Universidad de Valparaíso
CHILE
Facultad de Ciencias
Instituto de Estadística

Una revisión a la metodología de selección de variables utilizada en modelos de riesgos por el Banco de Crédito e Inversiones

Proyecto de titulación para optar al titulo de:
Ingeniero en Estadística
Investigado por:
Matías Rolando Barría Olivares

Profesor guía:
Harvey Rosas Q., Ph.D.
Universidad de Valparaíso

Co-guía:
Ing. Manuel Caro Riveros
Banco de Crédito e Inversiones

Valparaíso, Diciembre del 2018

Índice general

Índice de figuras

Índice de cuadros

Resumen

¿Como identificar si un nuevo cliente que realiza una solicitud de crédito va a hacer devolución del préstamo?, ¿es posible gestionar de mejor forma las operaciones realizadas por los clientes del banco?, ¿es factible tener un sistema que permita alertar sobre un posible atraso en los compromisos crediticios de los clientes? Todas estas preguntas pueden ser aclaradas entendiendo como se aplican los modelos estadísticos en las instituciones que otorgan créditos.

En este trabajo de título, se estudió las metodologías que tiene actualmente el BCI para construir modelos de riesgos, analizando en base a la teoría si son adecuadas dado el contexto y transformaciones de variables aplicadas en instancias preliminares. En este trabajo de título se logró eliminar ciertas metodologías que no se estaban utilizando de forma adecuada, y se propusieron métodos análogos para reducir dimensiones de variables.

Abstract

How to identify if a new client that makes a credit request is going to repay the loan? Is it possible to better manage the operations performed by the bank's clients? Is it possible to have a system that allows to alert about a possible delay in the credit commitments of the clients? All of these questions can be clarified by understanding how statistical models are applied in institutions that grant creeds.

In this title project, it was studied the methodologies that the BCI currently has to build risk models, analyzing based on the theory if they are appropriate given the context and transformations of variables applied in preliminary instances. In this project it was possible to eliminate certain methodologies that were not being used properly, and analogous methods were proposed to reduce event dimensions / or variable.

Capítulo 1

Introducción

Existen numerosas técnicas estadísticas avanzadas que permiten predecir el comportamiento futuro de las personas (Mays, 2001). Algunas de estás técnicas suelen ser más utilizadas que otras (por su facilidad de uso u otro aspecto), por lo que la elección de las técnicas quedan a criterio del especialista. Sin embargo, en la industria bancaria uno de los métodos más utilizados para predecir un comportamiento dicotómico de los clientes es la regresión logística (Cox y Shell, 1989). Por ejemplo, para predecir si un cliente cumplirá en los plazos establecidos con la devolución del crédito solicitado se utilizan modelos de puntuación crediticia (modelos estadísticos que asignan un puntaje determinado a los clientes según su historia crediticia). Todos estos tipos de modelos son llamados modelos de riesgos, ya que permiten mitigar de forma oportuna los riesgos asociados a una operación financiera.

Para el desarrollo de modelos de riesgos, las instituciones financieras siguen un proceso estandarizado por la literatura y organismos que validan este tipo de modelos, de modo que se pueda hacer un seguimiento a las metodologías utilizadas (SBIF, 2014). Una de las fases del proceso de construcción de los modelos de riesgos, es el proceso de selección de variables.

1.1. Objetivos

1.1.1. Objetivo General

- Revisar la metodología utilizada por el BCI en la etapa de reducción de dimensiones en modelos de riesgos del banco de crédito e inversiones.

1.1.2. Objetivos Específicos

- Definir la naturaleza teórica de las variables transformadas usando su peso de la evidencia.
- Comparar los distintos tipos de correlación para el tipo de variables a utilizar en los modelos.
- Proponer un valor critico de aceptación del tipo de correlación utilizada para ingresar o eliminar una variable del modelo.
- Decidir con base a la teoría, si el análisis de componentes principales es una metodología redundante en el proceso de selección de variables.
- Comparar las técnicas de reducción de dimensiones utilizada por el banco, con la metodología propuesta en el presente trabajo de título.

1.2. Hipótesis

Es posible hacer una revisión de la transformación de variables por el peso de la evidencia, para mejorar los procesos de selección de variables en modelos de riesgos del BCI.

1.3. Alcances

El desarrollo de este trabajo de titulación tiene por objetivo definir métodos estadísticos y matemáticos que cuenten con un sustento metodológico adecuado según los datos que se tienen, en ningún caso la construcción de un modelo de riesgo crediticio para que la

compañía ponga en producción. De igual forma, se explicarán temas asociados a la construcción de los modelos de riesgo de crédito con el fin de contextualizar al lector sobre el área e industria en el cual se aplicarán las metodologías de reducción de dimensiones propuestas en este trabajo.

1.4. Contexto BCI

Durante los años 2013 y 2014, BCI llevo a cabo una transformación completa de su área de riesgo, en donde se propuso cuestionarse la forma en la cual el banco hace su tareas, así como también la estructura organizacional que tenía en ese momento. El proyecto, básicamente consistía en poner especial atención en la forma en que el banco construye y explota los modelos de riesgos, tomando en cuenta la posibilidad de utilizar otras herramientas predictivas que ayudarán la gestión del riesgo. Dado lo anterior, el banco proporcionó una variedad de mejoras en cuanto a la generación de modelos y sus aplicaciones, sin embargo, un aspecto relevante que se detectó, fue el descubrir que no había un organismo o área interna del banco que validara los modelos generados para cuantificar el riesgo. Fue así, que nació la unidad de *Validación y Seguimiento de Modelos*, unidad encargada de validar los procesos metodológicos y documentales respectivos en la generación de un modelo de riesgo.

El comité de Basilea provee ciertas directrices en lo que se refiere a la validación de modelos estadísticos (Cruz, 2015):

Un proceso de validación se considera adecuado si abarca los siguientes aspectos

- Si se evalúa el poder predictivo de los modelos o estimaciones de riesgo.
- Si existe una evaluación de los procesos que estiman los parámetros de los modelos, así como si existen procesos que controlarán el poder predictivo futuro de los modelos.
- Si se evalúa la integración del modelo de riesgo a los procesos y gestión del Banco.

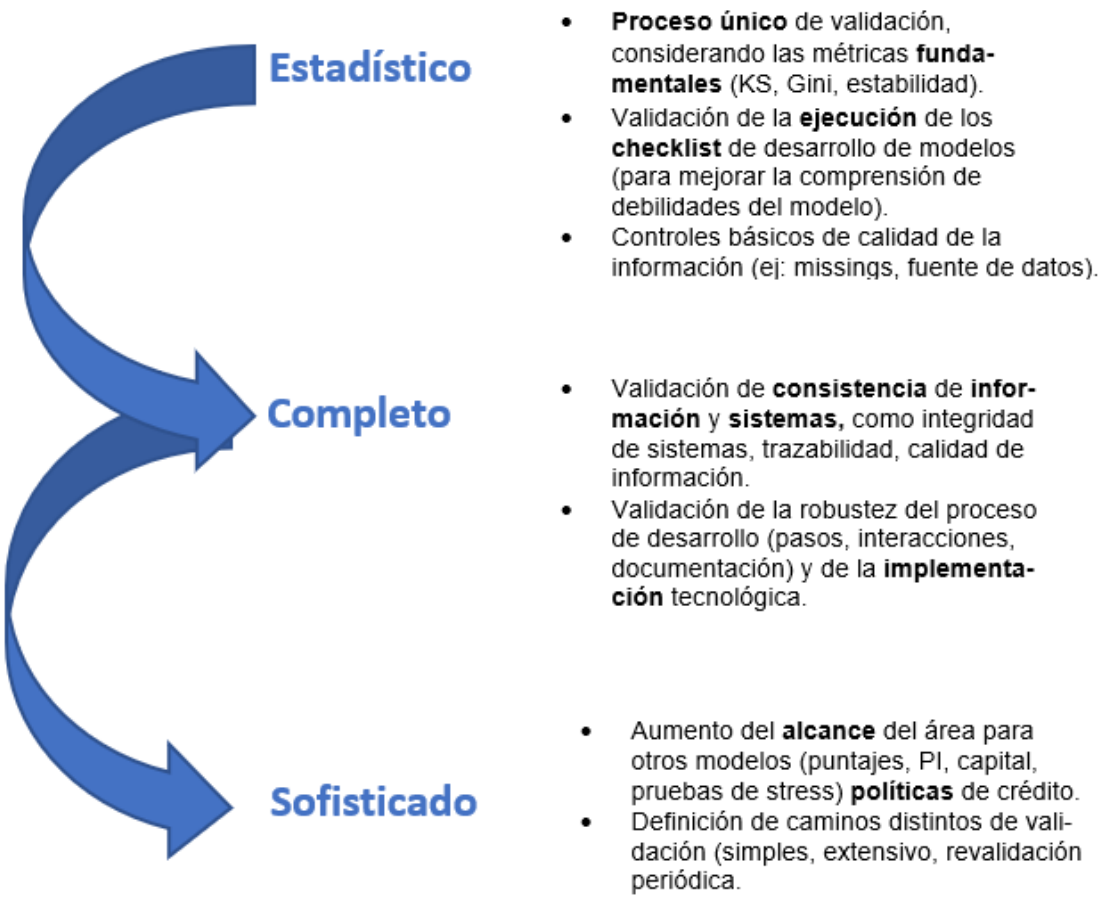

Figura 1.1: Tipos de validación de modelos de riesgos
Fuente: Elaboración propia.

Como se aprecia en la Figura 1, existen diferentes tipos de validación en los MDR, clasificándose según la valorización otorgada por los analistas (se cuantifica cuanto se "ahorra" el banco implementando el modelo). En las validaciones más exhaustivas, se debe por ejemplo, calcular nuevamente el puntaje otorgado por el modelo, tramificar las variables y estimar los parámetros, contrastándolo con los resultados obtenidos por el desarrollador.

Capítulo 2

Metodología para construir un modelo de riesgo

2.1. Riesgo Financiero

Para comprender de mejor forma el contexto financiero en el cual se encuentra desarrollado este trabajo de titulación es necesario estudiar los conceptos asociados al riesgo financiero, así como también los tipos de riesgos asociados a una operación financiera.

El riesgo puede tener muchos significados, sin embargo, las características que más se repiten en la literatura son la aleatoriedad e incertidumbre. Desde el punto de vista financiero, es fundamental tener cuantificados estos riesgos para poder implementar políticas que permitan reducir el impacto a la entidad financiera dado el incumplimiento de las obligaciones financiera de sus clientes. En general, se conocen cuatro tipo de riesgos financieros (SBIF,2009):

- Riesgo operacional: es el riesgo de sufrir un impacto negativo en la empresa debido a errores en los procedimientos internos por parte de los colaboradores. Estos errores internos se pueden deber a fraudes realizados por operadores, fallas en los sistemas de integración tecnológicos, entre otros.

- Riesgo de mercado: es el riesgo de una posible perdida dentro de un plazo en específico en el valor de un instrumento o portafolio financiero producto de variación

significativa en las variables de mercado. Un ejemplo de esto es la inflación, tasas de interés, tasas de cambios, entre otros.

- Riesgo de liquidez: Según el Banco Central Europeo, el concepto de liquidez se define como la facilidad que un activo financiero se pueda convertir en dinero en efectivo para cancelar un pasivo. Por lo tanto, el riesgo de liquidez se define como la dificultad de las entidades financieras para solventar un aumento en los activos cuando la institución es capaz de cumplir con sus obligaciones.

- Riesgo crediticio: Es la probabilidad que un cliente comience a estar en mora, es decir, que el solicitante del crédito no cumpla con sus obligaciones en los plazos estipulados, provocando una perdida para la institución bancaria. Es por este tipo de riesgos que se crearon métodos estadísticos capaces de predecir con datos históricos la capacidad de pago futuro de los clientes solicitantes de créditos. Para cuantificar este tipo de riesgo se utiliza generalmente la regresión logística binaria, dado que se espera estima/predecir (en caso de un modelo de admisión) si un cliente va a hacer la devolución del crédito solicitado.

2.2. Definición del problema

Esta etapa está orientada a resolver temas de fuentes de datos, lógica del negocio y metodología a implementar. Esta es una de las fases más relevantes, dado que son las definiciones iniciales que debe tener todo proyecto tecnológico.

- Metodología: Se revisan aspectos tales como: cambios normativos que conlleven a un cambio en la metodología, problemas con los datos que necesiten un filtro de exclusión, definición de las ventanas de observación.

- Fuentes de datos: Se enfocan principalmente en definir que tablas son de utilidad para construir el modelo, así como también la cantidad de registros que contienen estás tablas.

- Lógica del negocio: Reunión que sirve para definir si del punto de vista del negocio, el modelo aporta valor a la compañía. Ademas de estudiar la factibilidad de desarrollo del modelo.

2.3. Calidad de datos

Durante esta sección, se debe intentar recabar la información necesaria para generar una fuente de datos que tenga un cierto grado de calidad, para la construcción de un modelo estadístico robusto. La información contenida, debe ser consolidada como integra y además, tiene que reflejar el comportamiento del negocio estudiado.

2.3.1. Filtros de exclusión

Filtro de calidad de datos:

Basado en análisis descriptivos se pueden identificar registros anómalos que puedan impactar negativamente a la robustez del modelo. Si existen registros que no se pueden corregir a través de tratamientos, entonces, aquellos registros deben ser eliminados.

Filtro metodológico del modelado:

Este es un filtro, que se aplica dependiendo de el comportamiento que se quiere estudiar. En ocasiones, pueden haber clientes u operaciones financieras que no cumplen los requisitos mínimo para estudiarlos a través de un modelo estadístico.

Filtro de lógica del negocio:

Filtro utilizado para eliminar clientes, carteras, y operaciones que tengan un comportamiento muy distinto con respecto a los demás.

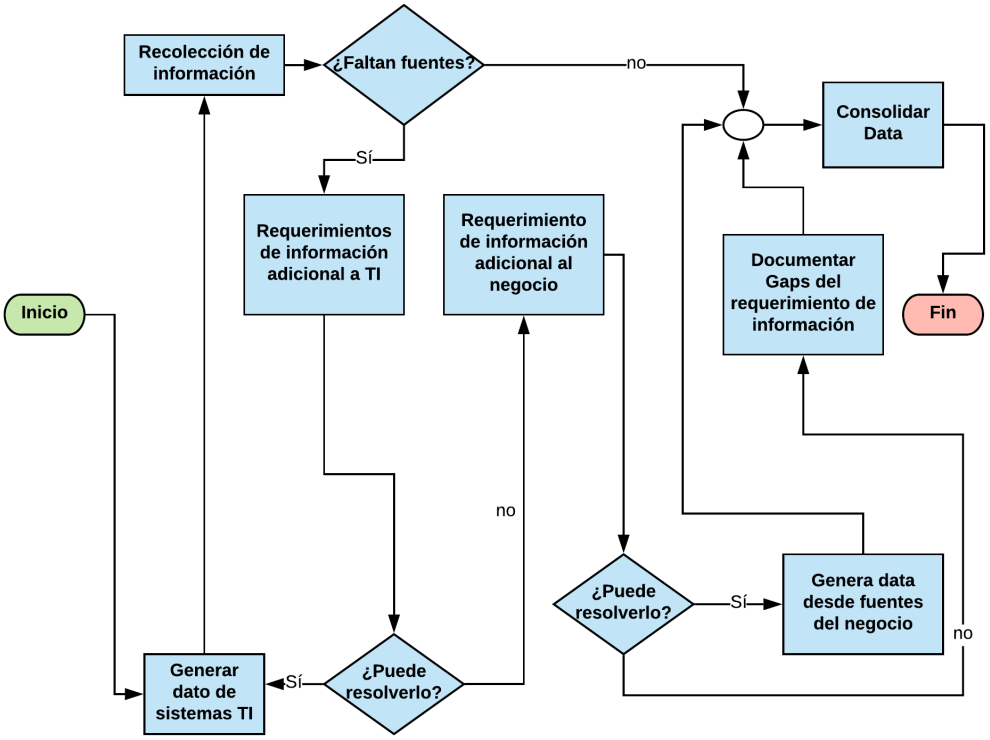

Figura 2.1: Diagrama que muestra como recopilar la información para construir el modelo
Fuente: Elaboración propia.

2.4. Muestreo

El banco a través de la herramienta llamada TeraData tiene una carga periódica de información en su bodega de datos. Dada la gran información que se maneja, no es viable hacer un desarrollo de un modelo con toda la información disponible, ya que para la generación de un modelo estadístico se pueden realizar cálculos muy complejos provocando lentitud en los procesos. Para resolver este problema existe una fase de muestreo, la cual busca encontrar un subconjunto de la información total que se tiene, de modo que se pueda construir el modelo con una muestra representativa, logrando extrapolar las conclusiones finales del modelo hacia la población. A continuación se presentan algunas de las metodologías que se utilizan actualmente en el banco para poder seleccionar la muestra más óptima y representativa posible.

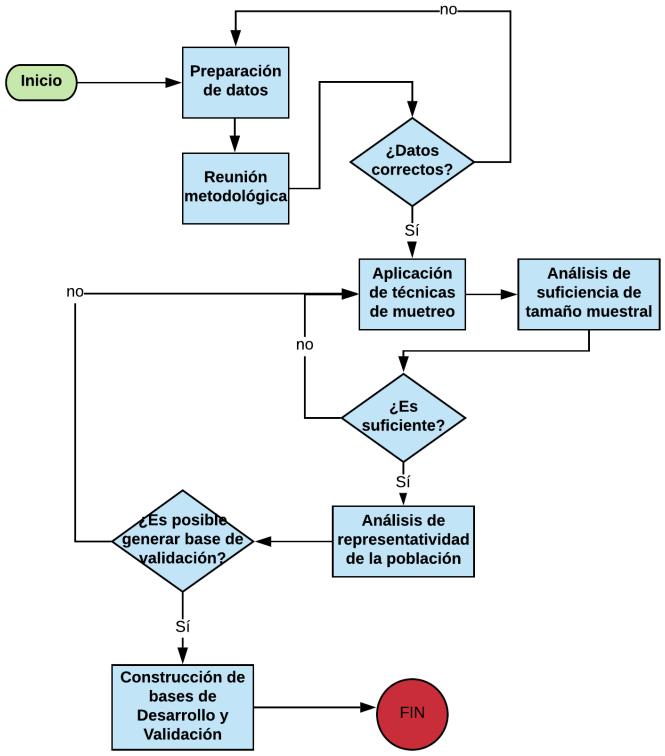

Figura 2.2: Diagrama que muestra las etapas del proceso de muestreo
Fuente: Elaboración propia.

2.4.1. Definición de la población objetivo

Lo primero que se tiene que resolver en un proceso de muestreo es definir la población objetivo. Durante esta fase se realizan filtros de exclusión sobre la cartera, con el objetivo de reducir la cantidad de clientes para obtener el grupo que se quiere estudiar.

Ejemplos de reducción de la cartera según modelo a construir:

- Modelo de probabilidad de incumplimiento (PI): Se filtra la cartera con clientes que tengan una mora inferior a los 90 días en el momento que son seleccionados como muestra.
- Modelo de probabilidad de perdida dado el incumplimiento (PDI): Se realiza un

filtro a la cartera para obtener a clientes con mora superior a 90 días, dado que se estudian clientes que ya están en incumplimiento.

2.4.2. Carteras pequeñas

Mediante un forzaje de aparición del cliente, se puede maximizar el uso de la información para carteras que tienen muy pocos clientes. Se considera sólo los meses en que los clientes aparecen, seleccionadolos en la muestra para el desarrollo.

2.4.3. Uso de ponderadores

Según el modelo estadístico a desarrollar, es posible aplicar distintos ponderadores sobre la muestra. Algunos de estos ponderadores se describen a continuación:

- Ponderador por aparición en la ventana de desarrollo: es el ratio entre el número de periodos que la unidad muestral se encuentra en la muestra y el total de periodos de la ventana de desarrollo.
$$p_i = \frac{M}{N}. \tag{2.1}$$

- Ponderador por aparición en la ventana de desempeño: corresponde al ratio entre el número de periodos que la unidad muestral se encuentra en la muestra y el total de periodos de la ventana de desempeño.
$$p_i = \frac{D}{K}. \tag{2.2}$$

- Ponderador por saldo IFRS: ratio entre la deuda de una operación y el total de deuda de un cliente para un periodo.
$$p_i = \frac{SALDOIFRS_i}{\sum_i^n (SALDOIFRS_i)}. \tag{2.3}$$

2.5. Transformación de Variables

2.5.1. Tramificación

Debido a que en etapas posteriores se realiza una transformación a las variables utilizando el peso de la evidencia, antes de este proceso se debe realizar una tramificación óptima a las variables continuas. Para esto, se realiza una tramificación a través de arboles de decisión CHAID. Es relevante destacar, que las categorías que presenten muy poca concentración de casos serán eliminadas dado que afectaría al poder predictivo de la variable. La metodología para tramificar variables continuas de presenta a continuación:

Árbol de decisión CHAID

El árbol de decisión CHAID es un algoritmo que permite crear segmentos en las variables de análisis. La técnica CHAID se genera a través del árbol AID y el test Chi Cuadrado con el objetivo de fundir clases. La metodología para construir este árbol de decisión se basa principalmente en tres pasos: fusión, división y detención (Bigss y Suen, 1991).

- El propósito de esta etapa es realizar una fusión a todas aquellas categorías que no sean estadísticamente significativas. Cada categoría considerada como diferente pasa a ser un nodo de división para el árbol que se entregará. Esta etapa en si consta de 9 pasos:
    - Si X tiene un sola categoría, entonces la etapa se acaba y se debe ajustar el valor-p a 1.
    - Si X tiene dos categorías, entonces ir al paso 5.
    - si X tiene más de dos categorías, entonces se debe encontrar el par de categorías que sean estadísticamente equivalentes (valor-p más grande).
    - Para el par de categorías que tenga el valor-p más grande se debe revisar si este supera el nivel asignado por el analista. Si esto ocurre, se deben fusionar las clases estudiadas.
    - El valor-p ajustado es calculado para las categorías fusionadas aplicando las correcciones de Bonferroni.

- División: la mejor división para cada predictor es encontrada en la etapa de Fusión. La etapa de división sirve para seleccionar al mejor predictor. La selección se basa en la comparación de los valores-p obtenidos en la etapa ocho de la Fusión.

    - Seleccionar el predictor que tenga mejor valor-p ajustado (es decir, el más significativo).
    - Si el valor-p ajustado es menor que el especificado por el usuario, entonces se divide un nodo usando este predictor. En caso contrario, no se divide el nodo y se considere como terminal.

- Detención: esta etapa verifica si el proceso de crecimiento del árbol debería parar de acuerdo a las siguientes reglas:

    - Si un nodo se convierte en puro, es decir, todos los casos son asignados a la misma clase de la variable respuesta el nodo no se divide.
    - Si todos los casos en un nodo tienen valores idénticos para cada predictor, el nodo no se divide.
    - Si la profundidad del árbol llega hasta lo especificado por el usuario, entonces el proceso de crecimiento termina.
    - Si el tamaño de un nodo resulta en un nodo hijo, se analiza su tamaño y se compara con el mínimo establecido por el especialista. Todos aquellos nodos que tengan tamaños menores al establecido por el usuario se fundirán con el nodo hijo más similar (en base al valor-p). Sin embargo, si el número de nodos hijos resultantes es 1, el nodo no se divide.

Test de hipótesis:

Para la etapa de fusión, el algoritmo CHAID necesita del valor-p de un par de categorías de X y a veces requiere de todas las categorías de X. Sea D el conjunto de datos donde se aplicarán los test, y supongamos que existen I categorías de X y J categorías de Y (cuando Y es categórica). El cálculo del valor-p usando los datos en D dependerá si la variable dependiente es continua, ordinal u nominal. Para el caso de una variable continua

se debe utilizar el siguiente estadístico de prueba:

$$F = \frac{\sum_{i=1}^{I} \sum_{n \in D} w_n f_n I(x_n = i)(\bar{y_n} - \bar{y})^2 (N_f - I)}{\sum_{i=1}^{I} \sum_{n \in D} w_n f_n I(x_n = i)(y_n - \bar{y_i})^2 (I-1)} \tag{2.4}$$

De donde se obtiene el valor-p mediante la siguiente ecuación:

$$p = P(F(I-1, N_f - I) > F). \tag{2.5}$$

donde,

$$\bar{y_i} = \frac{\sum_{n \in D} w_n f_n y_n I(x_n = i)}{\sum_{n \in D} w_n f_n I(x_n = i)} \quad , \quad \bar{y} = \frac{\sum_{n \in D} w_n f_n y_n}{\sum_{n \in D} w_n f_n} \quad , \quad N_f = \sum_{n \in D} f_n, \tag{2.6}$$

y $F(I-1, N_f - I))$ es una variable aleatoria que sigue una distribución F de Fisher con grados de libertad I y $N_f - I$, correspondiente al numerador y denominador respectivamente.

Si la variable dependiente es nominal, las pruebas cambian ya que lo que se prueba es la independencia de las variables X e Y. Para esto se crea la tabla de contingencia entre las clases de la variable Y y las clases de la variable X y si existen pesos se deben realizar las estimaciones apropiadas a los elementos de las intersecciones. Lo anterior se muestra a continuación:

$$X^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - \widehat{m}_{ij})^2}{\widehat{m}_{ij}}. \tag{2.7}$$

$$G^2 = 2 \sum_{j}^{J} \sum_{i=1}^{I} n_{ij} ln \left(\frac{n_{ij}}{\widehat{m}_{ij}} \right). \tag{2.8}$$

donde $n_{ij} = \sum_{n \in D} f_n I(x_n = i \wedge y_n = j)$ es la frecuencia observada y $\widehat{m}_{ij}$ es la frecuencia esperada para $(x_n = i, y_n = j)$. El correspondiente valor-p se obtiene de $p = P(X_d^2 > X^2)$ para la prueba Chi-cuadrado de Pearson o de $p = P(X_d^2 > G^2)$ para la prueba de verosimilitud de proporciones, donde X_d^2 sigue una distribución Chi-cuadrado con $d = (J-1)(I-1)$

grados de libertad.

Corrección de Bonferroni

El valor-p ajustado se calcula como el valor-p multiplicado por un multiplicador de Bonferroni. El multiplicador de Bonferroni se ajusta para múltiples pruebas.

Si se supone que una variable independiente (predictora) tiene I categorías, y estas se reducen a r categorías luego del la etapa de fusión. El multiplicador de Bonferroni B es el número de formas posibles que las I categorías se pueden combinar en r categorías. Para $r = I, B = 1$. Para $2 \leq r < I$, se utiliza la siguiente ecuación:

$$B = \begin{cases} \begin{pmatrix} I-1 \\ r-1 \end{pmatrix} & \text{Predictor Ordinal} \\ \sum_{v=0}^{r-1} (-1)^v \dfrac{(r-v)^I}{v!(r-v)!} & \text{Nominal Ordinal} \\ \begin{pmatrix} I-2 \\ r-2 \end{pmatrix} + r \begin{pmatrix} I-2 \\ r-1 \end{pmatrix} & \text{Predictor Ordinal} \end{cases} . \tag{2.9}$$

2.5.2. Peso de la evidencia

La fábrica en conjunto con el área de validación, realizan una transformación a las variables utilizando el peso de la evidencia, *WoE* (del inglés *weight of evidence*), discretizando todas las posibles variables a incluir en el modelo. El propósito de esta transformación, es proporcionar herramientas flexibles para recodificar los valores de las variables (ya sean continuos o discretos) en categorías discretas, asignando a cada categoría un único valor *WoE*. Según Leung, et. al. (2008), afirman en su investigación que la recodificación de las variables predictivas usando la transformación *WoE*, es particularmente utilizado para su posterior modelizado usando regresión logística. Específicamente, la regresión logística se ajustará a una ecuación de regresión lineal de predictores (*WoE* variables recodificadas), para predecir los valores binarios de la variable dependiente Y. Por lo tanto, al usar variables predictivas recodificadas por su peso de evidencia en la regresión logística, se está

preparando y codificando en la misma escala a las variables independientes del modelo. La transformación se puede observar en la ecuación 2:

$$\text{WoE} = ln\left(\frac{B_i}{M_i}\right). \tag{2.10}$$

El valor de WoE será 0 si B_i/M_i es igual a 1. Si la frecuencia relativa de clientes morosos en un grupo es mayor que la frecuencia relativa de clientes buenos, la razón de probabilidades será menor que 1 y el valor WoE será un número negativo. Además, Si la frecuencia relativa de los clientes no morosos es mayor que la frecuencia relativa de los clientes malos en un grupo, el valor WoE será un número positivo.

2.6. Métodos para reducir la dimensión

En el contexto de la modelización estadística, la selección de variables tiene por objetivo elegir las mejores variables candidatas a explicar los patrones presentes en el conjunto de datos a evaluar en los modelos. Esta selección, se hace cargo de un problema de eficiencia, el cual busca minimizar la cantidad de variables o atributos a estudiar, maximizando la información útil que éstos contienen (Anderson, 1958).

Las técnicas posibles de aplicar pueden ser de diversa índole, como por ejemplo, filtros asociados a algún indice de desempeño, e incluso indicadores asociados a la calidad de datos. Para efectos de este estudio, se abordarán técnicas de reducción de dimensionalidad asociadas a técnicas estadísticas, con su respectivo fundamento teórico. A continuación, se presentan los filtros más utilizados por el banco de crédito e inversiones para reducir la dimensionalidad de las variables candidatas a ingresar al modelo.

2.6.1. Análisis de componentes principales

El análisis de componentes principales es un método estadístico multivariante de simplificación o reducción de la dimensión de una tabla de variables con datos cuantitativos, para obtener otra de menor número de variables mediante una combinación lineal de las variables iniciales, que de denominan componentes principales o factores, cuya posterior

interpretación permitirá un análisis más simple del problema estudiado.

Este método, permite describir la estructura y las intercorrelaciones de las variables originales en el fenómeno que se estudia a partir de las componentes obtenidas que naturalmente habrá que interpretar (Smith, 2002). El mayor número posible de componentes coincide con el número total de variables. Quedarse con todas ellas no simplificará el problema, por lo que el investigador deberá seleccionar entre distintas alternativas que, siendo pocas e interpretables, expliquen una proporción aceptable de la varianza global e inercia de la nube de puntos que suponga una razonable pérdida de información. Smith en el año 2002, expone que una de las principales justificaciones para utilizar el análisis de componentes principales en un conjunto de datos es que las variables utilizadas esten correlacionadas, de lo contrario no tiene sentido utilizar este método.

Obtención de los componentes principales:

En el análisis de componentes principales se dispone de una muestra de tamaño n a p variables $X_1, X_2, ..., X_p$ (expresada en desviaciones respecto a su media) inicialmente correlacionadas, para posteriormente obtener a partir de ellas un número $k \leq p$ de variables intercorrelacionadas $Z_1, Z_2, ..., Z_p$ que sean combinación lineal de las variables iniciales y que expliquen la mayor parte de su variabilidad.

La primera componente principal, al igual que las restantes, se expresan como una combinación lineal de las variables originales como sigue:

$$Z_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + ... + u_{1p}X_{pi}. \tag{2.11}$$

Para el conjunto de las n observaciones muestrales esta ecuación puede expresarse matricialmente de la siguiente forma:

$$\begin{pmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1n} \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{13} & X_{2n} & X_{pn} & X_{pn} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{pmatrix}. \tag{2.12}$$

Se demuestra que al componente principal h-ésima se define como $Z_h = X_{uh}$, donde u_h

es el vector propio de V asociado a su h-ésimo mayor valor propio. Suele denominarse también a u_h eje factorial h-ésimo.

En el proceso de obtención de las componentes se constata que la varianza de la componente h-ésima es:

$$V(Z_h) = u_h V_{uh} = \lambda_k. \tag{2.13}$$

Es decir, la varianza de cada componente es igual al valor propio de la matriz V al que va asociada, en consecuencia, la medida de la variabilidad de las variables originales es la suma de sus varianzas:

$$\sum_{h=1}^{p} V(X_h) = traza(V). \tag{2.14}$$

En general, el objetivo de la aplicación de las componentes principales es reducir las dimensiones de las variables originales, pasando de p variables originales a $m < p$ componentes principales. El problema que se plantea es como fijar m, o dicho de otra forma, ¿qué número de componentes principales se debe retener? Aunque para la extracción de las componentes principales no hace falta plantear un modelo estadístico previo, algunos de los criterios para determinar cual debe ser el número óptimo de componentes a retener requieren la formulación previa de hipótesis estadísticas. Según el criterio de la media aritmética se seleccionan aquellas componentes cuya raíz característica λ_j excede la media de las raíces características. Cabe recordar que la raíz característica asociada a una componente es precisamente su varianza. Analíticamente este criterio implica retener todas aquellas componentes en que se verifique que:

$$\lambda_j > \bar{\lambda} = \frac{\sum_{j=1}^{p} \lambda_j}{p}. \tag{2.15}$$

Correlación de Pearson:

El coeficiente de correlación de Pearson, pensado para variables cuantitativas (escala mínima de intervalo), es un indice que mide el grado de covariación entre distintas variables relacionadas linealmente que poseen una distribución normal bivariada conjunta. Esto significa que puede haber variables fuertemente relacionadas, per no de forma lineal, en cuyo caso se sugiere no proceder a aplicarse la correlación de Pearson.

Dado el punto anterior, el ajuste será de la forma: $Y = a + bX + \varepsilon$ (recta de regresión de Y sobre X) o $X = c + dY + \varepsilon$ (recta de regresión de X sobre Y), donde:

$$a = \bar{y} - \frac{\bar{x}Sxy}{Sx^2}$$

$$b = \frac{Sxy}{Sx^2}$$

$$c = \bar{x} - \frac{\bar{y}Sxy}{Sy^2}$$

$$d = \frac{Sxy}{Sy^2}.$$

A los parámetros a y b, se les denomina coeficientes de regresión de Y sobre X, y a los parámetros de c y d se les llama coeficientes de regresión de X sobre Y. También se pueden expresar las rectas de regresión de Y sobre X y X sobre Y de la forma:

$$y - \bar{y} = \frac{(x - \bar{x})Sxy}{Sx^2} \quad \text{Y} \quad x - \bar{x} = \frac{(y - \bar{y})Sxy}{Sy^2}. \tag{2.16}$$

La expresión del coeficiente de correlación lineal de Pearson entre las variables X eY viene dad por la expresión:

$$r = \frac{S_{xy}}{S_x S_y}. \tag{2.17}$$

El coeficiente puede ser interpretado de la siguiente forma:

- Si $r = 1$ existe correlación perfecta positiva y la relación funcional entre ambas variables es exacta y positva, variando ambas variables en el mismo sentido.
- Si $r = -1$ existe correlación perfecta y negativa y la relación funcional entre ambas variables es exacta y negativa, variando ambas variables en el sentido opuesto.
- Si $r = 0$ la correlación es nula y las variables no estan asociadas, siendo imposible encontrar una relación funcional entre ellas.
- Si $0 < r < 1$ la correlación es positiva, pero el grado de asociación entre las dos variables será mayor a medida que r se acerca más a 1, y será menor a medida que r se acerca más a cero.

- Si $-1 < r < 0$ la correlación es negativa, pero el grado de asociación entre las dos variables será mayor a medida que r se acerca más a -1, y será menor a medida que r se acerca más a cero.

2.6.2. Correlación Rho de Spearman

La correlación Rho de Spearman, es la verisón no paramétrica del coeficiente de correlación de Pearson, que se basa en los rangos de los datos en lugar de hacerlo en los valores reales (Morales y Rodriguez, 2016). Esta técnica resulta apropiada para datos ordinales, o los de intervalo que no satisfagan el supuesto de normalidad. ρ mide la tendencia de X, Y a relacionarse en forma monótona creciente o decreciente. Al medir el grado de asociación de forma monótona entre las variables X y Y, ρ no se encuentra restringido a descubrir sólo una asociación lineal entre las variables.

El coeficiente de correlación por rangos también se utiliza para variables cuantitativas, con la aclaración de que el grado de asociación obtenido no es el de los valores de las variables, sino el de las clasificaciones por rangos de dichos valores. La expresión de este coeficiente viene dada por:

$$\rho = 1 - \frac{6\sum_i d_i^2}{N^3 - N}. \tag{2.18}$$

siendo $d_i = x_i - y_i$.

2.6.3. Correlación Tau-b de Kendall

Es una medida no paramétrica de asociación para variables ordinales o de rangos que tiene en consideración los empates (Morales y Rodriguez, 2016). El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la magnitud de la misma, de tal modo que los mayores valores absolutos indican relaciones más fuertes. Los valores posibles van de -1 a 1, pero un valor de -1 o +1 sólo se pude obtener a partir de tablas cuadradas.

Se define que dos observaciones (X_i, Y_i) y (X_j, Y_j) son concordantes, si están en el mismo orden con respecto a cada variable. Es decir, si $X_i < X_j$ y $Y_i < Y_j$, o si $X_i > X_j$ y $Y_i > Y_j$.

Por otra parte, se define que son disonantes, si están en el orden inverso de X e Y, o los valores están dispuestos en direcciones opuestas. Es decir si $X_i < X_j$ y $Y_i > Y_j$, o si $X_i > X_j$ y $Y_i < Y_j$.

También se definen empatadas si, $X_i = X_j$ o $Y_i = Y_j$. El número total de pares que se puede construir para un tamaño de muestra n es:

$$N = \binom{n}{2} = \frac{1}{2}n(n-1). \tag{2.19}$$

N puede ser descompuesto en cinco cantidades: $N = P + Q + X_0 + Y_0 + (XY)_0$.

Donde:
P: Número de pares concordantes.
Q: Número de pares disonantes.
X_0: Número de pares empatados sólo con la variable X.
Y_0: Número de pares empatados sólo con la variable Y.
$(XY)_0$: Número de pares empatados tanto en X e Y.

Índice de estabilidad poblacional:

El índice de estabilidad de la población (IEP) es un número positivo que indica en una escala continua el grado de similaridad o disimilaridad entre dos poblaciones diferentes o para una misma población, pero observadas en dos momentos distintos en el tiempo (Gadidov y McBurnett, 2015). Para su cálculo hay que proceder a agrupar la variable de interés en categorías (rangos para el caso de variables continuas o segmentos para el caso de variables discretas o categóricas). Su fómula se muestra en lo que sigue:

$$IEP = \sum_{i=1}^{n} DM_i = \sum_{i=1}^{n} Ln\left[\frac{D_i}{R_i}\right](D_i - R_i). \tag{2.20}$$

Donde,
DM_i es la divergencia marginal que corresponde a la comparación relativa para la cate-

goría i entre la muestra de desarrollo y la reciente.
R_i es la frecuencia relativa o porcentual de la categoría i, para el caso de la muestra reciente.
D_i es la frecuencia relativa o porcentual de la categoría i, para la muestra de referencia o de desarrollo.

2.6.4. Valor de la información (V.I.)

El Número de Información (del inglés *Information Value*), es una medida que indica la fuerza predictiva de una variable, y es un indicador que se deriva de la teoría de la información (Lund y Brotherton, 2013). La medida de divergencia de Kullback (también conocida como número de información, ganancia de la información o entropía relativa) es una medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad $B(x)$ (distribución de clientes buenos) y $M(x)$ (distribución de clientes malos). La divergencia de Kullback es una seudométrica o seudo distancia, dado que no es simétrica: La divergencia de Kullback de $B(x)$ a $M(x)$ no necesariamente es la misma divergencia de Kullback de $M(x)$ a $B(x)$.

Para el caso discreto, el cual es del interés de este estudio, bien puede definirse la distribución de Buenos clientes $(B(x))$ y de malos clientes $(M(x))$ usando los porcentajes de las respectivas poblaciones de buenos o malos. En particular, la medida de divergencia de Kullback puede ser definida en función de las frecuencias relativas como sigue:

$$IV = \sum_{i=1}^{n} (B_i - M_i) Ln\left(\frac{B_i}{M_i}\right). \tag{2.21}$$

2.6.5. Análisis de correspondencias

El análisis de correspondencias al igual que el de componentes principales, tiene por objetivo intentar reducir la cantidad de dimensiones de modo que se pueda resumir una gran cantidad de datos en un subconjunto más pequeño, intentando perder la menor cantidad de información posible (de la Fuente, 2011). A diferencia del análisis de componentes principales (que se utiliza en variables continuas) el análisis de correspondencias se aplica

en variables categóricas u ordinales.

Para conocer si existe alguna relación entre X e Y se utilizan pruebas de hipótesis sobre la independencia de estas variables. La prueba que generalmente se utiliza para determinar independencia entre las variables es la prueba de Chi-cuadrado de Pearson que se mencionó en anteriores métodos.

En lo que se basa esta prueba es en la comparación de los perfiles fila y columna con los perfiles marginales correspondientes. Si H_0 es verdadero (el valor-p no es suficientemente significativo), entonces todos los perfiles fila son iguales entre sí e iguales al perfil marginal de X con respecto a Y.
El estadístico que se pone a prueba se detalla a continuación:

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = X^2_{(k-1);(m-1)}. \tag{2.22}$$

donde, $e_{ij} = E[n_{ij}/H_0 \text{ es verdadera}] = \frac{N_{i\bullet}N_{\bullet j}}{N_{\bullet\bullet}}$.

Descomposición en valores singulares

La descomposición en valores singulares de la matriz sirve para construir un sistema basado en coordenadas relacionados a las filas y columnas de la tabla de contingencia definida anteriormente.

$$C = (r_{ij}). \tag{2.23}$$

donde,

$$r_{ij} = \frac{n_{ij}-e_{ij}}{\sqrt{e_{ij}}}. \tag{2.24}$$

Las distancias Chi-cuadrado entre perfiles son distancias pitagóricas ponderadas, en donde se calculan por lo que sigue:

Distancias entre perfiles filas

$$d_{ij} = \sum_{h=1}^{m}\frac{1}{N_{\bullet h}}\left(\frac{n_{ih}}{N_{i\bullet}} - \frac{n_{jh}}{N_{j\bullet}}\right)^2. \tag{2.25}$$

Distancias entre perfiles columnas

$$d_{ij}^{c} = \sum_{h=1}^{k} \frac{1}{N_{h\bullet}} \left(\frac{n_{hi}}{N_{\bullet i}} - \frac{n_{hj}}{N_{\bullet j}} \right)^2 . \tag{2.26}$$

2.7. Estimación de parámetros

La administración de riesgos a través de modelos de predicción, es fundamental para calcular el capital regulatorio de los bancos, llamado comúnmente en el rubro financiero: provisiones, que según el Banco Bilbao Vizcaya Argentaria (BBVA) se define como *"los fondos necesarios para cubrir tanto posibles pérdidas de valor del activo como para hacer frente a potenciales obligaciones que aún no se han materializado."*.

Dado que los modelos se encuentran estrictamente normados en Chile por la Súper Intendencia de Bancos e Instituciones Financiera (SBIF), se debe contar con un sistema interpretable y explicable a través de todos su componentes de estimación, por lo tanto, como metodología estándar los bancos en Chile ocupan la regresión logística para identificar el comportamiento futuro de los clientes (y no clientes), ya que mediante este método, se puede construir una tarjeta de puntuación, la cual se basa en constituir un puntaje para cada cliente según los parámetros que se presentan a continuación:

En un modelo logístico, la probabilidad de que un cliente sea bueno, es decir $y = 1$, viene dado por la siguiente expresión:

$$P(y = 0) = \frac{1}{1 + e^{L_x}}. \tag{2.27}$$

donde L_x corresponde a la multiplicación de las variables con sus respectivos coeficientes estimados por la regresión. Además, se sabe que la probabilidad de que un cliente posea un desempeño bueno se puede expresar de la siguiente manera:

$$P(y = 0) = \frac{Bueno}{Bueno + Malo}. \tag{2.28}$$

donde, Bueno: frecuencia relativa de clientes buenos y Malo, la frecuencia relativa de clientes malos. Por lo tanto, si se igualan las dos ecuaciones anteriores se puede obtener

lo que sigue:

$$\frac{Bueno}{Bueno+Malo} = \frac{1}{1+e^{L_x}}. \tag{2.29}$$

Luego si se despega la ecuación anterior, se obtiene:

$$1+e^{L_x} = \frac{Bueno+Malo}{Bueno} - 1. \tag{2.30}$$

$$1+e^{L_x} = \frac{Bueno+Malo-Bueno}{Malo}. \tag{2.31}$$

$$L_x = -Ln\left(\frac{Malo}{Bueno}\right). \tag{2.32}$$

$$L_x = Ln\left(\frac{Bueno}{Malo}\right) = Ln(Odds). \tag{2.33}$$

El objetivo de esta estimación mediante regresión logística es encontrar el puntaje denominado Score, de manera que la relación lineal se mantenga, es decir:

$$Ln(Odds) = \alpha + (\beta)Score. \tag{2.34}$$

El modelo anterior puede ser definido mediante dos parámetros conocidos:

- P_0 definido como el puntaje en el cual la relación de Odds es $\gamma : 1$
- PDO definido como los puntos necesarios para doblar los Odds.

dado los parámetros conocidos anteriores se tiene que:

$$Ln(\gamma) = \alpha + \beta(P_0) \qquad , \qquad Ln(2\gamma) = \alpha + \beta(P_0 + PDO). \tag{2.35}$$

y desarrollando las ecuaciones anteriores se pueden encontrar α y β como sigue:

$$\alpha = Ln(\gamma) - \beta(P_0) = Ln(\gamma) - \frac{Ln(2\gamma)P_o}{PDO} \qquad , \qquad \beta = \frac{Ln(2\gamma)}{PDO}. \tag{2.36}$$

también, se sabe que $Ln(Odds) = L_x$, y además que $Ln(Odds) = \alpha + \beta(Score)$, por lo tanto:

$$L_x = \alpha + \beta(Score) \qquad , \qquad Score = \frac{1}{\beta}L_x - \frac{\alpha}{\beta}. \tag{2.37}$$

$$Score = Factor(L_x) + Offset. \tag{2.38}$$

donde,

$$Factor = \frac{PDO}{Ln(2\gamma)} \qquad , \qquad Offset = P_0 - Factor \cdot Ln(\gamma). \tag{2.39}$$

de esta forma se obtiene una calibración del score en base a los parámetros conocidos P_0 y PDO, lo que resta ahora es aplicar estos valores para estimar la transformación lineal del puntaje.

2.7.1. Regresión Logística

La regresión logística, es un modelo estadístico de regresión, el cual posee una variable dependiente categórica (Rencher, 2008). Esta variable categórica, puede ser dicotómica en el caso de la regresión logística binaria o multinomial. La regresión logística estima la variable respuesta a través de una probabilidad de ocurrencia, la cual se ajusta mediante la función logística que se muestra a continuación:

$$p_i = E(y_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}. \tag{2.40}$$

Por sus características, los modelos de regresión logística pueden ser utilizados para los siguientes objetivos (Rencher, 2008):

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente.
- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables (variables explicativas).

Supuestros de la Regresión Logística

La regresión logística no asume los mismos supuestos de la regresión lineal, particularmente los supuestos de normalidad, linealidad y varianza constante en sus residuos (Mood

y Graybill, 1974). Este tipo de regresión puede manejar cualquier tipo de relación (no necesariamente lineal), ya que aplica una transformación logarítmica no lineal. Las variables explicativas pueden ser continuas o discretas (categóricas u ordinales) y no necesitan ser independientes pero de serlo, el modelo tiene mayor robustez. En tanto, las covariables de naturaleza discreta pueden adoptar un número limitado de categorías. También, se debe tener especial consideración en que la relación entre la variable independiente y la probabilidad del suceso no cambie de sentido, ya que en ese caso el modelo logístico deja de tener la interpretabilidad deseada.

Capítulo 3

Resultados

3.1. Ranking por el Valor de la Información

Para proceder a la etapa de reducción de dimensiones y posteriormente estimar los parámetros para construir el modelo, se debe primeramente calcular el valor de la información y su valor *WoE* para cada variable presente en el estudio. Esto se realiza, ya que para utilizar como método para reducir dimensiones la correlación (en este caso la correlación de Kendall), se debe antes hacer un ranking por IV para determinar que variable correlacionada se elimina del modelo. Algunas variables con su respectivo valor de la información y valor *WoE* se muestran a continuación.

Peso de le evidencia para la Variable 1 analizada

Tramos	WOE	IV
(-1,19;0]	-12,00	82,49
(0;1.19e+03]	37,59	258,39

Cuadro 3.1: Valor de la información y *WoE* para la la variable 1.

Peso de la evidencia para la Variable 2 analizada

Tramos	WOE	IV
0	-0,25	0,06
1	22,19	5,46

Cuadro 3.2: Valor de la información y *WoE* para la la variable 2.

Peso de la evidencia para la Variable 3 analizada

Tramos	WOE	IV
(-2,41;0]	-12,29	85,32
(0;2,41e+03]	36,36	252,41

Cuadro 3.3: Valor de la información y *WoE* para la la variable 3.

Peso de la evidencia para la Variable 4 analizada

Tramos	WOE	IV
(-0,001;0,682]	-17,95	30,98
(0,682;0,806]	-16,64	27,79
(0,806;0,875]	-12,10	17,32
(0,875;0,933]	-4,27	3,03
(0,933;1]	14,09	77,99

Cuadro 3.4: Valor de la información y *WoE* para la la variable 4.

De las tablas anteriores, se puede concluir que no todas las variables tienen las mismas cantidades de tramos. Esto se debe a que las categorías que presenten muy baja concentración de información se eliminan de la variable. Además, se observa que de las variables presentes en las tablas, las variables 1 y 3 son las que presentan mayor valor de la información, por lo que de estar correlacionadas - por ejemplo con la variable 2- se eliminaría la que tenga menor VI, dado que la variable que tenga un mayor IV posee mayor poder predictivo para el modelo de riesgo.

3.2. Indice de Estabilidad Poblacional (IEP)

Como se mencionó en el Capitulo 3 de este trabajo de título, uno de los primeros filtros para reducir la cantidad de variables para el modelo de riesgo, es el filtro de estabilidad poblacional, el cual compara dos poblaciones distintas, o bien una población pero observadas en tiempos distintos. Cabe destacar que este indicador se debe calcular despues de realizar la tramificación de variables continuas por arboles de decisión y posterior a la transformación de variables utilizando sus pesos de la evidencia.

El criterio de exclusión de variables según su valor se muestra en la Tabla que sigue:

Tramos	Interpretación
Hasta 10 %	Resultado Satisfactorio
Sobre 10 % y hasta 25 %	Zona de advertencia
Sobre 10 %	Zona de riesgo

Cuadro 3.5: Criterios para eliminación de variables según su indicador de estabilidad poblacional.

Los resultados del indice de estabilidad poblacional alguna de las variables estudiadas se muestra a continuación:

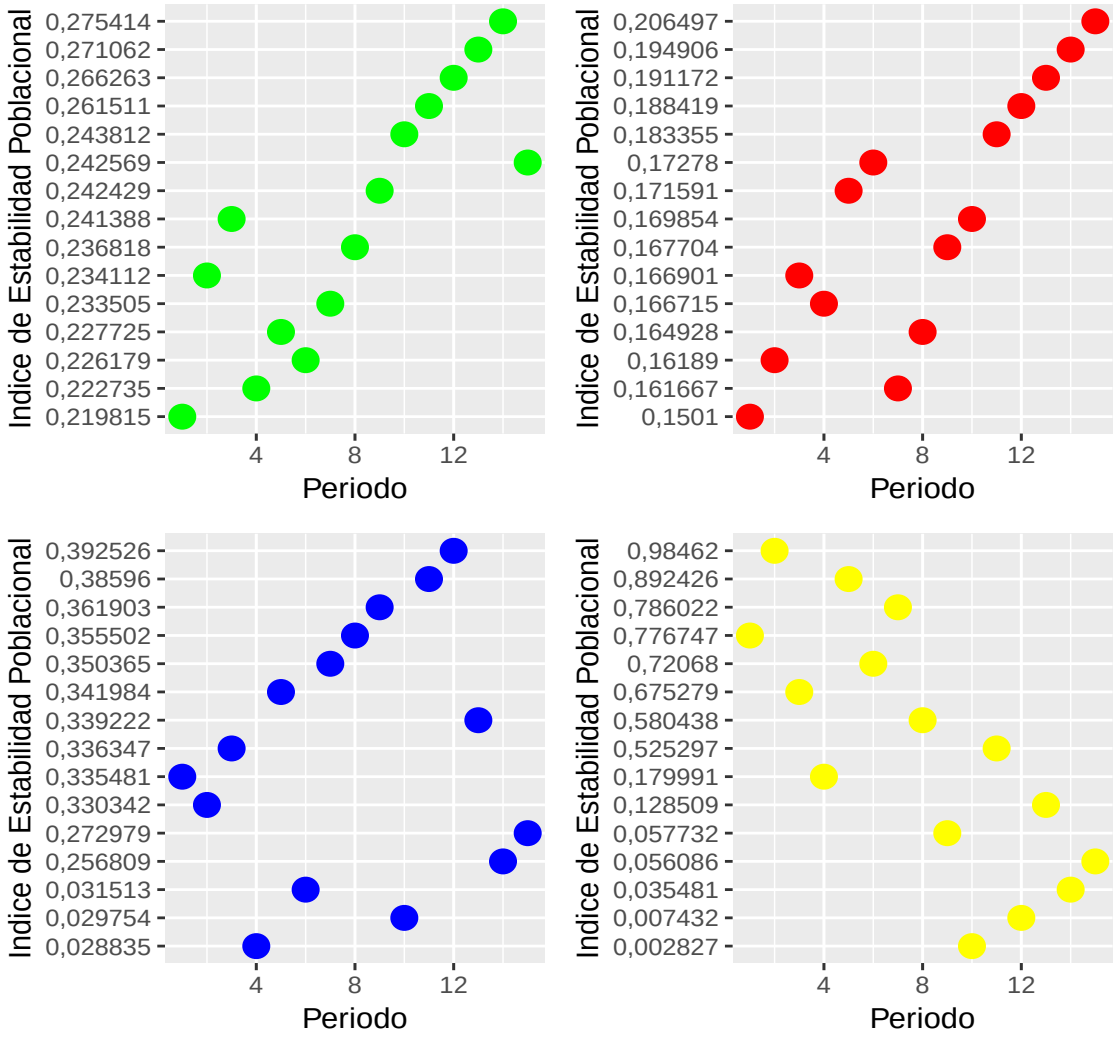

Figura 3.1: Indice de estabilidad poblacional de un estracto de sólo 4 variables en 14 periodos analizados.

Como se puede apreciar en la Figura anterior, se tiene un total de 15 periodos analizados obteniendo para cada uno de ellos un IEP. Las variables presentadas en el diagrama presentan una alta inestabilidad poblacional, por lo que se elimina todas aquellas que no cumplan con un IEP satisfactorio (según la Tabla 1 de este capitulo).

3.3. Filtro por Correlación de Kendall

Según lo profundizado en la etapa metodológica, se decidió no usar la correlación de Pearson dada la naturaleza teórica de las variables tramificadas y transformadas a través de su peso de la evidencia (*WoE*). Como método alternativo, se propone utilizar un filtro por correlación de Kendall, ya que este tipo de correlación está orientada para ser utilizada en variables discretas (que en el caso de este estudio sería lo metodológicamente correcto, dada la discretización de variables en instancias previas).

A continuación se presenta la matriz de correlación resultante posterior a la aplicación del método:

Variables	V1	V2	V3	V4	V5
V1	1,00	0,83	0,13	0,68	0,88
V2	0,83	1,00	0,15	0,61	0,78
V3	0,13	0,15	1,00	1,14	0,15
V4	0,68	0,61	0,14	1,00	0,82
V5	0,88	0,78	0,15	0,82	1,00

Cuadro 3.6: Extracto de la matriz de correlación de Kendall resultante.

Según lo expuesto en el Capitulo Metodológico de este trabajo, la correlación de Pearson es más estricta que la de Kendall al momento de calcular los coeficientes de correlación. Para mitigar el riesgo de ignorar una correlación alta dada la severidad moderada de la correlación de Kendall, se opta por eliminar todas aquellas variables que presenten un coeficiente de correlación superior al 80 %.

3.4. Selección de variables por análisis de correspondencias

Ya que el análisis de componentes principales es un método para reducir dimensiones que considera sólo variables continuas (utiliza la correlación de Pearson para extraer los componentes principales de la matriz de información) se concluyó que el mejor método para realizar una última reducción de campos es el de análisis de correspondencias. La

gráfica una vez aplicado esté método esta dado por lo que sigue a continuación:

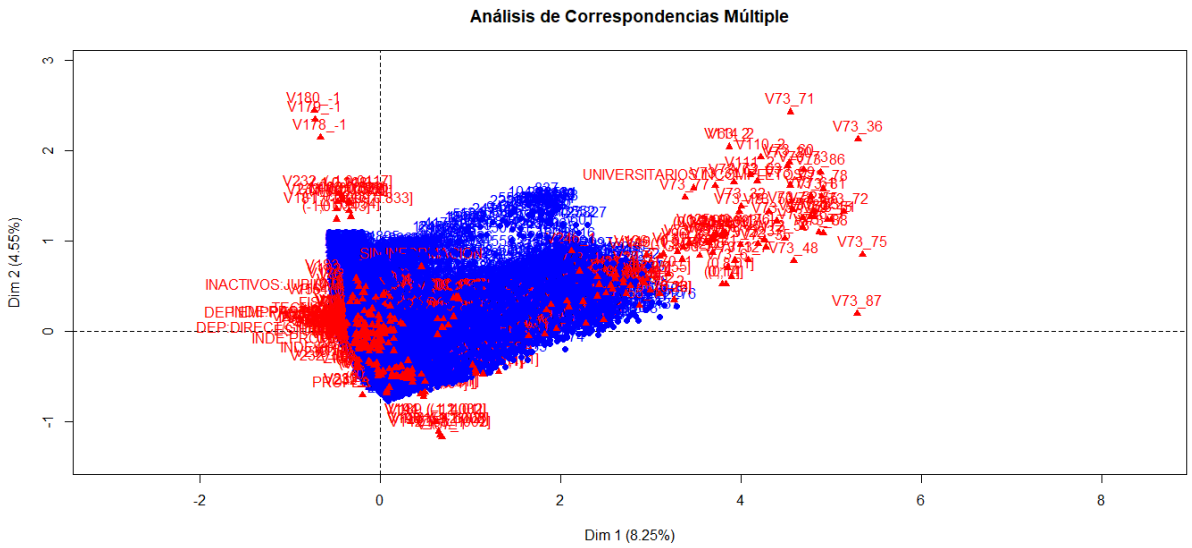

Figura 3.2: Análisis de Corerspondencias Múltiples sobre la matriz de información.

Ya que este método es sólo exploratorio, se verificó el valor del estadístico V (que sigue una distribución Gausseana), el cúal indica el siguiente criterio para observar la significancia de la variable en estudio: $V-test >= 1{,}96 \quad o, \quad V-test <= -1{,}96$.

3.5. Lista final de variables

Para finalizar con este estudio, es importante recalcar que luego de los filtros por etapas a los que se fueron sometiendo las variables, finalmente quedaron 14 campos para la modelización. Sin embargo, para que los modelos no presenten inestabilidad a través del tiempo (por la gran cantidad de información presente en los datos que puede ir variando), se realizó un último filtro por experiencia del negocio. La lista corta de variables para el modelo final es la siguiente:

- VAR EXTDEM INDICADOR COLABORADOR BCI
- VARABONOREM N ABONOREM

- VARD00 FLAG RENEG 06
- VARD00 FLAG RENEG CON D00 06
- VARD00 MAX N OPERACIONES SOBREGIRO 03
- VAR D00 PEOR SIT TDC
- VAREXTHIP PEOR SIT HIP FMES 06
- VARPROTESTOS N PROTESTOS FR 03
- VARSBIF EVO COM SBIF12
- VARSBIF EVO CONSCOM SBIF06
- VARSBIF RATIO DEUDA COMPROMETIDA SBIF03
- VARSBIF RATIO VCDA12
- VARSBIF USO12
- VAR ASICOM LoanToValue ORIGEN Min

En la siguiente Tabla se presenta las variables con su respectiva estimación de parámetros como sigue:

Variable	β	Error estándar	Sig
VAR EXTDEM INDICADOR COLABORADOR BCI	-0,855	0,269	0,001
VARABONOREM N ABONOREM	-0,619	-0,084	0,000
VARD00 FLAG RENEG 06	-0,284	0,098	0,004
VARD00 MAX N OPERACIONES SOBREGIRO 03	-0,535	0,064	0,000
VAREXTHIP PEOR SIT HIP FMES 06	-0,785	0,019	0,000
VARPROTESTOS N PROTESTOS FR 03	-0,342	0,107	0,001
VARSBIF EVO CONSCOM SBIF06	-0,528	0,082	0,000
VARSBIF RATIO DEUDA COMPROMETIDA SBIF03	-0,339	0,047	0,000
VARSBIF RATIO VCDA12	-0,287	0,41	0,000
VARSBIF USO12	-0,260	0,044	0,000
VAR ASICOM LoanToValue ORIGEN Min	-0,513	0,140	0,000
Constante	-3,668	0,053	0,000

Cuadro 3.7: Modelo de regresión logística con su respectiva estimación de parámetros.

Como se aprecia en el cuadro anterior, todos los coeficientes de los parámetros son negativos y significativos. El signo que tiene cada coeficiente es relevante a la hora de interpretar el modelo, ya que este representa la relación de la variable con el no pago del crédito. Si el coeficiente es positivo, significa que la variable tiene una relación directamente proporcional con el no pago, y inversamente proporcional en el caso de un coeficiente negativo.

Capítulo 4

Conclusiones

Para finalizar con este trabajo de título, cabe destacar que se realizó un estudio asociado a las metodologías para reducir dimensiones y principalmente para construir modelos predictivos en la industria bancaria. Para esto, se tuvo que estudiar inicialmente la naturaleza teórica de las variables transformadas por su peso de la evidencia, concluyendo que estas presentan una naturaleza discreta. Por lo tanto, la metodología aplicada actualmente en el banco para reducir dimensiones no está siendo bien utilizada. Es por ello que en este trabajo de titulación se propuso alternativas que se agregaron a las ya utilizadas en el banco, como por ejemplo: análisis de correspondencias, correlación de Kendall y la eliminación del análisis de componentes principales como método de selección de variables. Esta última metodología se eliminó principalmente por la justificación de su uso, y básicamente sus restricciones en variables discretas. Finalmente, se pudo construir un modelo de admisión para determinar la probabilidad que tiene un cliente del Banco de Crédito e Inversiones para cubrir la cobertura total del crédito solicitado. El modelo estadístico resultante contiene todos sus parámetros negativos y significativos, por lo que se puede concluir que cada una de las variables escogidas en este trabajo de titulación aportan suficiente poder de discriminación y estabilidad predictiva al modelo (dado su valor de la información y estabilidad poblacional).

Bibliografía

Bigss, D., Ville, B., and Suen, E. (1991). *A Method of Choosing Multiway Partitions for Classification and Decision Trees*. Journal of Applied Statistics, 18, 1, 49-62.

Cox, D., Snell, E. (1989). *Analysis of Binary Data*. (Segunda Edición). Londres, Inglaterra: Chapman and Hall.

Freed, N., F. Glover (1981), Simple but powerful goal programming formulations for the discriminant problem, European J. Oper. Res., 7, 44-60.

Gadidov, B., McBurnett, B. (2015). Population Stability and Model Performance Metrics Replication for Business Model at SunTrust Bank. *SESUG*.

Gonzales, A. (2015). *Selección de variables: Una revisión de métodos existentes* (Tesis de maestría). Recuperado de http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_ 1263.pdf

Goodman, L. A. (1979). *Simple Models for the Analysis of Association in CrossClassifications Having Ordered Categories*. Journal of the American Statistical Association, 74, 537-552.

Johnson, R. y Wichern, D. (2002). *Multivariate analysis*. (Quinta edición). New Jersey, Estados Unidos: PRENTICE HALL.

Kass, G. V. (1980). *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Applied Statistics, 20, 2, 119-127.

Leung, K., Cheong, F., Cheong, C., O'Farrell, S., Tissington, R., y Ou, C. M. (2008). A comparison of variable selection techniques for credit scoring. *In Proceedings of*

the 7th International Conference on Computational Intelligence in Economics and Finance, Volúmen(4).

Leung, K., Cheong, F., Cheong, C., O'Farrell, S., y Tissington, R. (2008). *Building a scorecard in practice*. In Proceedings 7th International Conference on Computational Intelligence in Economics and Finance.

Lund, B., Brotherton, D. (2013). Information Value Statistic. *MWSUG 2013 Conference Proceedings*.

Mays, E. (2001). *Handbook of credit scoring*. Chicago, Estados Unidos: Global Professional Publishi.

Morales, P., Rodríguez, L. (2016). Application of the Kendall correlation and Spearman coefficients.

Mood, A. M., y Graybill, F. A. (6). Boes, DC (1974). *Introduction to the theory of statistics*. (Tercera edición). New York, Estados Unidos: McGraw and Hill Book Company.

Parrado, E. (2014). *Hacia una mejor gestión del riesgo*. Web Superintendencia de Bancos e Instituciones Financieras. Recuperado de https://www.sbif.cl/sbifweb3/internet/archivos/DISCURSOS_10548.pdf

Rencher, A. y Schaalje, G. (2008). *Linear models in statistics*.(Segunda Edición). New Jersey, Estados Unidos: John Wiley and Sons.

Reyes, P. (2017). *Las provisiones bancarias ¿qué son y cuántos tipos hay?* Web Banco BBVA. Recuperado de https://www.bbva.com/es/las-provisiones-bancarias-cuantos-tipos/.

Trejo, J., Martínez, M. y Venegas, F. (2017). *Administración del riesgo crediticio al menudeo en México: una mejora econométrica en la selección de variables y cambios en sus características*. Contaduría y administración. Recuperado de https://www.sciencedirect.com/science/article/pii/S0186104217300037.