Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Estadística Aplicada

MODELO LOGÍSTICO BINOMIAL APLICADO A LA CANCELACIÓN ANTICIPADA DE PRÉSTAMOS EN UNA INSTITUCIÓN FINANCIERA

Lcda. Mildred Jenifer Chiquitó Burrión
Asesorado por el Mtro. Walter Arnoldo Bardales Espinoza

Guatemala, febrero de 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

MODELO LOGÍSTICO BINOMIAL APLICADO A LA CANCELACIÓN ANTICIPADA DE PRÉSTAMOS EN UNA INSTITUCIÓN FINANCIERA.

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

LCDA. MILDRED JENIFER CHIQUITÓ BURRIÓN
ASESORADO POR EL MTRO. WALTER ARNOLDO BARDALES ESPINOZA

AL CONFERÍRSELE EL TÍTULO DE

MAESTRA EN ESTADÍSTICA APLICADA

GUATEMALA, FEBRERO DE 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANA	Inga. Aurelia Anabela Cordova Estrada
EXAMINADOR	Mtro. Edwin Adalberto Bracamonte
EXAMINADOR	Mtro. William Eduardo Fagiani Cruz
EXAMINADOR	Dra. Aura Marina Rodríguez Pérez
SECRETARIO	Mtro. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

MODELO LOGÍSTICO BINOMIAL APLICADO A LA CANCELACIÓN ANTICIPADA DE PRÉSTAMOS EN UNA INSTITUCIÓN FINANCIERA.

Tema que me fuera asignado por la Dirección de Escuela de Estudios de Postgrado con fecha 20 de julio de 2021.

Lcda. Mildred Jenifer Chiquitó Burrión

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.242.2023

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **MODELO LOGÍSTICO BINOMIAL APLICADO A LA CANCELACIÓN ANTICIPADA DE PRÉSTAMOS EN UNA INSTITUCIÓN FINANCIERA,** presentado por: **Lcda. Mildred Jenifer Chiquitó Burrión,** que pertenece al programa de Maestría en artes en Estadística aplicada después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, febrero de 2023

AACE/gaoc

Guatemala, febrero de 2023

LNG.EEP.OI.242.2023

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

"MODELO LOGÍSTICO BINOMIAL APLICADO A LA CANCELACIÓN ANTICIPADA DE PRÉSTAMOS EN UNA INSTITUCIÓN FINANCIERA"

presentado por **Lcda. Mildred Jenifer Chiquitó Burrión** correspondiente al programa de **Maestría en artes en Estadística aplicada** ; apruebo y autorizo el mismo.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA
DIRECCIÓN
ESCUELA DE POSTGRADO

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

Guatemala 31 de mayo 2022.

M.A. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

M.A. Ingeniero Álvarez Cotí:

Por este medio informo que he revisado y aprobado el Informe Final del trabajo de graduación titulado **"MODELO LOGÍSTICO BINOMIAL APLICADO A LA CANCELACIÓN ANTICIPADA DE PRÉSTAMOS EN UNA INSTITUCIÓN FINANCIERA"** del estudiante **Mildred Jenifer Chiquitó Burrión** quien se identifica con número de carné **999002811** del programa de Maestría en Estadística Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014*. Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente,

MSc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestría en Estadística Aplicada
Escuela de Estudios de Postgrado

Guatemala, 21 de octubre de 2021.

M.A. Ing. Edgar Darío Álvarez Cotí

Director

Escuela de Estudios de Postgrado

Presente

Estimado M.A. Ing. Álvarez Cotí

Por este medio informo a usted, que he revisado y aprobado el Trabajo de Graduación y el Artículo Científico: "**MODELO LOGÍSTICO BINOMIAL APLICADO A LA CANCELACIÓN ANTICIPADA DE PRÉSTAMOS EN UNA INSTITUCIÓN FINANCIERA**" de la estudiante **Mildred Jenifer Chiquitó Burrión** del programa de Maestría en **Estadística Aplicada**, identificada con número de carné: **201111847.**

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

MSc. Ing. Walter Arnoldo Bardales Espinoza

Colegiado No. 4279

Asesor de Tesis

ACTO QUE DEDICO A:

Dios	Jehová y Jesucristo gracias por estar siempre con nosotros y brindarnos su amor. Gracias por protegernos y bendecirnos. Gracias, por hacerme cada vez mejor ser humano.
Mi madre	Esperanza Burrión gracias por todo, por estar siempre y ser mi soporte, te quiero.
Mis hermanos	Bryan Chiquitó y Beatriz Chiquitó, gracias por todo y por estar siempre, los quiero, gracias ser mi motivación. Dios siempre nos protegerá. Gracias por motivarme, Rudy Chiquitó te quiero.
Mi padre	Jacobo, muchas gracias por la motivación y esfuerzo.
A mi	Por el esfuerzo y dedicación.

AGRADECIMIENTOS A:

Dios	Gracias Jehová y Jesucristo, los amo.
Universidad de San Carlos de Guatemala	Por formar parte del crecimiento académico de la población guatemalteca.
Institución financiera	Por brindarme la confianza de realizar el presente estudio.
Ing. Schwartz	Por la confianza depositada en mí y el apoyo brindado.
Los docentes	Que aportaron su conocimiento
Asesor y coasesores	Por compartir sus conocimientos y guiarme.
Familia y amigos en general	Gracias por el apoyo

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
χ^2	Chi cuadrado
σ	Desviación estándar
z	Desviaciones estándar en relación con la media.
D	Devianza
β_0	Intercepto de la ecuación lineal
p	Probabilidad
R^2	R cuadrado
β_k	Regresores de la ecuación lineal
α	Significancia estadística
y	Variable dependiente
x	Variable independiente
σ^2	Varianza

GLOSARIO

Calificaciones internas	Variables generadas a partir de comportamiento de uso o pago de productos financieros.
Crédito revolvente	Línea de crédito que se basa en los ingresos mensuales, es utilizado en el financiamiento a través de tarjetas de crédito.
Impago	En un crédito, pagos pendientes de amortizarse
Microfinanzas	Créditos efectuados a microempresas o pymes
Missing values	Datos faltantes en las observaciones
Mora	Estado de un crédito que ha permanecido n cantidad de días con algún pago atrasado.
Outlier	Dato atípico con respecto a los datos
R	Ambiente y lenguaje de programación con un enfoque estadístico.
***Scoring* crediticio**	Riesgo de impago por parte del titular de un crédito

RESUMEN

Esta investigación se realizó con el propósito de brindar a una institución financiera una herramienta que le servirá de apoyo para actuar ágilmente en la elaboración de planes de acción que promuevan la fidelización de los clientes, y con ello la reducción de los créditos cancelados anticipadamente.

El objetivo principal consistió en elaborar un modelo matemático construido con variables relevantes y con un error estimado, que permitiera generar probabilidades para categorizar los préstamos en propensos o no propensos para ser cancelados anticipadamente.

Para ello, se realizó un estudio con enfoque cuantitativo, en donde se siguió una secuencia de pasos lógicos para describir y correlacionar las variables (alcance), mediante un diseño no experimental de corte transversal, en cuyos resultados no se intervino, únicamente se analizó el comportamiento del evento de interés durante el periodo de octubre de 2019. El proceso de investigación inicio al examinar la bibliografía disponible sobre el tema, en donde se encontró que el mismo ha sido poco incursionado a nivel académico, pero que existen muchas fuentes sobre un tema semejante el *scoring* crediticio. Con esta premisa, se procedió a identificar las variables que proveen explicación al evento analizado a través de pruebas de independencia y con ellas se generó un modelo de regresión logístico binario, del cual se evaluó su capacidad predictiva mediante validación cruzada.

El principal resultado de estudio fue conocer que el modelo generado con las variables analizadas en el presente estudio provee una sensibilidad del 65 %, con un área bajo la curva del 61 % para predecir créditos propensos a ser cancelados anticipadamente. El aporte al conocimiento sobre el tema fue comprobar que, cuando se trabajan con grupos desbalanceados, es necesario aplicar un balanceo a los datos, también que la medida de exactitud y precisión obtenida a través de una matriz de confusión puede ser inexacta al valorar los resultados del grupo de menor volumen.

Se concluyó a nivel general que el modelo de regresión logística binomial generado cuenta con una capacidad predictiva media, con un error estimado del 35 %, el cual puede ser mejorado al agregar otras variables comportamentales de tipo cuantitativo, identificadas mediante pruebas de Wald.

Por lo que se recomienda, que los planes de acción que de él se deriven, vayan acompañados de un análisis costo-beneficio y de pruebas piloto, para evaluar la efectividad de estos.

PLANTEAMIENTO DEL PROBLEMA

Contexto general

En una institución financiera uno de los principales productos disponibles para los tarjetahabientes es el préstamo monetario, el cual consiste en otorgar un monto de efectivo a clientes precalificados, el cual deberá retornarse en cuotas en un plazo acordado por ambas partes. Estas cuotas están compuestas de capital e intereses, el tipo de amortización de dicho crédito es de cuota constante y pueden ser cancelados de forma parcial o total antes del plazo estipulado.

En la práctica, a nivel nacional por temas comerciales generalmente no se aplica penalización monetaria por cancelar anticipadamente un préstamo. En específico, en la unidad estudiada se permite efectuar cancelación parcial o total de estos, por eso existen un alto volumen de cancelaciones cada mes. Esto afecta principalmente los ingresos financieros para la institución, porque los préstamos otorgados en realidad conllevan un grado de incertidumbre sobre cuáles de estos podrán seguir activos cada mes.

Descripción del problema

En el control de la cancelación anticipada de préstamos únicamente se efectúan estimaciones sobre cuál será el comportamiento de estas durante el año, pero no se tiene identificado que préstamos son propensos a ser cancelados. Con ello, el porcentaje de cancelación anticipada muestra una tendencia al alza. Esto tiene como efecto una reducción en la cartera e ingresos

del producto, así como el riesgo de migración de los clientes a otra institución financiera.

En lo concerniente a la medición, se ignoran que variables tienen incidencia significativa con este fenómeno, tampoco se ha aplicado ningún modelo matemático que estime la probabilidad que tiene cada crédito de ser cancelado antes del plazo establecido, y no se conoce cuál sería el error de estimación que tendría categorizar dichos créditos como propensos o no propensos a cancelar anticipadamente. Por eso surge la necesidad de efectuar un análisis estadístico para que el área comercial pueda promover estrategias para fidelizar al cliente y mantenerlo activo.

Formulación del problema

Pregunta central

¿Cuál es el modelo matemático elaborado con variables relevantes que genera probabilidades para categorizar los préstamos propensos a cancelarse anticipadamente y cuál es su error de estimación?

Preguntas auxiliares

- ¿Qué variables explican significativamente la cancelación anticipada de los préstamos?

- ¿Cuál es modelo matemático que permite estimar la probabilidad para identificar los préstamos propensos a ser cancelados anticipadamente?

- ¿Cuál es el error de estimación de un modelo que categoriza a los préstamos como propensos o no propensos a ser cancelados anticipadamente?

Delimitación del problema

En el estudio se consideraron solamente variables relacionadas a las características del producto, demográficas y del comportamiento crediticio que inciden en la cancelación anticipada por parte de los tarjetahabientes titulares de un préstamo de la institución dentro del territorio de Guatemala.

Esta información se trabajó bajo estricta confidencialidad con fines académicos. Para ello, se utilizaron las variables de forma categórica y los datos generales a través de porcentajes, debido a que únicamente se necesitó conocer que variables explican el fenómeno y su grado de influencia. La unidad de tiempo del estudio transversal se delimitó a los datos históricos del mes de octubre del año 2019, en donde se analizarán los préstamos tanto activos como cancelados anticipadamente en dicho periodo. Este mes fue seleccionado por ser el de mayor incidencia del evento investigado.

OBJETIVOS

General

Elaborar un modelo matemático construido con variables relevantes y con un error estimado, mediante regresión logística binomial, pruebas de independencia y validación cruzada, que permita generar probabilidades para categorizar los préstamos en propensos o no propensos a ser cancelados anticipadamente.

Específicos

1. Contrastar las variables independientes contra la variable dependiente, a través pruebas de independencia, para definir cuáles son necesarias en el modelo por generar.

2. Construir un modelo matemático utilizando la regresión logística binomial para estimar la probabilidad que tiene un préstamo de ser cancelado anticipadamente.

3. Calcular el error de estimación del modelo matemático por medio de la técnica de validación cruzada.

RESUMEN DEL MARCO METODOLÓGICO

Los objetivos de la investigación se alcanzaron al aplicar la metodología descrita en la presente sección.

Características del estudio

El enfoque del estudio fue cuantitativo porque siguió una secuencia de pasos lógicos para solventar el problema. La medición numérica y la conversión de los datos a categóricos fue la base para describir, analizar y modelar estadísticamente las variables que intervienen en la cancelación anticipada.

Con un alcance descriptivo correlacional se buscó conocer la relación entre el evento y las covariables, para poder definir cuales explican la ocurrencia este de forma significativa, con el objetivo de generar un modelo que permitió clasificar los préstamos según su probabilidad de ocurrencia.

Se aplicó un diseño no experimental de corte transversal, donde no se interactuó o influyó sobre el estado de las variables analizadas durante el periodo en específico del estudio, octubre de 2019.

Unidades de análisis

La población objeto de estudio se conformó de los préstamos que componían la cartera de la institución financiera durante el mes de octubre de 2019, los cuales estaban divididos en subpoblaciones dadas por el estado del préstamo al cierre de dicho mes, lo que permite ser activos o cancelados.

Con los datos 21,420 se realizó una división, 70 % para el entrenamiento y 30 % para la prueba del modelo. Sin embargo, debido a que la proporción de cancelaciones anticipadas es reducida, la muestra se tuvo que corregir mediante el algoritmo ROSE, que creó una submuestra que balanceó los grupos según el estado del crédito (activo-cancelado) y permitió generar un modelo más sensible a la detección de cancelaciones anticipadas.

Variables

La variable dependiente de la investigación fue el estado del préstamo, mientras que se consideraron 12 covariables referentes a las características del producto, información demográfica y comportamiento crediticio, las cuales se describen a continuación:

Tabla I. **Variables del estudio**

Variable	**Definición teórica**	**Definición operativa**
Estado del préstamo	Préstamo activo o cancelado anticipadamente.	Variable dicotómica – escala nominal 0 = activo 1 = cancelado
Género	Género del titular del préstamo.	Variable dicotómica – escala nominal F = femenino M = masculino
Estado civil	Estado civil del titular del préstamo.	Variable dicotómica – escala nominal C = Casado S = Soltero

Continuación tabla I.

Variable	Definición teórica	Definición operativa
Rango de edad	Edad del titular del préstamo en intervalos de 10 años, punto de inicio 25 años.	Variable politómica – escala ordinal 1 = 25-34 2 = 35-44 3 = 45-54 4 = 55-65
Región	Región del país de Guatemala en donde reside el titular del préstamo. La categorización se define según dos criterios región central e interior del país.	Variable dicotómica – escala nominal C = central I = interior del país
Ciclo de vida del producto	Meses de antigüedad del préstamo dividido el plazo.	Variable politómica - escala ordinal 1 = conversión (<0.25) 2 = crecimiento (<0.50) 3 = retención (<0.75) 4 = reactivación (<1)
Plazo	Plazo del préstamo	Variable politómica – escala nominal Corto < 1 año Mediano < 3 años Largo > 3 años
Préstamos	Uniproducto o multiproducto.	Variable dicotómica – escala nominal U = Uniproducto M = Multiproducto

Continuación tabla I.

Variable	Definición teórica	Definición operativa
Impagos	Pagos no realizados en fecha correspondiente a cada mes.	Variable dicotómica – escala nominal 0 = sin impago 1 = impago
Categoría crediticia	Calidad del record crediticio del titular del préstamo.	Variable politómica – escala nominal 1 = excelente 2 = bueno 3 = regular 4 = malo
Categoría de pago	Calidad en puntualidad de pagos.	Variable politómica – escala nominal 0 = sin pagos 1 = excelente 2 = bueno 3 = regular 4 = malo
Cancelación anticipada previa	Préstamos cancelados de forma anticipada anteriormente.	Variable dicotómica – escala nominal 0 = sin cancelación anticipada previa 1 = con cancelación anticipada previa
Cancelación natural previa	Préstamos concluidos de forma natural, es decir cumplido el plazo.	Variable dicotómica – escala nominal 0 = sin cancelación natural previa 1 = con cancelación natural previa

Fuente: elaboración propia.

Fases del estudio

La investigación y la propuesta de solución se desarrollaron de la siguiente forma:

- Fase 1: revisión de documentos. se consultaron estudios académicos relacionados con la cancelación anticipada de préstamos, donde se encontró que este tema tiene carácter innovador porque ha sido poco analizado. Sin embargo, existen múltiples investigaciones similares aplicadas al ámbito bancario como las que abarcan el *scoring* crediticio, las cuales se consideraron como premisa. Además, se realizó el marco conceptual que brindó el respaldo teórico al análisis de las variables y la propuesta de solución.

- Fase 2: gestión o recolección de la información. La información sobre las variables analizadas fue proporcionada por la institución financiera, los datos fueron manejados de forma confidencial con fines académicos. La información proporcionada corresponde al mes de octubre de 2019, por ser un mes de interés debido al aumento en el volumen de las cancelaciones. En esta parte del proceso se realizó una limpieza de los datos, para brindar un tratamiento a los datos nulos (*missing value*).

- Fase 3: análisis de información**.** Como primera parte se procedió a analizar de forma descriptiva los datos del evento cancelación anticipada. Donde, se evaluó la significancia estadística de las covariables consideradas en esta investigación a través de pruebas de independencia. Además, se conoció que las categorías del estado de un préstamo son desbalanceadas, lo que aporto complejidad al estudio.

- Fase 4: modelación y evaluación del modelo. En esta fase como primer paso se realizó una división de forma aleatoria de la base de datos para formar dos subbases una de entrenamiento 70 % y otra de prueba 30 %. Al tener una base de datos desbalanceada, se utilizó el algoritmo ROSE para equilibrar los dos grupos analizados. La modelación de la información se efectuó a través del software libre R, en donde se ingresó la información de la base de entrenamiento para calcular el modelo de regresión logística, que utiliza el método de máxima verosimilitud para generar los coeficientes, con los cuales se identificaron que variables son significativas para el fenómeno y posteriormente se depuraron, dejando únicamente las variables significativas. Mientras que para evaluar la significancia estadística y bondad de ajuste del modelo logístico se aplicó la prueba de *Likelihood Ratio*, el R cuadrado de Nagelkerke, la matriz de confusión y el área bajo la curva.

- Fase 5: interpretación de información. En esta fase se generaron conclusiones y recomendaciones, donde se destaca el conocimiento aportado al lograr los objetivos planteados.

Técnicas de análisis de información

Las técnicas que se aplicaron para cumplir los objetivos son las siguientes:

- Tabla de frecuencias relativas y graficas de barras: estas tablas sirvieron para medir la frecuencia de los eventos analizados, y las gráficas para presentar resultados visualmente.

- Pruebas Chi cuadrado de Pearson: mediante pruebas de independencia chi cuadrado de Pearson, aplicadas a los datos desbalanceados, se

evaluó de forma preliminar las variables explicativas para posteriormente corroborarlo en la generación del modelo.

- Pruebas Wald: se aplicaron en la generación de los coeficientes del modelo, además con ellas se corroboraron las variables que explican el evento de la cancelación anticipada.

- Devianza residual y nula*:* prueba de hipótesis basada en la devianza residual y nula que permitió conocer la significancia del modelo.

- Pseudo R^2 de Nagelkerke: permitió brindar un grado de explicación del modelo a nivel general.

- Matriz de confusión o validación cruzada: se utilizó para estimar el error del modelo de regresión logística aplicándolo a la base de datos de prueba para estimar que tan bien predice. El error en este tipo de modelos no se puede generalizar con exactitud, pero se puede obtener una buena estimación.

- Área bajo la curva y Curva ROC: al contar con datos desbalanceados se planteó la necesidad de conocer el área bajo la curva de las predicciones generadas por el modelo de prueba, por los sesgos en algunas métricas de la validación cruzada que se pueden presentar en estos casos.

INTRODUCCIÓN

El presente estudio es una sistematización, porque se aplicaron los conocimientos académicos previos sobre *scoring* crediticio y se adaptaron al evento cancelación anticipada, para corroborar o refutar la similitud entre ambos modelos que se presentan en el área financiera.

El problema abordado consiste en que en la institución financiera objeto de estudio las cancelaciones anticipadas de préstamos han ido en aumento, lo cual tiene un efecto negativo sobre el producto, por el riesgo de migración de los clientes a otra institución financiera.

La importancia de generar un modelo matemático que permita categorizar a los clientes según su probabilidad de cancelación radica en que aporta el conocimiento que a nivel estratégico hará factible desplegar planes de acción con los expertos en el producto para promover la fidelización del cliente.

Por ello, en este estudio se generó un modelo logístico binomial que permite categorizar los créditos como propensos o no a ser cancelados anticipadamente, el cual fue resultado de la identificación de las variables significativas a través de pruebas de independencia Chi cuadrado (χ^2) y Wald. Del cual se examinó su exactitud, precisión y sensibilidad por medio de validación cruzada.

Dentro de los principales aportes de la investigación fue identificar que las variables que más peso tienen en la explicación del evento analizado son las que

intervienen en el comportamiento del récord crediticio del titular del préstamo, así como aspectos propios del producto.

El estudio se realizó en cinco fases: investigación bibliográfica, recolección de la información, análisis de la información, generación y evaluación del modelo de regresión logística binaria, para concluir con la interpretación de los resultados y con la redacción del informe final.

La investigación fue factible de realizar, aunque durante la generación del modelo se tuvo que contemplar el factor de los grupos desbalanceados para poder ampliar y abordar la metodología que mejor se adaptará al evento analizado. Esto resulto ser un aporte adicional, debido a que al contemplar las posibles soluciones para manejar datos desbalanceados se logró identificar que efectivamente a nivel de explicación no darle un tratamiento correcto resta explicación al modelo, en especial al evento minoritario que en este caso es la cancelación anticipada.

El presente informe, se estructura de la siguiente forma:

Capítulo I, cuenta con el marco referencial en donde se incluyen estudios previos relacionados con modelos matemáticos de clasificación, que fueron la guía para la investigación y su posterior discusión de resultados.

Capitulo II, en él se desarrolla el marco teórico que contiene el conjunto de conceptos que respaldan las pruebas estadísticas aplicadas a la cancelación anticipada.

Capitulo III, en esta parte se presentan los resultados alcanzados al aplicar las técnicas de análisis de información. En donde se identificó que las variables

relacionadas con las características del producto y comportamentales son las más significativas para el modelo y que la capacidad predictiva es media.

Capitulo IV, en dicha sección se discuten los resultados mediante un análisis crítico interno y externo.

Posterior a estos capítulos, se enumeran las conclusiones y recomendaciones, y se enlista la bibliografía respectiva.

1. MARCO REFERENCIAL

El marco referencial del estudio y el conjunto de conceptos teóricos que respaldan la siguiente investigación se enuncian en este capítulo.

Marco referencial

Los estudios efectuados en instituciones financieras generalmente están enfocados en el tema de precalificación de clientes aptos para préstamos, para reducir el riesgo de impago, mientras que la cancelación anticipada es un elemento poco analizado. Sin embargo, las premisas financieras y estadísticas pueden ser aplicables en ambos casos.

Según los estándares internacionales “el sistema financiero se rige a los principios de Basilea II, que promueven la aplicación de modelos de medición de riesgos para administrarlos” (Superintendencia de Bancos de Guatemala, 2018, párr. 16-18). Por ello, en las instituciones financieras se aplican modelos predictivos internos que califican a los clientes y los categorizan para no arriesgar el capital. Con respecto a estos modelos, a nivel académico existen múltiples investigaciones sobre el tema, de los cuales se resumen algunas en este apartado que se consideraron como marco referencial.

Arango y Restrepo (2017) efectuaron un estudio sobre la aplicación de modelos matemáticos para establecer el *scoring* de un cliente, que básicamente es la medición del riesgo de impago. Dicha investigación principió con la identificación de las variables independientes que tuvieran significancia estadística sobre la

variable dependiente, luego plantearon la transformación y limpieza de la información para separarla en datos de entrenamiento y de prueba, y generar con los primeros un modelo logístico, del cual evaluaron sus resultados a través de validación cruzada. Esta secuencia proporcionó una guía de los pasos necesarios para efectuar un modelo estadístico aplicado a productos de crédito consumo.

Rayo, Lara y Camino (2010) en su investigación realizada en una institución de microfinanzas diferencian cuales son las principales metodologías estadísticas que se aplican en la obtención de un *scoring* de crédito:

- Análisis discriminante: es una técnica paramétrica usada para definir a los buenos y malos perfiles de crédito, en este método no se puede calcular las probabilidades de impago (Fisher, citado en Rayo *et al*., 2010). Además, que se necesitan cumplir los supuestos de normalidad.
- Modelos de probabilidad lineal: “la ecuación de regresión es una función lineal de las variables explicativas” (Rayo *et al*., 2010, párr. 17). La respuesta del modelo es una variable *dummy* que clasifica de forma binaria los resultados.
- Modelos logit: generan una probabilidad por cliente, para categorizarlos según su potencial de pago, se adaptan mejor a variables categóricas y estas no necesitan cumplir los supuestos de normalidad.
- Redes neuronales: funcionan procesando las variables independientes a través de nodos que simulan un sistema nervioso que proporciona como salida una probabilidad estimada de ocurrencia del evento. La desventaja que tienen radica en su complejidad.
- Árboles de decisión: al igual que los modelos de redes neuronales, la comprensión sobre cómo procesa la información es más difícil, sin embargo, tiene como ventaja que es una técnica no paramétrica.

Al considerar las características de las metodologías anteriormente expuestas se dedujo que el modelo que mejor se adapta a las necesidades del problema estudiado es el logit binomial, porque permitiría definir una probabilidad de cancelación anticipada a cada crédito, y con esta información se podrían realizar planes de acción.

Rayo *et al.* (2010) para seleccionar los datos partieron desde dos perspectivas: el inicio del crédito y su posterior evolución, mencionaron que las variables que más aportan a los modelos suelen ser cuantitativas. Sin embargo, por contar con mayor disponibilidad de variables categóricas eligieron las siguientes: antigüedad, con crédito preexistente o alguna denegación, sector de la actividad económica, destino de crédito, comportamiento de pago, información demográfica, garantías, tipo de préstamo solicitado, y como elemento innovador incluyeron variables macroeconómicas. Esta información la utilizaron para generar un modelo logit binomial, el cual consideraron el más adecuado a la situación, porque establece probabilidades y se adapta mejor a las variables categóricas. Mencionadas premisas se utilizaron para definir las posibles variables de la presente investigación adaptadas al evento analizado, además es importante destacar que el modelo logit binomial resulta ser uno de más utilizados en temas similares.

Existen dos tipos de modelos de *scoring*, el de precalificación y el de evaluación del comportamiento del crédito (Saunders y Allen, citados en Vargas y Mostajo, 2014). En su investigación aplicada al sector empresarial Vargas y Mostajo (2014) identificaron dos tipos de métodos para medir el riesgo, el primero es el estándar y el segundo es el de calificaciones internas, en donde se calculan varios indicadores propios de la naturaleza del producto. La aplicación estadística consistió en establecer variables, describir su comportamiento y crear un modelo de regresión logística para identificar la mora. Dentro de los resultados relevantes sobresale que

existen características específicas en el cliente que indican bajo riesgo: con créditos anteriores, personas casadas, de género femenino, con capacidad de pago alta y con garantía física. Esto demostró la factibilidad de generar un modelo a partir de calificaciones internas en una institución financiera.

Mientras que en un ámbito macro Fernández, Bejarano y Vicente (2019) en su investigación del sistema financiero español aplicaron análisis factorial para evaluar las variables y definir el nivel explicativo de estas. Posterior a ello, con la aplicación de *Data Mining* hicieron pruebas con tres tipos de modelo: los clásicos (logit y nominal), redes neuronales y árboles de decisión, luego confrontaron los verdaderos positivos contra los falsos positivos para comparar los modelos con una Curva ROC, y encontraron que no existía diferencia significativa entre ellos, aunque el modelo de árbol de decisión CHAID exhaustivo proporciona mayor precisión. Resulta relevante que, aunque los modelos generaron la misma capacidad predictiva, la selección del mejor fue mediante la evaluación del más preciso en los aciertos por ajustarse mejor a la realidad.

Cardona (2004), indica que “las premisas básicas para la construcción del modelo son la simplicidad, potencia y estabilidad” (p.146). Es decir, que el modelo pueda ser comprensible, que mantenga la calidad de predicción a través del tiempo y sea capaz de adaptarse a los cambios. En su estudio sobre riesgo crediticio trabajó con árboles de decisión, comprobó los resultados con las pruebas: “Kolmogorov-Smirnov para 2 muestras (K −S), la curva ROC (*Receive Operative Curve*), el coeficiente Gini y la prueba F” (Cardona, 2014, p. 147). En donde mencionó que las primeras son aplicadas usualmente a la regresión logística y son fáciles de comprender. Lo que muestra que la curva ROC al igual que en el estudio de Fernández *et al*., (2019), sirven de herramienta para medir diversos modelos.

Trejo, Ríos y Almagro (2016) aplicaron una metodología para proponer una optimización al modelo de riesgo crediticio revolventes de México, porque los indicadores teóricos de ingresos tienen cifras mayores a las reales, ellos tomaron como base un modelo logístico, que utiliza la máxima verosimilitud para generar los estimadores, las variables independientes basadas en el comportamiento crediticio que consideraron como base fueron: límite de crédito, historial de falta de pago, proporción de pago y pagos atrasados, donde el límite de crédito y el historial de falta de pagos fueron las variables que proporcionaron mayor explicación. Antes de generar el modelo evaluaron la multicolinealidad en las variables y en su posterior evaluación usaron una curva ROC y la prueba Kolmogorov Smirnov (K-S), las cuales demostraron que los resultados son mejores a los del modelo actual. Algo interesante de este estudio es que categorizaron las probabilidades en clases (A-1, A-2, B-1, B-2, B-3, C-1. C-2, D y E) según grados de riesgo. Esta técnica podría utilizarse para categorizar a los clientes propensos a cancelar en más de dos categorías.

Izquierdo (2000), con respecto a los modelos estadísticos de riesgo, hace una mención sobre los riesgos que paradójicamente tienen los modelos, explica varios casos reales en donde el ingreso incorrecto de variables puede acarrear errores de estimación que tengan un impacto financiero negativo. Es decir, que es primordial prestar mucha atención a la generación de un modelo matemático, mediante pruebas periódicas de su confiabilidad, porque puede afectar en vez de ayudar a solucionar los problemas.

Gámez (2016) realizó un estudio de aplicabilidad de la regresión logística, en donde explica que una vez obtenido el modelo se debe considerar el ajuste este, con tal motivo midió la efectividad al comparar la precisión de la clasificación

obtenida contra los datos que sirvieron para entrenar el modelo, así también recalcó que la calidad del modelo será dada por su ajuste global.

Vallejo, Guevara y Medina (2018) indican que la minería de datos es utilizada ampliamente en instituciones financieras para analizar los grandes volúmenes de datos, y es una herramienta que permite aplicar estrategias comerciales porque provee información de manera eficiente. Este aspecto es importante pues es una alternativa para automatizar procesos como la generación de modelos.

Beltrán (s.f.) en su estudio realizado en una compañía telefónica menciona que los algoritmos de aprendizaje automático se ven afectados por los datos desbalanceados, debido a que sesga los resultados y perjudica a la clase minoritaria. Las técnicas que analizó para corregir dicho desequilibrio fueron los métodos de muestreo: submuestreo, sobremuestreo, *SMOTE* y *ROSE*. Concluye que con los datos originales “es muy eficiente clasificando la clase negativa, que es mayoritaria, pero no así con respecto a la clase positiva minoritaria” (Beltrán, s.f., p.32). En la comparación de resultados, el método ROSE sobresalió con respecto a la sensibilidad, en tanto los datos desbalanceados presentaron exactitud y especificidad altas, pero sensibilidad menor. Es decir, estos últimos no brindan las condiciones para predecir el evento de interés.

En síntesis, después de haber analizado cada metodología aplicada a la medición del riesgo crediticio, se planteó que es viable utilizar estas bases para efectuar un análisis estadístico de los créditos propensos a la cancelación anticipada. Se consideró que según los estudios expuestos el modelo de regresión logístico binomial es el que mejor se adapta para resolver el problema planteado en la presente investigación, debido a que las variables analizadas en su mayoría son categóricas.

2. MARCO TEÓRICO

En seguida se presenta el conjunto de conceptos teóricos que respaldan la investigación.

2.1. Regresión logística

La regresión logística es ampliamente aplicada en las ciencias sociales para realizar estudios de respuesta binaria. Estos modelos utilizan como base la distribución de Bernoulli, en la cual 1 significa presencia de cierta característica y 0 la ausencia de esta. En dicha distribución la probabilidad de éxito (p) está en función de la variable regresora (x), con media igual a (p) y varianza de Bernoulli no constante igual a $p * (1 - p)$ (Walpole, Myers, Myers y Ye, 2012,). Lo cual se expresa en la siguiente ecuación:

$$p = f(x, \beta) \qquad \text{(Ec. 01)}$$

La regresión logística forma parte de los Modelos Lineales Generalizados (GLM), Celis y Labrada (2014) describen que dichos modelos son aplicados cuando los datos no cumplen los supuestos de normalidad, homocedasticidad de la varianza o independencia de los residuos. Este tipo de regresión tiene mucha semejanza con los modelos de regresión lineal múltiple, con los cuales comparten el fundamento de la ecuación lineal $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_{\mathrm{k}})$, la cual que es transformada mediante la función logit para cumplir el supuesto de normalidad.

> En la regresión logística el objetivo del análisis es poder efectuar predicciones del comportamiento, esto es, estimar las probabilidades de un suceso definido por la variable dependiente en función de un conjunto de variables predictoras o de pronóstico. La regresión logística tiene como ventaja sobre la regresión lineal el poder generar buenos predictores con base a variables categóricas. Los modelos de regresión logística se dividen en binarios (respuesta dicotómica) o multinomiales (respuesta politómica), mientras que las variables explicativas pueden ser tanto cuantitativas como cualitativas. (López y Fachelli, 2016, p.5)

Amat (2016) indica que según la cantidad de variables predictoras los modelos pueden ser simples (1) o múltiples (más de 1).

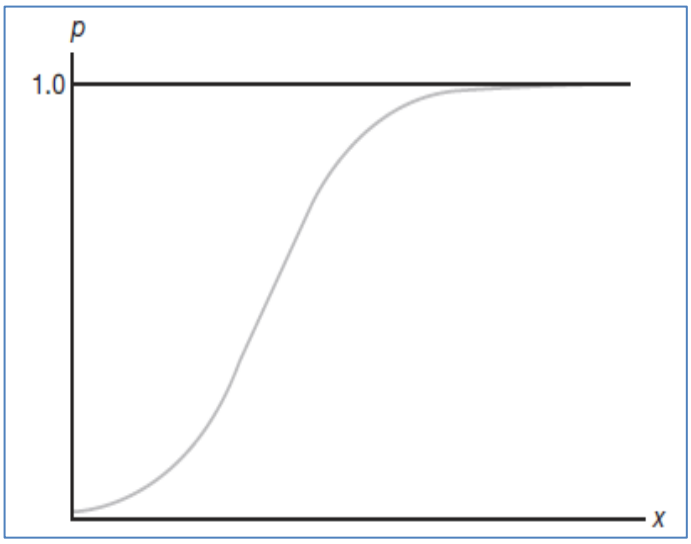
2.1.1. Función logística

Walpole *et al*., (2012) mencionan que la respuesta binaria de la regresión logística se mide mediante una probabilidad de éxito generada por la función logit:

$$P = \frac{1}{1+e^{-x\prime\beta\cdot}} = \frac{e^{x\prime\beta\cdot}}{1+e^{x\prime\beta\cdot}} \qquad \text{(Ec. 02)}$$

Donde, la e representa la función exponencial y "la porción $(x'\beta)$ se llama predictor lineal y, en el caso de un solo regresor (x), se puede escribir $(x'\beta) = \beta_0+\beta_1 x$ " (Walpole, et al., 2012, p. 498). Esta función adquiere forma de curva sigmoidal, en donde los posibles valores respuesta oscilan de 0 a 1 (véase figura 1).

Figura 1. **Función logística**



Fuente: Walpole, Myers, Myers y Ye, (2012). Probabilidad y estadística para ingeniería y ciencia.

Amat (2016) menciona que la función sigmoide básicamente efectúa una transformación de los coeficientes de la regresión lineal mediante exponenciación ($e^{-x\prime\beta\cdot}$) para determinar el efecto que tienen sobre la ocurrencia del evento. Donde, a mayor valor positivo en la función lineal ($x'\beta_0$) mayor probabilidad y a mayor valor negativo menor probabilidad, los resultados siempre oscilan entre el rango de 0 y 1.

Celis y Labrada (2014) explican que con la función logística se genera el modelo logístico, reemplazando $x'\beta\cdot$ ($\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_\mathrm{k}$) por los valores de los coeficientes, los cuales se calculan a partir del método de máxima verosimilitud (*maximum likelihood*), el cual busca maximizar la probabilidad de ocurrencia del evento.

2.1.2. Método de máxima verosimilitud

"Los coeficientes β son estimados mediante un procedimiento conocido como máxima verosimilitud (o maximum likelihood, en inglés). Cuando estos coeficientes son conocidos, podemos, mediante el modelo logístico, calcular la probabilidad de que un individuo presente un evento de interés" (Celis y Labrada, 2014, p. 219). En la regresión logística se aplica este método para calcular los coeficientes, de tal modo que se maximice la probabilidad de ocurrencia del evento.

2.1.3. Razón de probabilidad

"Una forma alternativa de representar la posibilidad de ocurrencia de un evento de interés es mediante el uso de *odds*, definidos como un cociente entre el número de eventos y el número de no eventos" (Cerda, Vera y Rada, 2013, p.1330). Los *odds* representan cuantos eventos verdaderos se tiene en función de los falsos, tal como se muestra en la ecuación 03:

$$odds = \frac{P}{1-P} \quad \text{(Ec. 03)}$$

A partir de los *odds* se puede hacer el cálculo inverso para encontrar la probabilidad, realizando la división de los *odds* entre la sumatoria de los *odds* más una unidad.

$$P = \frac{odds}{\sum odds+1} \qquad \text{(Ec. 04)}$$

Los *Odds Ratio (OR)* se define como un cociente entre dos *odds* (Cerda, *et al.,* 2013), que es equivalente a decir que el evento del numerador sucede *n* veces más que el evento del denominador. Brinda un panorama cuantificable de cuantas veces se espera que ocurra un evento considerando la presencia o ausencia de otra variable.

$$odds\ ratio\ (OR) = \frac{odds1}{odds2} \qquad \text{(Ec. 05)}$$

Celis y Labrada (2014) señalan la forma de generar el OR en la regresión logística a partir de la exponenciación del coeficiente de uno o varios regresores, considerando la siguiente ecuación base:

$$odds\ ratio\ (OR) = \frac{e^{\beta_0+\beta_1 x_1}}{e^{\beta_0}} = e^{\beta_1 x_1} \qquad \text{(Ec. 06)}$$

Amat (2016) indica que los *odds* oscilan en valores de $[0, \infty]$ y que para obtener un valor que represente la relación lineal entre el aumento en la variable explicativa y el aumento en la probabilidad de ocurrencia se aplica la transformación logit:

$$\beta_0 + \beta_1 x_1 = ln\left(\frac{P}{1-P}\right) = ln(odds) \qquad \text{(Ec. 07)}$$

En tanto, los intervalos de confianza permiten delimitar el efecto esperado de la variable regresora, la fórmula para su cálculo es la siguiente Celis y Labrada, (2014):

$$IC = e^{estimación\ puntual \pm z_{1-\alpha/2} * error\ estándar} \qquad \text{(Ec. 08)}$$

Cerda, et al. (2013). enlista las siguientes reglas para interpretar el OR:

- OR = 1 el regresor no tiene un efecto significativo en la ocurrencia del evento.
- OR > 1 = el regresor aumenta el efecto en la ocurrencia del evento.
- OR < 1 = el regresor disminuye el efecto en la ocurrencia del evento.

2.1.4. Supuestos de la regresión logística

Ferre (2015) resume los supuestos que deben cumplir un modelo de regresión logística son linealidad, independencia de errores y multicolinealidad, en seguida se detallan:

Linealidad: la ventaja de aplicar modelos de regresión logística radica en que la función logit permite el cumplimiento del supuesto de linealidad entre las variables regresoras y el logaritmo del efecto del OR (Ferre, 2015, Celis y Labrada, 2014).

Independencia de los errores: “los distintos casos de los datos no deben estar relacionados” (Ferre, 2015, párr. 30). Un elemento de la unidad de análisis no se debe repetir en el mismo estudio.

Multicolinealidad: “el concepto de colinealidad se refiere a las relaciones que existen entre las covariables, y no entre éstas y la variable dependiente” (Celis y Labrada, 2014, p.206). La colinealidad se da entre dos variables independientes, en tanto la multicolinealidad puede incluir la relación entre más de dos predictores. Es importante evaluar este aspecto para generar un modelo según el principio de parsimonia (el más simple).

Montgomery (citado en Hernández y Mazo, 2020), indica que para la multicolinealidad se evalúa el Factor de la Inflación de la Varianza (VIF) el cual se

define como el efecto combinado que tienen las dependencias entre los regresores sobre la varianza de ese término. Hay indicios de multicolinealidad si el VIF sobrepasa el valor 5 o 10.

2.1.5. Datos desbalanceados

Hernández, (2019) expresa que para la generación de modelos el manejar datos desbalanceados es un punto de interés debido a que pueden generar sesgo en los modelos. Algunos ejemplos, de este tipo de casos son casos de clientes con impago, en donde solamente una proporción de estos incurran en este.

Menciona que en el mundo del aprendizaje automático existen varias metodologías para solventar este inconveniente el sobremuestreo, submuestreo y la ponderación, los cuales se detallan a continuación:

- Sobremuestreo (*oversampling*): se obtiene una muestra mayor, equipara los datos del grupo mayoritario, al replicar los datos del minoritario, se puede tener una proporción 0.80-0.20.
- Submuestreo (*undersampling*): se genera una muestra menor, se reduce la de mayor volumen para lograr el equilibrio con la minoritaria.
- Ponderación (*weighting*): se procede a balancear el conjunto de datos asignando el peso en el software, esta opción no está contemplada en todos. (Hernández, 2019)

En tanto, Beltrán (s.f.) menciona en su estudio que existen dos algoritmos específicos para dicho fin:

- ROSE (*Randomly Over Sampling Examples*): “crea una muestra de datos sintéticos ampliando el espacio de características de la clase minoritaria y mayoritaria” (Lunardon, Menardi y Torelli, 2021, p. 9). Este algoritmo genera las muestras sintéticas a raíz de la estimación de densidad de Kernel condicional para las dos clases. Este algoritmo se puede usar tanto para datos categóricos como para cuantitativos.
- SMOTE (*Synthetic Minority Over-sampling Technique*): “genera nuevas instancias de la clase minoritaria interpolando los valores de las instancias minoritarias más cercanas a una dada” (Moreno *et al*., 2009, p.74). Los datos para implementar el algoritmo deber ser cuantitativos.

2.1.6. Tablas de contingencia y prueba chi cuadrado de Pearson

Al analizar más de una variable se utiliza la tabla de contingencia para agrupar en filas y columnas las observaciones según los niveles de la categorías (Mendenhall, Beaver y Beaver, 2010). “El objetivo es determinar si un método de clasificación es o no es contingente o dependiente del otro método de clasificación. Si no lo es, se dice que los dos métodos de clasificación son independientes” (Mendenhall *et al*., 2010, p.602).

Para probar la independencia entre la clasificación de las dos variables, una dependiente y otra independiente se utilizan pruebas chi o ji cuadrado, en donde se prueba la siguiente hipótesis:

$$Ho: los\ dos\ métodos\ de\ clasificación\ son\ independientes$$
$$Ha: los\ dos\ métodos\ de\ clasificación\ son\ dependientes$$

Mencionan que el estadístico de prueba se obtiene del promedio ponderado del cuadrado de las diferencias entre las cantidades observadas y esperadas:

$$\chi^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \text{(Ec. 09)}$$

Donde, las frecuencias observadas se calculan con:

$$frecuencias\ esperadas = \frac{total\ por\ columna * total\ por\ fila}{gran\ total} \qquad \text{(Ec. 10)}$$

Mientras los grados de libertad se obtienen a partir del número de filas (r) y columnas (c).

$$df = v = (r-1)(c-1) \qquad \text{(Ec. 11)}$$

Para la interpretación, se calcula el p-valor del ji cuadrado y el mismo se compara con la significancia, usualmente se utilizan valores 5 % para un 95 % de confianza, si el p-valor<0.05 entonces se rechaza la hipótesis nula y es posible indicar que no existe evidencia estadística para indicar que ambos métodos de clasificación son independientes.

Walpole, *et al*., (2012) resaltan un escenario particular: “en una tabla de contingencia de 2 × 2, donde sólo tenemos 1 grado de libertad, se aplica una corrección llamada corrección de Yates para continuidad” (p. 376).

$$\chi^2 = \Sigma \frac{(|\, O_{ij} - E_{ij} \,| - 0.5)^2}{E_{ij}} \qquad \text{(Ec. 12)}$$

En tanto, si los valores esperados son menores a 5 la prueba a aplicar es la prueba exacta de Fisher-Irwin.

2.1.7. Pruebas de significancia estadística

Walpole *et al*., (2012) indican que "las pruebas de los coeficientes individuales se calculan utilizando el estadístico χ^2 en lugar del estadístico t, puesto que no hay una varianza común σ^2" (p. 499). Para ello, se usa la siguiente fórmula:

$$\chi^2 = \frac{coeficiente}{error\ estándar}^2 = \frac{\beta_k}{SE(\beta_k)} \qquad \text{(Ec. 13)}$$

Esta prueba es conocida como de Wald (Wald chi-square) que "utiliza el estadístico Z y la distribución normal para probar la hipótesis nula de que un coeficiente en particular, β, es igual a 0" (Celis y Labrada, 2014, p. 221). Amat (2016) menciona que esta prueba es aplicada por el software estadístico R, mismo programa utilizado en el análisis de la información del estudio.

2.1.8. Bondad de ajuste de un modelo

Para evaluar la bondad de ajuste del modelo se aplican diversas pruebas estadísticas:

- Devianza: "De la función de verosimilitud se calculan dos momentos: uno inicial, V_I, y otro final, V_F. Ambos valores son menores de 1, y $V_I < V_F$. Con ambos números se puede calcular la devianza, o *D*" (Celis y Labrada, 2014, p.221). La fórmula es:

$$D = -2ln\left(\frac{V_I}{V_F}\right) \qquad \text{(Ec. 14)}$$

López y Fachelli (2016) interpretan la devianza como la "evaluación del ajuste del modelo a través del cambio o del incremento del estadístico $-2log(L)$ donde L es la razón de verosimilitud que varía entre 0 y 1" (p. 25).

Celis y Labrada (2014) mencionan que para comparar dos modelos, uno que no incluye predictores y uno que sí, se aplica la siguiente fórmula sobre las verosimilitudes finales de ambos:

$$D = [-2ln(V_{F*})] - [-2ln(V_F)] \quad \text{(Ec. 15)}$$

"Que se distribuye como χ^2 con $K - K^*$ grados de libertad, donde el asterisco identifica al modelo con el menor número de variables" (Celis y Labrada, 2014, p.221). Con una prueba chi cuadrado se puede obtener el p-valor, para contrastar la siguiente hipótesis:

$$H_o = los\ coeficientes\ de\ los\ predictores\ son\ igual\ a\ cero.$$
$$H_a = los\ coeficientes\ de\ los\ predictores\ son\ distintos\ a\ cero.$$

Cuando el p-valor>0.05 el modelo es significativo con un 95% de confianza.

- Pseudo R^2: "Se trata de medidas que evalúan el incremento de la verosimilitud del modelo: el cambio del estadístico -2log(L) o bien de L, la razón de verosimilitud que varía entre 0 y 1." (López y Fachelli, 2016, p. 25). Básicamente mide la variabilidad explicada por los coeficientes del modelo, con la siguiente fórmula (Celis y Labrada, 2014):

$$Pseudo\ R^2 = \frac{devianza\ nula - devianza\ residual}{devianza\ nula} * 100 \quad \text{(Ec. 16)}$$

López y Fachelli (2016) mencionan que existen también los Pseudo R^2 de Cox y Snell y el de Nagelkerke, siendo este último el que mayor valor provee, en cualquier caso es usual alcanzar valores entre el 0.20 y 0.3 en la regresión logística.

- Validación cruzada: Gil (2018) menciona que "puede aplicarse para estimar el test error asociado a un determinado método de aprendizaje estadístico (tanto regresión como clasificación) para evaluar el rendimiento del modelo (*model assessment*" (párr. 2). Las probabilidades generadas se evalúan según un punto de corte, que usualmente es del 0.5 (López y Fachelli, 2016). Las principales métricas que se pueden calcular para conocer la bondad de ajuste del modelo son:

$$exactitud = \frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{total\ de\ casos} \quad \text{(Ec. 17)}$$

$$presición = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos} \quad \text{(Ec. 18)}$$

$$sensibilidad = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos} \quad \text{(Ec. 19)}$$

$$tasa\ de\ falsos\ negativos = \frac{Falsos\ Negativos}{Falsos\ Negativos + Verdaderos\ Positivos} \quad \text{(Ec. 20)}$$

- Curva ROC: **"**representa la fracción de falsos positivos en abscisas frente a la fracción de verdaderos positivos en ordenadas" (Valle, s.f., p.17).

2.2. Cancelación anticipada de préstamos

Se define como la cancelación total de un préstamo antes del plazo acordado con una institución financiera (Barral, 2020).

2.2.1. Modelos de incumplimiento

Cardona (2004), menciona que existen dos metodologías para recabar las variables para medir un crédito:

- Basado en el inicio: se estudia el perfil al momento de iniciar el crédito para evaluar su incumplimiento futuro.,
- Basado en el comportamiento: analiza en un punto de corte la evolución del crédito considerando una cantidad determinada de meses históricos, y evalúa su comportamiento en un periodo de tiempo posterior.

2.2.2. Variables que intervienen en un préstamo

Pedrosa (s.f.) menciona que las variables que intervienen que intervienen en un préstamo son:

- Capital: monto de dinero proporcionado en calidad de préstamo.
- Tasa: el porcentaje de interés que el prestatario pagará sobre el capital adquirido en calidad de préstamo.
- Amortización: es una parte de la cuota mensual del préstamo, corresponde al aporte a capital.

- Intereses: es una parte de la cuota mensual del préstamo, pero en conceptos de intereses generados al aplicar la tasa de interés sobre el saldo de capital activo.
- Plazo: es el número de cuotas en las cuales se saldará el capital y los intereses correspondientes.

2.2.3. Segmentación de mercado

Según Kotler (citado en Ciribeli y Miquelito, 2014) a segmentación de un mercado se puede dividir según cuatro categorías:

- Geográfica: en Guatemala la segmentación se puede realizar por municipio, departamento o región.
- Demográfica: son datos propios del individuo, genero, edad, estado civil, ingresos, nivel educativo.
- Psicográficos: son aspectos subjetivos como gustos y preferencias.
- Comportamental: se adaptan más a aspectos de comportamiento, en el escenario especifico de la investigación pueden ser aspectos ya sea de consumo o de pago.

2.2.4. Marco legal guatemalteco aplicable a cancelaciones anticipadas

El Decreto Número 29-95 (1995. Art. 6) que reforma el artículo 87 de la Ley de Bancos indica que las instituciones financieras pueden cobrar un monto equivalente a dos meses de interés futuros por concepto de cancelación antes del plazo estipulado, cifra calculada con la tasa de interés acordada con el cliente al inicio del préstamo. Sin embargo, en la práctica a nivel nacional generalmente no

se aplica mencionada penalización, esta facilidad otorgada al cliente permite aumentar la competitividad en el mercado.

3. PRESENTACIÓN DE RESULTADOS

De acuerdo con los objetivos propuestos se presentan los siguientes resultados.

Objetivo 1: contrastar las variables independientes contra la variable dependiente, a través pruebas de independencia, para definir cuáles son necesarias en el modelo a generar.

3.1. Variables independientes del modelo

Con la finalidad de definir los factores que influyen de forma significativa en la cancelación anticipada, se dividieron las variables preliminares en tres categorías: características del producto, demográficas (incluye variable geográfica) y comportamentales. En seguida se muestran las pruebas de independencia realizadas sobre cada una de ellas.

3.1.1. Pruebas de independencia del ciclo de vida del producto

Para realizar las pruebas de independencia se categorizó, tanto la variable dependiente, como independiente. La variable de interés cancelación anticipada se codificó como un (1) y el estado activo de un crédito como un (0).

Las variables en esta categoría son: préstamos (uniproducto o multiproducto), plazo del crédito (corto, mediano y largo) y el ciclo de vida del

producto, el cual es el resultado de la división entre las cuotas transcurridas y el plazo del crédito, existen cuatro etapas:

- Conversión (1): menos del 25 % de la duración del crédito adquirido recientemente.
- Crecimiento (2): menos del 50 %, el crédito no ha cumplido el 50 % de su periodo.
- Retención (3): menos del 75 %, crédito que esta entre el 50 % y 75 % de su duración.
- Reactivación (4): crédito donde la mayoría de las cuotas ya han sido acreditadas.

Al aplicar la prueba chi cuadrado de Pearson se buscó conocer si estas variables son independientes del estado del crédito, en seguida se plantea la hipótesis general que incluye cada una de las variables:

Ho: *la cancelación anticipada es independiente de las características del producto*
Ha: *la cancelación anticipada no es independiente de las características del producto*

Con ello, a un nivel de significancia del 5 % se generó el estadístico de prueba chi cuadrado y su respectivo p-valor, en donde se obtuvieron los resultados siguientes:

Tabla II. **Pruebas de independencia caracteristicas del producto**

Dependiente	Independiente	x-squared	df	p-value	
Estado del préstamo	Ciclo de vida	38	3	3.E-08	<0.05
	Cat_plazo	68	2	2.E-15	<0.05
	Cat_préstamos	64	1	1.E-15	<0.05

Fuente: elaboración propia.

Con una significancia menor a 5 % se rechaza la Ho. Esto implica dependencia entre el evento y las variables ciclo del producto, plazo y tipo de préstamo. Lo cual muestra que dependiendo del ciclo del crédito, el cliente puede ser más o menos propenso a cancelar anticipadamente, resulta lógico porque a mayor etapa del ciclo los saldos se van reduciendo y es más accesible poder dar por finalizado el crédito. Sin embargo, los coeficientes del modelo corroborarán esta explicación previa.

3.1.2. Pruebas de independencia variables demográficas

Con respecto a las variables demográficas se consideraron el género, el estado civil, el rango de edad y la región de residencia. Estas son las principales covariables del perfil demográfico del titular del producto que se consideraron como factores para que un préstamo que sea más propenso a ser cancelado. La hipótesis que engloba las pruebas realizadas en cada una de las variables se escribe en seguida:

$$Ho: la\ cancelación\ anticipada\ es\ independiente\ de\ las\ variables\ demograficas.$$
$$Ha: la\ cancelación\ anticipada\ no\ es\ independiente\ de\ las\ variables\ demograficas.$$

El resultado obtenido es el siguiente:

Tabla III. **Prueba de independencia variables demográficas**

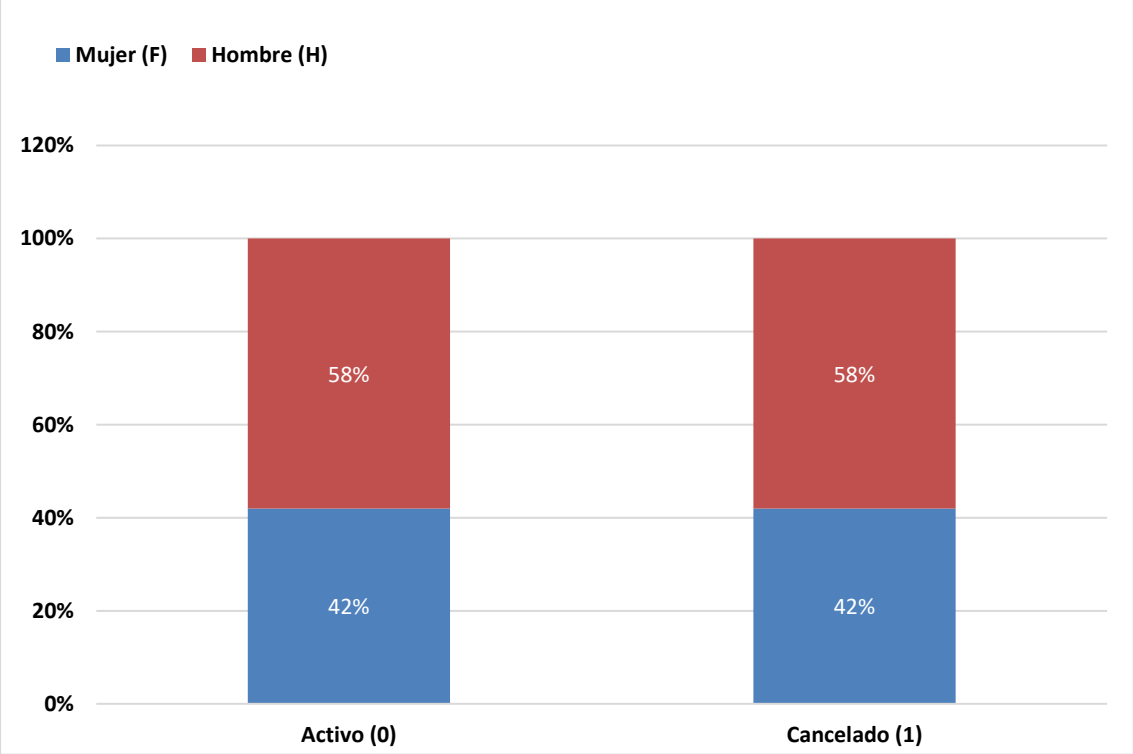
Dependiente	Independiente	x-squared	df	p-value	
Estado del préstamo	Género	0	1	6.E-01	>0.05
	Estado civil	0	1	7.E-01	>0.05
	Rango de edad	6	3	9.E-02	>0.05
	Región	1	1	3.E-01	>0.05

Fuente: elaboración propia.

Según las pruebas de independencia, con un nivel de significancia de 5 % existe evidencia estadística suficiente para no rechazar la Ho, lo que muestra que el evento cancelación anticipada se presenta independiente de las variables demográficas. Es la variable rango de edad la que más se acerca a la significancia con un p-valor del 0.09. Para visualizar los resultados anteriores a nivel gráfico, se examinaron las frecuencias relativas de las variables respectivas (a excepción de la variable región por motivos de confidencialidad).

Primero se examinó el covariable género, que se clasifica en dos categorías: mujer u hombre.

Figura 2. **Distribución de frecuencias variable género**

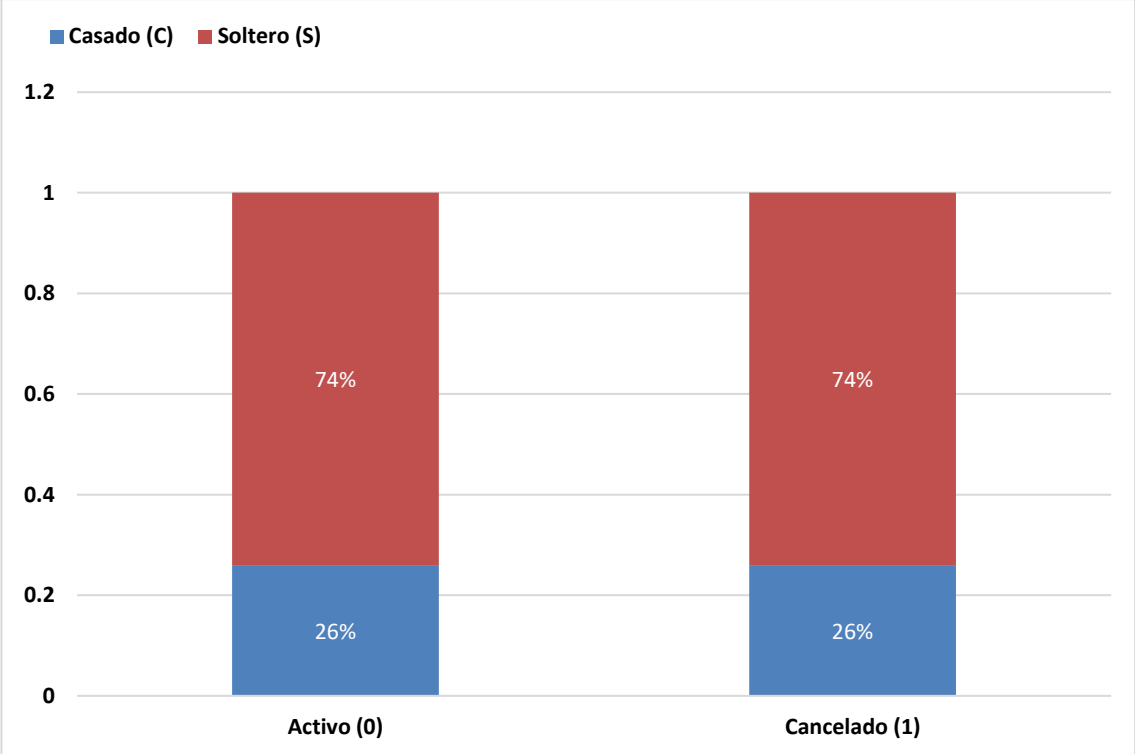


Fuente: elaboración propia.

La gráfica de barras que compara al segmento que permanece activo y al que cancela, enseña que efectivamente los grupos son bastante similares a nivel de proporción, es decir el evento cancelación anticipada ocurre independientemente del género.

En tanto, el estado civil clasificado en casado y soltero, de acuerdo con la posible carga familiar tiene el siguiente comportamiento:

Figura 3. **Distribución de frecuencias estado civil**



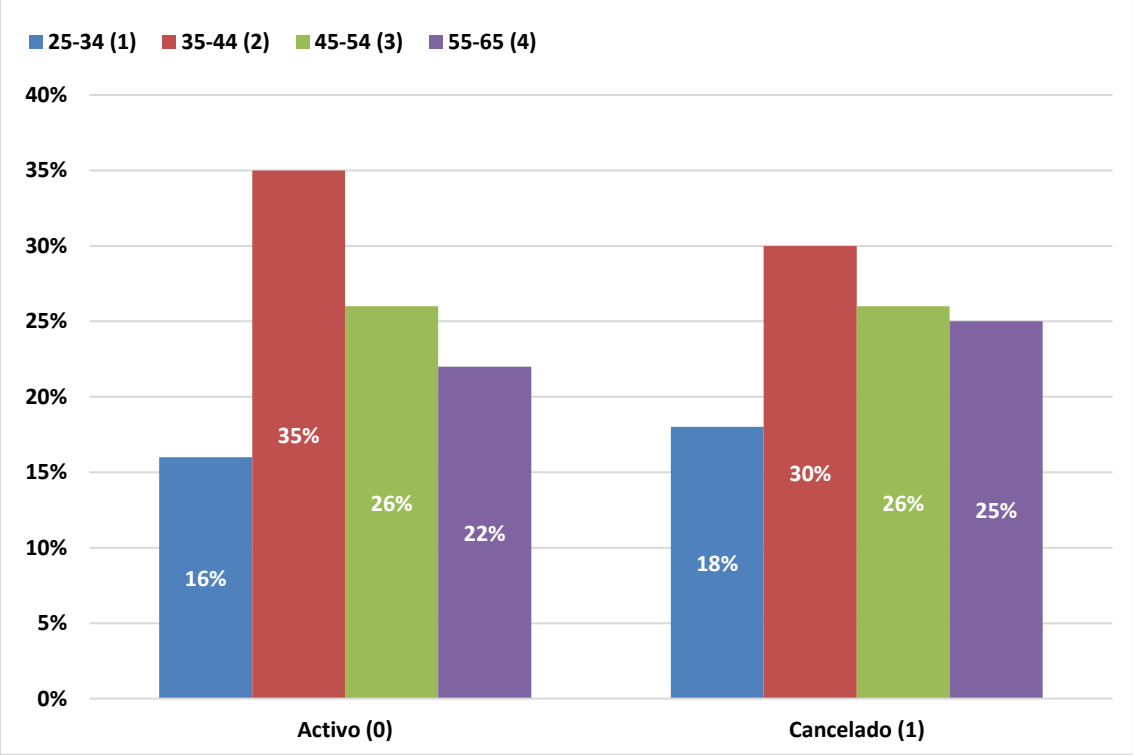
Fuente: elaboración propia.

De igual forma la distribución de las frecuencias relativas es exactamente la misma para ambos grupos observados, la cancelación anticipada es independiente

del covariable estado civil, es decir indistintamente si el estado civil es soltero o casado las condiciones de ambos estados de un crédito no se modifican.

Mientras que para el rango de edad que se subdivide en 4 categorías, de acuerdo con la edad presentada en el punto de corte del estudio, los resultados son los siguientes.

Figura 4. **Distribución de frecuencias del rango de edad**



Fuente: elaboración propia.

El rango de edad es similar entre las agrupaciones, a excepción del rango de 35 a 44 años, el cual presenta mayor proporción de clientes activos. Por esta razón dentro de los resultados de la prueba de independencia esta variable fue la más cercana a la significancia estadística.

3.1.3. Pruebas de independencia variables comportamiento del crédito

En esta sección se evaluaron las variables comportamentales del crédito, siendo las variables por examinar las siguientes:

- Impagos (evaluado durante los últimos 12 meses)
- Cancelación anticipada previa (durante los últimos 12 meses)
- Cancelación natural previa (durante los últimos 12 meses)
- Categoría de pago (de acuerdo con las historias crediticio)
- Categoría del titular (evaluado durante los últimos 12 meses)

Las hipótesis con cada una de las variables enlistadas trataron de validar lo siguiente:

Ho: *la cancelación anticipada es independiente de las variables comportamentales.*
Ha: *la cancelación anticipada no es independiente de las variables comportamentales.*

El resultado obtenido fue el siguiente:

Tabla IV. **Prueba de independencia variables comportamentales**

Dependiente	Independiente	x-squared	df	p-value	
Estado del préstamo	Impago	33	1	8.E-08	<0.05
	Cancelación anticipada previa	29	1	9.E-08	<0.05
	Cancelación natural previa	14	1	2.E-03	<0.05
	Categoría pago	68	4	8.E-14	<0.05
	Categoría titular	32	3	6.E-07	<0.05

Fuente: elaboración propia.

Con una significancia de 5%, es posible concluir que existe evidencia estadística suficiente para rechazar la hipótesis nula, la cancelación anticipada no es independiente de las variables comportamentales. Ahora bien, para determinar si la influencias de estas favorece o hace menos probable la cancelación se revisan los *odds ratio* (OR) del modelo a generar en el siguiente punto.

3.2. Regresión logística aplicada a la cancelación anticipada

Objetivo 2: construir un modelo matemático utilizando la regresión logística binomial para estimar la probabilidad que tiene un préstamo de ser cancelado anticipadamente.

3.2.1. Muestra del modelo

En primera instancia, se generó la proporción de cancelaciones anticipadas respecto al total de observaciones. Donde, se encontró que según el estado del préstamo clasificado como activo y cancelado, la distribución es del 90 %-10 % respectivamente, es decir los grupos a modelar son desbalanceados. Este es un factor que afecta la regresión logística, por tanto, se procedió a equilibrar la base de datos con la intensión de aportar las condiciones óptimas en el cálculo de las probabilidades y realizar una correcta interpretación de la bondad de ajuste de las predicciones.

Para generar el modelo primero se procedió a limpiar los datos, excluyendo los valores perdidos debido a que su proporción ascendía a menos del 1 % del total de registros, no impactando en la base de datos general. Para realizar el balanceo de las categorías se estudiaron cuatro métodos:

- Submuestreo: se consideró la cantidad de registros cancelados anticipadamente y se redujeron a la misma cantidad los datos de la clase predominante, por medio de selección aleatoria, se obtiene una muestra más reducida.
- Sobremuestreo: se partió del volumen de los créditos activos y se replica n cantidad de veces los datos de la clase minoritaria, hasta alcanzar el equilibrio.
- Combinación de submuestreo y sobremuestreo: se realizó por medio de algoritmos en r, que seleccionan de forma aleatoria la combinación óptima entre los dos tipos de muestreo expuestos anteriormente.
- ROSE: utiliza un algoritmo este nombre ROSE para crear muestras sintéticas que amplían las características de ambas clases, aplicando una estimación a partir de la densidad de Kernel condicionada para dos clases.

Al evaluar cada una de estas alternativas, en un modelo logístico se obtuvieron los siguientes resultados:

Tabla V. **Comparación de métodos para balancear datos**

	Método				
Variable	**Desbalanceado**	**submuestreo**	**sobremuestreo**	**combinado (sub-sobre)**	**ROSE**
Punto de corte	10 %	50 %	50 %	50 %	**50 %**
Precisión	10 %	6 %	7 %	0 %	**7 %**
Sensibilidad	5 %	64 %	70 %	63 %	**70 %**
AUC	53 %	64 %	67 %	64 %	**67 %**

Fuente: elaboración propia.

Con la premisa de que estas variables se utilizaron en cada una de las corridas del modelo, se obtuvieron los siguientes resultados:

- Datos desbalanceados: en este método la presión de los resultados fue la más alta entre todos los analizados, sin embargo, únicamente se tuvo una sensibilidad del 5 %, lo cual está directamente relacionado al desbalanceo porque debido a que la variable de interés pertenece al grupo minoritario básicamente el modelo no es capaz de predecir, lo cual se evidencia que se tuvo que utilizar un punto de corte del 10 % debido a que ningún valor predicho sobrepasaba del 50 % de probabilidad.

- Submuestreo, sobremuestreo y combinación: los resultados fueron bastante similares, en estos métodos ya se refleja una mejora en la sensibilidad que es al final de cuentas el indicador de mayor interés derivado a que brinda una medida de cuantos valores de la variable de interés se están prediciendo con exactitud.

- Algoritmo ROSE: los resultados fueron bastante similares a las técnicas de muestreo anteriores, sin embargo, en este método se tiene como beneficio que se crean muestras sintéticas derivadas de las muestras originales, sin recurrir a duplicar exactamente estas o eliminarlas como en los métodos anteriores. Dado que los resultados fueron bastante consistentes con el resto de los métodos aplicados para balancear los grupos, pero en este método su muestran las ventajas descritas anteriormente, se decidió aplicarlo en la generación del modelo de regresión logística final.

3.2.2. Generar coeficientes del modelo de regresión logística

Una vez generada la muestra mediante el algoritmo ROSE se generó el modelo a través del programa R, mediante un proceso de pasos hacia atrás y hacia adelante, es decir primero se incluyeron en el modelo todas las variables y

posteriormente se depuraron o agregaron para obtener el mejor modelo con únicamente las variables significativas. Los resultados se muestran en seguida:

Tabla VI. **Mejor modelo de regresión logística binaria**

Variable	Estimación	p-value	criterio	Odds ratio
Intercepto	-0.34	7.E-05	<0.05	0.71
cat_ciclo2	0.21	1.E-10	<0.05	1.23
cat_ciclo3	0.14	7.E-04	<0.05	1.15
cat_ciclo4	-0.21	4.E-05	<0.05	0.81
cat_plazolargo	-0.59	2.E-16	<0.05	0.55
cat_plazomediano	-0.31	9.E-09	<0.05	0.74
cat_generoM	-0.04	1.E-01	>0.05	0.96
cat_edad2	-0.03	4.E-01	>0.05	0.97
cat_edad3	0.03	4.E-01	>0.05	1.03
cat_edad4	-0.10	1.E-02	<0.05	0.90
cat_prestamoU	0.58	2.E-16	<0.05	1.79
cat_impago1	-1.08	2.E-16	<0.05	0.34
cat_categoria2	0.05	2.E-01	>0.05	1.05
cat_categoria3	0.10	7.E-02	>0.05	1.11
cat_categoria4	0.24	1.E-03	<0.05	1.28
cat_pago1	-0.04	1.E-05	<0.05	0.96
cat_pago2	0.01	9.E-01	>0.05	1.01
cat_pago3	0.10	2.E-01	>0.05	1.10
cat_pago4	0.40	4.E-13	<0.05	1.50
cat_cancelacion_nat1	-0.40	2.E-16	<0.05	0.67
cat_cancelacion_previa1	0.36	2.E-16	<0.05	1.44

Fuente: elaboración propia.

Resulta interesante que, al equilibrar los datos, la variable rango de edad en su categoría de 55 a 65 años es significativa con un 95 % de confianza, según la prueba de Wald que utiliza el software estadístico R para generar los coeficientes, a diferencia de la prueba de chi cuadrado de Pearson aplicada a los grupos desbalanceados.

A través de los coeficientes se identificó que las variables que aumentan la probabilidad de cancelación son:

- Ciclo del producto entre el 25 % al 75 %.
- Ser uniproducto
- Categoría general media a baja
- No pagador de contado
- Cancelación previa en los últimos 12 meses

Mientras que las que reducen la probabilidad son:

- Ciclo del producto del 75 % al 100 %
- Duración de mediano y largo plazo
- Edad de 55 a 65 años
- Con algún impago los últimos 12 meses
- Con cancelación natural previa

Posterior a la generación del modelo, se procede a validar la bondad de ajuste este en el siguiente punto.

3.3. Bondad de ajuste del modelo propuesto para monitorear la cancelación anticipada

Objetivo 3: calcular el error de estimación del modelo matemático por medio de la técnica de validación cruzada.

Con el modelo de regresión logístico binario generado, se procedió a validar el mismo, aplicando la función logit y la técnica de validación cruzada sobre los

datos de prueba. En esta comprobación no se balancearon los grupos, debido a que se requirió validar la exactitud de los resultados en un ambiente real, porque esa sería la forma de evaluar mes a mes cada crédito para inferir su probabilidad de cancelación.

Los resultados se evaluaron a través de una matriz de confusión en donde se obtuvieron los datos siguientes:

Tabla VII. **Validación cruzada del modelo de regresión logística binario**

	Predicción	
Real	**Cancelados**	**Activos**
Cancelados	65 %	35 %
Activos	50 %	50 %

Fuente: elaboración propia.

De la anterior tabla se desglosan las siguientes medias de bondad de ajuste:

$$exactitud = \frac{VP+VN}{total\ de\ casos} = 51\% \quad \text{(Ec. 09)}$$

$$presición = \frac{VP}{VP+FP} = 5\% \quad \text{(Ec. 10)}$$

$$sensibilidad = \frac{VP}{VP+FN} = 65.3\% \quad \text{(Ec. 11)}$$

$$tasa\ de\ falsos\ negativos = \frac{FN}{FN+VP} = 35\% \quad \text{(Ec. 12)}$$

La sensibilidad del modelo para el grupo de interés es del 65 %, con una tasa de falsos negativos del 35 % y un R^2 de Nagelkerke del 8 %**.** El modelo a nivel

general tiene una capacidad predictiva media, dentro de los motivos que originan la misma, radica que dentro de las variables explicativas analizadas en este modelo no existió una que marcará una diferencia explicita entre los dos grupos, si bien se encontraron variables significativas estas solamente explicaron el 8 % de la variación entre los dos grupos. A nivel general, si bien la sensibilidad del modelo del 65 % no se puede catalogar como buena, la misma supera la precisión brindada por un modelo desbalanceado, en el cual no se alcanzó a predecir ningún cliente propenso a cancelar por encima del 50 % de probabilidad.

En tanto, según la prueba likelihood ratio que consiste en evaluar la diferencia entre la devianza nula y residual a través de una prueba chi cuadrado, se obtuvo que el modelo es significativo con un 95 % de confianza. Mientras que al realizar un ejercicio de bootstrap con 100 iteraciones de la distribución del área bajo la curva de los datos, los resultados muestran las medidas de tendencia central homogéneas, con una media y mediana alrededor del 61.7 %. Con ello, el error considerando este criterio asciende a un 38.3 %.

3.4. Modelo de regresión logística

Objetivo general: elaborar un modelo matemático construido con variables relevantes y con un error estimado, mediante regresión logística binomial, pruebas de independencia y validación cruzada, que permita generar probabilidades para categorizar los préstamos en propensos o no propensos a ser cancelados anticipadamente.

Con un 95 % de confianza las variables que mayor explicación proveen del fenómeno, según pruebas de independencia de Wald sobre los datos balanceados, son las relacionadas al comportamiento de pago, record crediticio del préstamo y

características del producto. Las cuales, al introducirse en el modelo de regresión logística binaria, que calcula los coeficientes a través del método de máxima verosimilitud, predicen el 65 % de los nuevos eventos. El modelo tiene capacidad predictiva media, con un error estimado del 35 % (tasa de falsos negativos), este error puede ser reducido en futuras investigaciones al considerar otras variables comportamentales cuantitativas que puedan ser medidas de forma mensual.

La exactitud (*accuracy*) del modelo es del 51 %, al tratarse de datos desbalanceados se preveía este resultado debido a que se evalúa el beneficio de identificar los casos de interés (sensibilidad o recall) a costa de perder precisión y exactitud. Esto sucede particularmente porque el evento analizado es el que tiene la menor proporción o es un evento poco frecuente. La ecuación del modelo es la siguiente $P = (e^{x'\beta\cdot}/(1 + e^{x'\beta\cdot}))$ en donde $x'\beta\cdot$ es igual a:

$x'\beta\cdot$ = -0.34 + cat_ciclo2 * 0.21 + cat_ciclo3 * 0.14 + cat_ciclo4 * -0.21 + cat_plazolargo * -0.59 + cat_plazomediano * -0.31 + cat_generoM * -0.04 + cat_edad2 * -0.03 + cat_edad3 * 0.03 + cat_edad4 * -0.1 + cat_prestamoU * 0.58 + cat_impago1 * -1.08 + cat_categoria2 * 0.05 + cat_categoria3 * 0.1 + cat_categoria4 * 0.24 + cat_pago1 * -0.04 + cat_pago2 * 0.01 + cat_pago3 * 0.1 + cat_pago4 * 0.4 + cat_cancelacion_nat1 * -0.4 + cat_cancelacion_previa1 * 0.36

4. DISCUSIÓN DE RESULTADOS

Los resultados alcanzados en la investigación se discuten a continuación mediante análisis interno y externo.

4.1. Análisis interno

En el desarrollo de la investigación se presentaron dos dificultares, a raíz de las cuales surgió la necesidad de realizar una mayor investigación para solventar estas. Las cuales se describen en seguida:

La primera, el número de eventos de interés, en este caso la cancelación anticipada únicamente representa el 10 % del total de la población. Se comprobó mediante la generación del primer modelo de entrenamiento que este aspecto influía en la reducción de la capacidad predictiva del modelo. Debido a que al tener grupos desbalanceados el modelo de regresión logística tiende a favorecer en la predicción al evento de mayor proporción.

Para entender mejor el problema, se ejemplifica con la siguiente tabla de validación cruzada el efecto de los grupos desbalanceados.

Tabla VIII. **Efecto datos desbalanceados**

	Predicción de estado	
Real	Cancelados	Activos
Cancelados	44	377
Activos	10	11,345

Fuente: elaboración propia.

Las medidas de clasificación en este caso serían:

- Exactitud = 97 %
- Precisión = 81 %
- Sensibilidad = 10 %

En este caso al ser nuestra variable de interés la cancelación, es posible detectar que únicamente se están prediciendo 44 casos, pese a tener medias de exactitud y precisión alta. Lo cual ciertamente da una falsa medida de precisión, porque el modelo tiene menor error porque el grupo o la clase minoritaria no tienen un peso relevante en las proporciones.

Con ello, queda demostrado la necesidad de balancear los grupos al presentarse este escenario, el cual es muy usual encontrárselo a nivel práctico. Por ello, han surgido múltiples soluciones que pretenden realizar esta labor, y tal como se vio en la presentación de resultados una de los que proporciona mejores resultados y por ello se aplicó en este problema es el algoritmo ROSE. Con el cual, se alcanzó una sensibilidad del 65 %, es decir de cada 100 casos verdaderos el modelo predecirá 65. Es de observarse que, si bien la sensibilidad del modelo mejora con el balanceo, la exactitud se reduce, pero es por los argumentos mencionados.

En este punto surge el segundo problema, porque pese a balancear los grupos la capacidad predictiva del modelo es moderada con un 65 % y un área bajo la curva del 61 %. Esto se atribuye a que las variables independientes contempladas en la investigación no proveen un alto grado de explicación al evento, si bien mediante el contraste de pruebas de independencia se conoció que las variables comportamentales y las relacionadas a las características del producto generan un

moderado grado de explicación, esto abre la puerta para poder abarcar más variables este tipo que contengan información del comportamiento de consumo.

Ambos problemas identificados terminaron aportando un valioso conocimiento sobre el fenómeno de cancelación anticipada, porque en futuros estudios sobre modelos de clasificación se pueden considerar como premisa para realizar la planificación inicial.

En conjunto, el modelo de regresión logística binaria desarrollado en la presente investigación alcanzó un 65% de predicción sobre el grupo de interés, considerando para su elaboración variables significativas con un 95 % de confianza. Con una tasa de error del 35 % de error, esto es necesario recalcarlo y considerarlo al momento de realizar planes de acción para fidelizar al cliente.

4.2. Análisis externo

El marco referencial sobre *scoring* de crédito aplicado ante la escasez de estudios previos sobre cancelación anticipada de préstamos, resultó sumamente útil para vislumbrar la metodología y el tipo de variables a analizar, es interesante como ambos temas se asemejan. Rayo, Lara y Camino (2010) en su estudio aplicado al sector de las microfinanzas mencionan el aporte a su modelo de las variables que miden la evolución del crédito para determinar la probabilidad de impago, lo cual se ajusta con los resultados expuestos en esta investigación, pues las variables sobre el comportamiento de pago del cliente y las características del producto son las que proporcionaron mayor grado de explicación.

En concordancia con Vargas y Mostajo (2014) se evidenció que existe la posibilidad de crear un modelo a partir de calificaciones internas. Sin embargo, los

resultados de las variables significativas variaron en su interpretación, desde la perspectiva del *scoring* el cliente con mayor capacidad de pago fue el que menos riesgo tenía de caer en mora, y desde la perspectiva de la cancelación, precisamente este perfil es el menos propenso a cancelar antes del plazo un préstamo. Otra variable en común es la preexistencia de un crédito, estos clientes son menos probables en retrasarse en las cuotas y a cancelar su crédito.

En otro estudio muy enriquecedor Trejo, Ríos y Almagro (2016) en su tema aplicado a la medición de riesgo crediticio resaltan que las variables más significativas para dicho evento son las referentes al límite de crédito e historial de pagos. Siendo, esta última variable coincidentemente la que mayor explicación provee de forma positiva a la cancelación anticipada.

En un estudio muy alejado del ámbito bancario, con enfoque en la fuga de clientes en una empresa de telefonía móvil Beltrán (s.f.) con datos sobre frecuencia y consumos telefónicos, evaluó varios modelos entre ellos la regresión logística y otras técnicas como *random forest* y redes neuronales, sin embargo, antes de realizar cada uno de sus modelos procedió a balancear los datos. Es interesante mencionar en este estudio también se analizaron varios métodos de submuetreo, siendo el algoritmo ROSE el que proveyó una mejor métrica a nivel de sensibilidad, 70 % sin depuración de variables significativa. Sin embargo, la explicación del modelo es media, esto se atribuye a que todas las variables analizadas fueron categóricas, pero es factible en próximos estudios integrar variables cuantitativas porque proporcionen un mejor ajuste, tal como lo menciona. (Rayo *et al*., 2010)

Por último, se comparte la postura de Izquierdo (2000), con respecto a los riesgos de los modelos. Confiar excesivamente en una sola métrica o en el modelo en sí, puede llegar a causar errores en la implementación de las soluciones que

coadyuven a mejorar las condiciones del evento. Por ejemplo, los datos para el presente estudio por ser desbalanceados tuvieron que pasar por un proceso de balance, porque métricas como la exactitud y la precisión podrían brindar una falsa seguridad de la capacidad predictiva del modelo, pues al tener grupos con proporciones muy diferentes los coeficientes buscarán maximizar al grupo más grande provocando un sesgo en la predicción del menor.

Con la generación del modelo, se identificó el perfil básico de los préstamos propenso a ser cancelado anticipadamente, siendo este: uniproductos, que lleven del 25 % al 75 % de su ciclo, con cancelación previa en los últimos 12 meses, no pagador de contado y categoría crediticia media-baja. Con dicha premisa la institución financiera podrá realizar análisis costo-beneficio para evaluar los planes de acción que de este conocimiento se deriven.

CONCLUSIONES

1. Al contrastar la relación de las variables a través de pruebas de independencia Chi cuadrado sobre datos desbalanceados, se definió con un nivel de significancia de 5 % que las variables que aportan mayor explicación relacionadas a las características del producto y el comportamiento de pago. Las variables demográficas no influyen de forma significativa para predecir la cancelación anticipada de un préstamo.

2. El modelo matemático construido mediante regresión logística binomial utilizando variables comportamentales y del producto con una significancia del 5 %, permitió estimar la probabilidad que tiene cada préstamo de ser cancelado anticipadamente, utilizando un punto de corte del 50 %. Para ello, previo a la generación del modelo fue necesario realizar un balance de las clases mediante el algoritmo ROSE para aumentar la capacidad predictiva del modelo a un 65 % en el evento de interés.

3. El error de estimación del modelo calculado sobre los datos de prueba, por medio de la técnica de validación cruzada es del 35 %, así también mediante la técnica de bootstrap se infiere que el modelo genera un área bajo la curva del 62 %. Al aplicar la prueba ratio-likelihood con un 95 % de confianza se comprobó que el modelo es significativo con respecto a un modelo nulo, sin predictores.

4. El modelo de regresión logística binomial construido permite estimar probabilidades para categorizar los préstamos en propensos o no propensos a ser cancelados anticipadamente, con una sensibilidad en la predicción del 65 % y una tasa de error del 35 %. Las variables categóricas que aumentan la probabilidad de cancelación están relacionadas a la cantidad de productos activos, categoría de pago, cancelación anticipada previa y categoría crediticia. Mientras que las que reducen la probabilidad son la categoría de impago y el período del préstamo. En referencia a estas variables cualitativas, es viable para futuras investigaciones integrar al modelo variables de tipo cuantitativo este carácter para mejorar la predicción.

RECOMENDACIONES

1. En consideración a que las principales variables que resultaron significativas y explicativas en el modelo son las relacionadas con características comportamentales de pago y record crediticio, se sugiere integrar otras métricas de tipo cuantitativo que por motivos de confidencialidad no se consideraron en el presente estudio y evaluar si aportan un mayor grado de explicación.

2. Con el modelo de regresión logístico generado se sugiere a la entidad financiera actualizar mensualmente en los primeros días del mes la probabilidad de cancelación anticipada de cada préstamo, con base en el estado de las variables al cierre del mes anterior.

3. Resulta conveniente monitorear mensualmente el ajuste del modelo con las métricas de sensibilidad y exactitud por cada grupo, considerando que los datos son desbalanceados y el error máximo estimado asciende a un 38.3 %. Conjuntamente evaluar la evolución de los resultados a través de curvas ROC.

4. En general, se sugiere realizar planes de acción factibles para desplegar estrategias con base a la identificación de los préstamos posibles de ser cancelados mes a mes, pero que vayan acompañados con un análisis de costo-beneficio, que permita identificar qué acciones aportan más a la institución financiera. Además de realizar dichos planes mediante pruebas piloto.

REFERENCIAS

1. Arango, L., y Restrepo, D. (2017). *Diseño de un modelo de scoring para el otorgamiento de crédito de consumo en una compañía de financiamiento colombiana*. (Tesis de maestría). Universidad EAFIT. Colombia. ¿Recuperado de https://repository.eafit.edu.co/bitstream/handle/10784/12434/Laura_ArangoDuque_Daniel_RestrepoBaena_2017.pdf?sequence=2&isAllowed=y

2. Beltrán, V. (s.f.). *Predicción de fuga de clientes en empresas de telefonía móvil: el caso de estudio de Virgin Mobile.* Chile: Universidad de Desarrollo. Recuperado de: https://repositorio.udd.cl/handle/11447/2665

3. Cardona, P. (febrero 2004). Árboles de decisión en modelos de riesgo crediticio. *Revista Colombiana de estadística*, 27 (2), 139-151. Recuperado de http://www.kurims.kyoto-u.ac.jp/EMIS/journals/RCE/V27/V27_2_139 Cardona.pdf.

4. Celis, A., y Labrada, V. (2014). *Bioestadística* (3ª edición). México: Editorial El Manual Moderno, S.A. de C.V

5. Cerda, J., Vera, C. y Rada, G. (octubre 2013). Odds ratio: aspectos teóricos y prácticos. *Revista Médica de Chile*. 141 (10) 12-25-. Recuperado de

https://scielo. conicyt.cl/scielo.php?script=sci_arttext&pid=S0034-9887201300100 0014

6. Fernández, J., Bejarano, V. y Vicente J. (marzo 2019). Evaluación de riesgos con data *mining*: el sistema financiero español. *Revista mexicana de economía y finanzas*. 14 (23), 1-54. Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462019000300309

7. Gámez, J. (2016). *Modelización mediante regresión logística para estimación de proporciones en encuestas completas*. (Tesis de maestría). Universidad de Granada. España. ¡Recuperado de https://masteres.ugr.es/moea/pages/curso201516/tfm1516/gamezortiz_tfm/!

8. Hernández, F. y Mazo, M. (2020). *Análisis de regresión con R*. Colombia: Universidad Nacional de Colombia. Recuperado de https://fhernanb.github.io/libro_regresion/index.html#estructura-del-libro

9. Izquierdo, A. (julio 2000). Modelos estadísticos del riesgo y Riesgo de los Modelos Estadísticos. *Revista de la metodología de las ciencias sociales.* (3) 101-129. Recuperado de http://e-spacio.uned.es/fez/eserv/bibliuned:Empiria-2000-45232DBE-E087-44A6-75CE-3DD97DEE80B7/Documento.pdf

10. Mendenhall, W. Beaver, R. y Beaver, B. (2010). *Introducción a la probabilidad y estadística*. (13ª edición). México: Cengage Learning Editores, S.A. de C.V.

11. Rayo, S., Lara, J. y Camino, D. (enero 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), 1-133. Recuperado de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2077-18862010000100005

12. Superintendencia de Bancos de Guatemala. (2018). *Estándares internacionales.* Guatemala: Autor. Recuperado de https://www.sib.gob.gt/web/sib/faq/basilea

13. Trejo, J., Ríos, H. y Almagro, F. (agosto 2016). Actualización del modelo de riesgo crediticio, una necesidad para la banca revolvente en México. *Revista Finanzas y Política Económica.* 8 (1), 77-130. Recuperado de https://repository.ucatolica.edu.co/bitstream/10983/17162/1/1Actualizaci%c3%b3n%20del%20modelo%20de%20riesgo%20de%20credito%20un%20movimiento%20necesario%20para%20las%20lineas%20de%20credito%20resolventes%20em%20mexico.pdf

14. Valle, A. (s.f.). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones.* España: Universidad de Sevilla. Recuperado de: https://idus.us.es/bitstream/handle/11441/63201/Valle%20Benavides%20Ana%20Roc%C3%ADo%20del%20TFG.pdf?sequence=1

15. Vallejo, P., Guevara, E. y Medina, S. (agosto 2018). Minería de datos. *Revista Científica Mundo de la Investigación y el Conocimiento, 2 (Especial)*. 339-349. ¿Recuperado de https://dialnet.unirioja.es/servlet/articulo? codigo=6732870

16. Vargas, A. y Monstajo, S. (febrero 2014). Medición del riesgo crediticio mediante la aplicación de métodos basados en calificaciones internas. *Investigación & Desarrollo*, 2 (14), 1-138. Recuperado de http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2518-44312014000200002

17. Walpole, R., Myers, R., Myers, S. y Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencia*. (9ª edición). México: Pearson Educación de México, S.A. de C.V.

APÉNDICES

Apéndice 1. **Matriz de coherencia**

Descripción del problema	Preguntas de investigación	Objetivos	Metodología	Resultados	Conclusión	Recomendación
El porcentaje de cancelación anticipada de préstamos muestra una tendencia al alza. Esto tiene como efecto una reducción en la cartera e ingresos del producto, así como el riesgo de migración de los clientes a otra institución financiera. En lo concerniente a la medición, se ignoran que variables tienen incidencia significativa con este fenómeno, tampoco se ha aplicado ningún modelo matemático que estime la probabilidad que tiene cada crédito de ser cancelado antes del plazo establecido, y no se conoce cuál sería el error de estimación que tendría categorizar dichos créditos como propensos o no propensos a cancelar anticipadamente. Por eso surge la necesidad de	¿Cuál es el modelo matemático elaborado con variables relevantes que genera probabilidades para categorizar los préstamos propensos a cancelarse anticipadamente y cuál es su error de estimación?	Elaborar un modelo matemático construido con variables relevantes y con un error estimado, mediante regresión logística binomial, pruebas de independencia y validación cruzada, que permita generar probabilidades para categorizar los préstamos en propensos o no propensos a ser cancelados anticipadamente.	Pruebas de independencia Chi cuadrado, pruebas de Wald, validación cruzada, área bajo la curva, R2.	Fue posible generar el modelo de regresión logístico para conocer la probabilidad de cancelación anticipada de cada préstamo con base a variables significativas, el cual tiene un error estimado del 35% y capacidad predictiva media del 65%.	Se generó el modelo de regresión logística binomial que permite generar probabilidades para categorizar los préstamos en propensos o no propensos a ser cancelados anticipadamente, con una capacidad predictiva media (sensibilidad del 65% y área bajo la curva 61%), en donde las variables que tienen mayor relación son las que intervienen en el comportamiento de pago, record crediticio y características del producto.	En general, se sugiere realizar planes de acción factibles para desplegar estrategias con base a la identificación de los préstamos posibles de ser cancelados mes a mes, pero que vayan acompañados con un análisis de costo-beneficio, que permita identificar qué acciones aportan más a la institución financiera. Además de realizar dichos planes mediante pruebas piloto.
	¿Qué variables explican significativamente la cancelación anticipada de los préstamos?	Contrastar las variables independientes contra la variable dependiente, a través pruebas de independencia, para definir cuáles son necesarias en el modelo a generar.	Pruebas de independencia Chi cuadrado	Las variables relacionadas al producto y comportamentales tienen significancia estadística sobre el evento analizado, con un 95% de confianza. Mientras que las variables demográficas y geográficas no son significativas.	1. Al contrastar la relación de las variables a través de pruebas de independencia Chi cuadrado sobre datos desbalanceados, se definió con un 95% de confianza que las variables que aportan mayor explicación son las relacionadas al comportamiento de pago y características	1. En consideración a que las principales variables que resultaron significativas y explicativas en el modelo son las relacionadas a características comportamentales de pago y record crediticio, se sugiere integrar otras métricas de tipo cuantitativo que por motivos de confidencialidad no se consideraron en el presente estudio y evaluar si aportan un

Continuación apéndice 1.

efectuar un análisis estadístico para que el área comercial pueda promover estrategias para fidelizar al cliente y mantenerlo activo.					del producto. Las variables demográficas no influyen de forma significativa para predecir la cancelación anticipada de un préstamo.	mayor grado de explicación.
	¿Cuál es modelo matemático que permite estimar la probabilidad para identificar los préstamos propensos a ser cancelados anticipadamente?	Construir un modelo matemático utilizando la regresión logística binomial para estimar la probabilidad que tiene un préstamo de ser cancelado anticipadamente.	Regresión logística binomial, pruebas de Wald	Para generar el modelo con mejor ajuste, previo se balancearon los datos mediante el algoritmo ROSE y posterior a ello se generó el modelo de regresión logístico con el método de máxima verosimilitud.	2. El modelo matemático construido mediante regresión logística binomial utilizando variables comportamentales y del producto con una significancia del 5%, permitió estimar la probabilidad que tiene cada préstamo de ser cancelado anticipadamente, utilizando un punto de corte del 50%. Para ello, previo a la generación del modelo fue necesario realizar un balance de las clases mediante el algoritmo ROSE para aumentar la capacidad predictiva del modelo a un 65% en el evento de interés.	2. Con el modelo de regresión logístico generado se sugiere a la entidad financiera actualizar mensualmente en los primero días del mes la probabilidad de cancelación anticipada de cada préstamo, con base al estado de las variables al cierre del mes anterior.
	¿Cuál es el error de estimación de un modelo que categoriza a los préstamos como propensos o no propensos a ser cancelados anticipadamente?	Calcular el error de estimación del modelo matemático por medio de la técnica de validación cruzada.	Validación cruzada, área bajo la curva y pseudo R2.	Por tratarse de estados desbalanceados, se evaluó como principal media la sensibilidad que asciende a un 65% y el área bajo la curva que corresponde a un 61%. Con un error estimado del 35%, y uno máximo del 38.3%.	3. El error de estimación del modelo calculado por medio de la técnica de validación cruzada sobre los datos de prueba es del 35%, así también mediante la curva ROC se determinó que el modelo generado posee un área bajo la curva del 61%, y por medio de la técnica del bootstrap se determinó que el error asciende a un 38.3%. Al aplicar la prueba ratio-likelihood con un 95% de confianza se comprobó que el	3. Resulta conveniente monitorear mensualmente el ajuste del modelo con las métricas de sensibilidad y exactitud por cada grupo, considerando que los datos son desbalanceados y el error máximo estimado asciende a un 38.3%. Conjuntamente de evaluar la evolución de los resultados a través de curvas ROC.

Continuación apéndice 1.

					modelo es significativo con respecto a un modelo nulo, es decir sin predictores.	

Fuente: elaboración propia.

Apéndice 2. **Código R Modelo**

Código R para generación de modelo de regresión logística binaria, con una muestra sintética con base en el algoritmo ROSE.

```
# ============================================================
        # MODELO 2: ROSE con datos sintéticos
# ============================================================
         library(ROSE)
        # generar datos balanceados con ROSE
        # p = proporcion de la muestra
         datos.rose <- ROSE(estado ~ .
                    ,data=datos.entrenamiento, p=0.5, seed=1)$data
         summary(datos.rose)

        # plot datos desbalanceados y balanceados metodo ROSE.
         plot(datos.entrenamiento$estado)
         table(datos.entrenamiento$estado)

         plot(datos.rose$estado)
         table(datos.rose$estado)

# ============================================================
        # MODELO 2: regresión con datos balanceados método ROSE
# ============================================================
        modelo_rose <-  glm(estado ~ .
                    ,
```

Continuación apéndice 2.

```
            family=binomial (link="logit"),
                data = datos.rose)

# ======================================
#  step(object, scope, scale = 0,
#  direction = c("both", "backward", "forward"),
#  trace = 1, keep = NULL, steps = 1000, k = 2, ...)
step(modelo_rose, direction = "both")

# ======================================
# mejor modelo
 modelo_rose<- glm(formula = estado ~ cat_ciclo + cat_plazo + cat_edad + cat_prestamos+
       cat_impago + catpag1 + cat_cliente + cat_cancelacion_nat +
       cat_cancelacion_previa, family = binomial(link = "logit"), data = datos.rose)

# ======================================
# ver significancia de variables test Wald chi-test
# coeficiente < 0.05 = variable significativa,
# el coeficiente es distinto de cero
 summary(modelo_rose)

# intervalos de confianza, 95% de confianza
 confint(modelo_rose, level = 0.95)

# ======================================
# Calcular la probabilidad del modelo logístico binomial
 prob.modelo.rose <-
   predict.glm(modelo_rose, newdata = datos.prueba, type = "response")
 prob.modelo.rose
# threshold asignar estado de cancelacion segun probabilidad
 pred.modelo.rose <- rep(0, length(prob.modelo.rose))
```

Continuación apéndice 2.

```
pred.modelo.rose[prob.modelo.rose > 0.5] <- 1
 pred.modelo.rose
# Matriz de confusión
 matriz.confusion <- table(datos.prueba$estado,
                           pred.modelo.rose,
                           dnn=c("real","modelo"))
matriz.confusion

# pseudo R2 Nagelkerke, nivel explicativo variables independientes
 library(fmsb)
 NagelkerkeR2(modelo_rose)

# ======================================
# odds ratio: tienen n cantidad de veces más probabilidad de
# cancelar que los que no tienen la condicion.
# ======================================
 exp(modelo_rose$coefficients)

# intervalo de confianza odds
# li>1 y ls>1 = direccion inferible a la poblacion
 exp(confint(modelo_rose))

# ======================================
# Likelihood ratio: examinar modelo nulo
# no rechazar hipotesis nula >0.05 = variables independientes significativas
# rechazar hipotesis nula < 0.05 = por lo menos una variable no significativa
 dif_residuos <- modelo_rose$null.deviance - modelo_rose$deviance
 grados_libertad <- modelo_rose$df.null - modelo_rose$df.residual
 p_value <- pchisq(q = dif_residuos,df = grados_libertad, lower.tail = FALSE)
 # diferencia de residuos de deviance modelo nulo - modelo
 dif_residuos
```

Continuación apéndice 2.

```
# grados de libertad
 grados_libertad
# p-valor
 p_value

# =====================================
# devianza modelo < devianza modelo null = mejor modelo
# AIC: el menor indica el mejor modelo
# devianza nula < dif_residuos<
 summary(modelo_rose)
# =====================================
# curva roc: area bajo la curva
# =====================================
 library(ROCR)
 pred = ROCR::prediction(prob.modelo.rose,datos.prueba$estado)
 perf <- performance(pred, "tpr", "fpr")
 plot(perf)

 AUC_rose=performance(pred, measure = "auc")@y.values[[1]]
 cat("Area Bajo la Curva: ",AUC_rose,"n")

 library(ROSE)
 roc.curve(datos.prueba$estado, prob.modelo.rose)
 roc.curve(datos.prueba$estado, pred.modelo.rose)
 # acurracy
 table(datos.prueba$estado,pred.modelo.rose)
 accuracy.meas(datos.prueba$estado, prob.modelo.rose, threshold = 0.50)

# =====================================
# multicolinealidad: <10 sin multicolinealidad en las variables
 library(ROCR)
```

Continuación apéndice 2.

```
 vif(modelo_rose)

# bootstrap
boot <- ROSE.eval(estado ~ .,
                  data=datos.rose,
                  learner = glm,
                  method.assess = "BOOT",
                  control.learner = list(family=binomial),
                  trace=TRUE,
                  B=100)
summary(boot)
```

Fuente: elaboración propia.

Apéndice 3. **Código R Pruebas de Independencia**

```
#=============================================================
# pruebas de independencia
# =============================================================

# manta$cat_ciclo
# manta$cat_plazo
# manta$cat_genero
# manta$cat_estado_civil
# manta$cat_edad
# manta$cat_region

# Pearson's Chi-squared test:
 # bondad de ajuste (goodness of fit tests
 test <- table(manta$estado,manta$cat_ciclo)
 chitest <- chisq.test(test)
 chitest
```

Continuación apéndice 3.

```
# valores versus esperados
# si valor esperado < 5 entonces test de independencia Fischers test
# si x-squared > qchisq entonces se rechaza ho
chitest$observed
chitest$expected
chitest
qchisq(0.95,1)
chitest$parameter
chitest$p.value
round(chitest$p.value,4)

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
test <- table(manta$estado,manta$cat_plazo)
chitest <- chisq.test(test)
chitest

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
test <- table(manta$estado,manta$cat_genero)
chitest <- chisq.test(test)
chitest

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
test <- table(manta$estado,manta$cat_estado_civil)
chitest <- chisq.test(test)
chitest

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
```

Continuación apéndice 3.

```
test <- table(manta$estado,manta$cat_edad)
 chitest <- chisq.test(test)
 chitest

 # Pearson's Chi-squared test:
 # si x-squared > qchisq entonces se rechaza ho
 test <- table(manta$estado,manta$cat_region)
 chitest <- chisq.test(test)
 chitest



# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
 test <- table(manta$estado,manta$cat_impago)
 chitest <- chisq.test(test)
 chitest

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
 test <- table(manta$estado,manta$cat_prestamos)
 chitest <- chisq.test(test)
 chitest

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
 test <- table(manta$estado,manta$catpag1)
 chitest <- chisq.test(test)
 chitest

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
```

Continuación apéndice 3.

```
test <- table(manta$estado,manta$cat_cliente)
 chitest <- chisq.test(test)
 chitest

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
 test <- table(manta$estado,manta$cat_cancelacion_previa)
 chitest <- chisq.test(test)
 chitest

# Pearson's Chi-squared test:
# si x-squared > qchisq entonces se rechaza ho
 test <- table(manta$estado,manta$cat_cancelacion_nat)
 chitest <- chisq.test(test)
 chitest
```

Fuente: elaboración propia.

ANEXOS

Anexo 1. **Descripción paquete ROSE**

ROSE-package ROSE: Random Over-Sampling Examples

Description

Functions to deal with binary classification problems in the presence of imbalanced classes. Syn- thetic balanced samples are generated according to ROSE (Menardi and Torelli, 2014). Functions that implement more traditional remedies to the class imbalance are also provided, as well as dif- ferent metrics to evaluate a learner accuracy. These are estimated by holdout, bootrstrap or cross- validation methods.

Details

The package pivots on function ROSE which generates synthetic balanced samples and thus allows to strenghten the subsequent estimation of any binary classifier. ROSE (Random Over-Sampling Examples) is a bootstrap-based technique which aids the task of binary classification in the presence of rare classes. It handles both continuous and categorical data by generating synthetic examples from a conditional density estimate of the two classes. Different metrics to evaluate a learner accuracy are supplied by functions roc.curve and accuracy.meas. Holdout, bootstrap or cross-validation estimators of these accuracy metrics are computed by means of ROSE and provided by function ROSE.eval, to be used in conjuction with virtually any binary classifier. Additionally, function ovun.sample implements more traditional

Continuación anexo 1.

remedies to the class imbalance, such as over-sampling the minority class, under-sampling the majority class, or a combination of over- and under- sampling.

Author(s)

Nicola Lunardon, Giovanna Menardi, Nicola Torelli Maintainer: Nicola Lunardon <lunardon@stat.unipd.it>

References

Lunardon, N., Menardi, G., and Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learn- ing. R Jorunal, 6:82–92.

Menardi, G. and Torelli, N. (2014). Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery, 28:92–122.

See Also

nnet, rpart

Examples

```
# loading data
data(hacide)

# check imbalance
accuracy.meas 3
table(hacide.train$cls)
```

Continuación anexo 1.

```
# train logistic regression on imbalanced data
log.reg.imb <- glm(cls ~ ., data=hacide.train, family=binomial)

# use the trained model to predict test data
pred.log.reg.imb <- predict(log.reg.imb, newdata=hacide.test, type="response")

# generate new balanced data by ROSE
hacide.rose <- ROSE(cls ~ ., data=hacide.train, seed=123)$data

# check (im)balance of new data
table(hacide.rose$cls)

# train logistic regression on balanced data
log.reg.bal <- glm(cls ~ ., data=hacide.rose, family=binomial)

# use the trained model to predict test data
pred.log.reg.bal <- predict(log.reg.bal, newdata=hacide.test,
type="response")

# check accuracy of the two learners by measuring auc
roc.curve(hacide.test$cls, pred.log.reg.imb)
roc.curve(hacide.test$cls, pred.log.reg.bal, add.roc=TRUE, col=2)

# determine bootstrap distribution of the AUC of logit models

# trained on ROSE balanced samples
```

Continuación anexo 1.

```
# B has been reduced from 100 to 10 for time saving solely
boot.auc.bal <- ROSE.eval(cls ~ ., data=hacide.train, learner= glm,
method.assess = "BOOT", control.learner=list(family=binomial), trace=TRUE, B=10)
summary(boot.auc.bal)
```

Fuente: ROSE (2022). *Random Over-Sampling Examples*.

Anexo 2. **Descripción funciones paquete ROSE**

Funtion	Description
accuracy.meas	This function computes precision, recall and the F measure of a prediction.
hacide	Prediction of positive or negative labels depends on the classification threshold, here defined as the value such that observations with predicted value greater than the threshold are assigned to the positive class. Some caution is due in setting the threshold as well as in using the default setting both because the default value is meant for predicted probabilities and because the default 0.5 is not necessarily the optimal choice for imbalanced learning. Smaller values set for the threshold correspond to assign a larger misclassification costs to the rare class, which is usually the case.
ovun.sample	Simulated training and test set for imbalanced binary classification. The rare class may be described as a half circle depleted filled with the prevalent class, which is normally distributed and has elliptical contours.

Continuación anexo 2.

roc.curve	This function returns the ROC curve and computes the area under the curve (AUC) for binary classifiers.
ROSE	Creates a sample of synthetic data by enlarging the features space of minority and majority class examples. Operationally, the new examples are drawn from a conditional kernel density estimate of the two classes, as described in Menardi and Torelli (2013).
ROSE.eval	Given a classifier and a set of data, this function exploits ROSE generation of synthetic samples to provide holdout, bootstrap or leave-K-out cross-validation estimates of a specified accuracy measure.

Fuente: ROSE (2022). *Random Over-Sampling Examples*.