



UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA

Proyecto de Título:

*Modelo de scoring para clientes residenciales de
Essbio y Nuevosur*

Profesor Patrocinante	: Luisa Rivas Calabrán	Firma	
Profesor Colaborador	: Tamara Pardo Márquez	Firma	
Profesor Consejero	: Sebastián Niklitschek Soto	Firma	
Nombre Memorante	: Carolina Espinoza Zapata	Firma	

Concepción, Mayo 2020.



UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA

Proyecto de Título:
**Modelo de scoring para clientes residenciales de
Essbio y Nuevosur**

*Memoria para optar al título
profesional de Ingeniero Estadístico*

Carolina Elizabeth Espinoza Zapata

Prof. Guía: Dra. Luisa Rivas Calabrán

Concepción, Mayo 2020

Agradecimientos

Gracias a Dios, por haber trazado mi camino y eliminado cada obstáculo que se presentó, por haberme fortalecido y acompañado durante este proceso de aprendizaje intelectual y personal.

Gracias a mi familia, por su inconmensurable amor y comprensión, a mi madre por su apoyo incondicional, por las enseñanzas y valores entregados, y más aún, a mi padre, por su esfuerzo y sacrificio cada día en pos de otorgar lo mejor a nuestra familia con las pocas herramientas que la vida le ha entregado. Gracias por darme la oportunidad de estudiar, por luchar para que yo pueda ser alguien y logre salir adelante.

Gracias a mi compañero, Sebastián, por su admirable paciencia y amor, por su apoyo en cada decisión, y sobre todo, por cada momento de felicidad que me dio nuevas fuerzas para continuar cuando el camino se hacía difícil. Gracias a mi mejor amiga, Fernanda, por su apoyo y comprensión en este largo proceso, por ser la mejor compañía y la mejor consejera en momentos de duda y temor.

Gracias a cada uno de los profesores de los cuales tuve la oportunidad de aprender en la Universidad, gracias por su dedicación y por el conocimiento entregado. En especial, gracias a la jefa de carrera, Dra. Katia Sáez, por su preocupación y buena disposición, así como a mi profesora guía en este proyecto, Dra. Luisa Rivas, por su tiempo, su ayuda y apoyo en este trabajo.

Gracias a la empresa sanitaria Essbio, y en particular a la ingeniera estadística, Tamara Pardo, por darme la oportunidad de incursionar por primera vez en el trabajo de un estadístico a través de la práctica profesional. Gracias por la confianza y por permitirme fortalecer y aplicar mis conocimientos en beneficio de la empresa. Gracias por su comprensión, apoyo y buena disposición a lo largo del proceso. Gracias al departamento de cobranza y medios de pago por su cálida acogida.

A cada uno, mis más sinceros agradecimientos.

Resumen

El presente estudio se ha desarrollado con el objetivo de obtener un modelo de scoring para clientes residenciales de una empresa sanitaria. La idea es identificar a aquellos clientes que estén más propensos a convertirse en cuentas incobrables, lo cual ocurre cuando el cliente alcanza 720 o más días de atraso en el pago de su deuda. Para este propósito, se han considerado dos técnicas de aprendizaje automático, por un lado, la regresión logística, por ser el método más ampliamente conocido y utilizado bajo este contexto, y por otro, la potenciación del gradiente en árboles de decisión, el cual, si bien es un método menos reconocido, ha demostrado tener buenos resultados en diversos problemas de clasificación.

Se debe tener en cuenta que, el evento de llegar a ser una cuenta incobrable es bastante raro, lo cual genera un conjunto de datos no balanceados. Se analizará cómo esto afecta a los algoritmos propuestos y cuál es capaz de presentar mejores resultados en términos de precisión y poder de discriminación bajo estas circunstancias. Si bien, tanto el poder predictivo como la precisión alcanzada son atributos importantes para un modelo de clasificación, otras cuestiones, como el tiempo de ejecución y el poder computacional necesario se deben tener en consideración. El modelo seleccionado será aquel que mejor se adapte a las necesidades y características de la empresa. Finalmente, se propone una segmentación de riesgo, basada en las probabilidades estimadas, que permita clasificar a los clientes de acuerdo a su probabilidad de convertirse en incobrable.

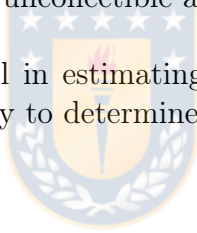
Es de esperar que, este modelo pueda ser utilizado para estimar la pérdida esperada debido a las cuentas por cobrar y que represente una manera confiable de determinar la provisión de incobrables.

Abstract

The present study has been developed in order to propose a scoring model for residential customers in a water and sanitation company. The aim is to identify those customers who are more likely to become uncollectible accounts, which happens when the customer is 720 or more days late in the payment of their debt. For this purpose, two machine learning algorithms have been approached, namely, logistic regression, as it is a widely known and used method for this sort of model, and gradient boosting decision tree which, though less famous, has proved remarkable results in classification problems.

Also, it must be noticed that the event of becoming an uncollectible account is a rather rare event, which in turn, generates an imbalanced dataset. We will see how this affects the algorithms and which one does a better job in discriminating and predicting each label under this circumstances. Although, the predictive power and precision achieved are important attributes for any classification model, other issues like execution time and computing power must be taken into account. The chosen model is the one which best suits the company needs and characteristics. Finally, a risk segmentation is proposed based on the estimated probabilities, which will allow to classify customers according to their probability of becoming an uncollectible account.

Hopefully, this model will be useful in estimating the expected loss due to accounts receivable and provide a reliable way to determine the provision for uncollectible accounts.



Índice general

1. Introducción	9
2. Contexto	11
2.1. Essbio y Nuevosur	11
2.1.1. Tipo de clientes	13
2.1.2. Gestión de cobranza	13
2.1.3. Convenios de pago	16
2.2. Conceptos relevantes	17
2.2.1. Provisión de incobrable	17
2.2.2. NIIF 9 Instrumentos Financieros	17
2.2.3. Modelo de Scoring de Clientes	18
3. Objetivos	20
3.1. Objetivo General	20
3.2. Objetivos Específicos	20
4. Metodología	21
4.1. Definición de la variable objetivo	21
4.2. Definición del período de observación y desempeño	23
4.3. Obtención del Universo	23
4.4. Obtención de la muestra	25
4.5. Obtención y depuración de los datos	27
4.6. Definición de variables explicativas	28
4.7. Análisis exploratorio	28
4.7.1. D de Somers corregida	29
4.7.2. Datos atípicos	29
4.7.3. Datos <i>missing</i>	29
5. Marco teórico	30
5.1. Técnicas Estadísticas	30
5.1.1. Regresión Logística	30
5.1.2. Potenciación del Gradiente	36
5.2. Criterios de evaluación de modelos	39

<i>ÍNDICE GENERAL</i>	6
5.2.1. Curva ROC	39
5.2.2. Matriz de confusión	40
5.2.3. Validación cruzada de q -capas	41
6. Implementación y Resultados	43
6.1. Regresión logística	43
6.1.1. Resultados	45
6.2. Potenciación del gradiente de árboles de decisión	53
6.2.1. Resultados	53
6.3. Segmentación del riesgo	59
6.4. Cálculo del puntaje	62
7. Consideraciones finales	63
7.1. Discusión	63
7.2. Conclusión	65
8. Anexo	66
8.1. Variables regresión logística	66
8.2. Variables potenciación del gradiente de árboles de decisión.	72



Índice de figuras

2.1. Estructura Subgerencia Gestión Comercial.	12
2.2. Proceso de cobranza	15
4.1. Períodos de observación y desempeño.	23
5.1. Curva de ROC.	39
5.2. Procedimiento validación cruzada.	42
6.1. Curva de ROC modelo de regresión logística.	50
6.2. Curva ROC modelo potenciación del gradiente.	57
6.3. Distribución de eventos por intervalo de probabilidad estimada.	60
6.4. Eventos por intervalo de probabilidad estimada mayor a 0,000829.	61
8.1. Distribución de eventos de acuerdo a la variable 1.	66
8.2. Distribución de eventos de acuerdo a la variable 2.	67
8.3. Distribución de eventos de acuerdo a la variable 3.	67
8.4. Distribución de eventos de acuerdo a la variable 4.	68
8.5. Distribución de eventos de acuerdo a la variable 5.	68
8.6. Distribución de eventos de acuerdo a la variable 6.	69
8.7. Distribución de eventos de acuerdo a la variable 7.	69
8.8. Distribución de eventos de acuerdo a la variable 8.	70
8.9. Distribución de eventos de acuerdo a la variable 9.	70
8.10. Distribución de eventos de acuerdo a la variable 10.	71
8.11. Distribución de eventos de acuerdo a la variable 11.	71
8.12. Distribución de eventos de acuerdo a la variable 1.	72
8.13. Distribución de eventos de acuerdo a la variable 2.	73
8.14. Distribución de eventos de acuerdo a la variable 3.	73
8.15. Distribución de eventos de acuerdo a la variable 4.	74
8.16. Distribución de eventos de acuerdo a la variable 5.	74
8.17. Distribución de eventos de acuerdo a la variable 6.	75
8.18. Distribución de eventos de acuerdo a la variable 7.	75
8.19. Distribución de eventos de acuerdo a la variable 8.	76
8.20. Distribución de eventos de acuerdo a la variable 9.	76
8.21. Distribución de eventos de acuerdo a la variable 10.	77

8.22. Distribución de eventos de acuerdo a la variable 11. 77



Capítulo 1

Introducción

La industria sanitaria en Chile es un monopolio regulado, que en sus inicios estuvo gestionado y dirigido por CORFO y por ende era una empresa pública. Sin embargo, actualmente, un 94 % de la industria está bajo la gestión privada, esto dado la alta inversión que se requería para modernizar la infraestructura existente y ampliar la capacidad y servicios prestados. Se distinguen dos modalidades de propiedad, empresas vendidas casi en su totalidad a capitales privados en las que el estado mantiene, a lo menos, un 5 % de participación y empresas privadas que participan de la industria mediante concesiones, generalmente a 30 años, que pueden ser o no renovables (Valenzuela, 2018).

Al ser un monopolio regulado, existe una activa participación del ente fiscalizador del estado que en este caso es la Superintendencia de Servicio Sanitarios (SISS), es este organismo quien debe velar por la correcta operación y entrega de un buen servicio a los clientes, también es la encargada de fijar las tarifas y exigir el cumplimiento de la regulación medioambiental.

Las empresas sanitarias encargadas de prestar sus servicios de producción y distribución de agua potable, recolección y tratamiento de aguas servidas y mantenimiento de alcantarillados en las regiones del Libertador General Bernardo O'Higgins (VI), Maule (VII), Biobío (VIII) y Ñuble (XVI) son Essbio S.A y Nuevosur S.A. Essbio es una empresa sanitaria constituida por venta de activos, mientras que la sanitaria Nuevosur es un arriendo de concesión a 30 años, ambas pertenecientes al fondo de inversión Ontario Teachers' Pension Plan (OTPP), donde Nuevosur S.A. es administrada por Essbio S.A. desde Concepción (Valenzuela, 2018).

Con el objetivo de modernizar sus procesos y prevenir pérdidas por el no pago de servicios prestados, la empresa Essbio-Nuevosur busca adaptarse al nuevo marco de las Normas Internacionales de Información Financiera (NIIF), en las cuales se propone un modelo de deterioro basado en pérdidas esperadas. Uno de los requisitos para aplicar

este modelo es conocer la probabilidad de incumplimiento de los clientes, esto es, la probabilidad de que la cuenta se vuelva incobrable y por lo tanto deba provisionarse.

En este marco, surge el objetivo de este proyecto, el cual consiste en desarrollar un modelo de scoring para clientes residenciales, que permita predecir la probabilidad de que el cliente caiga en incobrabilidad dentro de 30 meses a partir de una fecha de referencia y asignarle un puntaje a partir de ella. Se entenderá por cliente incobrable todo aquel que llegue a tener 720 o más días de atraso en el pago.

Dado que la condición antes descrita representa un caso extremo, ocurre en pocas ocasiones, es decir, el volumen de clientes residenciales incobrables es bastante pequeño. A septiembre de 2019, sólo un 2.74% de estos clientes llegan a cumplir esta condición, lo cual implica que el modelo es construido en base a datos no balanceados, donde los no eventos superan ampliamente a los eventos.

Con el propósito de cumplir el objetivo planteado, se propuso desarrollar un modelo de scoring mediante el método convencional y ampliamente utilizado de la regresión logística, y además se utilizó una herramienta de aprendizaje automático denominada potenciación del gradiente, la cual fue aplicada sobre árboles de decisión. Finalmente, se seleccionó el modelo que, además de tener un buen desempeño en términos de precisión, se adaptó mejor al contexto del problema y su utilización en la empresa. Las probabilidades de incumplimiento estimadas a partir del modelo seleccionado fueron utilizadas para calcular un puntaje y asignar al cliente dentro de una categoría de riesgo de incobrabilidad.

El presente documento se encuentra organizado tal como se describe a continuación. En el Capítulo 2, se presenta un resumen de aspectos referentes a la empresa que es necesario tener en consideración para entender su problemática y poder plantear una solución acorde a sus necesidades. En el Capítulo 3, se presentan los objetivos que se plantearon para este proyecto. El Capítulo 4, está destinado a presentar la metodología o proceso que se debió llevar a cabo para obtener la tabla de datos necesaria al momento de ajustar un modelo de clasificación. En el quinto Capítulo, se encuentra una descripción de las técnicas estadísticas a implementar para la obtención de los modelos de clasificación propuestos. Luego, en el Capítulo 6, se habla sobre la implementación de los algoritmos en el software R Studio, así como de los resultados obtenidos para cada método aplicado. En el Capítulo 7, se exponen las conclusiones finales y por último, se presenta un Anexo, donde se encuentra una visualización de la relación entre el evento de interés y las variables seleccionadas en los respectivos modelos.

Capítulo 2

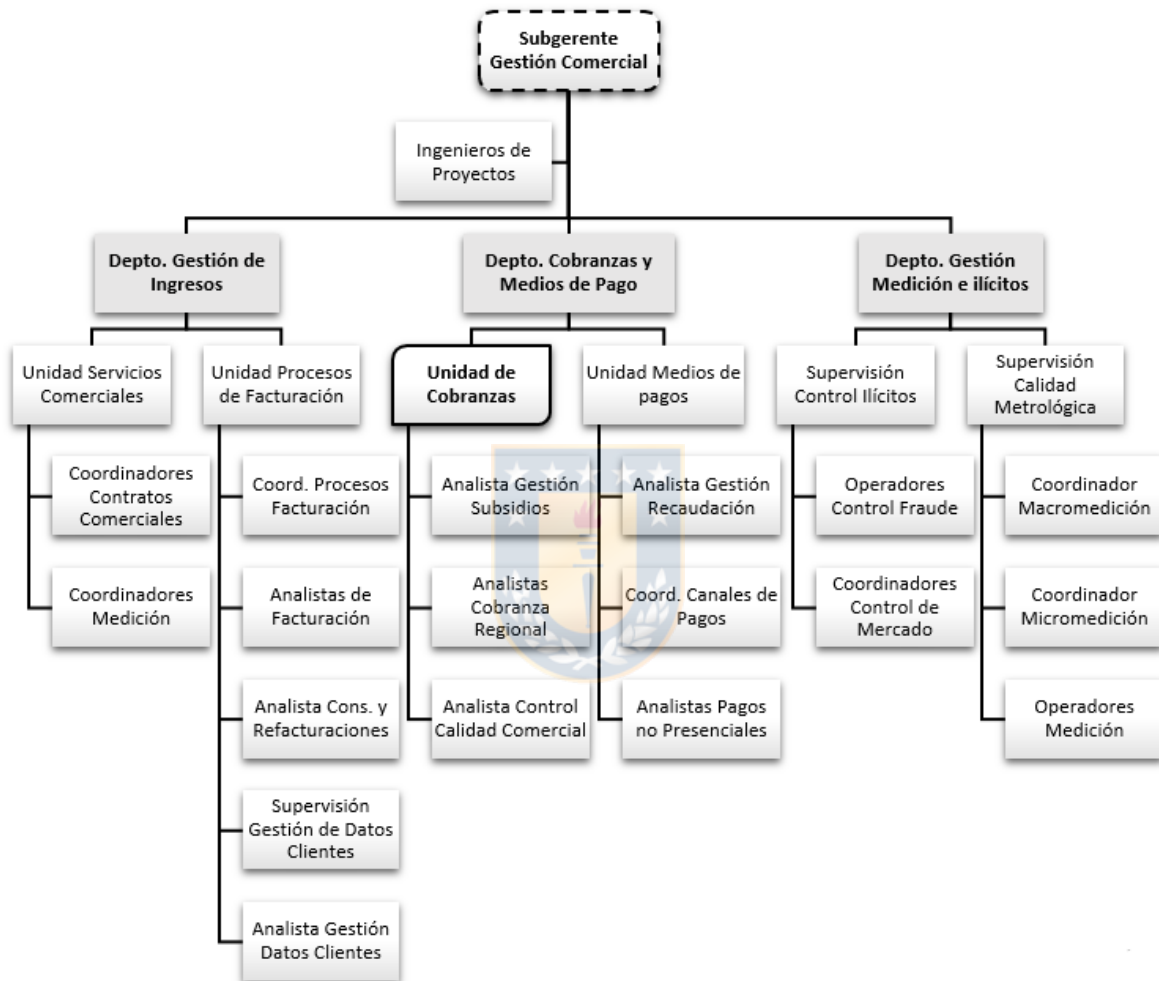
Contexto

2.1. Essbio y Nuevosur

Essbio S.A (Empresa de servicios sanitarios del Biobío) es la sanitaria encargada de proveer soluciones y servicios de alta calidad en el ciclo integral del agua (agua potable, alcantarillado y descontaminación de aguas servidas) y en la gestión integral de residuos en las regiones del Biobío, Maule, Ohiggins y actualmente Ñuble. Fue constituida en 1990, junto a otras 11 sociedades como sucesoras legales de Sendos (Servicio nacional de obras sanitarias) y fue privatizada el año 2000 a manos de Thames Water. En 2004, Essbio comienza a prestar servicios gerenciales a Nuevosur, sanitaria que opera en la Región del Maule.

Este proyecto se llevó a cabo en la Gerencia de Clientes de la empresa, subgerencia de Gestión Comercial, departamento de Cobranzas y Medios de Pagos. Este último es el encargado de asegurar el cobro de los servicios facturados, entregando a los clientes facilidad de pago con el objetivo de maximizar los flujos de caja y la sostenibilidad de la compañía. En la Figura 2.1 se puede ver la estructura organizacional de la subgerencia de gestión comercial con sus respectivas funciones por departamento.

Los servicios del área comercial son diversos y comienzan a prestarse una vez incorporado el cliente al sistema, el cual es atendido de forma periódica en función de lotes designados de acuerdo a la ubicación geográfica o sector al cual pertenece la instalación. Dicho proceso comercial consta de siete subprocesos que permiten entregar una atención completa al cliente por el servicio prestado: Medición, Lectura, Facturación, Impresión y Reparto, Recaudación, Cobranza y Control de Ilícitos (Valenzuela, 2018).



Fuente: Essbio (2014)

Figura 2.1: Estructura Subgerencia Gestión Comercial.

2.1.1. Tipo de clientes

De acuerdo al uso que se le da a la casa habitación, los clientes de servicios de agua potable y alcantarillado de Essbio-Nuevosur se clasifican de la siguiente manera:

- **Cliente Residencial:** es aquel cliente cuyo inmueble en el cual recibe el servicio sanitario, está destinada principalmente a fines residenciales (50 % o más de su construcción) o que los consumos facturados (m^3) sean mayoritariamente de este tipo.
- **Cliente Comercial:** es un cliente tal que la propiedad asociada al servicio sanitario está destinada principalmente a actividades comerciales (50 % o más de su construcción) o que los consumos facturados (m^3) sean mayoritariamente de este tipo.
- **Cliente Industrial:** es aquel cliente cuyo inmueble en el cual recibe el servicio sanitario de agua potable, alcantarillado o ambos, está destinada principalmente a actividades industriales (50 % o más de su construcción) o que los consumos facturados (m^3) sean mayoritariamente de este tipo.
- **Cliente Fiscal:** es un cliente tal que la propiedad asociada al servicio sanitario, está destinada principalmente a actividades desarrolladas por instituciones de tipo público o fiscal (50 % o más de su construcción) o que los consumos facturados (m^3) sean mayoritariamente de este tipo. Dentro de esta categoría deben ser considerados por ejemplo: bomberos, carabineros, policía de investigaciones, ejército y fuerzas armadas en general, instituciones pertenecientes al servicio público de salud (hospitales, consultorios, postas, etc.), instituciones educacionales (colegios municipales, universidades tradicionales, universidades privadas, etc.), instituciones pertenecientes a la administración pública (Municipalidades, Intendencias, Gobernaciones, Ministerios, Subsecretarías, Superintendencias, SEREMIS, etc.).
- **Cliente Fuente propia:** Son aquellos clientes que cuentan con su propia fuente de abastecimiento de agua potable (puntera, pozo, etc), y a los cuales se les presta por lo tanto, sólo el servicio de alcantarillado.

Tal como se indicó en la introducción, este proyecto está enfocado en los clientes de tipo residencial, los cuales constituyen más de un 90 % de los clientes de la empresa. A continuación, se detallan algunos aspectos del proceso de cobranza que se aplica a este tipo de clientes, así como de las opciones de pago que se ofrecen para ellos.

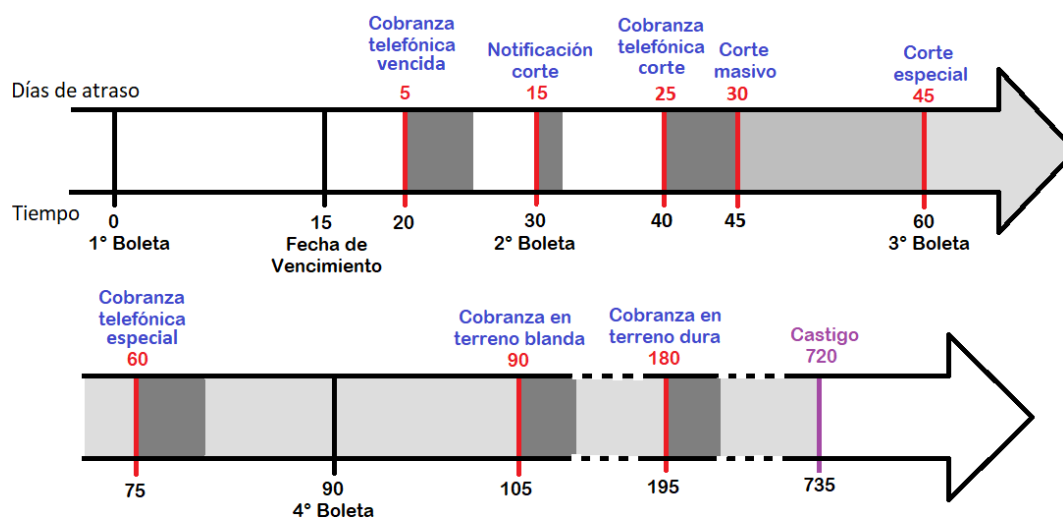
2.1.2. Gestión de cobranza

La gestión de cobranza es prácticamente el último eslabón de la cadena comercial, pero no por eso menos importante, ya que es la encargada de adelantar los flujos de caja y

disminuir las provisiones por deudores incobrables para no afectar negativamente los estados financieros de la empresa. A esta labor se designa un 31 % del presupuesto anual asignado a la subgerencia comercial, por lo que el foco que se le entrega a esta área es sumamente relevante.

El proceso de cobranza inicia cuando el cliente alcanza 5 días de atraso en el pago de su cuenta, contando a partir de la fecha de vencimiento especificada en su boleta, en esta instancia se recurre a la acción de cobranza telefónica, la cual consiste en la realización de una llamada o el envío de un SMS para recordar al cliente que tiene una deuda impaga. Esta gestión es realizada por el *Contact Center* Atento, el cual dispone de un plazo de entre 1 y 5 días para llevar a cabo la acción.

Posteriormente, cuando el cliente alcanza 15 días de atraso, se genera una orden de notificación, la cual consiste en un mensaje adjunto a la próxima boleta advirtiéndolo sobre la fecha de corte del suministro. Este último se lleva a cabo 15 días hábiles después de la notificación, y se conoce como la instancia de corte masivo, sin embargo, 5 días hábiles previo a ello, se realiza una llamada telefónica para recordar al cliente sobre su fecha de corte del suministro de agua potable. Cuando el cliente ha alcanzado 45 días de atraso pasa a formar parte de una cartera de corte especial, mientras que a los 60 días de atraso se realiza una nueva llamada telefónica. Por último, cuando las gestiones anteriores no han logrado la amortización de la deuda, se recurre a acciones más costosas, que consisten en la visita de un ejecutivo en terreno cuyo propósito es negociar el pago de la deuda, esto ocurre cuando el cliente ya ha alcanzado 90 días de atraso (cobranza en terreno blanda) y 180 días de atraso en su deuda (cobranza en terreno dura). En la Figura 2.2 se aprecia este proceso con mayor claridad, los índices numéricos corresponden a días hábiles.



Fuente: Elaboración propia.

Figura 2.2: Proceso de cobranza



2.1.3. Convenios de pago

Existen facilidades de pago que se otorgan a clientes con deuda vigente que cumplen ciertas características, algunas de las cuales se especifican en el Cuadro 2.1. De acuerdo a ellas, los denominados planes o convenios de pago pueden ser de distintos tipos y están destinados principalmente a clientes de bajo nivel socioeconómico, así como aquellos que acumulan una gran deuda por haber cometido algún ilícito, comúnmente la intervención de algún medidor de consumo de agua. Un convenio es desactivado cuando el cliente incurre en el impago de dos cuotas facturadas consecutivas.

Tipo	Modalidad	Cuotas	Orientación	Días de atraso	Nivel socio-económico
Carenciado	sin interés con pie	1 - 60	Clientes con incapacidad de pago	≥ 90	Medio, bajo
Pago Fácil	con interés con pie	1 - 24	Clientes con capacidad de pago	≥ 90	Todos
Instalación	sin interés con pie	1 - 12	Clientes con venta medidor	≥ 90	Medio, bajo
Control Mercado	sin interés sin pie	1 - 24	Clientes con ilícito	≥ 0	Todos
Renovación Control Mercado	sin interés con pie	1 - 24	Clientes que caduca un plan CM	≥ 60	Todos
Consumo	sin interés con pie	1 - 12	Exceso consumo	≥ 90	Todos

Cuadro 2.1: Tipo de convenios de pago.

2.2. Conceptos relevantes

2.2.1. Provisión de incobrable

Cuando una empresa vende a crédito o sin haber cobrado el dinero en ese momento, asume el riesgo de que un porcentaje de sus clientes no pague sus deudas. Ese dinero que no puede recuperar constituye una pérdida y por lo tanto debe aparecer en el ejercicio contable como gasto. Es inevitable que esto ocurra, luego, las empresas deben tener cubiertas dichas pérdidas, tanto las esperadas, a través de provisiones, como las inesperadas, con capital propio.

Una provisión para cuentas incobrables es un tipo de cuenta de salvaguardia establecida por muchas empresas. La función principal de este tipo de cuenta es proporcionar un colchón contra las facturas de clientes que estén pendientes de pago durante períodos prolongados de tiempo y cuya recuperación es dudosa e incierta.

Para calcular esta provisión existen distintos métodos, los que varían de una entidad a otra, sin embargo, lo más aceptado es aplicar un modelo basado en los porcentajes de incobrabilidad de períodos anteriores, complementado con el criterio de expertos en la empresa (Barrantes, 2017).

2.2.2. NIIF 9 Instrumentos Financieros

La NIIF 9 se originó a raíz de la crisis económica en EE.UU, cuando el comité de normas internacionales de contabilidad, IASB por su sigla en inglés, y el consejo de Normas de Contabilidad Financiera, FASB, coincidieron en que existían ciertas debilidades en las normas, las cuales de haberse identificado e implementado mejoras a tiempo, habrían permitido prevenir la crisis, específicamente en el cálculo de la provisión en base a pérdidas incurridas. Se evidencia el hecho de que, no se puede esperar a que la pérdida ya se haya materializado para establecer que hay un deterioro y por ende, concluyen que se hace necesario anticiparse a los posibles detrimentos que puedan suceder, dándole paso entonces al nuevo modelo de pérdida esperada planteado por la NIIF 9 (Ríos, 2017). En este contexto, se entiende por deterioro la pérdida que se produce cuando el importe en libros de un activo es superior a su importe recuperable.

La NIIF 9 Instrumentos Financieros, tiene como objetivo “establecer los principios para la información financiera sobre activos y pasivos financieros, de forma que se presente información útil y relevante para los usuarios” (IFRS Foundation, 2014). Es importante destacar que, la NIIF 9 reemplaza la NIC 39 (Norma internacional de contabilidad 39), presentando cambios en la clasificación de activos financieros, en el reconocimiento de las pérdidas por riesgo de crédito y en la contabilidad de coberturas, comenzando su implementación obligatoria en los ejercicios desde el 1 de enero de 2018.

De acuerdo con la norma, el nuevo mecanismo de medición del deterioro se basa en una estimación dual de las pérdidas de riesgo de crédito. La metodología apunta a estimar los parámetros de probabilidad de incumplimiento, acorde al plazo del instrumento, y a la pérdida dado el incumplimiento, con tal de constituir provisiones según la pérdida esperada. La ecuación (2.1) representa el cálculo de la pérdida esperada, la cual debe entenderse como la porción de pérdidas probables que debe constituir una provisión para resguardar el riesgo de crédito de los deudores. (Gutiérrez, Osorio y Romero, 2018).

$$PE = EAD \cdot PD \cdot LGD, \quad (2.1)$$

donde:

PE: Pérdida esperada por riesgo de crédito.

EAD: Exposición al incumplimiento o exposición en riesgo.

PD: Probabilidad de que un determinado deudor incumpla sus obligaciones de pago dentro de un período de tiempo determinado.

LGD: Tasa de pérdida incurrida una vez ocurrido el incumplimiento.

2.2.3. Modelo de Scoring de Clientes

En los últimos años, una de las herramientas más utilizadas para la medición y control del riesgo son los modelos de scoring, definidos por Hand y Henley (1997) como “métodos estadísticos utilizados para clasificar a los solicitantes de crédito, o incluso a quienes ya son clientes de la entidad evaluadora, entre las clases de riesgo bueno o malo”.

La utilización de modelos de scoring para la evaluación de riesgo de crédito, esto es, para estimar probabilidades de *default* y ordenar los deudores y solicitantes en función de su riesgo de incumplimiento, comenzó en la década de los años 70, pero se popularizó a partir de los 90, gracias al desarrollo de mejores recursos estadísticos y computacionales y de la creciente necesidad por parte de la industria bancaria de ser más eficaz y eficiente en el estudio de créditos. Esta idea de categorización de riesgo a través de probabilidades de *default* puede extenderse a otras entidades no bancarias que presten servicios a un gran número de clientes y que se vean propensas a incumplimientos por parte de los mismos.

Los modelos de scoring están basados en técnicas estadísticas, matemáticas, econométricas y de inteligencia artificial. Los métodos más usados son: análisis discriminante, regresión lineal, regresión logística, modelos Probit, modelos Logit, árboles de decisión y redes neuronales.

Por lo general, los modelos de scoring le asignan al evaluado un puntaje o *score*, o una calificación, clasificación o rating. Algunos métodos los asignan a grupos, en donde

cada grupo tiene un perfil de riesgo distinto; sin embargo, en la práctica, esto equivale a una calificación. A su vez, estos ordenamientos de los deudores permiten obtener estimaciones más concretas del riesgo; en general se busca obtener alguna estimación de la probabilidad de incumplimiento del deudor (PD, por probabilidad de *default*) asociada a su *score*, *rating* o calificación. Esta estimación se puede obtener directamente del *score* en el caso de los modelos econométricos, o también en función de la tasa de incumplimiento (TD, por tasa de *default*) histórica observada en el grupo de deudores con la misma calificación o puntaje similar (Girault, 2007).

El proceso de construcción de un modelo de scoring se puede resumir, de acuerdo a Wu (2008), en los siguientes pasos:

1. Definición de incumplimiento
2. Definición del horizonte de tiempo
3. Obtención de la muestra
4. Definición de variables explicativas
5. Análisis exploratorio
6. Ajuste del modelo
7. Validación del modelo
8. Obtención del puntaje



Una vez obtenido el modelo de scoring, este puede ser utilizado para definir segmentos de riesgo de incumplimiento, los cuales se definen a partir de la distribución de las probabilidades estimadas, por lo general, se denominan como clases de riesgo bajo, medio y alto.

Capítulo 3

Objetivos

3.1. Objetivo General

Desarrollar un modelo de scoring para clientes residenciales que permita categorizarlos de acuerdo a su riesgo de incumplimiento.

3.2. Objetivos Específicos

- Generar modelos de clasificación binaria para predecir las probabilidades de incumplimiento de los clientes mediante regresión logística y potenciación del gradiente de árboles de decisión.
- Identificar el modelo más robusto en términos de precisión y poder de discriminación.
- Definir categorías de riesgo de incumplimiento en base al modelo seleccionado para apoyar las gestiones de cobranza.

Capítulo 4

Metodología

4.1. Definición de la variable objetivo

El objetivo de este modelo es lograr estimar la probabilidad de que un cliente residencial se vuelva incobrable, este concepto no está formalmente definido dentro de la empresa, ya que no es claro cuándo un cliente ya no pagará su deuda. Sin embargo, el concepto más cercano a la incobrabilidad y la instancia que se quiere evitar es el proceso judicial que se comienza cuando el cliente alcanza los 720 días de atraso en el pago.

Se sabe de análisis anteriores llevados a cabo por la empresa, que menos de un 1 % de los clientes que llegan a esta condición cancelan parte o totalidad de su deuda. Es por esto que, inicialmente se propone como variable objetivo, que el cliente llegue a tener 720 días de atraso o más en un período de tiempo por determinar. Considerando que, la ventana de desempeño que se le debe otorgar a una cuenta para que tenga probabilidades de llegar a 720 debe ser de al menos 24 meses, y teniendo en cuenta el tiempo de descarga de los datos necesarios, se propone realizar un análisis para determinar si una antigüedad más baja es un buen representante de 720, por ejemplo, si un cliente llega a tener 360 días de atraso, ¿qué tan seguro es que llegue a 720?, si la probabilidad es alta, puede no haber gran diferencia en considerar 360 o 720 días de atraso como variable respuesta.

El análisis que se llevó a cabo para responder a la pregunta anterior consistió en revisar distintos períodos de tiempo e identificar el número de clientes que tiene 180, 360, 540 días de atraso en una fecha en particular, luego se analizan 36 meses posterior a eso, para determinar qué porcentaje de ese volumen llega a tener 720 o más días de atraso dentro de dicho período.

En el Cuadro 4.1 se puede ver que, a Enero de 2016, 32.425 clientes llegaron a tener 180 días de atraso, al analizar su evolución en los 36 meses posteriores, se observa que, 6.575 llegaron a tener 720 días de atraso, esto es, un 20.28 %, luego no se puede asegurar que, si un cliente llega a 180, muy probablemente llegará a 720, pues existe casi un 80 % que cancelará parte o la totalidad de su deuda. Incluso, al analizar aquellos clientes que

llegan a tener 540 días, no es posible asegurar que llegarán a 720. En las distintas fechas de referencia que se presentan en los Cuadros 4.2, 4.3 y 4.4 podemos observar que, las conclusiones son similares, ni siquiera el 50 % de los clientes que llega a tener 540 días de atraso alcanza a tener 720, al menos en un período de 3 años.

En vista de lo anterior, se decide definir la variable objetivo como 1, si el cliente llega a tener 720 o más días de atraso en el pago en los próximos 30 meses a partir de la fecha de referencia, y como 0 si no es así.

Tramo	Frecuencia	Llega a 720	%
180 - 720	32.425	6.575	20,28
240 - 720	23.208	5.887	25,37
360 - 720	12.991	4.454	34,29
540 - 720	5.228	2.509	47,99

Cuadro 4.1: Enero 2016

Tramo	Frecuencia	Llega a 720	%
180 - 720	31.879	6.401	20,08
240 - 720	22.779	5.671	24,90
360 - 720	12.878	4.251	33,01
540 - 720	5.028	2.246	44,67

Cuadro 4.2: Abril 2016

Tramo	Frecuencia	Llega a 720	%
180 - 720	30.948	5.956	19,25
240 - 720	22.233	5.305	23,86
360 - 720	12.585	3.953	31,41
540 - 720	4.757	1.862	39,14

Cuadro 4.3: Julio 2016

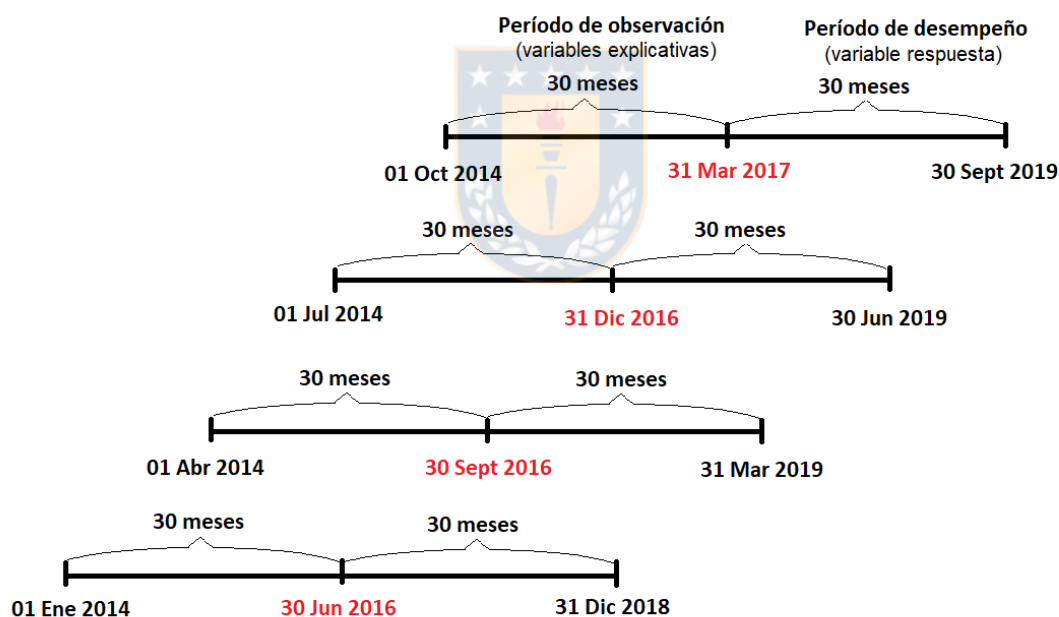
Tramo	Frecuencia	Llega a 720	%
180 - 720	30.337	5.951	19,62
240 - 720	21.827	5.284	24,21
360 - 720	12.329	3.942	31,97
540 - 720	4.839	1.964	40,59

Cuadro 4.4: Octubre 2016

4.2. Definición del período de observación y desempeño

Considerando la definición de la variable respuesta, esto es, que el cliente llegue a tener 720 días de atraso, lo cual equivale a 24 meses, se considera otorgar una ventana de desempeño de 30 meses, para que tenga sentido luego aplicar el modelo sobre aquellos clientes que no tienen atrasos, y que su probabilidad de llegar a 720 no sea nula. No se define una ventana mayor a 30, por el tiempo que implica la descarga de la información necesaria para cubrir tal período de tiempo. Además, la historia que se considera para medir las variables explicativas del modelo corresponde a los 30 meses previos a la fecha de referencia.

En vista de lo anterior y para evitar cualquier tipo de estacionalidad en el comportamiento de los clientes, se consideran cuatro fechas de referencia para desarrollar este modelo, en la Figura 4.1 se pueden observar más claramente los períodos de tiempo a analizar.



Fuente: Elaboración propia.

Figura 4.1: Períodos de observación y desempeño.

4.3. Obtención del Universo

El objetivo en esta etapa es determinar el número exacto de cuentas de clientes residenciales válidas durante el período de observación y desempeño considerado para

desarrollar el presente proyecto, el cual abarca desde Enero de 2014 hasta Septiembre de 2019. Para esto, se llevan a cabo el siguiente procedimiento:

1. Mediante la plataforma SAP, se descargan los conjuntos de datos estáticos mensuales para el período antes mencionado, se obtienen 69 conjuntos de datos, con 23 variables cada uno, de las cuales se mantienen las siguientes:

- Días de atraso en el pago: Corresponde a la antigüedad de la deuda en días.
- Tipo de cliente: Hace referencia al uso que se le da a la casa habitación, ya sea para fines residenciales, comerciales, fiscales o industriales.
- Fecha de creación: Se refiere a la fecha de creación de la cuenta del cliente.
- Motivo del bloqueo: Hace referencia a la situación de la cuenta, si está activa, castigada, eliminada o es ficticia.

2. Se realiza un filtro para mantener solo aquellas cuentas que son Residenciales durante todo el período, ya que algunas cuentas pasan de ser residenciales a comerciales o viceversa. Se obtienen 1.104.740 cuentas de clientes residenciales.

3. Se quieren mantener solo aquellas cuentas cuya fecha de creación es anterior a Enero de 2014, ya que las cuentas nuevas no entregarán información suficiente. De este filtro se obtienen 146.661 cuentas con fechas de creación posterior a Enero 2014, 17.152 cuentas sin fecha de creación y 940.927 cuentas creadas antes del 2014.

4. Se hace un filtro por la variable Motivo de bloqueo, se eliminan aquellas cuentas que son ficticias. En cuanto a las cuentas castigadas y eliminadas, estas se eliminan solo si presentan esta condición previo a la fecha de referencia, ya que si caen en condición de castigo o eliminación en el período de desempeño podrían representar un evento para el modelo. Se obtienen 935.991 cuentas.

El Cuadro 4.5 muestra el número de eventos, esto es, cuentas que presentan 720 o más días de atraso en el pago en las ventanas de desempeño para cada fecha de referencia.


Fecha de referencia	No eventos	eventos
30-06-2016	910.561	25.430
30-09-2016	910.602	25.389
31-12-2016	910.259	25.732
31-03-2017	909.920	26.071

Cuadro 4.5: Número de eventos por fecha de referencia.

Así, se tiene un total de 3.743.964 registros, correspondientes a las 935.991 cuentas observadas en cada fecha de referencia, de las cuales 102.622 llegan a tener 720 o más días de atraso en el pago, esto es un 2.74 %.

4.4. Obtención de la muestra

Para seleccionar a los clientes residenciales que finalmente formarían parte del análisis se utilizó el muestreo aleatorio estratificado, el cual consiste en obtener la muestra mediante la separación de los elementos de la población en grupos que no presenten traslapes, llamados estratos, y la selección posterior de una muestra aleatoria simple en cada estrato. Los individuos de un mismo estrato se comportan de manera similar respecto a la variable de interés, pero son heterogéneos entre ellos. Uno de los objetivos de la estratificación es producir estimadores con varianza pequeña, por lo cual, de acuerdo a Scheaffer, Mendenhall y Ott (2006) el mejor criterio para definir los estratos es el conjunto de valores que la respuesta puede tomar. En este caso la variable respuesta está relacionada con los días de atraso en el pago o antigüedad de la deuda, por lo tanto, los estratos a considerar corresponden a tramos de antigüedad, definidos de la siguiente manera:

- 
- Tramo 1: Entre 0 y 30 días de atraso en la deuda.
 - Tramo 2: Entre 30 y 90 días de atraso en la deuda.
 - Tramo 3: Entre 90 y 720 días de atraso en la deuda.
 - Tramo 4: Más de 720 días de atraso en la deuda.

Para asignar cada cuenta a un único estrato se analizó toda la información disponible, es decir, 69 meses comprendidos entre Enero de 2014 y Septiembre de 2019, a cada cuenta se le asignaron sus frecuencias por tramo, esto es, el número de meses en los cuales sus días de atraso caían dentro de cada uno de ellos, y se le asignó aquel que presentara la mayor frecuencia, en caso de empate, se asignó la cuenta al tramo con menos clientes. Así, los estratos se pueden definir como:

- Estrato 1: Clientes que entre enero 2014 y septiembre 2019 tienen días de atraso mayormente dentro del tramo de 0 a 30 días.
- Estrato 2: Clientes que entre enero 2014 y septiembre 2019 tienen días de atraso mayormente dentro del tramo de 30 a 90 días.
- Estrato 3: Clientes que entre enero 2014 y septiembre 2019 tienen días de atraso mayormente dentro del tramo de 90 a 720 días.

- Estrato 4: Clientes que entre enero 2014 y septiembre 2019 tienen días de atraso mayormente dentro del tramo de 720 o más días.

Finalmente la distribución de cuentas por estrato se presenta en el Cuadro 4.6.

	Frecuencia	Porcentaje (%)
Estrato 1	892.188	95,320
Estrato 2	16.305	1,742
Estrato 3	9.752	1,042
Estrato 4	17.746	1,896
Total	935.991	100,000

Cuadro 4.6: Distribución de cuentas por estrato.

Tamaño de muestra

Uno de los criterios más utilizados para estimar un tamaño de muestra mínimo en regresión logística es que debe haber al menos 10 eventos por variable predictora ($EPV \geq 10$), por ejemplo, si se consideran 100 variables predictoras, la muestra debería presentar al menos 1.000 eventos. Peduzzi, Concato, Kemper, Holford y Feinstein (1996) afinan esta idea, señalando que además se debe tomar en cuenta la proporción de éxitos (p), tal que : $n \cdot p = 10 \cdot k$, donde k es el número de variables y n el tamaño mínimo de la muestra. En este caso se propone $n=320.000$, esto es, 80.000 cuentas evaluadas en las cuatro fechas de referencia, para obtenerlas se realiza un muestreo aleatorio simple proporcional por cada estrato, como se indica en el Cuadro 4.7.

	Frecuencia	Porcentaje (%)
Estrato 1	76.256	95,320
Estrato 2	1.393	1,742
Estrato 3	834	1,042
Estrato 4	1.517	1,896
Total	80.000	100,000

Cuadro 4.7: Número de cuentas a extraer, por estrato.

Una vez obtenidas las 80.000 cuentas únicas, se analiza su comportamiento de pago en las cuatro fechas de referencia para determinar la cantidad de eventos (ver Cuadro 4.8).

Luego, de un total de 320.000 registros existen 8.806 eventos, esto es, un 2.75%. Finalmente, se eliminan aquellos registros que, a fecha de referencia, ya tienen más de 720 días de atraso en el pago, quedando un total de 317.611 cuentas, de los cuales 6.468 corresponden a eventos.

Fecha de referencia	No eventos	Eventos
30-06-2016	77.834	2.166
30-09-2016	77.821	2.179
31-12-2016	77.781	2.219
31-03-2017	77.758	2.242

Cuadro 4.8: Número de eventos en la muestra, por fecha de referencia.

4.5. Obtención y depuración de los datos

En esta etapa se procede a organizar la descarga de los datos necesarios para desarrollar el modelo. La plataforma de SAP utilizada por la empresa, permite descargar tablas de datos que están predefinidas con respecto a las variables que contienen y cuya información es mensual. Los tipos de tablas que se solicitaron fueron las siguientes:

1. **Fica:** Esta tabla de datos esta constituida por 42 variables que contienen información referente a consumo, facturación y pagos realizados por los clientes en un mes en particular, además de otras variables relacionadas a convenios de pago.
2. **Estático:** En esta tabla se encuentran variables estáticas como la región, la dirección, el nombre del cliente, entre otras, además de la variable no estática que representa los días de atraso del cliente en el mes de análisis.
3. **Corte:** Esta tabla está constituida por 6 variables dicotómicas respecto ha si los clientes han recibido o no algún tipo de notificación o si han registrado cortes y/o reposiciones del servicio durante el mes.
4. **ATC:** Este tipo de tabla corresponde a archivos *csv* que indican el número de llamadas por categoría que realiza cada cuenta, estas categorías son: atención deficiente, cobro indebido, consultas comerciales, corte y/o reposición, gestión de cobranza, gestión interna comercial, metros cúbicos facturados, recaudación y cobranza, reparto de documentos, situaciones irregulares y solicitudes comerciales.
5. **Refacturaciones:** Cada uno de los archivos mensuales de este tipo de tabla, contiene variables referentes a montos refacturados y motivos de refacturación.
6. **Convenios desactivados:** Esta tabla corresponde a un archivo *csv* que contiene variables referentes a todos los convenios que han sido desactivados desde Enero 2014 hasta Marzo de 2017.
7. **ODS:** Para obtener la información referente a los resultados de las diversas órdenes de corte generadas se desarrolló una consulta SQL que entrega variables asociadas a cada orden y lo reportado por el contratista, si fue efectiva o imprecendente y por qué razones.

Cabe señalar que, los archivos entregados por SAP se descargan en formato *txt*, ya que la mayoría tiene un número de filas mayor al que puede soportar *Excel*. Luego, se trabajó con archivos *txt* y *csv*, los cuales fueron utilizados directamente en el software R Studio, desde donde se usaron los códigos necesarios para obtener una tabla de datos ordenada y otorgar el formato adecuado a cada variable.

4.6. Definición de variables explicativas

Esta etapa es la que toma más tiempo, ya que consiste en trabajar con cada uno de los conjuntos de datos antes mencionados, darles el formato adecuado para su uso en el software R Studio y proceder a analizar la información que contienen de manera de poder definir las variables explicativas candidatas al modelo.

La mayoría de las variables calculadas se basan en variables que han sido propuestas previamente para otros modelos llevados a cabo en la empresa. Para seleccionar las variables en cuestión, el procedimiento usual consiste en que, a lo largo del proceso se presentan las variables candidatas y se van proponiendo otras nuevas como resultado de reuniones con los miembros de la gerencia de gestión comercial.

Finalmente, de este proceso, se obtienen un total de 520 variables candidatas al modelo, cabe notar que cada variable se calcula para períodos de tiempo de 3, 6, 12, 18, 24 y 30 meses. Por ejemplo, “máximo monto facturado en los últimos 6 meses”.

4.7. Análisis exploratorio

Como primer paso, es de esperar que, ya que una misma variable es calculada para distintos períodos de tiempo, pueda existir una relación entre ellas. Silva y Barroso (2004) consideran que una correlación mayor a 0,8 puede ocasionar problemas de multicolinealidad en el ajuste de un modelo de regresión logística. Basado en esto, se calcula el estadístico correspondiente para medir la relación o asociación entre las variables medidas en distintos horizontes. En su mayoría las variables continuas no se distribuyen normal, por lo cual se utilizó el test de correlación no paramétrico de Spearman. Se eliminaron todas aquellas variables que presentaran una correlación mayor a 0,8, por lo general, las variables elegidas fueron aquellas medidas en períodos de 3, 12 y 30 meses. Luego de llevar a cabo este procedimiento, el total de variables disminuye a 272.

Posteriormente, se procede a categorizar todas las variables para poder realizar una visualización de la proporción de eventos que se presentan en cada una de sus categorías o intervalos. Esto permite tener una idea de la relación de cada variable regresora con la variable respuesta. Para cuantificar dicha relación se utilizó el estadístico D de Somers,

el cual calcula la asociación entre dos variables categóricas ordinales, específicamente su fórmula corregida para el caso en que la variable dependiente es binaria.

4.7.1. D de Somers corregida

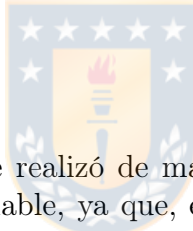
Calcula la asociación entre una variable categórica ordinal x con dos o más categorías y una variable binaria y . Ambas deben tener la misma longitud n , mientras que n_1 es el número de éxitos de la variable binaria. El estadístico viene dado por:

$$D_{xy} = 2 \cdot (c - 0,5),$$

donde,

$$c = \frac{\bar{r} - \frac{n_1 + 1}{2}}{n - n_1},$$

y \bar{r} es el promedio del vector de orden correspondiente a la posición que tendría cada elemento al ser ordenado de menor a mayor, en caso de empate se promedian las posiciones.



4.7.2. Datos atípicos

El tratamiento de datos atípicos se realizó de manera visual y utilizando un criterio acorde a la naturaleza de cada variable, ya que, es válido y común en el negocio que existan variables cuyos valores, si bien tienen una mayor concentración en 0 o valores pequeños también pueden tomar valores muy altos, por lo tanto, métodos establecidos como aquellos definidos a partir del rango intercuartílico resultan ser muy conservadores respecto al rango de valores típicos, lo que puede resultar en la pérdida de información relevante, especialmente en casos en que valores altos de ciertas variables se relacionan directamente con la ocurrencia del evento de interés.

4.7.3. Datos *missing*

Los casos en que el valor de una o más variables no era conocido resultó ser bastante bajo, un 98,72% de la tabla estaba completa, por lo que se decidió prescindir de esta información, siempre y cuando no representaran un evento para el modelo, en cuyo caso se analizaron las variables y las cuentas en el sistema SAP para determinar un valor válido de acuerdo a la situación de cada una de ellas. Finalmente, la tabla quedó compuesta por 313.541 observaciones.

Capítulo 5

Marco teórico

5.1. Técnicas Estadísticas

Para desarrollar el modelo de scoring para clientes residenciales se consideran dos métodos, la regresión logística y la potenciación del gradiente de árboles de decisión.

5.1.1. Regresión Logística

El modelo de regresión logística analiza la relación entre un conjunto de variables independientes y una variable dependiente categórica al estimar la probabilidad de ocurrencia de un evento ajustando los datos a una curva logística (Park, 2013). Específicamente, un modelo de regresión logística binario se utiliza cuando la variable dependiente es dicotómica, mientras que las variables independientes pueden ser continuas o categóricas. En este caso, la variable respuesta es codificada como 0 o 1, representando la ausencia o presencia de una determinada característica.

La *chance* de que ocurra un evento es la razón entre la probabilidad de que ocurra (p) y la probabilidad de que no ocurra ($1-p$). La regresión logística busca estimar esta *chance* mediante un conjunto de variables regresoras \mathbf{x} , como se aprecia en la siguiente ecuación:

$$\frac{p}{1-p} = \beta_0 + \beta \cdot \mathbf{x} \quad (5.1)$$

Sin embargo, este no es un buen modelo, ya que, para grandes valores de \mathbf{x} , la ecuación (5.1) entrega valores que no están contenidos entre 0 y 1.

La solución de la regresión logística es transformar las *chances* usando el logaritmo natural (Peng, Lee e Ingersoll, 2002). Luego, se modela el logaritmo de las *chances* como una función lineal de las variables explicativas.

$$\ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta \cdot \mathbf{x} \quad (5.2)$$

Sea una muestra de n observaciones independientes del vector $(y_i, x_{1i}, x_{2i}, \dots, x_{ki})$, donde y_i denota la variable respuesta dicotómica y $(x_{1i}, x_{2i}, \dots, x_{ki})$ son los valores de las k variables explicativas para el i -ésimo sujeto. La ecuación (5.2) se escribe como:

$$\ln \left(\frac{P(y_i = 1)}{1 - P(y_i = 1)} \right) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} \quad (5.3)$$

La ecuación (5.3) es conocida como el *logit*(y_i), reacomodando términos se obtiene la probabilidad de que y_i presente una determinada característica.

$$P(y_i = 1|x_i) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_{1i} - \dots - \beta_k x_{ki}}}$$

Entre los supuestos destacados por algunos autores como Bewick, Cheek y Ball (2005), además de Peng y So (2002), se tiene que, las observaciones deben ser independientes entre sí, el modelo debe presentar poca o nula multicolinealidad, esto es, las variables independientes no son funciones lineales entre ellas. Además, la relación entre las variables independientes y el logaritmo de las *chances* de un evento debe ser lineal, por último, la regresión logística requiere tamaños de muestra grandes, ya que, los estimadores por máxima verosimilitud son menos robustos que los estimadores por mínimos cuadrados ordinarios usados para estimar los parámetros de un modelo de regresión lineal.

Estimación de los parámetros

La estimación de los parámetros se obtiene mediante máxima verosimilitud (Menard, 2002), antes de obtener los parámetros es necesario definir algunos conceptos.

Sea Y una variable aleatoria con función de distribución dependiente de un único parámetro θ . Se dice que esta distribución pertenece a la familia exponencial si se puede escribir de la forma:

$$f(y; \theta) = e^{y \cdot b(\theta) + c(\theta) + d(y)}$$

Además, se puede demostrar que:

$$E[Y] = -\frac{c'(\theta)}{b'(\theta)} \quad (5.4)$$

y,

$$Var(Y) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{(b'(\theta))^3}$$

Considérese ahora un conjunto de variables aleatorias Y_1, \dots, Y_n independientes provenientes de la familia exponencial. Siendo $E[Y_i] = \mu_i$, se quiere encontrar el estimador por máxima verosimilitud de β , que satisface:

$$g(\mu_i) = x_i^T \beta = \eta_i, \quad (5.5)$$

donde x_i es un vector de elementos x_{ij} con $j = 1, \dots, k$ y $\beta = \beta_0, \dots, \beta_k$. La función de log-verosimilitud para cada Y_i viene dada por:

$$l_i = y_i \cdot b(\theta_i) + c(\theta_i) + d(y_i) \quad (5.6)$$

La función de log-verosimilitud para todas las Y_i corresponde a:

$$l = \sum_{i=1}^N l_i = \sum_{i=1}^N y_i \cdot b(\theta) + \sum_{i=1}^N c(\theta_i) + \sum_{i=1}^N d(y_i)$$

Para estimar los β_j se calcula:

$$\frac{\delta l}{\delta \beta_j} = U_j = \sum_{i=1}^N \frac{\delta l_i}{\delta \beta_j} = \sum_{i=1}^N \frac{\delta l_i}{\delta \theta_i} \cdot \frac{\delta \theta_i}{\delta \mu_i} \cdot \frac{\delta \mu_i}{\delta \beta_j}$$

Derivando desde las ecuaciones (5.4), (5.5), (5.6) y multiplicando, se obtiene la expresión:

$$U_j = \sum_{i=1}^N \frac{(y_i - \mu_i)}{\text{Var}(y_i)} x_{ij} \left(\frac{\delta \mu_i}{\delta \eta_i} \right) \quad (5.7)$$

La matriz de varianzas-covarianzas de U_j es de la forma:

$$J_{jk} = \sum_{i=1}^N \frac{x_{ij} x_{ik}}{\text{Var}(y_i)} \left(\frac{\delta \mu_i}{\delta \eta_i} \right)^2 \quad (5.8)$$

Para el caso de la regresión logística, se tiene que, la variable respuesta y_i sigue una distribución Bernoulli de parámetro p_i . Considerando que esta distribución pertenece a la familia exponencial, se pueden calcular las expresiones para U_j y J_{jk} . Se tiene que $E[y_i] = \mu_i = p_i$, y la función enlace g viene dada por:

$$g(p_i) = \ln \left(\frac{p_i}{1 - p_i} \right) = \eta_i,$$

despejando,

$$p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

Considerando que $Var(y_i) = p_i(1 - p_i)$ y que $\frac{\delta\mu_i}{\delta\eta_i} = Var(y_i)$, a partir de la ecuación (5.7) se tiene:

$$U_j = \sum_{i=1}^N (y_i - p_i) \cdot x_{ij}, \quad (5.9)$$

mientras que, de (5.8),

$$J_{jk} = \sum_{i=1}^N x_{ij} \cdot x_{ik} \cdot p_i(1 - p_i) \quad (5.10)$$

Uno de los métodos de estimación numérica más utilizados es el algoritmo de Newton-Raphson, el cual permite estimar la raíz x_0 de una función t , tal que $t(x_0) = 0$ (Dobson y Barnett, 2008). La pendiente de la función t en el punto x_1 está dada por:

$$t'(x_1) = \frac{t(x_0) - t(x_1)}{x_0 - x_1},$$

si $t(x_0) = 0$, despejando:

$$x_1 = x_0 - \frac{t(x_0)}{t'(x_1)} \quad (5.11)$$

Luego (5.11) entrega la solución para la ecuación $t(x) = 0$ a través de un proceso iterativo hasta encontrar el valor de x que hace converger la función t a 0.

Para la estimación por máxima verosimilitud, utilizando la función U , la ecuación (5.11) se transforma en:

$$b^{(m)} = b^{(m-1)} - \frac{U^{(m-1)}}{U'^{(m-1)}} \quad (5.12)$$

donde $b^{(m)}$ es el vector de estimadores de β_0, \dots, β_k en la m -ésima iteración.

Utilizando la siguiente igualdad:

$$J = Var(U) = -E[U'],$$

y aproximando U' por $E[U']$ se obtiene una nueva expresión para (5.12):

$$b^{(m)} = b^{(m-1)} + \frac{U^{(m-1)}}{J^{(m-1)}},$$

donde los valores que conforman U y J en cada iteración vienen dados por las ecuaciones (5.9) y (5.10) respectivamente, además esta ecuación puede ser expresada matricialmente como:

$$b^{(m)} = b^{(m-1)} + (\mathbf{x}^T A^{(m-1)} \mathbf{x})^{-1} \mathbf{x}^T (\mathbf{y} - \mathbf{p}^{(m-1)})$$

donde A es la matriz diagonal de dimensión $n \times n$, con elementos:

$$a_{ii} = p_i(1 - p_i) \quad i = 1, \dots, n$$

Significación de los estimadores

Test de Wald

Sea la prueba de hipótesis:

$$H_0 : \beta = \beta_0 \quad vs \quad H_1 : \beta \neq \beta_0,$$

se puede probar que (Wasserman, 2013),

$$\frac{\sqrt{n}(\hat{\beta} - \beta_0)}{\hat{se}(\hat{\beta})} \rightarrow N(0, 1),$$

donde $\hat{\beta}$ es el estimador por máxima verosimilitud de β y $\hat{se}(\hat{\beta})$ es el error estándar estimado de $\hat{\beta}$. El test de Wald de tamaño α consiste en rechazar H_0 cuando $|W| > z_{\alpha/2}$, donde,

$$W = \frac{\hat{\beta} - \beta_0}{\hat{se}(\hat{\beta})}$$

Por lo general, $\beta_0 = 0$, se quiere probar si el estimador es significativamente distinto de 0.

Bondad de ajuste

Test de Hosmer-Lemeshow

Este test se utiliza para examinar si las proporciones observadas de un evento son similares a las probabilidades de ocurrencia esperadas en subgrupos de la población. El test se aplica al dividir las probabilidades observadas en G grupos, por lo general, en deciles, y luego calcular un test Chi-cuadrado de Pearson que compara las frecuencias observadas con las esperadas. El valor del estadístico viene dado por la expresión (5.13).

$$H = \sum_{g=1}^G \frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}} \sim \chi_{G-2}^2, \quad (5.13)$$

donde O_{lg} y E_{lg} representan los eventos observados y esperados para $l = 0, 1$ en el g -ésimo grupo, respectivamente. Los grados de libertad corresponden al número de grupos menos 2.

Si bien este es el test más utilizado para probar la bondad de ajuste de un modelo de regresión logística, pierde poder cuando el tamaño de la muestra es demasiado grande. Aunque, el modelo ajustado sea bueno, siempre habrá diferencias con el modelo esperado, y al concentrar un gran número de observaciones en cada grupo estas diferencias comienzan a acumularse. Paul, Pennell y Lemeshow (2013) propusieron una estrategia interesante que consiste en aplicar el test repetidas veces usando 10 grupos sobre subconjuntos aleatorias del total de observaciones. Kramer y Zimmerman (2007) hacen la misma recomendación, sugiriendo subconjuntos de tamaño 5.000. Sin embargo, ninguno de estos autores propone una guía completa de implementación. Bartley (2014) lleva a cabo un estudio de simulación bajo distintos escenarios, simulando observaciones de distintos modelos y ajustando luego un modelo que debería presentar una buena aproximación, para esto consideró tamaños de muestra desde 25.000 a 250.000 y aplicó el test sobre 100 subconjuntos de tamaños 1000, 2000, 5000 y 2% del tamaño de la muestra. De su estudio concluyó que, esta estrategia es válida cuando se tienen más de 100.000 observaciones, se deben tomar 100 subconjuntos de tamaño 5.000 y si, más de 10 de ellos resultan en tests significativos ($p < 0,05$) se debe sospechar de un mal ajuste.

Pseudo R^2 de McFadden

Existen muchas propuestas de cálculo para un R^2 en regresión logística, y ningún consenso sobre cual es el mejor. Los dos métodos más comúnmente reportados en software estadísticos parecen ser el propuesto por McFadden (1974) y el desarrollado por Cox y Snell (1989). Menard (2000), en un estudio en el que comparó diversos índices en el contexto de regresión logística, concluyó que el pseudo R^2 de McFadden era preferible debido a su similitud conceptual con el R^2 de mínimos cuadrados ordinarios y gracias a su independencia relativa con la proporción de eventos en la variable respuesta.

Como se ha indicado, los parámetros de la regresión logística son aquellos que maximizan la verosimilitud de los datos observados. El pseudo R^2 de McFadden se define de acuerdo a la expresión (5.14).

$$R_M^2 = 1 - \frac{\ln(L_M)}{\ln(L_0)}, \quad (5.14)$$

donde L_M denota el valor de la verosimilitud maximizada del modelo estimado y L_0 denota el valor correspondiente al modelo nulo, es decir, el modelo con solo el intercepto y sin covariables. La función de verosimilitud en este contexto toma valores entre 0 y 1, luego la log-verosimilitud es menor a 0. Para un modelo con alta verosimilitud, el valor del logaritmo será muy pequeño y por lo tanto el pseudo R^2 será cercano a 1. Sin embargo, este índice no llega a ser igual a 1. McFadden (1977) afirma que un valor mayor a 0,4 ya representa un buen ajuste.

5.1.2. Potenciación del Gradiente

La potenciación del gradiente o *Gradient boosting* es una técnica de aprendizaje automático que construye modelos de regresión o clasificación aditivos al ajustar de manera secuencial una función parametrizada (aprendiz base o *base learner*) a los “pseudo-residuos” actuales a través de mínimos cuadrados en cada iteración. Los pseudo-residuos son el gradiente de la función de pérdida minimizada con respecto a los valores del modelo en cada dato de entrenamiento, evaluado en la iteración actual.

Sea y la variable respuesta y $\mathbf{x}=(x_1, \dots, x_k)$ el conjunto de variables explicativas, dado una muestra de entrenamiento de tamaño n el objetivo es encontrar una función $F^*(\mathbf{x})$ tal que el valor esperado (sobre la distribución conjunta de (y, \mathbf{x})) de una función de pérdida especificada $\Psi(y, F(\mathbf{x}))$ sea minimizado, esto es:

$$F^*(\mathbf{x}) = \underset{F(x)}{\operatorname{argmin}} E_{y, \mathbf{x}} \Psi(y, F(\mathbf{x}))$$

La función $F^*(\mathbf{x})$ se puede aproximar por una expansión de la forma:

$$F(\mathbf{x}) = \sum_{m=0}^M \gamma_m h(\mathbf{x}; \mathbf{a}_m)$$

donde la función $h(\mathbf{x}, \mathbf{a})$ (aprendiz base) es una función simple de \mathbf{x} de parámetros $\mathbf{a}=\{a_1, a_2, \dots\}$. Los coeficientes de expansión $\{\gamma_0, \dots, \gamma_M\}$ y los parámetros $\{\mathbf{a}_1, \dots, \mathbf{a}_M\}$ se estiman de manera iterativa, tal que, se comienza con una suposición inicial F_0 y luego para $m = 1, 2, \dots, M$:

$$(\gamma_m, \mathbf{a}_m) = \underset{\gamma, \mathbf{a}}{\operatorname{argmin}} \sum_{i=1}^n \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma h(\mathbf{x}_i; \mathbf{a})) \quad (5.15)$$

y,

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h(\mathbf{x}_i; \mathbf{a}_m)$$

La potenciación del gradiente de árboles de decisión especializa este enfoque al caso en que el aprendiz base $h(\mathbf{x}; \mathbf{a})$ es el nodo terminal de un árbol de decisión. Si se tienen L nodos terminales, en cada iteración m , un árbol particiona el espacio de \mathbf{x} en L regiones disjuntas $\{R_{lm}\}_{l=1}^L$ y predice un valor constante en cada una:

$$h(\mathbf{x}_i; \{R_{lm}\}_1^L) = \sum_{l=1}^L r_{lm} I(\mathbf{x} \in R_{lm})$$

donde r_{im} son los pseudo-residuos, de la forma:

$$r_{im} = - \left[\frac{\delta \Psi(y_i, F(\mathbf{x}_i))}{\delta F(\mathbf{x}_i)} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$$

Los parámetros del aprendizaje base son las variables segmentadoras y los puntos de división que definen el árbol, que, en consecuencia, definen la región $\{R_{lm}\}_{l=1}^L$ correspondiente. Con los árboles de decisión la ecuación (5.15) puede ser resuelta de manera independiente en cada región definida por el nodo terminal l del m -ésimo árbol.

$$\gamma_{lm} = \underset{\gamma}{\operatorname{argmin}} \sum_{\mathbf{x}_i \in R_{lm}} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma)$$

$F_m(\mathbf{x})$ se actualiza en cada región:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} I(\mathbf{x} \in R_{lm})$$

El parámetro de contracción $0 < \nu < 1$ controla la tasa de aprendizaje del procedimiento (cuánto aporta a la predicción final cada árbol). Empíricamente Friedman (2001) descubrió que, $\nu \leq 0,1$ lleva a cometer menos errores.

Finalmente, el algoritmo para la potenciación del gradiente de árboles de decisión, propuesto por Friedman (2001), queda definido como:

Algoritmo:

$$1) F_0 = \underset{F(x)}{\operatorname{argmin}} \sum_{i=1}^n \Psi(y_i, F(x))$$

2) Para m de 1 a M :

$$a. \text{ Calcular } r_{im} = - \left[\frac{\delta \Psi(y_i, F(\mathbf{x}_i))}{\delta F(\mathbf{x}_i)} \right]_{F(x)=F_{m-1}(x)} \quad i = 1, \dots, n$$

b. Ajustar un árbol de decisión a los pseudo residuos r_{im} y definir las regiones $\{R_{lm}\}_1^L$.

$$c. \text{ Calcular para cada región } \gamma_{lm} = \underset{\gamma}{\operatorname{argmin}} \sum_{\mathbf{x}_i \in R_{lm}} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma)$$

$$d. \text{ Actualizar } F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \sum_{l=1}^L \gamma_{lm} \cdot I(\mathbf{x} \in R_{lm})$$

Este método fue utilizado por Di Cellio, Forti y Witarsa (2018) para desarrollar un modelo de scoring sobre una base de créditos de Brasil. Al compararlo con un modelo desarrollado mediante regresión logística los resultados fueron levemente mejores en términos de coeficiente de Gini y curva ROC para el modelo de scoring obtenido mediante potenciación del gradiente de árboles de decisión.



Importancia relativa de las variables

Sea I_j la influencia relativa de la variable de entrada x_j , se define, en términos generales, como:

$$I_j = \left(E_{\mathbf{x}} \left[\frac{\delta F(\mathbf{x})}{\delta x_j} \right]^2 \cdot \text{var}_{\mathbf{x}}(x_j) \right)^{1/2} \quad (5.16)$$

Sin embargo, en el caso de árboles de decisión, (5.16) no existe, y debe ser aproximada por una medida subrogante que refleje sus propiedades. Breiman, Friedman, Olshen y Stone (1983) proponen:

$$\hat{I}_j^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 \mathbb{1}(v_t = j), \quad (5.17)$$

donde T denota el árbol con J nodos terminales, $\mathbb{1}$ es la función indicatriz, v_t es la variable divisora asociada al nodo t y \hat{i}_t^2 es la mejora en el error cuadrático que resulta de la división.

Para una colección de árboles $\{T_m\}_1^M$ obtenidos mediante potenciación, (5.17) se puede generalizar como el promedio sobre todos los árboles.

$$\hat{I}_j^2 = \frac{1}{M} \sum_{m=1}^M \hat{I}_j^2(T_m) \quad (5.18)$$

Para los problemas de clasificación en Λ clases, hay Λ funciones de regresión logística $\{F_{\lambda M}(\mathbf{x})\}_{\lambda=1}^{\Lambda}$, cada una descrita por una secuencia de M árboles. En este caso la expresión (5.18) se calcula como:

$$\hat{I}_{j\lambda}^2 = \frac{1}{M} \sum_{m=1}^M \hat{I}_j^2(T_{\lambda m}) \quad (5.19)$$

donde $T_{\lambda m}$ es el árbol inducido por la λ -ésima clase en la iteración m . $\hat{I}_{j\lambda}$ se puede interpretar como la relevancia de la variable predictora x_j al separar la clase λ de la(s) otra(s). Finalmente, la relevancia global de x_j , de acuerdo a Friedman (2001), se obtiene promediando sobre todas las clases:

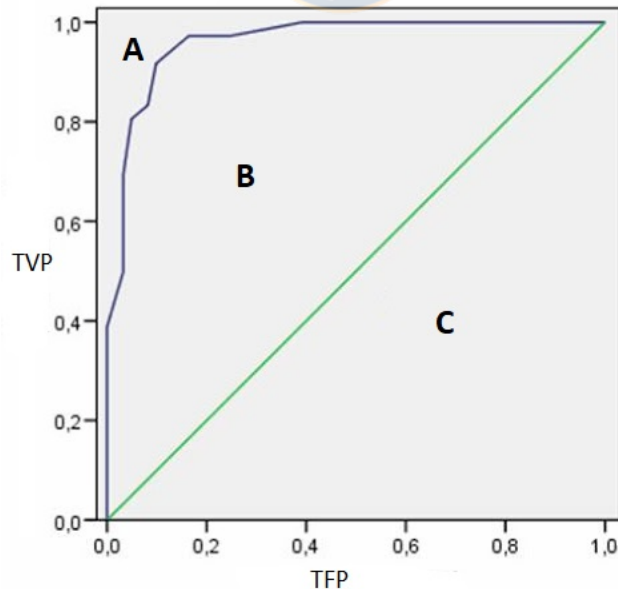
$$\hat{I}_j = \frac{1}{\Lambda} \sum_{\lambda=1}^{\Lambda} \hat{I}_{j\lambda} \quad (5.20)$$

5.2. Criterios de evaluación de modelos

5.2.1. Curva ROC

Un método común para representar el poder de discriminación de un modelo de clasificación es la curva ROC (*receiver operating characteristic*). Esta curva está constituida por los valores de la tasa de falsos positivos en el eje X y la tasa de verdaderos positivos en el eje Y para distintos puntos de corte o umbrales de probabilidad, esto es, aquel valor que determina cuando un sujeto pertenece a una u otra clase. Por lo tanto, esta curva ayuda a decidir para qué valor de corte se obtiene una tasa de verdaderos positivos alta pero también una tasa de falsos positivos baja (Hajian-Tilaki, 2013).

El peor caso es aquel en que el corte es 0 y por lo tanto todos los casos son clasificados como eventos, alcanzando una tasa de verdaderos positivos de 1 y una tasa de falsos positivos también de 1. Cualquier valor en la diagonal indica que la proporción de sujetos clasificados correctamente es igual a la proporción de sujetos clasificados erróneamente. Por otro lado, un modelo ideal sería aquel que discrimine perfectamente los eventos de los no eventos y por lo tanto presenta una tasa de verdaderos positivos igual a 1 y una tasa de falsos positivos igual a 0. En este caso el área bajo la curva de ROC (AUROC) sería igual a 1. Así, mientras más cerca a 1 esté el área bajo la curva mejor será el modelo, mientras que un AUROC de 0,5, es el que resulta de la diagonal, cuando el modelo clasifica aleatoriamente los casos.



Fuente: Elaboración propia.

Figura 5.1: Curva de ROC.

Coefficiente de Gini

El coeficiente o índice de Gini es el área comprendida entre la curva ROC y la línea diagonal, representada por B en la Figura 5.1. Se suele representar como una proporción (Lizares, 2017), tal que:

$$\text{Gini} = \frac{B}{A + B}$$

De la misma manera el AUROC se puede representar como:

$$\text{AUROC} = \frac{B + C}{A + B + C}$$

A modo de guía, en el Cuadro 5.1 se establece una escala de evaluación para un modelo basado en su AUROC.

AUROC	Modelo
0,5 - 0,6	Deficiente
0,6 - 0,7	Débil
0,7 - 0,8	Aceptable
0,8 - 0,9	Bueno
0,9 - 1	Excelente

Cuadro 5.1: Calidad de un modelo de acuerdo al valor de AUROC.

5.2.2. Matriz de confusión

Una vez que se ha determinado el umbral óptimo para distinguir entre las clases, se puede conseguir la matriz de confusión de un modelo sobre un conjunto de datos de prueba. Sea la matriz de confusión:

		Real	
		0	1
Predicho	0	VP	FP
	1	FN	VN

donde, en el contexto de riesgo de incumplimiento:

- VN: Verdaderos negativos, clientes que fueron correctamente clasificados como malos clientes.

- FP: Falsos positivos, clientes malos que fueron erróneamente clasificados como buenos clientes.
- FN: Falsos negativos, clientes buenos que fueron erróneamente clasificados como malos clientes.
- VP: Verdaderos positivos, clientes que fueron correctamente clasificados como buenos clientes.

La tasa de precisión (Pr) del modelo viene dada por:

$$Pr = \frac{VN + VP}{VN + VP + FN + FP} \cdot 100\%.$$

También es importante considerar, especialmente en modelos con datos no balanceados, la especificidad del modelo, esto es, la tasa de eventos predichos correctamente, a partir de la matriz de confusión se obtiene mediante:

$$Especificidad = \frac{VN}{VN + FP} \cdot 100\%.$$

Análogamente se puede obtener la proporción de no eventos predichos correctamente, denominada sensibilidad:

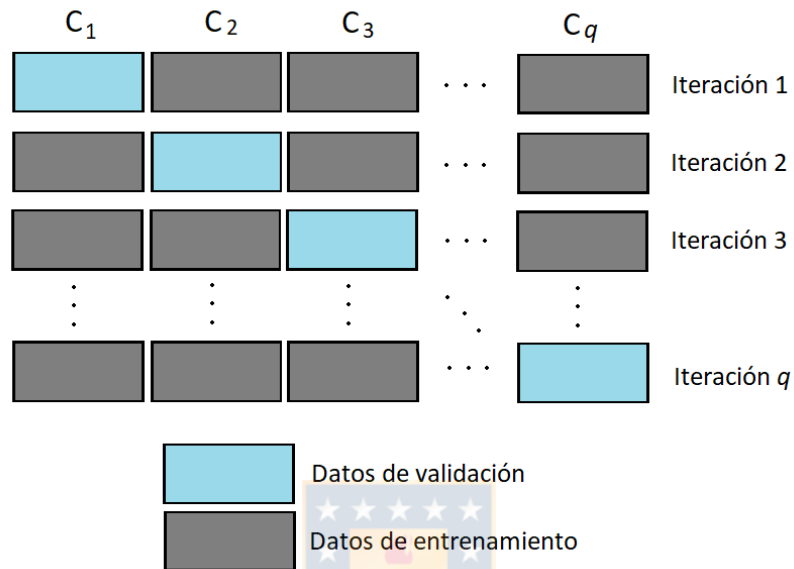
$$Sensibilidad = \frac{VP}{VP + FN} \cdot 100\%.$$

5.2.3. Validación cruzada de q -capas

Cuando se desarrolla un modelo de clasificación, se parte de un conjunto de datos inicial, el cual es separado en dos grupos, un conjunto de entrenamiento y un conjunto de validación o prueba. Una vez realizado el entrenamiento del modelo con el primer conjunto, este se aplica sobre el conjunto de prueba con el fin de evaluar la precisión en la predicción y su error. Sin embargo, dada la naturaleza estocástica del procedimiento, no necesariamente se van a obtener siempre los mismos resultados y, en consecuencia, las mismas precisiones y errores para distintos conjuntos de entrenamiento y prueba. En vista de esto, es recomendable la implementación de la validación cruzada.

La validación cruzada de q -capas consiste en dividir el conjunto de datos original en q grupos de igual tamaño, uno de ellos será utilizado para probar el modelo mientras que los restantes $q-1$ componen el conjunto de entrenamiento. Este proceso se repite q

veces, de manera que cada vez se cambia el conjunto de prueba. Una vez finalizadas las iteraciones, se calcula la precisión y el error para cada uno de los modelos producidos, y para obtener la precisión y el error final se calcula el promedio de los q modelos entrenados.



Fuente: Elaboración propia.

Figura 5.2: Procedimiento validación cruzada.

Capítulo 6

Implementación y Resultados

En este capítulo, se detalla el procedimiento llevado a cabo para obtener los resultados buscados, esto es, un modelo de clasificación para clientes residenciales de Essbio, que sea capaz de discriminar a quienes llegan a tener 720 o más días de atraso en el pago de su cuenta en un plazo de 30 meses a partir de la fecha de referencia.

El ajuste de los modelos se llevó a cabo mediante el software estadístico R Studio. En el Cuadro 6.3 se presentan las características de la muestra y de los conjuntos de entrenamiento y prueba utilizados, en este caso, se ha optado por usar un 80 % de la muestra para entrenar el modelo y un 20 % como datos de validación o prueba.

Conjunto de datos	Eventos	No eventos	Total
Muestra total	6.464	307.077	313.541
Muestra entrenamiento	5.182	245.651	250.833
Muestra prueba	1.282	61.426	62.708

Cuadro 6.1: Composición de los datos.

6.1. Regresión logística

Para ajustar un modelo de regresión logística en R Studio se utilizó la función `glm()`, para lo cual se especificaron los siguientes parámetros:

formula: Expresión de la forma $y \sim (\cdot)$, donde y es la variable dependiente y (\cdot) representa el conjunto de variables explicativas presente en la muestra de entrenamiento.

family: Especifica la distribución de la variable respuesta, en este caso, binomial. Además, se puede especificar la función de enlace, para nuestros efectos, el enlace logístico.

data : Representa el conjunto de datos de entrenamiento para el modelo.

El tiempo de ejecución de esta función al utilizar 272 variables, fue de 11,30 segundos.

Selección de variables

El procedimiento implementado para identificar las variables significativas, se conoce como “selección hacia atrás”, ya que, se parte desde el conjunto total de variables, para ir eliminando aquellas que no resultan significativas de acuerdo al test de Wald. El comando *summary()* nos permite conocer los resultados del ajuste, entregando la siguiente información:

1. Resumen de los residuos de la devianza: Consiste en el reporte de los valores mínimo y máximo, así como de los cuartiles de los residuos de la devianza (d_i), los cuales, en el caso de la regresión logística, se calculan como:

$$d_i = \begin{cases} \sqrt{-2 \cdot \ln(p_i)} & \text{si } y_i = 1 \\ -\sqrt{-2 \cdot \ln(1 - p_i)} & \text{si } y_i = 0 \end{cases}$$

2. Para cada variable explicativa y el intercepto, se obtienen:

- Coeficiente: Corresponde al estimador por máxima verosimilitud de β_k en la ecuación (4.3).
- Error estándar: Indica la variación muestral del coeficiente estimado.
- Valor de z : Hace referencia al valor del estadístico de Wald.
- Valor p: Indica si es posible rechazar la hipótesis nula de que el coeficiente es igual a 0. Rechazamos a un nivel de significancia de 0,05.

3. Devianza nula (D_N) y Devianza residual (D_R):

$$D_N = 2 \cdot (\ln(L_S) - \ln(L_0))$$

$$D_R = 2 \cdot (\ln(L_S) - \ln(L_M)),$$

donde L_S, L_0, L_M , denotan la verosimilitud del modelo saturado, nulo y propuesto, respectivamente. El modelo saturado es el que asume que cada observación tiene su propio parámetro (n), de manera opuesta, el modelo nulo otorga un sólo parámetro para los datos (intercepto), mientras que, el modelo propuesto es el que se evalúa, y tiene $k + 1$ parámetros.

4. AIC: Criterio de información de Akaike, representa una medida de la calidad relativa de un modelo, favoreciendo la bondad de ajuste pero penalizando por la complejidad del mismo, esto es, el número de variables explicativas. Se utiliza para comparar modelos, recomendándose optar por aquel que presente un menor AIC. Se calcula como:

$$AIC = 2 \cdot k - 2 \cdot \ln(L_M)$$

6.1.1. Resultados

Se logró ajustar dos modelos con resultados satisfactorios, los cuales son presentados a la empresa. En este informe se muestran los resultados del modelo que presentó una mayor precisión en su predicción mediante validación cruzada.

En el Cuadro 6.2, se presenta un resumen del modelo seleccionado, el cual está conformado por 11 variables explicativas más el término del intercepto, además se observa el valor de cada coeficiente estimado y su error estándar respectivo. Por último, el valor p para cada variable es menor a 0,01, lo cual nos indica que no existe evidencia suficiente en favor de la hipótesis nula, la cual planteaba que los coeficientes son iguales a 0.

Variable	Coficiente	Error estándar	valor p
Intercepto	-1,9580	0,0790	<0,01
Máximo saldo del mes anterior en los últimos 30 meses. (miles de pesos)	0,0010	0,0001	<0,01
Número de meses en que el importe total baja respecto al mes anterior en los últimos 30 meses	-0,1556	0,0055	<0,01
Ratio del importe total sobre el saldo del mes en los últimos 6 meses.	-0,0276	0,0011	<0,01
Realizó alguna llamada del tipo “Recaudación y cobranza” en los últimos 12 meses	-1,6224	0,3091	<0,01
Tuvo o no más de 38 días de atraso en el último mes	1,5660	0,0605	<0,01
Tuvo un atraso máximo mayor a 62 días en los últimos 12 meses	1,1760	0,0679	<0,01
Mínimo número de pagos consecutivos en los últimos 3 meses	-0,2799	0,0357	<0,01
Ha tenido un convenio activo en los últimos 12 meses	0,6843	0,0851	<0,01
Ha tenido el servicio cortado en alguno de los últimos 12 meses	0,5927	0,0442	<0,01
Ha recibido una carta por extinción de subsidio en los últimos 12 meses	0,5307	0,1077	<0,01
Ha tenido algún convenio del tipo CA o CI en los últimos 30 meses	0,2651	0,0841	<0,01

Cuadro 6.2: Resumen del modelo ajustado mediante regresión logística.

Contribución de cada variable

Se implementa un método de cálculo de la contribución de cada variable planteado por la consultora AIS, el cual está basado en el rango y peso de cada una de ellas. El peso mínimo de una variable corresponde al valor mínimo que toma dentro de la muestra multiplicado por su coeficiente estimado, análogamente se obtiene el peso máximo. Los resultados se observan en el Cuadro 6.3, además se calcula el total de cada peso, siendo -8,7798 el mínimo $\log(odds)$ que puede tener una cuenta y 8,6790 el máximo.

Luego, la contribución de cada variable x_i se calcula como:

$$\text{Contribución}(x_i) = \frac{\text{Peso máximo}(x_i) - \text{Peso mínimo}(x_i)}{8,6790 - (-8,7798)} \cdot 100\% \quad , i = 1, \dots, 11$$

Variable	Peso mínimo	Peso máximo	Contribución
Máximo saldo del mes anterior en los últimos 30 meses (miles de pesos)	-0,1316	3,8643	22,89 %
Número de meses en que el importe total baja respecto al mes anterior en los últimos 30 meses	-3,4244	0	19,61 %
Ratio del importe total sobre el saldo del mes en los últimos 6 meses	-2,7614	0	15,82 %
Realizó alguna llamada del tipo “Recaudación y cobranza” en los últimos 12 meses	-1,6224	0	9,29 %
Tuvo o no más de 38 días de atraso en el último mes	0	1,5660	8,97 %
Tuvo un atraso máximo mayor a 62 días en los últimos 12 meses	0	1,1760	6,74 %
Mínimo número de pagos consecutivos en los últimos 3 meses	-0,8400	0	4,81 %
Ha tenido un convenio activo en los últimos 12 meses	0	0,6843	3,92 %
Ha tenido el servicio cortado en alguno de los últimos 12 meses	0	0,5927	3,39 %
Ha recibido una carta por extinción de subsidio en los últimos 12 meses	0	0,5307	3,04 %
Ha tenido algún convenio del tipo CA o CI en los últimos 30 meses	0	0,6843	1,52 %
Total	-8,7798	8,6790	

Cuadro 6.3: Porcentaje de contribución de cada variable al modelo.

Del Cuadro 6.3 podemos observar que, la variable que tiene una mayor contribución al modelo es aquella que representa el máximo saldo del mes anterior en los últimos 30 meses a partir de la fecha de referencia, la cual se relaciona de manera positiva con el evento de interés, como se ve en el Cuadro 6.2, es decir, aquellas cuentas que presenten saldos altos tendrán una mayor probabilidad de llegar a ser clientes incobrables. La segunda variable con más peso en el modelo se relaciona con el comportamiento de pago de los clientes en los últimos 30 meses, la cual disminuye la probabilidad de llegar a tener 720 días de atraso para cuentas que presentan un comportamiento monótono con pocas fluctuaciones. La tercera variable corresponde a la razón entre el importe total cancelado en los últimos 6 meses, sobre la suma de los respectivos saldos en los últimos 6 meses, multiplicado por 100. Como es de esperar, mientras más cercano a 100 sea su valor, mejor será el comportamiento de pago del cliente, ya que está cancelando gran parte o la totalidad de su saldo mensual, el cual está conformado por la facturación del mes más el saldo del mes anterior, y por ende, es menos probable que llegue a ser incobrable. En cuanto a las llamadas de recaudación y cobranza, estas son realizadas

generalmente por clientes que tienen un monto adeudado y pretenden pagarlo, luego, son clientes que regularizan su situación y no deberían llegar a ser incobrables. En cuanto a los días de atraso, se observó un aumento de eventos cuando la deuda tiene más de 38 días en el último mes, y algo similar ocurre con 62 días de atraso en una ventana de un año, esto se puede verificar en el Anexo de este documento. Por otro lado, cuando el cliente tiene pagos consecutivos es un buen pagador, mientras que la solicitud de convenios de pago se relaciona con un cliente que no ha tenido un buen comportamiento de pago y por ende, ha debido convenir su deuda. En específico, son malos pagadores aquellos clientes acreedores de convenios de tipo carenciado o de instalación. Igualmente, tiene sentido que aquellos clientes que han tenido alguna vez un corte en el servicio sean más propensos a ser incobrables que aquellos que no.

Prueba de multicolinealidad

La prueba de multicolinealidad se lleva a cabo a través de la función $vif()$ de la librería *car*, la cual calcula el factor de inflación de la varianza propuesto por Fox y Monette (1992), de acuerdo a la ecuación (6.1). La raíz cuadrada del VIF nos indica que tanto más grande es el error estándar del coeficiente estimado de una variable en específico respecto al caso en que dicha variable es independiente de las demás.

$$VIF(x_i) = \frac{\det(r_{ii}) \cdot \det(R_{ii})}{\det(R)}, \quad (6.1)$$

donde R es la matriz de correlación de las variables explicativas, de elementos r_{ij} , $i = 1, \dots, 11$, $j = 1, \dots, 11$. Mientras que R_{ii} es la matriz que resulta de eliminar la i -ésima fila y la i -ésima columna.

Una guía general es que un VIF mayor a 5 o 10 es indicador de problemas de multicolinealidad. En el Cuadro 6.4 se observa que ninguna variable tiene VIF mayor a 3, por lo tanto, se puede asumir que, el modelo ajustado no presenta multicolinealidad.

Variable	VIF
Máximo saldo del mes anterior en los últimos 30 meses (miles de pesos)	1,31
Número de meses en que el importe total baja respecto al mes anterior en los últimos 30 meses	1,44
Ratio del importe total sobre el saldo del mes en los últimos 6 meses.	2,33
Realizó alguna llamada del tipo “Recaudación y cobranza” en los últimos 12 meses	1,01
Tuvo o no más de 38 días de atraso en el último mes	2,01
Tuvo un atraso máximo mayor a 62 días en los últimos 12 meses	1,70
Mínimo número de pagos consecutivos en los últimos 3 meses	2,50
Ha tenido un convenio activo en los últimos 12 meses	2,10
Ha tenido el servicio cortado en alguno de los últimos 12 meses	1,14
Ha recibido una carta por extinción de subsidio en los últimos 12 meses	1,02
Ha tenido algún convenio del tipo CA o CI en los últimos 30 meses	2,05

Cuadro 6.4: Valores del factor de inflación de la varianza para las variables explicativas del modelo ajustado.

Prueba de Hosmer Lemeshow

Tal como se especificó en la Sección 4.1.1, el test de Hosmer y Lemeshow no es confiable para muestras demasiado grandes, por lo cual, se optó por seguir el método de implementación sugerido por Bartley (2014), el cual consiste en tomar 100 subconjuntos aleatorios de tamaño 5.000 de la muestra de entrenamiento, y aplicar el test a cada uno de ellos. En un primer intento se obtiene que 80 de los 100 tests son no significativos ($p > 0,05$), sin embargo, dada la naturaleza estocástica del proceso se obtienen resultados diferentes al repetir la ejecución. Al iterar el procedimiento anterior 1.000 veces, en promedio, se obtienen 78 tests no significativos. De acuerdo a Bartley (2014), más de 10 tests significativos pueden ser un mal indicio, sin embargo, es una propuesta conservadora, tal como él lo admite, además en este contexto, el hecho de que los datos sean no balanceados también afecta el valor del estadístico H , ya que, como se verá en la curva de ROC, probabilidades estimadas mayores a 0,1 pueden resultar en eventos, lo cual incrementa la diferencia entre los valores observados y esperados.

Pseudo R^2 de McFadden

El valor del R^2 de McFadden para este modelo fue calculado a través de la función $PseudoR2()$ de la librería *DescTools*, la cual permite calcular diferentes pseudo R^2 propuestos en la literatura. El resultado fue el siguiente:

$$R_M^2 = 0,6125$$

Como se puede deducir de la ecuación (4.14), el pseudo R^2 de McFadden no está acotado por 1, y de acuerdo al autor, un valor mayor a 0,4 es indicador de un buen ajuste.

Curva de ROC

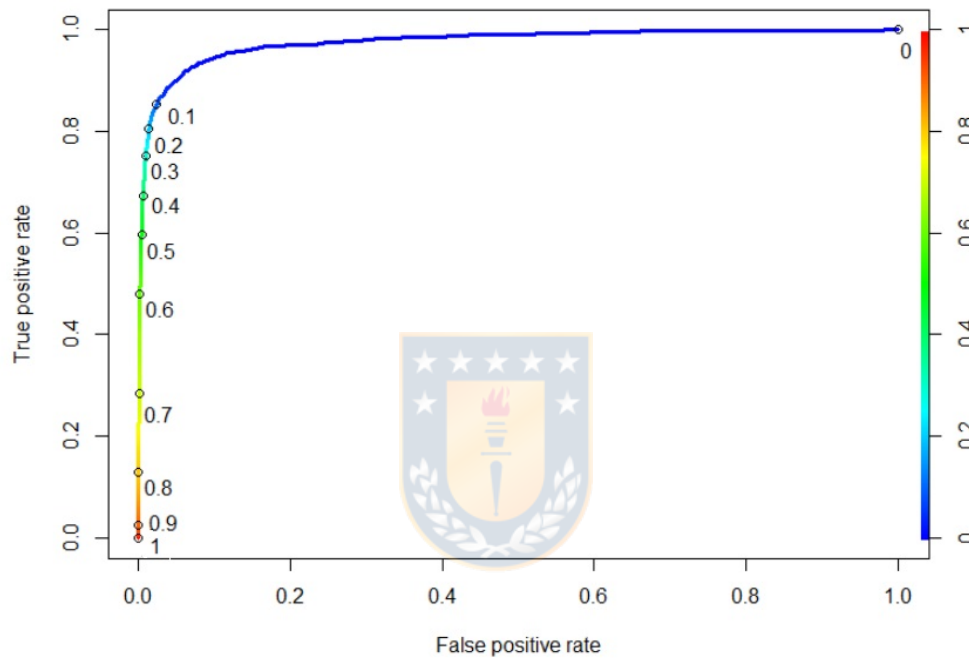


Figura 6.1: Curva de ROC modelo de regresión logística.

En la Figura 6.1, podemos ver la curva de ROC resultante del modelo ajustado, los números sobre la curva representan los umbrales de probabilidad que separan las clases, al igual que los colores. Se observa claramente que el valor que mejor discrimina las clases es el 0,1, ya que para este valor se obtiene una tasa de falsos positivos baja y una tasa de verdaderos positivos superior al 80%. Esto quiere decir que, si un individuo tiene una probabilidad estimada mayor a 0,1 es considerado un evento. El umbral puede parecer bajo, sin embargo, considerando lo raro del evento, tiene sentido que el modelo sea más preciso al identificar los no eventos, asignándole probabilidades muy cercanas a 0, mientras que, en el caso de los eventos, existen algunos que presentan probabilidades más bajas de lo esperado. De todas formas, se debe destacar el hecho de que el modelo es capaz de discriminar las clases con gran precisión. En el Cuadro 6.5, se presentan los valores del AUROC y el coeficiente de Gini, ambos son cercanos a 1, lo cual es indicador

de un buen poder de discriminación.

Índice	Valor
AUROC	0,9758
GINI	0,9516

Cuadro 6.5: Poder de discriminación.

Matriz de Confusión

En el Cuadro 6.6 se presenta la matriz de confusión que se obtuvo a partir de la muestra de prueba. Además, en el Cuadro 6.7 se especifican los porcentajes de precisión, especificidad y sensibilidad definidos en la sección 4.2.2. El modelo predice correctamente un 85,26 % de los clientes que llegan a tener 720 o más días de atraso en el pago, mientras que un 97,62 % de los clientes buenos son clasificados correctamente. De manera global, el modelo tiene un 97,36 % de precisión, de acuerdo a los resultados obtenidos sobre la muestra de prueba.

		Real	
		0	1
Predicho	0	59.962	189
	1	1.464	1.093

Cuadro 6.6: Matriz de confusión muestra de prueba.

Índice	Valor
Precisión	97,36 %
Especificidad	85,26 %
Sensibilidad	97,62 %

Cuadro 6.7: Medidas de precisión sobre la muestra de prueba.

Validación cruzada

Para llevar a cabo la validación cruzada en q -capas se define $q=10$, esto es, el conjunto de datos inicial formado por 313.541 registros es dividido en 10 grupos, manteniendo el porcentaje de eventos en cada uno. En el Cuadro 6.8 se pueden ver los valores de precisión, especificidad y sensibilidad promedio entre los 10 modelos entrenados.

Índice	Valor
Precisión	97,24 %
Especificidad	84,42 %
Sensibilidad	97,51 %

Cuadro 6.8: Medidas de precisión mediante validación cruzada.

Por último se mide, además, el desempeño del modelo sobre la muestra de entrenamiento, obteniendo la matriz de confusión que se observa en el Cuadro 6.9. A partir de ella, se obtienen los índices de precisión presentados en el Cuadro 6.10. Se puede verificar que, si bien los índices decrecen levemente en la validación cruzada, siguen siendo altos y similares, por lo que no se sospecha de un sobreajuste del modelo sobre la muestra de entrenamiento.

		Real	
		0	1
Predicho	0	239.488	818
	1	6.163	4.364

Cuadro 6.9: Matriz de confusión muestra de entrenamiento.

Índice	Valor
Precisión	97,22 %
Especificidad	84,21 %
Sensibilidad	97,49 %

Cuadro 6.10: Medidas de precisión sobre la muestra de entrenamiento.

6.2. Potenciación del gradiente de árboles de decisión

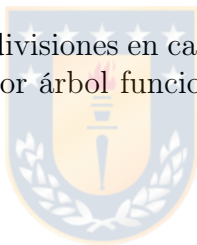
Este algoritmo se encuentra disponible en la librería *gbm*. Contiene dos funciones que permiten aplicar la potenciación del gradiente de árboles de decisión, *gbm()* y *gbm.fit()*. La principal diferencia entre ambas es que la primera utiliza una fórmula para especificar las variables tal como *glm()*, mientras que la segunda requiere un vector *y* y una matriz de variables explicativas *x*. Esta última es más eficiente cuando se trata con muchas variables. Para efectos de este modelo, la función *gbm()* presentó buenos resultados. Esta función tiene muchos parámetros y opciones de aplicación, en este caso, se especifican los siguientes:

distribution: Si es un modelo de regresión, se especifica *gaussian*, mientras que si es un modelo de clasificación, será *bernoulli*, también tiene la opción *multinomial* para modelos de clasificación entre más de una clase.

shrinkage: Se refiere a la tasa de aprendizaje del modelo.

n.trees: Número de árboles a ajustar.

interaction.depth: El número de divisiones en cada árbol, controla la complejidad del modelo, por lo general, una rama por árbol funciona bien, raramente se escoge mayor a 10.



6.2.1. Resultados

Los parámetros por defecto de la función son una tasa de aprendizaje de 0,001 y 100 árboles de decisión, sin embargo, una tasa tan pequeña requiere un mayor número de árboles (Friedman, 2002), en este caso, para probar el total de variables se escoge *shrinkage*=0,01 y *n.trees*=1.000, mientras que las divisiones por árbol se mantienen en 1. El primer modelo, a partir de las 272 variables, tomó 15,612 minutos en correr y su ajuste indica que 29 variables tienen una influencia mayor a 0. Al ir extrayendo las variables con menos influencia, se obtienen las medidas de precisión que se presentan en el Cuadro 6.11. Sin embargo, las variables seleccionadas son muy similares entre sí, por ejemplo, selecciona la variable “Ratio del importe pagado sobre el monto facturado” en todos los períodos de tiempo (3,6,12,18,24 y 30 meses), además de las variables generadas a partir de los días de atraso. Considerando que es importante para la empresa que exista diversidad entre las variables, se extraen del conjunto de entrenamiento aquellas variables que son muy similares a otras, y se vuelve a correr el modelo, esta vez con 231 variables, de las cuales, 28 tienen una influencia mayor a 0. El modelo demoró 15,168 minutos en correr, tiempo similar al primer modelo que tenía más variables, probablemente al quitar algunas de las variables que resultaban influyentes, el algoritmo tuvo más trabajo para converger. Aún así, se logra obtener una especificidad mayor, para

un modelo constituido por 14 variables explicativas. Una vez aceptadas las variables que forman el modelo, se prueba el desempeño del algoritmo al cambiar los valores de la tasa de aprendizaje y el número de árboles, los resultados se observan en el Cuadro 6.12.

N° variables	Tiempo (s)	Precisión	Especificidad	Sensibilidad
29	3,27	98,77 %	80,19 %	99,17 %
24	3,01	98,78 %	80,11 %	99,17 %
19	2,67	98,77 %	80,11 %	99,16 %
11	1,97	98,67 %	80,27 %	99,06 %

Cuadro 6.11: Medidas de precisión al sacar variables.

Shrinkage	N° árboles	Tiempo	Precisión (%)	Especificidad (%)	Sensibilidad (%)	Variables influyentes
0,01	100	12,50 s	98,63	70,83	99,22	3
0,01	1.000	1,94 m	98,55	78,24	98,98	13
0,01	5.000	9,63 m	98,35	84,63	98,64	14
0,05	100	12,52 s	98,72	75,98	99,19	8
0,05	1.000	1,93 m	98,33	84,87	98,61	14
0,05	5.000	9,58 m	98,30	85,41	98,57	14
0,10	100	12,50 s	98,63	77,61	99,07	12
0,10	1.000	1,95 m	98,31	85,49	98,58	14
0,10	5.000	9,57 m	98,29	85,80	98,55	14

Cuadro 6.12: Medidas de precisión al variar la tasa de aprendizaje y el número de árboles.

El algoritmo es capaz de mantener su precisión utilizando variables que en un primer momento no habían sido seleccionadas, pero que ahora se presentan como una buena opción. Del Cuadro 6.12 se puede deducir que, al aumentar el número de árboles, disminuye la sensibilidad y por ende, la precisión, sin embargo, hay un notable aumento de la especificidad. La misma situación se evidencia al aumentar la tasa de aprendizaje y mantener el número de árboles. De acuerdo a lo anterior, tiene sentido que los mejores resultados se presenten para las tasas de aprendizaje 0,05 y 0,1, con un número de árboles de 1.000 a 5.000. Para estas combinaciones, se procede a variar el número de divisiones por árbol, que hasta ahora era igual a 1. Veremos qué ocurre al aumentar este valor a 3 y 5. En el Cuadro 6.13 se observan los resultados.

Shrinkage	N° árboles	N° divisiones	Tiempo	Precisión (%)	Especificidad (%)	Sensibilidad (%)
0,05	1.000	3	4,49 m	98,29	85,96	98,54
0,05	1.000	5	6,60 m	98,32	85,41	98,59
0,05	5.000	3	22,09 m	98,37	84,79	98,62
0,05	5.000	5	32,70 m	98,46	83,70	98,77
0,10	1.000	3	4,46 m	98,27	85,73	98,53
0,10	1.000	5	6,64 m	98,36	84,63	98,64
0,10	5.000	3	22,19 m	98,42	83,78	98,73
0,10	5.000	5	34,04 m	98,43	81,75	98,78

Cuadro 6.13: Medidas de precisión al variar el número de divisiones por árbol.

Se observa a raíz del Cuadro 6.13 que, para ambos valores de tasa de aprendizaje, cuando el número de árboles es 1.000, al aumentar el número de divisiones de 1 a 3 la especificidad aumenta, mientras que, para el caso en que se usan 5.000 árboles, el aumentar el número de divisiones hace que la especificidad disminuya. Por otro lado, tanto para 1.000 como para 5.000 árboles, el aumentar las divisiones de 3 a 5, empeora la especificidad del modelo. Luego, se puede concluir que, 3 divisiones es lo óptimo en este caso, mientras que el número de árboles y la tasa de aprendizaje que presentaron mejores resultados, en términos de especificidad, y a un costo de tiempo más bajo, fueron 1.000 árboles con una tasa de aprendizaje igual a 0,05. Así, estos fueron los parámetros seleccionados para aplicar el modelo, ya que interesa reconocer la mayor cantidad de eventos posibles.

En cuanto a las variables seleccionadas por el modelo, estas se presentan en el Cuadro 6.14, junto con sus respectivas influencias relativas sobre el algoritmo. Podemos ver que, existen algunas similitudes en las variables seleccionadas con respecto al modelo obtenido mediante regresión logística, por ejemplo, el máximo entre los días de atraso de los últimos 12 meses, la cual se presentó como dicotómica en el primer modelo, además de la variable que hace alusión a si el servicio ha sido cortado en los últimos 12 meses. También comparten variables relativas a la razón entre el importe pagado y el saldo del mes, pero para distintos períodos de tiempo. Además, ambos modelos seleccionaron una variable relacionada con la tenencia de algún convenio del tipo carenciado o de instalación. De todas formas, este modelo presenta mayor redundancia en sus variables. En el Anexo de este documento se presentan gráficos que permiten ver la relación entre el evento de interés y las variables seleccionadas por este método de potenciación del gradiente en árboles de decisión.

Variable	Influencia relativa
Máximo días de atraso en el pago en los últimos 12 meses	34,57 %
Ratio del importe total pagado sobre la facturación de los últimos 6 meses	27,10 %
Tuvo más de 38 días de atraso en el último mes	20,36 %
Ratio del importe total pagado sobre el saldo del mes de los últimos 3 meses	6,80 %
Ratio del importe total pagado sobre la facturación de los últimos 30 meses	4,49 %
Días de atraso en el pago promedio en los últimos 30 meses	2,45 %
Ha tenido un convenio CA o CI en los últimos 3 meses	0,80 %
Saldo del mes anterior del último mes	0,68 %
Ratio de la facturación sobre el saldo del mes de los últimos 12 meses	0,63 %
Saldo del mes anterior promedio en los últimos 24 meses	0,60 %
Ratio del saldo del mes anterior sobre el saldo del mes en los últimos 30 meses	0,54 %
Máximo monto convenido de un convenio desactivado en los últimos 12 meses.	0,53 %
Ha tenido el servicio cortado en los últimos 12 meses	0,29 %
Ha tenido una orden de corte improcedente por ausencia del cliente los últimos 6 meses	0,15 %

Cuadro 6.14: Resumen del modelo ajustado por potenciación del gradiente.

Curva de ROC

A simple vista, la curva de ROC para este modelo, ilustrada en la Figura 6.2, es bastante similar a la obtenida para el modelo logístico en la Figura 6.1, sin embargo, cabe notar que la distribución de los puntos de corte de probabilidad sobre la curva difieren. A partir de un umbral de 0,6, el modelo ya es capaz de discriminar sobre el 60 % de los eventos, lo cual en el primer modelo ocurre para un umbral de 0,4. Incluso, la distancia entre 0,9 y 1 es muy superior, lo que indica que a una cantidad considerable de eventos se le ha asignado una probabilidad estimada de incumplimiento mayor o igual a 0,9. Para el primer modelo se había destacado el hecho de que al ser raro el evento, el algoritmo aproxima mejor a los no eventos, asignándoles probabilidades muy cercanas a 0. En este caso, el modelo es capaz de identificar mejor los eventos y asignarles probabilidades más cercanas a 1, de todas formas, el umbral óptimo para discriminar las clases continúa siendo igual a 0,1. En el Cuadro 6.15, se presentan los valores del área bajo la curva de ROC así como el índice de Gini, ambos son cercanos a 1, por lo que el modelo tiene una buena capacidad de discriminación, además, son levemente superiores a lo obtenido para el modelo de regresión logística.

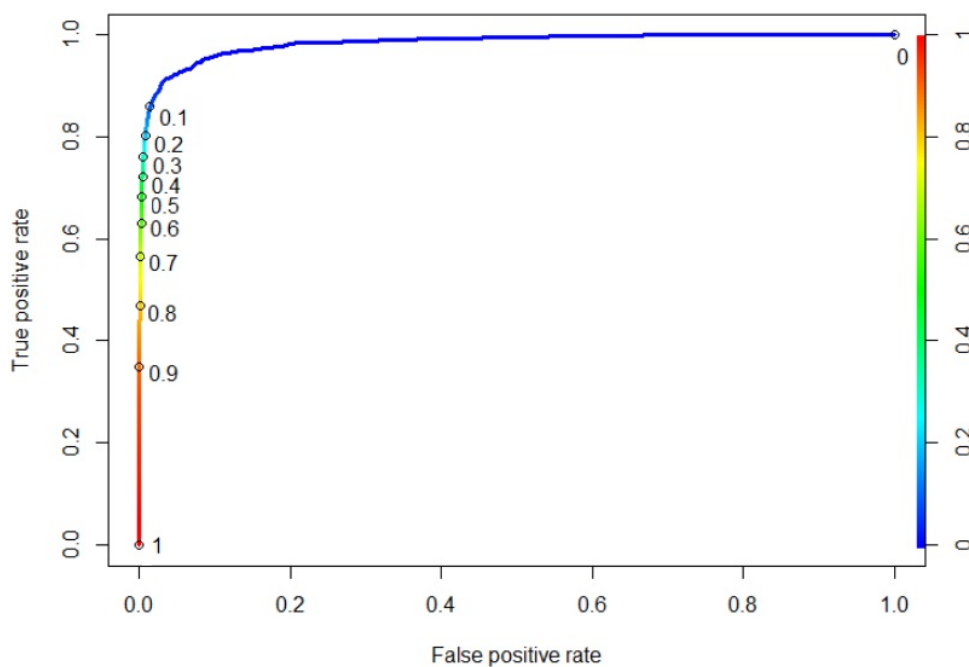


Figura 6.2: Curva ROC modelo potenciación del gradiente.

Índice	Valor
AUROC	0,9826
GINI	0,9652

Cuadro 6.15: Poder de discriminación.

Matriz de confusión

En el Cuadro 6.16, se presenta la matriz de confusión que se obtuvo para la muestra de prueba, cuyos porcentajes de precisión se pueden encontrar en el Cuadro 6.13.

		Real	
		0	1
Predicho	0	60.531	180
	1	895	1.102

Cuadro 6.16: Matriz de confusión muestra de prueba.

También se presenta el desempeño del modelo sobre la muestra de entrenamiento, cuya matriz de confusión se observa en el Cuadro 6.17. Además, se puede ver, a partir del

Cuadro 6.18, que la precisión baja levemente, pero hay un aumento en el porcentaje de eventos que son correctamente clasificados.

		Real	
		0	1
Predicho	0	241.977	659
	1	3.674	4.523

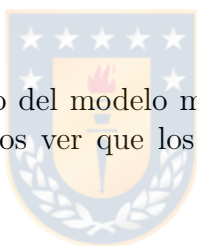
Cuadro 6.17: Matriz de confusión muestra de entrenamiento.

Índice	Valor
Precisión	98,27 %
Especificidad	87,28 %
Sensibilidad	98,50 %

Cuadro 6.18: Medidas de precisión sobre la muestra de entrenamiento.

Validación cruzada

Por último, se prueba el desempeño del modelo mediante la validación cruzada en 10 grupos. En el Cuadro 6.19, podemos ver que los resultados son consistentes con los obtenidos en la muestra de prueba.



Índice	Valor
Precisión	98,21 %
Especificidad	86,20 %
Sensibilidad	98,46 %

Cuadro 6.19: Medidas de precisión mediante validación cruzada.

En este capítulo, se han presentado dos propuestas de modelos de clasificación entrenados a partir de dos técnicas de aprendizaje diferentes, una convencional y otra más innovadora. A continuación, se analizan las ventajas y desventajas de cada una de estas técnicas en el contexto de aplicación a clientes residenciales de las sanitarias Essbio y Nuevosur. Las ventajas del modelo de regresión logística son, en primer lugar, su popularidad y amplia aplicación en el contexto de riesgo crediticio, lo que conlleva un mayor conocimiento sobre su desarrollo. Además, esta metodología permite inferir acerca del impacto de cada variable seleccionada sobre la probabilidad de incumplimiento, lo cual es de gran importancia para la empresa. En cuanto a la potenciación del gradiente para árboles de decisión, este modelo es capaz de alcanzar una mayor precisión sin disminuir la especificidad, presentando un mejor desempeño que el modelo logístico en términos

netamente predictivos, sin embargo, el cálculo de la probabilidad es más complejo y difícil de replicar, además conlleva un mayor esfuerzo computacional y tiempo de ejecución. Considerando esto, el modelo seleccionado es el obtenido mediante regresión logística, a partir del cual se procederá a definir los segmentos de riesgo y calcular el puntaje o *score*.

6.3. Segmentación del riesgo

Para definir los tramos de riesgo, se analiza la distribución de eventos sobre intervalos de probabilidad estimada a partir de la muestra de entrenamiento. Una vez obtenida la curva, se aplica una función que permite identificar los puntos de quiebre o cambios estructurales a lo largo de ella, permitiendo así identificar a partir de qué probabilidad comienzan a aumentar los eventos.

La función utilizada es *breakpoints()* de la librería *strucchange*. Sea el modelo de regresión lineal de y sobre el conjunto de variables explicativas \mathbf{x} :

$$y = \mathbf{x} \cdot \beta + \epsilon, \quad (6.2)$$

donde ϵ es el término del error y β el vector de parámetros de la regresión. Se asume que existen m puntos de quiebre a lo largo de la ecuación, es decir, los coeficientes cambian de una recta a otra. Luego, hay $m + 1$ segmentos en los cuales los coeficientes son constantes, y el modelo se puede escribir como:

$$y = \mathbf{x} \cdot \beta_j + \epsilon, \quad j = 1, \dots, m + 1 \quad (6.3)$$

La función *breakpoints()* estima los puntos de quiebre al minimizar la suma de cuadrados de los residuos de la ecuación (6.3). El algoritmo implementado fue desarrollado por Bai y Perron (2003). La función permite especificar el número de quiebres que se quiere encontrar.

En la Figura 6.3, se observa como se distribuyen los eventos entre intervalos de probabilidad con frecuencias similares. El número de intervalos fue determinado de manera que las barras de frecuencias fuesen lo más parecidas posibles. Se puede ver como el porcentaje de eventos se comienza a diferenciar de 0 para una probabilidad de 0,005 aproximadamente, pero continúa siendo muy bajo. Como es de esperar la mayor proporción de eventos se encuentra en el último intervalo, en este caso, desde 0,317 a 1. El primer quiebre estructural es indicado para el punto 18, que corta en una probabilidad de 0,000829. Sin embargo, los intervalos posteriores siguen presentando tasas de eventos bajas. Para poder particionar de mejor manera los segmentos de riesgo, se consideran las probabilidades estimadas entre 0,000829 y 1, esto da origen al gráfico que se presenta en la Figura 6.4, conformado por 25 intervalos de probabilidad.

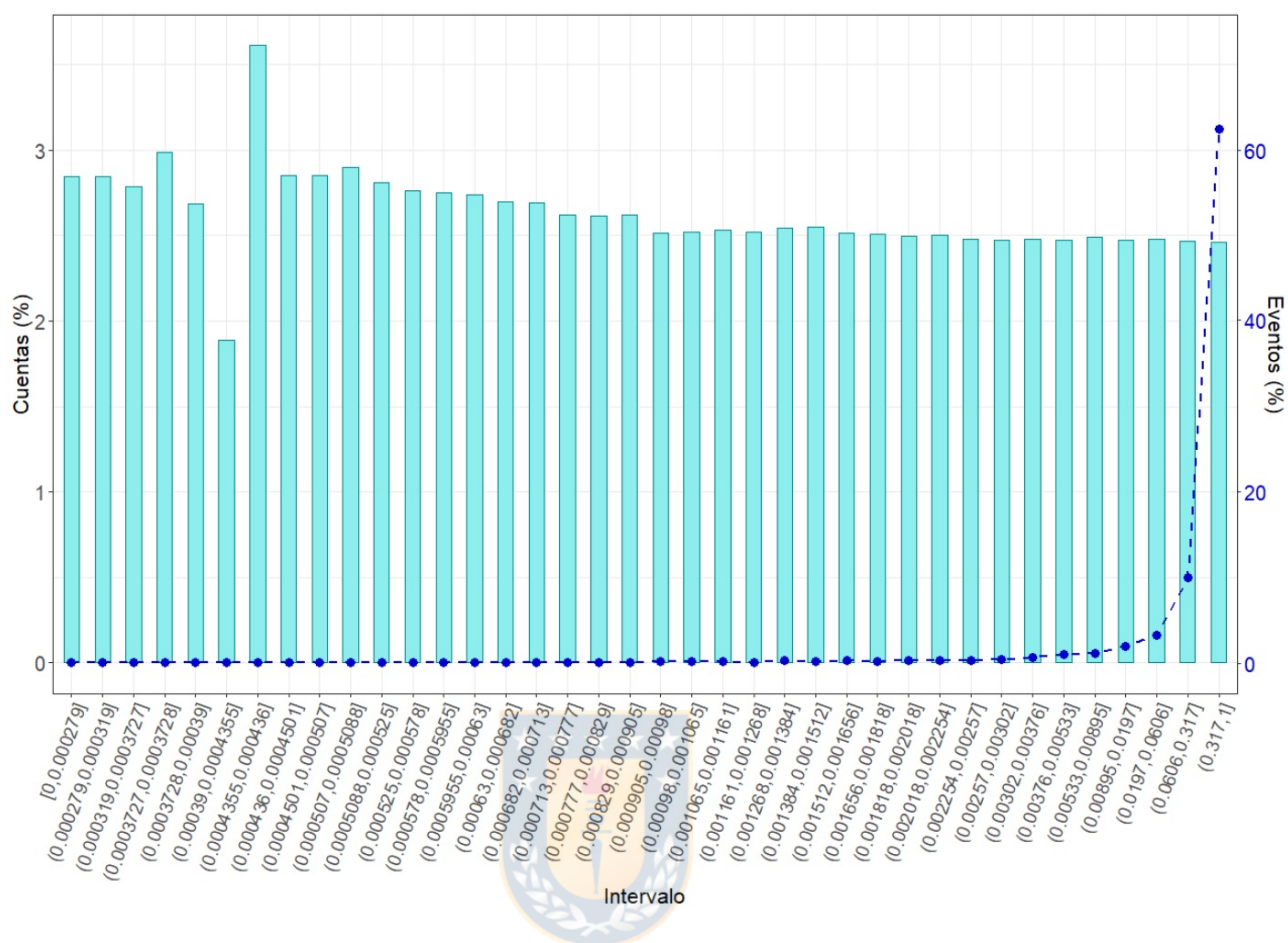


Figura 6.3: Distribución de eventos por intervalo de probabilidad estimada.

Los puntos de quiebre obtenidos corresponden a las posiciones 13, 16, 19 y 22. Para no subestimar la categoría de riesgo medio, se propone seleccionar los puntos 13 y 22, dando origen a la definición de segmentos de riesgo presentada en el Cuadro 6.20. Se observa que, sólo un 6,01 % de los clientes es clasificado dentro del intervalo de riesgo alto, lo cual tiene sentido considerando que, el porcentaje de incobrables en la población era de un 2,74 %. Por otro lado, cerca de un 76 % de los clientes es muy poco probable que llegue a convertirse en incobrable en los próximos 30 meses.

Riesgo	Intervalo	Frecuencia	Proporción	Incobrables	Proporción
Bajo	[0 - 0,00208]	190.532	75,96 %	157	0,082 %
Medio	(0,00208 - 0,0331]	45.226	18,03 %	442	0,977 %
Alto	(0,0331 - 1]	15.076	6,01 %	4.583	30,399 %

Cuadro 6.20: Segmentos de riesgo.

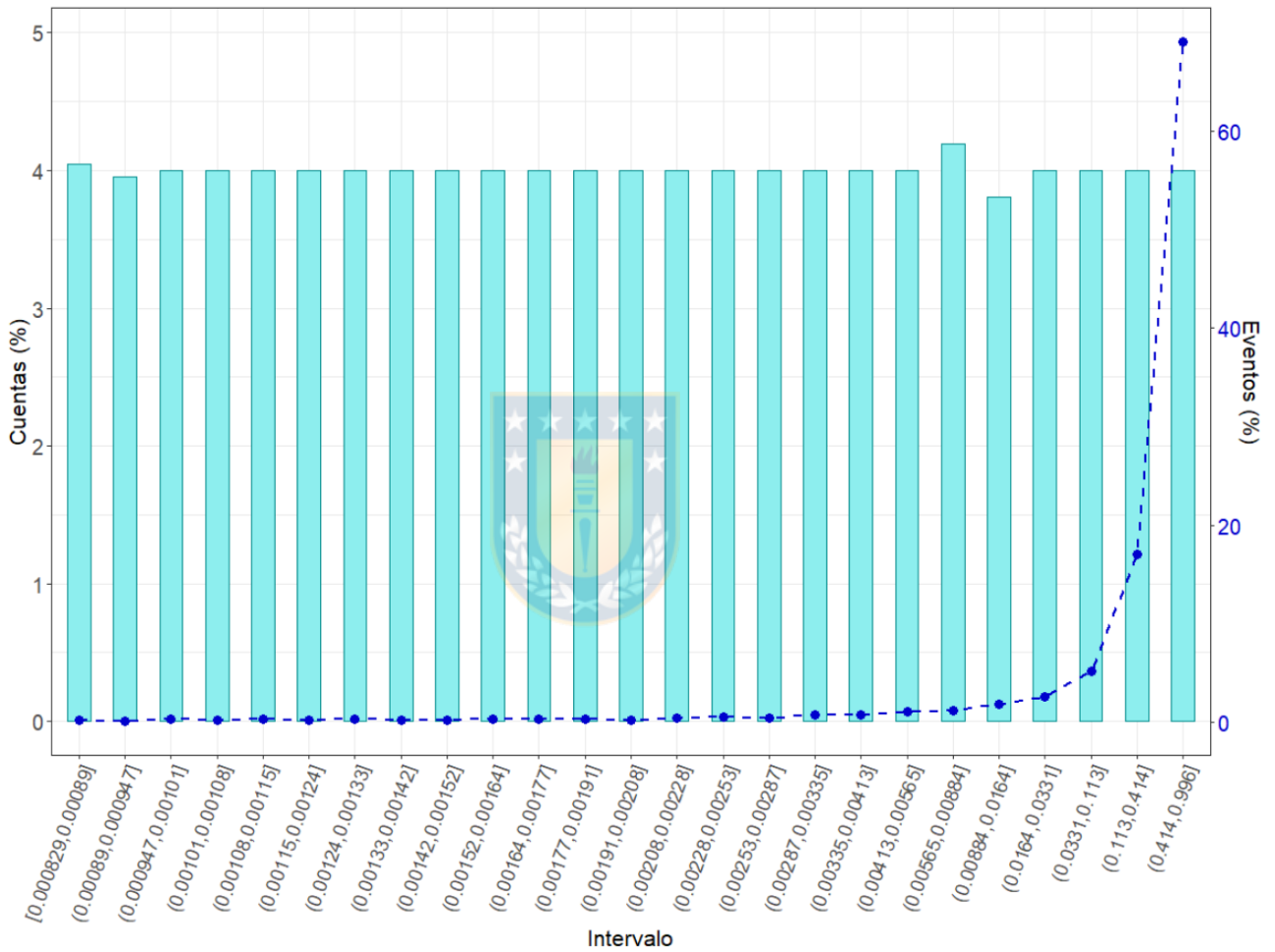


Figura 6.4: Eventos por intervalo de probabilidad estimada mayor a 0,000829.

6.4. Cálculo del puntaje

El *score* se puede calcular de distintas maneras, una de ellas es mediante una transformación lineal del logaritmo de las *chances* tal como propone Siddiqi (2012), quien plantea una relación directa entre este y el *score*, como se indica en la expresión (6.4).

$$score = a + b \cdot \ln \left(\frac{\hat{p}}{1 - \hat{p}} \right) \quad (6.4)$$

Para encontrar la magnitud de a y b , basta con determinar dos pares de valores para el logaritmo de las *chances* o \hat{p} y el puntaje, de manera de resolver un sistema de ecuaciones. La idea es que el puntaje se encuentre entre 0 y 1.000, otorgando mayor puntaje a los buenos clientes, además, sus valores deben ser acordes con la distribución del riesgo. Sabemos que, por ejemplo, clientes con probabilidad estimada mayor a 0,0331 son muy riesgosos, por lo que se les debe asignar un puntaje bajo. De acuerdo a lo anterior, se asignará un puntaje de 0 a aquellos clientes que alcancen una probabilidad de 0,9 y un puntaje de 1.000 a aquellos que tengan una probabilidad estimada igual a 0,00005. Al resolver el sistema de ecuaciones se obtiene:

$$b = -82,640, \quad a = 181,579$$

Luego, el puntaje se calcula de acuerdo a la expresión (6.5).

$$score = 181,579 - 82,640 \cdot \ln \left(\frac{\hat{p}}{1 - \hat{p}} \right), \quad (6.5)$$

$$score = 0 \quad \text{si } score < 0,$$

$$score = 1.000 \quad \text{si } score > 1.000$$

Así, la distribución del puntaje de acuerdo a los segmentos de riesgo equivale a los intervalos especificados en el Cuadro 6.21.

Riesgo	Puntaje
Bajo	692 - 1.000
Medio	460 - 692
Alto	0 - 460

Cuadro 6.21: Segmentos de riesgo.

Capítulo 7

Consideraciones finales

7.1. Discusión

El propósito de este proyecto fue desarrollar un modelo de clasificación binaria para predecir la incobrabilidad de clientes residenciales de Essbio y Nuevosur, otorgarles un puntaje y asignarles una categoría de riesgo a través de él. Para esto, se planteó utilizar técnicas de aprendizaje supervisado, como son la regresión logística y la potenciación del gradiente en árboles de decisión.

La mayor dificultad de este modelo reside en la naturaleza no balanceada de los datos, ya que, el evento de interés es poco común dentro de la población, apenas un 2,74 % de los clientes de la empresa cumple con la condición de tener 720 o más días de atraso en el pago de su deuda. De acuerdo a King y Zeng (2001), esto puede hacer que el modelo de regresión logística subestime la probabilidad de ocurrencia del evento, a raíz de lo cual, proponen una corrección al modelo logístico usual para este tipo de situaciones. Dicha corrección se aplica sobre los coeficientes estimados por máxima verosimilitud, restando a cada uno de ellos su sesgo estimado a causa de un tamaño de muestra pequeño o la ocurrencia de eventos raros. Además, se aplica un factor de corrección sobre el intercepto, basado en la verdadera proporción de eventos sobre la población.

Por otro lado, otros autores, como Allison (2013), afirman que el problema no es la proporción de eventos sino el número de observaciones, “si tienes un tamaño muestral de 1.000, pero sólo 20 eventos, tienes un problema. Si tienes un tamaño muestral de 10.000 con 200 eventos, puedes estar bien. Si tu muestra tiene 100.000 casos con 2.000 eventos, estas de maravilla”.

El algoritmo planteado por King y Zeng (2001) está disponible en el software R Studio, mediante la función *zelig()*, al aplicarla en este caso, los resultados son muy similares a los obtenidos mediante *glm()*, los estimadores varían apenas en el tercer o cuarto decimal, y al hacer predicciones sobre la muestra de prueba, las matrices de confusión para ambos casos, resultan idénticas. Luego, se puede deducir que el sesgo de los coeficientes

era muy pequeño, y que el problema abordado por los autores es más común en casos de muestras con menor tamaño al utilizado en este problema.

De todas formas, como se puede deducir de los resultados presentados, es cierto que, el modelo de regresión logística tiende a subestimar la probabilidad de ocurrencia de los eventos, un 60 % obtuvo una probabilidad estimada mayor a 0,5. Es por esto, que la segmentación del riesgo juega un papel importante, ya que, cada categoría de riesgo será la que indique a partir de qué probabilidad el cliente debe ser considerado como altamente propenso a llegar a ser incobrable. Por otro lado, en este sentido, el modelo de potenciación del gradiente de árboles de decisión, presenta mejores resultados, ya que, un 68 % de los eventos tiene una probabilidad estimada mayor a 0,5.

Cabe destacar que, además de los modelos presentados, se intentó aplicar un modelo de redes neuronales, con 10 capas ocultas y una función de pérdida logística. Sin embargo, la capacidad computacional no fue suficiente para ejecutar el modelo sobre la muestra total, por lo que se aplicó sobre un subconjunto de ella, de tamaño 150.000, el cual tomó 5,16 horas en concluir. Finalmente, para un umbral de probabilidad estimada de 0,1, se obtiene apenas un 61 % de especificidad, por lo cual se descarta este modelo. Se cree que una de las razones por las cuales el desempeño de este modelo no es óptimo es por la baja proporción de eventos, por lo cual sería aconsejable modificar la arquitectura de la red y aplicar alguna técnica que permita mejorar sus resultados, como puede ser el remuestreo. Sin embargo, considerando que los objetivos planteados por la empresa ya fueron alcanzados, no se ahonda en esta línea de investigación, pero resulta interesante proponerlo como un trabajo futuro.

En vista de lo anterior, se recomienda el siguiente procedimiento para mejorar el desempeño de algoritmos de clasificación bajo el escenario de datos no balanceados:

- Hacer un submuestreo de la clase dominante, esto es, seleccionar subconjuntos al azar, de tamaño similar o igual al número de eventos disponibles.
- Entrenar distintos modelos utilizando los eventos y cada uno de los subconjuntos obtenidos del ítem anterior.
- Aplicar un método de consenso apropiado, como el promedio u otro, a partir de las probabilidades estimadas por cada modelo.
- Evaluar el desempeño del modelo y comparar con los resultados obtenidos a partir de la muestra no balanceada.

7.2. Conclusión

Respecto al desempeño de las técnicas utilizadas en este proyecto, cabe señalar que, ambas presentaron resultados satisfactorios, en términos de precisión y poder de discriminación. Por un lado, la regresión logística, seleccionó variables de interés para la empresa y permite una interpretación más directa respecto a la contribución de cada una de ellas sobre la probabilidad de ocurrencia del evento de incobrabilidad. Por otro lado, el modelo obtenido mediante la potenciación del gradiente de árboles de decisión fue capaz de igualar la especificidad obtenida mediante el modelo logístico, pero manteniendo la sensibilidad, es decir, disminuyendo la tasa de falsos negativos. Además, cabe recordar que, este último modelo fue bastante flexible en la elección de las variables, ya que, al extraer de la muestra aquellas que inicialmente resultaron significativas pero que eran demasiado similares a otras, fue capaz de identificar nuevas variables como influyentes, sin perjudicar su precisión en la predicción.

Cabe recordar también que, como se indicó en la discusión, el modelo logístico tiende a subestimar la probabilidad de ocurrencia del evento. En cuanto a la segmentación del riesgo, la metodología aplicada tiene más sentido al utilizar las probabilidades estimadas mediante la regresión logística, ya que es posible distribuirlas en intervalos con frecuencias aproximadamente iguales, lo que permite evaluar el comportamiento de la proporción de eventos. En cuanto a lo obtenido por potenciación del gradiente, más de un 60 % de las probabilidades estimadas se acumulan en el intervalo $[0; 0,00137]$, lo que hace más complicada la segmentación de riesgo.

Desde un punto de vista netamente estadístico, el modelo obtenido a través de la potenciación del gradiente de árboles de decisión, presentó mejores resultados, sin embargo, la decisión final sobre el modelo a utilizar para calcular el *score* recae sobre la empresa, y en este sentido, resulta importante la naturaleza de las variables seleccionadas, así como la interpretabilidad de los resultados.

Capítulo 8

Anexo

8.1. Variables regresión logística

A continuación, se presenta un resumen gráfico de cada una de las variables que resultaron significativas para el modelo de regresión logística, el cual permite identificar su relación con el evento de interés. Cada variable continua fue categorizada en intervalos que presentasen una frecuencia lo más similar posible, representada por las barras. Mientras que la curva azul representa el porcentaje de eventos presente en cada intervalo.

1. Máximo saldo del mes anterior en los últimos 30 meses (en miles de pesos).

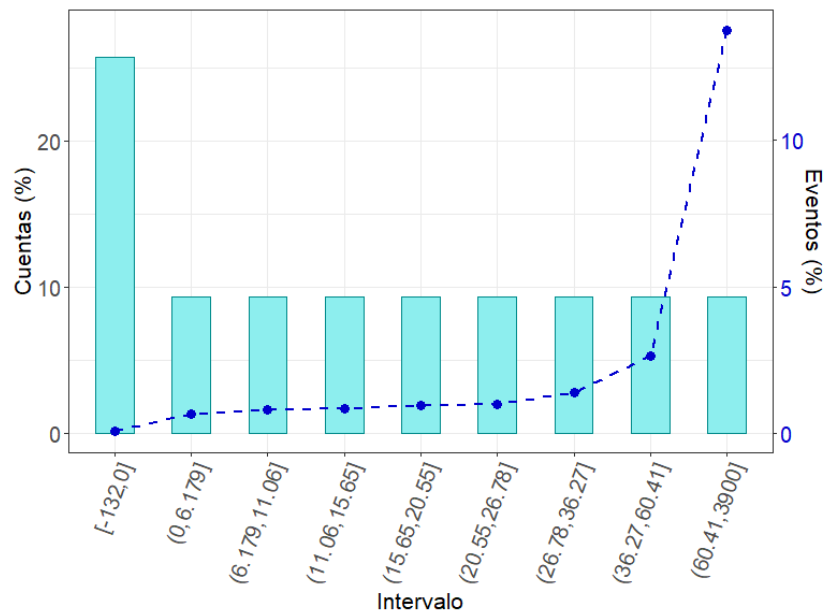


Figura 8.1: Distribución de eventos de acuerdo a la variable 1.

2. Número de meses en que el importe total baja respecto al mes anterior en los últimos 30 meses.

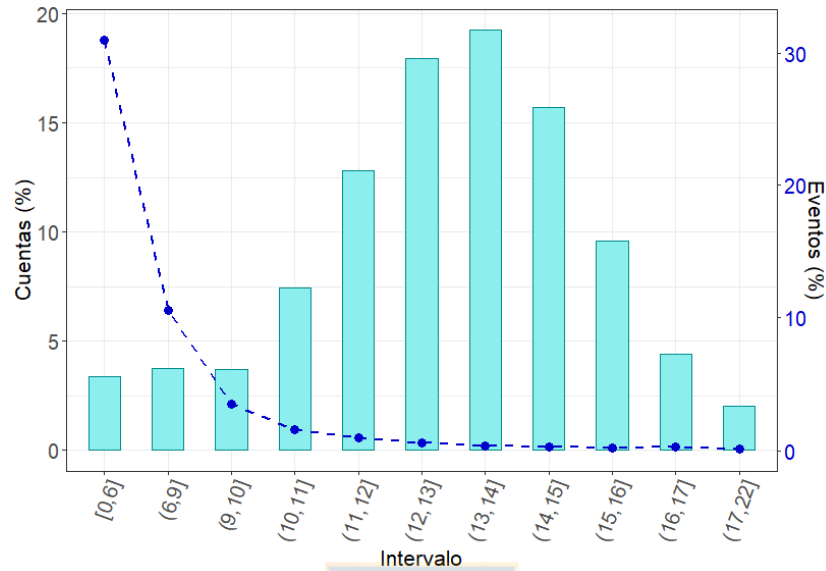


Figura 8.2: Distribución de eventos de acuerdo a la variable 2.

3. Ratio del importe total sobre el saldo del mes en los últimos 6 meses.

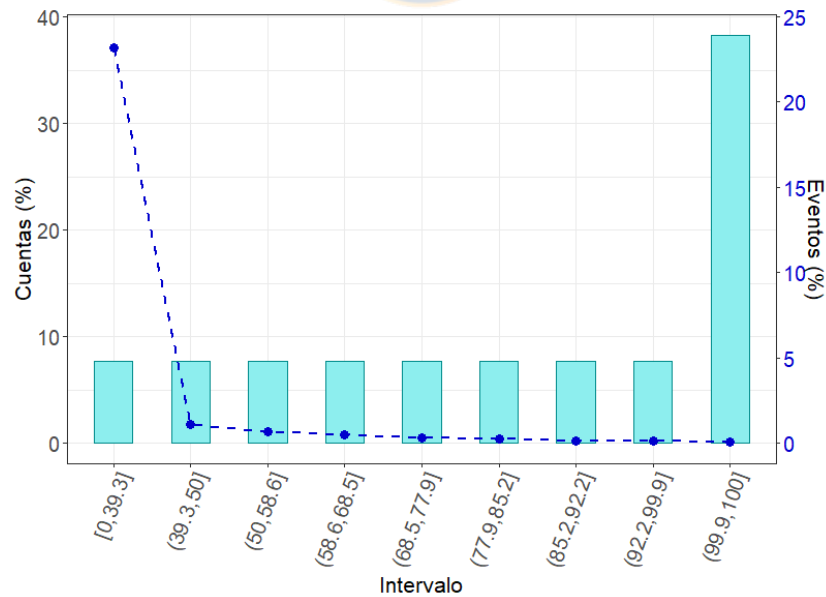


Figura 8.3: Distribución de eventos de acuerdo a la variable 3.

4. Realizó o no alguna llamada del tipo “Recaudación y cobranza” en los últimos 12 meses.

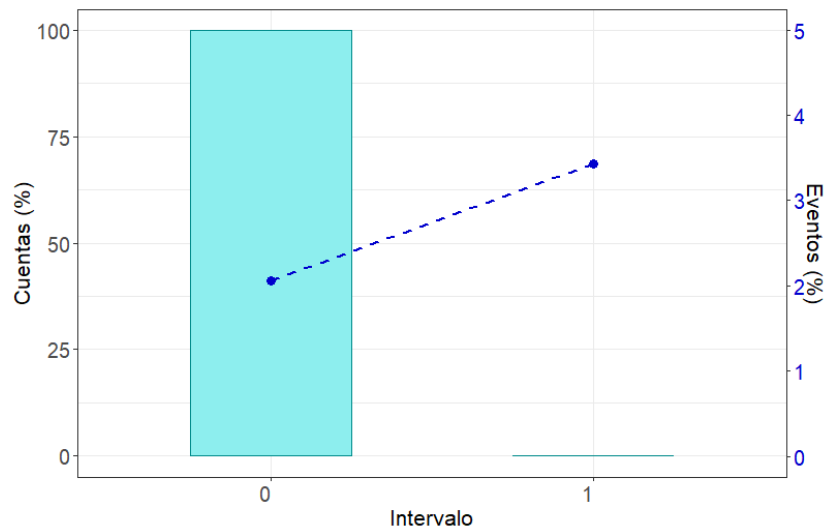


Figura 8.4: Distribución de eventos de acuerdo a la variable 4.

5. Tuvo o no más de 38 días de atraso en el último mes.

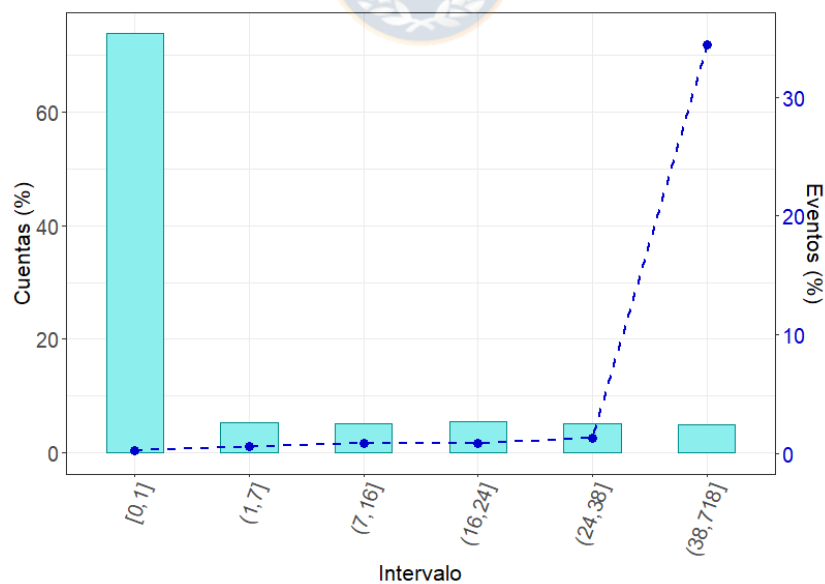


Figura 8.5: Distribución de eventos de acuerdo a la variable 5.

6. Tuvo o no un atraso máximo mayor a 62 días en los últimos 12 meses.

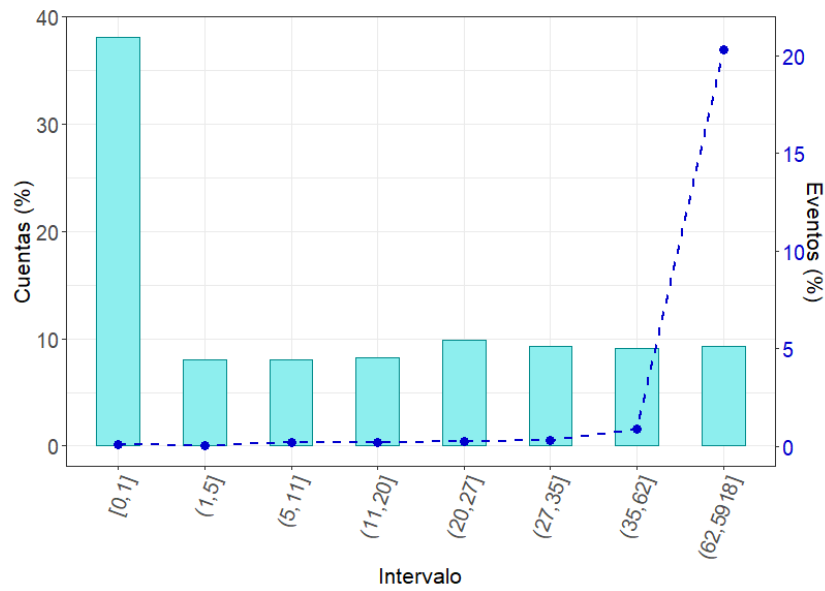


Figura 8.6: Distribución de eventos de acuerdo a la variable 6.

7. Mínimo número de pagos consecutivos en los últimos 3 meses.

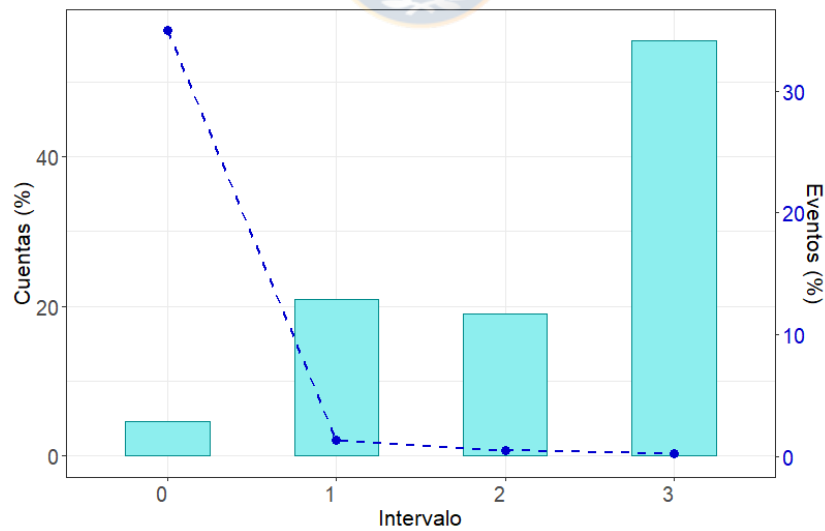


Figura 8.7: Distribución de eventos de acuerdo a la variable 7.

8. Ha tenido un convenio activo en los últimos 12 meses.

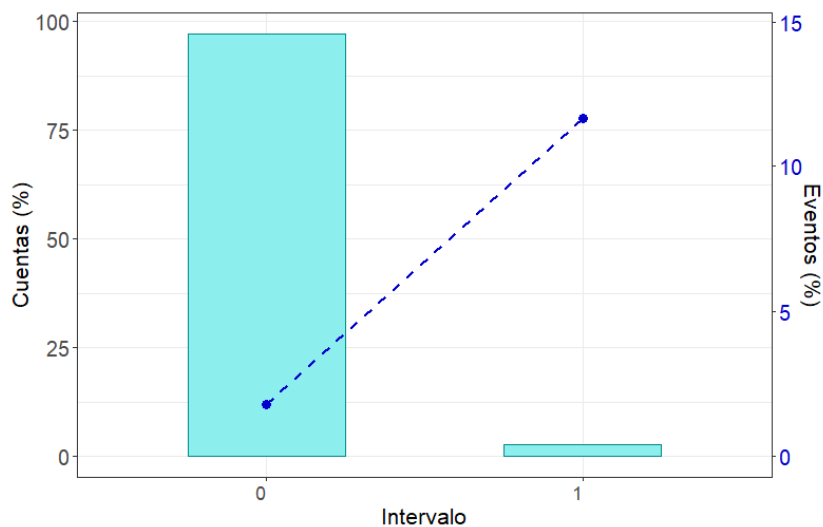


Figura 8.8: Distribución de eventos de acuerdo a la variable 8.

9. Ha tenido el servicio cortado en alguno de los últimos 12 meses.

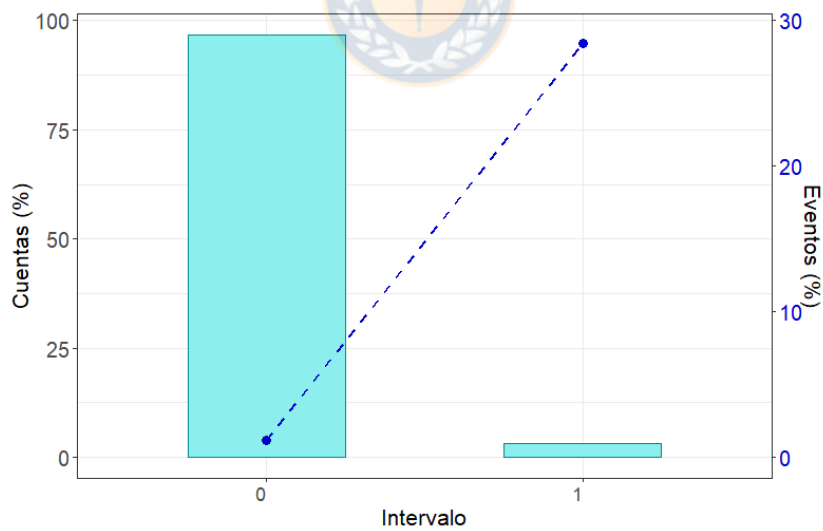


Figura 8.9: Distribución de eventos de acuerdo a la variable 9.

10. Ha recibido una carta por extinción de subsidio en los últimos 12 meses.

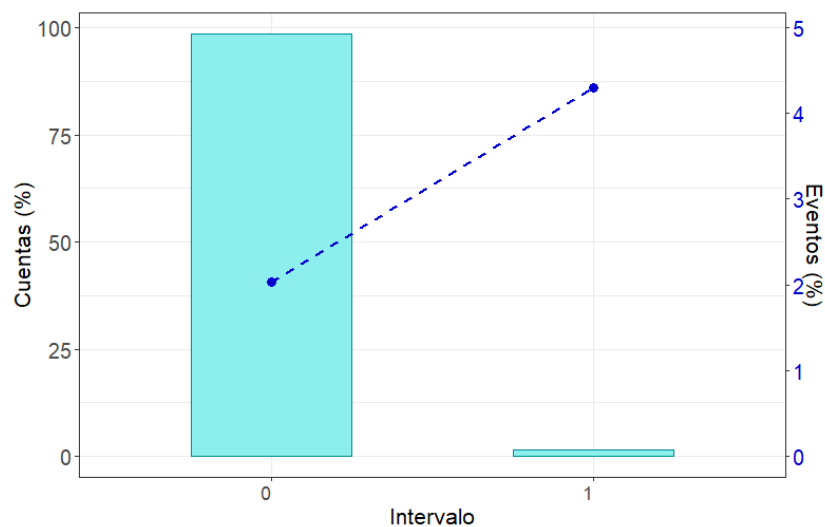


Figura 8.10: Distribución de eventos de acuerdo a la variable 10.

11. Ha tenido algún convenio del tipo CA o CI en los últimos 30 meses.

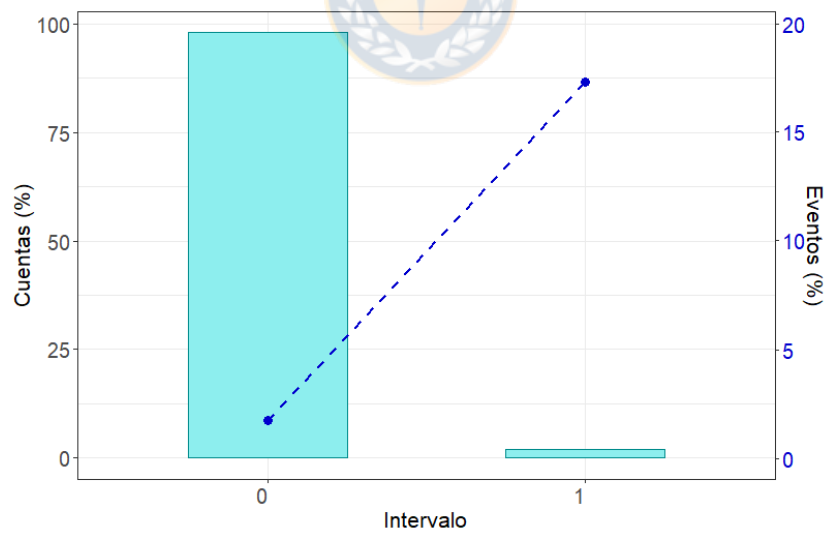


Figura 8.11: Distribución de eventos de acuerdo a la variable 11.

8.2. Variables potenciación del gradiente de árboles de decisión.

En esta sección se presenta el análisis correspondiente a las variables seleccionadas por el modelo ajustado mediante potenciación del gradiente en árboles de decisión. La interpretación de cada gráfico es análoga al caso anterior.

1. Ratio del importe total pagado sobre la facturación de los últimos 6 meses.

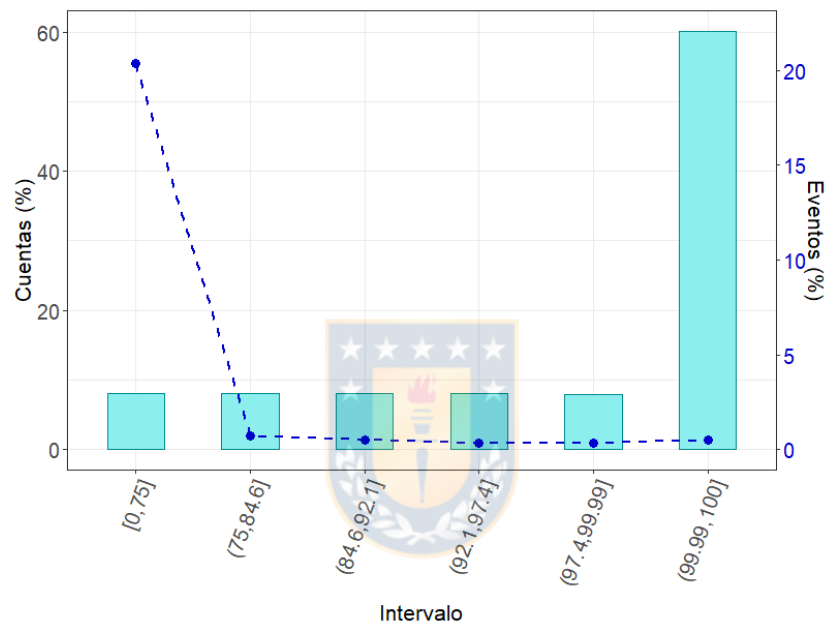


Figura 8.12: Distribución de eventos de acuerdo a la variable 1.

2. Ratio del importe total pagado sobre la facturación de los últimos 30 meses.

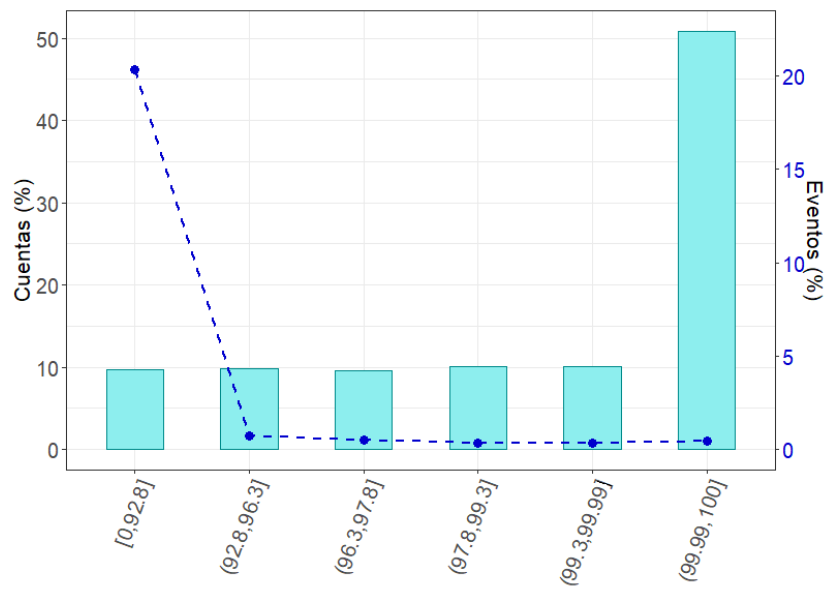


Figura 8.13: Distribución de eventos de acuerdo a la variable 2.

3. Ratio del importe total pagado sobre el saldo del mes de los últimos 3 meses.

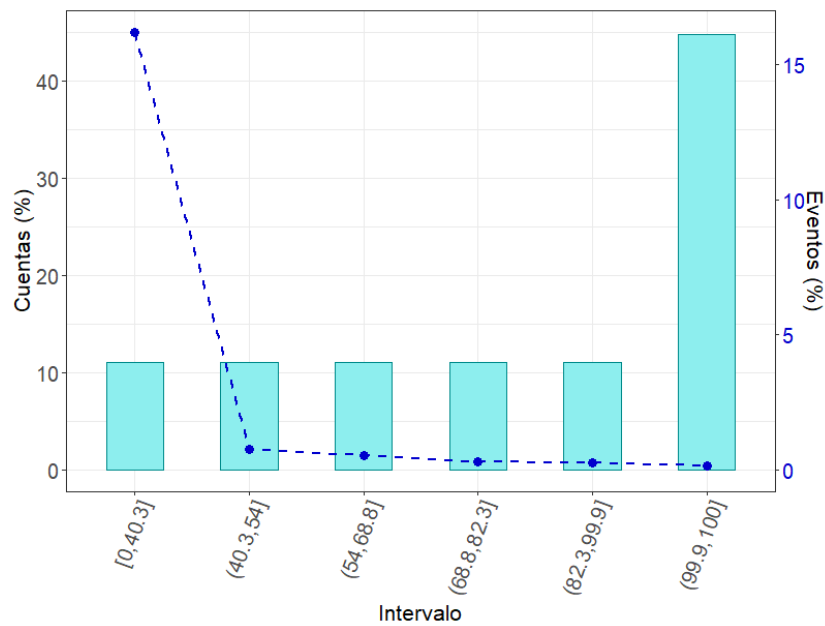


Figura 8.14: Distribución de eventos de acuerdo a la variable 3.

4. Días de atraso en el pago promedio en los últimos 30 meses.

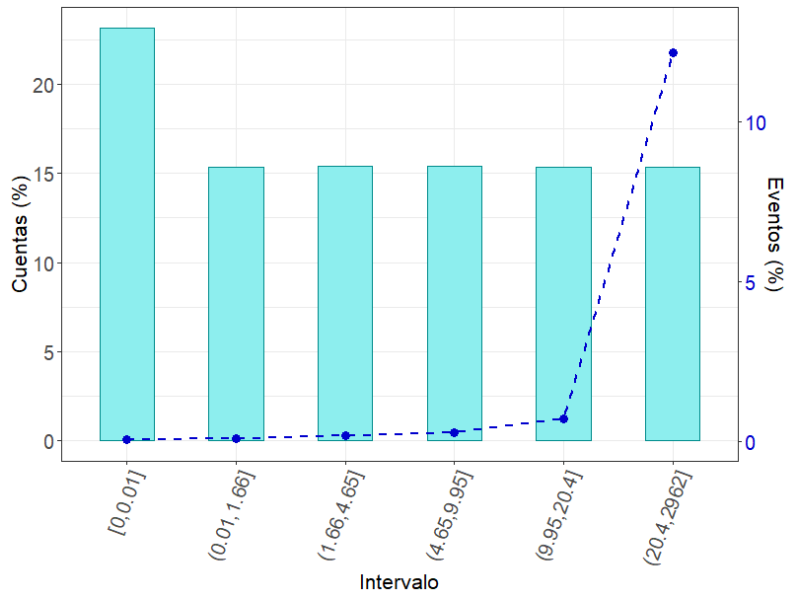


Figura 8.15: Distribución de eventos de acuerdo a la variable 4.

5. Saldo del mes anterior del último mes.

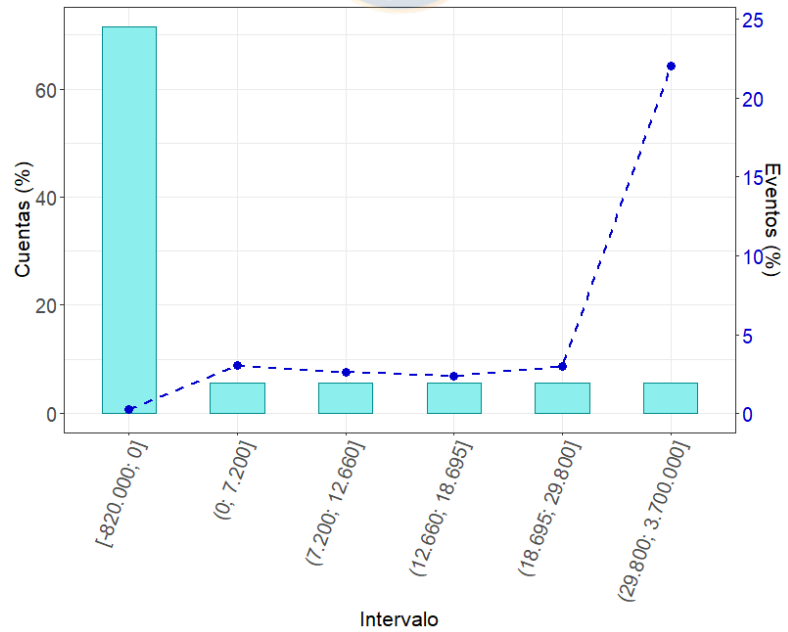


Figura 8.16: Distribución de eventos de acuerdo a la variable 5.

6. Máximo monto convenido de un convenio desactivado en los últimos 12 meses.

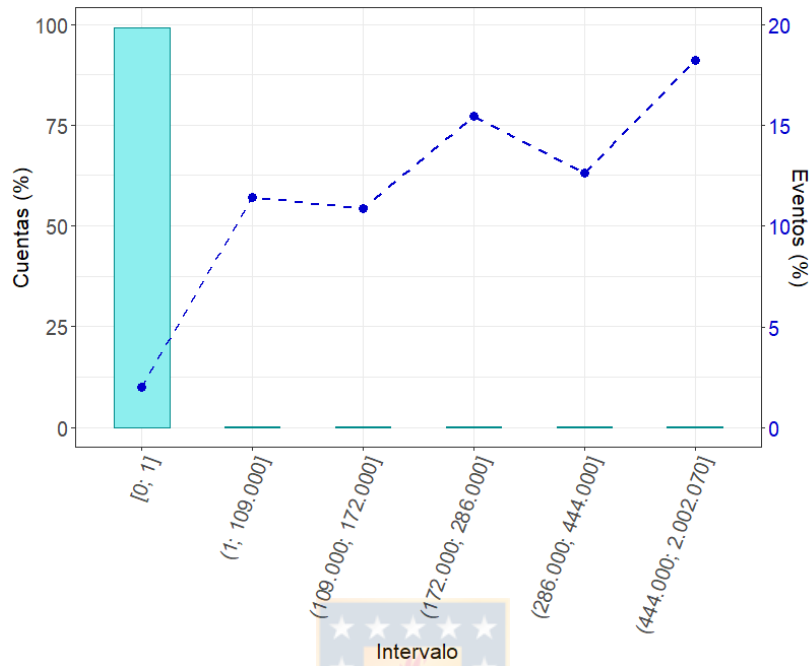


Figura 8.17: Distribución de eventos de acuerdo a la variable 6.

7. Ha tenido un convenio CA o CI en los últimos 3 meses.

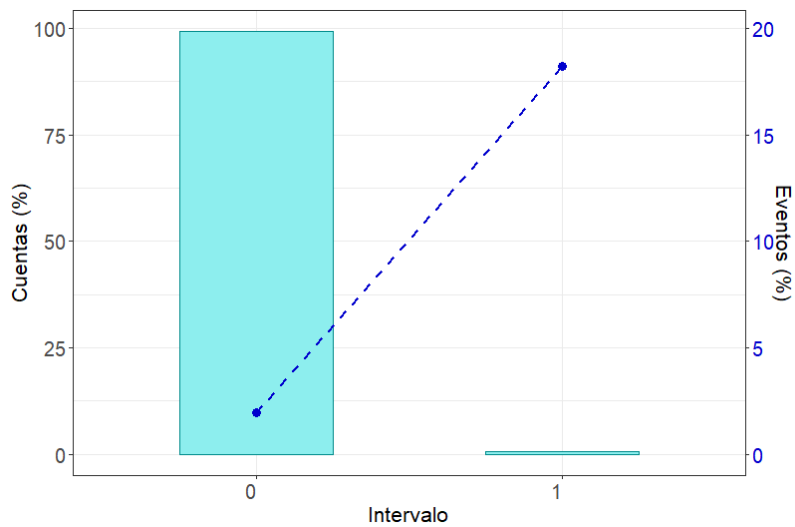


Figura 8.18: Distribución de eventos de acuerdo a la variable 7.

8. Ratio de la facturación sobre el saldo del mes de los últimos 12 meses.

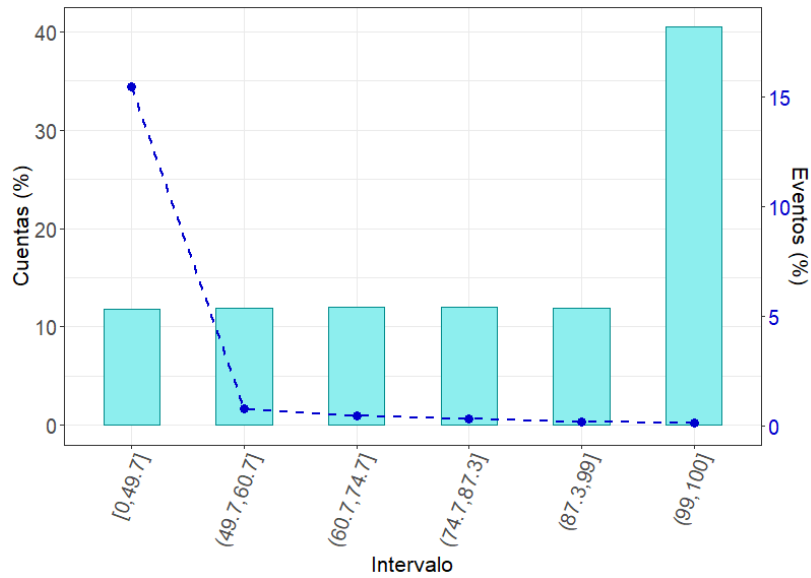


Figura 8.19: Distribución de eventos de acuerdo a la variable 8.

9. Saldo del mes anterior promedio en los últimos 24 meses.

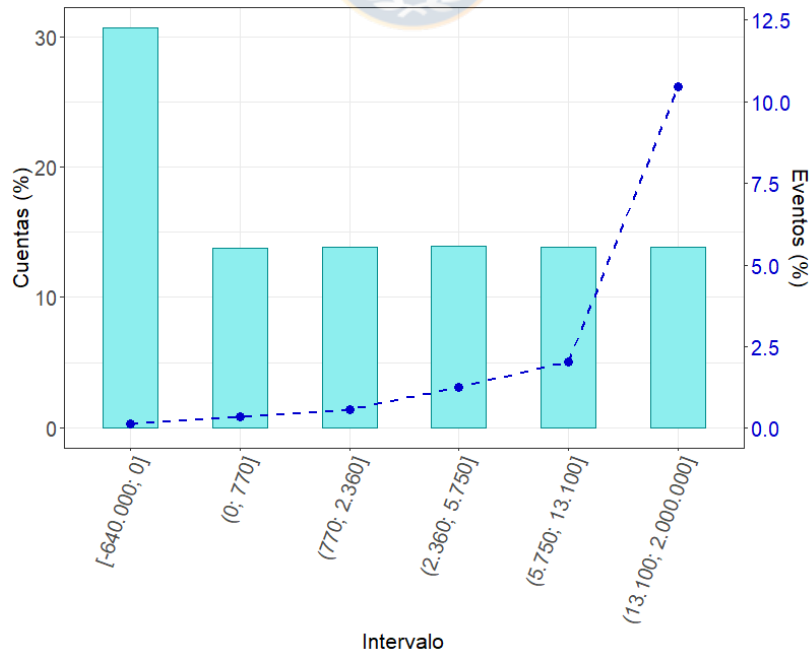


Figura 8.20: Distribución de eventos de acuerdo a la variable 9.

10. Ha tenido una orden de corte improcedente por ausencia del cliente los últimos 6 meses.

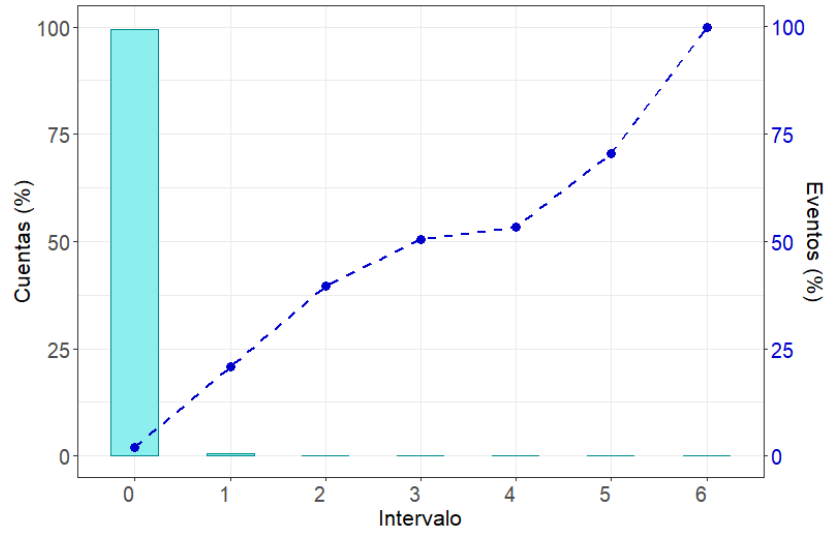


Figura 8.21: Distribución de eventos de acuerdo a la variable 10.

11. Ratio del saldo del mes anterior sobre el saldo del mes en los últimos 30 meses.

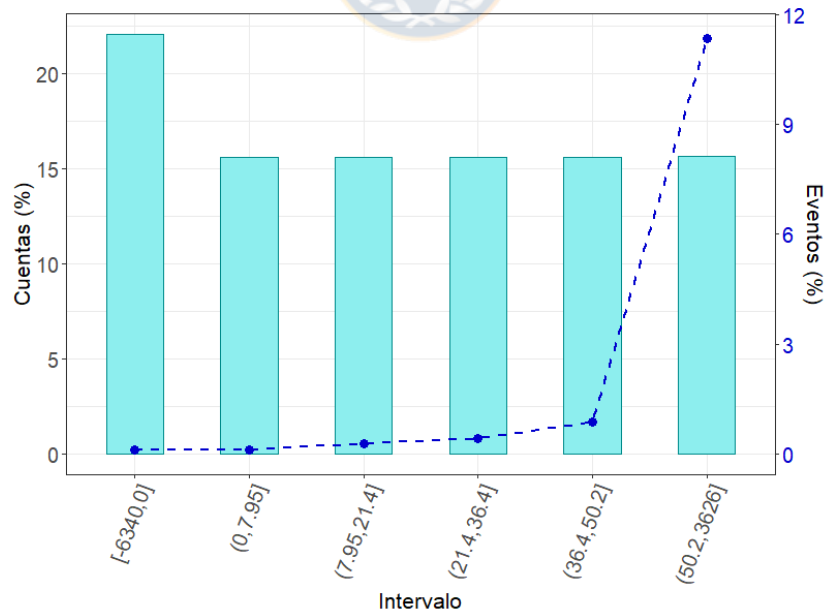


Figura 8.22: Distribución de eventos de acuerdo a la variable 11.

Bibliografía

Allison, P., 2013. *Logistic Regression For Rare Events*. Statisticalhorizons.com. Recuperado de: <https://statisticalhorizons.com/logistic-regression-for-rare-events>.

Bai, J., & Perron, P. (2003). *Computation and analysis of multiple structural change models*. Journal of applied econometrics, 18(1), 1-22.

Barrantes, L. D. (2017). *Propuesta de backtesting para comparar la estimación de provisión por incobrables contra la pérdida real de la cartera de préstamos hipotecarios, según establecen los principios de basilea para Ticobank* (Tesis de Posgrado). Universidad de Costa Rica, San José, Costa Rica.

Bartley, A. C. (2014). *Evaluating goodness-of-fit for a logistic regression model using the Hosmer-Lemeshow test on samples from a large data set* (Tesis de posgrado). Universidad de Ohio, Ohio, Estados Unidos.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: *Logistic regression*. Critical care, 9(1), 112.

Breiman, L., Friedman, J. H., Olshen, R. & Stone, C. (1983). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). *A systematic study of the class imbalance problem in convolutional neural networks*. Neural Networks, 106, 249-259.

Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). London: Chapman and Hall.

Di Cellio, P. C., Forti, M., & Witarsa, M. (2018). *A comparison of gradient boosting with logistic regression in practical cases*. SAS Support.

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed.). Chapman and Hall/CRC.

Essbio-Nuevosur. (2014). *Depto. Cobranza y medios de pago*.

Fox, J., & Monette, G. (1992). *Generalized collinearity diagnostics*. Journal of the American Statistical Association, 87(417), 178-183.

Friedman, J. H. (2001). *Greedy function approximation: a gradient boosting machine*. Annals of statistics, 1189-1232.

Girault, M. (2007). *Modelos de credit scoring: ¿Qué, cómo, cuándo, y para qué?*. Buenos Aires, Argentina: Gerencia de Investigación y Planificación Normativa, Subgerencia General de Normas, BCRA.

Gutiérrez, S., Osorio, H., & Romero, R. (2018). *NIIF 9 (IFRS 9) Instrumentos Financieros: Aplicación Práctica para Determinación de Pérdida Esperada de Carteras de Activos Financieros*.

Hajian-Tilaki, K. (2013). *Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation*. Caspian journal of internal medicine, 4(2), 627.

Hand, D. J., & Henley, W. E. (1997). *Statistical classification methods in consumer credit scoring: a review*. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.

IFRS Foundation. (2014). *IFRS 9 Instrumentos Financieros*. Recuperado de: https://www.mef.gob.pe/contenidos/conta_public/vigentes/NIIF9_2014_v12112014.pdf

King, G., & Zeng, L. (2001). *Logistic regression in rare events data*. Political analysis, 9(2), 137-163.

Kramer, A. A., & Zimmerman, J. E. (2007). *Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited*. Critical care medicine, 35(9), 2052-2056.

Lizares, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico* (Tesis de Pregrado). Universidad Nacional Mayor de San Marcos, Lima, Perú.

McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior*. Frontiers in econometrics (pp. 104-142). New York: Academic Press.

McFadden, D. (1977). *Quantitative methods for analyzing travel behaviour of individuals: Some recent developments (Cowles Foundation Discussion Papers No. 474)*. Cowles Foundation for Research in Economics, Yale University.

Menard, S. (2000). *Coefficients of determination for multiple logistic regression analy-*

sis. *The American Statistician*, 54(1), 17-24.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

Park, H. (2013). *An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain*. *Journal of Korean Academy of Nursing*, 43(2), 154-164.

Paul, P., Pennell, M. L., & Lemeshow, S. (2013). *Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets*. *Statistics in medicine*, 32(1), 67-80.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). *A simulation study of the number of events per variable in logistic regression analysis*. *Journal of clinical epidemiology*, 49(12), 1373-1379.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). *An introduction to logistic regression analysis and reporting*. *The journal of educational research*, 96(1), 3-14.

Peng, C. Y. J., & So, T. S. H. (2002). *Logistic regression analysis and reporting: A primer*. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 1(1), 31-70.

Ríos, J. (2017). *NIIF 9-Implementación y principales impactos*. KPMG. Recuperado de: <https://home.kpmg/co/es/home/media/Notas%20de%20prensa/2017/08/niif9-implementacion-y-principales-impactos.html>

Scheaffer, R. L., Mendenhall, W., & Ott, L. (2006). *Elementos de muestreo*(6ta ed.). Editorial Paraninfo.

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring* (Vol. 3). John Wiley Sons.

Silva, L. C., & Barroso, I. M. (2004). *Regresión logística*. Madrid: La Muralla.

Valenzuela, P.(2018). *Diseño e implementación de un sistema de apoyo a la gestión de cobranza, una solución para las sanitarias Essbio y Nuevosur* (Tesis de Pregrado).Universidad de Talca, Curicó, Chile.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science Business Media.

Wu, X. (2008). *Credit Scoring Model Validation*. Faculty of Science, Korteweg-de Vries Insitute for Mathematics.