UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DESARROLLO DE UN MODELO DE MEDICIÓN DE SCORING PARA LA ADMISIÓN DE DOCUMENTOS A OPERAR DENTRO DE UNA FINTECH DE FACTORING ONLINE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

ALFONSO ESTEBAN ORDOÑEZ ORMEÑO

PROFESOR GUÍA:

CARLOS PULGAR ARATA

MIEMBROS DE LA COMISIÓN:

HUGO SÁNCHEZ RAMÍREZ

RONALD FISCHER BARKAN

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: Ingeniero Civil Industrial
POR: Alfonso Esteban Ordoñez Ormeño
FECHA: 2023
PROFESOR GUÍA: Carlos Pulgar Arata

DESARROLLO DE UN MODELO DE MEDICIÓN DE SCORING PARA LA ADMISIÓN DE DOCUMENTOS A OPERAR DENTRO DE UNA FINTECH DE FACTORING ONLINE

El incumplimiento de pago en las instituciones financieras se manifiesta en un impacto significativo en su rentabilidad y estabilidad financiera de esta. Cuando los clientes o deudores no cumplen con sus obligaciones de pago, las instituciones financieras enfrentan pérdidas financieras, aumento de los costos de recuperación de los créditos y deterioro de su reputación.

Con el objetivo de reducir el riesgo asociado al incumplimiento, las instituciones financieras utilizan diferentes estrategias, como el análisis de la calidad de la cartera de clientes, el establecimiento de límites de crédito y la implementación de sistemas de Scoring y monitoreo de crédito. Estas estrategias son fundamentales para la disminución de la probabilidad de incumplimiento y, por ende, para mejorar la gestión del riesgo de crédito de las instituciones financieras.

El presente trabajo de memoria tiene como objetivo desarrollar un modelo predictivo que estime un puntaje basado en la probabilidad de default que trae asociado consigo un documento o factura a operar por la empresa, a fin de apoyar el proceso de admisión de documentos permitiendo decisiones rápidas y efectivas en el área de riesgo y operaciones de la empresa.

En cuanto a la metodología escogida para llevar a cabo el desarrollo del modelo de Credit Scoring, se actuará de acuerdo con el proceso KDD (Knowledge Discovery in Databases). El modelo desarrollado asigna un score a cada documento que debe ser operado en la empresa de factoring, con un desempeño de aproximadamente 80%. Para tomar decisiones estratégicas sobre el documento, se han establecido puntos de corte específicos. Los documentos con puntajes por debajo de 367 son rechazados automáticamente, mientras que aquellos con puntajes entre 630 y 1000 son aceptados automáticamente. Los puntajes intermedios, entre 367 y 630, son sometidos a evaluación por el comité de riesgo de la empresa.

El modelo propone beneficios significativos, como la reducción del 33% en la cantidad de documentos que el área de operaciones y riesgo debe evaluar, lo que permite liberar recursos y mejorar la eficiencia del área. Además, se ha logrado un ahorro considerable en el monto a provisionar gracias a la implementación del modelo de Scoring, debido a la esperable disminución de la perdida incurrida por la empresa. Se deja además una serie de recomendaciones y futuras directrices a seguir para la correcta implementación y constante actualización del modelo desarrollado.

Para todos mis seres queridos,
“Lo que con mucho trabajo se obtiene, más se ama”

AGRADECIMIENTOS

Agradezco a mis padres, Alfonso e Ilse, por darme todas las oportunidades y facilidades que he tenido hasta el día de hoy. En especial a mi madre, quien me ha apoyado incondicionalmente durante todos estos años tanto de estudios básicos como superiores. A mi hermano Iván, por su compañía y preocupación durante todos estos años, y por supuesto al miembro más joven de mi familia, mi amigo y mascota Mick, por su compañía fiel e incondicional.

A todos mis amigos de la infancia que han estado para brindar apoyo en todo momento, y nos ayudan a disfrutar la vida a full.

A todos los amigos que he generado durante mi transcurso en la universidad, en especial a mis mejores amigos, ellos saben quién son, con quien hemos cursado una carrera llena de alegrías y derrotas, pero sin sacrificio no hay victoria.

Agradecer a los profesores Carlos Pulgar, Hugo Sánchez y Carlos Reyes, quienes me han apoyado y acompañado durante un largo año en este proceso de titulación, siempre mostrando una gran disposición y buena onda.

TABLA DE CONTENIDO

1. Introducción 1
1.1 Características de la empresa 1
1.2 Justificación 5
1.3 Objetivos 10
1.3.1 Objetivo general 10
1.3.2 Objetivos específicos 10
1.4 Alcances, resultados esperados y limitaciones 10
2. Marco teórico 11
3. Metodología 14
3.1 Proceso KDD 14
3.2 Metodología propuesta en base a proceso KDD 15
3.2.1 Selección de datos 15
3.2.2 Procesamiento y limpieza 16
3.2.3 Transformación de datos 16
3.2.4 Data Mining 17
3.3 Métricas de evaluación de modelos de clasificación 18
3.3.1 Matriz de confusión 19
3.3.2 Curva ROC 19
3.3.3 ROC AUC 20
3.4 Selección de puntajes de corte 20
4. Selección de datos 21
4.1 Datos iniciales y estudio de variables a utilizar 21
4.2 Creación de variable objetivo 22
5. Selección, limpieza y transformación de las variables 23
5.1 Selección de variables 23
5.2 Limpieza de variables 25
5.3 Análisis de correlación entre las variables 25
5.4 Análisis exploratorio 25
5.5 Transformación de variables 28
5.5.1 Variables Categóricas 28
5.5.2 Categorización de variables continuas 30

6. Desarrollo del modelo 33
6.1 Primera Iteración 33
6.1.1 Selección del modelo 33
6.1.2 Selección de variables 34
6.1.3 Resultados del modelo 34
6.1.4 Evaluación de desempeño del modelo 38
6.2 Segunda Iteración 39
6.2.1 Selección del modelo 39
6.2.2 Selección de variables 40
6.2.3 Resultados del modelo 43
6.2.4 Evaluación de desempeño del modelo 45
7. Interpretación y evaluación de resultados 46
7.1 Escalamiento de los betas a score 46
7.2 Validación del score 52
7.3 Definición de puntajes de corte 54
8. Estimación de beneficios del modelo 60
9. Conclusiones 62
Bibliografía 66
Anexos 68
Anexo A. Reglas modelo LINCE 68
Anexo B. Resumen variables base “Documents” 70
Anexo C. Tratamiento de valores NA 72
Anexo D. Matriz de correlación 73
Anexo E. Clasificación tamaño empresas según ventas anuales (SII) 74
Anexo F. Resultados modelo Logit 75

ÍNDICE DE TABLAS

Tabla 1: Matriz de pérdidas esperada actual. (Fuente: elaboración propia) 14

Tabla 2: Betas calculados por el modelo logit para cada categoría y variable definida.....35

Tabla 3: Betas calculados por el modelo logit para cada categoría y variable definida (segunda iteración).....43

Tabla 4: Score asignado a cada variable…….48

Tabla 5: Puntajes de corte con variables acumuladas.....54

Tabla 6: Promedio de variables para el cálculo de ingresos y costos……57

Tabla 7: Utilidades y costos promedio calculados…….....58

Tabla 8: Utilidades promedio por rango de score…….....58

Tabla 9: Distribución final del score y puntajes de corte.....59

Tabla 10: Tasa de operación para diferentes intervalos de respuesta de operaciones…….....61

INDICE DE IMÁGENES

Ilustración 1: Proceso de Factoring de la empresa (Elaboración personal).....1

Ilustración 2: Etapas del proceso KDD ……15

Ilustración 3: Visualizaciones posibles valores curva AUC……..20

1. Introducción

1.1 Características de la empresa

Chita Spa es una Fintech fundada en la segunda mitad del año 2016, corresponde a la primera plataforma de factoring online en Chile, enfocada principalmente en el sector de micros, pequeñas y medianas empresas en busca de financiamiento**.** La empresa brinda servicios financieros de factoring con un modelo de negocios Business To Business (B2B), otorgándole liquidez a las diferentes empresas con las que opera, logrando así posicionarse en el mercado financiero chileno como una Fintech de factoring online. De acuerdo con datos entregados por la empresa en su sitio web, Chita ha financiado más de $255.000 millones, ha operado más de 150.000 facturas y posee más de 6.800 clientes. Actualmente la empresa cuanta con un total de 54 trabajadores. Por motivos de confidencialidad no es posible explicitar los balances de la empresa.

Según lo define la Comisión para el Mercado Financiero [1], el factoring o factoraje es una alternativa de financiamiento que se orienta de preferencia a pequeñas y medianas empresas y consiste en un contrato mediante el cual una empresa traspasa el servicio de cobranza futura de los créditos y facturas existentes a su favor y a cambio obtiene de manera inmediata el dinero a que esas operaciones se refiere, aunque con un descuento. Celebrándose de esta manera un contrato de 3 partes:

1. **Empresa de factoring**: Parte que anticipa un monto de dinero con total liquidez al cliente.

2. **Cliente**: Parte que solicita un anticipo de dinero ofreciendo sus cuentas por cobrar a la empresa, principalmente facturas, a cambio de liquidez inmediata.

3. **Deudo**r: Parte que cancela la deuda de los documentos emitidos por el cliente directamente con la empresa de factoring, ahora propietario de las cuentas por cobrar.

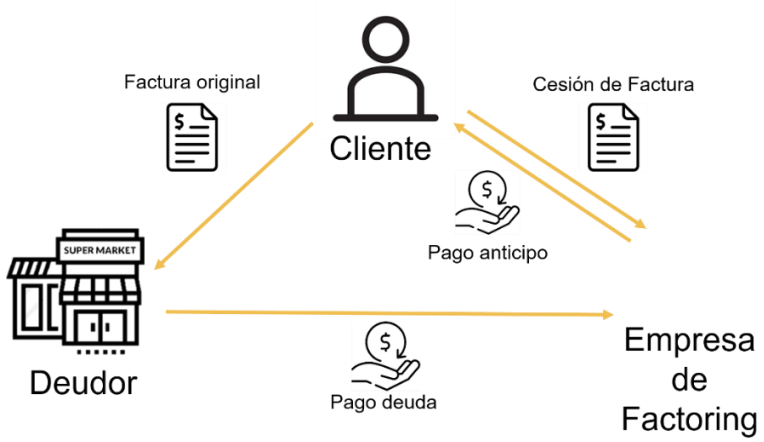

Ilustración 1: Proceso de Factoring de la empresa (Elaboración personal).

Como se detalla en la ilustración 1, se genera un flujo entre la empresa de Factoring, el cliente y el deudor. El proceso comienza cuando el cliente contacta a la empresa de factoring a través de su sitio web para recibir sus servicios. La empresa de factoring a la vez solicita información de este cliente evaluando su historial con la empresa y su situación en el Servicios de Impuestos Internos a fin de determinar el riesgo asociado a dicho cliente, que permita decidir si operar o no sus facturas. Posteriormente, una vez realizado este análisis del cliente se hace efectiva la cesión de la factura y el cliente recibe su monto de dinero anticipado. Es importante destacar que uno de los requisitos solicitados al cliente corresponde a la firma del Contrato Marco, estrategia de contratación basada en un acuerdo de voluntades que celebra una dependencia o entidad con uno o más posibles proveedores, mediante los cuales se establecen las especificaciones técnicas y de calidad, alcances, precios y condiciones que regularán la adquisición o arrendamiento de bienes muebles, o la prestación de servicios. [2] La firma de este contrato confiere a la empresa de factoring la autoridad para asumir una posición legal en caso de incumplimiento en el pago de las facturas. Esto implica que tanto el cliente como el deudor pueden ser considerados responsables desde una perspectiva legal. Una vez finalizada esta parte del proceso, el deudor deberá cancelar el monto de la factura dentro de los plazos y condiciones acordadas en el contrato realizado por el cliente y la empresa.

Los diferentes casos de pago en factoring están relacionados con el momento y la forma en que se realiza el pago por parte del deudor o del cliente del cliente de la empresa de factoring. La empresa de factoring asume diferentes niveles de riesgo y obtiene su rentabilidad a través de la comisión y/o intereses que ha acordado con su cliente. A continuación, se detallan los diferentes casos de pago que se pueden presentar en esta industria:

- **Pago del deudor:** En este caso el deudor realiza el pago del documento o factura directamente a la empresa de factoring dentro de la fecha de vencimiento establecida al momento de la operación. La empresa de factoring obtiene su rentabilidad a través de la comisión e intereses acordados con el cliente.

- **Pago del cliente:** En este caso el cliente realiza el pago directamente a la empresa de factoring. Se da principalmente cuando el deudor incumple sus funciones como pagador, donde el cliente, debido a firma del contrato marco previo a la operación, acuerda asumir toda responsabilidad jurídica del crédito.

- **Pago anticipado por el deudor:** En este caso, el deudor realiza el pago de la factura antes de la fecha de vencimiento establecida en la factura. La empresa obtiene su rentabilidad a través de la comisión y/o intereses que ha acordado con su cliente.

- **Impago de la factura:** Se produce cuando el deudor no realiza el pago antes de la fecha de vencimiento.

- **Mal pago:** En este caso, el deudor comete un error y realiza el pago del documento al cliente. Donde el cliente debe transferir dicho pago a la empresa de factoring.

- **Pago post fecha vencimiento:** En este caso, el deudor realiza el pago del documento posterior a la fecha de vencimiento de este. La empresa obtiene su rentabilidad a través de la comisión e intereses que ha acordado con su cliente, además de una tasa asociada a los días de mora que acumuló el documento.

Chita Spa brinda sus servicios principalmente a empresas que son categorizadas por el Servicio de Impuestos Internos (SII) como microempresas, pequeñas empresas y medianas empresas. Esta categorización se basa en la evaluación del volumen de ventas de dichas empresas.

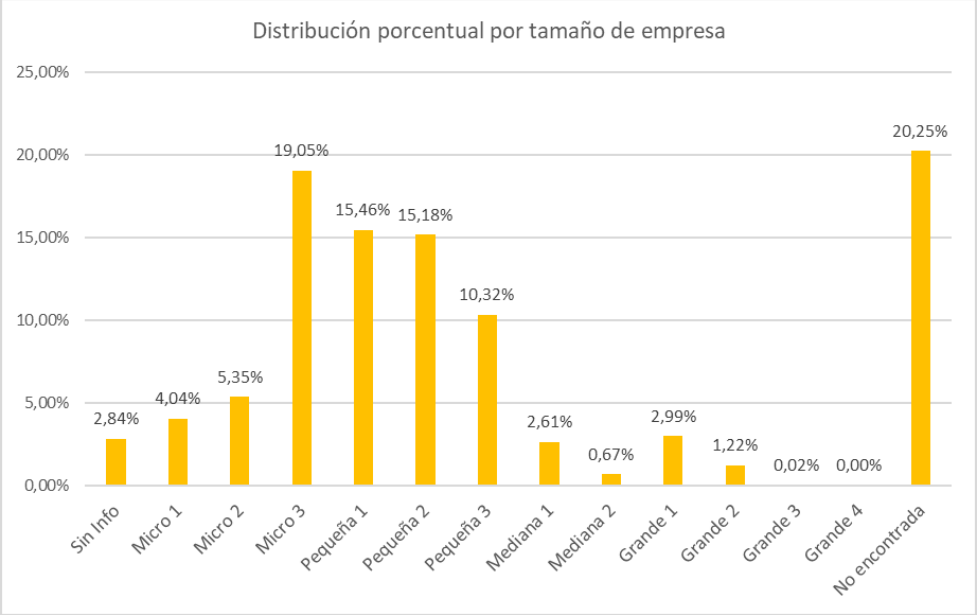

Gráfico 1: Distribución por tamaño de clientes operados por Chita (Elaboración personal)

Como se aprecia en el gráfico 1, la mayor parte de los clientes de la empresa se encuentran entre los rangos micro, pequeña y mediana empresa, la columna definida como “No encontrada” corresponde a clientes de la empresa cuyos datos no se encuentran en la base del SII (considerando que la base entrega la nómina de personas jurídicas año comercial 2020) y el gráfico se elaboró con datos de la empresa correspondientes al cierre del mes de noviembre del año 2022.

En el gráfico 2, se presenta la evolución en el último año del volumen YTD (Year To Date) en monto operado por la empresa.

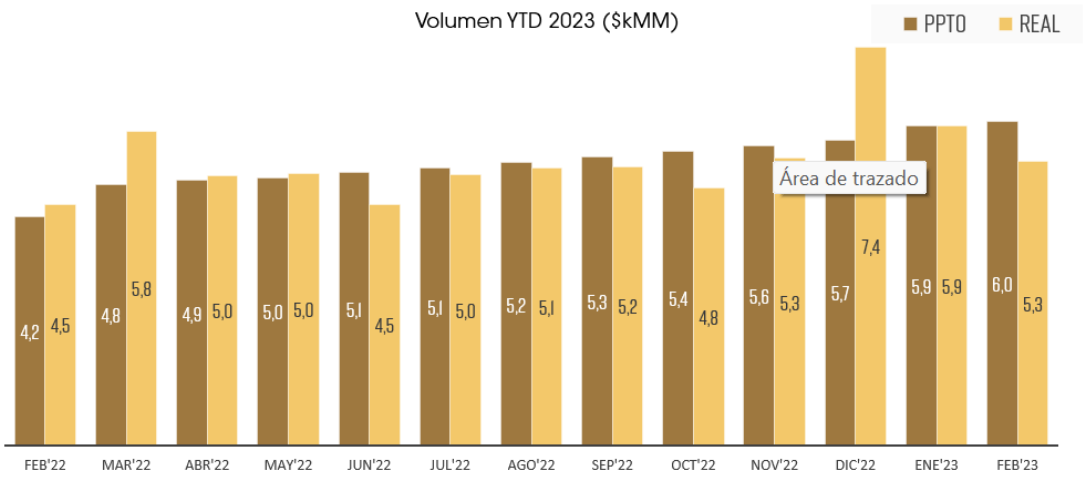

Gráfico 2: Evolución del volumen de colocaciones de la empresa Chita (Elaboración personal)

La empresa, al ser una Fintech, no tiene una sucursal física. En cambio, opera de manera cien por ciento remota, llevando a cabo todas sus transacciones y contactos con los clientes a través del sitio web de la empresa y mediante vía telefónica por parte del equipo del área comercial.

Dentro del marco institucional y regulatorio del país para empresas que pertenecen al sector financiero, existen varias regulaciones instauradas por la Comisión del Mercado Financiero (CMF) y la Superintendencia de Valores. Más en específico, la empresa se ve principalmente regulada por la Ley Nº 19.983, la cual regula la transferencia y otorga mérito ejecutivo a la copia de facturas para pequeños y medianos empresarios [3].

Además, es relevante destacar que el 12 de octubre de 2022, la Cámara de Diputados aprobó el Proyecto de Ley Fintech y Open-Banking, faltando únicamente su promulgación y publicación. Esta ley representa un avance significativo para promover la innovación financiera y una mayor competencia en el sistema financiero, así como el desarrollo de nuevos productos y servicios financieros para los consumidores. Entre los principales aspectos que contempla dicha ley se encuentran:

- Se establece un marco regulatorio para ciertos servicios financieros de base tecnológica que no contaban con un marco jurídico propio.

- Las entidades que realizan operaciones de intermediación y custodia, provisión de plataformas de transacción y asesoría sobre instrumentos financieros pasarán a estar reguladas por la Comisión para el Mercado Financiero y deberán acreditar el cumplimiento de los requisitos que la autoridad fije para que puedan operar.

- Se crea un Sistema de Finanzas Abiertas (Open Banking) que posibilitará que los proveedores de servicios financieros intercambien información financiera de clientes.

- Regula a los proveedores de servicios de iniciación de pagos, quienes podrán prestar servicios para efectuar transferencias electrónicas desde la cuenta de los clientes a cuentas de terceros.

- Se reconoce el uso de cripto activos como medios de pagos.

- Se modifican distintas leyes que rigen a instituciones financieras tradicionales a fin de lograr simetría regulatoria en la prestación de servicios financieros similares. [4]

1.2 Justificación

Todas las entidades financieras que ofrecen créditos masivos a sus clientes deben abordar el problema del incumplimiento de manera efectiva para garantizar la estabilidad financiera y proteger sus intereses y los de sus clientes. Por ende, es fundamental que las instituciones financieras implementen medidas efectivas de gestión de riesgos, así como políticas claras de concesión de créditos y de recuperación de deudas.

Actualmente en Chita Spa, de la totalidad de documentos o facturas que llegan a la empresa con la intención de ser operadas, solo el 25% se opera de manera automática a través del modelo **Lince** existente en la empresa, de la siguiente manera:

- Lince corresponde a un motor que actualmente presenta un total de 52 reglas definidas por el área de riesgo y operaciones de la empresa (“Anexo A”), las cuales se dividen en reglas de aceptación automática y bloqueo.

- Toda operación que cumpla con alguna de las reglas de aceptación automática definidas por el motor se acepta automáticamente, siempre y cuando no haga match con una regla de bloqueo.

- Toda operación que haga match con alguna regla de bloqueo o no haga match con ninguna regla de aceptación automática establecida por el motor pasa directamente a ser evaluada por el área de riesgo y operaciones de la empresa.

Cada analista del área de riesgo y operaciones de la empresa se encarga de analizar y evaluar el restante 75% de las facturas que no fueron aceptadas automáticamente por las reglas LINCE a fin de determinar si aceptar o rechazar la operación, análisis principalmente basado en la experiencia y por lo tanto en la subjetividad del analista, lo que deriva en un proceso que se desarrolla con lentitud. Además, considerando el número de operaciones que llegan por día a la empresa, el hecho de no existir una automatización de este proceso deriva en un aumento del costo para la empresa, manifestado en la necesidad de mantener un considerable número de trabajadores en el área de operaciones encargados de dicho análisis.

El trabajo de memoria se desarrolló en el área de Business Intelligence (BI) más en específico, se llevó a cabo en conjunto con el área de Riesgo y Operaciones de la empresa lideradas por Patrick Real, COO (Chief Operating Officer) de la empresa. Entre

las principales funciones del área de BI, se encuentra la creación y automatización de reportes qué combinando análisis, minería, visualización, herramientas e infraestructura de datos, ayudan a las distintas áreas de la empresa a tomar decisiones basadas en dichos datos. Por otro lado, el área de riesgo se encarga de balancear los riesgos con las oportunidades que estos presentan, su retorno sobre la inversión y su impacto en el crecimiento y en la supervivencia de la empresa.

El trabajo de memoria fue solicitado por Christian Real y Patrick Real, CEO y COO de la empresa, respectivamente, con el propósito de desarrollar un Modelo de Scoring que evalúe directamente los documentos o facturas recibidos por la empresa, pues el modelo actual utilizado en la institución solo asigna una clusterización a clientes y deudores por separado. Dicho modelo tiene como objetivo asignar una puntuación o Scoring a cada documento o factura que llega a la empresa con la intención de ser procesada. El Scoring se basa en diversas características del documento, ya que cada documento que llega a la empresa con la intención de ser aceptado y procesado está asociado a dos entidades: un cliente y un deudor. La solicitud de elaboración de un modelo de Scoring para la empresa se debe al alto porcentaje de facturas o documentos no pagados por diferentes clientes y deudores que la empresa ha experimentado en sus seis años de operación.

Un problema clásico que se da en la industria del factoring consiste en el default o no pago de los montos anticipados por parte de las empresas. En específico, en la empresa Chita existe el concepto denominado como “castigos”, los castigos corresponden a documentos o facturas que no se han pagado luego de 360 días desde la fecha de vencimiento del documento. De acuerdo con el modelo de riesgo actual presente en la empresa se asume que este monto castigado por factura no se recuperará, por lo que se necesita provisionar el 100% del monto anticipado.

De acuerdo con un análisis realizado por el memorista, al cierre del mes de marzo del año 2023 el monto total castigado por la empresa en sus 6 años de existencia es de aproximadamente $462.421.795. Más en específico, en el año 2020 hubo un total de monto castigado de $45.899.416, monto que al compararlo con el resultado anual en ventas del año 2020 que equivale a $234.535.324 corresponde a aproximadamente un 20% de este, por lo que el problema encontrado tiene una gran relevancia y presenta una gran oportunidad de mejora.

Actualmente, la empresa posee un modelo simple de clasificación de clientes y deudores según su probabilidad de default, que analizando diversas características de estas empresas les asigna una calificación que va desde la A+ para las empresas que son muy buenos pagadores hasta la letra F para aquellas que presentan una gran probabilidad de default. Sin embargo, al ser un modelo simple sin un mayor desarrollo estadístico este presenta varias falencias al momento de determinar la probabilidad de no pago de los clientes y deudores.

El modelo que presenta actualmente la empresa considera 5 aspectos fundamentales para evaluar y asignar la clasificación a cada empresa, ya sea cliente o deudor, estas son:

- **Cumplimiento:** Capacidad de una persona o entidad para cumplir con sus obligaciones financieras en tiempo y forma. Obtenida a través de la consulta al Sistema Nacional de Comunicaciones Financieras (SINACOFI) y a DICOM.

  Información consultada de cada empresa:

  - Deuda total con distintas instituciones financieras.
  - Monto de documentos impagos.
  - Infracciones previsionales.

- **Respaldo:** Se refiere a la garantía que se ofrece para respaldar el cumplimiento de una obligación financiera, como el pago de una deuda o la emisión de una garantía.

  Información consultada:

  - Patrimonio, automóviles y bienes raíces (información que la empresa obtiene de manera manual)
  - Tamaño (obtenida a través de la Nómina de empresas personas jurídicas reportada cada año por el Servicio de Impuestos Internos)

- **Historial:** Corresponde al registro de la actividad crediticia de una entidad en el tiempo con la empresa de factoring.

  Información utilizada:

  - Número de documentos facturas que se han operado con la empresa.
  - Porcentaje de la totalidad de documentos operados que han sido pagados por el deudor.

- **Estabilidad:** Corresponde a la capacidad de una empresa o entidad para mantener una situación financiera saludable y sostenible en el tiempo.

  Información consultada:

  - Venta promedio mensual.
  - Antigüedad en el SII.

- **Trazabilidad:** Corresponde a la capacidad de seguir el rastro o la ruta de una transacción financiera desde su origen hasta su destino final.

  Información consultada:

  - Sitio web confiable (consultado de forma manual por el equipo de riesgos y operaciones de la empresa) a fin de determinar si efectivamente el cliente presta los servicios informados.
  - Dirección de la empresa

Se presenta un análisis detallado del desempeño histórico del modelo, que se aplica tanto a las empresas que operan como clientes, como a aquellas que operan como deudores.

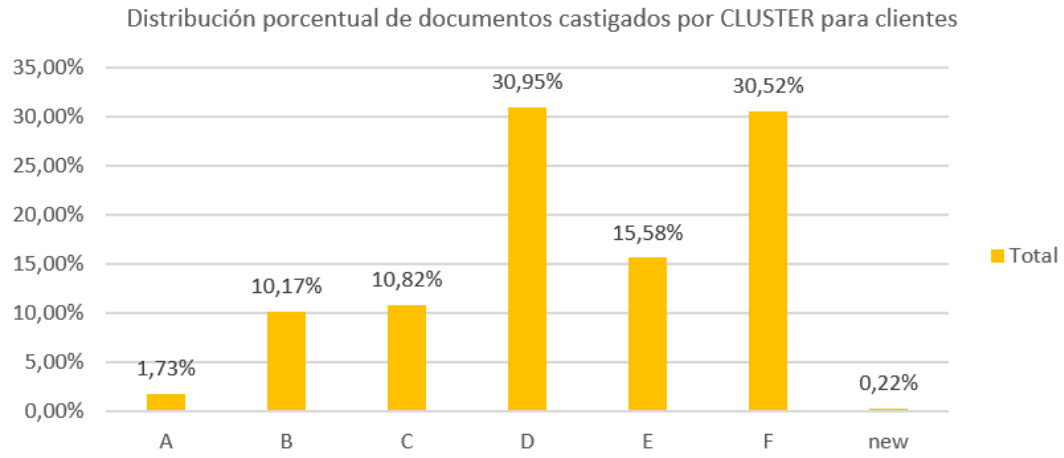

Gráfico 3: Distribución porcentual de documentos castigados para clientes de la empresa (elaboración propia).

Como se puede apreciar en el gráfico 3, el mayor porcentaje de documentos castigados pertenece a los clientes asignados por el modelo actual en el Clúster B, C, D, E y F, no existiendo documentos castigados para el Clúster A+ y solo un 1,73% para clientes con cluster A. Se realizó el mismo análisis para los deudores, lo cual se presenta en el siguiente gráfico.

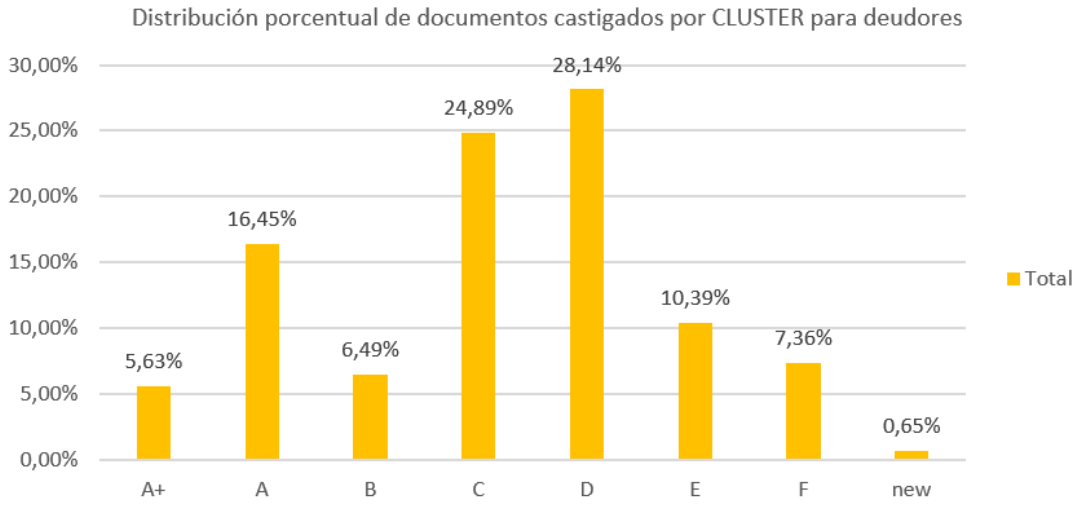

Gráfico 4: Distribución porcentual de documentos castigados para deudores de la empresa (elaboración propia).

Del gráfico 4, es posible observar que, para los deudores, la mayor parte de los documentos castigados corresponden a facturas operadas con deudores que según el modelo actual presentan un clúster C y D. Sin embargo, es posible observar que del total de documentos castigados un 16,45%, corresponden a documentos operados por deudores clasificados por el modelo actual como A, superando incluso al porcentaje de documentos castigados operados por deudores pertenecientes al cluster E y F. Además,

un 5,63% del total de documentos castigados pertenecen a deudores clasificados como A+,

Lo anterior, es una evidencia de que el modelo actual no clasifica de manera correcta a cada cliente y deudor que opera con la empresa, lo que se ha traducido en el monto de $443.911.890 total castigado hasta el 28 de febrero del año 2023.

Considerando que el modelo actual presentado por la empresa fue construido y ha estado en uso desde su inicio en el año 2016, existe una alta probabilidad de que su efectividad se haya visto disminuida debido a cambios en la distribución de los factores de riesgo que este utiliza y considera.

Ante esto, se plantea como alternativa de solución la creación de un modelo de Scoring, herramienta analística que tiene el objetivo de predecir, modelar y discriminar entre buenos y malos pagadores, basado en el historial de pagos de los clientes y deudores con la empresa y sus características, con el fin de caracterizar perfiles prestatarios con alta y baja probabilidad de impago. Estos modelos de análisis de Scoring se han estudiado a partir del año 1970 en el análisis de otorgamiento de crédito con Altman (1968), pero generalizados e instaurados a partir de los años 90 debido al desarrollo estadístico y tecnológico con Beaver (1996).

Ricardo Schliebener, presidente de la EFA (Empresas de servicios Financieros), que corresponde a un gremio que agrupa a empresas de servicios financieros no bancarios enfocados en factoring, señaló en un artículo publicado en el diario La Tercera como: “Construcción y comercio: los sectores que mantienen en alerta a los factoring no bancarios”, que a marzo del año 2022 el número de colocaciones de factoring aumentó en un 23% y la mora subió un 5 %. “En épocas como esta, cuando hay crisis, los bancos restringen sus líneas y retroceden, y quedan fuera empresas que no tenían problemas, que son buenas, entonces nos llegan más clientes que los que teníamos antes, como el industrial” (Ricardo Schliebener, 2022) [5]. Lo anteriormente mencionado, indica que la industria del factoring sigue creciendo, por lo que contar con una herramienta capaz de clasificar rápidamente entre clientes y deudores rentables para la compañía de los que no lo son, se alinea de manera correcta con los objetivos de la empresa.

La toma de decisiones financieras es crucial para cualquier empresa. En la actualidad, cada vez son más las instituciones financieras que utilizan modelos de Scoring para evaluar el riesgo crediticio de sus clientes y prevenir pérdidas financieras. En este sentido, la implementación de un modelo de Scoring en la empresa puede traer numerosos beneficios en términos de gestión de riesgos, toma de decisiones y eficiencia operativa. A continuación, se describe detalladamente los beneficios que este modelo puede aportar a la empresa:

- Aceleración y automatización del proceso de evaluación de documentos o facturas realizado de forma manual por el área de riesgo y operaciones de la empresa.

- La automatización en la evaluación de los documentos que no cumplen con las reglas de LINCE y que son remitidos directamente al área de riesgos y operaciones, permitirá acelerar el proceso de evaluación de operaciones

disminuyendo los tiempos de respuestas de estos, aumentando la eficiencia del área.

- La implementación de un modelo de Scoring puede conducir a una disminución significativa de la pérdida incurrida por la empresa. Como resultado, la empresa puede reducir el porcentaje a provisionar por tramo de mora definido por ella, lo que puede ser un beneficio significativo en términos de eficiencia operativa y gestión de riesgos.

1.3 Objetivos

1.3.1 Objetivo general

El objetivo general del trabajo de memoria se define como:

“Desarrollar un modelo predictivo que estime un puntaje basado en la probabilidad de default que trae asociado consigo un documento o factura a operar por la empresa, a fin de apoyar el proceso de admisión de documentos permitiendo decisiones rápidas y efectivas en el área de riesgo y operaciones de la empresa”

1.3.2 Objetivos específicos

Para lograr a cabalidad el objetivo general definido, se definen los siguientes objetivos específicos:

1. Identificar variables significativas que influyen en la probabilidad de que un documento o factura caiga en default o no pago por parte de clientes y deudores.

2. Estudiar y modelar un algoritmo o modelo que permita estimar la probabilidad de que un documento o factura caiga en default, analizando tanto características del cliente y deudor, como de las facturas y la transaccionalidad con el cliente.

3. Definir puntajes de corte que permitan asignar a cada documento o factura dentro de una categoría definida de acuerdo con su probabilidad de pago, permitiendo definir umbrales de aceptación y rechazo.

4. Realizar una estimación de los beneficios tantos financieros como operacionales que traería a la empresa la implementación de este modelo de Scoring.

1.4 Alcances, resultados esperados y limitaciones

Se espera la construcción de un modelo de Scoring que permita la admisión o rechazo del 75% de las operaciones que llegan a la empresa y no son aprobadas automáticamente por el modelo lince, con el objetivo de apoyar al área de riesgo y operaciones de la empresa en este proceso. Es importante recalcar que dentro de los alcances del trabajo no está contemplado reemplazar las reglas LINCE, pues estas

corresponden a 52 reglas que han sido pulidas por el área de riesgo de la empresa, las cuales consideran además muchos otros factores que no son aplicables a un modelo de Scoring, como procesos internos de la empresa, plazos en las facturas y diferentes alertas de la plataforma de la empresa. Además, de acuerdo con un análisis realizado por el memorista, del total de documentos aceptados automáticamente por LINCE, solo un 1,89% ha caído en incumplimiento. Se espera que el modelo desarrollado se utilice en conjunto con las reglas LINCE, como un segundo filtro posterior a las reglas LINCE utilizado por el área de operaciones.

Utilizando un modelo probabilístico se estimará la probabilidad de default o incumplimiento en el pago de un documento. Definido por el evento de no-pago en el portafolio en cuestión, el que se materializa transcurridos 90 días de atraso o en una reestructuración forzosa, dentro de un horizonte de un año.

Se espera que la construcción e implementación de un prototipo inicial de un modelo de Scoring dentro de la empresa se traduzca en una disminución significativa de la pérdida esperada y por ende en una futura disminución de las pérdidas incurridas por la empresa en relación con periodos anteriores. Todo esto derivará en una disminución de las reservas que la empresa mantiene en forma permanente para cubrir pérdidas futuras, es decir, en una disminución de los porcentajes a provisionar definidos por la empresa.

Otro beneficio importante que puede derivarse de la implementación de un modelo de Scoring es la reducción significativa de los costos de tiempo y económicos de la empresa. Al automatizar los procesos de evaluación de documentos o facturas realizados por el área de riesgo y operaciones, la empresa puede agilizar el proceso de toma de decisiones y reducir los costos de personal y los errores humanos. En consecuencia, la empresa puede obtener una mayor eficiencia operativa y una mejor gestión de los recursos.

Es importante destacar que un modelo de Scoring establece dos umbrales, uno para aceptación y otro para rechazo, con el objetivo de reducir al mínimo el área que queda entre ambos umbrales, que representa la zona evaluada por el comité de riesgo. La decisión sobre qué porcentaje de documentos será abarcada por esta área dependerá del apetito por riesgo tanto del patrocinador como de la entidad financiera involucrada.

Es importante considerar que los modelos de Scoring se basan en las condiciones actuales del mercado y la economía. Si se producen cambios significativos en la economía o el mercado, el modelo puede no ser capaz de predecir el comportamiento futuro de los clientes de manera precisa. Además, este tipo de modelos presenta limitaciones ante eventos imprevistos, como desastres naturales o cambios en la política gubernamental, que pueden afectar la capacidad de pago de los clientes y que no se pueden prever a través de los datos históricos.

2. Marco teórico

Los modelos de Scoring, introducidos en los años 70’, pero generalizados en los años 90’ gracias al avance existente en cuanto a recursos estadísticos y computacionales, son hoy en día una herramienta utilizada por la gran mayoría de las entidades financieras

para estimar la probabilidad de default o no pago, ordenando a los clientes y deudores solicitantes de financiamiento en función de su probabilidad de incumplimiento. No obstante, actualmente la utilización de los modelos Credit Scoring aún va de la mano con el juicio humano al momento de decidir si entregar o no financiamiento a diferentes entidades, comúnmente siguiendo un conjunto de reglas ya predefinidas por la entidad financiera utilizadas para filtrar las solicitudes de operación de clientes y deudores.

El objetivo de este modelo es obtener una estimación de la probabilidad de incumplimiento del cliente o deudor por probabilidad de default (PD) asociada a su score o clasificación mediante modelos estocásticos, o también en función de su tasa de default (TD) histórica observada en el grupo de clientes y deudores con la misma clasificación o score.

El modelo de Scoring plantea que la variable dependiente toma valores discretos, más en específico se trata como una variable binaria que toma los valores 0 y 1, que corresponde a la ocurrencia de un fenómeno binario que hace referencia a si el cliente o deudor cae o no en default. Presentando la variable dependiente definida como Y_i con $i = 1 \dots n$ donde i corresponde al cliente o deudor en cuestión. Y_i puede tomar el valor 1 si el cliente o deudor paga correctamente el monto anticipado de las facturas y 0 si el deudor entra en mora, o bien se podría definir 0 si entra en castigo. Además, se define X_i con $i = 1 \dots m$ al conjunto de características relevantes del cliente o deudor i que expliquen el valor de Y_i.

Se tiene así que el modelo de Credit Scoring, resuelve la probabilidad de ocurrencia de un evento Y_i condicional al conjunto de características X_i del cliente o deudor.

$$P_I = P(Y_i = 1|X_i)$$

Al tratarse Y_i de una variable binaria, se tiene que su esperanza condicional es la probabilidad condicional de ocurrencia del evento [6].

Es importante definir además como varía la probabilidad de incumplimiento estimada, al cambiar las diferentes variables macroeconómicas del país y las características de las políticas de crédito que establece la institución financiera (variables sistemáticas). Para esto se analizan dos tipos de clasificación de riesgo: PIT (Point in the time) o TTC (Through the cycle), los cuales consideran estas características definidas y además las características propias de los clientes y deudores que acuden a la institución financiera (variables idiosincráticas).

El sistema de clasificación PIT utiliza toda la información disponible de los clientes y deudores para asignarlos a grupos de riesgo. Este sistema de clasificación utiliza tanto variables sistémicas como idiosincráticas, variando el comportamiento de la PD (Probabilidad de default) de acuerdo con las fluctuaciones macroeconómicas, por lo que este sistema de clasificación tiende a ajustarse rápidamente a un entorno económico cambiante.

Por otro lado, el sistema de clasificación TTC solo utiliza variables idiosincráticas considerando las características estáticas y dinámicas de los clientes y deudores, por lo

que el comportamiento de la PD no depende estrictamente de características macroeconómicas del entorno, ni de políticas de crédito, manteniéndose de esta manera constante aun cuando las condiciones macroeconómicas del país varían en el tiempo.

Es importante considerar además que existen sistemas de clasificación híbridos que mezclan características de los sistemas de clasificación PIT y TTC.

Considerando un periodo de estabilidad económica, que el horizonte temporal es de corto plazo, y que el trabajo se desarrolla dentro de una empresa de factoring, donde se busca asegurar el pago por parte de las empresas independiente de las diferentes condiciones económicas, el sistema de clasificación que se utilizará para el trabajo de título corresponde al sistema Through the cycle (TTC). Cabe mencionar que este tipo de clasificación va a castigar aquellas empresas que han tenido un alto riesgo crediticio, pues analiza la calidad crediticia de esta en diferentes ciclos económicos.

Un concepto clave a considerar son las provisiones por riesgo de crédito, las cuales se definen como un resguardo que poseen las entidades financieras colocadoras permanentemente en el tiempo, con el objetivo de cubrir pérdidas futuras.

En línea con los estándares internacionales de regulación basada en riesgos, promovidos por el Comité de Basilea de Supervisión Bancaria (BCBS) y el Financial Stability Board (FSB), las provisiones por riesgo de crédito deben determinarse en base a:

1. La estimación de las pérdidas esperadas, según el enfoque de distribución de pérdidas incurridas y latentes.

2. La filosofía de clasificación de riesgo Through-the-cycle (TTC), que se funda en estimaciones de largo plazo para los factores de riesgo [7]**.**

Siguiendo el enfoque de distribución de pérdidas, la Pérdida Esperada (PE) puede ser expresada a través de la fórmula:

$$PE = PD * LGD * EAD$$

Donde PD corresponde a la probabilidad de default o incumplimiento del cliente o deudor, definido por el evento de no-pago en el portafolio en cuestión, el que se materializa transcurridos 90 días de atraso o en una reestructuración forzosa, disminuyendo la obligación o postergando el pago del principal o los intereses dentro del horizonte de un año. La LGD corresponde a la pérdida dado el incumplimiento, se define como el porcentaje sobre la exposición en riesgo que no se espera recuperar en caso de incumplimiento [7]. La EAD corresponde a la exposición al incumplimiento, definida como el importe de deuda pendiente de pago en el momento del incumplimiento del cliente [7].

Actualmente Chita SPA, se rige por los siguientes porcentajes a provisionar de acuerdo con el rango de días de mora que alcanza un documento operado.

Rango temporal	Pérdida Esperada
0 días	0,48%
1-30 días	0,98%
31-60 días	5,11%
61-90 días	23,38%
91-120 días	49,63%
121-180 días	72,27%
181-240 días	72,27%
241-300 días	86,71%
301-359 días	100,00%

Tabla 1: Matriz pérdida esperada actual

3. Metodología

3.1 Proceso KDD

Para el desarrollo del modelo de Credit Scoring se seguirá el proceso KDD (Knowledge Discovery on Databases) que corresponde a un proceso automático que consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice [8]. El proceso KDD consta de diferentes etapas que involucran numerosos pasos con la intervención del usuario en la toma de muchas decisiones, estas etapas son:

- **Selección:** En esta etapa, se crea un conjunto de datos objetivos de acuerdo con las metas definidas al momento de utilizar del proceso KDD, seleccionando todo el conjunto de datos o una muestra representativa de estos, sobre los que se realizará el proceso de descubrimiento.

- **Preprocesamiento/limpieza:** etapa también denominada data cleaning, en esta se analiza la calidad de los datos, se remueven aquellos datos que pueden generar ruido al momento de análisis, se seleccionan estrategias para el manejo de datos desconocidos, datos duplicados y nulos, es decir, se manipulan y transforman los datos en bruto, de manera que la información contenida en el conjunto de datos pueda ser descubierta.

- **Transformación/reducción:** En esta etapa del proceso se eligen características útiles para representar los datos dependiendo de la meta del proceso, se utilizan técnicas de reducción como agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía, muestreo, entre otras [6].

- **Data Mining (minería de datos):** El uso de técnicas de minería de datos tiene como objetivo crear modelos que son predictivos o descriptivos. Los modelos predictivos buscan estimar valores futuros de variables de interés, denominadas variables objetivos o dependientes, usando variables independientes o predictivas, acá podemos encontrar los subprocesos de regresión y clasificación. En cambio,

los modelos descriptivos buscan identificar patrones que explican o resumen los datos, estudian las propiedades de los datos analizados. Por lo tanto, la escogencia de un algoritmo de minería de datos incluye la selección de los métodos por aplicar en la búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos (categóricos, numéricos) por utilizar [6].

- **Interpretación/evaluación:** En esta etapa se visualizan los patrones de datos extraídos y se interpretan de acuerdo con el objetivo del proceso impuesto, se realizan distintas pruebas como análisis de sensibilidad y validación con distintas muestras para probar la robustez del modelo.

  El cumplimento de los pasos del proceso KDD permite llegar a modelos más robustos y evita incurrir en errores de modelación, por ende, será utilizado como guía para el desarrollo de este proyecto. En la ilustración 4 se presenta el diagrama con las etapas del proceso KDD [9].

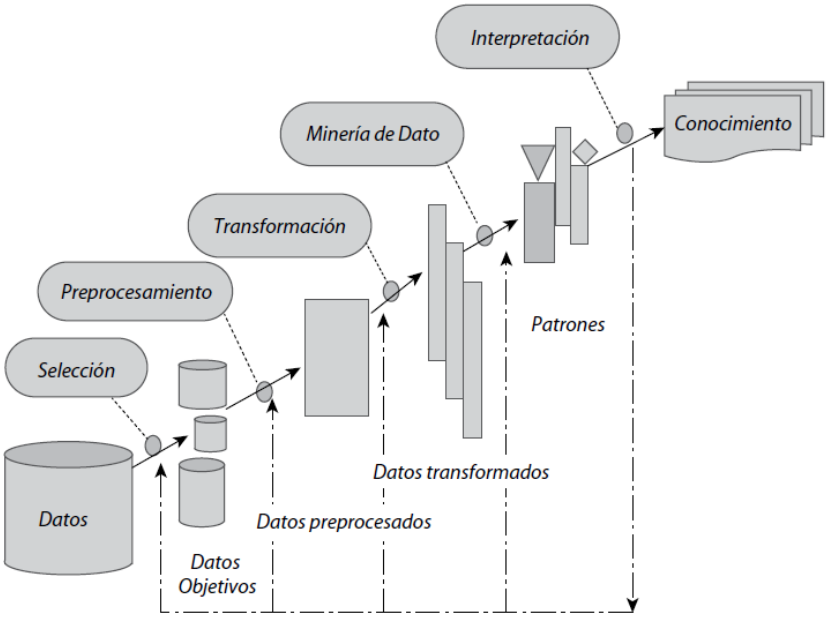

Ilustración 2: Etapas del proceso KDD [6]

3.2 Metodología propuesta en base a proceso KDD

3.2.1 Selección de datos

Siguiendo las etapas del proceso KDD, lo primero será seleccionar el conjunto de datos a utilizar, para esto las fuentes desde donde se extraerán los datos serán de distinta naturaleza:

- **Bases de datos internas:** Se extraerán las bases de datos desde el sitio web de la empresa Chita.cl, donde es posible acceder a una base histórica de los clientes y deudores que han operado con la empresa.

- **Bases de datos externas:** Se extraerán bases de datos desde fuentes externas o ajenas a la entidad financiera, como es el caso del Servicio de Impuestos Internos**.**

3.2.2 Procesamiento y limpieza

Una vez obtenidos todos los datos a utilizar desde fuentes tanto internas como externas, el siguiente paso consiste en la limpieza de dichos datos y la selección de variables. La limpieza de datos involucra el llenado de datos faltantes, suavizar los errores de los datos, corregir datos inconsistentes, y eliminar redundancia entre datos causada por la integración de estos.

Para el manejo de los datos faltantes, se puede utilizar el algoritmo K-Medias, el cual corresponde a un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos por agrupar. K-Medias es uno de los algoritmos de clustering más utilizados. La “K” se refiere al hecho de que el algoritmo funciona para un número fijo de clústeres, los cuales son definidos en términos de la proximidad entre los puntos de datos.

Una extensión de este algoritmo es el denominado K-Modas en el cual se sustituye la media por la moda, para aplicarlo a datos categóricos, ya que K-Medias está orientado a datos numéricos [10].

Para el manejo de los datos con ruidos se puede utilizar el método Binning, en el cual una serie de valores ordenados se agrupan en porciones y luego se suaviza cada porción. De esta forma, lo que se hace es un tratamiento local del ruido ya que se actúa de manera individual en cada porción [10].

3.2.3 Transformación de datos

La transformación de datos involucra:

- **Normalización:** los atributos de la base de datos son escalados dentro de un rango pequeño de valores como -1 y 1 o entre 0 y 1.

- **Suavizado:** Remover el ruido de los datos.

- **Agregación**: operaciones de síntesis u operación son agregadas a los datos**.**

- **Generalización:** Datos de bajo nivel son reemplazados por conceptos de un mayor nivel, haciendo uso del concepto de jerarquía.

La reducción de datos es utilizada para obtener una representación reducida de los datos manteniendo la integridad de los datos originales. Para esto se plantea la siguiente estrategia de reducción de datos:

- **Reducción de dimensión:** Atributos o dimensiones poco relevantes o redundantes son detectados y eliminados. Para la selección de atributos se pueden utilizar métodos heurísticos básicos como la selección hacia atrás, eliminación hacia atrás y la inducción por árbol de decisión.

3.2.4 Data Mining

Para la etapa de minería de datos, existe una gran variedad de metodologías disponibles como lo son el análisis discriminante, modelos de regresión lineal, modelos de regresión logística, modelos probit, modelos logit, multilogit, métodos no paramétricos de suavizado, métodos de programación matemática, modelos basados en cadenas de Markov, algoritmos genéticos y redes neuronales.

Desde un enfoque econométrico, los modelos de regresión lineal presentan varias desventajas técnicas, por lo que han caído en desuso, mientras que por otro lado los modelos probit, logit, y la regresión logística son superiores al análisis discriminante ya que entregan para cada cliente y deudor una probabilidad de default mientras que el análisis discriminante sólo clasifica a estas entidades en grupos de riesgo. Por su parte, los modelos no paramétricos y de inteligencia artificial como las redes neuronales y los algoritmos genéticos son poco intuitivos y de difícil implementación.

- **Modelo Logit**: La regresión Logit se utiliza para predecir un resultado binario. Esta regresión binaria es un tipo de análisis de regresión donde la variable dependiente es una variable dummy: código 0 (Buen Cliente) o 1 (Mal Cliente) (Fernández Castaño, Horacio; Pérez Ramírez, Fredy Ocaris, 2005). En esta Regresión se relaciona la variable dependiente con las variables independientes $\boldsymbol{X_1}$,…., $\boldsymbol{X_i}$ a través de la siguiente ecuación:

$$\boldsymbol{Y_i} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{exp}\,(-\boldsymbol{z})} + \boldsymbol{u_i}$$

Y_i = Variable independiente.
$\boldsymbol{z}$ = Scoring logístico. $\boldsymbol{z} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} + \ldots\ldots + \boldsymbol{\beta_k X_k}$
$\boldsymbol{u_i}$ = Variable aleatoria que se distribuye normalmente N (0,σ^2)

De esta manera la probabilidad de default está dada por:

$$\boldsymbol{P}(\boldsymbol{Y} = \mathbf{1}|\boldsymbol{X}) = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{exp}\,(-(\boldsymbol{\beta_0} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{\beta_i x_i}))}$$

Donde los β_0 … … . . $\boldsymbol{\beta_k}$ son los parámetros estimados a partir de los datos [11].

- **Modelo multinomial:** El modelo de regresión multinomial es muy similar al modelo de regresión Logit, pero la diferencia radica en que la variable dependiente $\boldsymbol{Y_i}$ no se encuentra restringida a solamente dos categorías.

- **Modelo Probit:** El modelo Probit es un modelo estadístico utilizado para predecir variables binarias o dicotómicas, al igual que el modelo de regresión Logit. Ambos

modelos son métodos comúnmente utilizados en el análisis de regresión para estimar la probabilidad de ocurrencia de un evento o resultado binario. La principal diferencia entre el modelo Probit y el modelo de regresión logit radica en la función de distribución utilizada para modelar la relación entre las variables predictoras y la variable objetivo. Mientras que el modelo de regresión logit se basa en la función de distribución logística, el modelo Probit se basa en la función de distribución normal acumulada.

En este contexto, para el propósito de este trabajo de memoria, se utilizará el modelo de regresión logit. Este modelo estadístico es ampliamente utilizado para predecir variables binarias o categorías dicotómicas, como es el caso de la probabilidad de incumplimiento.

Para la selección de variables a utilizar en la creación del modelo, en primer lugar, al permitir este la admisión o el rechazo de documentos a operar, es necesario descartar de la base de datos inicial todos aquellos campos que correspondan a características y comportamientos del documento ya ingresado al sistema de la empresa, como, por ejemplo, el gestionador, el monto anticipado y la diferencia de precio, pues estas características del documento no influyen en la probabilidad de que este caiga en default, pues corresponde a variables que se le atribuyen una vez el documento ya fue aceptado dentro de la empresa. En una primera iteración se seleccionan variables que describen al cliente, deudor y a los documentos que buscan operar, como, por ejemplo: tamaño del cliente, tamaño del deudor, antigüedad SII cliente, antigüedad SII deudor, localidad del cliente, localidad del deudor, monto documento, tasa documento.

Es importante considerar que esta primera iteración de selección de variables se evaluará en el software R-Studio, donde a partir de los betas obtenidos de los modelos se evaluará la significancia e impacto que tiene cada una de las variables definidas en la probabilidad de incumplimiento o default (PD) de un documento.

3.2.5 Interpretación y evaluación

Se interpretarán los valores obtenidos de los modelos logit y probit desde donde se obtendrá una variable Z_i que represente el score estimado para cada cliente y deudor, es decir, una métrica de su calidad crediticia a partir de los parámetros estimados y de su propia información. Este score Z_i obtenido, aplicado a las funciones de distribución de probabilidades acumuladas normal o logística, permite conocer la probabilidad de incumplimiento o default y por ende el riesgo que trae consigo cada cliente y deudor a la empresa.

3.3 Métricas de evaluación de modelos de clasificación

Las métricas de evaluación de modelos de clasificación explican el rendimiento de un modelo. Con el objetivo de determinar si los resultados predichos por el modelo coinciden con los resultados reales, existen diferentes técnicas de evaluación de modelos de clasificación.

3.3.1 Matriz de confusión

Una matriz de confusión es una representación matricial de los resultados de las predicciones de cualquier prueba binaria que se utiliza a menudo para describir el rendimiento del modelo de clasificación sobre un conjunto de datos de prueba cuyos valores reales se conocen. Cada predicción puede clasificarse dentro de uno de los cuatro siguientes resultados presentados por la matriz, basándose en su coincidencia con el resultado real.

- **True Negatives (TN):** valores que el modelo predice como negativos y efectivamente son negativos.
- **True Positives (TP):** valores que el modelo predice como positivos y efectivamente son positivos.
- **False Negatives (FN):** valores que el modelo predice como negativos, pero son positivos.
- **False Positives (FP):** valores que el modelo predice como positivo, pero son negativos.

A partir de la matriz de confusión se pueden obtener las siguientes métricas:

- **Accuracy**: medida que muestra la suma de predicciones correctas sobre el total de predicciones

  ***Accuracy* = (TP + TN) /(TP + TN + FP + FN)**

- **Tasa de error:** corresponde al número de predicciones incorrectas, sobre el total de predicciones.

  ***Error Rate* = 1 – Accuracy = (FP + FN) /(TP + TN + FP + FN)**

- **Sensibilidad:** corresponde al porcentaje de observaciones positivas (TP + FN), que fueron clasificadas como positivas por el modelo (TP). Un modelo con alta sensibilidad es un modelo con pocos falsos negativos, es decir, un modelo que clasifica correctamente la mayoría de los casos efectivamente positivos.

  ***Sensibilidad* = TP / (TP + FN)**

3.3.2 Curva ROC

La curva ROC (Receiver Operating Characteristic) corresponde a un gráfico muy utilizado en la evaluación de modelos de machine learning. Esta gráfica se construye graficando

el ratio de verdaderos positivos (recall o sensibilidad) sobre el ratio de falsos positivos, para distintos umbrales de clasificación, permitiendo así visualizar la capacidad de un modelo de clasificación en distintos umbrales de discriminación. Los algoritmos de clasificación entregan una probabilidad, es decir, un valor continuo entre 0 y 1, se define como umbral de clasificación al valor a partir del cual se considera la predicción como positiva o negativa. Generalmente se utiliza el valor 0.5.

3.3.3 ROC AUC

La métrica AUC (Area Under the Curve), mide el área bidimensional bajo la curva ROC. El AUC proporciona una medida agregada del rendimiento en todos los umbrales de clasificación posibles. Una forma de interpretar el AUC es como la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio. El AUC varía en valor de 0 a 1, de esta forma, un modelo cuyas predicciones son un 100% incorrectas tiene un AUC de 0, en cambio uno cuyas predicciones son un 100% correctas tiene un AUC de 1.

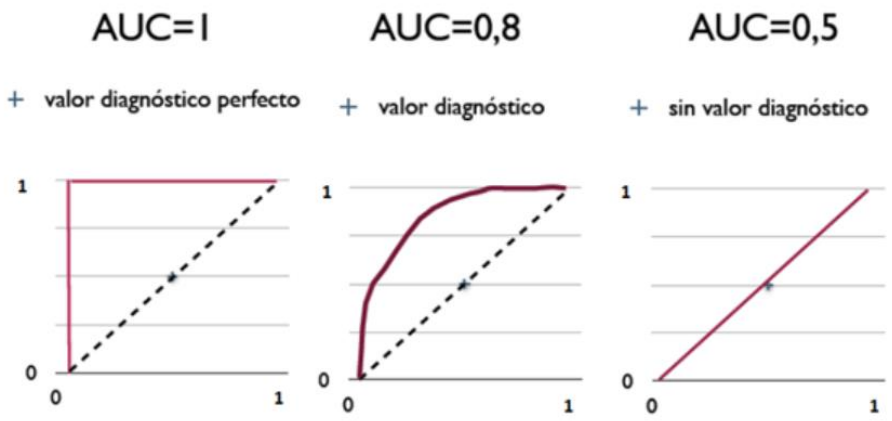

Ilustración 3: Visualizaciones posibles valores curva AUC.

Como se aprecia en la ilustración 3, cuando el AUC es igual a 1, el modelo predice perfectamente entre positivos y negativos. Por otro lado, cuando el AUC es igual aproximadamente 0,5, significa que el modelo carece de capacidad de discriminación para distinguir entre resultados positivos y negativos [12].

Para la evaluación de los modelos de Scoring, se espera que el AUC sea igual o superior a 0.75, es decir, que el modelo de clasificación posea al menos un 75% de probabilidad de discriminar correctamente entre positivos y negativos, esto considerando que la empresa opera principalmente con PYMES, las cuales tiene asociadas un alto riesgo.

3.4 Selección de puntajes de corte

Una vez que se ha calculado el puntaje que contribuirá cada variable seleccionada en la construcción del modelo de Scoring, se debe establecer el umbral de puntaje o score a

partir del cual se considerará que una factura o documento es aceptable o rechazable. En otras palabras, se debe definir el puntaje mínimo requerido para aprobar una factura o documento y el puntaje máximo permitido para rechazarla. Esta es una etapa crucial en el proceso de Scoring, ya que determinará la eficacia del modelo en la identificación de facturas o documentos que podrían incurrir en incumplimiento. Existen varias metodologías para definir los puntajes de corte, entre ellas destacan:

- **Costo-beneficio:** Este método implica considerar los costos y beneficios asociados a los falsos positivos (FP) y los falsos negativos (FN) del modelo. A partir de esto, se busca establecer un puntaje de corte que maximice los beneficios y minimice los costos asociados a los errores de clasificación.

- **Distribución de puntajes:** Este método implica analizar la distribución de los puntajes de las facturas o documentos en el muestreo de entrenamiento del modelo y seleccionar los umbrales de corte de acuerdo con la esperanza de no pago o perdida asociada.

- **Curva ROC:** Este método implica el análisis de la curva ROC del modelo realizado, esta se construye considerando el ratio de verdaderos positivos (recall o sensibilidad) sobre el ratio de falsos positivos, para distintos umbrales de clasificación, permitiendo así visualizar la capacidad de un modelo de clasificación en distintos umbrales de discriminación. A partir de esto, se busca encontrar el umbral de discriminación que maximice el área bajo la curva ROC (ROC AUC)

- **Sensibilidad:** Este método busca establecer el puntaje de corte de tal manera que se maximice tanto la sensibilidad (la capacidad del modelo para identificar los casos positivos) como la especificidad (la capacidad del modelo para identificar los casos negativos).

4. Selección de datos

4.1 Datos iniciales y estudio de variables a utilizar

En el proceso de desarrollo del trabajo se emplearán las bases de datos de la empresa, las cuales ofrecen una amplia variedad de tablas que contienen diversas características de las operaciones, clientes, deudores y documentos. Se ha elegido la tabla principal "Documents", la cual abarca todas las facturas y documentos gestionados por la empresa hasta la fecha. Además, se realizan consultas cruzadas con otras tablas de la base de datos con el fin de obtener información adicional relevante asociada al documento en cuestión, así como datos pertinentes del cliente y del deudor correspondiente. Para esto también se utiliza la base “Nómina de empresas personas jurídicas” actualizada cada año por el Servicio de Impuestos Internos

En el “Anexo B” se presenta un resumen de todas las variables presentes en la tabla “Documents”.

4.2 Creación de variable objetivo

Una vez obtenido lo anterior, se procede a la creación de la variable objetivo definida como “Default” o “Incumplimiento”, la cual se define de la siguiente manera:

$$SI\ "Días\ de\ mora"\ \geq\ 90\ \ ------>\ "Default"\ =\ 1$$

Por el contrario:

$$SI\ "Días\ de\ mora"\ <\ 90\ \ ------>\ "Default"\ =\ 0$$

Se define la variable objetivo “Default” como una variable binaria que toma el valor 1 cuando un documento ha alcanzado los 90 o más días de mora, y 0 cuando no. El tiempo que se considera como incumplimiento de pago depende de los términos y condiciones acordados entre las partes. En general, las empresas establecen una política de pago de mora que incluye un plazo para el pago de facturas, el cual suele ser de 30, 60 o 90 días a partir de la fecha de vencimiento. Para este trabajo se ha establecido que el incumplimiento en los pagos se considerará a partir de los 90 días de mora, pues la empresa tiene como objetivo minimizar en la medida de lo posible la cantidad de morosidad, ya que el modelo de riesgo actual de la empresa contempla altos porcentajes de provisiones por tramos de mora.

Es importante mencionar que se utilizará la base de datos correspondiente al cierre del mes marzo del 2023 para el análisis, ya que se requiere un período de al menos 90 días posteriores a la fecha de vencimiento de los documentos para determinar si se consideran en estado de default.

Como resultado, la base de datos actual cuenta con un total de 141.803 documentos operados hasta la fecha. No obstante, es crucial excluir los documentos que se encuentran en estado "Vigente" y "Giro pendiente", ya que corresponde a aquellos documentos que han sido gestionados recientemente y por lo tanto aún no han alcanzado su fecha de expiración.

Después de analizar la variable objetivo establecida, se tiene que un porcentaje del 4,44% del total de documentos operados hasta la fecha definida anteriormente, corresponde a documentos que han presentado una morosidad igual o superior a los 90 días.

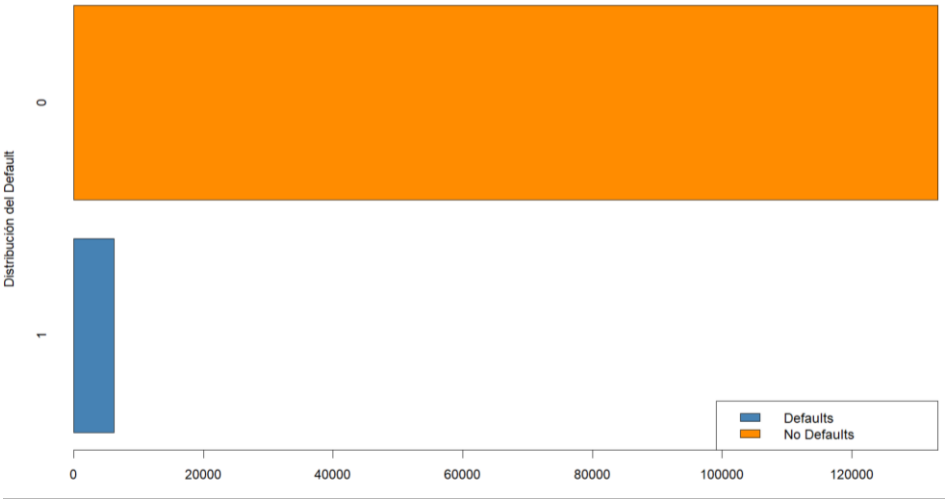

Gráfico 5: Porcentaje de documentos definidos en “Default”

5. Selección, limpieza y transformación de las variables

5.1 Selección de variables

Una vez se ha realizado el análisis y se han identificado los documentos que pertenecen a esta variable “Default”, es importante contar con información adicional para la creación del modelo de Scoring. Por esta razón, se decidió a traer a la base de datos todos los datos relevantes en relación con el cliente y al deudor asociados a cada documento, a través del cruce con otras bases de la empresa y del servicio de impuestos internos.

Esto incluye información como el historial crediticio del cliente y del deudor, la relación comercial entre las partes, la antigüedad de la deuda, el monto adeudado, entre otros datos relevantes que permiten tener una visión más completa de la situación financiera de cada cliente y deudor.

Con esta información, se podrán llevar a cabo análisis más profundos y detallados que permitan tomar decisiones informadas y adecuadas para minimizar los riesgos asociados con los documentos en mora. Asimismo, la información obtenida puede ser utilizada para establecer políticas de crédito y cobranza más eficientes y efectivas en el futuro.

A continuación, se presentan de manera detallada todas las variables que han sido incorporadas en la base de datos generada, las cuales serán consideradas en la primera iteración del análisis.

Variables asociadas a las empresas, tanto para clientes como deudores**:**

- **Tamaño:** Corresponde al tamaño de la empresa en ventas anuales agrupado en dígitos del 1 al 13, donde el valor 1 corresponde a empresas sin información, el 2, 3 y 4 corresponden a microempresas, 5, 6 y 7 a pequeñas empresas, 8 y 9 a

medianas empresas y 10, 11, 12 13 a grandes empresas, de acuerdo a los datos entregados por el SII.[14]

- **Zona:** Corresponde a la zona donde se encuentra ubicada la empresa, agrupado por región.

- **Sector de actividad económica:** Representa el sector al cual pertenece la principal actividad económica realizada por la empresa

- **Tipo empresa:** Corresponde al tipo de propiedad de la empresa, y puede ser de dos tipos: estatal o privada.

- **Tipo persona:** Corresponde al tipo de entidad a la que pertenece una persona, y puede ser de dos tipos: natural o jurídica.

- **Historial de facturas total operadas:** Corresponde al total de documentos históricos que ha operado la empresa en Chita.

- **Historial de facturas operadas en los últimos 12 meses:** Corresponde al total de documentos que ha operado la empresa en Chita en los últimos 12 meses.

- **Historial de facturas operadas en los últimos 6 meses:** Corresponde al total de documentos que ha operado la empresa en Chita en los últimos 6 meses.

- **Historial de facturas operadas en los últimos 3 meses:** Corresponde al total de documentos que ha operado la empresa en Chita en los últimos 3 meses.

- **Historial de facturas saldadas:** Corresponde al total de documentos históricos que la empresa ha saldado en Chita.

- **Historial de facturas saldadas en los últimos 12 meses:** Corresponde al total de documentos que la empresa ha saldado en Chita en los últimos 12 meses.

- **Historial de facturas saldadas en los últimos 6 meses:** Corresponde al total de documentos que la empresa ha saldado en Chita en los últimos 6 meses.

- **Historial de facturas saldadas en los últimos 3 meses:** Corresponde al total de documentos que la empresa ha saldado en Chita en los últimos 3 meses.

- **Antigüedad SII:** Corresponde a la antigüedad o inicio de actividades de la empresa en el registro del Servicio de Impuestos Internos.

- **Antigüedad cliente en Chita:** Corresponde a la antigüedad del cliente dentro de la empresa, obtenido a través de su primera operación.

Variables asociadas al documento:

- **Monto:** Corresponde al monto del documento operado.

- **Tasa del Documento:** Corresponde a la tasa acordada que tendrá el documento al operar**.**

- **Porcentaje anticipado:** Corresponde al porcentaje del monto del documento operado que le será anticipado al cliente, generalmente superior al 90%.

5.2 Limpieza de variables

El tratamiento de valores atípicos o outliers es clave para garantizar la integridad, precisión y validez de los análisis de datos, en especial para la creación de un modelo de Scoring, los cuales asumen ciertas distribuciones y supuestos sobre los datos. Los valores atípicos pueden violar estos supuestos y afectar la precisión y la confiabilidad del modelo. El tratamiento de los outliers ayuda a garantizar que los modelos se ajusten de manera adecuada y sean representativos de los datos subyacentes.

Se llevó a cabo el tratamiento y análisis de los valores nulos presentes en cada una de las variables explicativas, detallado en el “Anexo C”. Con el objetivo de fortalecer el análisis y el desarrollo del modelo, y dado el bajo porcentaje de missing values presentes en la base de datos, las filas que contenían estos valores nulos fueron eliminadas de los datos. Este procedimiento asegura la integridad y la confiabilidad de los resultados al trabajar con datos completos, lo que contribuye a una mayor precisión y fiabilidad en el análisis del modelo de Scoring propuesto.

5.3 Análisis de correlación entre las variables

En el “Anexo D” se observa la matriz de correlación de las variables seleccionadas, se aprecia que la correlación entre la cantidad de documentos operados y la cantidad de documentos saldados es bastante cercana a 1 tanto para deudores, como para clientes, es por eso por lo que se decidió trabajar estas variables como una ratio de facturas saldadas, el cual se explica en la **Sección 6.1.2.** Por otro lado, las demás variables no presentan en su mayoría una gran correlación, por lo que inicialmente no es necesaria su eliminación del modelo.

5.4 Análisis exploratorio

El análisis exploratorio de datos o EDA (Exploratory Data Analysis) es una fase crucial en el proceso de análisis de datos que consiste en examinar y comprender la estructura y características de los datos antes de aplicar técnicas estadísticas más avanzadas o modelos predictivos. Esta etapa se realiza con el objetivo de obtener una visión general de los datos, identificar patrones, tendencias, relaciones y anomalías que puedan ser relevantes para la investigación o el problema en cuestión.

Se presenta un EDA de las principales variables continuas a considerar en el desarrollo del modelo.

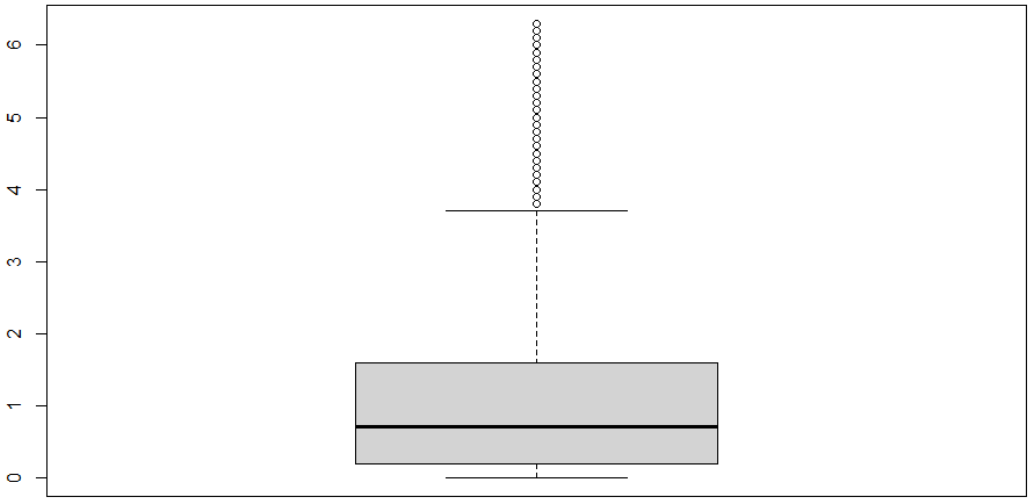

Gráfico 6: Boxplot antigüedad de cliente en la empresa

En el Gráfico 6 se observa que la antigüedad de los clientes que operan con la empresa se concentra mayormente en el rango de 0 a 2 años.

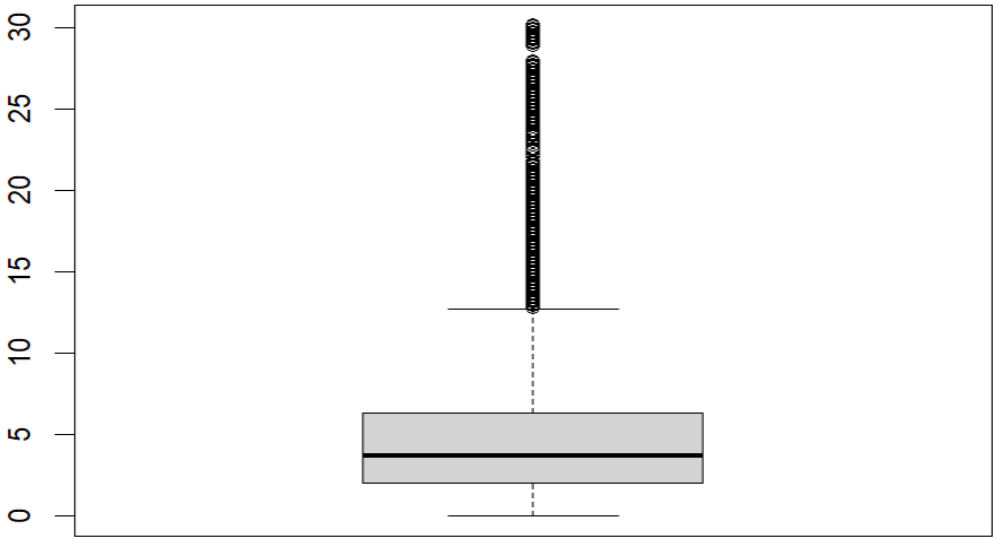

Gráfico 7: Boxplot antigüedad de cliente en el SII

En el Gráfico 7 se observa que los clientes que operan con la empresa presentan una antigüedad en el SII concentrada mayormente entre los 0 a 6 años. Sin embargo, también hay varios clientes que se encuentran entre los 15 y los 30 años de antigüedad en el SII. Por lo que la mayoría corresponden a PYMES que no han superado el llamado “Valle de la muerte”.

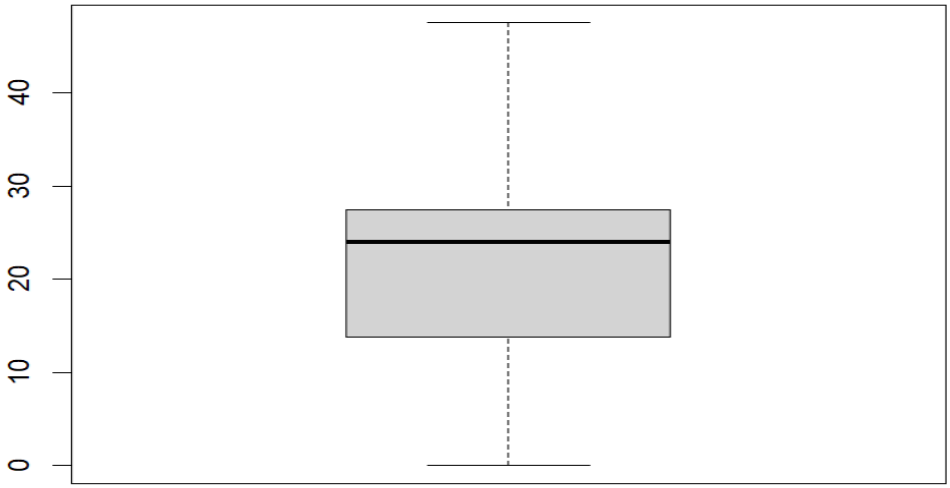

Gráfico 8: Boxplot antigüedad del deudor en el SII

En el Gráfico 8 se observa que los deudores que operan con la empresa presentan una antigüedad en el SII concentrada mayormente entre los 12 a 30 años.

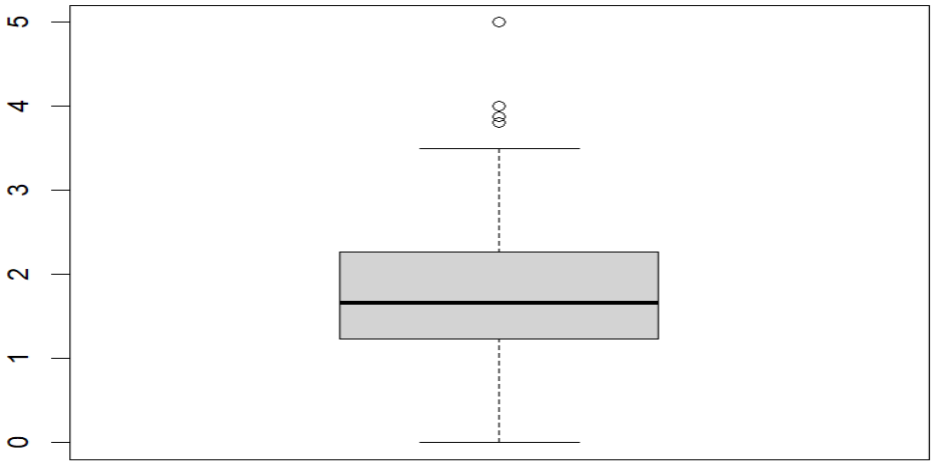

Gráfico 9: Boxplot tasa de documentos

El Gráfico 9 muestra que las tasas de operación (definida también como tasa del documento) impuestas por la empresa de factoring se concentran principalmente en el rango del 1,2% al 2,2%.

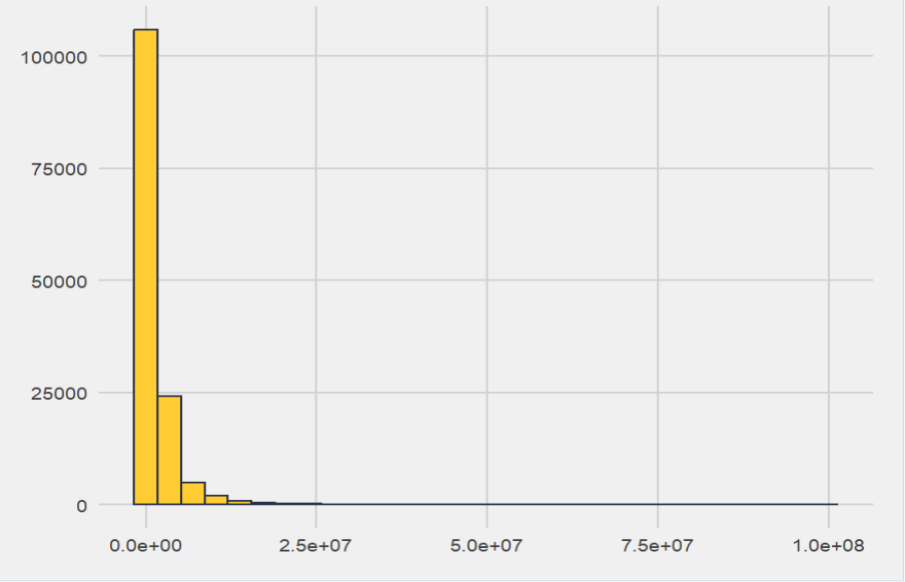

Gráfico 10: Histograma monto de documentos operados

El Gráfico 10 muestra la distribución en monto de los documentos operados por la empresa. El monto promedio por documento es de $1.655.075.

5.5 Transformación de variables

Para el desarrollo de un modelo de Scoring, es necesario realizar una categorización de las variables que se utilizarán. Esto se debe a que se busca asignar un puntaje en función del rango en el que se encuentre cada variable al momento de la solicitud. En este proceso, es importante categorizar las variables continuas y modificar las variables categóricas que presenten una amplia variedad de categorías.

La categorización de variables continuas permite agrupar los valores en rangos o intervalos específicos, lo cual facilita la interpretación y el cálculo del puntaje en el modelo de Scoring. Al mismo tiempo, las variables categóricas que presentan múltiples categorías se pueden simplificar o reducir a fin de evitar problemas como la multicolinealidad y la pérdida de grados de libertad en el modelo.

Al reducir el número de categorías en las variables categóricas, se evita la correlación excesiva entre las variables predictoras y se mejora la capacidad del modelo para estimar coeficientes confiables y significativos, permitiendo una mejor interpretación.

5.5.1 Variables Categóricas

Las variables Tamaño cliente y Tamaño deudor, se encuentran categorizadas a través de un número entre 1 y 13, el cual representa su rango de ventas anuales que define el tamaño de la empresa, el cual se encuentra definido por el Servicio de Impuestos Internos (SII) como se puede apreciar en el “Anexo E”

La variable categórica de tamaño se agrupó de la siguiente manera:

- 1 = Empresa sin información
- [2,4] = Micro empresa
- [5,7] = Pequeña empresa
- [8,9] = Mediana empresa
- [10,13] = Gran empresa

Las variables Zona cliente y Zona deudor originalmente representaban la región geográfica de ubicación de la empresa. Con el fin de utilizarlas en un modelo de análisis, se decidió realizar una categorización de estas variables de la siguiente manera:

- **Zona norte:** Región de Arica y Parinacota, Región de Tarapacá, Región de Antofagasta, III Región de Atacama, IV Región de Coquimbo
- **Zona Centro:** Región de Valparaíso, Región Metropolitana, Región del Libertador General Bernardo O'Higgins, Región de Maule.
- **Zona Sur:** Región de Ñuble, Región del Biobío, Región de La Araucanía, Región de Los Ríos, Región de Los Lagos, Región de Aysén del General Carlos Ibáñez del Campo, Región de Magallanes y de la Antártica Chilena.

Las variables Sector actividad económica cliente y Sector actividad económica deudor, presentaban varias categorías las cuales fueron agrupadas en 3 de acuerdo con el sector económico al que pertenecen:

- **Sector Primario:** Agricultura, ganadería, pesca, explotación de minas y canteras y actividades relacionadas con los recursos naturales.
- **Sector Secundario:** Industrias manufactureras metálicas, Industrias manufactureras no metálicas, construcción, suministros de electricidad, agua y gas, gestión de desechos y descontaminación y evacuación de aguas residuales.
- **Sector Terciario:** Comercio al por mayor y por menor, administración pública y defensa, intermediación financiera, transporte, almacenamiento y comunicaciones, actividades inmobiliarias y de alquiler, enseñanza, actividades profesionales, científicas y técnicas, actividades de alojamiento y comidas, actividades artísticas, de entretenimiento y recreativas y otras actividades de servicios.

Las variables Tipo empresa cliente y Tipo empresa deudor, se encuentran correctamente categorizadas en públicas y privadas.

Las variables Tipo de persona cliente y Tipo de persona deudor, se encuentran correctamente categorizadas en naturales y jurídicas.

5.5.2 Categorización de variables continuas

En una primera iteración, la categorización de las variables continuas se realizó utilizando una metodología de distribución de datos, este tipo de categorización implica agrupar o clasificar a los valores de las variables continuas en función de cómo se distribuyen en la población o en la muestra de datos. Categorizar a través de la distribución de datos considera características de la distribución como la forma, la dispersión y los valores atípicos presentes en los datos.

Las variables asociadas al historial de facturas operadas por el cliente y por el deudor y al historial de facturas saldadas que presenta el cliente y el deudor, ya sea históricas, en los últimos 12 meses, en los últimos 6 meses o en los últimos 3 meses, se agruparon para considerarse como una ratio de facturas saldadas, de la siguiente manera:

1. Ratio de facturas saldadas cliente:

$$\frac{Historial\ de\ documentos\ saldados\ que\ presenta\ el\ cliente}{Historial\ de\ documentos\ operados\ por\ el\ cliente}$$

2. Ratio de facturas saldadas cliente últimos 12 meses:

$$\frac{Historial\ de\ documentos\ saldados\ que\ presenta\ el\ cliente\ en\ los\ últimos\ 12\ meses}{Historial\ de\ documentos\ operados\ por\ el\ cliente\ en\ los\ últimos\ 12\ meses}$$

3. Ratio de facturas saldadas cliente últimos 6 meses:

$$\frac{Historial\ de\ documentos\ saldados\ que\ presenta\ el\ cliente\ en\ los\ últimos\ 6\ meses}{Historial\ de\ documentos\ operados\ por\ el\ cliente\ en\ los\ últimos\ 6\ meses}$$

4. Ratio de facturas saldadas cliente últimos 3 meses:

$$\frac{Historial\ de\ documentos\ saldados\ que\ presenta\ el\ cliente\ en\ los\ últimos\ 3\ meses}{Historial\ de\ documentos\ operados\ por\ el\ cliente\ en\ los\ últimos\ 3\ meses}$$

5. Ratio de facturas saldadas por el deudor:

$$\frac{Historial\ de\ documentos\ saldados\ por\ el\ deudor}{Historial\ de\ documentos\ operados\ que\ presenta\ el\ deudor}$$

6. Ratio de facturas saldadas deudor por los últimos 12 meses:

$$\frac{Historial\ de\ documentos\ saldados\ por\ el\ deudor\ en\ los\ últimos\ 12\ meses}{Historial\ de\ documentos\ operados\ que\ presenta\ el\ deudor\ en\ los\ últimos\ 12\ meses}$$

7. Ratio de facturas saldadas por el deudor últimos 6 meses:

$$\frac{Historial\ de\ documentos\ saldados\ por\ el\ deudor\ en\ los\ últimos\ 6\ \ meses}{Historial\ de\ documentos\ operados\ que\ presenta\ el\ deudor\ en\ los\ últimos\ 6\ \ meses}$$

8. Ratio de facturas saldadas por el deudor últimos 3 meses:

$$\frac{Historial\ de\ documentos\ saldados\ por\ el\ deudor\ en\ los\ últimos\ 3\ \ meses}{Historial\ de\ documentos\ operados\ que\ presenta\ el\ deudor\ en\ los\ últimos\ 3\ \ meses}$$

Inicialmente, la distribución de datos se realizó en cuartiles, lo que implica dividir los datos en cuatro categorías de igual tamaño o proporción. Las variables continuas quedan categorizadas de la siguiente manera:

1. Tasa del documento:

    - Categoría 1: [0, 1.22]
    - Categoría 2: (1.22, 1.66]
    - Categoría 3: (1.66, 2.26]
    - Categoría 4: (2.26, 3.5]

2. Porcentaje anticipado:

    - Categoría 1: [0,97%]
    - Categoría 2: (97%,98%]
    - Categoría 3: (98%,99%]
    - Categoría 4: (99%,100%]

3. Antigüedad del cliente en el SII:

    - Categoría 1: [0 años, 2 años]
    - Categoría 2: (2 años, 3.6 años]
    - Categoría 3: (3.6 años, 6.2 años]
    - Categoría 4: (6.2 años, 30.2 años]

4. Antigüedad del deudor en el SII:

    - Categoría 1: [1.7 años,14.4 años]
    - Categoría 2: (14.4 años, 24.5 años]
    - Categoría 3: (24.5 años, 28.1 años]
    - Categoría 4: (28.1 años, 35.8 años]

5. Antigüedad del cliente en la empresa:

    - Categoría 1: [0 años, 0.2 años]
    - Categoría 2: (0.2 años, 0.7 años]
    - Categoría 3: (0.7 años, 1.6 años]

- Categoría 4: (1.6 años, 5.7 años]

6. Ratio de facturas saldadas que presenta el cliente:

    - Categoría 1: [0%, 25%]
    - Categoría 2: (25%, 50%]
    - Categoría 3: (50%, 75%]
    - Categoría 4: (75%, 100%]

7. Ratio de facturas saldadas que presenta el cliente en los últimos 12 meses:

    - Categoría 1: [0%, 25%]
    - Categoría 2: (25%, 50%]
    - Categoría 3: (50%, 75%]
    - Categoría 4: (75%, 100%]

8. Ratio de facturas saldadas que presenta el cliente en los últimos 6 meses:

    - Categoría 1 : [0%, 25%]
    - Categoría 2: (25%, 50%]
    - Categoría 3: (50%, 75%]
    - Categoría 4: (75%, 100%]

9. Ratio de facturas saldadas que presenta el cliente en los últimos 3 meses:

    - Categoría 1: [0%, 25%]
    - Categoría 2: (25%, 50%]
    - Categoría 3: (50%, 75%]
    - Categoría 4: (75%, 100%]

10. Ratio de facturas saldadas por el deudor en los últimos 12 meses:

    - Categoría 1: [0%, 25%]
    - Categoría 2: (25%, 50%]
    - Categoría 3: (50%, 75%]
    - Categoría 4: (75%, 100%]

11. Ratio de facturas saldadas por el deudor en los últimos 6 meses:

    - Categoría 1: [0%, 25%]
    - Categoría 2: (25%, 50%]
    - Categoría 3: (50%, 75%]
    - Categoría 4: (75%, 100%]

12. Ratio de facturas saldadas por el deudor en los últimos 3 meses:

    - Categoría 1: [0%, 25%]
    - Categoría 2: (25%, 50%]

- Categoría 3: (50%, 75%]
- Categoría 4: (75%, 100%]

Para la categorización de la variable monto se utilizó el algoritmo K-means, el cual corresponde a un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster, usando generalmente la distancia cuadrática.

El algoritmo k-means resuelve un problema de optimización, siendo la función de optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster.[16]

Así, la variable monto queda categorizada de la siguiente manera:

1. Monto:

- Categoría 1: [0, 605.000)
- Categoría 2: [605.500, 1.125.086)
- Categoría 3: [1.125.086,1.817.130)
- Categoría 4: [1.817.130, 2.820.120)
- Categoría 5: [2.820.120, 4.124.908)
- Categoría 6: [4.124.908, 5.808.179)
- Categoría 7: [5.808.179, 7.953.960)
- Categoría 8: [7.953.960, 10.591.000)
- Categoría 9: [10.591.000 o más)

6. Desarrollo del modelo

6.1 Primera Iteración

6.1.1 Selección del modelo

El modelo seleccionado para el desarrollo del modelo de Scoring corresponde a un modelo de regresión logística o modelo logit, ya que corresponden a modelos estadísticos utilizados para predecir variables binarias o categorías dicotómicas, en este caso, la probabilidad de incumplimiento. Dentro de los enfoques econométricos, los modelos de probabilidad lineal han caído en desuso por sus desventajas técnicas, en tanto que los modelos probit, logit y la regresión logística son superiores al análisis discriminante ya que proveen para cada deudor una probabilidad de default, en tanto que este sólo clasifica a los deudores en grupos de riesgo. Por otro lado, modelos de inteligencia artificial como redes neuronales y algoritmos genéticos, a pesar de ser superiores cuando se desconoce la forma de la relación funcional y se presume que no es lineal, son poco intuitivos y de difícil implementación.[6]

Se utilizó un modelo lineal generalizado (GLM de las siglas en inglés de Generalized Linear Models), el cual corresponde a una extensión de los modelos de regresión lineales

que permite utilizar distribuciones no normales de los errores (Binomiales, Poisson y Gamma) y varianzas no constantes. La razón fundamental es que para poder utilizar un modelo de regresión lineal es necesario que la variable de respuesta sea continua, y que cumpla las hipótesis estándar del modelo lineal (datos normales y varianza constante). Si la variable de interés es binaria y se ignora este hecho, al estimar los parámetros utilizando el procedimiento lm () (Regresión lineal) se cometen dos grandes errores:

1. Los valores predichos de la probabilidad podrían estar fuera del intervalo (0, 1).

2. Los intervalos de confianza y los test para evaluar qué variables son realmente significativas están basados en la hipótesis de que los datos vienen de una distribución normal, lo que es incorrecto al trabajar con datos binarios.[17]

6.1.2 Selección de variables

Para el desarrollo del modelo de Scoring, en una primera iteración se consideran todas las variables previamente definidas en la Sección 3.4, las cuales ya han sido previamente transformadas y categorizadas para su posterior análisis. En esta primera iteración se busca evaluar el impacto que presenta cada una de las variables utilizadas en la probabilidad de incumplimiento por parte de las empresas. Durante esta etapa, se realiza un análisis exhaustivo de las variables consideradas, con el fin de determinar cuáles de ellas serán utilizadas en la creación del modelo final. Se busca identificar las variables que presentan una influencia significativa en la predicción del incumplimiento, descartando aquellas que no aporten un valor sustancial al modelo.

Este proceso de evaluación permite seleccionar las variables más relevantes y significativas para la construcción del modelo de Scoring. Una vez finalizada esta primera iteración, se realizarán ajustes y refinamientos a las variables con el objetivo de mejorar la precisión y eficacia del modelo, utilizando solamente aquellas variables que han demostrado ser más predictivas en cuanto al incumplimiento empresarial.

6.1.3 Resultados del modelo

Previo al desarrollo del modelo, se realizó una partición aleatoria de la base de datos considerando un 70% de estos como una base de entrenamiento. La cual se utiliza para ajustar los parámetros del modelo y permitirle aprender los patrones y relaciones en los datos, entrenando de esta manera el modelo. Por otro lado, se consideró un 30% de los datos totales como una base de testeo o prueba, conjunto de datos que se reserva exclusivamente para evaluar el rendimiento del modelo entrenado.

Los betas estimados y calculados para cada variable por el modelo de regresión se observan en la Tabla 2:

Variable	Categoría	Beta estimado
Monto	[0, 605.000)	
	[605.500, 1.125.086)	0,32286
	[1.125.086,1.817.130)	0,27011
	[1.817.130, 2.820.120)	0,57076
	[2.820.120, 4.124.908)	0,45598
	[4.124.908, 5.808.179)	0,59039
	[5.808.179, 7.953.960)	0,26939
	[7.953.960, 10.591.000)	0,1786
	[10.591.000 o más)	-0,04251
Tamaño Cliente	Grande	
	Mediana	-1,55668
	Micro	-0,78399
	Pequeña	-1,09859
	Sin info	-0,82034
Ratio de facturas saldadas por el cliente en los últimos 6 meses	[0%, 25%]	
	(25%, 50%]	-0,45184
	(50%, 75%]	-0,53642
	(75%, 100%]	-1,04595

Tasa del documento	[0, 1.22]	
	(1.22, 1.66]	-0,09968
	(1.66, 2.26]	0,04621
	(2.26, 3.5]	0,19302
Zona cliente	Zona Centro	
	Zona Norte	-0,09397
	Zona Sur	-0,12416
Sector económico cliente	Sector Primario	
	Sector Secundario	0,03616
	Sector Terciario	-0,18167
Antigüedad SII cliente	[0 años, 2 años]	
	(2 años, 3.6 años]	-0,24095
	(3.6 años, 6.2 años]	-0,37696
	(6.2 años, 30.2 años]	-0,33919
Antigüedad Chita Cliente	[0 años, 0.2 años]	
	(0.2 años, 0.7 años]	-0,27372
	(0.7 años, 1.6 años]	-0,53618
	(1.6 años, 5.7 años]	-0,73076
Tipo empresa cliente	Estatal	

	Privada	7,076
Tipo persona cliente	Jurídica	
	Natural	-0,05513
Tamaño deudor	Grande	
	Mediana	0,54694
	Micro	0,63661
	Pequeña	0,76674
	Sin información	0,81812
Ratio de facturas saldadas por el deudor en los últimos 6 meses	[0%, 25%]	
	(25%, 50%]	-0,03723
	(50%, 75%]	-0,52246
	(75%, 100%]	-0,96883
Zona Deudor	Zona Centro	
	Zona Norte	0,15563
	Zona Sur	0,34365
Tipo persona deudor	Jurídica	
	Natural	0,77978
Antigüedad SII deudor	[1.7 años,14.4 años]	

	(14.4 años, 24.5 años]	-0,24575
	(24.5 años, 28.1 años]	0,03584
	(28.1 años, 35.8 años]	0,169
Sector económico deudor	Sector Primario	
	Sector Secundario	0,39521
	Sector Terciario	0,59199
Tipo empresa deudor	Estatal	
	Privada	-0,56453

Tabla 2: Betas calculados por el modelo logit para cada categoría y variable definida

6.1.4 Evaluación de desempeño del modelo

Para la evaluación del desempeño del modelo realizado, se utilizó la curva ROC y el área bajo dicha curva (AUC). La curva ROC es un gráfico que representa la relación entre la tasa de verdaderos positivos (Sensibilidad) y la tasa de falsos positivos (1-Especificidad), definidos en la **Sección 3.3**, en diferentes umbrales de clasificación.

Por otro lado, el AUC (Área Bajo la Curva) es una medida numérica que resume la curva ROC en un solo valor. Representa el área total bajo la curva ROC y proporciona una medida de qué tan bien el modelo puede distinguir entre las dos clases. El AUC varía entre 0 y 1, donde un valor de 0.5 indica un modelo que es puramente aleatorio en sus predicciones, comparable a la probabilidad de un lanzamiento de moneda, y un valor cercano a 1 indica un modelo con una alta capacidad de discriminación.

En general, una curva ROC más cercana a la esquina superior izquierda y un AUC más cercano a 1 indican un mejor rendimiento del modelo en términos de sensibilidad y especificidad. El AUC también puede interpretarse como la probabilidad de que el modelo clasifique correctamente una instancia aleatoria de la clase positiva por encima de una instancia aleatoria de la clase negativa.

En el Gráfico 11 se puede observar la curva ROC obtenida del modelo, junto al valor del AUC.

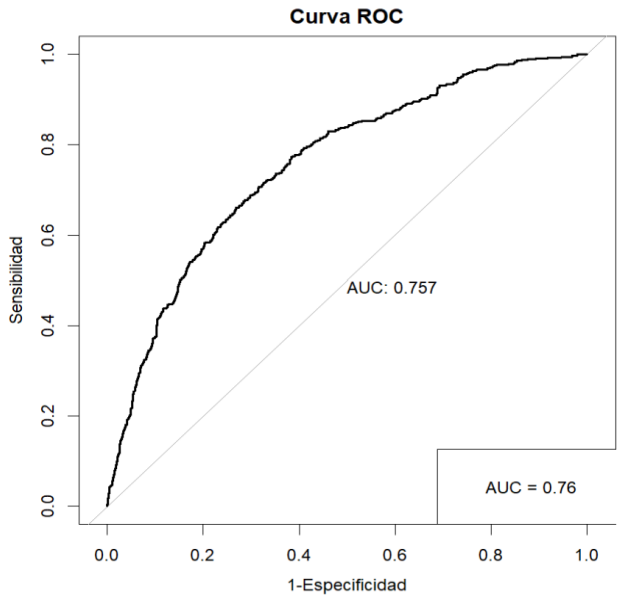

Gráfico 11: Curva ROC con estadístico AUC para primera iteración del modelo

Para esta primera iteración, se evaluó el desempeño de los datos predichos por el modelo, comparándolos con el porcentaje de la data definida para el testeo y se obtuvo un valor del área bajo la curva (AUC) de 0.756, con un intervalo de confianza del 95%. Este resultado indica que el modelo tiene una capacidad de predicción considerablemente mejor que la probabilidad al lanzar una moneda al azar, donde se obtendría un AUC de 0.5.

Un valor de AUC de 0.756 sugiere que el modelo tiene un rendimiento satisfactorio en la clasificación de las instancias, superando el azar y mostrando una habilidad para discriminar entre las clases. Esto indica que el modelo es capaz de realizar predicciones más precisas y útiles que una simple elección aleatoria.

El intervalo de confianza del 95% proporciona una estimación de la incertidumbre asociada al AUC. Al estar dentro de este intervalo, el modelo presenta un rendimiento razonablemente bueno en términos de su capacidad para distinguir entre las clases.

Sin embargo, es importante recalcar que ciertas variables incorporadas en esta primera iteración del modelo son muy poco significativas, por ejemplo, variables como el tipo de persona y tipo de empresa del cliente y el tipo de persona del deudor, variables que al ser poco significativas afectan el desempeño del modelo, las cuales deben ser eliminadas en próximas iteraciones. Cabe destacar que el resultado obtenido se basa en la primera iteración del modelo y puede requerir ajustes y refinamientos adicionales para mejorar aún más su rendimiento.

6.2 Segunda Iteración

6.2.1 Selección del modelo

El modelo seleccionado para la elaboración del modelo es el mismo estipulado en la **Sección 6.1.1** del presente informe.

6.2.2 Selección de variables

Es importante destacar que, en esta nueva selección de variables, se agregaron además las variables Rechazados cliente y Rechazados deudor, definidas como:

- **Rechazados cliente:** Variable binaria que corresponde a 1 si el cliente presenta documentos que han sido rechazados por la empresa, 0 si no.

- **Rechazados deudor:** Variable binaria que corresponde a 1 si el deudor presenta documentos que han sido rechazados por la empresa, 0 si no.

Además, con el objetivo de obtener monotonía, propiedad deseable que puede tener una relación entre la variable independiente (X) y la variable dependiente (Y). Una relación monótona implica que a medida que los valores de la variable independiente aumentan (o disminuyen), los valores de la variable dependiente también siguen la misma dirección, es decir, aumentan (o disminuyen) de manera consistente, se realizó una nueva categorización de las variables continuas:

1. Tasa del documento:

- Categoría 1: [0, 1.22]
- Categoría 2: (1.22, ∞)
- Categoría 3: (1.66, ∞)
- Categoría 4: (2.26, ∞)

2. Porcentaje anticipado:

- Categoría 1: [0,97%]
- Categoría 2: (97%, ∞)
- Categoría 3: (98%, ∞)
- Categoría 4: (99%, ∞)

3. Antigüedad del cliente en el SII:

- Categoría 1: [0 años, 2 años]
- Categoría 2: (2 años, 3.6 años]
- Categoría 3: (3.6 años, 6.2 años]
- Categoría 4: (6.2 años, 30.2 años]

4. Antigüedad del deudor en el SII:

- Categoría 1: [1.7 años,14.4 años]
- Categoría 2: (14.4 años, 24.5 años]
- Categoría 3: (24.5 años, 28.1 años]
- Categoría 4: (28.1 años, 35.8 años]

5. Antigüedad del cliente en la empresa:

- Categoría 1: [0 años, 0.6 años]
- Categoría 2: (0.6 años, ∞)
- Categoría 3: (1 año, ∞)
- Categoría 4: (3 años, ∞)

6. Ratio de facturas saldadas que presenta el cliente:
-
- Categoría 1: [0%, 25%]
- Categoría 2: (25%, 50%]
- Categoría 3: (50%, 75%]
- Categoría 4: (75%, 100%]

7. Ratio de facturas saldadas que presenta el cliente en los últimos 12 meses:
-
- Categoría 1: [0%, 25%]
- Categoría 2: (25%, 50%]
- Categoría 3: (50%, 75%]
- Categoría 4: (75%, 100%]

8. Ratio de facturas saldadas que presenta el cliente en los últimos 6 meses:

- Categoría 1: [0%, 25%]
- Categoría 2: (25%, 50%]
- Categoría 3: (50%, 75%]
- Categoría 4: (75%, 100%]

9. Ratio de facturas saldadas que presenta el cliente en los últimos 3 meses:

- Categoría 1: [0%, 25%]
- Categoría 2: (25%, 50%]
- Categoría 3: (50%, 75%]
- Categoría 4: (75%, 100%]

10. Ratio de facturas saldadas por el deudor en los últimos 12 meses:
-
- Categoría 1: [0%, 25%]
- Categoría 2: (25%, 50%]
- Categoría 3: (50%, 75%]
- Categoría 4: (75%, 100%]

11. Ratio de facturas saldadas por el deudor en los últimos 6 meses:

- Categoría 1: [0%, 25%]
- Categoría 2: (25%, 50%]
- Categoría 3: (50%, 75%]

- Categoría 4: (75%, 100%]

12. Ratio de facturas saldadas por el deudor en los últimos 3 meses:

- Categoría 1: [0%, 25%]
- Categoría 2: (25%, 50%]
- Categoría 3: (50%, 75%]
- Categoría 4: (75%, 100%]

13. Monto:

- Categoría 1: [0, 111.657]
- Categoría 2: (111.657, ∞)
- Categoría 3: (201.125, ∞)
- Categoría 4 :(314.520, ∞)
- Categoría 5: (463.769, ∞)
- Categoría 6: (666.246, ∞)
- Categoría 7: (940.395, ∞)
- Categoría 8: (1.370.809, ∞)
- Categoría 9: (2.140.756, ∞)
- Categoría 10: (3.861.473, ∞)

Para la selección de variables a incorporar en el modelo, se utilizó la regresión paso a paso o regresión Stepwise, la cual corresponde a una iteración iterativa paso a paso de un modelo de regresión que implica la selección automática de variables independientes de acuerdo con su relevancia e impacto en la variable dependiente. De esta manera, el objetivo de la regresión stepwise es encontrar un conjunto de variables independientes que influyan significativamente en la variable dependiente.

El enfoque empleado para realizar la regresión paso a paso es conocido como forward selection. En este método, el modelo comienza sin incluir ninguna variable y luego se evalúa cada una de ellas para determinar su significancia estadística al agregarlas al modelo. Solo las variables más significativas son retenidas, repitiendo este proceso hasta alcanzar los resultados óptimos.[18]

Las variables seleccionadas por la regresión stepwise son:

- Monto del documento
- Ratio de facturas saldadas por el cliente en los últimos 6 meses
- Antigüedad del cliente en el SII
- Antigüedad del cliente en la empresa
- Rechazadas cliente
- Rechazadas deudor
- Ratio de facturas saldadas por el deudor en los últimos 6 meses
- Tamaño del cliente
- Tamaño del deudor
- Zona del deudor
- Sector económico del deudor.

6.2.3 Resultados del modelo

Previo al desarrollo del modelo, se realizó una partición aleatoria de la base de datos considerando un 70% de estos como una base de entrenamiento. La cual se utiliza para ajustar los parámetros del modelo y permitirle aprender los patrones y relaciones en los datos, entrenando de esta manera el modelo. Por otro lado, se consideró un 30% de los datos totales como una base de testeo o prueba, conjunto de datos que se reserva exclusivamente para evaluar el rendimiento del modelo entrenado.

Los betas estimados y calculados para cada variable por el modelo de regresión se observan en la Tabla 3:

Variable	Categoría	Beta estimado
Monto	<= 201.125	
	> 201.125	0,24970
	> 314.520	0,35814
	> 1.370.809	0,62773
	> 2.140.756	0,91688
	> 3.861.473	0,57581
Tamaño Cliente	Grande	
	Mediana	-0,45602
	Micro	-0,24391
	Pequeña	-0,48380
	Sin info	0,17888
Ratio de facturas saldadas por el cliente en los últimos 6 meses	[0%, 24%]	
	[25%, 49%]	-0,22055

	[50%, 74%]	-0,68329
	[75%, 100%]	-1,01414
Antigüedad SII cliente	[0 años, 2 años]	
	(2 años, 3.7 años]	-0,26578
	(3.7 años, 6.3 años]	-0,48299
	(6.3 años, 30.2 años]	-0,51028
Antigüedad Chita Cliente	Menor o igual a 1 año	
	Mayor a un 1 año	-0,19858
Tamaño deudor	Grande	
	Mediana	0,53585
	Micro	0,75333
	Pequeña	0,60213
	Sin info	0,89887
Ratio de facturas saldadas por el deudor en los últimos 6 meses	[0%, 24%]	
	[25%, 49%]	-0,00535
	[50%, 74%]	-0,61519
	[75%, 100%]	-0,93771
Zona Deudor	Zona Centro	

	Zona Norte	0,16778
	Zona Sur	0,20097
Rechazadas cliente	No	
	Si	0,59916
Sector económico deudor	Sector Primario	
	Sector Secundario	0,33391
	Sector Terciario	0,53613
Rechazadas deudor	No	
	Si	0,445261
Constante		-2,59940

Tabla 3: Betas calculados por el modelo logit para cada categoría y variable definida (segunda iteración)

6.2.4 Evaluación de desempeño del modelo

Para la evaluación del desempeño del modelo realizado, se utilizó la curva ROC y el área bajo dicha curva (AUC). La curva ROC es un gráfico que representa la relación entre la tasa de verdaderos positivos (Sensibilidad) y la tasa de falsos positivos (1-Especificidad), definidos en la **Sección 3.3**, en diferentes umbrales de clasificación.

En el Gráfico 12 se puede observar la curva ROC obtenida del modelo, junto al valor del AUC.

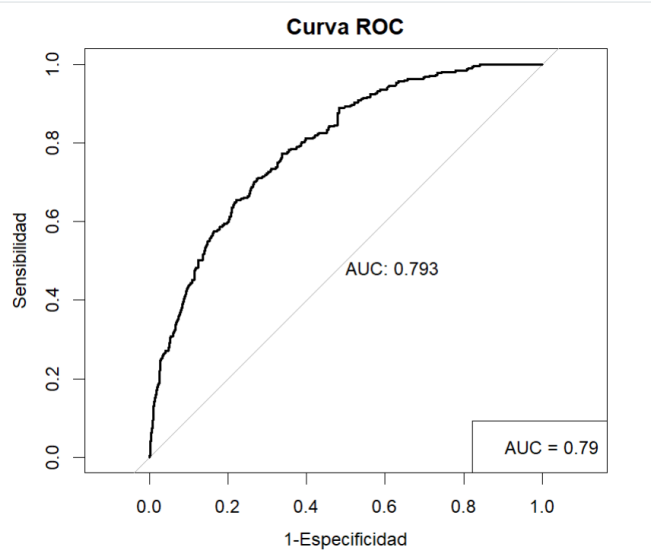

Gráfico 12: Curva ROC con estadístico AUC para segunda iteración del modelo

Para esta segunda iteración, se evaluó el desempeño de los datos predichos por el modelo, comparándolos con el porcentaje de la data definida para el testeo y se obtuvo un valor del área bajo la curva (AUC) de 0.793, con un intervalo de confianza del 95%. Este resultado indica que el modelo tiene una capacidad de predicción considerablemente mejor que la probabilidad al lanzar una moneda al azar, donde se obtendría un AUC de 0.5 y superior al AUC obtenido en la primaria iteración del modelo. Esto indica que el modelo es capaz de realizar predicciones más precisas y útiles que una simple elección aleatoria.

El intervalo de confianza del 95% proporciona una estimación de la incertidumbre asociada al AUC. Al estar dentro de este intervalo, el modelo presenta un rendimiento razonablemente bueno en términos de su capacidad para distinguir entre las clases.

7. Interpretación y evaluación de resultados

7.1 Escalamiento de los betas a score

Entre los resultados proporcionados por el modelo de regresión logística binomial, además de los betas asociadas a cada categoría de cada una de las variables explicativas, se encuentra la probabilidad de que un documento o factura caiga en default, probabilidad representada por un número entre 0 y 1.

Utilizando los resultados obtenidos del modelo de regresión logística se realiza un re-escalamiento a score donde a cada categoría de las variables explicativas a utilizar se le asocia un puntaje positivo o negativo de acuerdo con su incidencia en la probabilidad de default del documento, cuya suma corresponde a un puntaje definido entre 0 y 1000 donde el riesgo de default es menor a medida que aumenta el puntaje y viceversa.

Para la regresión logística, la probabilidad de que ocurra un suceso (en este caso el default del documento) se aproxima mediante la función logística:

$$P(x) = \frac{1}{1+ e^{-(\Sigma \beta_i x_i + \beta_0)}} \qquad (1)$$

Donde despejando $P(x)$ se tiene:

$$\Rightarrow P(x) * (1 - P(x))^{-1} = e^{(\Sigma \beta_i x_i + \beta_0)}$$

$$\Rightarrow \frac{P(x)}{1- P(x)} = e^{(\Sigma \beta_i x_i + \beta_0)} \qquad (2)$$

Para facilitar el cálculo de los puntajes se ocupó la ecuación (2) de tal manera que:

$$\Rightarrow ln(\frac{P(x)}{1- P(x)}) = \Sigma \beta_i x_i + \beta_0 => ln(\frac{1-P(x)}{P(x)}) = -(\Sigma \beta_i x_i + \beta_0) \quad (3)$$

Una vez definida la ecuación (5) se definen los mínimos y máximos para $ln(\frac{1-P(x)}{P(x)})$ donde:

- Valor mínimo: -0,61151
- Valor máximo: 5,741733

Para el cálculo de los valores máximos y mínimos se utilizaron los valores obtenidos de la función logística $P(x)$ para el modelo desarrollado, dichos valores fueron nuevamente calculados en $ln(\frac{1-P(x)}{P(x)})$, donde se obtuvo un vector de valores, desde el cual se eligió el valor máximo y el mínimo.

La metodología utilizada para realizar la transformación de los betas obtenidas a puntajes es convertir los mínimos y máximos de $ln(\frac{1-P(x)}{P(x)})$ en cero y 1.000 respectivamente. Para ello se suma el valor mínimo a ambos extremos, quedando los límites:

- Valor mínimo: 0
- Valor máximo: 6,35327

Una vez definidos los extremos, se busca encontrar un valor x que cumpla:

$$\Rightarrow (ln(\frac{1-P(x)}{P(x)}) + 0{,}61151) * x = 0 \quad (4)$$
$$\Rightarrow (ln(\frac{1-P(x)}{P(x)}) + 0{,}61151) * x = 1000 \quad (5)$$

De la ecuación (4) es simple deducir que $(ln(\frac{1-P(x)}{P(x)}) + 0{,}61151) = 0$, por lo que (4) se cumple para cualquier valor de x. En cambio, para la ecuación (4) se tiene que:

$$\Rightarrow (ln(\frac{1-P(x)}{P(x)}) + 0{,}61151) * x = 1000 \quad (5)$$

$$\Rightarrow x = \frac{1000}{(ln(\frac{1-P(x)}{P(x)}) + 0{,}61151)} \quad (5)$$

Reemplazando en la ecuación (5) $ln(\frac{1-P(x)}{P(x)})$ por el máximo obtenido:

$$\Rightarrow x = \frac{1000}{(5{,}74173 + 0{,}61151)} = \frac{1000}{6{,}25324} = 157{,}4$$

Reemplazando el valor de x obtenido en la ecuación (5):

$$\Rightarrow score = (ln(\frac{1-P(x)}{P(x)}) + 0{,}61151) * 157{,}4 \quad (6)$$

Utilizando la ecuación (6) y reemplazando los valores por lo estipulado en la ecuación (5):

$$\Rightarrow score = ((-(\Sigma \beta_i x_i + \beta_0) + 0{,}61151) * 157{,}4 \quad (7)$$

Por simplicidad, considerando que el score total es la suma de los scores asociados a cada variable del documento por separado, es necesario dividir el valor del beta constante y el incremento de 0,61151 constante, por el número de variables presentes en el modelo, para este caso, 11 variables. Además, para el cálculo es importante considerar que $\beta_i x_i$ corresponde a los betas asociados a cada variable utilizada en el modelo, de esta manera, la fórmula de score asociado a cada variable presenta la siguiente estructura:

$$\Rightarrow score = ((-(\Sigma \beta_i x_i + \frac{\beta_0}{11}) + \frac{0{,}61151}{11}) * 157{,}4 \quad (8)$$

Los scores obtenidos para cada variable se observan en la Tabla 4:

Variable	Categoría	Score calculado
Monto	<= 201.125	46
	> 201.125	7
	> 314.520	-10
	> 1.370.809	-53
	> 2.140.756	-98
	> 3.861.473	-45

Tamaño Cliente	Grande	46
	Mediana	118
	Micro	84
	Pequeña	122
	Sin info	18
Ratio de facturas saldadas por el cliente en los últimos 6 meses	[0%, 24%]	46
	[25%, 49%]	81
	[50%, 74%]	153
	[75%, 100%]	206
Antigüedad SII cliente	[0 años, 2 años]	46
	(2 años, 3.7 años]	88
	(3.7 años, 6.3 años]	122
	(6.3 años, 30.2 años]	126
Antigüedad Chita Cliente	Menor o igual a 1 año	46
	Mayor a un 1 año	77
Tamaño deudor	Grande	46
	Mediana	-38
	Micro	-73
	Pequeña	-49

	Sin info	-96
Ratio de facturas saldadas por el deudor en los últimos 6 meses	[0%, 24%]	46
	[25%, 49%]	47
	[50%, 74%]	143
	[75%, 100%]	194
Zona Deudor	Zona Centro	46
	Zona Norte	20
	Zona Sur	14
Rechazadas cliente	No	46
	Si	-48
Sector económico deudor	Sector Primario	46
	Sector Secundario	-7
	Sector Terciario	-38
Rechazadas deudor	No	46
	Si	-24

Tabla 4: Score asignado a cada variable

En el contexto del análisis de regresión, el beta (β) es un coeficiente que indica la magnitud y la dirección de la relación entre una variable independiente y la variable dependiente. Cuando se evalúan las variables mediante un puntaje, se considera el signo del coeficiente beta para asignar los valores.

Cuando una variable tiene un coeficiente beta positivo, significa que un aumento en su valor se relaciona con un aumento en la variable dependiente. En este caso, para el puntaje, se asigna un valor menor.

Por otro lado, si una variable tiene un coeficiente beta negativo, indica que un aumento en su valor se relaciona con una disminución en la variable dependiente. En este escenario, la variable obtiene un puntaje más alto.

Además, de la tabla 4 es importante destacar:

- Al aumentar el monto de la factura, el score asociado a la variable disminuye, esto se da hasta aproximadamente los $ 4MM, ya que de acuerdo con los betas obtenidos, al superar dicho monto disminuye el beta asociado al incumplimiento, principalmente debido a que el promedio en monto de las facturas operadas por la empresa se encuentra alrededor de los 1.6 MM.

- Al aumentar el ratio de facturas saldadas en los últimos 6 meses, tanto para el cliente como para el deudor, el score asociado a la variable aumenta. Por ende, el score premia a los clientes y deudores con un alto historial de pago.

- Al aumentar la antigüedad del cliente, tanto en el Servicio de Impuestos Internos como en la empresa de factoring, aumenta el score asociado a la variable. Por ende, el score premia a aquellos clientes que presentan una mayor antigüedad.

- En cuanto a los rechazos, el score premia a aquellos cliente y deudores que no presentan documentos rechazados por la empresa.

- En cuanto a la zona del deudor, el score favorece a aquellos deudores localizados en la zona central del país, siendo menor para aquellos deudores localizados en las zonas norte y sur.

- En cuanto al tamaño del deudor, el score favorece a aquellos deudores que presentan un mayor tamaño en ventas, castigando en gran medida a aquellos deudores cuya información respecto a tamaño en ventas se desconoce.

- En cuanto al sector económico del deudor, el score favorece a aquellos deudores cuya actividad económica se sitúa en el sector primario. Esto debido a principalmente a que la base de datos presenta un bajo porcentaje de deudores pertenecientes al sector primario, agrupándose la mayoría en el sector secundario y terciario.

- En cuanto al tamaño del cliente, considerando que aproximadamente el 80% de las empresas que operan con el factoring corresponden a micro, pequeña y medianas empresas, estas presentan un alto score, castigando únicamente a aquellos clientes cuyo tamaño en ventas se desconoce.

7.2 Validación del score

El gráfico 13 muestra la distribución del score para la base de entrenamiento, donde el máximo y el mínimo puntaje obtenido fueron 1000 y 0, respectivamente, con un promedio de 598.428 y una desviación estándar de 166,85.

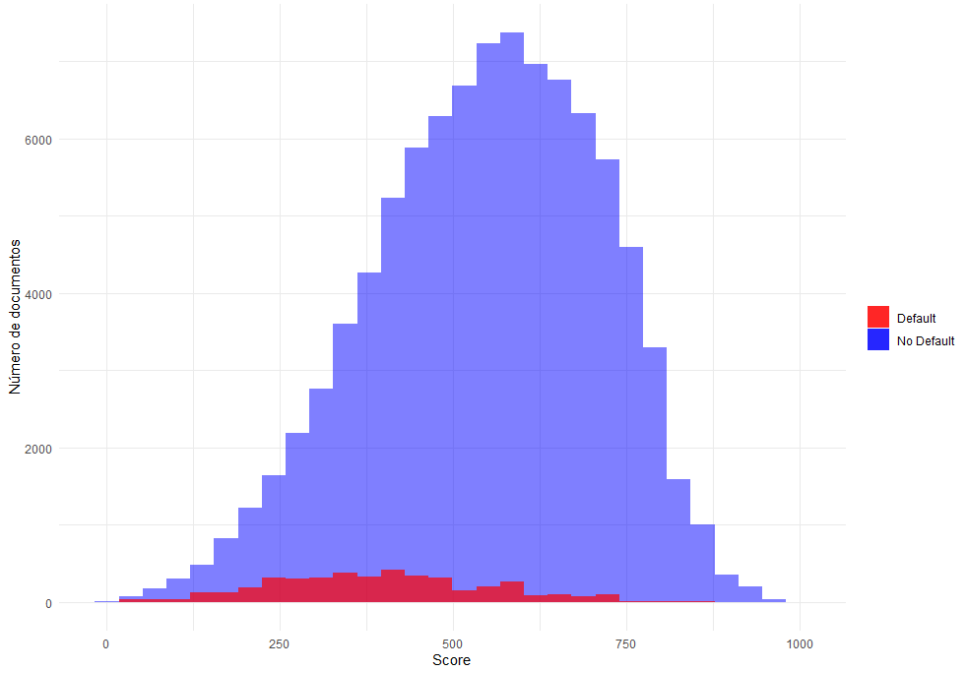

Gráfico 13: Distribución del score en base de entrenamiento

El gráfico 14 muestra la distribución del score para la base de testeo, donde el máximo y el mínimo puntaje obtenido fueron 997 y 8, respectivamente, con un promedio de 599,55 y una desviación estándar de 166,80.

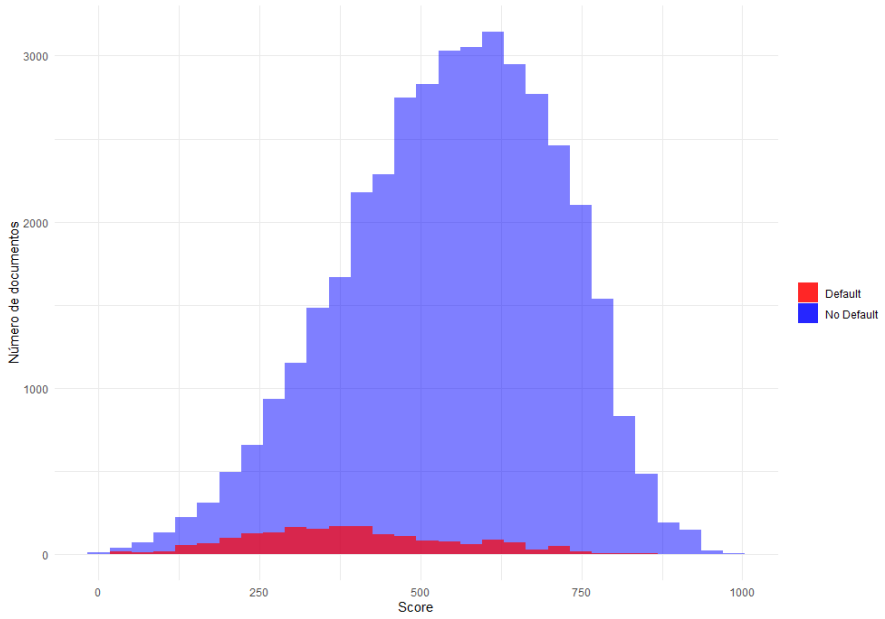

Gráfico 14: Distribución del score en base de testeo

El gráfico 15 muestra la distribución del score para la base completa, donde el máximo y el mínimo puntaje obtenido fueron 1000 y 0, respectivamente, con un promedio de 598,75 y una desviación estándar de 166,84.

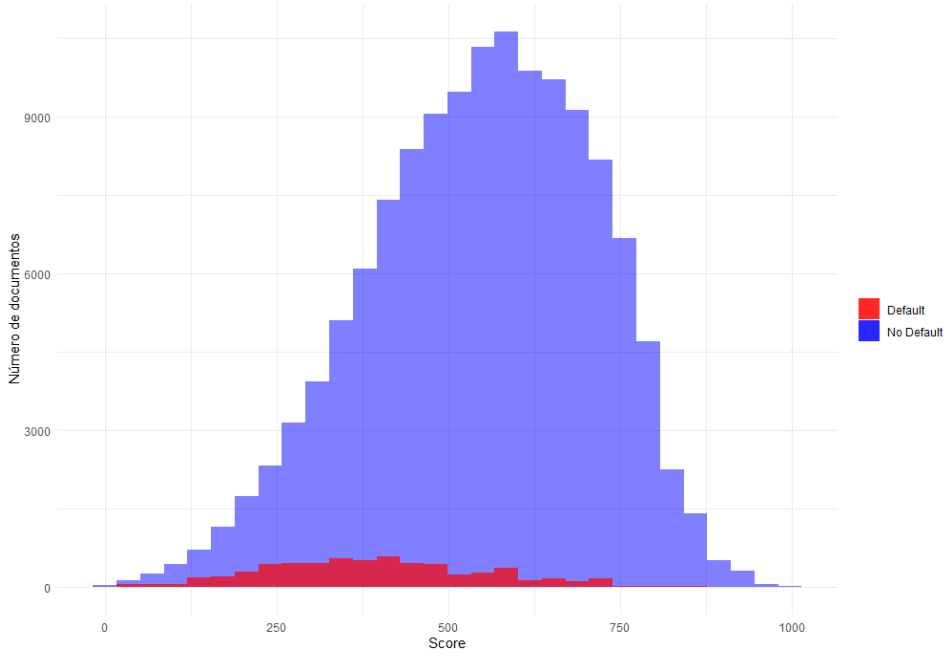

Gráfico 15: Distribución del score en la base completa

De los gráficos 13, 14 y 15 es posible observar que al aumentar el puntaje o score el número de documentos que incurre en default disminuye significativamente, cumpliendo el objetivo del escalamiento de los betas obtenidos a score.

7.3 Definición de puntajes de corte

En los modelos de Scoring, la noción de puntajes de corte surge como un elemento esencial para la toma de decisiones y la asignación de categorías de riesgo o probabilidad. Un puntaje de corte, también conocido como "threshold" en inglés, representa un valor crítico que demarca la separación entre diferentes clasificaciones o resultados deseados en el modelo. En esencia, es el punto de referencia que determina si un individuo, entidad o solicitud será clasificado como positivo o negativo con respecto al objetivo de la evaluación.

La definición adecuada de puntajes de corte constituye un desafío fundamental en la construcción y aplicación de modelos de Scoring, ya que impacta directamente en la precisión y efectividad de dichos modelos. Establecer un puntaje de corte óptimo requiere un equilibrio delicado entre minimizar los falsos positivos y falsos negativos, garantizando así que el modelo pueda identificar de manera certera a aquellos casos que cumplen con ciertos criterios predefinidos. Esta es una decisión estratégica que debe estar alineada con los objetivos de la empresa y con las solicitudes del sponsor.

Con el objetivo de encontrar los puntos de corte óptimos, se realizó una división del score en 20 cortes equidistantes, presentes en la Tabla 5:

Score Mínimo	Score Máximo	Total, de documentos	Porcentaje de buenos aceptados	Porcentaje de malos rechazados	Cobertura	Cobertura acumulada
945	1.000	58	0,04%	100%	0,04%	0,04%
892	945	570	0,47%	99,91%	0,41%	0,45%
840	892	1.760	1,78%	99,69%	1,26%	1,71%
787	840	5.014	5,54%	99,29%	3,60%	5,31%
735	787	9.306	12,51%	98,54%	6,68%	11,99%
682	735	13.613	22,58%	94,98%	9,77%	21,77%
630	682	14.689	33,47%	91,56%	10,54%	32,31%
577	630	15.896	45,15%	85,61%	11,41%	43,72%
525	577	16.272	57,10%	79,54%	11,68%	55,40%
472	525	14.842	67,92%	72,28%	10,65%	66,05%
420	472	13.665	77,53%	58,41%	9,81%	75,86%

367	420	10.625	84,93%	46,05%	7,63%	83,69%
315	367	8.447	90,64%	32,77%	6,06%	89,55%
262	315	6.098	94,67%	21,34%	4,38%	93,93%
210	262	4.003	97,23%	12,02%	2,87%	96,80%
157	210	2.465	98,82%	6,55%	1,77%	98,57%
105	157	1.245	99,59%	2,99%	0,89%	99,47%
52	105	534	99,89%	0,91%	0,38%	99,85%
0	52	207	100%	0%	0,15%	100%

Tabla 5: Puntajes de corte con variables acumulada

La Tabla 5 presenta un análisis detallado del rango de puntajes, donde las dos primeras columnas establecen los límites del score. La columna "Total de documentos" muestra la cantidad total de documentos que se encuentran dentro de cada intervalo de puntajes. Para ofrecer una perspectiva más clara, las columnas "Porcentaje de buenos" y "Porcentaje de malos" expresan el porcentaje de documentos considerados buenos y malos respectivamente, en relación con el total de documentos buenos y malos presentes en la base.

Finalmente, la columna "Cobertura" adquiere gran relevancia al indicar el porcentaje de documentos que abarca cada rango respecto al total de la base de datos. Este indicador es de suma importancia, ya que permite entender qué proporción de la población total se encuentra representada en cada intervalo de puntajes, brindando así una visión integral del rendimiento del modelo de Scoring.

El análisis proporcionado en la Tabla 5 resulta fundamental para comprender la distribución de puntajes y la calidad del modelo, permitiendo una evaluación exhaustiva de su capacidad de predicción y desempeño en la clasificación de los documentos según su riesgo crediticio o de default. Estos hallazgos contribuyen significativamente a la toma de decisiones informadas en distintos contextos financieros y crediticios, respaldando la efectividad y fiabilidad del modelo de Scoring implementado.

Actualmente, el área de operaciones de la empresa acepta aproximadamente un 75% del total de documentos que no son aceptados automáticamente por el modelo LINCE. Sin embargo, la directiva de la empresa desea que este porcentaje de aceptación aumente a un 80%. Esta decisión tiene como propósito asegurar que el modelo de Scoring sea lo suficientemente amplio y efectivo en la clasificación de los documentos, permitiendo abarcar la mayoría de las solicitudes que cumplen con los criterios de aceptación de la empresa.

Al utilizar esta información como guía para definir el puntaje de corte de rechazos, se busca lograr una mayor alineación entre las decisiones manuales y las del modelo,

optimizando así los resultados y garantizando que las solicitudes de mayor riesgo sean adecuadamente identificadas y atendidas. Esta estrategia contribuirá a mejorar la eficiencia y precisión del proceso de evaluación, brindando una mayor confianza en el sistema de Scoring y facilitando la toma de decisiones informadas en el área de operaciones y riesgo de la empresa.

Al analizar detenidamente la Tabla 5, podemos destacar que el porcentaje de aceptación deseado del 80% se alcanza en el intervalo de puntajes que va desde 367 hasta 420. En este rango específico, se logra abarcar un significativo 84.93% del total de documentos evaluados.

Tomando en cuenta esta información, se ha tomado la decisión de establecer un puntaje de corte de rechazos en 367. Esto implica que todas las solicitudes de operación que obtengan un puntaje por debajo de este valor serán automáticamente rechazadas por el modelo de Scoring.

Al implementar este puntaje de corte, se asegura que el modelo mantenga una cobertura satisfactoria, capturando al menos el 75% de las solicitudes que actualmente son aceptadas manualmente por el área de riesgos y operaciones. Además, se garantiza una mayor coherencia entre las decisiones del modelo y las tomadas de forma manual, optimizando así el proceso de evaluación y mejorando la precisión en la identificación de documentos de alto riesgo.

Con el objetivo de establecer los puntajes de aceptación óptimos, se procederá a seleccionar aquellos rangos que maximicen las utilidades de la empresa. Para lograr esto, se llevará a cabo un análisis detallado de los datos obtenidos, considerando tanto la cobertura alcanzada como el rendimiento financiero asociado a cada intervalo de puntajes.

El proceso de selección se basará en encontrar el equilibrio adecuado entre la cantidad de solicitudes aceptadas y el riesgo crediticio asociado a cada decisión. Se busca identificar los rangos de puntajes que permitan capturar un porcentaje significativo de documentos, manteniendo al mismo tiempo una tasa de aceptación de préstamos que sea coherente con los objetivos de rentabilidad y seguridad de la empresa.

Este enfoque permitirá ajustar los puntajes de aceptación de manera estratégica, considerando la información proporcionada en la Tabla 5 y las metas establecidas junto con el sponsor de la empresa. De esta forma, se garantiza una toma de decisiones fundamentada en datos cuantitativos, maximizando las oportunidades de negocio sin comprometer la salud financiera de la organización.

Con el propósito de calcular los ingresos asociados a documentos correctamente operados y costos de aquellos que incurren en default, se ha creado la Tabla 6, donde se presenta un promedio de los datos más relevantes de la data. Esta tabla busca proporcionar un panorama consolidado y representativo de los resultados financieros obtenidos por ambos tipos de documentos.

Al tomar en consideración los datos más significativos y aplicar un enfoque promedio, se obtiene una estimación más precisa de los ingresos generados por cada categoría de documentos. Esto incluye tanto el rendimiento financiero de los documentos que son operados de manera adecuada, como las pérdidas asociadas a aquellos que incurren en default.

Dato	Promedio por documento
Monto factura	$1.655.075
Tasa del documento	1,74%
Plazo (días)	38
Porcentaje anticipado	97%
Tasa fondeo	1,12%
Días transcurridos entre la fecha de operación y el pago.	40
Tasa Mora	3,26%
Días de mora acumulados para los casos buenos	14
Comisión	$8.431

Tabla 6: Promedio de variables para el cálculo de ingresos y costos

Las utilidades asociadas a documentos pagados correctamente corresponden a:

$$\Rightarrow Anticipo = (Monto\ factura * porcentaje\ anticipado)$$

$$\Rightarrow Utilidades\ documento\ bueno = ((Anticipo * (\frac{Tasa\ del\ documento}{30}) * Plazo) + comisión)$$
$$-(Anticipo * (\frac{Tasa\ fondeo}{30}) * Días\ transcurridos\ entre\ la\ fecha\ de\ operación\ y\ el\ pago)$$
$$+ (Anticipo * (\frac{Tasa\ mora}{30}) * Días\ de\ mora)$$

Por otro lado, las utilidades asociadas a un documento que incurre en default corresponden a:

$$\Rightarrow Utilidades\ documento\ malo = -(Anticipo + \left(Anticipo * \left(\frac{Tasa\ fondeo}{30}\right) * 360\right))$$

Donde los días transcurridos entre la fecha de operación hasta la fecha de pago se topa en 360, que corresponde a los días a partir del cual se castiga el documento, se considera además como supuesto una LGD del 100%. Además, es relevante mencionar que los

días de mora no se consideran para los costos, ya que se asume que no serán pagados. En este sentido, se prevé una provisión del 100% de lo anticipado, teniendo en cuenta que los costos asociados a estos días de mora no serán recuperados.

En la Tabla 7 se muestran los ingresos generados por los documentos pagados correctamente y los costos asociados a aquellos que incurren en default. Los cuales se obtuvieron ocupando las cifras presentadas en la Tabla 6, en las fórmulas de utilidades de los documentos buenos y malos presentadas en los párrafos anteriores.

Utilidades de aceptar un documento bueno	Utilidades de aceptar documento malo
$44.264	$-1.821.192

Tabla 7: Utilidades y costos promedio calculados

A partir del análisis de la Tabla 7, se destaca que el ratio entre el costo de aceptar una operación mala y el beneficio de aceptar una operación buena es aproximadamente de un 2,4%. En otras palabras, por cada documento que incurre en default, es necesario aceptar y operar alrededor de 41 documentos buenos para lograr obtener utilidades positivas.

Este indicador es de gran relevancia, ya que proporciona una perspectiva clara sobre la relación entre los costos asociados a los documentos que incurren en default y los ingresos generados por aquellos que son operados correctamente y cumplen con sus obligaciones de pago.

A partir de esta información, se construye la tabla 8:

Score Mínimo	Score Máximo	Total de documentos	Nº buenos	Nº malos	Utilidades documento bueno	Utilidades documentos en default	Utilidades Totales (Utilidades buenos - Utilidades malos)
945	1.000	58	58	0	$2.567.314	$0	$2.567.314
892	945	570	564	6	$24.964.916	$-10.927.149	$14.037.767
840	892	1.760	1.746	14	$77.285.006	$-25.496.682	$51.788.325
787	840	5.014	4.988	26	$220.789.010	$-47.350.981	$173.438.030
735	787	9.306	9.258	48	$409.796.443	$-87.417.195	$322.379.248
682	735	13.613	13.383	230	$592.385.591	$-418.874.061	$173.511.530
630	682	14.689	14.468	221	$640.412.070	$-402.483.336	$237.928.733
577	630	15.896	15.512	384	$686.623.723	$-699.337.562	$-12.713.839
525	577	16.272	15.880	392	$702.912.888	$-713.907.094	$-10.994.206

472	525	14.842	14.373	469	$636.206.986	$-854.138.845	$-217.931.859
420	472	13.665	12.769	896	$565.207.473	$-1.631.787.645	$-1.066.580.172
367	420	10.625	9.827	798	$434.982.680	$-1.453.310.871	$-1.018.328.191
315	367	8.447	7.589	858	$335.919.768	$-1.562.582.365	$-1.226.662.597
262	315	6.098	5.360	738	$237.255.232	$-1.344.039.377	$-1.106.784.145
210	262	4.003	3.401	602	$150.541.986	$-1.096.357.324	$-945.815.338
157	210	2.465	2.112	353	$93.485.644	$-642.880.623	$-549.394.980
105	157	1.245	1.015	230	$44.927.996	$-418.874.061	$-373.946.064
52	105	534	470	134	$17.705.614	$-244.039.670	$-226.334.056
0	52	207	148	59	$6.551.077	$-107.450.302	$-100.899.225

Tabla 8: Utilidades promedio por rango de score

El análisis detallado de la Tabla 8 revela que los puntos de corte óptimos para maximizar las utilidades de la empresa se encuentran en el rango de 630 a 1000. En estos intervalos de puntajes, se reportan utilidades positivas, incluso cuando se aceptan operaciones que incurren en default. Esto representa un valioso 32.31% de cobertura de la cartera.

Además, se considera que los rangos de puntajes entre 367 y 630 continúen siendo evaluados por el área de operaciones y riesgos de la empresa, tal como se realiza actualmente.

De esta manera la asignación final de los puntajes de corte se puede apreciar en la Tabla 9:

Score Mínimo	Score Máximo	Total de documentos	Porcentaje de buenos aceptados	Porcentaje de malos rechazados	Cobertura	Cobertura acumulada
945	1.000	58	0,04%	100%	0,04%	0,04%
892	945	570	0,47%	99,91%	0,41%	0,45%
840	892	1.760	1,78%	99,69%	1,26%	1,71%
787	840	5.014	5,54%	99,29%	3,60%	5,31%
735	787	9.306	12,51%	98,54%	6,68%	11,99%
682	735	13.613	22,58%	94,98%	9,77%	21,77%
630	682	14.689	33,47%	91,56%	10,54%	32,31%
577	630	15.896	45,15%	85,61%	11,41%	43,72%

525	577	16.272	57,10%	79,54%	11,68%	55,40%
472	525	14.842	67,92%	72,28%	10,65%	66,05%
420	472	13.665	77,53%	58,41%	9,81%	75,86%
367	420	10.625	84,93%	46,05%	7,63%	83,69%
315	367	8.447	90,64%	32,77%	6,06%	89,55%
262	315	6.098	94,67%	21,34%	4,38%	93,93%
210	262	4.003	97,23%	12,02%	2,87%	96,80%
157	210	2.465	98,82%	6,55%	1,77%	98,57%
105	157	1.245	99,59%	2,99%	0,89%	99,47%
52	105	534	99,89%	0,91%	0,38%	99,85%
0	52	207	100%	0%	0,15%	100%

Tabla 9: Distribución final del score y puntajes de corte

En la Tabla 9, los puntajes destacados en verde representan aquellos documentos que serán aceptados automáticamente por el modelo de Scoring. Estos intervalos de puntajes cumplen con los criterios establecidos para una operación segura y rentable, y no requerirán una revisión adicional por parte del área de operaciones y riesgos de la empresa.

Los valores resaltados en amarillo indican aquellos documentos que pasarán a ser evaluados minuciosamente por el área de operaciones y riesgos de la empresa. Estos rangos de puntajes se encuentran en una zona intermedia, donde es necesario realizar una revisión más detallada para determinar su viabilidad y nivel de riesgo asociado.

Por otro lado, los puntajes resaltados en rojo corresponden a documentos que serán rechazados automáticamente por el modelo de Scoring. Estos intervalos de puntajes no cumplen con los criterios de aceptación y presentan un riesgo elevado, por lo que se tomará la decisión de rechazarlos sin necesidad de una evaluación adicional.

Al tomar en cuenta estos hallazgos, la empresa podrá ajustar su proceso de evaluación y aceptación de documentos, asegurando que se mantenga un equilibrio entre el riesgo y la rentabilidad. Esta estrategia permitirá fortalecer la cartera de documentos y optimizar los resultados financieros, garantizando un desempeño sólido y sostenible en el mercado.

8. Estimación de beneficios del modelo

Al analizar la base de testeo, se identificaron un total de 1,910 documentos que incurrieron en default. Si estos documentos hubiesen pasado por el modelo de Scoring

desarrollado, un total de 898 facturas hubiese sido rechazadas por tener un Scoring menor al definido como umbral de rechazo. Como se detalló en la sección 7.3, específicamente en la Tabla 7, se determinó que el costo asociado a estos documentos rechazados asciende a $1.821.192. Sin embargo, gracias a la implementación del modelo propuesto, este rechazo de los 898 documentos que incurrieron en impago hubiese generado un ahorro significativo de $1.635.430.028.

La aplicación del modelo de Scoring permitió una selección más precisa y acertada de los documentos, evitando la operación de aquellos con un alto riesgo de default. Esta estrategia de rechazo basada en el puntaje de corte establecido resultó en un ahorro considerable para la empresa, reduciendo sustancialmente los costos asociados a la morosidad.

Por otro lado, en los alcances del trabajo, también se mencionó una posible reducción de los tiempos de respuestas al momento de responder operaciones. En la Tabla 10 se presenta la tasa de operación para distintos intervalos de respuesta de operaciones:

Tiempo de respuesta	Tasa de Operación
1 a 10 min	65,38%
11 a 30 min	50,59%
31 a 60 min	50,44%
61 a 120 min	47,09%
Más de 120 min	45,11%

Tabla 10: Tasa de operación para diferentes intervalos de respuesta de operaciones

En base a los datos presentados en la Tabla 10, se aprecia una tendencia de disminución en la tasa de operación a medida que aumenta el tiempo de respuesta de las operaciones por parte de la empresa. Es decir, a medida que transcurre más tiempo desde la recepción de las solicitudes hasta su evaluación y aceptación, la cantidad de operaciones realizadas disminuye. Esto principalmente porque el cliente desiste de operar con la empresa si esta demora mucho en responder sus solicitudes de operación.

Es en este contexto que el modelo de Scoring propuesto cobra una gran relevancia, ya que su implementación tiene como objetivo agilizar el proceso de respuesta de las operaciones por parte del equipo de riesgos y operaciones de la empresa. Al utilizar este modelo, se espera optimizar la eficiencia y rapidez en la toma de decisiones, lo que se traduce en una mayor tasa de operación en un tiempo reducido. La agilización del proceso de respuesta de las operaciones brinda numerosos beneficios para la empresa, tales como una mejora en la experiencia del cliente.

El modelo de Scoring propuesto se posiciona como una herramienta estratégica que contribuirá a fortalecer la posición competitiva de la empresa y a alcanzar un alto nivel de eficiencia en su gestión operativa. Al proporcionar una respuesta más rápida y precisa en la evaluación de documentos, facilitará la toma de decisiones informadas en un tiempo óptimo, lo que impactará positivamente en la rentabilidad y el rendimiento financiero de la organización.

Adicionalmente, se espera que el modelo propuesto contribuya al crecimiento y desarrollo del área de operaciones y riesgo de la empresa sin la necesidad de contratar más personal. Gracias a la implementación del modelo de Scoring, se busca optimizar la eficiencia y la precisión en la toma de decisiones, lo que permitirá afrontar un mayor volumen de solicitudes sin aumentar la carga laboral.

El modelo actúa como una herramienta estratégica que agiliza y automatiza gran parte del proceso de evaluación y aceptación de documentos, liberando tiempo y recursos humanos para enfocarse en tareas de mayor valor añadido y análisis más profundos.

La posibilidad de manejar y evaluar un mayor flujo de solicitudes sin aumentar la plantilla de personal brindará una ventaja significativa a la empresa al permitirle adaptarse rápidamente a las demandas del mercado y aprovechar oportunidades de negocio con agilidad y eficacia. Con el modelo de Scoring actuando como un aliado estratégico, la empresa estará en una posición más sólida para enfrentar los desafíos del entorno económico y alcanzar sus objetivos de crecimiento y rentabilidad.

9. Conclusiones

El objetivo general de este trabajo se ha cumplido de manera correcta al lograr el desarrollo de un modelo predictivo capaz de asignar a cada documento o factura que la empresa maneja un score basado en la probabilidad de incumplimiento asociado a dicho documento. Mediante la implementación de este modelo, se ha logrado automatizar la aceptación de alrededor del 33% de los documentos que la empresa debe operar, lo que representa un significativo alivio en la carga de trabajo del área de operaciones y riesgo. Al limitar la evaluación manual solo a una parte de los documentos, se ha optimizado el proceso, permitiendo una gestión más eficiente y enfocada en aquellos casos que requieren mayor atención y análisis detallado. Esta mejora operativa brinda a la empresa una ventaja competitiva al agilizar sus procedimientos y tomar decisiones más acertadas en un entorno financiero siempre cambiante.

En cuanto al primer y segundo objetivo específico definido, mediante un modelo de regresión logística GLM (Generalized Linear Model) el cual es aplicado específicamente a problemas de clasificación binaria, se logró observar y concluir que:

- Las variables relacionadas con el cumplimiento histórico del cliente o deudor con la empresa, más en específico, el ratio de facturas saldadas por sobre las facturas operadas en un período determinado, han demostrado una alta relevancia e impacto significativo en la probabilidad de default. A medida que este ratio de facturas saldadas aumenta, es decir, cuando el cliente o deudor tiene un historial

de pago más sólido con la empresa, el coeficiente asociado a esta variable (beta) disminuye considerablemente, lo que se traduce en una menor probabilidad de incurrir en default. Estos hallazgos resaltan la importancia de considerar el comportamiento histórico de los clientes al evaluar el riesgo crediticio y proporcionan una valiosa información para la toma de decisiones financieras informadas.

- Las variables relacionadas con la antigüedad del cliente, tanto en el Servicio de Impuestos Internos como en la empresa, muestran una tendencia clara, a medida que la antigüedad del cliente aumenta, el coeficiente (beta) asociado a estas variables disminuye, lo que indica una reducción en la probabilidad de default. En otras palabras, cuanto más tiempo lleva el cliente con la empresa y registrado en el SII, menor es la probabilidad de que incurran en impagos. Sin embargo, es importante destacar que, durante el proceso de selección de variables mediante el método stepwise, la variable que representa la antigüedad del deudor en el SII fue descartada. Esto sugiere que dicha variable no aportó significativamente a la explicación del modelo en términos de predecir la probabilidad de default.

- Las variables relacionadas con la localidad del deudor revelan un hallazgo significativo, cuando el deudor se encuentra ubicado en la zona central del país, el coeficiente asociado a estas variables es menor. Esto implica que la probabilidad de default disminuye considerablemente en comparación con aquellos deudores ubicados en el norte o sur del país. En otras palabras, la ubicación geográfica juega un papel importante en el riesgo crediticio, y los deudores situados en la zona central tienen una mayor probabilidad de cumplir con sus obligaciones financieras en comparación con aquellos en otras regiones del país. Es importante señalar que la variable relacionada con la localidad del cliente fue excluida del modelo stepwise. La principal razón detrás de esta exclusión es que, en última instancia, es el deudor quien tiene la responsabilidad de cumplir con las obligaciones de pago. Por lo tanto, resulta más factible llegar a un deudor ubicado en la zona central del país, lo que podría haber afectado la relevancia de esta variable en la predicción del riesgo crediticio.

- Las variables tasa de documento y el porcentaje anticipado fueron excluidas del análisis, dado que estas son variables que la empresa determina internamente según el riesgo asociado a cada documento a operar. En relación con el monto del documento, se observó que, al aumentar el valor del mismo, también se incrementa el coeficiente asociado y, por ende, la probabilidad de default. Sin embargo, esta tendencia se mantiene solo hasta aproximadamente los 4 millones. A partir de ese punto, el coeficiente estimado disminuye. Esto puede explicarse principalmente por el hecho de que el promedio de facturas operadas por la empresa ronda alrededor de 1.6 millones. Estos resultados sugieren que existe una relación no lineal entre el monto del documento y el riesgo de incumplimiento. A medida que el monto se incrementa, el riesgo crediticio también aumenta, pero solo hasta cierto punto, después del cual la relación se invierte. Esto puede deberse a las características operativas de la empresa de factoring y a su capacidad para gestionar montos más grandes de manera más efectiva.

- Cuando el cliente o deudor presenta documentos rechazados para su operación por parte de la empresa, se observa un incremento en el coeficiente asociado (beta) y, por consiguiente, un aumento en su probabilidad de incumplimiento de pago.

- En relación con el tamaño de la empresa, se observan diferencias significativas en el coeficiente asociado para el caso del deudor. En particular, el coeficiente (beta) de una empresa catalogada como grande es considerablemente menor que el de empresas clasificadas como micro, pequeñas o medianas. Por otro lado, el coeficiente asociado a empresas que no presentan información sobre su tamaño en ventas es significativamente mayor. Estos resultados indican que el tamaño de la empresa del deudor juega un papel relevante en la evaluación del riesgo crediticio. Las empresas grandes tienden a presentar un menor riesgo, mientras que aquellas sin información disponible sobre su tamaño representan un riesgo más elevado para la empresa de factoring. Estos hallazgos destacan la importancia de considerar el tamaño de la empresa como un factor clave en el análisis de riesgo y la toma de decisiones adecuada

Con respecto a la evaluación de desempeño del modelo de Scoring, donde se logró un AUROC del 0,793. Es importante destacar que un gran porcentaje de las empresas que operan con el factoring corresponde a PYMES, las cuales bordan entre los 2-3 años de antigüedad, como es posible apreciar en el Gráfico 7. Una gran parte de estas empresas no han superado el llamado “Valle de la muerte” el cual es parte del ciclo de vida de una pyme en donde, por un período de tiempo, algunos proyectos son insolventes, pues los costos son mayores que los beneficios. Esto ocurre por varios factores, sobre todo durante los primeros meses o el primer año, donde se invierte, se contrata personal y servicios, pero no se logran las ventas para poder conseguir el equilibrio. Estas empresas presentan una gran probabilidad de incumplimiento debido a problemas de solvencia, por lo que un AUROC del 0,793 es una métrica de desempeño bastante alta considerando que gran parte de las empresas que operan con el factoring se encuentran en este ciclo de su vida.[19]

El tercer objetivo específico planteado se cumple satisfactoriamente pues se han definido puntajes de corte tanto de aceptación, comité y rechazo, permitiendo una cobertura del 83,49% de los documentos, que va en concordancia con el actual 75% de aceptación de los documentos por parte de la empresa sumándole además el apetito por riesgo que esta presenta.

Finalmente, el cuarto objetivo específico también ha sido cumplido satisfactoriamente, ya que se lograron especificar claramente tanto los beneficios monetarios como las ventajas operacionales que están asociados al modelo de Scoring propuesto. Al identificar y comunicar de manera precisa estos beneficios, se brinda una perspectiva más clara sobre el impacto positivo que el modelo puede generar en la empresa de factoring. Estos resultados permiten destacar el valor agregado del modelo y su relevancia en la mejora de la eficiencia operativa y en la toma de decisiones financieras más acertadas. En general, el cumplimiento de los objetivos planteados demuestra el éxito y la efectividad del proyecto en su totalidad.

Con base en el modelo de Scoring propuesto, se presentan las siguientes recomendaciones:

Integración del modelo a plataforma de la empresa: Se recomienda integrar el modelo directamente en la plataforma de la empresa, implementando un panel que facilite su uso de manera accesible y práctica. Al hacerlo, los usuarios podrán aprovechar el modelo de Scoring de manera eficiente, permitiendo una evaluación rápida y efectiva del riesgo crediticio de los clientes y deudores. Esta integración agilizará los procesos de toma de decisiones financieras y optimizará sus operaciones de factoring de manera más efectiva.

Capacitación del personal de operaciones y riesgo: Se recomienda brindar una capacitación al personal del área de operaciones y riesgo de la empresa sobre la utilización del modelo de Scoring. Esta capacitación permitirá que el equipo adquiera un conocimiento profundo sobre el funcionamiento del modelo propuesto, cómo interpretar sus resultados y cómo aplicarlos de manera efectiva en la toma de decisiones crediticias. La formación garantizará una implementación adecuada y una utilización óptima del modelo. Además, una mayor comprensión del modelo por parte del personal fortalecerá la confianza de este en la toma de decisiones, respaldando así la eficiencia operativa y la toma de decisiones financieras más informadas y fundamentadas.

Actualización continua del modelo: Se recomienda ir actualizando constantemente el modelo de Scoring, ya que los factores que determinan el incumplimiento varían constantemente, ya que el modelo cumple con un sistema de clasificación de riesgo PIT, el cual utiliza toda la información disponible de los clientes y deudores para asignarlos a grupos de riesgo. Este sistema de clasificación utiliza tanto variables sistémicas como idiosincráticas, variando el comportamiento de la PD (Probabilidad de default) de acuerdo con las fluctuaciones macroeconómicas, por lo que este sistema de clasificación tiende a ajustarse rápidamente a un entorno económico cambiante.

Mejora continua del modelo: Se recomienda mejorar la robustez del modelo utilizando una estrategia conjunta que combine el modelo de Scoring con llamados a la API de Equifax. La integración de datos adicionales provenientes de fuentes externas como Equifax permitirá enriquecer la información utilizada en el modelo, proporcionando una visión más completa y precisa del perfil crediticio de los clientes y deudores. Esta sinergia aumentará la capacidad de evaluación del riesgo y facilitará la toma de decisiones financieras más sólidas y fundamentadas. La combinación de ambas fuentes de información fortalecerá la efectividad del modelo y contribuirá a una gestión más eficiente del riesgo crediticio en la empresa de factoring.

Bibliografía

[1] C.M.F. (2017). ¿Qué es el Factoring? CMF [En línea] https://www.cmfchile.cl/educa/621/w3-article-27145.html [Fecha consulta: 7 de marzo del 2023]

[2] Secretaría de la Función Pública (2015) "¿Qué son los contratos Marco?". [En línea] https://www.gob.mx/sfp/documentos/que-son-los-contratos-marco#:~:text=Los%20Contratos%20Marco%20son%20una,regular%C3%A1n%20la%20adquisici%C3%B3n%20o%20arrendamiento [Fecha consulta: 10 de marzo del 2023]

[3] Biblioteca del Congreso Nacional de Chile (2004). Ley N°199983. https://bcn.cl/2f9gu

[4] Barros & Errazuriz (19 de octubre del 2022). "Conoce los detalles de la nueva Ley Fintech de Chile". [En línea] https://www.bye.cl/conoce-los-detalles-de-la-nueva-ley-fintech-de-chile/#:~:text=La%20Ley%20Fintech%20representa%20un,servicios%20financieros%20para%20los%20consumidores. [Fecha consulta: 10 de marzo del 2023]

[5] Maximiliano Villena (13 de octubre del 2022). "Construcción y comercio: los sectores que mantienen en alerta a los factoring no bancarios " [En línea] https://www.latercera.com/pulso-pm/noticia/construccion-y-comercio-los-sectores-que-mantienen-en-alerta-a-los-factoring-no-bancarios/KYIQAB5T6VB77IGD745MWHDS5E/ [Fecha consulta: 15 de marzo del 2023]

[6] Gutierrez Girault, Matias Alfredo (2007). *"Credit Scoring models: what, how, when and for what purposes"*. Banco Central de la República Argentina.

[7] Jaime Forteza S., Víctor Medina O., Carlos Pulgar A. (2018). *"Marco general para el diseño de métodos estándar de provisiones por riesgo de crédito"*. Superintendencia de Bancos e Instituciones Financieras Chile.

[8] Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado- Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia.

[9] Bravo, C., Maldonado, S., Weber, R. (2010). "Experiencias Prácticas en la Medición de Riesgo Crediticio de Microempresarios utilizando Modelos de Credit Scoring". Revista Ingeniería de Sistemas, Vol. 24, (pp. 69-88).

[10] Claudia L. Hernández G., Jorge E. Rodríguez R. (2008). *Preprocesamiento de datos estructurados, Structured Data Preprocessing.*

[11] Fernández Castaño, Horacio; Pérez Ramírez, Fredy Ocaris. (2005) *El modelo logístico: una herramienta estadística para evaluar el riesgo de crédito.* Revista Ingenierías Universidad de Medellín, vol. 4, núm. 6 (pp. 55-75).

[12] Guillermo Acuña. “Machine learning: cómo evaluar modelos de clasificación”. [En línea] https://www.percepcioneseconomicas.cl/data-science/categoria-machine-learning/ [Fecha consulta: 10 de abril del 2023]

[13] Blanca Avella Miravet (2021) “Mejora de las predicciones en muestras desbalanceadas” Universidad Autónoma de Madrid. (pp. 15-21).

[14]SII. “Personas Jurídicas y Empresas” [En línea] https://www.sii.cl/sobre_el_sii/nominapersonasjuridicas.html [Fecha consulta: 28 de marzo del 2023]

[15]Oracle® Fusion Cloud EPM Trabajo con Planning. “IQR (Rango intercuartílico)” [En línea]https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_IQR.htm#PFUSU-GUID-CF37CAEA-730B-4346-801E-64612719FF6B [Fecha consulta: 10 de mayo del 2023]

[16] “El algoritmo k-means aplicado a clasificación y procesamiento de imágenes” [En línea]
https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html#:~:text=K%2Dmeans%20es%20un%20algoritmo,suele%20usar%20la%20distancia%20cuadr%C3%A1tica. [Fecha consulta: 12 de mayo del 2023]

[17] María Durban. “Modelos Lineales Generalizados”. (pp. 14-25).

[18]Jorge Méndez González (2019). “Stepwise Regresión” [En línea] https://rpubs.com/jorge_mendez/609253#:~:text=La%20regresi%C3%B3n%20paso%20a%20paso,modelos%20con%20cientos%20de%20variables. [Fecha consulta: 26 de mayo del 2023]

[19] Soledad Araya. “Valle de la muerte: Cómo evitar un emprendimiento fracasado”. [En línea] https://blog.nubox.com/empresas/valle-de-la-muerte. [Fecha consulta: 22 de julio del 2023]

Anexos

Anexo A. Reglas modelo LINCE

ID	Regla	Tipo	Prioridad	% Match	Activa
2	Bloquea todo lo que no tenga Cliente o Deudor existente en Chita	Bloqueo	0	0	Si
3	Bloquea todo lo que Cliente o Deudor se encuentren bloqueados por Riesgo	Bloqueo	1	0,1187	Si
5	Acepta todo lo que sea Pronto Pago	Aceptación	2	0,0008	Si
4	Bloquea todo lo que tenga Alerta de Glosa	Bloqueo	3	0,0103	Si
6	Bloquea todo lo que sobrepase el límite de exposición al aceptar kyc	Bloqueo	4	0,1764	Si
7	Bloquea todo lo que tenga Cliente en Cluster new	Bloqueo	5	0,1416	Si
8	Bloquea todo lo que tenga rating Cordada C	Bloqueo	6	0,0212	Si
9	Bloquea todo lo que tenga menos de 15 días de plazo	Bloqueo	7	0,0615	Si
10	Bloquea todo lo que Deudor tenga documentos morosos Deudor sobre 10 días	Bloqueo	7	0,0443	Si
11	Bloquea documentos del Deudor que su ultima factura pagada con mora sobrepase la mediana de dias de mora de lo saldado	Bloqueo	8	0,028	Si
12	Bloquea todo lo que Deudor o Cliente se encuentren en quiebra	Bloqueo	9	0,0094	Si
13	Bloquea todo lo que Deudor sea Cliente en Chita	Bloqueo	10	0,0062	Si
14	Bloquea todo lo que el plazo es mayor a 120 días	Bloqueo	11	0,0016	Si
15	Bloquea todo lo que tiene más de 25 días de emisión y (mayor al 20% del stock cliente OR Mayor a 5MM)	Bloqueo	12	0,0303	Si
16	Bloquea todo lo que tiene plazo remanente (plazo promedio de pago deudor del indicador tooltip - días emisión) < 15 días	Bloqueo	13	0,0457	Si
17	Bloquea todo lo que el Deudor no ratifica y necesita confirmación de documentos	Bloqueo	14	0,0255	Si
18	Bloquea todo lo que Cliente tenga normalización o problemas en la gestión de cobranza	Bloqueo	15	0,0571	Si
19	Bloquea todo lo que el monto de Kyc es mayor al 500% de la mediana, mayor a $10.000.000 y no exista historial Cliente y Deudor mayo o igual a la factura en cuestión	Bloqueo	16	0,0098	Si
20	Bloquea todo lo que Deudor es Cluster F	Bloqueo	17	0,0094	Si
21	Bloquea todo lo que Deudor tiene mas de (máximo entre 20 y mediana dias de mora] dias de mora y no tiene documentos saldados en los últimos 20 días	Bloqueo	18	0	Si
22	Bloquea clientes C o inferior con mora mayor a 30 días, o Clientes B o superior con mora mayor a 45 días	Bloqueo	19	0,028	Si
23	Bloquea todo lo que sobrepase línea dinámica con operación aceptada	Bloqueo	20	0,1569	Si
24	Bloquea todo lo que sobrepase la concentración dinámica límite	Bloqueo	20	0,1166	Si
28	Bloquea todo clientes C o inferior con deudores C o inferior, donde ambos tienen trazabilidad 10% saldado historico cliente	Bloqueo	24	0,1139	Si
29	Bloquea todo lo que Cliente ha traído facturas en el mes por más de 5X veces su venta promedio registrada y la factura es mayor a $10.000.000	Bloqueo	25	0,0042	Si
41	Bloquea todo lo que tiene más de 14 días de emisión y el deudor es cluster NEW o es II.EE.	Bloqueo	26	0,0286	Si
42	No pasa directo nada donde hayan rechazos o anulaciones con el deudor y no tenga méritos.	Bloqueo	27	0,0017	Si
57	No pasa directo nada donde hay un rechazo o anulación con el deudor, posterior al último mérito.	Bloqueo	28	0,0266	Si
44	No pasa nada directo que en los últimos 3 meses de facturación disponibles tenga más del 20% de los últimos documentos facturados (DTE) rechazados o anulados.	Bloqueo	29	0,0449	Si
45	No pasa directo nada en que la última fecha de operación del cliente haya alguna rechazada o anulada.	Bloqueo	30	0,0172	Si
46	No pasa nada directo en que más del 20% de los últimos 25 documentos operados por el cliente tenga NC de anulación o Rechazo.	Bloqueo	31	0,0152	Si
47	No pasa nada directo que tenga un socio bloqueado.	Bloqueo	32	0,0021	Si
49	No pasa nada directo que tenga deudor con mal pago igual o superior al 50%	Bloqueo	33	0,0144	Si
56	No pasa nada directo cuando mediana de mora de todos los pagos del deudor es mayor a 20 días.	Bloqueo	34	0,0192	Si
58	No pasa nada directo que tendencia de pago del deudor (plazo de pago mínimo entre último documento saldado cliente-deudor y últimos 5 saldados deudor) sea más de 20 días superior al plazo operado.	Bloqueo	34	0,0777	Si
66	No pasa nada directo donde plazo es superior a 20 ds de la mediana pago deudor	Bloqueo	35	0,0062	Si
68	Bloquea toda compañia que este con Cobranza Judicial (Iniciación, Notificación, Embargo)	Bloqueo	36	0,0001	Si
67	No pasa directo nada que tenga mas del 50% de facturas saldadas con fondos provenientes del cliente	Bloqueo	37	0,0014	Si
30	Fórmula de escalamiento. Acepta automáticamente por monto deudor KYC + documentos vigentes, morosos y giro pendiente Deudor hasta X valor calculado por la fórmula en base a parámetros de par cliente-deudor	Aceptación	38	0,5711	Si
50	Acepta aumáticamente cuando monto no sobrepasa monto historico saldado cliente-deudor, deudor es Grande 1 o superior, más de 10 facturas saldadas Cliente-Deudor y monto 6mes OR 6-10 facturas saldadas cliente-deudor y monto 3mes OR 3-5 facturas saldadas cliente-deudor, monto 2mes OR 2 facturas saldadas cliente-deudor y monto 1mes OR 1 factura saldada cliente-deudor y monto <2.5MM.	Aceptación	39	0,1926	Si

51	Acepta automáticamente cuando Cliente y Deudor B o superior, monto no sobrepasa monto historico saldado cliente-deudor, deudor es Grande 1 o superior, más de 10 facturas saldadas Cliente-Deudor y monto 6mes OR 6-10 facturas saldadas cliente-deudor y monto 3mes OR 3-5 facturas saldadas cliente-deudor y monto 2mes OR 2 facturas saldadas cliente-deudor y monto 1mes OR 1 factura saldada cliente-deudor y monto <10MM OR monto inferior al 30% del saldado histórico del cliente.	Aceptación	40	0,0244	Si
52	Acepta automáticamente cuando Cliente y Deudor C o superior, monto no sobrepasa monto historico saldado cliente-deudor, deudor es Grande 1 o superior, más de 10 facturas saldadas Cliente-Deudor y monto 6mes OR 6-10 facturas saldadas cliente-deudor y monto 3mes OR 3-5 facturas saldadas cliente-deudor y monto 2mes OR 2 facturas saldadas cliente-deudor y monto 1mes OR 1 factura saldada cliente-deudor y monto <3MM.	Aceptación	41	0,1053	Si
53	Acepta automáticamente Deudores B o superior, monto inferior al 20% del saldado histórico (independiente del deudor) del Client, deudor es Grande 1 o superior	Aceptación	43	0,1779	Si
54	Acepta lo que Suma Cliente-Deudor inferior a 2% del saldado histórico total Cliente, cliente con documento saldado por Deudor hace menos de 60 días y deudor con calificacion B o superior y deudor Grande 1 o superior en SII.	Aceptación	46	0,0587	Si
69	Acepta automaticamente Deudor Grande 1 o superior , Clientes B o superior, Deudores A o superior, monto < 5MM OR Clientes y Deudores B o superior, monto < 3MM OR Clientes C y Deudores A o superior, monto < 3MM OR Clientes C y Deudores B o superior, monto < 2MM OR Clientes D y Deudores A o superior, monto < 2MM OR Clientes D y Deudores B o superior, monto < 1MM OR Clientes E y Deudores A o superior, monto < 1MM OR Clientes E y Deudores B o superior, monto < 0,5MM OR Clientes E, Deudores A o superior, Monto < 0,5MM (requiere historial)	Aceptación	47	0,0078	Si
70	Acepta automáticamente cuando Cliente B o superior, Deudor IIEE TOP [con historial en Chita, mal pago 6mes OR 6-10 facturas saldadas cliente-deudor y monto 3mes, OR 3-5 facturas saldadas cliente-deudor y monto 2mes OR 2 facturas saldadas cliente-deudor y monto 1mes OR 1 factura saldada cliente-deudor y monto <10MM OR monto inferior al 30% del saldado histórico del cliente.	Aceptación	48	0,0001	Si
71	Acepta automáticamente cuando Cliente C o superior, Deudor IIEE TOP, monto no sobrepasa monto historico saldado cliente-deudor, más de 10 facturas saldadas Cliente-Deudor y monto 6mes OR 6-10 facturas saldadas cliente-deudor y monto 3mes OR 3-5 facturas saldadas cliente-deudor y monto 2mes OR 2 facturas saldadas cliente-deudor y monto 1mes OR 1 factura saldada cliente-deudor y monto <3MM.	Aceptación	49	0,0005	Si
72	Acepta lo que Suma Cliente-Deudor inferior a 2% del saldado histórico total Cliente, cliente con documento saldado por Deudor hace menos de 60 días y deudor IIEE TOP	Aceptación	50	0,0002	Si
73	Acepta automáticamente Deudores IIEE TOP, monto inferior al 20% del saldado histórico (independiente del deudor) del Client.	Aceptación	51	0,0019	Si
74	Acepta automáticamente cuando monto no sobrepasa monto historico saldado cliente-deudor, Deudor IIEE TOP, más de 10 facturas saldadas Cliente-Deudor, monto 6mes OR 6-10 facturas saldadas cliente-deudor, monto 3mes OR 3-5 facturas saldadas cliente-deudor, monto 2mes OR 2 facturas saldadas cliente-deudor, monto 1mes OR 1 factura saldada cliente-deudor, monto <2.5MM."	Aceptación	52	0,0011	Si
75	Acepta automaticamente Deudor Grande 1 o superior, Cliente F, Deudores A o superior, Monto 0.	Aceptación	53	0	Si
77	Acepta automaticamente Deudor IIEE TOP, Clientes B o superior, monto < 5MM OR Clientes C, monto < 3MM OR Clientes D, monto < 2MM OR Clientes E, monto < 1MM	Aceptación	64	0,0014	Si

Anexo B. Resumen variables base “Documents”

Nombre de variable	Tipo de variable
# Op	numeric
# Doc	numeric
Período Operación	character
Fecha Operación	Date
Estado	character
Folio	numeric
Días Emitido	numeric
Plazo	numeric
Días Mora	numeric
Fecha Exp	POSIXct
Cliente	character
Rut Cliente	character
Tamaño Cliente	character
Deudor	character
Rut Deudor	character
Tamaño Deudor	character
Monto	numeric
% Anticipado	numeric
Rechazada_cliente	numeric
Anticipo Insoluto	numeric
Tasa Doc	numeric
Tasa Media Efectiva	numeric
Tasa Mora	numeric
Diferencia de Precio	numeric
Comisión Afecta	numeric
Comisión Exempta	numeric
IVA	numeric
Excedentes	numeric
Giro	numeric
Mora	numeric
Fecha pago excedente	Date
Fecha giro	Date
Fecha Saldado	Date
Fecha Saldado Deudor	Date
Monto Pagado	numeric
Saldo Adeudado	numeric
Excedente Pagado	character
Giro Pagado	character
Canal	character
Model Cluster Cliente	character
Static Cluster Cliente	character
Model Cluster Deudor	character
Static Cluster Deudor	character
Descuento Giro	numeric
Giro Pendiente	numeric
Descuentos Excedentes	numeric
Excedentes Remanentes	numeric
Fecha Castigo	logical
Fecha Primera Operacion Cliente	Date
Primera Operacion Cliente	character
Ejecutiv@ Comercial	character
Operador	logical
Rechazada_deudor	numeric
Ratificador@	logical
Tipo	character
Migración Tanner	logical
Enrolado	character
Fecha enrolado	character
Contrato Reparado	character
Contrato Express	character
Nota de crédito	character
Rechazado	character
Reintegro	character
Usuario	character
Fecha Facturación	POSIXct
Fecha Factura	POSIXct
Rating Cordada Deudor	character
Fecha Ratificación	character

Facturas saldadas	character
Historial_operadas_cliente	numeric
Operadas_12_meses_cliente	numeric
Operadas_6_meses_cliente	numeric
Operadas_3_meses_cliente	numeric
Historial_saldadas_cliente	numeric
Saldadas_12_meses_cliente	numeric
Saldadas_6_meses_cliente	numeric
Saldadas_3_meses_cliente	numeric
Ratio_saldadas_cliente	numeric
Ratio_12_meses_cliente	numeric
Ratio_6_meses_cliente	numeric
Ratio_3_meses_cliente	numeric
Historial_operadas_deudor	numeric
Operadas_12_meses_deudor	numeric
Operadas_6_meses_deudor	numeric
Operadas_3_meses_deudor	numeric
Historial_saldadas_deudor	numeric
Saldadas_12_meses_deudor	numeric
Saldadas_6_meses_deudor	numeric
Saldadas_3_meses_deudor	numeric
Ratio_saldadas_deudor	numeric
Ratio_12_meses_deudor	numeric
Ratio_6_meses_deudor	numeric
Ratio_3_meses_deudor	numeric
Exp_cliente	numeric
Exp_deudor	numeric

Anexo C. Tratamiento de valores NA

Variable	Tipo de variable	Porcentaje de valores nulos	Acción
Tamaño cliente	Continua	2,29%	Eliminar missing values
Zona cliente	Categórica	2,29%	Eliminar missing values
Sector de actividad economica cliente	Categórica	0,26%	Eliminar missing values
Tipo empresa cliente	Categórica	0,04%	Eliminar missing values
Tipo persona cliente	Categórica	0,04%	Eliminar missing values
Historial de facturas total operadas por el cliente	Continua	0%	Ninguna
Historial de facturas operadas en los últimos 12 meses por el cliente	Continua	0%	Ninguna
Historial de facturas operadas en los últimos 6 meses por el cliente	Continua	0%	Ninguna
Historial de facturas operadas en los últimos 3 meses por el cliente	Continua	0%	Ninguna
Historial de facturas saldadas por el cliente	Continua	0%	Ninguna
Historial de facturas saldadas en los últimos 12 meses por el cliente	Continua	0%	Ninguna
Historial de facturas saldadas en los últimos 6 meses por el cliente	Continua	0%	Ninguna
Historial de facturas saldadas en los últimos 3 meses por el cliente	Continua	0%	Ninguna
Antigüedad cliente SII	Continua	2,29%	Eliminar missing values
Antigüedad cliente en Chita	Continua	0%	Ninguna
Tamaño deudor	Continua	2,29%	Eliminar missing values
Zona deudor	Categórica	2,29%	Eliminar missing values
Sector de actividad economica deudor	Categórica	0,65%	Eliminar missing values
Tipo empresa deudor	Categórica	0,002%	Eliminar missing values
Tipo persona deudor	Categórica	0,002%	Eliminar missing values
Historial de facturas total operadas por el deudor	Continua	0%	Ninguna
Historial de facturas operadas en los últimos 12 meses por el deudor	Continua	0%	Ninguna
Historial de facturas operadas en los últimos 6 meses por el deudor	Continua	0%	Ninguna
Historial de facturas operadas en los últimos 3 meses por el deudor	Continua	0%	Ninguna
Historial de facturas saldadas por el deudor	Continua	0%	Ninguna
Historial de facturas saldadas en los últimos 12 meses por el deudor	Continua	0%	Ninguna
Historial de facturas saldadas en los últimos 6 meses por el deudor	Continua	0%	Ninguna
Historial de facturas saldadas en los últimos 3 meses por el deudor	Continua	0%	Ninguna
Antigüedad deudor SII	Continua	2,29%	Eliminar missing values
Monto	Continua	0%	Ninguna
Tasa Documento	Continua	0%	Ninguna
Porcentaje anticipado	Continua	0%	Ninguna
Rechazada cliente	Categórica	0%	Ninguna
Rechazada deudor	Categórica	0%	Ninguna

Anexo D. Matriz de correlación

Columna1	Monto	Tasa_doc	Historial_operadas_cliente	Operadas_12_meses_cliente	Operadas_6_meses_cliente	Operadas_3_meses_cliente	Historial_saldadas_cliente	Saldadas_12_meses_cliente	Saldadas_6_meses_cliente	Saldadas_3_meses_cliente
Monto	1,000	-0,033	-0,116	-0,156	-0,173	-0,185	-0,106	-0,146	-0,162	-0,166
Tasa_doc	-0,033	1,000	-0,087	-0,155	-0,149	-0,129	-0,078	-0,148	-0,135	-0,084
Historial_operadas_cliente	-0,116	-0,087	1,000	0,841	0,747	0,675	0,997	0,847	0,749	0,652
Operadas_12_meses_cliente	-0,156	-0,155	0,841	1,000	0,953	0,880	0,811	0,992	0,945	0,822
Operadas_6_meses_cliente	-0,173	-0,149	0,747	0,953	1,000	0,960	0,704	0,925	0,980	0,894
Operadas_3_meses_cliente	-0,185	-0,129	0,675	0,880	0,960	1,000	0,624	0,834	0,912	0,921
Historial_saldadas_cliente	-0,106	-0,078	0,997	0,811	0,704	0,624	1,000	0,826	0,719	0,624
Saldadas_12_meses_cliente	-0,146	-0,148	0,847	0,992	0,925	0,834	0,826	1,000	0,940	0,815
Saldadas_6_meses_cliente	-0,162	-0,135	0,749	0,945	0,980	0,912	0,719	0,940	1,000	0,917
Saldadas_3_meses_cliente	-0,166	-0,084	0,652	0,822	0,894	0,921	0,624	0,815	0,917	1,000
Historial_operadas_deudor	-0,092	-0,017	0,372	0,383	0,361	0,341	0,371	0,395	0,384	0,381
Operadas_12_meses_deudor	-0,113	-0,075	0,395	0,524	0,512	0,478	0,389	0,543	0,552	0,542
Operadas_6_meses_deudor	-0,120	-0,076	0,347	0,485	0,512	0,493	0,337	0,495	0,547	0,558
Operadas_3_meses_deudor	-0,121	-0,063	0,303	0,431	0,469	0,484	0,291	0,434	0,493	0,544
Historial_saldadas_deudor	-0,088	-0,014	0,367	0,374	0,349	0,327	0,368	0,387	0,374	0,369
Saldadas_12_meses_deudor	-0,107	-0,071	0,394	0,523	0,505	0,464	0,390	0,545	0,551	0,537
Saldadas_6_meses_deudor	-0,113	-0,069	0,349	0,487	0,511	0,484	0,342	0,504	0,558	0,569
Saldadas_3_meses_deudor	-0,119	-0,046	0,325	0,455	0,495	0,507	0,318	0,470	0,542	0,620
antiguedad_chita_cliente	0,004	0,168	0,456	0,245	0,177	0,136	0,470	0,262	0,198	0,167
antiguedad_SII_cliente	-0,206	-0,159	0,160	0,185	0,184	0,158	0,155	0,190	0,199	0,179
antiguedad_SII_deudor	0,059	0,104	-0,110	-0,147	-0,144	-0,138	-0,106	-0,147	-0,147	-0,141
Porcentaje_anticipado	-0,027	0,018	-0,009	-0,045	-0,046	-0,040	-0,009	-0,050	-0,057	-0,058
Rechazada_cliebte	0,003	0,151	0,025	0,087	0,057	0,128	0,025	0,103	0,194	0,051
Rechazada_deudor	-0,037	0,136	0,133	0,239	0,222	0,297	0,131	0,260	0,375	0,258

Columna1	Historial_operadas_deudor	Operadas_12_meses_deudor	Operadas_6_meses_deudor	Operadas_3_meses_deudor	Historial_saldadas_deudor	Saldadas_12_meses_deudor	Saldadas_6_meses_deudor	Saldadas_3_meses_deudor	Antiguedad_chita_cliente	Antiguedad_SII_cliente	Antiguedad_SII_deudor Porcen	Porcentaje_anticipado	Rechazada_cliente	Rechazada_deudor
Monto	-0,092	-0,113	-0,120	-0,121	-0,088	-0,107	-0,113	-0,119	0,004	-0,206	0,059	-0,027	0,003	-0,037
Tasa_doc	-0,017	-0,075	-0,076	-0,063	-0,014	-0,071	-0,069	-0,046	0,168	-0,159	0,104	0,018	0,151	0,136
Historial_operadas_cliente	0,372	0,395	0,347	0,303	0,367	0,394	0,349	0,325	0,456	0,160	-0,110	-0,009	0,025	0,133
Operadas_12_meses_cliente	0,383	0,524	0,485	0,431	0,374	0,523	0,487	0,455	0,245	0,185	-0,147	-0,045	0,087	0,239
Operadas_6_meses_cliente	0,361	0,512	0,512	0,469	0,349	0,505	0,511	0,495	0,177	0,184	-0,144	-0,046	0,057	0,222
Operadas_3_meses_cliente	0,341	0,478	0,493	0,484	0,327	0,464	0,484	0,507	0,136	0,158	-0,138	-0,040	0,128	0,297
Historial_saldadas_cliente	0,371	0,389	0,337	0,291	0,368	0,390	0,342	0,318	0,470	0,155	-0,106	-0,009	0,025	0,131
Saldadas_12_meses_cliente	0,395	0,543	0,495	0,434	0,387	0,545	0,504	0,470	0,262	0,190	-0,147	-0,050	0,103	0,260
Saldadas_6_meses_cliente	0,384	0,552	0,547	0,493	0,374	0,551	0,558	0,542	0,198	0,199	-0,147	-0,057	0,194	0,375
Saldadas_3_meses_cliente	0,381	0,542	0,558	0,544	0,369	0,537	0,569	0,620	0,167	0,179	-0,141	-0,058	0,051	0,258
Historial_operadas_deudor	1,000	0,715	0,632	0,576	0,998	0,702	0,614	0,576	0,121	0,077	-0,045	-0,066	0,020	0,212
Operadas_12_meses_deudor	0,715	1,000	0,914	0,824	0,694	0,990	0,902	0,835	0,065	0,156	-0,133	-0,117	0,120	0,398
Operadas_6_meses_deudor	0,632	0,914	1,000	0,913	0,598	0,877	0,971	0,894	0,036	0,167	-0,138	-0,114	0,036	0,334
Operadas_3_meses_deudor	0,576	0,824	0,913	1,000	0,529	0,752	0,815	0,912	0,019	0,135	-0,129	-0,101	0,023	0,327
Historial_saldadas_deudor	0,998	0,694	0,598	0,529	1,000	0,689	0,594	0,551	0,124	0,069	-0,041	-0,064	0,101	0,285
Saldadas_12_meses_deudor	0,702	0,990	0,877	0,752	0,689	1,000	0,896	0,813	0,071	0,150	-0,133	-0,118	0,034	0,305
Saldadas_6_meses_deudor	0,614	0,902	0,971	0,815	0,594	0,896	1,000	0,888	0,045	0,165	-0,144	-0,117	0,046	0,342
Saldadas_3_meses_deudor	0,576	0,835	0,894	0,912	0,551	0,813	0,888	1,000	0,040	0,145	-0,149	-0,109	0,032	0,365
antiguedad_chita_cliente	0,121	0,065	0,036	0,019	0,124	0,071	0,045	0,040	1,000	0,097	0,030	0,057	0,023	0,036
antiguedad_SII_cliente	0,077	0,156	0,167	0,135	0,069	0,150	0,165	0,145	0,097	1,000	-0,201	-0,068	0,102	0,065
antiguedad_SII_deudor	-0,045	-0,133	-0,138	-0,129	-0,041	-0,133	-0,144	-0,149	0,030	-0,201	1,000	0,033	0,023	-0,027
Porcentaje_anticipado	-0,066	-0,117	-0,114	-0,101	-0,064	-0,118	-0,117	-0,109	0,057	-0,068	0,033	1,000	0,098	0,062
Rechazada_cliebte	0,020	0,120	0,036	0,023	0,101	0,034	0,046	0,032	0,023	0,102	0,023	0,098	1,000	0,570
Rechazada_deudor	0,212	0,398	0,334	0,327	0,285	0,305	0,342	0,365	0,036	0,065	-0,027	0,062	0,570	1,000

Anexo E. Clasificación tamaño empresas según ventas anuales (SII)

1. **Sin Información:** corresponde a contribuyentes cuya información tributaria declarada, no permite determinar un monto estimado de ventas.
2. **1er Rango Micro Empresa:** 0,01 a 200,00 UF Anuales
3. **2do Rango Micro Empresa:** 200,01 a 600,00 UF Anuales
4. **3ro Rango Micro Empresa:** 600,01 a 2.400,00 UF Anuales
5. **1er Rango Pequeña Empresa:** 2.400,01 a 5.000,00 UF Anuales
6. **2do Rango Pequeña Empresa:** 5.000,01 a 10.000,00 UF Anuales
7. **3er Rango Pequeña Empresa:** 10.000,01 a 25.000,00 UF Anuales
8. **1er Rango Mediana Empresa:** 25.000,01 a 50.000,00 UF Anuales
9. **2do Rango Mediana Empresa:** 50.000,01 a 100.000,00 UF Anuales
10. **1er Rango Gran Empresa:** 100.000,01 a 200.000,00 UF Anuales
11. **2do Rango Gran Empresa:** 200.000,01 a 600.000,00 UF Anuales
12. **3er Rango Gran Empresa:** 600.000,01 a 1.000.000,00 UF Anuales
13. **4to Rango Gran Empresa:** más de 1.000.000,01 UF Anuales

Anexo F. Resultados modelo Logit

Columna1	Estimate S	td. Error	z value	Pr(>|z|)	Columna2
(Intercept)	-2,5994	0,089294	-29	< 2e-16	***
Monto 3	0,2497	0,054685	5	0,00000498	***
Monto 4	0,358146	0,047045	2	0,021116	*
Monto 8	0,627737	0,042596	6,329	2,47E-10	***
Monto 9	0,916882	0,051163	6	1,59E-08	***
Monto 10	0,575818	0,052654	-6	9,33E-11	***
Ratio_6_meses_cliente_categorizado25-49 %	-0,220556	0,033645	-7	5,55E-11	***
Ratio_6_meses_cliente_categorizado50-74 %	-0,683294	0,039055	-17	< 2e-16	***
Ratio_6_meses_cliente_categorizado75-100%	-1,014135	0,055476	-18	< 2e-16	***
antiguedad_SII_cliente_categorizado2	-0,265789	0,03505	-8	3,38E-14	***
antiguedad_SII_cliente_categorizado3	-0,482995	0,038607	-13	< 2e-16	***
antiguedad_SII_cliente_categorizado4	-0,510275	0,041226	-12	< 2e-16	***
antiguedad_chita_cliente_categorizado3	-0,198582	0,031709	-6	3,79E-10	***
R_cliente1	0,599156	0,030498	20	< 2e-16	***
R_deudor1	0,445261	0,030137	15	< 2e-16	***
Ratio_6_meses_deudor_categorizado25-49 %	-0,005349	0,038904	-0,137	0,890641	
Ratio_6_meses_deudor_categorizado50-74 %	-0,615192	0,039068	-16	< 2e-16	***
Ratio_6_meses_deudor_categorizado75-100%	-0,93771	0,038199	-25	< 2e-16	***
Tamaño_deudormediana	0,53585	0,045443	12	< 2e-16	***
Tamaño_deudormicro	0,753336	0,061457	12	< 2e-16	***
Tamaño_deudorpequeña	0,602135	0,048654	12	< 2e-16	***
Tamaño_deudorsin info	0,898877	0,047466	19	< 2e-16	***
Tamaño_clientemediana	-0,456022	0,089529	-14	< 2e-16	***
Tamaño_clientemicro	-0,243917	0,06869	-4	0,000384	***
Tamaño_clientepequeña	-0,483808	0,066184	-9	< 2e-16	***
Tamaño_clientesin info	0,178887	0,083222	2	0,031594	*
Zona_deudorNorte	0,167784	0,051331	3	0,001081	**
Zona_deudorSur	0,200971	0,038879	5	0,000000235	***
Sector_actividad_deudorSector secundario	0,333918	0,049981	7	2,37E-11	***
Sector_actividad_deudorSector terciario	0,536129	0,04908	11	< 2e-16	***