



UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

**DESARROLLO DE MODELO CREDIT SCORING PARA ADMISION DE
FACTURAS EN FACTORING**

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

PEDRO PABLO GONZÁLEZ LÓPEZ

PROFESOR GUÍA:

CARLOS EDUARDO PULGAR ARATA

MIEMBROS DE LA COMISIÓN:

JAVIER SUAZO SÁEZ

PATRICIO VALENZUELA AROS

SANTIAGO DE CHILE

2018

**RESUMEN DE LA MEMORIA PARA
OPTAR AL TÍTULO DE:** Ingeniero Civil
Industrial
POR: Pedro Pablo González López
FECHA: 29/01/2018
PROFESOR GUÍA: Carlos Pulgar Arata

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

La presente memoria consiste en el desarrollo de un modelo estadístico que permita predecir el default en el pago de créditos del producto factoring. La empresa solicitante del proyecto es una institución financiera no bancaria líder en el segmento de Pymes. En toda empresa que presta servicios financieros, incluyendo todo tipo de créditos, la gestión del riesgo es de los aspectos más relevantes. Por eso es pertinente contar con una herramienta que permita discriminar rápidamente y de manera automática entre buenos y malos pagadores.

De lo anterior se define como objetivo del estudio *“Construir un modelo predictivo para Factoring que estime la probabilidad de que un documento no sea pagado por el cliente o el deudor, para apoyar el proceso de admisión de operaciones”*.

La metodología para afrontar el problema corresponde a un entendimiento del negocio, obtención y preparación de la información, modelamiento y finalmente interpretación de los resultados. Para el proceso de modelación se utiliza la técnica de la industria bancaria llamada *credit scoring*, modificando la teoría para poder ser aplicada al factoring. Se cuenta con una base trabajada de datos, que contiene 1.2 MM de observaciones aproximadamente y 20 campos. Esta fue extraída y trabajada de los datos consolidados de las diferentes bases de datos con que cuenta la empresa.

El modelo final estimado entrega las probabilidades de no pago de cada documento, con una bondad de ajuste AUROC de 0,727 es bueno considerando un rango entre 0,5 siendo al azar y 1 que predice completamente bien. Este modelo integra 7 variables, las cuales son “porcentaje de descuento”, “contabilidad completa”, “Número de trabajadores”, “monto documento”, “tasa de operación”, “monto aprobado” y “tasa de operación”. Estas variables se encuentran categorizadas, con un puntaje asociado de acuerdo que tan riesgosas son.

El puntaje de cada variable se suma para lograr un score total del documento, distribuido entre 0 y 1000 para facilitar la interpretación, donde a mayor puntaje menos probabilidades de default tiene la factura. Este score se representa a través de una scorecard, donde se determina que todos los documentos bajo 378 puntos se rechazan, y los scores sobre 807 puntos se aceptan inmediatamente, enviado los documentos con puntaje entre estos dos cortes a un análisis más profundo para cada caso.

Los beneficios de esta herramienta contemplan una disminución en el número de operaciones analizadas directamente en un 28%, pudiendo liberar recursos para destinarlos en otras áreas. Además, para un mismo porcentaje de aceptación de solicitudes se rechazan un 15% de documentos riesgosos en comparación al modelo actual, lo que se impacta directamente en las provisiones, disminuyendo en \$4.800 millones anuales.

AGRADECIMIENTOS

Quiero agradecer a mi madre, de paciencia gigante y de corazón aún mayor, a mi padre, hombre justo del que he aprendido a apreciar las cosas simples de la vida. Ellos plantaron esta familia, de la cual empiezan a florecer ya en su totalidad los primeros frutos.

A mi hermanos, familia y amigos, que hacen de esta vida algo que valga la pena disfrutarla a concho.

A Pablo y Sebastián, que sin su ayuda este momento nunca hubiese llegado.

A mi familia putativa, que me acogió como uno más desde el día uno, gracias por su apoyo, paciencia, cariño y consejo.

A Javiera, compañera dentro y fuera de las trincheras, fue un largo camino, ipero lo logramos!!

A mis profesores Carlos Pulgar y Javier Suazo, además de José Manuel González y Osvaldo López, fundamentales para el desarrollo de esta memoria, gracias por las oportunidades, consejos, ayudas y buena disposición durante todo este proceso.

Finalmente agradezco a la Universidad de Chile, por formarme como persona y como ingeniero.

TABLA DE CONTENIDO

1. Introducción	1
1.1 Exposición del tema	1
1.1.1 Definiciones importantes	2
1.2 Justificación	4
1.3 Objetivos	8
1.3.1 Objetivo general	8
1.3.2 Objetivos específicos	8
1.4 Alcances y limitaciones del estudio	8
2. Marco teórico	9
3. Metodología	10
3.1 Tipos de modelos	14
3.2 Comparación de resultados	15
3.2.1 Matriz de confusión	15
3.2.2 Estadístico C o AUC	16
3.3 Tecnología	17
4. Recopilación y entendimiento de los datos	18
4.1 Datos iniciales	18
4.2 Estudio de variables existentes	18
4.3 Construcción de variable de objetivo	19
5. Preparación, transformación y selección de datos	20
5.1 Limpieza de datos	20
5.2 Análisis de los datos	22
5.3 Selección de variables	23
5.4 Transformación de variables	23
6. Aplicación del modelo	25
6.1 Análisis de variables	25
6.1.1 Primera iteración	26
6.1.2 Segunda iteración	30
7. Evaluación de resultados	35
7.1 Escalamiento a score	35

7.2	Generación del scorecard	41
7.2.1	Validación del scorecard	41
7.2.2	Generación de la herramienta de admisión	42
7.3	Cálculo de los beneficios	44
7.3.1	Determinación de puntaje de corte	44
8.	Conclusiones.....	52
8.1	Conclusiones generales.....	52
8.2	Conclusiones del trabajo realizado.....	53
9.	Recomendaciones	54
10.	Bibliografía	56
11.	Anexos	57
11.1	Anexo 1: Campos seleccionados y tratamiento de outliers	57
11.2	Anexo 2: NAAO.....	59
11.3	Anexo 3: Resumen Análisis Univariado	62
11.4	Anexo 4: Filtro de variables.....	63
11.5	Anexo 5: Matriz de covarianza	65
11.6	Anexo 6: Resultados Logit	68
11.7	Anexo 7: Distribución categorización variables continuas	69
11.8	Anexo 8: Tramos por punto de corte.....	70

ÍNDICE DE TABLAS

Tabla 1: Selección de tablas consolidado Tanner S.A (Fuente: elaboración propia en base a consolidado de Tanner S.A.)	18
Tabla 2: Porcentaje de Outliers y Extremos	22
Tabla 3: Variables seleccionadas. (Fuente: elaboración propia).....	23
Tabla 4: Betas calculados para primer modelo	27
Tabla 5: Betas calculados para cada variable y categoría	33
Tabla 6: Extremos ln(Odds) originales.....	36
Tabla 7: Nuevos extremos ln(odds)	36
Tabla 8: Puntaje calculado para cada variable de acuerdo a su categoría correspondiente	38
Tabla 9: Tabla de puntos de corte por rango	45
Tabla 10: Tabla de puntos de corte	47
Tabla 11: Promedio variables operación, ingresos y costos calculados	48
Tabla 12: Utilidades promedio por rango de Score	49
Tabla 13: Scorecard final	50

ÍNDICE DE ILUSTRACIONES

Ilustración 1: Diagrama de vínculos entre actores. (Fuente: elaboración propia).....	4
Ilustración 2: Proceso Knowledge Discovery in Databases (KDD). (Fuente: Elaboración propia en base a Fayyad, 1996)	11
Ilustración 3: Ejemplos de visualización de curva ROC.....	17
Ilustración 4: Flujo del proceso de Mora con sus posibles escenarios.....	20
Ilustración 5: Resumen de las distribuciones de las variables categorizadas	24
Ilustración 6: Prototipo funcional scorecard	43
Ilustración 7: Prototipo funcional scorecard	43
Ilustración 8: Prototipo funcional scorecard.....	44

1. Introducción

1.1 Exposición del tema

En el presente trabajo se enseña el trabajo realizado en conjunto con la empresa Tanner Servicios Financieros S.A, bajo el contexto del factoring, servicio ofrecido por la institución desde sus inicios.

Tanner es la institución financiera no bancaria líder en Chile en el segmento de pequeñas y medianas empresas (PYMES). La empresa inició sus actividades a principios de la década del 90 bajo el nombre de Bifactoring S.A., cuando comienza a desarrollarse el negocio de factoring en Chile. La compañía durante las décadas del 90 y comienzos del 2000 logra un crecimiento y expansión importante dentro del mercado financiero chileno, ofreciendo nuevos productos como crédito automotriz y leasing, instando a la empresa a que en 2005 cambie su nombre a FactorlineS.A. Finalmente en diciembre de 2011 cambian su nombre a Tanner Servicios Financieros.

Dentro de la empresa, el trabajo se desarrolla en la Sub Gerencia de Proyectos, área encargada de explorar e innovar en nuevas maneras de operar los negocios financieros, teniendo siempre como objetivo aumentar la rentabilidad sobre el capital (ROE) de los productos de la empresa.

La industria del factoring, al menos en sus aspectos más formales, tiene aproximadamente 20 años y desde su inicio se ha caracterizado por un continuo crecimiento. Las empresas de factoring se dividen en tres grandes grupos: i) las sociedades filiales de bancos, agrupadas en la Asociación Chilena de Factoring (Achef), a la que, por su origen y tamaño, está afiliada Tanner; ii) las independientes, estructuradas profesionalmente, muchas de ellas agrupadas bajo la Asociación Nacional de Empresas de Factoring (ANFAC) y algunas acogidas a la fiscalización de la Superintendencia de Valores y Seguros (SVS, en transición a la Comisión para el Mercado Financiero), y iii) sociedades con alto nivel de informalidad.

Desde el punto de vista legal, las empresas de factoring, puesto que carecen de un marco jurídico especial para operar, se sustentan básicamente en los códigos Civil y de Comercio, así como en la ley N° 19.983 del 15 de diciembre del 2004, que le otorga mérito ejecutivo a la factura.

En términos de participación, Tanner cuenta con un 8.3% del mercado de factoring, correspondiente a \$232 mil millones (repartidos en 2,882 clientes) de un mercado total

de \$3.17 billones¹, siendo la líder en instituciones no bancarias. Gráficamente se puede apreciar la participación de los actores principales en el “**Gráfico 1**”:

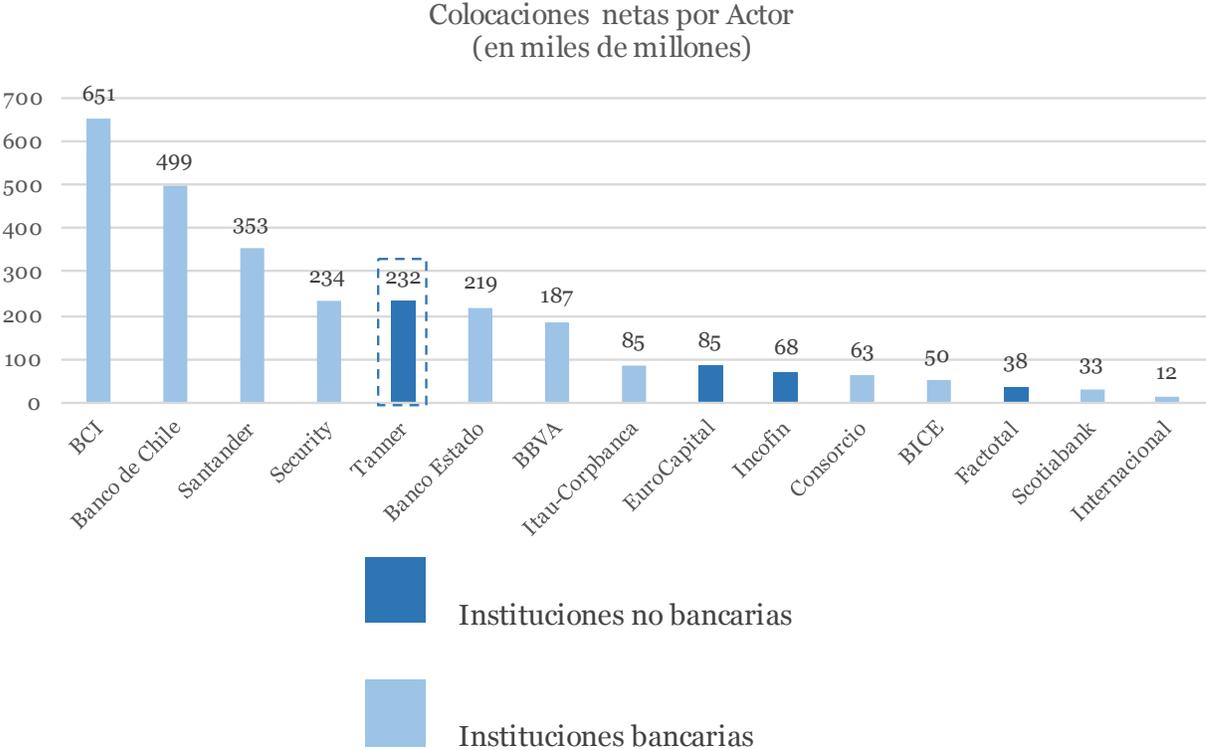


Gráfico 1: Colocaciones netas por actor. (Fuente: elaboración propia con datos de resultados públicos, junio 2017)

1.1.1 Definiciones importantes

Previo a la exposición del tema de investigación del producto financiero es importante definir los términos claves de este:

¹ Millones de millones.

- **Factoring:** Herramienta financiera que permite al cliente transformar en dinero efectivo sus cuentas por cobrar (generalmente cheques, facturas y letras), a través de lo que se denomina “cesión” de los documentos.
- **Ciente:** Persona natural o jurídica que solicita factorizar (vender) su cuenta por cobrar a la empresa financiera, con el objetivo de obtener liquidez inmediata.
- **Deudor:** Persona natural o jurídica al cual fue emitido el documento por el cliente, responsable del pago del documento en el plazo acordado con su proveedor, este que a su vez sería el cliente de la empresa de factoring.
- **Porcentaje de anticipo:** Parte del monto total del documento que se anticipa a través del factoring. Sobre este monto se calcula el interés de la operación, además de restar las comisiones y otros gastos. El resultado es el monto transferido al cliente.
- **Excedentes:** Monto restante del valor del documento que no se anticipó. Se devuelve al cliente una vez saldado el préstamo.
- **Tasa de operación:** Tasa de interés aplicada sobre el monto de anticipo del documento.
- **Cesión de responsabilidad:** El cliente otorga (cede) el mérito ejecutivo del documento a la empresa de factoring. Esto da seguridad jurídica al cobro de los documentos.
- **Fecha de vencimiento:** Fecha pactada con el cliente para que el préstamo quede saldado.
- **Línea de crédito del cliente:** Monto que puede factorizar el cliente. Esta línea es de uso interno (el cliente no conoce su cupo) y tiene como objetivo administrar de mejor manera el riesgo de cada cliente.
- **Mora:** Atraso en el pago del crédito, ya sea por no pago del cliente o del deudor, de acuerdo a la fecha de vencimiento acordada con la empresa.
- **Default:** Incumplimiento del pago. Generalmente se considera cuando la mora es mayor o igual a 90 días.

De manera ilustrativa se presenta a continuación en la “**Ilustración 1**” la relación entre los actores que participan del proceso del factoring:

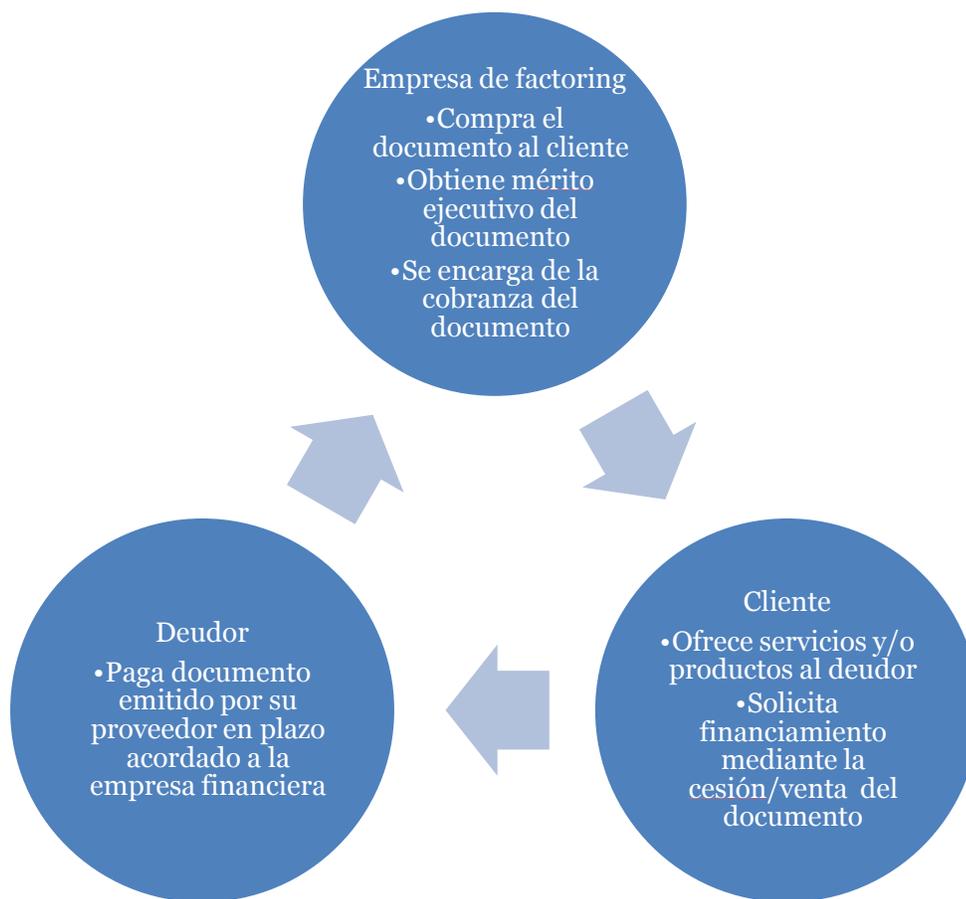


Ilustración 1: Diagrama de vínculos entre actores. (Fuente: elaboración propia)

En resumen, el cliente cede estos documentos a la empresa de factoring, quien le anticipa un porcentaje del monto indicado en el documento hasta la fecha de pago pactada previamente con la empresa deudora, obteniendo de esta forma liquidez inmediata.

1.2 Justificación

El principal desafío de las entidades financieras que ofrecen créditos masivos es administrar adecuadamente el riesgo de crédito, que corresponde a la posibilidad de sufrir una pérdida como consecuencia de un impago por parte de sus clientes en una operación financiera, dicho de otra manera, es el riesgo de que no pague el préstamo.

De lo anterior, se desprende la importancia de contar con una herramienta que permita discriminar rápidamente un gran número de solicitudes, manteniendo un porcentaje de default que sea coherente con el apetito por riesgo de la empresa.

Estas herramientas o modelos analíticos intentan predecir, modelar y discriminar entre buenos y malos pagadores, tanto de personas que como en empresas, de acuerdo a un puntaje estimado a partir de información de los clientes. Estos modelos han sido un problema estudiado ampliamente en la literatura, donde los primeros trabajos remontan a los años 60, con las investigaciones de Beaver (1996) y Altman (1968).

Aunque si bien los modelos de Credit Scoring han sido estudiados repetidamente en la literatura, su aplicación en factoring no es directa, debido principalmente a tres puntos:

- El pago de una cuota única (en la mayoría de los casos), a diferencia de otros créditos estudiados como pueden ser créditos de consumo o hipotecarios.
- La responsabilidad de dos actores distintos en el pago de la cuota, al existir un deudor, y a su vez, un cliente que tiene responsabilidad de pagar en caso de que el deudor no cancele la deuda.²
- La evaluación de la solicitud se hace de acuerdo a la información otorgada por el cliente. No existen métodos fidedignos de obtener información del deudor para evaluarlo.

Este segundo punto es de suma importancia para el estudio, ya que a diferencia de la metodología estándar de credit scoring, la herramienta o modelo de admisión no discriminará solamente de acuerdo a si el cliente es buen o mal pagador, si no, más bien lo hará de acuerdo a un documento en específico. Este tipo de modelos de scoring no ha sido muy investigado en la literatura de riesgo crédito.

Para ejemplificar lo anterior de mejor forma, un documento puede tener un cliente con alto grado de riesgo, pero el deudor del documento puede tener historia crediticia sólida, por lo que no es trivial rechazar la operación analizando al cliente solamente. Dicho de

² Esto es considerando el caso de factoring con responsabilidad del cedente. En caso de ser una operación sin responsabilidad, esta debe analizarse como un crédito de una sola cuota con el deudor del documento.

otro modo, los factores con un mayor poder discriminatorio utilizados en la literatura podrían no tener el mismo peso en el desenlace del otorgamiento del crédito.

Además de poder distinguir entre buenos casos y malos casos, como se menciona en el párrafo anterior, la discriminación debe considerar la variable tiempo, debido a que generalmente esta última es relevante en la ecuación de utilidad de los clientes, lo que se sustenta en el hecho de que el valor agregado del factoring como producto sea entregar una rápida liquidez, por lo que una demora en el otorgamiento, mayor a la esperada por parte del solicitante, puede incurrir en la fuga de un buen cliente, empujándolo a la competencia.

Ligado también con la variable tiempo, en la legislación chilena, el documento se vuelve título ejecutivo y se ratifica de forma automática pasado los 8 días de emitida. Previo a ese plazo, el receptor de dicha factura puede rechazarla sin ningún tipo de impedimento.

Volviendo el foco a Tanner, el otorgamiento de los créditos en la actualidad funciona de la siguiente manera:

- NAAO:
 - Motor de 17 reglas establecidas por el departamento de riesgo. (anexo 9.4)
 - Si la operación cumple con todas las reglas, se acepta automáticamente la operación.
 - Si no cumple la condición anterior, la operación pasa a ser analizada por un analista de riesgo.
 - Analista de Riesgo:
 - Determina si la operación es aceptada o, en su defecto, rechazada. Es un análisis subjetivo de acuerdo con la experiencia que tenga el analista.
 - En caso de que se requiera un mayor grado de análisis, la operación es estudiada por un comité de riesgo, en donde ejecutivos con más experiencia en la empresa determinan si proceden a aceptar o no la solicitud.

Sobre el motor de reglas NAAO es importante mencionar que todas las solicitudes ingresan al motor, siendo solo el 15% de las solicitudes aceptadas por la herramienta, donde el 85% restante queda con estado de “rechazada” o en “observación” debiendo ser estudiadas por analistas. Esto tiene su beneficio, debido a que la intención del sistema es

aceptar las operaciones muy buenas inmediatamente de acuerdo a las reglas estrictas, pero limita la liberación de recursos del área de riesgo, principalmente de horas hombre.

De lo anterior se desprenden los siguientes problemas:

- Subjetividad del analista en el estudio de riesgo de la operación.
- Proceso lento, mermando parte del valor del producto de cara al cliente.
- Proceso costoso, debido a la necesidad de mantener un importante número de mano de obra calificada encargada del análisis de riesgo.

Si a lo anterior se le suma las expectativas de crecimiento que tiene el mercado del factoring, según una noticia presentada en la página web de la cámara de diputados, “Comisión de Economía analizó la industria del factoring con asociación gremial”, es esperable que aumente el número de solicitudes hacia la empresa, por lo que es posible enfrentarse a un escenario en el cual habrá que evaluar un trade off entre aumentar el tiempo de respuesta de las operaciones (mayor número de solicitudes por analista), o un aumento en la cuota de analistas, aumentando los costos, a costa de mantener el tiempo de respuesta.

Es por esto, que contar con una herramienta capaz de discriminar rápidamente entre operaciones rentables para la compañía de las que no, se alinea de buena forma con los objetivos de la empresa, pudiendo llegar a ser una importante herramienta de gestión.

Finalmente se menciona como beneficio extra del modelo, poder determinar con fundamentos una tasa mínima de interés que debe ser considerada para cada operación, ajustándola empíricamente al riesgo.

1.3 Objetivos

1.3.1 Objetivo general

Se define el objetivo general como:

- Construir un modelo predictivo para Factoring que estime la probabilidad de que un documento no sea pagado por el cliente o el deudor, para apoyar el proceso de admisión de operaciones.

1.3.2 Objetivos específicos

- Identificar variables de mayor incidencia en la probabilidad de default de un documento.
- Obtener un modelo o un conjunto de reglas que permitan estimar la probabilidad de default de una operación, de acuerdo a características del cliente, deudor y otras variables operacionales.
- Cuantificar beneficios proyectados con implementación de nuevo modelo.

1.4 Alcances y limitaciones del estudio

Este trabajo considera la construcción de un modelo de admisión para el producto de factoring en una empresa no bancaria.

Los datos a utilizar corresponden a los datos que se encuentran en el consolidado de la gerencia de sistemas, obtenidos de las distintas bases de datos con las que cuenta la empresa.

Los documentos a estudiar corresponden a facturas y facturas electrónicas, no incluyendo cheques y letras.

Las operaciones para el estudio de la memoria son con cesión de la factura con responsabilidad del cliente, es decir, el cliente es el responsable final del pago del crédito

en caso de que la empresa deudora no pague. Esto se hace oficial mediante la firma de contrato marco previa a la operación, donde el cliente acuerda asumir responsabilidad jurídica del crédito. Una de las características que tiene este contrato es que debe ser firmado por los representantes de las empresas como avales. Este contrato se firma solamente una sola vez, y tiene validez legal durante varios años.

Cabe destacar que para el caso de factoring sin responsabilidad de pago por parte del cedente, se debe estudiar considerando la operación como un crédito de una sola cuota con el deudor como contrapartida.

Las probabilidades de default que se obtengan como resultado tendrán que ser actualizadas durante el tiempo. Si se quiere predecir la probabilidad para otros períodos de tiempo futuro, será necesario modificar el modelo de acuerdo a la nueva información, ya que se espera un aumento en el número de empresas que obtengan acceso al factoring.

2. Marco teórico

Hasta recientemente, la decisión de entregar créditos se basaba (y se basa) en el juicio humano para determinar el riesgo de que un cliente pagara o no un crédito de acuerdo a las distintas características de este. No obstante, el aumento en la demanda de créditos ha obligado a desarrollar herramientas y modelos para facilitar la decisión de los agentes otorgadores sobre quién es aceptable de recibirlo y quién no.

En los últimos años, el aumento de la competencia en la industria financiera, los avances tecnológicos (sobre todo en computación) y el crecimiento acelerado de las bases de datos, han llevado a estos métodos a transformarse en herramientas relevantes en la industria, y en caso de ser aplicadas correctamente, elementos diferenciadores de la competencia.

Credit Scoring se define formalmente como un método cuantitativo que se utiliza para predecir la probabilidad de que un aspirante a crédito o un cliente de la entidad crediticia existente deje de pagar el crédito o bien, no lo haga una vez que lo reciba (4). Su objetivo es ayudar a los proveedores de créditos a cuantificar y manejar el riesgo financiero relacionado con el otorgamiento de los préstamos, para así tomar decisiones de forma rápida y objetiva.

Junto a lo anterior, el Credit Scoring tiene beneficios también para el solicitante de crédito, ya que le permite entender el por qué su solicitud fue aceptada o rechazada, disminuyendo la discriminación por características de la persona (como por ejemplo sexo o religión) y reduciendo la subjetividad en el análisis.

El problema que resuelve el Credit Scoring corresponde a definir si un cliente que presenta un conjunto de características X va a caer en default. Es decir, se busca predecir la ocurrencia de un fenómeno binario y , el cual sería si el cliente cae en default o no, con un margen de error razonable. (5)

De forma matemática:

$$p(y = falla|X) = f(X) \quad (1)$$

Donde $p(y = falla|X) = f(X)$ corresponde a la probabilidad que la empresa caracterizada por X no cumpla su compromiso financiero y $f(X)$ corresponde a una función que aproxima la probabilidad a partir de los datos disponibles.

3. Metodología

La teoría tras el contexto mencionado en el punto anterior radica en el desarrollo de modelos estadísticos que permitan dividir a los clientes que solicitan créditos en clientes “buenos” y “malos” dependiendo de la tendencia que puedan tener en presentar un retraso en el pago de alguna cuota correspondiente. Para ello es necesario realizar una estimación de la probabilidad de default de cada cliente para luego segmentarlo según alguna regla que determine el corte de la clasificación.

Para poder estimar esta probabilidad de default (PD) es necesario construir un modelo que pueda analizar grandes cantidades de datos, los cuales pueden incluir historial de comportamiento del cliente o características demográficas, por ejemplo.

Se utiliza en primera instancia la metodología Knowledge Discovery in Databases (KDD) el cual es un proceso para identificar patrones útiles y comprensibles, pudiendo interpretar grandes cantidades de información y encontrar relaciones que faciliten o aporten en la toma de decisiones (Fayyad, 1996).

A continuación, se presenta el diagrama del proceso KDD en la “**Ilustración 2**”:

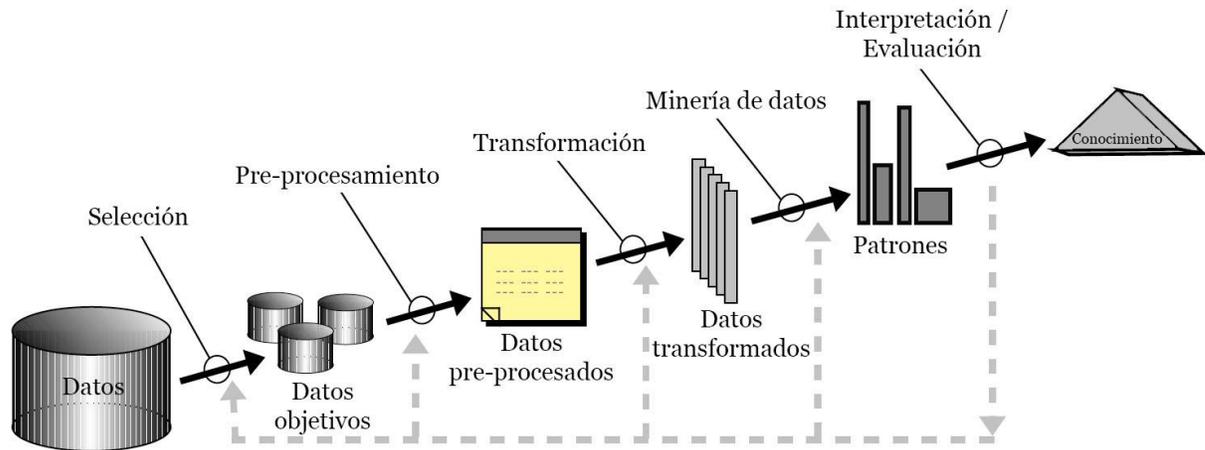


Ilustración 2: Proceso Knowledge Discovery in Databases (KDD). (Fuente: Elaboración propia en base a Fayyad, 1996)

En primer lugar, es importante tener un entendimiento del negocio, considerando el conocimiento relevante y objetivos de la investigación.

A continuación, se presentan los detalles en las etapas de la ilustración 2:

Selección: Consiste en la selección del conjunto de datos, del subconjunto de variables y de la muestra de datos sobre la cual se va a realizar el descubrimiento.

Pre procesamiento: Involucra la limpieza y pre procesamiento de los datos. Se deciden las estrategias sobre la forma en que se van a manejar los campos de datos no disponibles o vacíos, se realizan cambios fundamentales a la base de datos y se analiza la información.

Transformación: Consiste en encontrar otras representaciones de los datos y reducir el número de variables a ser consideradas, encontrando las características más significativas.

Data Mining: Se debe decidir el objetivo del modelamiento (regresión, clasificación o agrupamiento), esto con la finalidad de definir el algoritmo que se aplicará. Incluye tanto la selección de métodos a ser utilizados para la búsqueda de patrones en los datos, como la decisión sobre que parámetros son los más apropiados para cada modelo.

Interpretación/ Evaluación: Interpretación de los patrones encontrados, determinando su impacto para el negocio y como se puede aplicar el nuevo conocimiento. Incluye encontrar puntos de corte, cotas y reglas de decisión.

Este proceso puede iterarse durante el transcurso del proyecto, y se pone especial atención en la preparación de los datos, fase clave para asegurar la veracidad de los resultados.

El siguiente paso es definir una clasificación de los pagadores en dos grupos, los “buenos” y “malos”. Esta clasificación tiene que ser de acuerdo a un planteamiento estratégico acorde con los objetivos de la empresa, como, por ejemplo, que los casos “malos” sean aquellos que no sean rentables para la compañía.

La definición debe ser precisa para que sea posible identificar correctamente el comportamiento de los clientes. Si estas clasificaciones son muy restringidas pueden ocasionar que las muestras de este segmento sean muy pocas, o una clasificación muy amplia no generaría diferenciación, provocando que el score sea inútil.

Además, la definición debe ser fácil de entender, interpretar y monitorear, ya que no hay que perder el foco que esto es una herramienta que simplifica la toma de decisiones.

Finalmente, Basilea II, para el mercado bancario, establece que el default corresponde a 90 días de atraso en el pago de una cuota. Si bien esto sirve como una referencia inicial, la empresa, al no tener carácter bancario, no se ve afectado de la misma forma por este acuerdo, por lo que esta definición de default puede ser ampliada y probada para distintos escenarios, con el fin de obtener evaluar el efecto que esta pueda tener sobre el modelo.

Una vez con los datos seleccionados, estos se pueden reprocesar, el cual incluye eliminar variables que no tienen impacto en el problema, ya sea por razones de poca varianza, muchos valores nulos u otros. Este paso suele ser el de mayor duración en los proyectos de Data Mining (Fayyad, 1996). Las etapas que sigue este paso son:

Descripción de los datos: Todas las variables deben ser examinadas y descritas utilizando herramientas estadísticas. Esto para ayudar en el entendimiento y evaluar la calidad de la data.

Limpieza de datos: cada variable debe ser explorada y normalizada, eliminando los datos que no corresponden a las variables.

Inconsistencias en datos adicionales: es posible que en la base de datos existan algunos datos que presentan inconsistencias y deben ser considerados, o poseen ciertas variables que necesitan ser consolidadas.

Una vez con la base limpia debe definirse las variables que tengan mayor impacto en el proyecto. Los siguientes son opciones para poder discriminar entre variables, minimizando el riesgo de eliminar una variable importante:

- Análisis de la concentración de variables en un valor y variables con valores perdidos.
- Análisis univariado. Esto para eliminar las variables que no puedan discriminar entre los casos “buenos” de los “malos”
- Análisis Multivariado, con el objetivo de determinar cuáles variables no aportan en caso de incluirlas en el modelo final.

El último paso antes de analizar los modelos de Data Mining corresponde a la transformación de variables. La mayoría de los modelos necesitan que los inputs sean numéricos. Este paso permite que la data esté en el formato correcto. Algunas transformaciones comunes son:

Normalización: Atributos continuos como por ejemplo la edad o los ingresos necesitan ser re escalados a un intervalo común, para que cada variable tenga el mismo peso en el modelo.

WOE (Weight of Evidence): Se asigna a una variable el peso según la cantidad de información que aporta. El rango varía entre 0 y 1.

Ajuste de escala: Un ejemplo es la utilización del logaritmo de la variable, ya que permite expandir la escala de medida, como por ejemplo en ingresos, pudiendo observar diferencias en variables con datos muy condensados.

Agregación: Algunas variables pueden aumentar su impacto cuando se agregan junto a otras. Como por ejemplo los costos totales cuando se le agrega a la variable de costos fijos.

Ya con este paso completo se puede aplicar las técnicas de Data Mining. Este paso usualmente implica iteraciones utilizando distintas configuraciones de modelos, parámetros y variables.

Existen dos tipos principales de modelos: aprendizaje supervisado, cuyo objetivo corresponde a crear una función que permita predecir si un nuevo caso es “bueno” o malo”, utilizando datos de casos conocidos. El segundo tipo corresponde al aprendizaje no supervisado, donde el objetivo es descubrir estructuras o patrones en la data misma, sin ningún evento en particular. Existen también nuevas aplicaciones que mezclan estos dos tipos, que utiliza información adicional para enriquecer data no supervisada (6).

Con esta división, el proceso de escoger un modelo puede ser dividido en:

Selección de atributos finales: No todas las variables seleccionadas en los pasos previos entregan un aumento en la significancia en un modelo multivariado. Un método clásico corresponde a la selección de todas las variables y elimina una a la vez hasta que no ocurre un mejoramiento de acuerdo a test estadísticos (Bravo, Maldonado, Weber, 2010).

Ajuste de parámetros y selección de modelo: Se escoge el modelo de acuerdo a la dificultad del problema, tomando en cuenta la complejidad requerida, tiempo computacional disponible, entre otras. Es común dejar un 20% de los datos fuera del modelo para ver cómo se comporta después.

Finalmente, un modelo es validado cuando los resultados de las pruebas finales son consistentes, la varianza de los parámetros es robusta, y que las variables son razonables dentro del negocio (Fayyad, 1996).

3.1 Tipos de modelos

Existen muchas alternativas de modelos. Entre ellas una de las más utilizadas es la Regresión Logit. Su uso radica en la ventaja de medir la probabilidad de incumplimiento al mantener la variable explicada siempre dentro de un rango de variación entre cero y uno. La principal ventaja del modelo de regresión logística radica en que no tiene requerimientos sobre la distribución de las variables continuas de entrada. (Siddiqui, 2006)

Esta busca construir una función que determine la probabilidad de default para cada operación, considerando un set de variables independientes, y una variable dependiente binaria que es observable de la forma $y_i = 1$ si cliente i hace default, e $y_i = 0$ si no. Luego la probabilidad de default está dada por:

$$p(y_i = 1|x_i) = \frac{1}{1+e^{\beta_0+\sum \beta_j x_{ij}}} \quad (2)$$

donde β_0 es el intercepto, y β_j son los coeficientes asociados a la variable j . Dado que los parámetros son desconocidos, deben ser estimados usando, por ejemplo, el estimador de máxima verosimilitud, o información histórica. (7)

Este modelo tiene la ventaja de medir la probabilidad de incumplimiento al mantener la variable explicativa dentro de un rango entre 0 y 1, siendo de interpretación intuitiva.

Para el caso de estudio particular de factoring, otro modelo a aplicar donde se podrían tener resultados interesantes es la regresión logística multinomial. Esta se asemeja al logit tradicional, pero con la particularidad que permite el estudio de problemas con más de dos posibles resultados discretos. Por ejemplo, se modelaría:

$$y_i = 2 \text{ si deudor } i \text{ hace default, } y_i = 1 \text{ si cliente } i \text{ hace default, e } y_i = 0 \text{ si no}$$

dado que dentro del vector X de características de la operación existen variables del cliente, variables del deudor y variables del documento en sí, como puede ser la tasa del documento, el plazo, el porcentaje de adelanto, etc.

3.2 Comparación de resultados

Existen diversas técnicas que tienen como objetivo medir el resultado de los modelos con la finalidad de compararlos. El principal input que consideran son la cantidad de aciertos (verdaderos malos y verdaderos buenos) y minimizar el error de clasificación, representados por clasificar un caso “bueno” siendo “malo” (falso positivo) y un caso “malo” siendo “bueno” (falso negativo).

A continuación, se presentan las técnicas más utilizadas.

3.2.1 Matriz de confusión

Muestra la cantidad de verdaderos malos, verdaderos buenos, falsos malos y falsos buenos, es decir, los aciertos y desaciertos. A partir de esta matriz se pueden obtener 4 medidas de performance (4):

- Precisión = Verdaderos negativos y verdaderos positivos / Total de casos
- Tasa de error = Falsos negativos y falsos positivos / Total de casos
 - Tasa de error tipo II = Falsos negativos/ (Falsos negativos + verdaderos negativos)
 - o Tasa de error tipo I = Falsos positivos/ (Falsos positivos + verdaderos positivos)
 - Sensibilidad= Verdaderos positivos / Total de casos positivos.

Especificidad= Verdaderos negativos/ Total de casos negativos. Curva ROC Método de visualización de performance. Es el gráfico de la frecuencia acumulada de casos “malos” y casos “buenos”. Mientras más diferencia haya entre las curvas, mejor es el poder predictivo del modelo.

3.2.2 Estadístico C o AUC

Medida que representa el promedio de distancias entre las curvas de frecuencia acumulada de casos “buenos” y “malos”. Se calcula como el área bajo la curva ROC (o integral). Es la medida más efectiva ya que mide el espectro completo del scoring.

En palabras simples, AUC representa la probabilidad que el método de discriminación clasifique más alto a una observación aleatoria “buena” que una observación aleatoria “mala”.

La curva ROC se obtiene al calcular la sensibilidad y 1-especificidad de las observaciones para cada umbral de corte, siendo estas últimas los mismos valores calculados para la matriz de confusión presentada en la “**Sección 3.2.1**”.

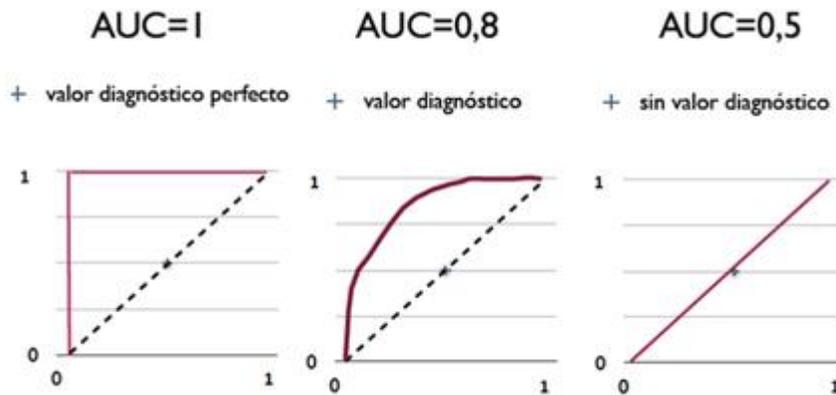


Ilustración 3: Ejemplos de visualización de curva ROC.

Como se aprecia en la “**Ilustración 3**”, el área bajo la curva puede tomar valores entre 0,5 y 1. Siendo 0,5 el análogo a lanzar una moneda para determinar la discriminación y 1 una discriminación perfecta. Para el caso de modelos de scorings como el de esta memoria se espera que la capacidad discriminante sea superior a 0.7.

3.3 Tecnología

Para realizar este estudio se utilizan los softwares estadísticos de la empresa IBM, los cuales son *SPSS Modeler 18* y *SPSS Statistics 23*. SPSS es uno de los programas más utilizados para el trabajo estadístico, dada su capacidad para trabajar con gran cantidad de datos y una interfaz sencilla para los análisis. Sus principales competidores son SAS, Stata, Matlab y R, este último lenguaje libre y gratuito.

La consulta de los datos hacia el “data wharehouse” de la empresa se realiza utilizando el software *SQLPro Studio1.0.91* de la empresa Hankinssoft Development, Inc.

4. Recopilación y entendimiento de los datos

4.1 Datos iniciales

Se cuenta con acceso al consolidado de los datos de la empresa, los cuales corresponden a un vaciado de la información ingresada desde las distintas bases que maneja la compañía. El consolidado se actualiza una vez al día, entre las 21 y 22 horas y contiene con más de 130 tablas distintas, las cuales se encuentran separadas de acuerdo a los distintos productos ofrecidas por la compañía.

Para el modelo de admisión existen 37 tablas pertenecientes al producto factoring, donde con la ayuda de personas con experiencia en la empresa se seleccionaron las que entregan un mayor valor de información. El detalle se puede ver en la siguiente tabla (“**Tabla 1**”):

Tabla 1: Selección de tablas consolidado Tanner S.A (Fuente: elaboración propia en base a consolidado de Tanner S.A.)

Tablas	Descripción
FACTORING.TDOCUMENTO	Características de la factura
FACTORING.ABONO_DOCUMENTO	Montos abonados por cliente /deudor según documento
FACTORING.ABONO_DOCUMENTO_CIERRE_PAGO	Características del pago del abono
FACTORING.TOPERACION	Características de cada operación
FACTORING.TLINEA_HISTORICA	Línea de crédito de uso interno de la empresa
FACTORING.TPROTESTO_HISTORICA	Información de protestos por documento
FACTORING.TPROTESTO_MOTIVO	Características del protesto

4.2 Estudio de variables existentes

En una segunda etapa de extracción, se analizan los campos contenidos en las tablas. La dificultad de esta etapa radica en la comprensión del contenido de cada campo, dado que la terminología puede tener un grado técnica y específica alta. Para sortear esta etapa se preguntó a las personas de las empresas encargadas de este consolidado el significado de los campos.

Se seleccionan en un principio 67 campos dentro de un universo aproximado de 1.000. El detalle se muestra en anexo 9.3 en el apartado correspondiente.

Finalmente se obtiene una base de datos inicial de 6.841.210 registros, correspondientes al número de documentos históricos operados por Tanner, considerando facturas, facturas electrónicas, cheques, letras y otros.

4.3 Construcción de variable de objetivo

Una vez con lo anterior, se construye la variable objetivo como “default”, que se define como si un documento particular presenta 90 días o más de mora en el abono de un documento.

Se definen los días de mora de la siguiente manera:

$$\begin{aligned} & \text{Si } saldo_documento = 0 \\ & \Rightarrow Mora = Max\left(fecha_{abono_documento} - fecha_{vencimiento_prórroga}, 0 \right) \end{aligned} \quad (3)$$

$$\begin{aligned} & \text{Si } saldo_de_documento \neq 0 \\ & \Rightarrow Mora = Max\left(fecha_{hoy} - fecha_{vencimiento_prórroga}, 0 \right) \end{aligned} \quad (4)$$

Donde *saldo_documento* corresponde a la resta del monto adeudado por el par cliente, deudor menos el monto que abona el deudor o cliente, es decir, la operación queda zanjada una vez que el saldo es igual a 0.

El campo *fecha_abono_documento* corresponde a la fecha en que cliente/deudor paga un abono. Por otro lado, *fecha_vencimiento_prórroga* se considera como la fecha de vencimiento para el cálculo de mora³. Se utiliza este campo ya que en caso de que un cliente solicite un mayor

³ En otros casos, la solicitud de prórroga por parte de los clientes podría ser considerada como un indicio de que caerán en incumplimiento. Esta vez se determinó lo contrario (considerarlo como una nueva fecha de vencimiento) luego de conversaciones con el sponsor dentro de la empresa, ya que, por experiencia, el cliente que pide prórroga generalmente paga.

plazo para pagar, la empresa lo evalúa y considera un nuevo plazo para la operación. En caso de que el cliente no solicite prórroga, el campo de $fecha_{vencimiento\text{prórroga}}$ tendrá el mismo valor que el campo de fecha de vencimiento, la cual fue pactada inicialmente.

Las ecuaciones (3) y (4) responden al siguiente flujo (“**Ilustración 4**”):

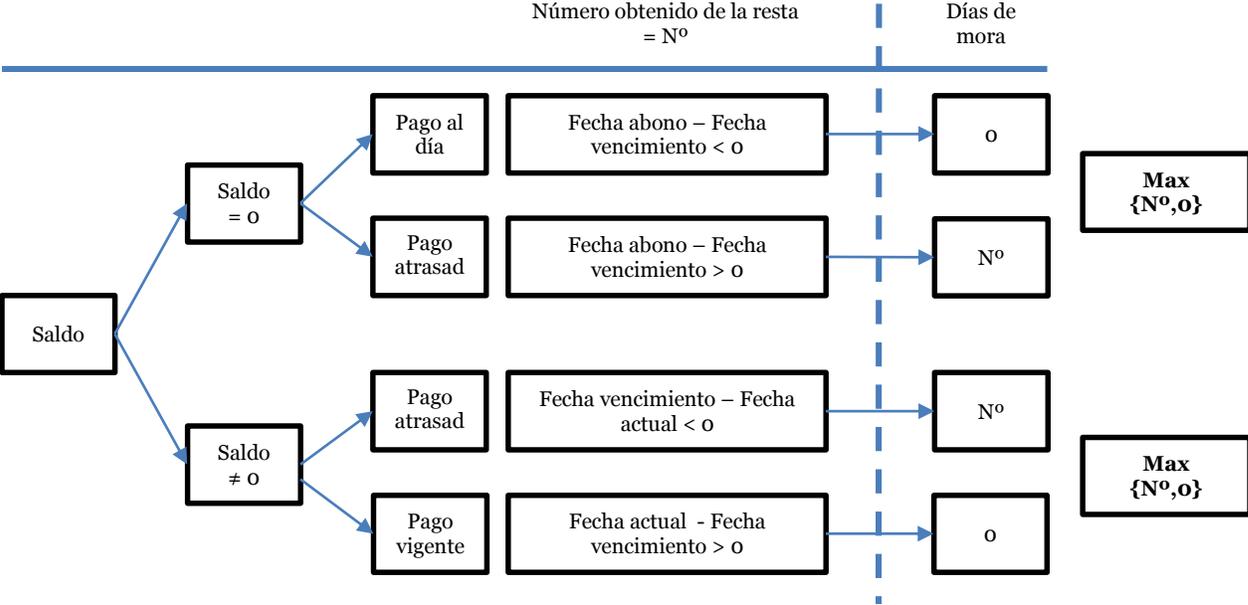


Ilustración 4: Flujo del proceso de Mora con sus posibles escenarios

De la base, se obtienen 717.911 documentos que no han recibido nunca un abono.

Esto corresponde al 10,4% de los documentos totales, los cuales, una parte corresponden a documentos vigentes al día de hoy, y otros, a documentos que efectivamente nunca tuvieron un pago, ya sea por el cliente o por el deudor.

5. Preparación, transformación y selección de datos

5.1 Limpieza de datos

En primer lugar, se filtra la variable $id_producto$ de acuerdo a tipo de producto para obtener los documentos de factoring correspondientes a facturas y facturas electrónicas,

dado que el negocio se puede realizar también con otro tipo de documentos, los que quedan fuera del alcance de este trabajo. Como resultado, el número de documentos se reduce de 6.841.210 a 1.234.432.

En segundo lugar, se procede a eliminar los documentos que se encuentran en estado vigente de la base, dado que se requiere una ventana de 90 días luego de la fecha de vencimiento para poder observar si el documento incurrió en default o no. En la base se encontraban 70.954 facturas con estas características, lo que corresponde a un 5.75% de la base y que fueron eliminados.

Finalmente se eliminan las variables *monto_protesto*, *saldo_protesto* y *monto_solicitado* ya que no contienen valores, solamente valores nulos.

Con los arreglos mencionados anteriormente la base final queda con 1.163.479 registros, donde la distribución de los documentos no pagados, marcados con un “1” se muestra en el “Gráfico 2”.

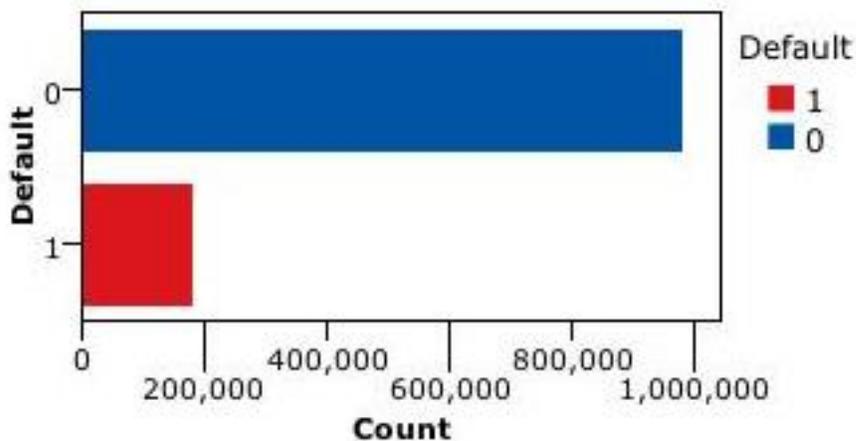


Gráfico 2: Distribución de Default en la base de datos. Fuente: Elaboración propia

El porcentaje de default con respecto al total de facturas corresponde al 15,69%.

5.2 Análisis de los datos

Una vez con la base limpia, se analiza cada variable por separado, para obtener una comprensión de los datos. El resumen se puede observar en el “**Anexo 11.3**”.

Para el tratamiento de los outliers se utiliza la técnica de reemplazar el valor por el valor más próximo que se encuentre dentro del rango. Los valores extremos se reemplazan por valores nulos, dado que pueden deberse a equivocaciones humanas en el ingreso de la información.

A modo de resumen se muestra la “**Tabla 2**”:

Tabla 2: Porcentaje de Outliers y Extremos

Extremos anulados		
Porcentaje de la base	12,7%	4,8%

Los valores nulos no se eliminan, dado que podrían contener información valiosa que podría caracterizar a la variable de respuesta. El detalle de cada acción tomada para cada variable se puede observar en el “**Anexo 11.1**”.

Otra observación es que existen campos que muestran valores que no tienen sentido para el negocio, en este caso valores negativos en los atributos *valor_nominal* y *monto_aprobado* de la línea, que significaría que el valor de la factura es negativo, o que la empresa definió una línea negativa al cliente⁴ respectivamente. Para la variable *valor_nominal* existen 10 documentos con esta característica, por lo que independiente de lo que se haga no afectará la distribución de la base. Para el caso de la variable *monto_aprobado* son 15 documentos con esta característica, por lo que igualmente no afectará la distribución.

⁴ Esta variable menciona el cupo que se define a priori que tiene el cliente para factorizar. No describe la posibilidad de que este cupo sea superado con las operaciones, en donde sí se puede obtener tener una línea negativa del cliente.

La decisión que se toma con respecto a estos casos es transformar todos los valores negativos en valores nulos, dado que se puede desprender que son errores en el ingreso de los datos o en que la operación se hizo o mal o definitivamente no se hizo e igual quedo registrada.

5.3 Selección de variables

Si bien los objetivos de la presente memoria tratan sobre un modelo de admisión, existen en la base variables que corresponden a observaciones del comportamiento que tuvo el documento una vez ya ingresado al sistema, por lo tanto, no es pertinente que sean analizadas de manera directa, si no que deben ser transformadas para obtener variables históricas de ese cliente y deudor en particular.

En una primera iteración del modelo, las variables que serán seleccionadas se muestran en la “**Tabla 3**”.

Tabla 3: Variables seleccionadas. (Fuente: elaboración propia)

Variables Utilizadas	
valor_nominal	interes_documento
tasa_operacion	tasa_mora
porcentaje_descuento	tipo_pagare
con_notificacion	con_custodia
glosa_zona	tipo_cambio
tipo_operacion	monto_aprobado
spread_minimo	numero_trabajadores
contabilidad_completa	glosa_motivo_protesto

En el “**Anexo 11.4**” se puede ver la tabla con las variables filtradas.

5.4 Transformación de variables

Para la implementación de un score, se requiere que las variables que integren el modelo estén categorizadas, con el fin de asignar un puntaje de acuerdo al rango en que

se encuentra la variable al momento de la solicitud. En una primera iteración las variables categóricas no sufren modificaciones, por lo que el trabajo se centra en las variables continuas. Para la categorización de las variables continuas se utilizan árboles de decisión con el algoritmo CHAID. Este algoritmo permite establecer cortes óptimos de acuerdo a la variable de respuesta, utilizando el estadístico chi-cuadrado.

Otra característica de este algoritmo es la utilización de los valores nulos como una categoría aparte, por lo que se alinea de buena manera con las intenciones con respecto a estos valores que se mencionan en la “**Sección 5.2**”.

A continuación, en la “**Ilustración 5**” se presenta el resumen de las nuevas distribuciones de las variables transformadas.

Field	Sample Graph	Measurement
⚠ Tipo_cambio		🚩 Flag
⚠ chaidcontabilidad_completa		🎯 Nominal
⚠ chaiddias_documento		🎯 Nominal
⚠ chaidinteres_documento		🎯 Nominal
⚠ chaidmonto_aprobado		🎯 Nominal
⚠ chaidnumero_trabajadores		🎯 Nominal
⚠ chaidporcentaje_descuento		🎯 Nominal
⚠ chaidspread_minimo		🎯 Nominal
⚠ chaidtasa_mora		🎯 Nominal
⚠ chaidtasa_operacion		🎯 Nominal

Ilustración 5: Resumen de las distribuciones de las variables categorizadas

Finalmente se realiza una partición la base en dos segmentos, la cual se muestra en el “**Gráfico 3**”. Un 70% de los datos se asignan a la base de entrenamiento y el 30 % restante a la base de prueba.

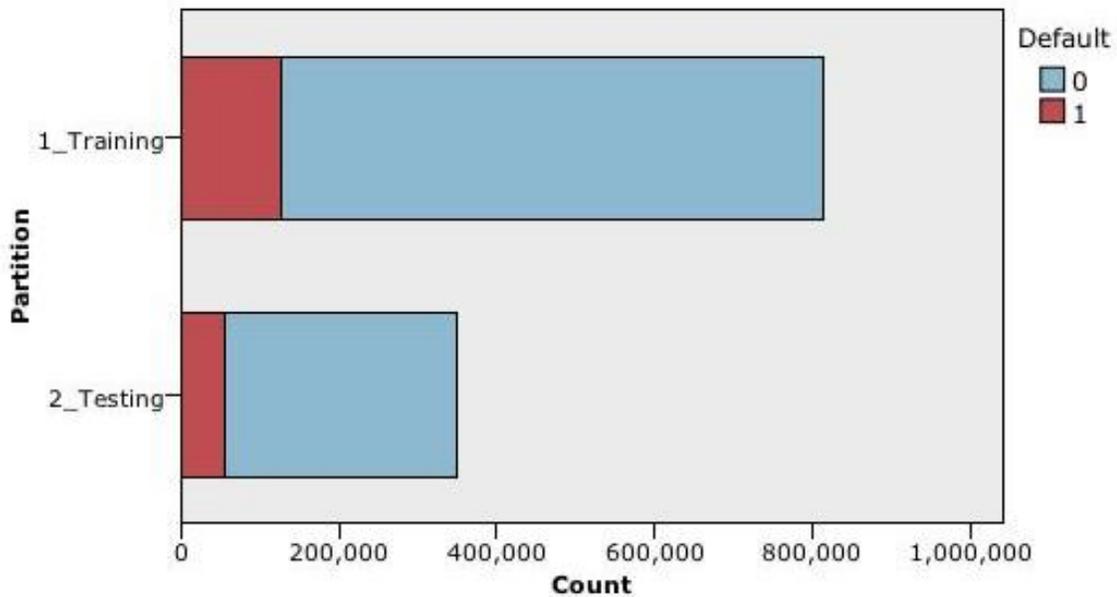


Gráfico 3: Base dividida en base de entrenamiento y base de prueba.

6. Aplicación del modelo

6.1 Análisis de variables

Se realiza un análisis de covarianzas para los campos de la base filtrada, con el fin de no incluir en el modelo variables con un nivel de correlación muy alto. De las variables seleccionadas, solamente muestra una correlación elevada entre si las variables *valor_nominal* e *interés_documento*, que es bastante lógico debido a que, manteniéndose una tasa de interés constante, mayor monto de dinero implica un mayor interés. El resto del detalle se puede observar en el “Anexo 11.5”.

Es importante destacar que el modelo debe ser lo más simple posible, para entregar un mayor valor de cara a los ejecutivos y/o analistas que utilizarán el producto. Al no contar con este tipo de herramientas actualmente en la empresa, se espera que se muestren un poco incrédulos con los beneficios del scoring. Esto decanta en el uso de variables obvias para el negocio, de fácil interpretación y rápida obtención, de manera que el modelo analice la operación de manera que le haga sentido al usuario.

6.1.1 Primera iteración

- **Selección de variables**

Para un primer modelo se seleccionan las variables indispensables para el producto de acuerdo a conversaciones con expertos de la empresa. Estas corresponden al valor nominal del documento, tasa de operación y porcentaje de descuento/adelanto. Además, se le agrega la variable de zona dada lo fácil de obtener desde el punto de vista de la admisión.

- **Calibración del modelo**

Para la regresión logística de la primera iteración se utiliza la configuración implementada por SPSS sin realizar configuraciones. Esto implica un umbral de clasificación de 0.5 y utilización del método “enter”, el cual significa que se ingresan todas las variables al mismo tiempo al modelo.

En una segunda iteración se modificará el método para observar si existen diferencias.

- **Evaluación del modelo**

Los betas calculados para cada variable se muestran en la siguiente “**Tabla 4**”:

Tabla 4: Betas calculados para primer modelo

Variables en la ecuación	Beta
chaidvalor_nominal	
chaidvalor_nominal (1)	-0,597
chaidvalor_nominal (2)	-0,688
chaidvalor_nominal (3)	-0,526
chaidvalor_nominal (4)	-0,663
chaidvalor_nominal (5)	-0,601
chaidvalor_nominal (6)	-0,546
chaidvalor_nominal (7)	-0,638
chaidvalor_nominal (8)	-0,473
chaidvalor_nominal (9)	-0,587
chaidvalor_nominal (10)	-0,427
chaidtasa_operacion	
chaidtasa_operacion (1)	0,548
chaidtasa_operacion (2)	0,583
chaidtasa_operacion (3)	0,947
chaidtasa_operacion (4)	0,865
chaidtasa_operacion (5)	-0,037
chaidtasa_operacion (6)	0,53
chaidtasa_operacion (7)	0,454
chaidtasa_operacion (8)	-0,097
chaidtasa_operacion (9)	0,024
chaidtasa_operacion (10)	-0,072
chaidporcentaje_descuento	
chaidporcentaje_descuento (1)	-9,72
chaidporcentaje_descuento (2)	-9,409
chaidporcentaje_descuento (3)	-8,828
chaidporcentaje_descuento (4)	-9,025
chaidporcentaje_descuento (5)	-8,37
chaidporcentaje_descuento (6)	-8,493
chaidporcentaje_descuento (7)	-8,618
glosa_zona	
glosa_zona (1)	-0,625
glosa_zona (2)	-2,903
glosa_zona (3)	-2,472
glosa_zona (4)	-1,747
glosa_zona (5)	-2,096
glosa_zona (6)	-1,289
glosa_zona (7)	-3,805
glosa_zona (8)	-0,261
glosa_zona (9)	-2,913
glosa_zona (10)	-6,228
glosa_zona (11)	-3,391
Constant	9,658

Se puede observar que las categorías relacionadas a la variable de porcentaje de descuento tienen un beta en magnitud muy superior al resto de las variables, esto podría implicar que está prediciendo sospechosamente bien.

En el “**Gráfico 4**” se muestra la curva ROC obtenida para el modelo ejecutado previamente.

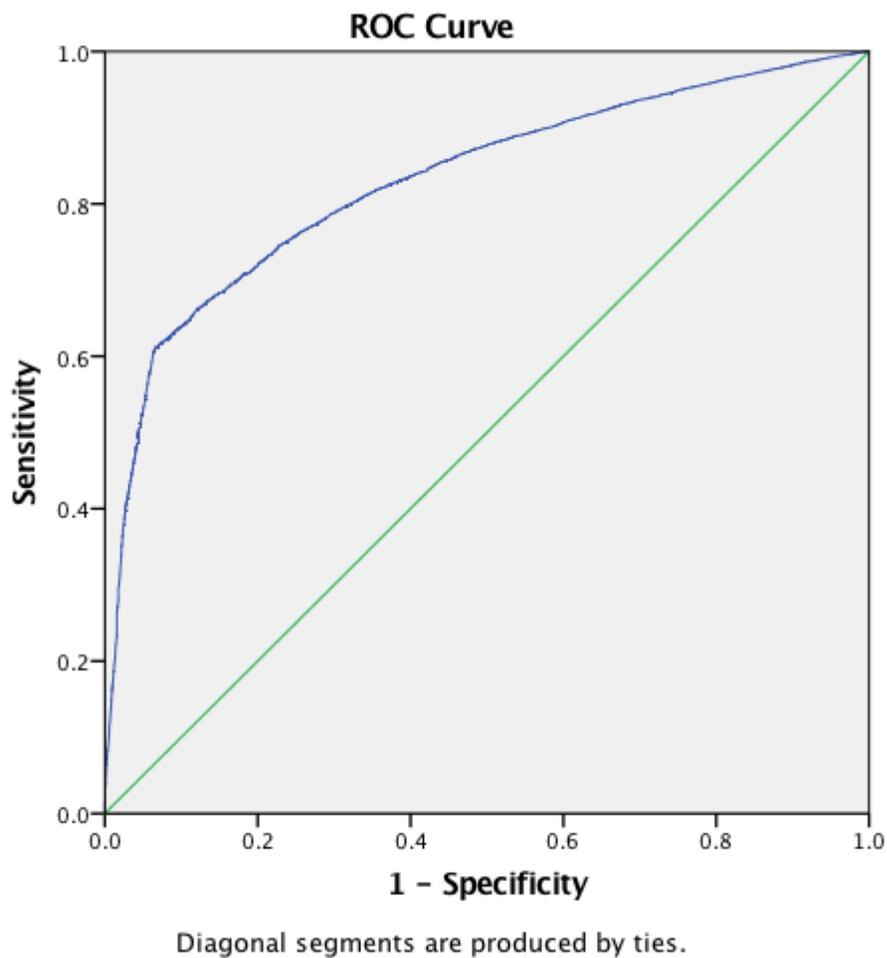


Gráfico 4: Curva ROC para primera iteración del modelo.

El área bajo la curva (AUROC) para este primer modelo es igual a 0.830 con un intervalo de confianza de 95%, lo que se considera muy bueno en comparación a un 0.5 que equivale a predecir utilizando el lanzamiento de una moneda. Considerando lo mencionado anteriormente, sobre los betas de la variable *porcentaje_descuento*, se realiza un histograma para evaluarla nuevamente, el cual se ve reflejado en el “**Gráfico 5**”.

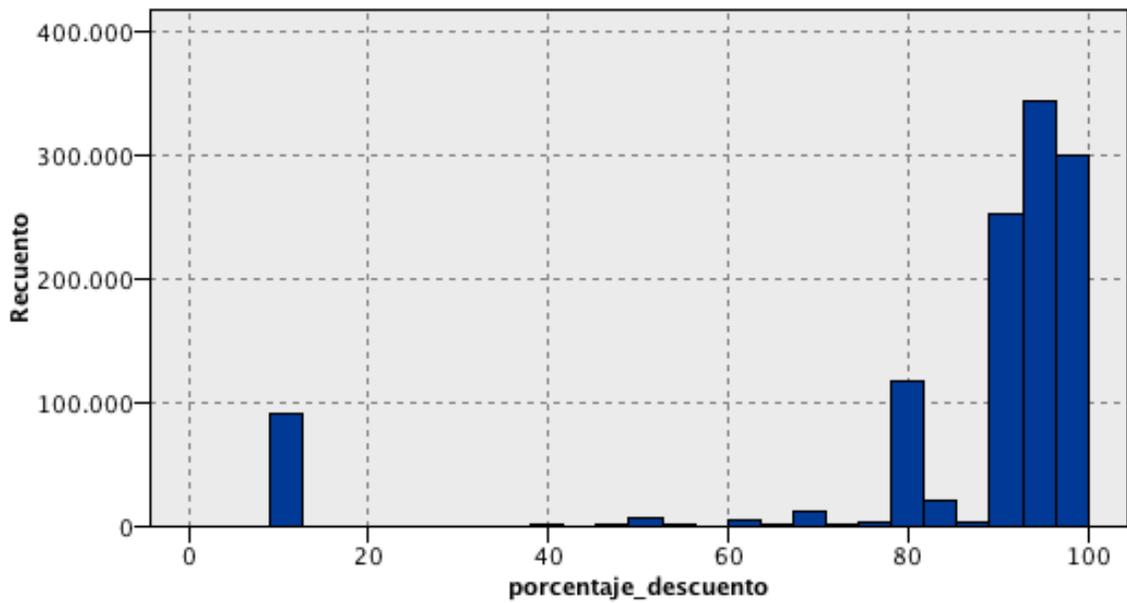


Gráfico 5: Histograma porcentaje de descuento

Del “**Gráfico 5**” se puede observar que existen 91.598 operaciones que se realizaron con un porcentaje de descuento del 11% en promedio.

Se decide eliminar estos documentos dado que condicionan de sobremanera los resultados de la regresión, además, desde un punto de vista estratégico del negocio no se deben entregar adelantos tan pequeños, lo cual se ve reflejado en el histograma del “**Gráfico 6**”.

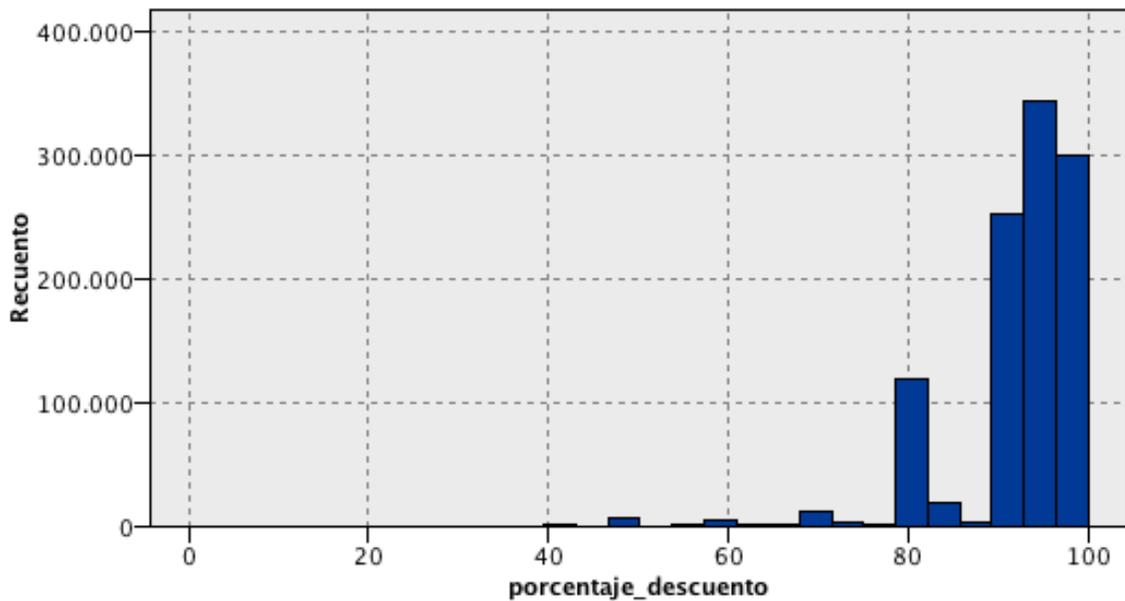


Gráfico 6: Histograma porcentaje de descuento arreglado

6.1.2 Segunda iteración

- **Selección de variables**

Luego, para esta segunda iteración del modelo se realiza una nueva transformación de variables, utilizando el algoritmo CRT de los árboles de decisión, cuyo objetivo es categorizar las variables continuas en segmentos internamente homogéneos y heterogéneos entre sí. Esta nueva categorización responde a que la categorización previamente realizada con el algoritmo CHAID establece muchas categorías por variable, con lo que si bien, se obtiene un mayor poder de predicción, se pierde interpretación. Además, existe posibilidad de que, al existir un mayor número de opciones a elegir, el usuario del scoring tenga más posibilidades de equivocarse, aumentando el riesgo operacional.

Esto da a lugar un total de 32 variables, cuyo resumen de categorización puede verse en el apartado de “**anexos 11.7**”

Luego, se realiza nuevamente una partición en un 70% de entrenamiento y un 30% de testeo, obteniéndose el “**Gráfico 7**”:

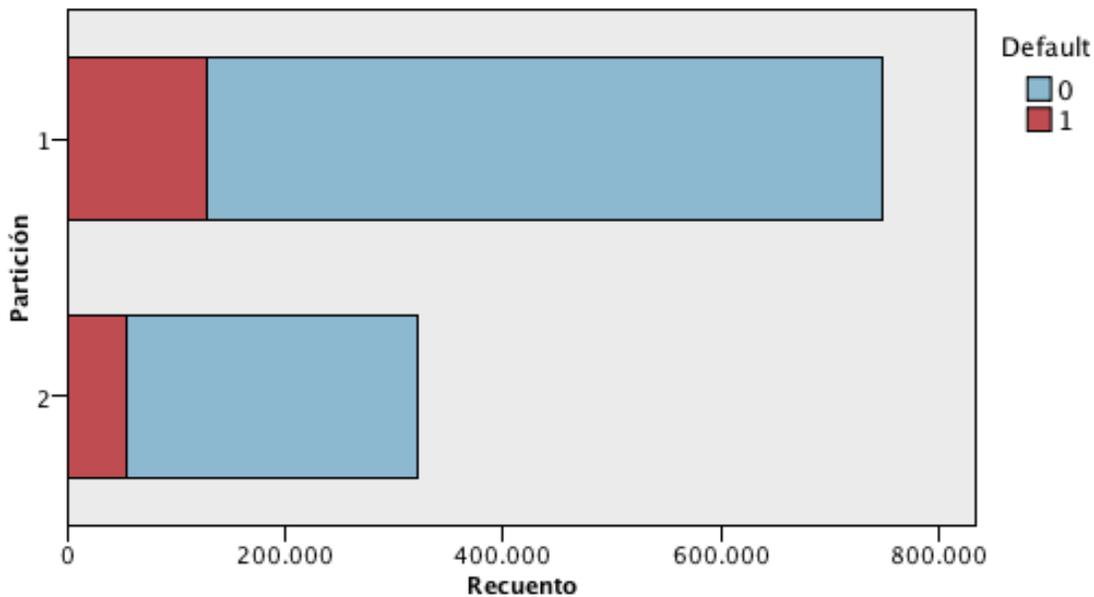


Gráfico 7: Distribución de los datos en base de entrenamiento y base de testeo.

- **Calibración del modelo**

Luego se ingresan estas variables en el modelo logit utilizando procedimientos de regresión “stepwise”, el cual es un método de ajuste del modelo, en donde en cada iteración del modo se va ingresando o eliminando variables explicativas basadas en un criterio especificado previamente. El software SPSS contiene los métodos de regresión *Forward* y *Backward*, cada uno de estos con la posibilidad de 3 criterios, los cuales son **condicional, ratio de verosimilitud, y estadístico de Wald**.

El método Foward lo que hace es en cada iteración ir sumando, el modelo agrega las variables una a la vez. En cada paso, cada variable que no está presente en el modelo se analiza para incluirla, en donde la más significativa se agrega al modelo siempre que tenga un p-valor menor a 0,05. De esta forma, se agregan variables hasta que no que ninguna de las variables restantes sea significativa.

El método Backward tiene un algoritmo similar, pero en sentido contrario, es decir, se ingresan todas las variables, y paso a paso se van eliminando las variables con el p-valor más alto, hasta la variable menos estadísticamente significativa.

Para cada método existen tres criterios posibles para la regresión logística: el criterio condicional, el cual testea de acuerdo al ratio de verosimilitud basado en un parámetro condicional estimado; el criterio de verosimilitud, el cual testea de acuerdo a maximizar la verosimilitud; y el criterio de Wald, el cual determina si la variable predictora es significativa o no.

Dado los objetivos del modelo, se realizan dos pruebas, utilizando los métodos Forward Wald y Backward Wald, de los cuales no se obtuvieron diferencias significativas. A continuación, se muestran los resultados obtenidos para la regresión logística utilizando Backward con criterio de Wald, y las 7 variables seleccionadas por el modelo.

- **Evaluación del modelo**

En la “**Tabla 5**” se muestran los betas obtenidos para el nuevo modelo:

Tabla 5: Betas calculados para cada variable y categoría

Variable	Categoría	Beta
Zona	Zona Sur	
	Sin zona	0,765
	Casa Matriz	-0,104
	Zona Centro	0,101
	Zona Metropolitana	0,491
	Zona Norte	-0,095
Contabilidad completa	Nulo	
	No	-0,956
	Si	-0,874
Línea aprobada	Nulo	
	Menos de 25 MM	-0,228
	Entre 25 MM y 850 MM	-0,637
	Más de 850MM	-0,49
Numero trabajadores	Nulo	
	Menos de 44	0,404
	Entre 44 y 528	-0,208
	Más de 528	0,633
Porcentaje descuento	Más de 96%	
	Menos de 84%	-1,071
	Más de 84% y menos de 94%	-0,81
	Más de 94% y menos de 96%	-0,05
Tasa de operación	Nulo	
	Menos de 1.09	0,476
	Entre 1.09 y 1.15	0,957
	Entre 1.15 y 1.5	0,102
	Mas de 1.5	-0,196
Valor del documento	Nulo	
	Menos de 68 M	-0,37
	Entre 1.2 MM y 2 MM	-0,538
	Entre 141 M y 250 M	-0,26
	Entre 2 MM y 3.8 MM	-0,543
	Entre 250 M y 430 M	-0,396
	Entre 3.8 MM y 9.5 MM	-0,405
	Entre 430 M y 710 M	-0,394
	Entre 68 M y 141 M	-0,198
	Entre 710 M y 1.2 MM	-0,432
Mas de 9.5 MM	-0,307	
	Constant	0,047

Su bondad de ajuste de acuerdo a AUROC se puede ver a continuación en el “**Gráfico 8**”:

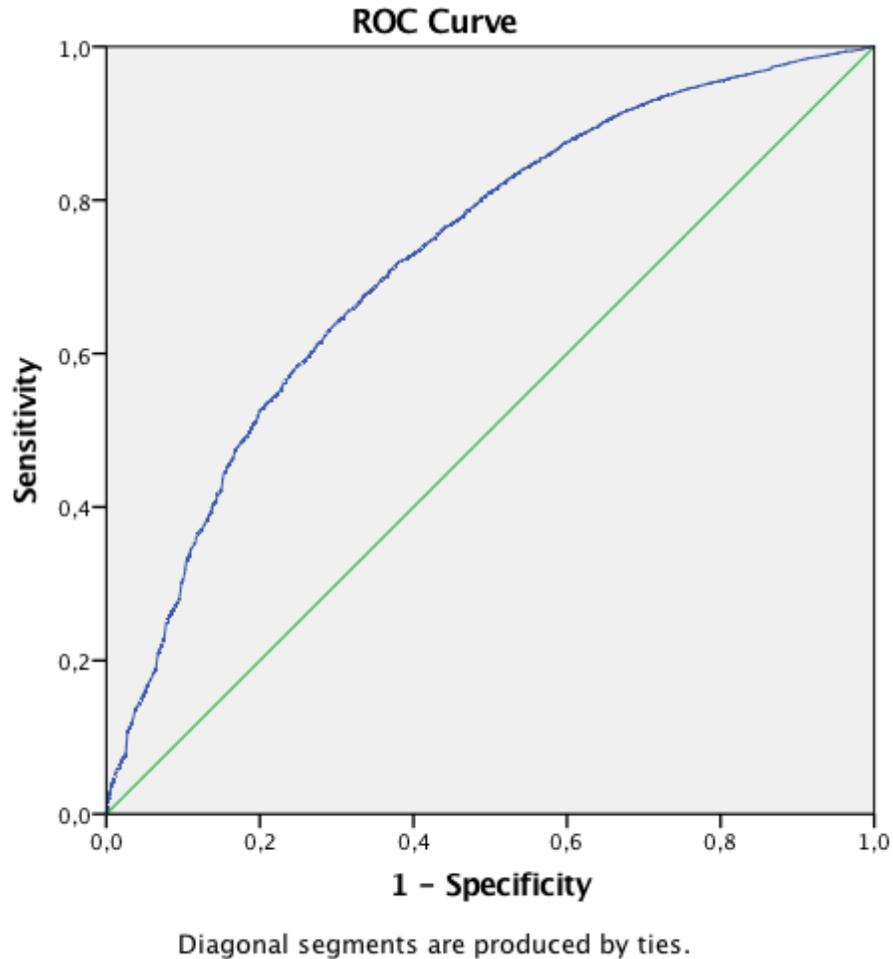


Gráfico 8: Curva ROC para segunda iteración del modelo. (Fuente: elaboración propia)

El AURCOC calculado para el modelo anterior es de 0,727 con un 95% de intervalo de confianza. Este resultado muestra que, si bien es menor en términos de predicción que el anterior, es más fiable debido a la corrección en las variables. Además, se sacrifica predicción por una mayor interpretación, asignando especial importancia de esta última, dado que la herramienta tiene que ser lo más intuitiva y fácil de usar posible.

7. Evaluación de resultados

7.1 Escalamiento a score

La regresión logística, además de entregar los betas asociados a cada categoría por variable, entrega como output un número entre 0 y 1 interpretable como la probabilidad de que cada documento incurra en default. Para mejorar el entendimiento de estas probabilidades se utiliza un re-escalamiento a un score que cumple:

- Puntaje entre 0 y 1000, donde el riesgo de no pago disminuye a medida que aumenta el puntaje.
- Cada rango de las variables tendrá asociado un puntaje positivo o negativo
- El puntaje final es la suma de los puntajes particulares.
- Definiremos “Odds” como $\frac{1-P(x)}{P(x)}$ que corresponde a la posibilidad de observar un evento.
- Donde $P(x)$ corresponden a las probabilidades calculadas con la regresión logística.

Por la fórmula del logit se tiene:

$$P(x) = \frac{1}{1+e^{-(\sum \beta_i x_i + \beta_0)}} \quad (6)$$

Desarrollando la ecuación:

$$\begin{aligned} \Rightarrow 1 + e^{-(\sum \beta_i x_i + \beta_0)} &= \frac{1}{P(x)} \\ \Rightarrow e^{-(\sum \beta_i x_i + \beta_0)} &= \frac{1 - P(x)}{P(x)} \quad / \ln() \\ \Rightarrow -(\sum \beta_i x_i + \beta_0) &= \ln\left(\frac{1-P(x)}{P(x)}\right) \quad (7) \end{aligned}$$

Donde $\ln\left(\frac{1-P(x)}{P(x)}\right)$ se conoce como la función de enlace logit. Gráficamente $\ln(Odds)$ se distribuye de la siguiente manera en el “**Gráfico 10**” y en la “**Tabla 6**”:

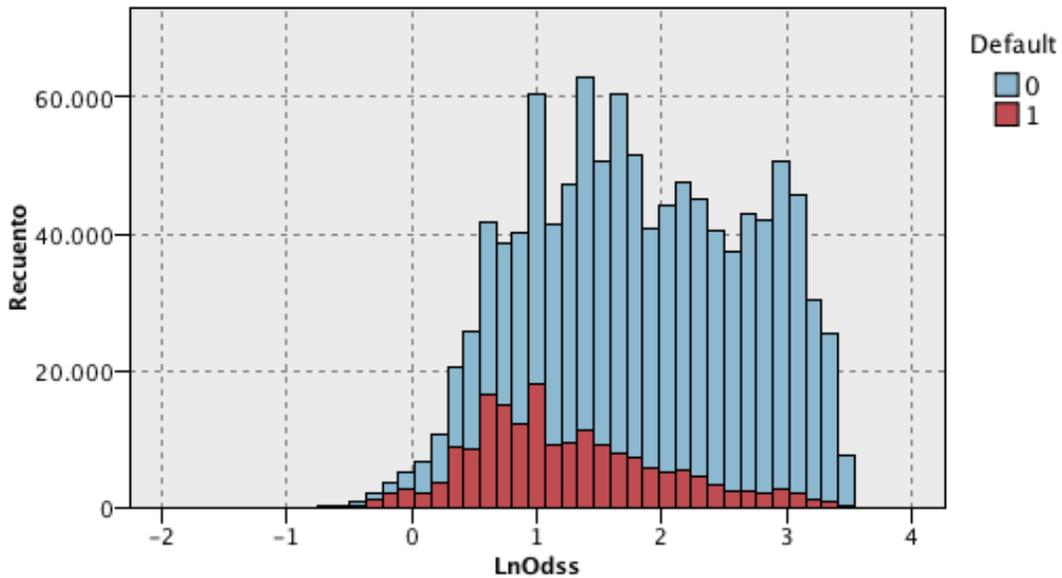


Gráfico 9: Distribución de la función enlace del logit $\ln(Odds)$. (Fuente: elaboración propia)

Tabla 6: Extremos $\ln(Odds)$ originales

Mínimo	Máximo
-1,41	3,668

La metodología a utilizar para realizar la transformación es convertir los mínimos y máximos de $\ln(Odds)$ en cero y 1.000 respectivamente. Para ello se suma 1,41 a los dos extremos, quedando los nuevos límites:

Tabla 7: Nuevos extremos $\ln(odds)$

Mínimo	Máximo
0	5,078

Con estos nuevos extremos se busca calcular el factor que cumpla:

$$\underset{\min}{\implies} (\ln(Odds) + 1,41) \cdot factor = 0 \quad (8)$$

$$\underset{\max}{\implies} (\ln(Odds) + 1,41) \cdot factor = 1000 \quad (9)$$

Es fácil ver que la **ecuación 8** se cumple para cualquier *factor*, dado que $(\ln(Odds) + 1,41)=0$.

Despejando la **ecuación 9**:

$$\underset{\max}{\implies} factor = \frac{1000}{\ln(Odds) + 1,41}$$

Reemplazando por el máximo obtenido:

$$\underset{\max}{\implies} factor = \frac{1000}{5,078} = 196,93$$

Utilizando este resultado en **(9)** entonces queda la transformación:

$$Score = (\ln(Odds) + 1,41) \cdot 196,93$$

Reemplazando por la función de enlace **(7)**:

$$Score = ((-\sum \beta_i x_i + \beta_0) + 1,41) \cdot 196,93 \quad (10)$$

Donde β_i corresponden a los betas calculados previamente para cada categoría *i*.

Ahora, dado que el score es la suma de los scores particulares para cada variable, se debe dividir el valor de la constante y el aumento de 1,41 por el número de variables presentes en el modelo (en este caso 7), quedando:

$$Score = \left(\left(-\sum \beta_i x_i + \frac{\beta_0}{7} \right) + \frac{1,41}{7} \right) \cdot 196,93 \quad (11)$$

De esta forma, los scores para las categorías de las 7 variables quedan:

Tabla 8: Puntaje calculado para cada variable de acuerdo a su categoría correspondiente

Variable	Categoría	Score
Zona	Sin zona	-112
	Zona Metropolitana	-58
	Zona Centro	18
	Zona Sur	38
	Zona Norte	57
	Casa Matriz	59
Contabilidad completa	Nulo	38
	Si	210
	No	227
Línea Aprobada	Nulo	38
	Menos de 25 MM	83
	Entre 25 MM y 850 MM	164
	Mas de 850MM	135
Numero trabajadores	Nulo	38
	Menos de 44	-41
	Entre 44 y 528	79
	Mas de 528	-86
Porcentaje descuento	Menos de 84%	249
	Mas de 84% y menos de 94%	198
	Mas de 94% y menos de 96%	48
	Mas de 96%	38

Variable	Categoría	Score
Porcentaje descuento	Menos de 84%	249
	Mas de 84% y menos de 94%	198
	Mas de 94% y menos de 96%	48
	Mas de 96%	38
Tasa Operación	Nulo	38
	Menos de 1.09	-55
	Entre 1.09 y 1.15	-150
	Entre 1.15 y 1.5	18
	Más de 1.5	77
Valor Documento	Nulo	38
	Menos de 68 M	111
	Entre 68 M y 141 M	77
	Entre 141 M y 250 M	90
	Entre 250 M y 430 M	116
	Entre 430 M y 710 M	116
	Entre 710 M y 1.2 MM	123
	Entre 1.2 MM y 2 MM	144
	Entre 2 MM y 3.8 MM	145

En primer lugar, la variable “**zona**” hace referencia al lugar geográfico donde se encuentra la sucursal de la empresa en la cual fue realizada la operación. En términos del score, destaca el caso donde un documento no tenga asociada una zona, que indica que sería más riesgoso.

“**Contabilidad completa**” indica si el cliente entrega información financiera al momento de solicitar financiamiento, normalmente sus balances y a su carpeta tributaria. El score estaría premiando de manera positiva a aquellos clientes que entreguen sus datos financieros, ya sean completos o en parte, a diferencia de no otorgar esta información.

“**Línea aprobada**” indica el monto aprobado que tiene el cliente para factorizar. Cabe destacar que la línea es de uso interno, es decir, el cliente no tiene conocimiento de cuál es el monto máximo que puede operar. Dentro de la empresa han mostrado ciertos intentos de cambiar la forma de determinar el monto de la línea de los clientes, dado que la empresa frecuentemente opera con clientes que tienen esta línea copada, siendo de uno de los elementos pasados a llevar al momento de decidir si aceptar una operación o no.

“**Número de trabajadores**” indica el número de trabajadores de la empresa solicitante de factoring. El score castiga empresas pequeñas y empresas muy grandes. Esto tiene lógica dado que los clientes pequeños tienden a tener una menor estabilidad financiera. En caso del score negativo para grandes empresas, se explica dado el poder de negociación que tienen en comparación a la empresa de factoring. Generalmente estas empresas tienen fechas particulares para saldar facturas además de estar constantemente trabajando con altos niveles de deuda, dado el modelo de negocios que los altos directivos determinen para su operación.

La variable “**porcentaje de descuento**” indica el porcentaje del monto total del documento que anticipa la empresa al cliente, quedando el resto retenido como excedente que se devuelve una vez saldada la factura. El score indica que mientras menor sea este porcentaje menos riesgo de default existe.

“**Tasa de operación**” indica la tasa mensual asociada a la operación de factoring. El score indica un grave problema en el pricing realizado por la empresa, dado que tasas más grandes indicarían operaciones menos riesgosas, porque las tasas no están bien asociadas al riesgo de las operaciones.

Finalmente “**valor documento**” indica el monto nominal de los documentos. Para favorecer la predicción de los resultados se decide establecer un mayor número de categorías. El score indica que el monto óptimo para operar se encuentra entre los \$700.000 y \$3.800.000, donde facturas muy pequeñas o muy grandes son más riesgosas. Se puede deducir de esta variable que operaciones con varios documentos con montos entre los mencionados anteriormente es preferible desde el punto de vista del riesgo que operaciones del mismo monto, pero de un solo documento.

7.2 Generación del scorecard

7.2.1 Validación del scorecard

Para evaluar la validez del scorecard se determina si dos sets de datos no son significativamente diferentes. Usualmente una examinación visual es suficiente para este propósito (Siddiqui, 2).

El “**Gráfico 11**” muestra la distribución del scorecard para la base de entrenamiento, donde el máximo puntaje obtenido es 1000 y el mínimo 0, con media 633 y desviación estándar 172,8

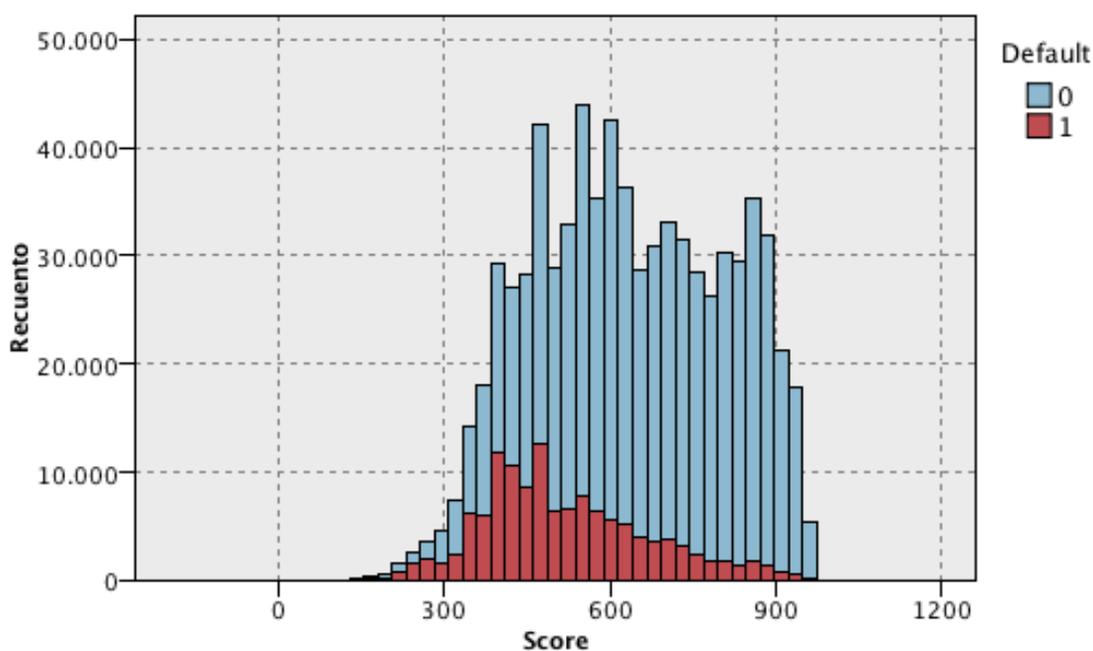


Gráfico 10: Distribución scorecard base de entrenamiento

El “**Gráfico 12**” muestra la distribución del scorecard para la base de prueba, donde el máximo puntaje obtenido fue 981 y el mínimo fue 27. La media fue 632,8 y la desviación estándar 172,9.

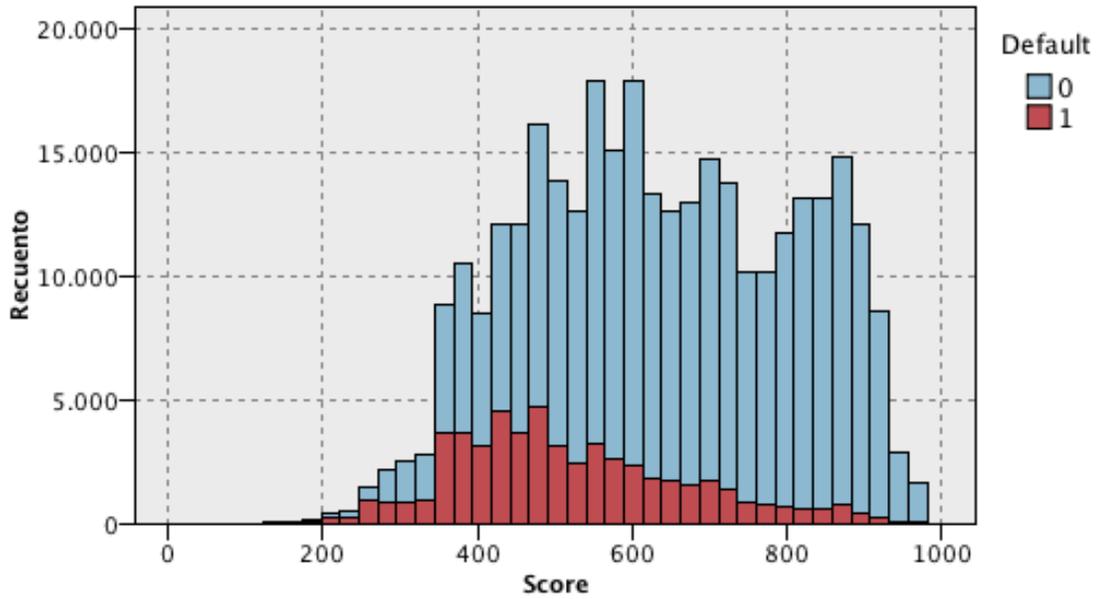


Gráfico 11: Distribución de score base de prueba

Además, se puede observar que a medida que el puntaje aumenta, el número de documentos que hicieron default disminuye, cumpliendo el objetivo del re-escalamiento.

7.2.2 Generación de la herramienta de admisión

Con los scores ya calculados, se genera un prototipo funcional de la herramienta, donde el objetivo, como se mencionaba anteriormente, es que sea sencillo de usar y de fácil de comprender.

Variables	Ingrese la opción	Score
Zona	Casa Matriz	59
Contabilidad	No	227
Monto aprobado	Nulo 1M y 850 MM	164
Número de trabajadores	No 528	79
Porcentaje de descuento	Si 84%	249
Tasa operacional		77
Valor documento	Entre 2 MM y 3.8 MM	145
Score Total		1000

Ilustración 6: Prototipo funcional scorecard

Variables	Ingrese la opción	Score
Zona	Casa Matriz	59
Contabilidad	No	227
Monto aprobado	Entre 25 MM y 850 MM	164
Número de trabajadores	Entre 44 y 528	79
Porcentaje de descuento	Menos de 84%	249
Tasa operacional	Mas de 96%	
Valor documento	Menos de 84%	
	Mas de 84% y menos de 94%	
	Mas de 94% y menos de 96%	

Ilustración 7: Prototipo funcional scorecard

Variables	Ingrese la opción	Score
Zona	Sin zona	-112
Contabilidad	Si	210
Monto aprobado	Entre 25 MM y 850 MM	164
Número de trabajadores	Menos de 44	-41
Porcentaje de descuento	Menos de 84%	249
Tasa operacional	Mas de 1.5	77
Valor documento	Entre 2 MM y 3.8 MM	145
Score Total		692

Ilustración 8: Prototipo funcional scorecard

Aquí se explica la relevancia de poder categorizar las variables continuas en los pasos anteriores, ya que además de reducir el riesgo operacional (equivocaciones del usuario al ingresar la información manualmente), la interpretación de variables continuas con un score no sería fácil de comprender.

7.3 Cálculo de los beneficios

7.3.1 Determinación de puntaje de corte

Previo a la cuantificación de los beneficios de la scorecard se debe determinar el puntaje con el cual una solicitud será rechazada o aceptada. Esta es una decisión estratégica que debe estar correctamente alineada con los objetivos de la empresa.

Para observar los puntos óptimos de corte se construye la “**Tabla 9**”.

Tabla 9: Tabla de puntos de corte por rango

Min Score	Max Score	Total	Buenos	Malos	%Buenos	%Malos	Cobertura
807	1.000	224.824	214.661	10.163	24%	6%	21%
704	807	164.599	150.216	14.383	17%	8%	15%
632	704	126.229	110.009	16.220	12%	9%	12%
575	632	124.023	106.073	17.950	12%	10%	12%
527	575	102.389	84.277	18.112	10%	10%	10%
485	527	76.153	59.112	17.041	7%	9%	7%
447	485	78.345	55.250	23.095	6%	13%	7%
411	447	55.958	35.192	20.766	4%	11%	5%
378	411	48.130	29.523	18.607	3%	10%	4%
346	378	35.283	20.941	14.342	2%	8%	3%
315	346	12.166	8.791	3.375	1%	2%	1%
284	315	9.313	5.889	3.424	1%	2%	1%
253	284	6.397	3.138	3.259	0%	2%	1%
222	253	4.065	1.487	2.578	0%	1%	0%
190	222	1.568	758	810	0%	0%	0%
157	190	759	473	286	0%	0%	0%
123	157	205	156	49	0%	0%	0%
86	123	43	28	15	0%	0%	0%
45	86	3	1	2	0%	0%	0%
-	45	6	3	3	0%	0%	0%
		1.070.458	885.978	184.480	100%	100%	100%

Las primeras dos columnas construyen el rango del puntaje, las columnas “buenos” y “malos” indican el número de documentos en ese rango que son 0 y 1 respectivamente para la variable dependiente. Las columnas de “%Buenos” y “%Malos” se crean dividiendo el número de casos buenos en el rango por el total de casos buenos en la base de datos. De igual manera para los documentos incurridos en no pago. Finalmente, la columna de

cobertura indica el porcentaje de documentos del total de la base que se encuentran en ese rango.

Utilizando lo anterior se obtiene el siguiente gráfico:

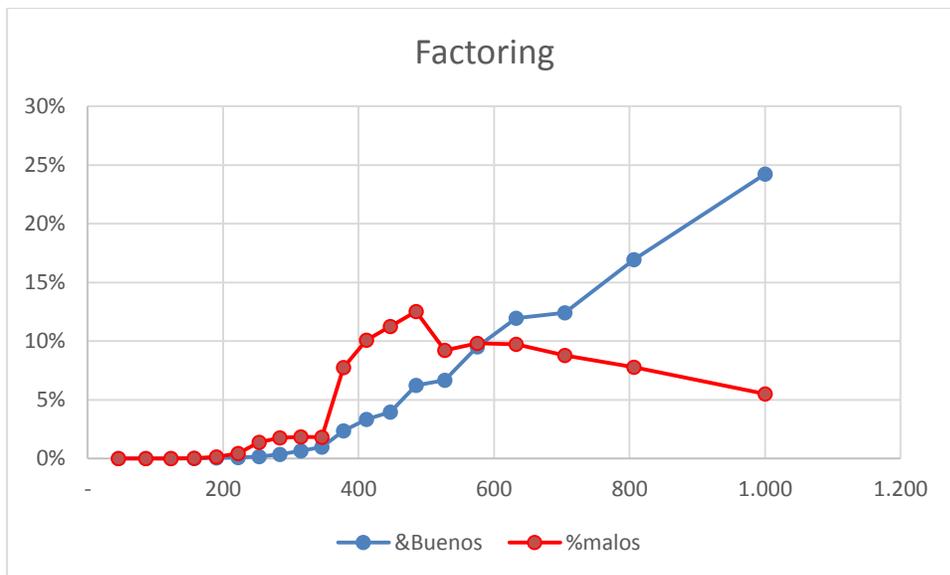


Gráfico 12: Score versus porcentaje de casos

Se observa que a medida que aumenta el score el porcentaje de casos buenos aumenta, mientras que los casos malos disminuyen desde el puntaje 360 aproximadamente en adelante. Esto indica que el score calculado tiene lógica con los datos trabajados, mientras mejor es un documento, más puntaje le asigna el modelo (menos riesgoso). En la parte izquierda de la figura los bajos porcentajes se deben a la menor frecuencia de documentos con scores más pequeños. Esto podría indicar que las variables actuales presentes en el modelo no predicen de manera lapidaria el default de una operación.

Los cortes posibles son variados y dependerán exclusivamente de la intención y/o el uso que se le quiera dar al scorecard. Es usual ver puntos de corte que maximicen la verosimilitud, o el punto que maximice la efectividad total de las predicciones. (Siddiqui, 2006)

Para esta memoria la decisión se toma en conjunto con el sponsor utilizando en primera instancia el porcentaje de aceptación de operaciones que tiene actualmente la empresa, el cual es un 95% aproximadamente.

Luego, se construye la “**Tabla 10**” generando sumas acumuladas para cada una de las columnas previas.

Tabla 10: Tabla de puntos de corte

Min Score	Max Score	%Buenos Aceptados	% Malos rechazados	%Acierto total	Cobertura Acu
807	1.000	24%	94%	27%	21%
704	807	41%	87%	45%	36%
632	704	54%	78%	57%	48%
575	632	66%	68%	66%	60%
527	575	75%	58%	72%	69%
485	527	82%	49%	74%	76%
447	485	88%	37%	73%	84%
411	447	92%	25%	67%	89%
378	411	95%	15%	64%	93%
346	378	98%	7%	61%	97%
315	346	99%	6%	73%	98%
284	315	99%	4%	64%	99%
253	284	100%	2%	50%	99%
222	253	100%	1%	37%	100%
190	222	100%	0%	48%	100%
157	190	100%	0%	62%	100%
123	157	100%	0%	76%	100%
86	123	100%	0%	65%	100%
45	86	100%	0%	33%	100%
-	45	100%	0%	50%	100%

Los tramos de puntaje que están más cerca de una cobertura del 95% son entre 346 y 411 puntos. Se propone como punto de corte el tramo entre 378-411 puntos, dado que en ese rango el modelo permite que, frente a un porcentaje de 92% de aceptación, pueda rechazar un 15% de documentos malos que hoy en día ingresan al sistema. Con esto, solicitudes que obtengan un puntaje menor a 378 puntos serán rechazadas automáticamente.

El segundo corte a determinar corresponde al puntaje que optimice las utilidades de la empresa. Para ello se define previamente:

Tabla 11: Promedio variables operación, ingresos y costos calculados

Item	Promedio por documento
Comisión	\$ 48.964
Días mora caso bueno	15
Monto Documento	\$4.438.147
Días documento	54
Tasa Operación	1,35%
Tasa Mora	3,78%
% descuento	92%
Tasa Costo de fondo	0,35%
Interés	\$73.823
Interés mora	77.282
Utilidad bueno	200.069
Costo aceptar malo	4.088.998

Los números proporcionados corresponden a los promedios calculados para los ítems que participan en el cálculo de los ingresos y costos para una operación de factoring. Se observa que el ratio entre aceptar una operación buena y aceptar una mala es $\frac{\text{Bueno}}{\text{Malo}} = 4,8\%$, es decir, por cada documento que incurre en default se deben aceptar al menos 20 documentos buenos para no tener pérdidas.

Con esto es posible generar la siguiente “**Tabla 12**”:

Tabla 12: Utilidades promedio por rango de Score

Min Score	Max Score	Buenos	Malos	Ingresos	Costos aceptar malo	Utilidades
807	1.000	214661	10163	42.946.975.371	41.556.486.425	1.390.488.946
704	807	150216	14383	30.053.539.546	58.812.057.882	-28.758.518.336
632	704	110009	16220	22.009.372.050	66.323.547.163	-44.314.175.113
575	632	106073	17950	21.221.901.130	73.397.513.660	-52.175.612.530
527	575	84277	18112	16.861.200.886	74.059.931.332	-57.198.730.446
485	527	59112	17041	11.826.468.749	69.680.614.501	-57.854.145.752
447	485	55250	23095	11.053.802.923	94.435.408.244	-83.381.605.321
411	447	35192	20766	7.040.822.307	84.912.131.959	-77.871.309.652
378	411	29523	18607	5.906.632.103	76.083.985.330	-70.177.353.227
346	378	20941	14342	4.189.641.394	58.644.408.965	-54.454.767.571
315	346	8791	3375	1.758.805.095	13.800.368.167	-12.041.563.072
284	315	5889	3424	1.178.205.347	14.000.729.068	-12.822.523.721
253	284	3138	3259	627.815.992	13.326.044.402	-12.698.228.410
222	253	1487	2578	297.502.352	10.541.436.781	-10.243.934.429
190	222	758	810	151.652.174	3.312.088.360	-3.160.436.186
157	190	473	286	94.632.557	1.169.453.421	-1.074.820.864
123	157	156	49	31.210.738	200.360.901	-169.150.163
86	123	28	15	5.601.927	61.334.970	-55.733.042
45	86	1	2	200.069	8.177.996	-7.977.927
-	45	3	3	600.206	12.266.994	-11.666.787

De la tabla anterior (“**Tabla 12**”) se puede determinar que el punto de corte que maximiza las utilidades para este modelo es el rango 807-1000, dado que aun cuando se acepten operaciones malas, se obtienen utilidades, no así en el resto de los rangos, donde se debe estudiar de manera más profunda cada solicitud. Luego, solicitudes con un puntaje sobre 807 el modelo las aceptará automáticamente, lo que significa un 21% de cobertura del total de los documentos en el sistema.

Finalmente se propone que los rangos entre 378 y 807 sean analizados de la forma actual, con analistas de riesgo con experiencia discriminando operaciones.

El scorecard final con la determinación de los puntos de corte se puede ver en la “**Tabla 13**” y el “**Gráfico 14**”.

Tabla 13: Scorecard final

Min Score	Max Score	%Buenos Aceptados	% Malos rechazados	%Acierto total	Cobertura Acu
807	1.000	24%	94%	27%	21%
704	807	41%	87%	45%	36%
632	704	54%	78%	57%	48%
575	632	66%	68%	66%	60%
527	575	75%	58%	72%	69%
485	527	82%	49%	74%	76%
447	485	88%	37%	73%	84%
411	447	92%	25%	67%	89%
378	411	95%	15%	64%	93%
346	378	98%	7%	61%	97%
315	346	99%	6%	73%	98%
284	315	99%	4%	64%	99%
253	284	100%	2%	50%	99%
222	253	100%	1%	37%	100%
190	222	100%	0%	48%	100%
157	190	100%	0%	62%	100%
123	157	100%	0%	76%	100%
86	123	100%	0%	65%	100%
45	86	100%	0%	33%	100%
-	45	100%	0%	50%	100%

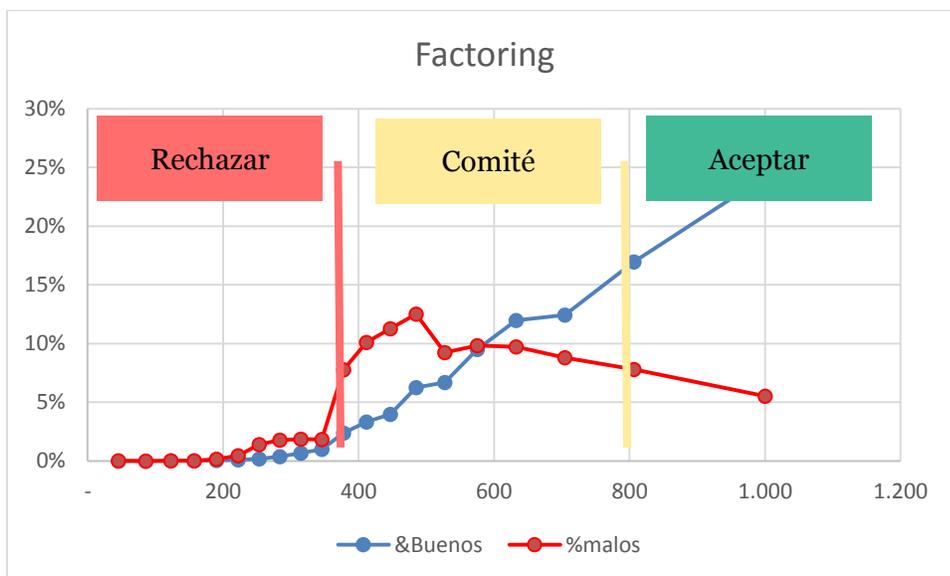


Gráfico 13: Puntos de cortes finales

Para el cálculo de beneficios se utilizarán los resultados obtenidos anteriormente en la base de testeo:

En primer lugar, se tiene que en la base de testeo existen 321.252 documentos, de los cuales 55.201 son documentos que incurren en situación de no pago, con el modelo propuesto, un 15% de estos documentos malos se rechazan manteniendo los niveles de aceptación que hoy en día tiene la empresa. Como fue mencionado anteriormente, en promedio un documento malo tiene un costo cercano a los 4 millones de pesos, por lo que, con el modelo planteado en esta memoria, esta reducción de un 15% genera un ahorro de aproximadamente 33 mil millones para la base de datos analizada.

Extrapolando lo anterior a beneficios anuales, de acuerdo a la presentación de resultados de 2016, las colocaciones por factoring fueron de \$230.375 MM, con un retorno del 13,9%. En promedio esto significa alrededor de 50.000 documentos aproximadamente. Manteniendo la proporción de un 14,88% de documentos en default, 8.120 documentos tendrían mora sobre 90 días. Utilizando el modelo de scoring construido, se hubieran rechazado un 15% de los documentos malos, que significan 1.200 documentos aproximadamente, que se traducen en **\$4.800 MM** menos que se deben provisionar. Esto es un beneficio importante dado que corresponde al **12,6% del margen bruto**.

Por otro lado, se tiene una liberación de recursos en cuanto al tiempo invertido en el procesamiento de aquellos documentos que decantan en comité, en donde anteriormente utilizando la herramienta NAAO un 85% de las solicitudes decantaban en este, mientras

que, con el modelo propuesto en esta memoria, tan solo un 57% cae en este estado de comité, **liberando un 28% de los recursos** destinados al análisis de riesgo.

8. Conclusiones

8.1 Conclusiones generales

En términos generales, el desarrollo del modelo durante la memoria presentó varias dificultades con respecto a los datos. La información que requiere un modelo de admisión para poder ser verdaderamente una ventaja competitiva, debe permitir recrear el contexto en el cual se aceptó o rechazó un crédito, es decir, si un cliente solicitó un crédito hace unos años atrás, el escenario de ese momento, representada en las variables asociadas a la operación, conocida como la “foto”, de ser guardado para así poder compararlo con las solicitudes del presente.

Esta información tiene mayor dificultad y costo de almacenamiento, a diferencia de las variables utilizadas en un modelo de comportamiento, en el cual el cliente permanece dentro del sistema de la empresa y es más sencillo realizar un seguimiento.

El valor que agrega un modelo de admisión es la posibilidad de discriminar correctamente a un cliente nuevo de manera rápida, sobretodo tomando en cuenta el aumento de la población que tiene acceso al crédito, donde el número de participantes que previamente no eran rentables para el sistema, aumenta todos los días y solicitan financiamiento

Ahora más específicamente en el producto estudiado en la memoria, el problema anterior toma una segunda derivada, al entrar en juego un tercer participante, el deudor, el cual, como se presentó en la motivación de esta memoria, es el que por definición debe responder por la factura, y el cliente toma una figura similar a la de un aval. Esto genera un desafío mayor de frente al desarrollo de un modelo de admisión, dado que no instancia en la cual la empresa de factoring pueda solicitar información al deudor, dado que su negocio fue con proveedor de este, el cliente de la institución financiera otorgante de factoring.

Además, los clientes de factoring no poseen el número de operaciones suficientes como para poder estimar modelos paramétrico que calculen el riesgo de default de un cliente, por eso la desviación de la literatura general del riesgo crédito que toma esta memoria, poniendo el foco en las facturas. Esto trae consecuencias, que pueden ser observadas en la “**tabla 16**” donde existe un 6% de facturas que incurren en default que obtuvieron un puntaje sobre 807. Esto indica que hay operaciones que incurren en default que no pueden explicarse con las variables actuales del modelo.

Tomando en cuenta todo lo anterior, es remarcable que siendo la primera iteración de un modelo de admisión para factoring, se hayan obtenido insights y resultados interesantes del negocio, presentados en las conclusiones del trabajo en el siguiente apartado.

8.2 Conclusiones del trabajo realizado

Se concluye que se cumple el objetivo general, el cual se definió como la construcción de un modelo que estimara las probabilidades del default de un documento. Durante el desarrollo del trabajo se calcularon las probabilidades para cada documento y luego fueron re-escaladas para facilitar su interpretación. Además, se obtiene una mejora importante en la admisión de facturas con respecto a la herramienta en uso actualmente NAAO. Se mejora en un 6% el porcentaje de aceptación inmediata, obteniendo un 21% total de solicitudes admitidas automáticamente. Independiente de la mejora en términos de estadísticas y números, lo que se logra verdaderamente es un modelo que tiene un respaldo técnico detrás, donde existen argumentos estadísticos para rechazar o aceptar operaciones, más allá de si no cumplió con algunos de los 17 parámetros del modelo anterior. Además, esto genera un beneficio para los ejecutivos, dado que al ser un modelo simple de entender, pueden observar y comprender de manera rápida, lógica y objetiva, riesgo de una operación.

Con respecto al primer objetivo específico, se pudieron determinar las variables con mayor incidencia en la probabilidad de default dentro del número acotado de variables utilizables que existen en la base de datos de la empresa. Destacan los resultados referentes al porcentaje de descuento, tasa de operación, monto del documento y línea aprobada

La primera variable determina que un menor porcentaje disminuye de manera importante el riesgo de no pago de un documento. Si bien es verdad que en el mercado del factoring es común observar porcentajes de descuento de un 100% dada la competencia que existe en la industria, empezar a generar estrategias de segmentación para ir comprendiendo a los clientes del factoring y eventualmente ofrecerle promociones más personalizadas podrían generar buenos resultados. Si bien la intuición indica que una empresa acude al factoring por problemas de liquidez, cabe la posibilidad que en determinadas instancias no necesite ese 100% de adelanto, si no más bien con la mitad de ese monto esté satisfecho, por ejemplo. En caso de investigar esta arista se debe tener cuidado en que el cliente comprenda que este mayor porcentaje le pertenece a él, y que la empresa de factoring devolverá este excedente una vez saldada la operación.

Los resultados obtenidos para la variable **tasa de operación** indican de manera que la empresa tiene mal determinadas las tasas ya que estas no están alineadas de acuerdo al riesgo de las operaciones. Si bien existen trabajos actualmente en proceso en el producto

de factoring relacionados a este tema, es recomendable que se haga una revisión de las tasas de colocación a nivel general de la empresa.

La variable **monto del documento** se destaca dado que cumple con uno de los principales principios de las finanzas, el cual es la diversificación. Documentos de montos más pequeños asociados a una misma operación, la cual puede o no tener más de un deudor, disminuye la probabilidad de no pago con respecto a facturas que por los montos más grandes pueden parecer más atractivas desde el punto de vista del retorno.

Por último, la variable **línea aprobada** muestra, por una parte, que empresas pequeñas poseen un mayor riesgo, resultado esperado dado que poseen una menor estabilidad financiera, pero además, empresas grandes presentan un menor puntaje. Esto se interpreta como el poder de negociación que tiene una empresa de mayor envergadura con respecto a la empresa de factoring, dado que pueden permitirse saldar la deuda cuando sea más conveniente, además de tener experiencia en cobranzas judiciales, por lo que se encuentran bien preparados en caso de recibir demandas por partes de clientes y/o entes financieros.

Se debe mencionar también que no se pudo obtener variables históricas del cliente y deudor con calidad aceptable para ser incluidas en el modelo. Este tipo de variables son las que mejor predicen el default de acuerdo a la literatura revisada de credit scoring.

Con respecto al segundo objetivo específico, se cumple el modelo de estimación para el cálculo de las probabilidades utilizando variables del cliente y variables operacionales. En la base de la empresa variables relacionadas al deudor no contenían información utilizable, en su mayoría datos nulos, por lo que no se completa a cabalidad este objetivo.

Finalmente se concluye que se cumple con el objetivo de la evaluación cuantitativa del modelo de acuerdo a como se planteó dentro de los alcances. Este primer modelo genera beneficios tanto en términos monetarios como en la liberación de recursos. Queda propuesto poder estimar los beneficios en términos de provisiones que puede derivar de este trabajo, dada las probabilidades de default ya calculadas. Es posible estimar una Pérdida dado el incumplimiento (PDI) utilizando las estimaciones generadas por la SBIF en su compendio de normas contables (2008).

9. Recomendaciones

En primer lugar, se sugiere implementar el modelo de scoring directamente a la base de datos a través del software SQL server que utiliza la empresa. Este software, creada por la empresa Microsoft, cuenta con una herramienta llamada SQL Server Managment Studio,

que permite especificar una expresión para una columna calculada. De esta manera, es posible llevar un registro de los scores históricos, y estudiar cómo se va comportando la variable dependiente con respecto a ellos. De esta forma el puntaje se obtendría a medida que se ingresan los datos de la solicitud los ejecutivos comerciales en la base de datos y se iría calculando de manera automática el score y guardándose en una columna extra en la tabla de la base en tiempo real. Esto permite además obtener un registro histórico de los scores obtenidos en el pasado y poder ir mejorándolo.

Dado el bajo número de variables de admisión en comparación a variables de comportamiento, se recomienda guardar información consultada de bases externas con sus respectivas fechas de consulta. Esto para poder tener varias “fotos” de un cliente y deudor en particular, permitiendo realizar modelos más sofisticados a futuro, ya sea utilizando datos de panel o alguna otra metodología.

Además, se recomienda establecer relaciones con empresas externas que manejen información de empresas como lo son Equifax, Datamart y NETPAG, para que no solamente comprarles información, sino también, en analizar esa información y convertirla en valiosos inputs para futuros modelos. Esto puede ser fundamental en la obtención de variables relacionadas a los deudores, dado la imposibilidad de obtenerla directamente.

Con respecto al modelo en si, como consideración a futuro, el modelo debe volver a ser calibrado utilizando la información nueva cada 3 a 6 meses, dado que la distribución de los casos malos puede variar en un futuro, ya sea por variables externas como los ciclos económicos como también variables internas de la empresa, como puede ser la modificación de la definición de default, ya que al no ser una entidad bancaria, no cae bajo la supervisión del ente regulador en la gestión del riesgo.

Finalmente, se propone continuar este estudio realizando una estimación de la **pérdida dado el incumplimiento** para el caso del factoring, de manera de poder generar un modelo de provisiones utilizando las probabilidades de incumplimiento calculadas en esta memoria. Además, se propone extender el estudio desde el punto de vista del pricing, pudiendo establecer un piso mínimo para la tasa de operación que se debe utilizar acorde con el riesgo de la solicitud.

10. Bibliografía

1. **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).
2. **Bravo, C.** (2012) "*Métodos para estimar riesgo crediticio en base a minería de datos y teoría de juegos*". Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Santiago.
3. **Fayyad, U.** (1996) "*Data mining and knowledge discovery- making sense out of data*". IEEE Expert-Intelligent Systems and their Applications, Vol. 11 (pp. 20-25).
4. **Beaver, W.H.** (1966) "*Financial ratios as Predictor of Failure*". Accounting Research, Vol. 4, (pp. 71-111).
5. **Altman, E.I.** (1968) "*Finantial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*". Finance, Vol. 23, (pp. 589-609).
6. **Camara de Diputados.** (2017) "Comisión de Economía analizó la industria del factoring con asociación gremial". [En línea] https://www.camara.cl/prensa/noticias_detalle.aspx?prmId=131865 [Fecha consulta: 4 de Junio 2017]
7. **Martens, D., Baesens, B., Van Gestel, T. & Vanthienen, J.** (2006) "*Comprehensible Credit Scoring Models using Rule Extraction from Support Vector Machines*". European Journal of Operational Research, Vol. 183, (pp. 1466-1476).
8. **Siddiqi, N.** (2006) "*Credit Risk Scorecards: Developing and implementing intelligent Credit Scoring*". John Wiley & Sons Inc, New Jersey.
9. **SBIF.** (2008) "*Compendio de normas contables*". [En línea] https://www.sbif.cl/sbifweb/internet/archivos/norma_6545_1.pdf [Fecha consulta: 10 de Noviembre 2017.]
10. **Tanner.** (2016) "*Presentación corporativa 2016*". Santiago, Chile.

11. Anexos

11.1 Anexo 1: Campos seleccionados y tratamiento de outliers

Campo	Tipo de medida	Outliers	Extremos	Acción	%Completo
id_operación	Continua	0	0	Ninguna	100
id_documento	Continua	0	0	Ninguna	100
valor_nominal	Continua	1.720	1.552	Forzar outliers / anular extremos	100
valor_futuro	Continua	1.774	1.625	Forzar outliers / anular extremos	100
días_documento	Continua	1	271	Forzar outliers / anular extremos	99,68
glosa_estado_documento	Catagórica	--	--	--	100
saldo	Continua	526	565	Forzar outliers / anular extremos	100
fecha_vencimiento_documento	Continua	0	0	Ninguna	97,43
fecha_vencimiento_prorroga	Continua	0	0	Ninguna	100
fecha_emisión	Continua	0	0	Ninguna	19,38
fecha_ingreso_documento	Continua	0	0	Ninguna	99,76
abono_deudor	Continua	5.207	4.916	Forzar outliers / anular extremos	99,76
saldo_deudor	Continua	631	797	Forzar outliers / anular extremos	99,76
saldo_doc_deudor	Continua	19	169	Forzar outliers / anular extremos	58,70
fecha_ultimo_pago	Catagórica	--	--	--	100
estado_deposito	Continua	0	0	Ninguna	99,76
interes_documento	Continua	697	520	Forzar outliers / anular extremos	99,76
fecha_deposito	Catagórica	--	--	--	100
fecha_protesto	Catagórica	--	--	--	100
comision_cobranza	Continua	277	300	Forzar outliers / anular extremos	99,66
Est_doc_deudor_	Continua	0	0	Ninguna	99,76
tasa_operacion	Continua	3149	1.200	Forzar outliers / anular extremos	100
tasa_mora	Continua	377	2.058	Forzar outliers / anular extremos	41,21
monto_comision_total	Continua	7698	1.427	Forzar outliers / anular extremos	92,61
tipo_cambio_operacion	Continua	0	929	Ninguna	100
porcentaje_descuento	Continua	91.471	0	Forzar outliers / anular extremos	99,88
diferencia_prepago	Continua	205	744	Forzar outliers / anular extremos	41,29
tipo_pagare	Continua	0	0	Ninguna	41,00

Campo	Tipo de medida	Outliers	Extremos	Acción	%Completo
tipo_cobranza	Continua	0	2.416	Forzar outliers / anular extremos	99,88
con_notificacion	Continua	0	2.468	Forzar outliers / anular extremos	41,21
con_custodia	Continua	0	572	Forzar outliers / anular extremos	99,88
tipo_operacion	Catagórica	--	--	--	100
devolución_intereses	Catagórica	--	--	--	100
monto_abono	Continua	4.394	3.812	Forzar outliers / anular extremos	84,86
fecha_abono	Continua	0	0	Ninguna	84,86
abono_mora	Continua	4.625	4.147	Forzar outliers / anular extremos	84,86
abono_capital	Continua	4.268	3.696	Forzar outliers / anular extremos	84,86
abono_interes	Continua	3.556	3.182	Forzar outliers / anular extremos	84,86
dias_mora	Continua	4.529	5.157	Forzar outliers / anular extremos	84,86
tipo_imputacion	Catagórica	--	--	--	100
tipo_cierre	Catagórica	--	--	--	100
recaudado	Continua	1.241	1.104	Forzar outliers / anular extremos	26,72
amortizado	Continua	1.226	1.088	Forzar outliers / anular extremos	26,72
reajustes	Continua	4	34	Forzar outliers / anular extremos	26,72
interes_penal	Continua	1.117	1.042	Forzar outliers / anular extremos	26,72
interes_reembolsar	Continua	154	309	Forzar outliers / anular extremos	26,72
total_reembolso	Continua	451	402		26,72
gastos	Continua	4	35	Forzar outliers / anular extremos	26,72
saldo_cliente	Continua	191	307	Forzar outliers / anular extremos	26,72
tipo_pago	Catagórica	--	--	--	100
iva_debito_bienes	Catagórica	--	--	--	100
iva_debito_intereses	Catagórica	--	--	--	100
iva_debito_mora	Continua	0	0	Ninguna	19,00
monto_aprobado	Continua	29.115	2.148	Forzar outliers / anular extremos	100
fecha_ingreso_linea	Continua	0	0	Ninguna	100
fecha_vencimiento_linea	Continua	0	0	Ninguna	100
glosa_estado_linea	Catagórica	--	--	--	100
spread_minimo	Continua	682	6.837	Forzar outliers / anular extremos	98,85
fecha_otorgamiento	Continua	0	0	Ninguna	40,14
monto_solicitado	Catagórica	--	--	--	100

Campo	Tipo de medida	Outliers	Extremos	Acción	%Completo
numero_trabajadores	Continua	925	417	Forzar outliers / anular extremos	19,95
contabilidad_completa	Catagórica	--	--	--	100
glosa_motivo_protesto	Catagórica	--	--	--	100
glosa_estado_prorroga	Catagórica	--	--	--	100
monto_protesto	Continua	0	0	Ninguna	0
saldo_protesto	Continua	0	0	Ninguna	0
glosa_zona	Catagórica	--	--	--	100
mora	Continua	50.678	63	Ninguna	99,79
Default	Catagórica	--	--	--	100
Tipo_cambio	Continua	--	--	--	100

11.2 Anexo 2: NAAO

NAAO: Reglas Externas



Punto de Análisis	Prueba	Criterio	Var. parámetro	Parámetros junio	Mensajes de Error
Externo Cliente	Protestos	Protestos Cheques / Protestos Letras	Q Protestos	0	Cliente con protestos en Bureaus
	Morosidad	"Mora 30 días" - "Mora 30 a 60 días" - "Mora 60 a 90 días" - "Mora > 90 días" - D. directa morosa - D. indirecta vencida - D. castigada indirecta - D. castigada directa - protestos pagarés - otros protestos - Deuda Total %	% de Deuda Total	5%	Cliente con mora en Bureaus
	Mora Previsional	$\sum \text{Monto Deuda Impos} / \text{Deuda Total} \times 100\%$	% de Deuda Total	10%	Cliente con mora previsional en Bureaus
Externo Deudor	Protestos	Protestos Cheques / Protestos Letras	Q Protestos	0	*RUT*: Deudor con protestos en Bureaus
	Morosidad	"Mora 30 días" - "Mora 30 a 60 días" - "Mora 60 a 90 días" - "Mora > 90 días" - D. directa morosa - D. indirecta vencida - D. castigada indirecta - D. castigada directa - protestos pagarés - otros protestos - Deuda Total %	% de Deuda Total	0%	*RUT*: Deudor con mora en Bureaus
	Mora Previsional	$\sum \text{Monto Deuda Impos} / \text{Deuda Total} \times 100\%$	% de Deuda Total	20%	*RUT*: Deudor con mora previsional en Bureaus

Punto de Análisis	Prueba	Criterio	Var. parámetro	Parámetros junio	Mensajes de Error
Línea: Monto	Exceso de línea	$(\text{Línea Ocupada} + \text{Operación} - \text{Línea Aprobada}) / \text{Línea Aprobada} \leq X\%$	Exceso de Línea	0%	Exceso de Línea
Línea: Vencimiento	Línea Vencida	$\text{Fecha Hoy} - \text{Fecha Venc.} \leq X \text{ días}$	Días vencida	0	Línea Vencida
Línea: Anticipo	% de Anticipo	$\% \text{ Anticipo} - \text{Operación} - \% \text{ Antic. Aprobado (T. Documento)} \leq X\%$	Diferencia	0%	% de Anticipo
Línea: Plazos	Plazo de pago de documento	$\text{Fecha Venc. Doc. Oper.} - \text{Fecha Hoy} \leq X \text{ días}$ $\text{Fecha Venc. Doc. Oper.} - \text{Fecha Hoy} > 0$	Días	180	Documento *#doc*: Plazo de pago de documento
Línea: Exposición por producto	Exposición de producto	$\text{FOR: Saldo Cartera Vigente (PR)} / \text{Línea Aprobada} \leq \text{Concentración global Aprobada (PR)} \leq X\%$	Diferencia %	0%	Exposición de Producto
Línea: Concentración	Concentración (CxD)	$\text{FOR: Concentración Operación} - \text{O (Concentración Normal; Concentración Especial)} \leq X\%$	Diferencia %	0%	*RUT*: Deudor con Exceso de Concentración

NAAO: Reglas relativas al Comportamiento Histórico



Categoría	Punto de Análisis	Prueba	Criterio	Casos	Var. parámetro	Parámetros junio	Mensajes de Error
Comportamiento de pago: Cliente	Cliente: Mora	Morosidad Interna Cliente	$\bullet \text{Max}(\text{días} \text{ mora } \text{ Cartera } \text{ morosa}; \text{ fecha } \text{ hoy} - \text{min}(\text{fecha} \text{ vencimiento } \text{ Cartera } \text{ morosa} \text{ CR})) \leq \text{DD} \text{ días}$ $\bullet \text{Monto } \text{ Cartera } \text{ morosa} \leq \text{Monto } \text{ Cartera } \text{ Aprobada}$ $\bullet \text{Monto } \text{ Cartera } \text{ morosa} \leq \text{Monto } \text{ Cartera } \text{ CR} \text{ Excedentes } \text{ disponibles} \leq \text{Excedentes } \text{ pendientes}$	Caso 1	Días Mora	30	Mora Interna Cliente
					% Línea	30%	
	Caso 2	Días Mora	15				
		% Línea	50%				
Cliente: CxC	Cheques y Letras protestadas en Cartera	Cheques Protestados + CxC en Cartera ≤ X		Monto	0	Cliente con cheques y/o letras protestadas en Cartera CxC	
Cliente: Protestos	Protestos Internos	$\text{Fecha } \text{ hoy} - \text{Max}(\text{fecha} \text{ de } \text{ último } \text{ protesto } \text{ cheques}; \text{ fecha} \text{ de } \text{ último } \text{ protesto } \text{ letras}) \leq \text{X} \text{ meses}$		Meses	1	Protesto Interno Cliente	
Comportamiento de pago: Deudor	Deudores: Protestos	Protestos Internos	$\text{Fecha } \text{ hoy} - \text{Max}(\text{fecha} \text{ de } \text{ último } \text{ protesto } \text{ cheques}; \text{ fecha} \text{ de } \text{ último } \text{ protesto } \text{ letras}) \leq \text{X} \text{ meses}$		Meses	3	*RUT*: Deudor con Protestos Internos
	Deudores: Alertas	Deudores bloqueados, en lista negra, con alertas de otros factoring, etc.	$\text{Montos } \text{ anticipados } \text{ de } \text{ Deudores } \text{ con } \text{ alertas} \leq \text{Monto } \text{ operación } \leq \text{X} \%$		% Operación	25%	*RUT*: Deudor/Cliente con Alertas

11.3 Anexo 3: Resumen Análisis Univariado

Field	Sample Graph	Measurement	Field	Sample Graph	Measurement
numero_trabajador...		Continuous	id_operacion		Continuous
contabilidad_comp...		Categorical	id_documento		Continuous
glosa_motivo_prot...		Categorical	valor_nominal		Continuous
glosa_estado_prorr...		Categorical	valor_futuro		Continuous
monto_protesto		Continuous	dias_documento		Continuous
saldo_protesto		Continuous	glosa_estado_docu...		Categorical
glosa_zona		Categorical	saldo		Continuous
mora		Continuous	fecha_vencimiento...		Continuous
Default		Categorical	fecha_vencimiento...		Continuous
Tipo cambio		Flag	fecha_emision		Continuous

Field	Sample Graph	Measurement	Field	Sample Graph	Measurement
fecha_ingreso_doc...		Continuous	est_doc_deudor		Continuous
abono_deudor		Continuous	tasa_operacion		Continuous
saldo_deudor		Continuous	tasa_mora		Continuous
saldo_doc_deudor		Continuous	monto_comision_t...		Continuous
fecha_ultimo_pago		Categorical	tipo_cambio_opera...		Continuous
estado_deposito		Continuous	porcentaje_descue...		Continuous
interes_documento		Continuous	diferencia_prepago		Continuous
fecha_deposito		Categorical	tipo_pagare		Continuous
fecha_protesto		Categorical	tipo_cobranza		Continuous
comision_cobranza		Continuous	con_notificacion		Continuous

Field	Sample Graph	Measurement	Field	Sample Graph	Measurement
con_custodia		Continuous	tipo_cierre		Categorical
tipo_operacion		Categorical	recaudado		Continuous
devolucion_intereses		Categorical	amortizado		Continuous
monto_abono		Continuous	reajustes		Continuous
fecha_abono		Continuous	interes_penal		Continuous
abono_mora		Continuous	interes_reembolsar		Continuous
abono_capital		Continuous	total_reembolsar		Continuous
abono_interes		Continuous	gastos		Continuous
dias_mora		Continuous	saldo_cliente		Continuous
tipo_imputacion		Categorical	tipo_pago		Categorical

11.4 Anexo 4: Filtro de variables

Campo	
id_operación	Filtrada
id_documento	Filtrada
valor_nominal	Utilizada
valor_futuro	Filtrada
días_documento	Utilizada
glosa_estado_documento	Filtrada
saldo	Filtrada
fecha_vencimiento_documento	Filtrada
fecha_vencimiento_prorroga	Filtrada
fecha_emisión	Filtrada
fecha_ingreso_documento	Filtrada
abono_deudor	Filtrada
saldo_deudor	Filtrada
saldo_doc_deudor	Filtrada
fecha_ultimo_pago	Filtrada

estado_deposito	Filtrada
interes_documento	Utilizada
fecha_deposito	Filtrada
fecha_protesto	Filtrada
comision_cobranza	Filtrada
Est_doc_deudor_	Filtrada
tasa_operacion	Utilizada
tasa_mora	Utilizada
monto_comision_total	Filtrada
tipo_cambio_operacion	Filtrada
porcentaje_descuento	Utilizada
diferencia_prepago	Filtrada
tipo_pagare	Utilizada
tipo_cobranza	Filtrada
con_notificacion	Utilizada
con_custodia	Utilizada
tipo_operacion	Utilizada
devolución_intereses	Filtrada
monto_abono	Filtrada
fecha_abono	Filtrada
abono_mora	Filtrada
abono_capital	Filtrada
abono_interes	Filtrada
días_mora	Filtrada
tipo_imputacion	Filtrada
tipo_cierre	Filtrada
recaudado	Filtrada
amortizado	Filtrada
reajustes	Filtrada
interes_penal	Filtrada
interes_reembolsar	Filtrada
total_reembolsar	Filtrada
gastos	Filtrada
saldo_cliente	Filtrada
tipo_pago	Filtrada
iva_debito_bienes	Filtrada
iva_debito_intereses	Filtrada
iva_debito_mora	Filtrada
monto_aprobado	Utilizada

fecha_ingreso_linea	Filtrada
fecha_vencimiento_linea	Filtrada
glosa_estado_linea	Filtrada
spread_minimo	Utilizada
fecha_otorgamiento	Filtrada
monto_solicitado	Utilizada
numero_trabajadores	Utilizada
contabilidad_completa	Utilizada
glosa_motivo_protesto	Utilizada
glosa_estado_prorroga	Filtrada
monto_protesto	Filtrada
saldo_protesto	Filtrada
glosa_zona	Utilizada
mora	Filtrada
Default	Utilizada
Tipo_cambio	Utilizada

11.5 Anexo 5: Matriz de covarianza

	valor_nominal	dias_documento	intereses_documento	tasas_operacion	tasas_mora	porcentaje_descuento	tipo_pagare	comunicacion	com custodia	monto_probado	spread_minimo	numero_trabajadores	contabilidad_completa	Default
valor_nominal	1													

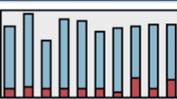
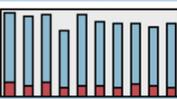
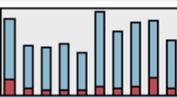
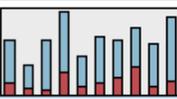
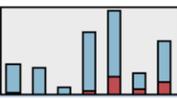
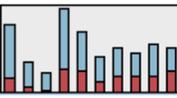
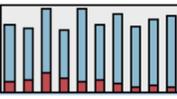
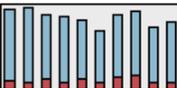
dia_s_documento	-.013**	1												
interes_documento	.799**	.102**	1											
tas_a_operacion	-.082**	-.041**	-.011**	1										
tas_a_mora	-.048**	-.019**	-.019**	.013**	1									
porcentaje_descuento	.016**	-0.001	.036**	-.014**	-.215**	1								
tipo_pagare	--	--	--	--	--	--	1							
con_notificacion	-.009**	-.006**	-.012**	-.033**	-.005**	-.041**	--	1						
con_cust	.006**	-.003**	.004**	-.026**	.016**	-.011**	--	.379**	1					

odi_a														
monto_a_probado	.024**	.020**	0.001	-.160**	-.242**	.125**	--	.003*	-.001	1				
spread_m_inimo	.015**	.015**	.007**	-.033**	.012**	-.108**	--	-.005**	-.002**	-.003**	1			
numero_trabajadores	-.056**	.011**	-.051**	-.141**	-.139**	.146**	--	-.038**	-.016**	.876**	--	1		
contabilidad_completa	.042**	.012**	.041**	-.014**	-.116**	.116**	--	-.011**	.014**	-.066**	-.040**	-.250**	1	
Default	.016**	.024**	.025**	-.133**	.155**	.149**	--	-.001	.028**	.106**	-.038**	.091**	.069**	1
** Correlation is significant at the 0.01 level (2-tailed).														
* Correlation is significant at the 0.05 level (2-tailed).														
b Cannot be computed because at least one of the variables is constant.														

11.6 Anexo 6: Resultados Logit

Variables en la ecuación							
Step 1a		B	S,E,	Wald	df	Sig,	Exp(B)
	chaidvalor_nominal			714,36	10	0	
	chaidvalor_nominal(1)	-0,597	0,069	75,03	1	0	0,55
	chaidvalor_nominal(2)	-0,688	0,069	99,50	1	0	0,503
	chaidvalor_nominal(3)	-0,526	0,069	58,29	1	0	0,591
	chaidvalor_nominal(4)	-0,663	0,069	92,41	1	0	0,516
	chaidvalor_nominal(5)	-0,601	0,069	75,97	1	0	0,549
	chaidvalor_nominal(6)	-0,546	0,069	62,90	1	0	0,579
	chaidvalor_nominal(7)	-0,638	0,069	85,84	1	0	0,528
	chaidvalor_nominal(8)	-0,473	0,069	47,27	1	0	0,623
	chaidvalor_nominal(9)	-0,587	0,069	72,76	1	0	0,556
	chaidvalor_nominal(10)	-0,427	0,069	38,69	1	0	0,652
	chaidtasa_operacion			13007,9	10	0	
	chaidtasa_operacion(1)	0,548	0,097	31,78	1	0	1,73
	chaidtasa_operacion(2)	0,583	0,097	35,83	1	0	1,791
	chaidtasa_operacion(3)	0,947	0,097	94,89	1	0	2,578
	chaidtasa_operacion(4)	0,865	0,097	79,00	1	0	2,376
	chaidtasa_operacion(5)	-0,037	0,098	0,14	1	0,704	0,964
	chaidtasa_operacion(6)	0,53	0,097	29,61	1	0	1,699
	chaidtasa_operacion(7)	0,454	0,097	21,72	1	0	1,575
	chaidtasa_operacion(8)	-0,097	0,098	0,98	1	0,322	0,908
	chaidtasa_operacion(9)	0,024	0,098	0,062	1	0,804	1,025
	chaidtasa_operacion(10)	-0,072	0,098	0,55	1	0,46	0,93
	chaidporcentaje_descuento			13173,79	7	0	
	chaidporcentaje_descuento(1)	-9,72	0,501	376,30	1	0	0
	chaidporcentaje_descuento(2)	-9,409	0,501	352,61	1	0	0
	chaidporcentaje_descuento(3)	-8,828	0,501	310,04	1	0	0
	chaidporcentaje_descuento(4)	-9,025	0,501	324,63	1	0	0
	chaidporcentaje_descuento(5)	-8,37	0,501	279,24	1	0	0
	chaidporcentaje_descuento(6)	-8,493	0,501	287,42	1	0	0
	chaidporcentaje_descuento(7)	-8,618	0,501	296,06	1	0	0
	glosa_zona			158368,66	11	0	
	glosa_zona(1)	-0,625	0,043	208,27	1	0	0,535
	glosa_zona(2)	-2,903	0,019	24453,97	1	0	0,055
	glosa_zona(3)	-2,472	0,019	17503,73	1	0	0,084
	glosa_zona(4)	-1,747	0,04	1914,739	1	0	0,174
	glosa_zona(5)	-2,096	0,059	1252,56	1	0	0,123
	glosa_zona(6)	-1,289	0,08	257,70	1	0	0,276
	glosa_zona(7)	-3,805	0,019	38960,85	1	0	0,022
	glosa_zona(8)	-0,261	0,018	200,80	1	0	0,77
	glosa_zona(9)	-2,913	0,021	20152,15	1	0	0,054
	glosa_zona(10)	-6,228	0,303	422,18	1	0	0,002
	glosa_zona(11)	-3,391	0,02	29623,63	1	0	0,034
	Constant	9,658	0514	352,60	1	0	15645,90

11.7 Anexo 7: Distribución categorización variables continuas

 chaidcontabilidad_co...		 Nominal
 chaiddias_documento		 Nominal
 chaidinteres_docume...		 Nominal
 chaidmonto_aprobado		 Nominal
 chaidnumero_trabaja...		 Nominal
 chaidporcentaje_desc...		 Nominal
 chaidspread_minimo		 Nominal
 chaidtasa_mora		 Nominal
 chaidtasa_operacion		 Nominal
 chaidvalor_nominal		 Nominal

11.8 Anexo 8: Tramos por punto de corte

Tramo	P(x)	Odd	Ln(odds)	Score
Tramo1	0,02489	39,18	3,67	1.000
Tramo2	0,06383	14,67	2,69	807
Tramo3	0,10278	8,73	2,17	704
Tramo4	0,14173	6,06	1,80	632
Tramo5	0,18067	4,53	1,51	575
Tramo6	0,21962	3,55	1,27	527
Tramo7	0,25856	2,87	1,05	485
Tramo8	0,29751	2,36	0,86	447
Tramo9	0,33646	1,97	0,68	411
Tramo10	0,37540	1,66	0,51	378
Tramo11	0,41435	1,41	0,35	346
Tramo12	0,45330	1,21	0,19	315
Tramo13	0,49224	1,03	0,03	284
Tramo14	0,53119	0,88	-0,12	253
Tramo15	0,57013	0,75	-0,28	222
Tramo16	0,60908	0,64	-0,44	190
Tramo17	0,64803	0,54	-0,61	157
Tramo18	0,68697	0,46	-0,79	123
Tramo19	0,72592	0,38	-0,97	86
Tramo20	0,76486	0,31	-1,18	45