0.0352
'PROMEDIO_SALDOS_MES_3_ULTIMOS'	0.0292	0.0299	0.0322	**0.0456**	**0.0541**
'PROMEDIO_SALDOS_REV_3_ULTIMOS'	0.0126	0.0141	0.0144	0.0145	0.0133
'PROMEDIO_SALDOS_NREV_3_ULTIMOS'	0.0354	0.0342	0.0354	0.0380	**0.0497**
'PROMEDIO_SALDOS_MES_6_ULTIMOS'	0.0346	0.0347	0.0375	**0.0462**	**0.0564**
'PROMEDIO_SALDOS_REV_6_ULTIMOS'	0.0162	0.0165	0.0184	0.0256	0.0203
'PROMEDIO_SALDOS_NREV_6_ULTIMOS'	0.0372	0.0403	0.0393	0.0351	0.0377
'MAX_SALDOS_MES_3_ULTIMOS'	0.0306	0.0308	0.0351	0.0466	0.0384
'MAX_SALDOS_REV_3_ULTIMOS'	0.0189	0.0184	0.0211	0.0310	0.0329
'MAX_SALDOS_NREV_3_ULTIMOS'	0.0373	0.0404	0.0402	**0.0512**	0.0392
'MAX_SALDOS_MES_6_ULTIMOS'	0.0321	0.0334	0.0374	0.0467	0.0443
'MAX_SALDOS_REV_6_ULTIMOS'	0.0210	0.0221	0.0242	0.0417	**0.0480**
'MAX_SALDOS_NREV_6_ULTIMOS'	**0.0433**	**0.0460**	**0.0438**	**0.0468**	0.0425
'PROMEDIO_SALDOS_3_MESES_TOTALES'	**0.0613**	**0.0634**	**0.0590**	0.0351	0.0351
'PROMEDIO_SALDOS_6_MESES_TOTALES'	**0.0649**	**0.0640**	**0.0627**	0.0349	0.0353
'MAXIMO_SALDOS_3_MESES_TOTALES'	**0.0600**	**0.0606**	**0.0610**	0.0280	0.0280
'MAXIMO_SALDOS_6_MESES_TOTALES'	**0.0634**	**0.0638**	**0.0650**	0.0289	0.0290
'VARIACION_UTIL_REV_6_VS_6MESES'	0.0166	0.0159	0.0128	0.0129	0.0137
'VARIACION_UTIL_REV_3_VS_9_MESES'	0.0135	0.0140	0.0132	0.0133	0.0124
'VARIACION_LINEA_REV_6_VS_6MESES'	0.0240	0.0224	0.0216	0.0208	0.0202
'VARIACION_LINEA_REV_3_VS_9_MESES'	0.0243	0.0234	0.0231	0.0258	0.0249
'DIF_ENTIDADES_3_MESES'	0.0093	0.0087	0.0088	0.0431	0.0361
'DIF_ENTIDADES_6_MESES'	0.0169	0.0132	0.0117	**0.0888**	**0.0875**

Nota. Se presenta los valores de importancia relativa obtenidos por Random Forest.

El proceso de elección de las variables se realizó de la mano de la Gerencia de Riesgos, con la finalidad de poder recibir las expectativas y opiniones asociadas. De dicha presentación se recibió la preocupación de verificar si el tipo de crédito No Revolvente permitía discriminar a los clientes morosos y no morosos en la campaña Pre Aprobado.

La elección de las variables se realizó tomando como principal referencia las importancias obtenidas mediante las técnicas de SMOTE-Tomek y SMOTE, pero sin desestimar los valores resultantes de las otras técnicas. A continuación, se detalla la elección de las variables candidatas:

- Se identificó que la variable "DIF_ENTIDADES_6_MESES" es la más relevante identificada mediante SMOTE-Tomek y SMOTE.
- Las siguientes variables más importantes son PROMEDIO_SALDOS_MES_6_ULTIMOS y PROMEDIO_SALDOS_MES_3_ULTIMOS, las cuales no pueden ser ingresadas en conjunto dado que ambas representan la misma definición (diferenciándose únicamente con la ventana de observación). Por ello, se preseleccionó la variable con vista del último semestre debido a que es ligeramente más relevante.
- La tercera variable que fue probada es MAX_SALDOS_REV_6_ULTIMOS, la cual fue de mucha relevancia en las técnicas SMOTE y SMOTE-Tomek.
- Se incluyó la variable PROMEDIO_SALDOS_NREV_3_ULTIMOS, por ser de gran importancia, y a la vez, responde a las expectativas del negocio.
- Por último, se seleccionó a VAR_6_MES_TOTALES como variable candidata, debido a que en las técnicas Undersampling, Oversampling y Tomek Link presentó un alto valor de importancia relativa y en las técnicas SMOTE y SMOTE-Tomek tenían un valor aceptable para no ser descartada.

Es importante mencionar que, las variables asociadas a deudas totales en el sistema financiero, tales como MAXIMO_SALDOS_6_MESES_TOTALES y PROMEDIO_SALDOS_3_MESES_TOTALES presentaron alta importancia en las técnicas Undersampling, Oversampling y Tomek Link. No obstante, se decidió no utilizarlas, debido a que dichas variables totalizan la deuda y no permite identificar cual es el tipo de deuda en particular que tiene mayor impacto en la detección de clientes en FPD.

Luego, se corroboró que las variables revisadas no presentan alta correlación, por lo que se verificó que no se incluyó información redundante. Como se puede observar en la tabla 8, ninguna cantidad, en valor absoluto, supera el 0.2.

Tabla 8

Matriz de correlaciones de variables preseleccionadas

CORRELACIÓN	VAR_6_MESES_TOTALES	PROMEDIO_SALDOS_MES_6_ULTIMOS	PROMEDIO_SALDOS_NREV_3_ULTIMOS	MAX_SALDOS_REV_6_ULTIMOS	DIF_ENTIDADES_6_MESES
VAR_6_MESES_TOTALES	1.000	0.123	0.041	0.034	0.127
PROMEDIO_SALDOS_MES_6_ULTIMOS	0.123	1.000	-0.150	-0.098	-0.011
PROMEDIO_SALDOS_NREV_3_ULTIMOS	0.041	-0.150	1.000	0.168	0.001
MAX_SALDOS_REV_6_ULTIMOS	0.034	-0.098	0.168	1.000	0.034
DIF_ENTIDADES_6_MESES	0.127	-0.011	0.001	0.034	1.000

Según lo mencionado anteriormente, en primera instancia se ajustó el modelo 1 de regresión logística con las 5 variables candidatas en la base balanceada mediante la técnica SMOTE-Tomek y se observó que la variable VAR_6_MESES_TOTALES presenta un p-value de 0.09 (tabla 9), por lo que generó la idea de ajustar un segundo modelo, pero sin dicha variable. Al realizar el ajuste del modelo 2 de regresión logística, se puede observar en la tabla 10 que todas las variables en cuestión contribuyen significativamente. Cabe resaltar que, los coeficientes en ambos modelos no varían de sentido ni de magnitud en ambos modelos. Para evaluar cuál de los modelos es más adecuado para predecir los clientes que caen en FPD, se analiza el performance considerando la base de entrenamiento, utilizando las métricas de accuracy, especificidad y sensibilidad.

Tabla 9

Coeficientes estimados del modelo 1 de regresión logística

Variable	Coeficiente	Error estándar	Z	P>|z|
Constante	0.0518	0.021	2.523	0.012
VAR_6_MESES_TOTALES	0.127	0.075	1.695	0.09
PROMEDIO_SALDOS_MES_6_ULTIMOS	-1.0971	0.075	-14.621	0
PROMEDIO_SALDOS_NREV_3_ULTIMOS	-1.4403	0.078	-18.581	0
MAX_SALDOS_REV_6_ULTIMOS	-0.7841	0.098	-8.03	0
DIF_ENTIDADES_6_MESES	1.7221	0.078	22.17	0

Nota. Modelo con todas las variables candidatas.

Tabla 10

Coeficientes estimados del modelo 2 de regresión logística

Variable	Coeficiente	Error estándar	Z	P>|z|
Constante	0.0755	0.015	5.034	0
PROMEDIO_SALDOS_MES_6_ULTIMOS	-1.0798	0.074	-14.532	0
PROMEDIO_SALDOS_NREV_3_ULTIMOS	-1.4324	0.077	-18.525	0
MAX_SALDOS_REV_6_ULTIMOS	-0.7782	0.098	-7.976	0
DIF_ENTIDADES_6_MESES	1.7384	0.077	22.552	0

Nota. Modelo sin la variable VAR_6_MESES_TOTALES.

Se analizó la matriz de confusión del modelo 1 de la tabla 11, de donde se obtuvo que la métrica de accuracy es 70%, la sensibilidad es 69% y la especificidad 70%. Por otro lado, al calcular la matriz de confusión del modelo logístico 2 (tabla 12) se obtuvo que la accuracy es 66%, la sensibilidad es 64% y la especificidad es 68%.

Tabla 11

Matriz de confusión del modelo logístico 1

Clase predicha	Clase observada	
	NO FPD	FPD
NO FPD	13577	5989
FPD	5789	13377

De lo anterior, se eligió el modelo que incluye la variable VAR_6_MESES_TOTALES (modelo logístico 1), debido a que aquel modelo presenta mejores valores en las métricas de interés, sobre todo el foco se centró en el valor de la sensibilidad (5% de mejora respecto al modelo logístico 2), dado que el objetivo principal fue poder detectar la clase que caen en incumplimiento de pago.

Tabla 12

Matriz de confusión del modelo logístico 2

Clase predicha	Clase observada	
	NO FPD	FPD
NO FPD	13110	7031
FPD	6256	12335

En base a las variables finalistas se tiene que el modelo logístico utiliza información referente a los tres tipos de créditos de interés de la entidad financiera (Microempresa, Revolvente y No Revolvente). Al respecto, se compara con las variables utilizadas en otras investigaciones dirigidas a entidades financieras peruanas como Chuquipul y David(2008) que, en el estudio para predecir la morosidad en tarjetas de crédito de un banco peruano, utilizaron variables sociodemográficas como el sexo, el estado civil, edad, continuidad laboral, antigüedad de su empleo, saldo medio en cuenta; por otro lado en la construcción de un modelo logístico para la predicción del incumplimiento de portafolio, Miranda (2021) hace uso de las variables de promedio de utilización de línea en tarjetas de crédito, máximo atraso en últimos 6 meses, deuda del último mes respecto al máximo de deuda en el último año, número de decrementos de deuda, tipo de ingresos, disposición de efectivo, máxima calificación, proporción de deuda revolvente.

La primera diferencia con las investigaciones mencionadas es la familia de variables utilizadas en el modelamiento, esto se debe a que el modelo construido para la entidad en estudio debía ser utilizado en la admisión de clientes nuevos y el score asignado sería asignado previamente a que la persona presenta una solicitud de crédito, con la finalidad de cumplir con el objetivo de la campaña de ofrecer créditos que se encuentren en condición de aprobados sin la necesidad de una evaluación, brindándoles una mejor tasa de interés en comparación al resto de planes financieros y beneficios como promociones comerciales; por tanto, no fue posible para el presente estudio utilizar variables sociodemográficas, que se recopilan por analistas de créditos, como antigüedad laboral o estado civil. Por otra parte, respecto a las variables asociadas a montos de deudas, se observó que en los otros estudios se ha utilizado netamente información referente a utilización de tarjeta de crédito (perteneciente al tipo de crédito Revolvente) y saldo de préstamos revolventes; sin embargo, en el presente estudio se consideró pertinente incluir los otros tipos de créditos, pese a que el tipo de crédito ofrecido por la entidad en el actual estudio es exclusivamente revolvente, ya que se planteó que el nivel de endeudamiento en los otros tipos de crédito impacta en la capacidad de pago.

Tras la elección de las variables, se evaluó la bondad de ajuste del modelo finalista mediante el test de significancia denominado Likelihood ratio (LLR), en el cual se prueba si el modelo completo de 5 variables es mejor que un modelo nulo. En dicha prueba se obtuvo que el valor

del log-Likelihood del modelo completo es -26276 y el valor del log-Likelihood del modelo nulo es -26847, y el test resultó significativo $P - value = 1.017 \times 10^{-244}$. Por tanto, se concluye que al menos uno de los coeficientes estimados contribuye significativamente en el modelo.

Como siguiente paso en la metodología seguida, se procedió a evaluar el modelo construido en una muestra fuera de la base de datos utilizada en la etapa de entrenamiento, siendo la base de prueba conformada por 11,419 registros. En la tabla 13 se muestra la matriz de confusión en tal muestra de prueba y se evalúo las métricas de ajuste y se obtuvo un valor de accuracy de 69%, sensibilidad de 67% y especificidad de 69%.

Tabla 13

Matriz de confusión en muestra de prueba

Clase predicha	Clase observada	
	NO FPD	FPD
NO FPD	7572	174
FPD	3311	362

Además, se presenta la curva ROC en la figura 23 y el valor bajo la curva, el cual es de 69%

Al evaluar el modelo en una muestra de prueba, se obtuvo métricas similares a las obtenidas en la muestra de entrenamiento, por lo que el modelo es confiable para realizar estimaciones para los clientes potenciales de campaña Pre Aprobado.

Posteriormente, mediante la formula presentada en Siddiqi, se realizó la transformación de las probabilidades obtenidas mediante el modelo de regresión logística hacia una escala del 0 al 1000, denominado score, donde aquellos potenciales clientes que cuenten con valores más altos representaban los perfiles con buen comportamiento de pago; caso contrario, los que presenten un bajo score eran los de mayor riesgo de incumplimiento de pago.

Figura 23

Curva ROC para muestra de prueba

Adicionalmente de las métricas ya presentadas, se comparó el modelo desarrollado versus el modelo anterior. Es importante aclarar que, previo a la construcción del modelo interno se utilizaba un modelo genérico, desarrollado externamente, el cual era utilizado para todos los nuevos clientes de la entidad; sin embargo, al ser la Campaña Pre Aprobado un caso en particular, este antiguo score no era determinante para ofrecerles beneficios financieros, pues solo era un valor que se asignaba por default al crearse una nueva cuenta de cliente.

Para tal comparación, se utilizaron las cosechas de noviembre y diciembre del 2016 y se revisó el indicador Kolmogorov–Smirnov (KS) utilizado en el seguimiento de modelos; para ello se examinaron las cosechas mencionadas en un año de desempeño, es decir, se observó si después de un año las cuentas habían caído en mora mayor a 30 días. En la tabla 14 se presenta los resultados para clientes admitidos en la Campaña Pre Aprobado usando el score antiguo, distribuido en deciles, en el cual se puede observar que el valor de KS es bajo, lo que indicaría que dicho score no permite discriminar adecuadamente entre clientes buenos y malos.

Tabla 14

Seguimiento modelo score antiguo

Decil	% Acumulados buenos	% Acumulados malos	Diferencia Porcentual (Malos - Buenos)
1	7.9%	12.9%	5.0%
2	15.7%	23.3%	7.6%
3	27.1%	37.0%	9.9%
4	42.3%	54.3%	12.0%
5	59.2%	71.8%	12.6%
6	75.1%	85.5%	10.4%
7	87.3%	93.7%	6.4%
8	95.2%	98.2%	3.0%
9	99.5%	99.9%	0.4%
10	100.0%	100.0%	0.0%
Kolgomorov Smirnov (K-S)			12.6%

Por el contrario, en la tabla 15 se muestra los resultados del back testing, simulando la aplicación del modelo interno propuesto, con el cual se puede observar una mejora notable en el indicador. Por lo cual, se concluyó que el modelo propuesto sí permitiría identificar a aquellas personas con alta posibilidad de incumplimiento y por ende se podría depurar la campaña, obteniéndose mejores resultados en indicadores de morosidad y rentabilidad.

Tabla 15

Resultados Back testing

Decil	% Acumulados buenos	% Acumulados malos	Diferencia Porcentual (Malos - Buenos)
1	5.6%	18.3%	12.7%
2	12.5%	32.1%	19.6%
3	20.6%	44.4%	23.8%
4	30.8%	57.0%	26.2%
5	38.7%	65.4%	26.8%
6	50.9%	76.0%	25.1%
7	61.5%	83.6%	22.1%
8	73.7%	91.0%	17.3%
9	86.7%	96.4%	9.6%
10	100.0%	100.0%	0.0%
Kolgomorov Smirnov (K-S)			26.8%

V. CONCLUSIONES

1. Es posible crear un ramillete de nuevas variables basándose en la información proporcionada en el RCC. Gran parte de las variables se construyó en base a la explotación del campo "Cuenta contable", ya que permitió diferenciar entre los montos de saldo (deuda directa), línea no utilizada (deuda indirecta), y como consecuencia el monto de línea crediticia. De este proceso de reingenería, se logró pasar de 4 campos input ("Entidades", "Tipo de crédito", "Saldo", "Cuenta contable") a poder probar la contribución 30 variables candidatas de índole crediticio. Asimismo, posterior a la construcción de variables, fue necesario realizar acotamiento y escalamiento de datos, a fin de evitar sesgos en el proceso de aprendizaje del modelo.

2. Fue importante incluir dentro del diseño del modelamiento la etapa de balanceo de la variable respuesta, FPD, debido a que inicialmente presentaba una proporción de 4.7% FPD y 95.3% NO FPD, y como consecuencia el modelo podía no ser preciso al predecir la clase de interés. Al probar diferentes técnicas de submuestreo y sobremuestreo, tales como Undersampling, Oversampling, Tomek Link, SMOTE, SMOTE- Tomek, se garantizó que posteriormente se seleccionen aquellas variables relevantes para realizar la predicción de clientes morosos.

3. De las 30 variables candidatas, se identificó que 5 variables eran las que contribuían para el aprendizaje del modelo. Dichas variables finalistas fueron seleccionadas tomando como referencia los resultados de modelos Random Forest ajustados en las bases de entrenamiento balanceadas con las técnicas antes mencionadas. De esta forma, se obtuvo 1 variable asociada a la variación de deuda, 1 variable asociada al número de entidades y 3 variables asociadas a los tipos de crédito de interés (Microempresa, Revolvente y No Revolvente).

4. Se estimaron los coeficientes de regresión logística utilizando las 5 variables finalistas y se observó que todas contribuían significativamente. Adicionalmente, se realizó el test Likelihood ratio, el cual indicó que el modelo construido es mejor a un modelo nulo.

5. Se corroboró en la muestra de prueba que el modelo construido realiza predicciones robustas, puesto que se utilizaron métricas como el accuracy, la sensibilidad y la especificidad para evaluar que no se exista un sobreajuste con la base de entrenamiento. Al respecto, al comparar los resultados entre la muestra de construcción y con la muestra de prueba se obtuvo que el accuracy y la especificidad varían en 1 punto porcentual, y la sensibilidad varía en 2 puntos porcentuales. Esta situación es favorable dado que no existe un gran cambio en los indicadores, principalmente en la sensibilidad ya que el objetivo principal de la entidad fue poder detectar correctamente a los morosos.

VI. RECOMENDACIONES

- Se sugiere incluir en el set de variables candidatas algunas con concepto de penalización, esto es, identificar cuentas contables específicas y crear nuevas variables asociadas a “Castigos” en el sistema financiero (situación que se presenta cuando el deudor superó en algún momento los 120 días de mora en los últimos 5 años), así como aquellas referidas a deudores con créditos pignoraticios (aquellos en los que se deja empeñado algún objeto a cambio de efectivo).

- Se sugiere que, de contar con diferentes fuentes de información además del RCC, poder incorporar variables referentes a deudas comerciales, pues a pesar de que estos datos no son reportados por entidades financieras supervisadas por la SBS, indican la capacidad y comportamiento de pago.

- Se recomienda considerar procedimientos en la construcción de modelos que sigan los lineamientos de la regulación financiera, puesto que hasta el momento del desarrollo del presente trabajo sugiere utilizar modelos tradicionales sobre los de caja negra, a fin de que pueda realizarse la trazabilidad y seguimiento de los modelos y de las variables utilizadas.

VII. REFERENCIAS

Ai-jun, L., & Peng, Z. (2020). Research on Unbalanced Data Processing Algorithm Base Tomeklinks-Smote. *Proceedings of the 2020 3rd International Conference on Artificial Intelligence and Pattern Recognition*, 13-17. https://doi.org/10.1145/3430199.3430222

Aridas C., Lemaitre, G., Victor, D. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research, 18* (17), 1-5. http://jmlr.org/papers/v18/16-365.html

Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. John Wiley & Sons.

Borkin, D., Némethová, A., Michal'conok, G. y Maiorov, K. (2019). Impact of data normalization on classification model accuracy. *Research Papers Faculty of Materials Science and Technology Slovak University of Technology, 27* (45), 79-84.

Cai, Z., Gu, Q., Zhu, L., & Huang, B. (2008). Data Mining on Imbalanced Data Sets. *2008 International Conference on Advanced Computer Theory and Engineering*, 1020-1024. https://doi.org/10.1109/ICACTE.2008.26

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

Chuquipul, E. & David, V. (2008). *Aplicación del modelo del credit scoring a las tarjetas de crédito en Lima Metropolitana: Caso de un banco peruano*. [Tesis de maestría. Universidad del pacífico]. Archivo digital. http://repositorio.up.edu.pe/handle/11354/3384

Dinov, I. (2018) *Data Science and Predictive Analytics*. Springer. https://doi.org/10.1007/978-3-319-72347-1

Elizondo, A., & Altman, E. I. (2004). *Medición integral del riesgo de crédito*. Editorial Limusa.

Galván, M. & Medina, F. (2007). *Imputación de datos: teoría y práctica*. Cepal.

González, R. S.-, Naranjo-Silva, E., Varela-Lorenzo, P., & Oñate-Andino, A. (2018). La gestión de riesgo: El ausente recurrente de la administración de empresas. *Revista Ciencia Unemi, 11*(26), 51-62.

Hastie, T., James, G., Tibshirani, R.& Witten, D. (2013). *An introduction to statistical learning: with applications in R.* Spinger.

Hoyos, J. (2019) *Metodología de clasificación de datos desbalanceados basado en métodos de submuestreo.* [Tesis de maestría. Universidad tecnológica de Pereira]. Archivo digital.

Khemais, Z., Nesrine, D., & Mohamed, M. (2016). Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression. *International Journal of Economics and Finance*, *8*(4), 39.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.

Medina, R. P., & Selva, M. L. M. (2013). Análisis del credit scoring. *Revista de Administração de Empresas*, *53*, 303-315.

Mester, L. J. (1997). What's the point of credit scoring. *Business review*, *3*, 3-16.

Miranda, A. (2021), *Predicción del riesgo de incumplimiento en el pago de los créditos del portafolio de una entidad financiera utilizando regresión logística.* [Tesis. Universidad Nacional Agraria La Molina]

Miyamoto, M. (2014). Credit risk assessment for a small bank by using a multinomial logistic regression model. *International Journal of Finance and Accounting*, *3*(5), 327-334.

Muñoz Rosas, J. F., & Álvarez Verdejo, E. (2009). Métodos de imputación para el tratamiento de datos faltantes: Aplicación mediante R/Splus. https://digibug.ugr.es/handle/10481/29627

Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. https://doi.org/10.48550/arXiv.1805.00801

Rayo Cantón, S., Lara Rubio, J., & Camino Blasco, D. (2010). Un Modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea II. *Journal of Economics, Finance and Administrative Science, 15*(28), 89-124.

Siddiqi, N. (2017). *Calificación crediticia inteligente: Creación e implementación de mejores tarjetas de puntuación de riesgo crediticio*. John Wiley & Sons.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524