UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
FINANZAS MATEMÁTICAS

Modelación del comportamiento de una cartera de crédito de prestamos personales

TESIS
QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias

PRESENTA:
Eric Martínez Maldonado

TUTOR PRINCIPAL:

Dr. Erick Treviño Aguilar
Instituto de Matemáticas Unidad Cuernavaca

CIUDAD UNIVERSITARIA, CDMX, FEBRERO 2024

A mis padres, el Instituto de Matemáticas y a la Universidad, por la formación que me han dado. Es gracias a ustedes que es posible el presente trabajo.
En verdad, gracias.
Eric Martínez Maldonado.

Reconocimientos

Trabajo realizado gracias al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM, proyecto TA101322.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Eric Martínez Maldonado. Ciudad Universitaria, CDMX, Febrero 2024

Resumen

La calificación de riesgo es una herramienta utilizada para evaluar el nivel de riesgo de incumplimiento asociado con los solicitantes de un crédito (modelos de originación), o incluso de clientes que ya forman parte de la cartera (modelos de comportamiento), en este sentido los modelos de clasificación pueden ser de ayuda para predecir el comportamiento crediticio de un cliente, por lo cual pueden ser utilizados para la toma de decisiones de la empresa emisora, y de esta forma aplicar las acciones necesarias para la oportuna recuperación de dicho crédito.

La presente tesis tiene como finalidad el hacer una comparación de metodologías para generar modelos de clasificación. Este análisis comparativo se basó en datos de una cartera de crédito especializada en microcréditos. Se deseaba determinar si un crédito ya aprobado y con una cierta cantidad de tiempo activo podría ser clasificado como un buen crédito durante el resto de tiempo que durara, teniendo como referencia algunas variables explicativas. La implementación de los modelos de clasificación a los datos reales se hizo mediante scripts del software libre R. Para todos los modelos se usaron los mismos datos de entrenamiento y los mismos datos de prueba para sus evaluaciones. Entre las principales conclusiones se puede señalar que en general todos los modelos fueron competitivos, destacando el modelo de Bosques Aleatorios.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Este trabajo tiene por objetivo la comparación del desempeño de diferentes modelos de clasificación para observar su efectividad al ser aplicados en la generación de un modelo de comportamiento crediticio. Así, el tema principal del presente trabajo de tesis es la calificación crediticia para el comportamiento de las personas cuyo crédito ya ha sido aprobado. Para el objetivo de la comparación se ajustan con datos reales cuatro modelos:

- Modelo de regresión logística (Logit),
- Modelo discriminante lineal (LDA),
- Modelo de maquinas de soporte vectorial (SVM) y
- Bosques aleatorios.

Se utilizan datos reales de una cartera de crédito especializada en microcréditos destinados a la compra de equipo electrónico con montos que en promedio no superan los 20 mil pesos.

El preprocesamiento de los datos fue una parte escencial del ejercicio de ajustar los modelos y representó un porcentaje significativo en términos de tiempo y en la cantidad de código que requirió. Dicho preprocesamiento va en el sentido de la metodología presentada en el Capítulo 4.

Además de esta introducción, el trabajo está organizado de la siguiente forma. En el Capítulo 2, se presentan los fundamentos de los modelos que se utilizan para el ajuste de datos, estos son, Modelo Logit, LDA, SVM y Bosques aleatorios. Así mismo, se presenta el indicador de ajuste dado por la curva ROC que se utiliza para medir el desempeño de los modelos ajustados. También en este capítulo se hace una breve exposicion de lo que la literatura reporta con respecto a la modelación en calificación crediticia.

En el Capítulo 3 se hace una breve reseña de antecedentes relativos al proceso de originar y administrar carteras de créditos, se señalan las diferentes etapas por las que pasa un crédito, particularmente la de originación.

El Capítulo 4 es crucial para el ejercicio de ajustar el Modelo logit, LDA, SVM y Bosques aleatorios con datos reales. En él se describen los pasos a seguir para el procesamiento de los datos. Estos consisten en (i) identificar las exclusiones de observaciones, (ii) identificar ventana de tiempo para realizar cortes generacionales, (iii) identificar la correcta definición del evento crediticio. Finalmente y no menos importante, (iv) identificar la necesidad de segmentar la cartera de crédito.

El Capítulo 5 reporta los resultados obtenidos de ajustar los modelos. El Capítulo 6 concluye el trabajo.

Capítulo 2

Revisión de modelos de datos

Una tarea fundamental es la así llamada clasificación supervisada. De ahí el interés en los modelos desarrollados en estadística (por ejemplo la regresión logística o el análisis discriminante) o en los así llamados modelos de aprendizaje de maquina (redes neuronales, árboles de decisión, bosques aleatorios, etc.).

En este capítulo se introducen modelos de clasificación que serán aplicados a un conjunto de datos reales. Estos modelos son, la regresión logística, el análisis discriminante lineal, maquina de soporte vectorial, y bosques aleatorios. También se presenta a detalle una métrica con la que se puede diagnosticar el desempeño de cada modelo, la métrica denotada por sus siglas en inglés como AUC (del término en inglés Area Under the Curve).

Denotemos mediante X un vector de entrada (que reúne las variables explicativas) y mediante G una variable de salida que representa una clasificación (también llamada variable a explicar). La tarea se puede describir como: dado el valor de un vector de entrada X, hacer una predicción de la salida G, denotada por $\hat{G}$. En los problemas de clasificación G es una variable categórica y entonces el pronostico $\hat{G}$ también deberá ser categórica con las mismas clases. El éxito del modelo de clasificación se mide en función de métricas que se construyen a partir del número de veces que $\hat{G}$ coincide con G.

2.1. Regresión Logística

Aquí discutiremos un enfoque de clasificación donde algunas o todas las variables son cualitativas. Este enfoque se llama regresión logística. En su versión más simple, la variable de respuesta G es dicotómica, es decir toma solamente dos valores. Por ejemplo, G puede registrarse como “hombre” o “mujer” o “empleado” y “no empleado” [9].

Aunque la variable respuesta puede ser una variable cualitativa de dos resultados, siempre podemos codificar los dos casos como 0 y 1. Por ejemplo, podemos tomar hombre = 0 y mujer = 1. La probabilidad p de la categoría 1 es un parámetro de interés. Representa la proporción de la población que presenta la etiqueta 1. La media de la distribución de ceros y unos también es p, ya que

$$\text{media} = 0 \times (1-p) + 1 \times p = p$$

La proporción de ceros es $1-p$ la cual a veces es denotada por q. La varianza de la distribución es

$$\text{varianza} = 0^2 \times (1-p) + 1^2 \times p - p^2 = p(1-p).$$

Es claro que la varianza depende de p y que tiende a 0 cuando p tiende a 0 o 1.

Sea G la variable respuesta que toma valores en el conjunto $\{0, 1\}$. Si tuviéramos que modelar la probabilidad del evento $\{G = 1\}$ con un modelo lineal de un solo predictor, escribiríamos

$$p = E(G|X = x) = \beta_0 + \beta_1 x$$

y luego agregariamos un término de error ϵ. Lo anterior presenta los siguientes inconvenientes:

- Los valores pronosticados de la respuesta G podrían salir del intervalo $[0, 1]$ debido a que la expresión lineal de su valor esperado no tiene límites.
- Uno de los supuestos de un análisis de regresión es que la varianza de G es constante en todos los valores de la variable predictora X. Este no es el caso.

Necesitamos otro enfoque para introducir variables predictoras o covariables X en el modelo (ver [11]). En especial, mediante el modelo logístico.

2.1.1. Modelo Logit

En lugar de modelar directamente la probabilidad p con un modelo lineal, se considera una transformación especial dada por la tasa de momios (*odds ratio*) definida por:

$$\text{momios} = \frac{p}{1-p}.$$

Cuando p se aproxima a uno los momios diverge a infinito y cuando la probabilidad de aproxima a cero también lo harán los momios. Es decir que los momios toman valores en el intervalo $(0, \infty)$ para $p \in (0, 1)$ lo cual será muy conveniente.

Aunado a lo anterior, al aplicar el logaritmo, se tendrá una variable valuada en todo el intervalo $(-\infty, \infty)$. A manera de ejemplo, si una proporción de 0.8 de personas pasaran por la aduana sin que se revise su equipaje, entonces $p = 0.8$ pero los momios de no ser revisado son $0.8/0.2 = 4$, o dicho de otro modo, 4 a 1 de no ser revisado. Hay una falta de simetría en tanto que las probabilidades de ser revisado son $0.2/0.8 = 1/4$. Tomando los logaritmos naturales, encontramos que $\ln(4) = 1.386$ y $\ln(1/4) = -1.386$ son de signo contrario. Lo cual tiene su conveniencia.

En la regresión logística para una variable binaria, modelamos el logaritmo natural de la tasa de momios, que se denomina *logit*(p). Es decir

$$\text{logit}(p) = \ln(momios) = \ln\left(\frac{p}{1-p}\right).$$

El logit es una función de la probabilidad p. En la especificación más simple el logit depende linealmente de la variable predictora. Escribiendo la dependencia en x obtenemos

$$\text{logit}(p(x)) = \ln(momios) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x.$$

La transformación logística es invertible. Exponenciando se tiene

$$\phi(x) := \frac{p(x)}{1-p(x)} = \exp(\beta_0 + \beta_1 x).$$

A continuación, resolviendo para $p(x)$, obtenemos

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

que describe una curva logística. La relación entre p y el predictor x no es lineal sino que tiene un gráfico en forma de S como se ilustra en la Figura 2.1.

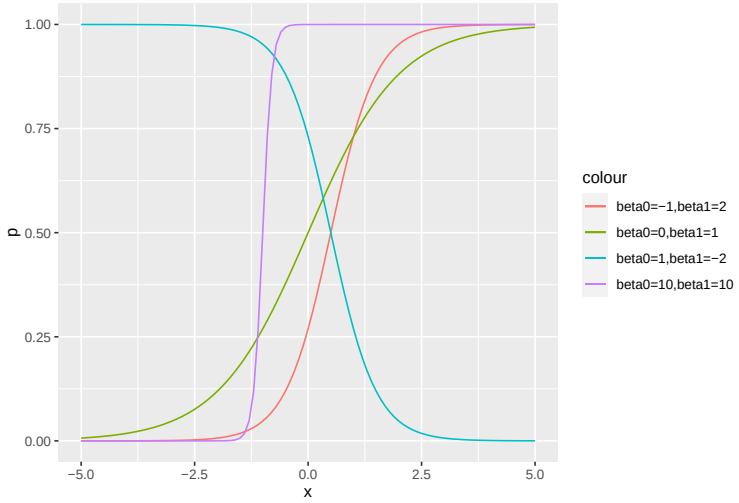

Figura 2.1: Curva logística para diferentes valores de β_0 y β_1

El parámetro β_1 en la curva logística determina qué tan rápido cambia p con x pero su interpretación no es tan simple como en la regresión lineal ordinaria porque la relación no es lineal, ni en x ni en β_1.

En resumen, la curva logística se puede escribir como

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad \text{o bien} \quad p(x) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}.$$

Ahora vamos a considerar el modelo con varias variables predictoras. Sean $(x_{i1}, x_{i2}, \ldots, x_{ip})$ los valores de los p predictores para la i-ésima observación. Es común fijar la primera entrada igual a 1 para incluir un intercepto y en este caso tenemos $x_i = [1, x_{i1}, x_{i2}, \ldots, x_{ip}]'$. Condicional a estos valores, suponemos que la observación G_i es Bernoulli con probabilidad de éxito $p(x_i)$, dependiendo de los valores de las covariables. Entonces

$$p(G_i = g_i) = p^{g_i}(x_i)(1 - p(x_i))^{1-g_i} \quad \text{para } g_i = 0, 1,$$

así que

$$\mathbb{E}(G_i) = p(x_i) \quad \text{y} \quad Var(G_i) = p(x_i)(1 - p(x_i)).$$

Observemos que no es la media la que sigue un modelo lineal sino el logaritmo natural de la tasa de momios. Con toda precisión, la especificación del modelo es la siguiente:

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta' \mathbf{x}_j,$$

donde $\beta = [\beta_0, \beta_1, \ldots, \beta_p]'$ y se incluye al intercepto.

Las estimaciones de los β_i pueden obtenerse mediante el método de máxima verosimilitud. La verosimilitud L viene dada por la distribución de probabilidad conjunta evaluada en los recuentos observados y_j. Entonces la función de verosimilitud es:

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{j=1}^{n} p^{y_j}(x_j)(1 - p(x_j))^{1-y_j} = \frac{\prod_{j=1}^{n} e^{y_j(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}}{\prod_{j=1}^{n} \left(1 + e^{y_j(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}\right)}.$$

Los valores de los parámetros que maximizan la probabilidad no se pueden expresar en una solución de forma cerrada como en el caso de los modelos lineales de mínimos cuadrados ordinarios. En cambio, deben determinarse numéricamente comenzando con una suposición inicial e iterando hasta el máximo de la función de verosimilitud. Técnicamente, este procedimiento se denomina método de mínimos cuadrados iterativamente reponderados (ver [11]). Es común el denotar mediante $\hat{\beta}$ al estimador del vector desconocido β.

2.2. Análisis Discriminante Lineal (LDA)

El análisis discriminante lineal (LDA) es un modelo clásico y bien conocido en problemas de calificación crediticia. Se ha utilizado ampliamente en muchas aplicaciones, como reconocimiento facial, recuperación de imágenes, clasificación de datos de microarrays, etc [16]. El LDA clásico proyecta los datos en un espacio vectorial de dimensiones inferiores de modo que la relación maximiza la distancia entre-clase con respecto a la distancia dentro de clase, logrando así la máxima discriminación.

Para el análisis es conveniente etiquetar las clases π_1 y π_2. Los objetos normalmente se separan y clasifican sobre la base de mediciones de, por ejemplo, p variables aleatorias asociadas $X' = [X_1, X_2, \ldots, X_p]$. Los valores observados de X difieren hasta cierto punto de una clase a otra. Podemos pensar en la totalidad de los valores de la primera clase como la población de valores x para π_1 y los de la segunda clase como la población de valores de x para π_2. Estas dos poblaciones se pueden describir mediante funciones de densidad de probabilidad $f(x)$ y $h(x)$, y en consecuencia, podemos hablar de asignar indistintamente observaciones a poblaciones u objetos a clases [9, Capítulo 11].

Las reglas de asignación o clasificación generalmente se desarrollan a partir de muestras. Las características de los objetos en la muestra para los cuales se tienen las etiquetas son analizadas en busca de diferencias. Por ejemplo en una clasificación binaria, el conjunto de todos los posibles resultados de la muestra se

divide en dos regiones, R_1 y R_2 de algún espacio $\mathbb{R}^n$. Si una nueva observación cae en R_1, se asigna a la población π_1, y si cae en R_2, se asigna a la población π_2. Por lo tanto, un conjunto de valores observados favorece a π_1, mientras que el otro conjunto de valores favorece a π_2. En símbolos, para un vector de características $x \in \mathbb{R}^n$ se asigna la etiqueta $\pi(x)$ de acuerdo a la siguiente regla:

$$\pi(x) = \begin{cases} \pi_1 & \text{si } x \in R_1 \\ \pi_2 & \text{si } x \in R_2. \end{cases}$$

Las regiones R_1 y R_2 se darán con mayor detalle en la Sección 2.2.2 en el caso especial del enfoque de Fisher.

2.2.1. Enfoque de Fisher para la clasificación con dos poblaciones

La idea de Fisher de clasificación era transformar las observaciones multivariadas x en observaciones univariadas y de modo que la variable y derivada de la población π_1 y π_2 estuvieran separadas tanto como fuera posible. Fisher sugirió tomar combinaciones lineales de $\mathbf{x}$ para crear la variable y. La elección es debido a la sencillez de la transformación lineal. El enfoque de Fisher no supone que las poblaciones tengan distribución normal. Sin embargo, asume implícitamente que las matrices de covarianza de la población son iguales, porque se utiliza una estimación conjunta de la matriz de covarianza común. Una combinación lineal fija de $\mathbf{x}$ toma los valores $y_{11}, y_{12}, \ldots, y_{1n_1}$ para las observaciones de la primera población y los valores $y_{21}, y_{22}, \ldots, y_{2n_2}$ para las observaciones de la segunda población (donde n_1 y n_2 es el número de observaciones de la población π_1 y π_2 respectivamente). La separación de estos dos conjuntos de y univariadas se evalúa en términos de la diferencia entre $\bar{y}_l$ y $\bar{y}_2$, expresado en unidades de desviación estándar. Esto es:

$$\text{separación} := \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \quad \text{donde } s_y^2 := \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2},$$

es la estimación agrupada de la varianza. El objetivo es seleccionar la combinación lineal de $\mathbf{x}$ para lograr la máxima separación de las medias muestrales $\bar{y}_1$ y $\bar{y}_2$. Para continuar introducimos algunas definiciones. Supongamos que tenemos n_l observaciones de la variable aleatoria multivariada $X' = [X_1, X_2, ..., X_p]$ de π_1 y n_2 medidas de esta cantidad de π_2, con $n_1 + n_2 - 2 \geq p$. Entonces las respectivas

matrices de datos son

$$\underset{(n_1\times p)}{X_1} = \begin{pmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \vdots \\ \mathbf{x}'_{1n_1} \end{pmatrix}$$

$$\underset{(n_2\times p)}{X_2} = \begin{pmatrix} \mathbf{x}'_{21} \\ \mathbf{x}'_{22} \\ \vdots \\ \mathbf{x}'_{2n_1} \end{pmatrix}.$$

A partir de estas matrices de datos, los vectores de media muestral y las matrices de covarianza se determinan mediante

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}, \quad \underset{(p\times p)_1}{S} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x_{1j}}-\bar{\mathbf{x}}_1)(\mathbf{x_{1j}}-\bar{\mathbf{x}}_1)'$$

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_2} = \frac{1}{n_2}\sum_{j=1}^{n_1}\mathbf{x}_{2j}, \quad \underset{(p\times p)_2}{S} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x_{2j}}-\bar{\mathbf{x}}_2)(\mathbf{x_{2j}}-\bar{\mathbf{x}}_2)'.$$

Dado que se supone que las poblaciones originales tienen la misma matriz de covarianza Σ, las matrices de covarianza muestral S_1 y S_2 se combinan (agrupan) para obtener una única estimación no sesgada de Σ. En particular, el promedio ponderado

$$S_{\text{pooled}} := \left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]S_1 + \left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]S_2.$$

es una estimador insesgado de Σ si las matrices de datos X_l y X_2 contienen muestras aleatorias de las poblaciones π_1 y π_2, respectivamente. Lo que estamos buscando es un vector $\mathbf{a}$ tal que $y = \mathbf{a}'\mathbf{x}$, es decir, $\mathbf{a}$ es el vector que contiene los coeficientes de la combinación lineal de $\mathbf{x}$ que se quiere obtener. Los vectores de medias y matrices de covarianzas son desconocidas y se deben utilizar estimadores. Estos estimadores se distinguen mediante un "gorro".

Resultado: Sea $\hat{\mathbf{a}}' := (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'S^{-1}_{\text{pooled}}$. La combinación lineal $\hat{y} = \hat{\mathbf{a}}'\mathbf{x}$ maximiza la relación

$$\frac{(\bar{y}_1-\bar{y}_2)^2}{s_y^2} = \frac{(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 - \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'S_{\text{pooled}}\hat{\mathbf{a}}} = \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'S_{\text{pooled}}\hat{\mathbf{a}}}$$

sobre todos los vectores de coeficientes posibles $\hat{\mathbf{a}}$ donde $d = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. El máximo de la razón es $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'S^{-1}_{\text{pooled}}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

Demostración. Ver Johnson and Wichern [9, Result 11.3]. □

A la combinación lineal $\hat{y} = \hat{\mathbf{a}}'\mathbf{x}$ se le conoce como la función discriminante y corresponde al discrimante lineal de Fisher bajo el supuesto de distribución normal.

2.2.2. Una regla de asignación basada en la función discriminante de Fisher

Gracias a la función discriminante de Fisher se puede construir una regla de asignación para el objetivo de clasificación. Recuerde la definición $\hat{\mathbf{a}}' := (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S_{\text{pooled}}^{-1}$. La regla es la siguiente: Asigne $\mathbf{x}_0$ a π_1 si

$$\bar{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 \geq \hat{m} = \frac{1}{2}\hat{\mathbf{a}}'(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2),$$

o bien

$$\hat{y}_0 - \hat{m} \geq 0.$$

Asigne $\mathbf{x}_0$ a π_2 si

$$\hat{y}_0 < \hat{m},$$

o bien

$$\hat{y}_0 - \hat{m} < 0.$$

Note que estamos asumiendo que $n_1 + n_2 - 2 \geq p$ ya que de no ser el caso S_{pooled} es singular y por lo tanto la inversa S_{pooled}^{-1} no existe. El término, $\bar{y} = \hat{\mathbf{a}}'\mathbf{x}$, en la regla de clasificación, es la función lineal obtenida por Fisher que maximiza la variabilidad univariada "entre" muestras en relación con la variabilidad "dentro" de las muestras. La expresión completa

$$\hat{w} = \hat{\mathbf{a}}'\mathbf{x} - \frac{1}{2}\hat{\mathbf{a}}'(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \hat{\mathbf{a}}'[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)].$$

se denomina con frecuencia función de clasificación de Anderson.

2.3. Maquina de Soporte Vectorial (SVM)

En esta sección describimos el problema de optimización que define la técnica conocida como máquina de soporte vectorial (abreviado como SVM por sus siglas en inglés). Esta técnica se basa en una frontera de decisión lineal. Es posible definir SVM tanto para problemas de regresión como para problemas de

clasificación siendo la diferencia en la especificación de la función objetivo o de pérdida y se presentan ambos. Existen versiones no lineales al transformar la variable de entrada. No aplicaremos dichas transformaciones no lineales. Sin embargo, en términos de presentación no representa un aumento en el costo y por ello presentamos también la formulación no lineal.

2.3.1. El clasificador de soporte vectorial

Consideremos una variable de salida G con clases 1 y -1. El clasificador de soporte vectorial requiere definir una frontera lineal a través de un hiperplano:

$$\{x : x'\beta + \beta_0 = 0\}$$

donde β es un vector unitario $||\beta|| = 1$ que debe ser seleccionado optimamente. Una regla de clasificación inducida por dicho hiperplano es

$$G(x) = \text{sign}[x'\beta + \beta_0].$$

En el caso en que las clases sean linealmente separables, existirá al menos una función $f(x) = x'\beta + \beta_0$ que satisfaga la condición

$$y_i f(x_i) > 0, \forall i.$$

En este caso pueden existir varias funciones que satisfagan la condición y se debe seleccionar óptimamente. Especialmente al encontrar el hiperplano que crea el mayor margen entre los puntos de entrenamiento para la clase 1 y -1. Este concepto es capturado por el siguiente problema de optimización

$$\max_{\beta,\beta_0,||\beta||=1} M \text{ sujeto a } y_i(x_i'\beta + \beta_0) \geq M, i = 1, \ldots, N. \tag{2.1}$$

Este problema se puede formular de la siguiente forma

$$\min_{\beta,\beta_0} ||\beta||^2 \text{ sujeto a } y_i(x_i'\beta + \beta_0) \geq 1, i = 1, \ldots, N. \tag{2.2}$$

Obsevar que en (2.2) se ha eliminado la restricción de la norma sobre β. El problema (2.2) es la forma habitual de escribir el criterio de soporte vectorial para datos separados. Este es un problema de optimización convexo. La siguiente proposición verifica la afirmación acerca de la relación en los problemas (2.1) y (2.2).

Proposición 2.3.1. *Sea (β_0^*, β^*) la solución al problema (2.2). Para el problema (2.1) una solución está dada por $||\beta^*||^{-1}(\beta_0^*, \beta^*)$ y se satisface*

$$\max_{\beta,\beta_0,||\beta||=1} M(\beta_0, \beta) = \frac{1}{||\beta^*||}. \tag{2.3}$$

Demostración. Sea (β^*, β^*) la solución al problema (2.2). Para cualquier par (α_0, α) con α que satisface la restricción $M(\alpha_0, \alpha) \geq 1$ en el problema (2.2) se tiene

$$y_i(\alpha \cdot x + \alpha_0) \geq 1 \Leftrightarrow y_i(\tilde{\alpha} \cdot x + \tilde{\alpha}_0) \geq \frac{1}{||\alpha||}$$

en donde $\tilde{\alpha}_0 = \frac{1}{||\alpha||}\tilde{\alpha}_0$ y $\tilde{\alpha} = \frac{1}{||\alpha||}\tilde{\alpha}$. Es decir que si α es admisible para el problema (2.2) se tiene para el problema (2.1) lo siguiente

$$\max_{\beta, \beta_0, ||\beta||=1} M(\beta_0, \beta) \geq \frac{1}{||\alpha||}.$$

Verifiquemos que se satisface la igualdad (2.3). Por contradicción suponga la desigualdad estricta en (2.3). Entonces para $\epsilon > 0$ suficientemente pequeño y (β_0, β) admisible para (2.1) se tiene que

$$M(\beta_0, \beta) > (1 + \epsilon)\frac{1}{||\beta^*||}.$$

Como consecuencia de la anterior desigualdad se tiene que $(1 + \epsilon)^{-1}||\beta^*||(\beta_0, \beta)$ es admisible para el problema (2.2). Esto es una contradicción a la minimalidad de la norma de β^* ya que $(1 + \epsilon)^{-1}||\beta^*||\beta$ es admisible y tiene norma menor. La contradicción generada prueba la igualdad deseada (2.3). □

El vector solución β se puede expresar en términos de una combinación lineal de los puntos de apoyo x_i (ver [7]).

En el caso en que las observaciones no sean linealmente separables, se relaja el problema introduciendo variables de holgura $\xi = (\xi_1, \xi_2, ..., \xi_N)$. La restricción en (2.1) se modifica de la siguiente forma

$$y_i(x_i'\beta + \beta_0) \geq M(1 - \xi_i),$$

$\forall i$, $\xi_i \geq 0$, $\sum_{i=1}^{N} \xi_i \leq$ constante. El problema es convexo y permite traslapes (penalizados) de las observaciones.

El valor ξ_i en la restricción $y_i(x_i'\beta + \beta_0) \geq M(1 - \xi_i)$ es la cantidad proporcional por la cual la predicción $f(x_i) = x_i'\beta + \beta_0$ está en el lado equivocado de su margen. Acotar la suma $\sum \xi_i$ lleva a acotar también la cantidad proporcional total por la cual las predicciones caen en el lado equivocado de su margen. Es posible de manera análoga a lo verificado en la Proposición 2.3.1 que es posible eliminar la

restricción de la norma sobre β. Definamos $M = 1/||\beta||$ y escribamos el problema de optimización

$$\text{mín}\,||\beta||,\ \text{sujeto a}\ \begin{cases} y_i(x'\beta + \beta_0) \geq 1 - \xi_i\ \forall i \\ \xi_i \geq 0, \sum \xi_i \leq \text{constante} \end{cases} \tag{2.4}$$

Esta es la forma habitual en que se define el clasificador de soporte vectorial para el caso no separable.

El problema (2.4) es cuadrático con restricciones de desigualdad lineal, por lo que es un problema de optimización convexo. Es posible caracterizar la solución mediante programación cuadrática utilizando multiplicadores de Lagrange. Computacionalmente es conveniente reexpresar (2.4) en la forma equivalente:

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N} \xi_i \text{ sujeto a } \xi_i \geq 0, y_i(x_i'\beta + \beta_0) \geq 1 - \xi_i\ \forall i$$

donde el parámetro de "costo" C reemplaza la constante en (2.4); el caso separable corresponde a $C = \infty$. El Lagrangiano del problema es

$$L_P = \frac{1}{2} + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i[y_i(x_i'\beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{N} \mu_i\xi_i \tag{2.5}$$

que minimizamos con respecto a β, β_0 y ξ_i. Igualando a cero las derivadas respectivas, obtenemos

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i$$

$$\alpha_i = C - \mu_i, \forall i,$$

así como las restricciones de positividad $\alpha_i, \mu_i, \xi_i \geq 0 \forall i$. Sustituyendo las ecuaciones anteriores en (2.5), obtenemos la función objetivo dual Lagrangiana (muchas veces llamada función objetivo de Wolfe).

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j x_i'x_j. \tag{2.6}$$

La solución a dicho sistema se puede implementar númericamente.

$$K(x, x') = \langle h(x), h(x')\rangle$$

2.4. Bosques aleatorios

2.4.1. Árboles de regresión y clasificación

Antes de definir un bosque aleatorio definiremos un arbol de regresión y de clasificación. Un árbol de decisión no es más que declaraciones condicionales que pueden usarse para predecir un resultado basado en datos, en la Figura 2.2 se observa un ejemplo de un árbol de clasificación bastante simple, en el cual se intenta clasificar el género de un individuo basado en su altura y su peso, en la primera declaración condicional if-else se pregunta si su altura es mayor a 180 cm, si cumple la condición se clasifica como género masculino, de lo contrario se pregunta si su peso es mayor a 80 kg, clasificando como masculino si esto es verdadero y femenino si es falso. Por otro lado un árbol de regresión se refiere a un algoritmo donde está la variable objetivo y el algoritmo se utiliza para predecir su valor. Como ejemplo de un problema de tipo regresión, es posible que desee predecir los precios de venta de una casa residencial, que es una variable dependiente continua.

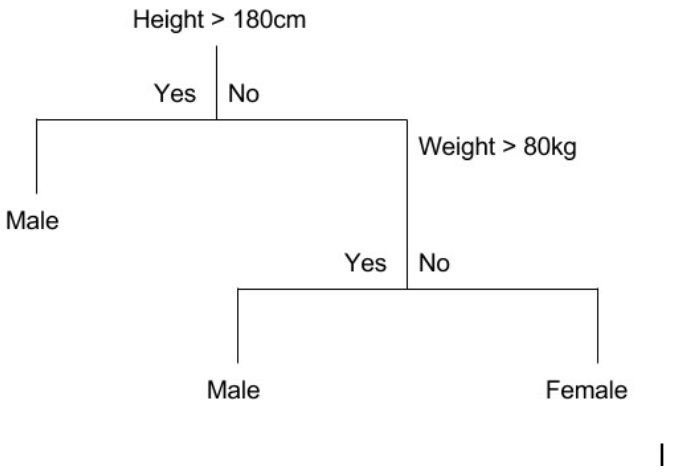

Figura 2.2: Ejemplo de árbol de clasificación

De manera más rigurosa consideremos un problema de regresión con variable de respuesta continua Y y p entradas $X_1, X_2, \ldots, X_p$, para cada una de las N observaciones: es decir, (x_i, y_i) para $i = 1, 2, ..., N$, con $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. El algoritmo debe decidir automáticamente las variables de división y los puntos de división, y también qué topología (forma) debe tener el árbol. Supongamos

primero que tenemos una partición en M regiones $R_1, R_2, ..., R_M$, y modelamos la respuesta como una constante c_m en cada región:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m).$$

Entonces los árboles de decisión dividen el espacio de todos los valores de las variables predictoras conjuntas en regiones separadas $R_j, j = 1, 2, ..., J$ (ver [7]), representadas por los nodos terminales del árbol. A partir de eso se asigna una constante c_m a cada región y la regla predictiva es

$$x \in R_m \Rightarrow \hat{f}(x) = c_m.$$

Así, un árbol se puede expresar formalmente como

$$T(x; \Theta) = \sum_{m=1}^{J} c_m I(x \in R_m)$$

con parámetros $\Theta = \{R_j, c_j\}_1^J$ y donde $I(S)$ es la función indicadora del conjunto S. La constante J generalmente se trata como un metaparámetro. Podemos encontrar los parámetros minimizando el riesgo empírico

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{x_i \in R_j} L(y_i, \gamma_j). \tag{2.7}$$

Encontrar la mejor partición binaria en términos de la suma de mínimos cuadrados es, generalmente, computacionalmente inviable, por lo tanto, se procede con el siguiente algoritmo [7]:

Comenzando con todos los datos, considere una variable de división j y un punto de división s, y defina el par de semiplanos

$$R_1(j, s) = \{X | X_j \leq s\} \quad \text{y} \quad R_2(j, s) = \{X | X_j > s\}$$

Luego buscamos la variable de división j y el punto de división s que resuelven

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

Para cualquier elección j y s, la minimización interna se resuelve mediante

$$\hat{c}_1 = \text{promedio}(y_i | x_i \in R_1(j, s)) \quad \text{y} \quad \hat{c}_2 = \text{promedio}(y_i | x_i \in R_2(j, s))$$

Para cada variable de división, la determinación del punto de división s se puede realizar muy rápidamente y, por lo tanto, mediante la exploración de todas las entradas, la determinación del mejor par (j, s) es factible.

Habiendo encontrado la mejor división, dividimos los datos en las dos regiones resultantes y repetimos el proceso de división en cada una de las dos regiones. Luego, este proceso se repite en todas las regiones resultantes.

Sin embargo, (Breiman, 2001) se pregunta ¿Qué tan grande debemos hacer crecer el árbol?, a lo cual llega a la respuesta de que claramente, un árbol muy grande podría sobreajustar los datos, mientras que un árbol pequeño podría no capturar la estructura importante. El tamaño del árbol es un parámetro de ajuste que rige la complejidad del modelo, y el tamaño óptimo del árbol debe elegirse de forma adaptativa a partir de los datos. Un enfoque sería dividir los nodos del árbol solo si la disminución en la suma de los cuadrados debido a la división supera algún umbral. Sin embargo, no es muy adecuada, ya que una división aparentemente sin valor podría conducir a una división muy buena debajo de ella.

La estrategia más común es hacer crecer un árbol grande T_0, deteniendo el proceso de división solo cuando se alcanza un tamaño mínimo de nodo. Luego, este árbol grande se poda utilizando la poda de complejidad de costos, descrita a continuación.

Definimos un subárbol $T \subset T_0$ como cualquier árbol que se puede obtener podando T_0, es decir, colapsando cualquier número de sus nodos internos (no terminales). Indexamos los nodos terminales por m, donde el nodo m representa la región R_m. Sea $|T|$ el número de nodos terminales en T. Sean:

$$\begin{aligned} N_m &= \#\{x_i \in R_m\} \\ \hat{c}_m &= \frac{1}{N_m} \sum_{x_i \in R_m} y_i \\ Q_m(T) &= \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2 \end{aligned} \tag{2.8}$$

definimos el criterio de complejidad de costes

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

La idea es encontrar, para cada α, el subárbol $T_\alpha \subseteq T_0$ para minimizar $C_\alpha(T)$. El parámetro de ajuste $\alpha \geq 0$ gobierna el equilibrio entre el tamaño del árbol y su bondad de ajuste a los datos. Los valores grandes de α dan como resultado árboles T_α más pequeños y, a la inversa, para valores más pequeños de α. Como sugiere la notación, con $\alpha = 0$ la solución es el árbol completo T_0.

Para cada α se puede mostrar que existe un único subárbol más pequeño T_α que minimiza $C_\alpha(T)$ [7]. Para encontrar T_α usamos la poda de enlace más débil: colapsamos sucesivamente el nodo interno que produce el aumento más pequeño por nodo en $\sum_m N_m Q_m(T)$, y continuamos hasta que producimos el árbol de un solo nodo (raíz). Esto da una secuencia (finita) de subárboles, y uno puede mostrar que esta secuencia debe contener T_α [7]. La estimación de α se logra mediante una validación cruzada de cinco o diez veces: elegimos el valor $\hat{\alpha}$ para minimizar la suma de cuadrados validada de forma cruzada. Nuestro último árbol es $T_{\hat{\alpha}}$.

Para el caso de los árboles de clasificación, es decir, cuando la variable objetivo es tal que toma los valores $1, 2, ..., K$, los únicos cambios necesarios en el algoritmo del árbol pertenecen a los criterios para dividir los nodos y podar el árbol. Para la regresión usamos la medida de impureza del nodo de error cuadrático $Q_m(T)$ definida en 2.8, pero esto no es adecuado para la clasificación. En un nodo m, que representa una región R_m con N_m observaciones, sea

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

la proporción de observaciones de clase k en el nodo m. Clasificamos las observaciones en el nodo m en la clase $k(m) = \arg\max_k \hat{p}_{mk}$, la clase mayoritaria en el nodo m. Las diferentes medidas $Q_m(T)$ de la impureza del nodo incluyen las siguientes [7]:

- Error de clasificación errónea: $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk}(m)$.
- Índice de Gini: $\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$.
- Entropía cruzada o desviación: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

2.4.2. Definición de bosque aleatorio

El modelo predictivo de bosques aleatorios consiste como su nombre sugiere en un conjunto de arboles de decisión los cuales lleván en su conjunto a una clasificación. La construcción de dicho modelo se basa en un método conocido en inglés como bagging. Siguiendo [7, Capítulo 15] veamos en que consiste dicho método. La idea esencial es promediar muchos modelos ruidosos pero aproximadamente imparciales y, por lo tanto, reducir la varianza. Los árboles son candidatos ideales para el bagging, ya que pueden capturar estructuras de interacción complejas en los datos y en el límite tienen un sesgo relativamente bajo. Dado que los árboles son notoriamente ruidosos, se benefician enormemente del promedio. Además,

dado que cada árbol generado en el bagging son idénticamente distribuidos (i.d.), la esperanza de un promedio de B de dichos árboles es la misma que la esperanza de cualquiera de ellos. Esto significa que el sesgo de los árboles de bagging es el mismo que el de los árboles individuales, y la mejora del modelo es posiblemente medinate la reducción de la varianza. Veamos este último punto con mayor detalle.

Un promedio de B variables aleatorias identicamente distribuidas e independientes (iid), cada una con varianza σ^2, tiene varianza $(1/B)\sigma^2$. Si las variables son simplemente i.d. (idénticamente distribuidas, pero no necesariamente independientes) con correlación positiva por pares ρ, la varianza del promedio es [7, Capítulo 15]:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{2.9}$$

A medida que B aumenta, el segundo término desaparece, este es el beneficio del método bagging. Ahora bien, para controlar el primer término $\rho\sigma^2$, la idea en los bosques aleatorios es reducir la correlación entre los árboles, sin aumentar demasiado la varianza. Esto se logra en el proceso de crecimiento de árboles a través de la **selección aleatoria** de las variables de entrada.

El modelo lo formalizamos mediante la descripción del algoritmo empleado para crear un bosque aleatorio.

El algoritmo de un bosque aleatorio para regresión o clasificación reportado en [7, Algoritmo 15.1 p. 588] es el siguiente:

1. Para $b = 1$ hasta B:

   a) Extraiga una muestra Z^* de tamaño N a partir de los datos de entrenamiento.

   b) Haga crecer un árbol de bosque aleatorio T_b a los datos Z^* repitiendo recursivamente los siguientes pasos para cada nodo terminal del árbol, hasta alcanzar el tamaño mínimo de nodos n_{min} (i.e. repetir los siguientes pasos mientras el numero de nodos terminales no alcance un cierto umbral especificado):

      I Seleccionar m variables aleatoriamente de las p variables totales.

      II Seleccione la mejor variable de clasificación de las m variables muestreadas.

      III Divida el nodo.

2. La salida del algoritmo será el conjunto de árboles $\{T_b\}_1^B$.

Para hacer una predicción en un nuevo punto x:

- Regresión:

$$\hat{f}_{\mathrm{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b). \tag{2.10}$$

- Clasificación: Sea $\hat{C}_b(x)$ la predicción de clase del b-ésimo árbol de bosque aleatorio. Entonces $\hat{C}_{\mathrm{rf}}^{B}(x) = \text{voto mayoritario}\{\hat{C}_b(x)\}_1^B$.

Se hacen las siguientes recomendaciones con respecto a los hiperparametros:

- Para la clasificación, el valor predeterminado para m es $\lfloor\sqrt{p}\rfloor$ y el tamaño mínimo de nodo es uno.
- Para la regresión, el valor predeterminado de m es $\lfloor p/3\rfloor$ y el tamaño mínimo de nodo es cinco.

2.5. Indicadores de Ajuste

Entendemos por un indicador de bondad de ajuste aquel indicador que nos proporciona información sobre el grado de acoplamiento que existe entre nuestros datos originales y los valores teóricos que se obtienen a través de el modelo que estemos realizando. Obviamente cuanto mejor sea el ajuste, más útil será el modelo en la pretensión de obtener los valores de la variable objetivo utilizando nuestras variables dependientes.

2.5.1. Curva ROC

Ahora analizaremos uno de los indicadores de bondad de ajuste que utilizaremos en nuestro análisis, la curva ROC [12], formalmente podemos plantear el siguiente problema: cada instancia I se asigna a un elemento del conjunto $\{p, n\}$ de clases positivas (correctas) y negativas. Un modelo de clasificación (o clasificador) es un mapeo de instancias a clases predichas. Algunos modelos de clasificación producen una salida continua (por ejemplo, una estimación de la probabilidad de pertenencia a una clase de una instancia) a la que se pueden aplicar diferentes umbrales para predecir la pertenencia a una clase. Para distinguir entre la clase real y la clase predicha de una instancia, usaremos las etiquetas $\{Y, N\}$ para las clasificaciones producidas por un modelo. Para nuestra discusión, sea $c(\text{clasificación}, \text{clase})$ una función de costo de error de dos posiciones donde $c(Y, n)$ es el costo de un error falso positivo y $c(N, p)$ es el costo de un error falso negativo. Representamos las distribuciones de clase por las probabilidades

a priori de las clases $\mathbb{P}(p)$ y $\mathbb{P}(n) = 1 - \mathbb{P}(p)$.
La tasa de verdaderos positivos (TPR), o tasa de aciertos, de un clasificador es:

$$TPR = \mathbb{P}(Y|p) \approx \frac{\text{positivos correctamente clasificados}}{\text{positivos totales}} = \frac{TP}{P}.$$

La tasa de falsos positivos (FPR), o tasa de falsas alarmas, de un clasificador es:

$$FPR = \mathbb{P}(Y|n) \approx \frac{\text{negativos clasificados incorrectamente}}{\text{negativos totales}} = \frac{FP}{N}.$$

Algunos términos adicionales asociados son:

$$\begin{aligned} \text{precisión} &= \frac{TP}{TP + FP} \\ \text{sensibilidad} &= \frac{TP}{P} \\ \text{exactitud} &= \frac{TP + TN}{P + N} \\ \text{especificidad} &= \frac{TN}{FP + TN}. \end{aligned}$$

Al error que se comete al predecir un falso positivo también se le conoce como error tipo I o error tipo alfa (α), mientras que al error de predecir un falso negativo se le denomina también como error tipo II o error tipo beta (β).

Sin embargo, el enfoque experimental tradicional es frágil porque elige un modelo como "mejor" con respecto a un conjunto específico de funciones de costo y distribución de clases. Si las condiciones del objetivo cambian, es posible que este sistema ya no funcione de manera óptima o incluso aceptable. Para ilustrar la idea anterior veamos el siguiente ejemplo, supongamos que tenemos una tasa máxima de falsos positivos FPR, que no debe ser excedida. Queremos encontrar el clasificador con la tasa de verdaderos positivos más alta posible, TPR, que no exceda el límite de FPR. Este es el criterio de decisión de Neyman-Pearson [4]. En la Figura 2.3 se muestran tres clasificadores, bajo tres límites de FPR. Un clasificador diferente es mejor para cada límite de FPR; cualquier sistema creado con un solo clasificador "mejor" es frágil si el requisito de FPR puede cambiar.

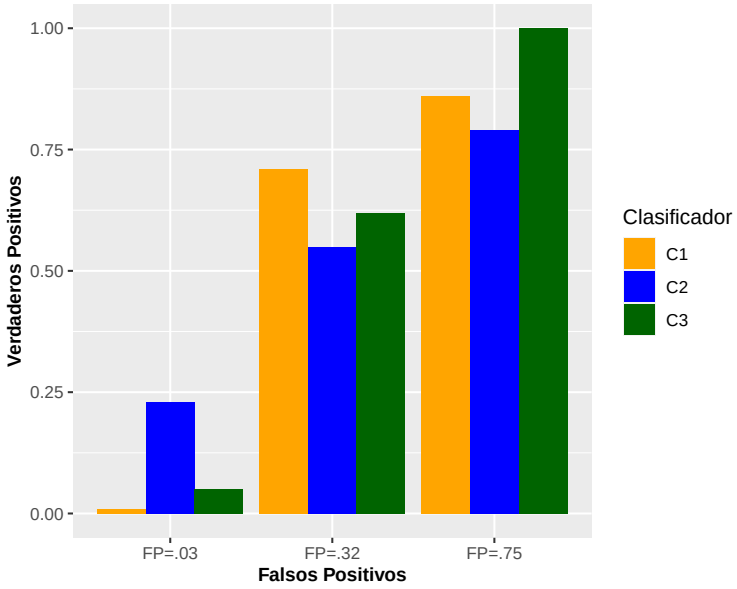

Figura 2.3: Tres clasificadores bajo tres criterios de decisión diferentes de Neyman-Pearson. Elaboración propia basado en [12]

La mayoría de los trabajos sobre la construcción de clasificadores utilizan la precisión de la clasificación (o, de manera equivalente, la tasa de error no diferenciada) como la métrica de evaluación principal. El uso de la precisión supone que las clases previas en el entorno de destino serán constantes y relativamente equilibradas. En el mundo real esto raramente es el caso. Los clasificadores a menudo se utilizan para filtrar una gran población de entidades normales o sin interés para encontrar un número relativamente pequeño de entidades inusuales.

A medida que la distribución de clases se vuelve más sesgada, la evaluación basada en la precisión se desmorona. Considere un dominio donde las clases aparecen en una proporción de 999:1. Una regla simple, clasificar siempre como la clase de máxima probabilidad, da una precisión del 99.9 %. Esta precisión puede ser bastante difícil de superar para un algoritmo de inducción, aunque presumiblemente la regla simple es inaceptable si se busca una solución no trivial. Los sesgos de 10^2 son comunes en la detección de fraude y se han informado sesgos superiores a 10^6 en otras aplicaciones [12].

La evaluación por precisión de clasificación también asume costos de error iguales: $c(Y, n) = c(N, p)$. En el mundo real, las clasificaciones conducen a acciones que tienen consecuencias. Las acciones pueden ser tan diversas como negar un cargo de crédito, descartar una pieza fabricada, mover una superficie de con-

trol en un avión o informar a un paciente sobre un diagnóstico de cáncer. Las consecuencias pueden ser graves y realizar una acción incorrecta puede ser muy costoso. Rara vez los costos de los errores son equivalentes. En la clasificación de hongos, por ejemplo, juzgar que un hongo venenoso es comestible es mucho peor que juzgar que un hongo comestible es venenoso. De hecho, es difícil imaginar un dominio en el que un sistema de clasificación pueda ser indiferente a si comete un error falso positivo o falso negativo. En tales casos, la maximización de la precisión debe reemplazarse por la minimización de costos.

Provost & Fawcett [12] mencionan que los problemas de costos de error desiguales y distribuciones de clases desiguales están relacionados. Se ha sugerido que, para el entrenamiento, las instancias de alto costo pueden compensarse aumentando su prevalencia en un conjunto de instancias. Desafortunadamente, se ha publicado poco trabajo sobre cualquiera de los dos problemas. Existen varias docenas de artículos en los que se sugieren técnicas para el aprendizaje sensible a los costos, pero pocos estudios las evalúan y comparan. La literatura brinda aún menos orientación en situaciones donde las distribuciones son imprecisas o pueden cambiar.

Dada una estimación de $\mathbb{P}(p|I)$, la probabilidad posterior de la pertenencia a una clase de una instancia, el análisis de decisión nos brinda una forma de producir clasificaciones sensibles al costo [15]. Las frecuencias de error del clasificador se pueden utilizar para aproximar tales probabilidades [12]. Para una instancia I, la decisión de emitir una clasificación positiva desde un clasificador particular es:

$$[1 - \mathbb{P}(p|I)]c(Y, n) < \mathbb{P}(p|I)c(N, p).$$

Independientemente de si un clasificador produce clasificaciones probabilísticas o binarias, su costo normalizado en un conjunto de prueba se puede evaluar empíricamente (una esperanza) como:

$$\text{Costo} = FPRc(Y, n) + (1 - TPR)c(N, p).$$

La mayoría de los trabajos publicados sobre la clasificación sensible a los costos utilizan una ecuación como esta para clasificar los clasificadores. Dado un conjunto de clasificadores, un conjunto de ejemplos y una función de costo precisa, se calcula el costo de cada clasificador y se elige el clasificador de costo mínimo. Sin embargo, como se discutió anteriormente, dichos análisis asumen que las distribuciones son estáticas y conocidas con precisión.

Es importante mencionar que Provost & Fawcett [12] también hacen mención de que se pueden hacer comparaciones más generales con el análisis de características operativas del receptor (Receiver Operating Characteristic, ROC), una metodología clásica de la teoría de detección de señales que es común en el diagnóstico médico y comenzó a usarse de manera más general en el trabajo de

clasificación de IA. Los gráficos ROC muestran las compensaciones entre la tasa de aciertos y la tasa de falsas alarmas.

Usamos el término espacio ROC para indicar el sistema de coordenadas utilizado para visualizar el rendimiento del clasificador. En el espacio ROC, TPR se representa en el eje Y y FPR se representa en el eje X. Cada clasificador está representado por el punto en el espacio ROC correspondiente a su par (FPR, TPR). Para modelos que producen una salida continua, por ejemplo, probabilidades posteriores, TPR y FPR varían juntos como un umbral en la salida varía entre sus extremos (cada umbral define un clasificador); la curva resultante se denomina curva ROC. Una curva ROC ilustra las compensaciones de error disponibles con un modelo dado.

La elección se realiza mediante la comparación del área bajo la curva de las respectivas curvas ROC de los clasificadores. Esta área posee un valor comprendido entre 0.5 y 1, donde 1 representa un valor de una predicción perfecta y 0.5 es un clasificador sin capacidad discriminatoria. Es decir, si área bajo la curva para un clasificador es 0.8 significa que existe un 80 % de probabilidad de que la predicción realizada sea correcta contra un 20 % de que sea incorrecta. Por esto, siempre se elige el clasificador que presente un mayor área bajo la curva.

Como orientación, se deben anotar varios puntos en un gráfico ROC. El punto inferior izquierdo (0, 0) representa la estrategia de nunca alarmar, el punto superior derecho (1, 1) representa la estrategia de siempre alarmar, el punto (0, 1) representa la clasificación perfecta y la recta $y = x$ representa la estrategia de adivinar aleatoriamente la clase. Informalmente, un punto en el espacio ROC es mejor que otro si está al noroeste (TPR es más alto, FPR es más bajo o ambos). Un gráfico ROC permite una comparación visual informal de un conjunto de clasificadores.

Los gráficos ROC ilustran el comportamiento de un clasificador sin tener en cuenta la distribución de clases o el costo del error, por lo que desvinculan el rendimiento de la clasificación de estos factores. Desafortunadamente, si bien un gráfico ROC es una técnica de visualización valiosa, no ayuda en la elección de los clasificadores. Solo cuando un clasificador domina claramente a otro en todo el espacio de actuación puede declararse mejor.

Muchos modelos de clasificadores son discretos: están diseñados para producir solo una etiqueta de clase de cada instancia de prueba. Sin embargo, a menudo queremos generar una curva ROC completa a partir de un clasificador en lugar de un solo punto. Con este fin, queremos generar puntajes a partir de un clasificador en lugar de solo una etiqueta de clase. Hay varias formas de producir tales partituras.

Muchos modelos de clasificadores discretos se pueden convertir fácilmente en clasificadores de puntuación "mirando dentro" (looking inside) de ellos en las estadísticas de instancias que mantienen. Por ejemplo, un árbol de decisión de-

termina una etiqueta de clase de un nodo hoja a partir de la proporción de instancias en el nodo; la decisión de clase es simplemente la clase más predominante. Estas proporciones de clase pueden servir como puntaje. Un aprendiz de reglas mantiene estadísticas similares sobre la confianza de la regla, y la confianza de una regla que coincide con una instancia se puede usar como puntaje [5].

Incluso si un clasificador solo produce una etiqueta de clase, se puede usar una agregación de ellos para generar una puntuación. MetaCost, el cual es un método para hacer a los clasificadores sensibles a costos [3], emplea bagging para generar un conjunto de clasificadores discretos, cada uno de los cuales produce un voto. El conjunto de votos podría utilizarse para generar una puntuación. Finalmente, se puede emplear alguna combinación de puntuación y votación. Por ejemplo, las reglas pueden proporcionar estimaciones de probabilidad básicas, que luego pueden usarse en la votación ponderada.

Dado un conjunto de prueba, a menudo queremos generar una curva ROC de manera eficiente a partir de él. Podemos explotar la monotonicidad de las clasificaciones con umbral: cualquier caso que se clasifique como positivo con respecto a un umbral determinado también se clasificará como positivo para todos los umbrales inferiores. Por lo tanto, podemos simplemente ordenar las instancias de prueba decreciendo por puntajes f y movernos hacia abajo en la lista, procesando una instancia a la vez y actualizando TP y FP a medida que avanzamos. De esta forma, se puede crear un gráfico ROC a partir de un escaneo lineal.

El algoritmo es el siguiente:
Entradas: L, el conjunto de ejemplos de prueba; $f(i)$, la estimación del clasificador probabilístico de que el ejemplo i es positivo; P y N, el número de ejemplos positivos y negativos.
Salidas: R, una lista de puntos ROC que aumenta por FPR.
Requerir: $P > 0$ y $N > 0$

1. $L_{\text{sorted}} = L$ ordenados decrecientes por puntajes f
2. $FP \leftarrow TP \leftarrow 0$
3. $R \leftarrow \langle\rangle$
4. $f_{\text{prev}} \leftarrow -\infty$
5. $i \leftarrow 1$
6. **while** $i \leq |L_{\text{sorted}}|$ **do**
7. **if** $f(i) \neq f_{\text{prev}}$ **then**
8. push $\left(\frac{FP}{N}; \frac{TP}{P}\right)$ en R

9. $f_{\text{prev}} \leftarrow f(i)$
10. **end if**
11. **if** $L_{\text{sorted}}[i]$ es un ejemplo positivo, **then**
12. $TP \leftarrow TP + 1$
13. **else** $/* \, i$ es un ejemplo negativo $* /$
14. $FP \leftarrow FP + 1$
15. **end if**
16. $i \leftarrow i + 1$
17. **end while**
18. push$\left(\frac{FP}{N}; \frac{TP}{P}\right)$ en R $/*$ Esto es $(1, 1) * /$
19. **end**

Sea n el número de puntos en el conjunto de prueba. Este algoritmo requiere una ordenación $\mathrm{O}(n \log n)$ seguida de un escaneo $O(n)$ hacia abajo en la lista, lo que da como resultado una complejidad total $\mathrm{O}(n \log n)$.

Las líneas 7 a 10 necesitan alguna explicación. Estos son necesarios para manejar correctamente secuencias de instancias con la misma puntuación. Suponga que tenemos un conjunto de prueba en el que hay una secuencia de instancias, cuatro negativas y seis positivas, todas puntuadas por igual por f. La clasificación en la línea 1 del algoritmo no impone ningún orden específico en estas instancias ya que sus puntajes f son iguales. ¿Qué sucede cuando creamos una curva ROC? En un caso extremo, todos los positivos terminan al comienzo de la secuencia y generamos el segmento L superior "optimista". En el extremo opuesto, todos los negativos terminan al comienzo de la secuencia y obtenemos la L inferior "pesimista" que se muestra en la Fig. 6. Cualquier orden mixto de las instancias dará un conjunto diferente de segmentos de paso dentro del rectángulo formado por estos dos extremos. Sin embargo, la curva ROC debe representar el rendimiento esperado del clasificador, que, a falta de otra información, es el promedio de los segmentos pesimista y optimista. Este promedio es la diagonal del rectángulo y se puede crear en el algoritmo de la curva ROC al no emitir un punto ROC hasta que se hayan procesado todas las instancias de valores f iguales. Esto es lo que logran la variable f_{prev} y la instrucción **if** de la línea 7.

Como ya se mencionó, los gráficos ROC permiten visualizar y organizar el rendimiento del clasificador sin tener en cuenta las distribuciones de clase o los

costos de error. Esta habilidad se vuelve muy importante cuando se investiga el aprendizaje con distribuciones sesgadas o aprendizaje sensible a los costos. Un investigador puede graficar el desempeño de un conjunto de clasificadores, y ese gráfico permanecerá invariable con respecto a las condiciones de operación (sesgo de clase y costos de error). A medida que cambian estas condiciones, la región de interés puede cambiar, pero el gráfico en sí no lo hará.

Provost & Fawcett [12] muestran que un conjunto de condiciones operativas puede transformarse fácilmente en una llamada línea de rendimiento iso (iso-performance line) en el espacio ROC. Dos puntos en el espacio ROC, (FP_1, TP_1) y (FP_2, TP_2), tienen el mismo rendimiento si

$$\frac{TP_2 - TP_1}{FP_2 - FP_1} = \frac{c(Y,n)\mathbb{P}(n)}{c(N,p)\mathbb{P}(p)} = m.$$

Esta ecuación define la pendiente de una línea de rendimiento iso. Todos los clasificadores correspondientes a puntos en una línea de pendiente m tienen el mismo costo esperado. Cada conjunto de distribuciones de clases y costos define una familia de líneas de desempeño iso. Las líneas "más al noroeste" (que tienen una intercepto TP más grande) son mejores porque corresponden a clasificadores con un costo esperado más bajo. De manera más general, un clasificador es potencialmente óptimo si y solo si se encuentra en el casco convexo del conjunto de puntos en el espacio ROC. La envolvente convexa del conjunto de puntos en el espacio ROC se denomina envolvente convexa ROC (ROC convex hull, ROCCH) del correspondiente conjunto de clasificadores.

Esta formulación de ROCCH tiene una serie de implicaciones útiles. Dado que solo los clasificadores en la envolvente convexa son potencialmente óptimos, no es necesario retener otros. Las condiciones de funcionamiento del clasificador pueden traducirse en una línea de rendimiento iso, que a su vez puede utilizarse para identificar una parte del ROCCH. A medida que cambian las condiciones, el casco mismo no cambia; sólo la parte de interés lo hará.

2.6. Hallazgos de la literatura acerca del desempeño de modelos de calificación crediticia

El desarrollo de modelos analíticos precisos de calificación crediticia se ha convertido en un enfoque importante para las instituciones financieras. Para ello, se han propuesto numerosos algoritmos de clasificación para la calificación crediticia. Es por ello que en esta sección se comentara acerca de hallazgos acerca del desempeño de algunos modelos de clasificación crediticia encontrados en algunos

artículos de investigación. En particular registramos la jerarquización de modelos acorde a su desempeño de ajuste.

Gunnarsson, et. al. [6] mencionan que los bosques aleatorios, debe considerarse como un método de referencia para la calificación crediticia. Más recientemente, se han propuesto métodos de boosting (ver Capítulos 10, 15 y 16 de [7]) para la puntuación crediticia, donde se ha demostrado que supera a los bosques aleatorios en algunos casos. Sin embargo, la investigación sobre algoritmos de clasificación para la calificación crediticia ha ignorado en gran medida el desarrollo de arquitecturas de aprendizaje profundo (ver Capítulo 11 de [7]) . También menciona que lo anterior requiere una mayor actualización de la investigación al considerar algoritmos de aprendizaje profundo para la calificación crediticia. Con este fin, Gunnarsson construye dos arquitecturas de aprendizaje profundo, a saber, redes de creencias profundas y redes de perceptrones multicapa, y se los compara con métodos convencionales para calificación crediticia, regresión logística y árboles de decisión, y dos métodos conjuntos para calificación crediticia, bosques aleatorios y XGBoost. Los diferentes clasificadores se compararon en base a cuatro indicadores de rendimiento sobre diez conjuntos de datos. Por último, los procedimientos de prueba estadística bayesiana se introdujeron en el contexto de la calificación crediticia y se compararon con los métodos NHST frecuentistas, que según Gunnarsson, tradicionalmente se han considerado las mejores prácticas en la calificación crediticia. Esta comparación destacó los muchos beneficios de los procedimientos de prueba estadísticos bayesianos y aseguró hallazgos empíricos.

Principalmente se obtuvieron dos conclusiones de la comparación de los diferentes clasificadores. En primer lugar, XGBoost es el mejor clasificador general de clasificación de todos los clasificadores considerados en [6] y es el clasificador con mejor rendimiento en función de todas las medidas de rendimiento consideradas. En segundo lugar, las redes profundas con varias capas ocultas, es decir, el aprendizaje profundo, no superan a las redes menos profundas con una capa oculta. También se debe tener en cuenta en esta comparación que las redes profundas tienen un costo computacional mucho mayor que los otros clasificadores que se consideraron, ya que la cantidad de modelos que se necesita construir para ajustar adecuadamente los hiperparámetros de los modelos crece exponencialmente con el número de capas ocultas. Por lo tanto, se concluye que los algoritmos de aprendizaje profundo no parecen ser métodos apropiados para la calificación crediticia y que, en general, se debe preferir un método conjunto, XGBoost, a los otros métodos de calificación crediticia considerados cuando el rendimiento de la clasificación es el objetivo principal de las actividades de calificación crediticia.

Además, se menciona que una explicación plausible de por qué las redes profundas no superan a los otros métodos considerados es probablemente proporcionada por el hecho de que se ha demostrado que el aprendizaje profundo es muy bueno para descubrir estructuras complejas, dado que hay muchas instan-

cias disponibles para aprender, lo que podría no ser el caso para la mayoría de los conjuntos de datos de riesgo crediticio.

También se debe tener en cuenta que los modelos de conjuntos son los llamados modelos de "caja negra", lo que indica que es difícil interpretar por qué estos modelos alcanzan un determinado resultado o hacen una determinada predicción. Si la interpretabilidad de la predicción del modelo es la principal preocupación, es posible que se desee recurrir a los métodos convencionales para la calificación crediticia, por ejemplo, regresión logística. Sin embargo, si el rendimiento predictivo es el enfoque principal de la construcción del modelo, XGBoost parece ser la mejor opción en general. Además, señala que incluso cuando la interpretabilidad es una condición previa para las actividades de calificación crediticia, los modelos de caja negra, que han demostrado un buen desempeño para la calificación crediticia, siguen siendo una valiosa herramienta de evaluación comparativa. Esto se debe al hecho de que estos modelos se pueden utilizar para identificar importantes efectos no lineales y/o interacciones en conjuntos de datos de calificación crediticia.

Gunnarsson llega a estos resultados incluyendo una cantidad considerable de conjuntos de datos de la vida real con una configuración empírica. Estos conjuntos de datos eran bastante variados tanto en términos del número de observaciones como de entradas utilizadas, por lo que afirma que la investigación brinda una buena indicación del desempeño general de los clasificadores considerados para la calificación crediticia.

Por su parte Louzada, et. al. [10] analizan y clasifican 187 artículos sobre credit scoring publicados en revistas científicas durante el periodo de 1992 a 2015. Allí observaron que independientemente del período de tiempo, el objetivo principal más común de los artículos revisados es proponer un nuevo método para la calificación crediticia, especialmente con técnicas híbridas, y se observó una similitud entre el desempeño predictivo de los métodos. Además, la comparación con las técnicas tradicionales rara vez se realizó en períodos de tiempo recientes al artículo (es decir en el año 2016). Este hecho demuestra que, aunque los investigadores están renunciando a comparar técnicas, continúa la búsqueda de un método general con un alto rendimiento predictivo.

Con base en sus resultados Louzada menciona que, por lo menos en ese momento, las redes neuronales, la máquina de soporte vectorial, las técnicas híbridas y combinadas aparecen como las herramientas principales más comunes. La regresión logística, los árboles y también las redes neuronales se utilizan mayoritariamente en las comparaciones de técnicas como estándares a superar. En general, la máquina de soporte vectorial aparece como un método de alto rendimiento predictivo y bajo esfuerzo computacional que otros métodos. También se menciona que con respecto a los conjuntos de datos para la calificación crediticia, el número había ido en aumento, así como la presencia de una mezcla de

variables continuas y discretas. Sin embargo, la mayoría de los conjuntos de datos son privados y existe un amplio uso de los conocidos conjuntos de datos alemanes y australianos. Este hecho muestra cuán difícil es obtener conjuntos de datos en el escenario de calificación crediticia, ya que existen problemas relacionados con el mantenimiento de la confidencialidad de las bases de datos de calificación crediticia. *La anterior consideración es de relevancia al material presentado mas adelante en el Capítulo 5 en donde se reportan los resultados de ajustar modelos con datos reales de una empresa privada.*

Otro punto que menciona Louzada es que aunque la revisión sistemática de la literatura es exhaustiva, aún persisten algunas limitaciones. Primero, los hallazgos se basaron en artículos publicados en inglés y en revistas científicas dentro de las siguientes bases de datos: Sciencedirect, Engineering Information, Reaxys y Scopus. Aunque tales bases de datos cubren más de 20.000 títulos de revistas, en lo sucesivo se pueden incluir otras bases de datos en la encuesta. En segundo lugar, no se incluyen en la encuesta otras formas de publicación, como documentos de trabajo inéditos, tesis de maestría y doctorado, libros, conferencias en actas, libros blancos y otros. A pesar de estas limitaciones, la revisión sistemática proporciona información importante sobre la literatura de investigación sobre las técnicas de clasificación aplicadas a la calificación crediticia y cómo esta área se ha movido con el tiempo.

Finalmente, Baesens, et al. [1], en una actualización de su artículo de 2003 se propone explorar la efectividad relativa de algoritmos de clasificación alternativos en la calificación crediticia minorista. Con ese fin, compararon 41 clasificadores en términos de seis medidas de rendimiento en ocho conjuntos de datos de calificación crediticia del mundo real. Los resultados sugieren que varios clasificadores predicen el riesgo de crédito con mucha más precisión que la regresión logistica estándar de la industria. Los clasificadores de conjuntos especialmente heterogéneos funcionan bien. También se proporciona alguna evidencia de que las scorecards más precisas facilitan rendimientos financieros considerables. Además, se muestra que varias medidas comunes de desempeño dan señales similares sobre qué scorecard es más efectiva, y recomiendan el uso de dos medidas raramente empleadas que aportan información adicional.

Baensens et. al. [1] encontró que las redes neuronales artificiales funcionan mejor que las máquinas de aprendizaje extremo (ver [8]), bosques aleatorios mejor que bosques de rotación y los conjuntos selectivos dinámicos son peores que casi todos los demás clasificadores. Esto puede indicar que el progreso en el campo se ha estancado, y que el foco de atención debe pasar de los modelos de probabilidad de incumplimiento a otros problemas de modelado en la industria crediticia, incluida la calidad de los datos, la recalibración de la scorecard, la selección de variables y el modelado de la pérdida dado el incumplimiento/exposición al incumplimiento.

También Baensens recomienda bosques aleatorios como punto de referencia contra el cual comparar nuevos algoritmos de clasificación. Afirman que HCES-Bag puede ser aún más difícil de superar, pero no está tan fácilmente disponible en el software estándar. Además, advierten contra la práctica de comparar un clasificador recién propuesto con regresión logística (o algún otro clasificador individual) únicamente. Regresión logística es el estándar de la industria y es útil examinar cómo se compara un nuevo clasificador con este enfoque.

Se comenta que desde una perspectiva de gestión, es importante razonar si el rendimiento superior que se observó para algunos clasificadores se generaliza a las aplicaciones del mundo real y en qué medida su adopción aumentaría los rendimientos. A partir de esto destacan algunos puntos.

Primero, muestran que los avances en el poder de las computadoras, el aprendizaje de clasificadores y las pruebas estadísticas facilitan las comparaciones rigurosas de clasificadores. Esto no garantiza la validez externa. Varias preocupaciones sobre por qué los experimentos de laboratorio pueden sobrestimar la ventaja de los clasificadores avanzados siguen siendo válidas. Sin embargo, los diseños experimentales con varias repeticiones de validación cruzada, diferentes medidas de desempeño y procedimientos apropiados de comparación múltiple superan algunas limitaciones de estudios previos y, por lo tanto, brindan un apoyo más sólido de que los clasificadores avanzados tienen el potencial de aumentar la precisión predictiva no solo en el laboratorio sino también en la industria.

En segundo lugar, los resultados facilitan algunos comentarios relacionados con la aceptación organizacional de clasificadores avanzados. En particular, la falta de aceptación puede deberse a la preocupación de que se necesita mucha experiencia para manejar dichos clasificadores. Sin embargo, mostraron que no era el caso. Las diferencias de rendimiento observadas son el resultado de un enfoque de modelado totalmente automático. En consecuencia, algunos clasificadores avanzados están bien preparados para predecir de manera significativamente más precisa que las alternativas más simples sin intervención manual. Además, el interés actual en Big Data y conceptos relacionados indica un cambio hacia un paradigma de toma de decisiones basado en datos entre los gerentes. Esto podría aumentar aún más la aceptabilidad de los métodos de puntuación avanzados.

Tercero, el valor comercial de las predicciones más precisas de la scorecard es un tema crucial. La simulación preliminar proporcionó alguna evidencia de que la ecuación "mayor precisión (estadística) es igual a más ganancias" podría sostenerse. Además, las scorecards minoristas respaldan una gran cantidad de decisiones comerciales. Esto lo sostienen a partir del ejemplo de la industria de las tarjetas de crédito o las tareas de calificación en entornos en línea (préstamos entre pares, ofertas de planes de pago en el comercio electrónico, etc.). En tales entornos, las inversiones únicas (por ejemplo, para hardware, software y capacitación de usuarios) en una técnica de puntuación más elaborada darán sus

frutos a largo plazo cuando las mejoras de precisión pequeñas pero significativas se multipliquen por cientos de miles de aplicaciones de scorecards. Las dificultades de introducir métodos de puntuación avanzados, incluidos los modelos de conjuntos, son más psicológicas que comerciales. El uso de una gran cantidad de modelos, una minoría significativa de los cuales da respuestas contradictorias, es contradictorio para muchos líderes empresariales. Dichas organizaciones deberán experimentar completamente antes de aceptar un cambio de los procedimientos estándar históricos de la industria.

Los marcos regulatorios y la aceptación organizacional restringen y, en ocasiones, prohíben el uso de técnicas de puntuación avanzadas en la actualidad; al menos para los productos de crédito clásicos. Sin embargo, considerando el interés actual en las ayudas para la toma de decisiones centradas en los datos y la riqueza de las formas de otorgamiento de crédito mediadas en línea, prevén un futuro brillante para los métodos de calificación avanzada en la calificación crediticia.

Capítulo 3

El proceso de otorgar y administrar un crédito

En este capítulo se presenta una breve exposición del proceso de otorgar y administrar un crédito. Lo anterior ayudará a comprender mejor el contexto de la calificación crediticia y en donde se sitúa en todo este proceso. Se hablará también de como es que los modelos de clasificación se han convertido en una herramienta muy importante para las instituciones financieras. Es necesario señalar que las principales fuentes para este capítulo son Siddiqi [13] y USAID [14].

Es pertinente hacer algunas aclaraciones con respecto a la referencia USAID [14]. Primero, esta referencia esta auspiciada por la organización US-AID. Esta referencia es un reporte generado bajo la sombrilla del proyecto "USAID-Funded Economic Governance II Project" y en la que el banco central de Iraq aparece como autor. Segundo, los créditos considerados en el reporte están dirigidos a las empresas, es decir, se consideran créditos empresariales. Además el reporte se enfoca en un proceso manual en el paso de asignar una calificación crediticia. Por supuesto, estas son diferencias fundamentales al contexto de la presente tesis ya que aquí estamos interesados en créditos a las personas y en generar un sistema que automatice la asignación de calificación crediticia. No obstante, el ejercicio de señalar analogías y diferencias es el estilo de exposición del presente capítulo.

3.1. Las ventajas de tarjetas de calificación internas

Siddiqi [13] señala algunas de los circunstancias que contribuyeron a la popularidad de los modelos automatizados de calificación crediticia. Primero, el aumento de la competencia ha llevado a las instituciones financieras y de concesión de crédito a buscar formas más eficaces de atraer nuevos clientes solventes y al mismo tiempo, a controlar las pérdidas. En la evolución de las instituciones financieras se ven esfuerzos significativos de marketing para captar mas clientes. Resultado que ha generado la necesidad de procesar ágilmente las solicitudes recibidas. Surge así el requerimiento de procesos automáticos de adjudicación y solicitud de crédito. Los sistemas que resuelvan dicha automatización deben ser eficientes: deben ser capaces de minimizar la denegación de crédito a clientes solventes, al mismo tiempo que excluye a tantas solicitudes como sea posible que sean potencialmente morosas.

En el pasado las instituciones financieras solventaban dicha necesidad mediante la adquisición de scorecards de riesgo crediticio de un conjunto muy pequeño de proveedores de riesgo. Por ejemplo, es popular el nombre de Fair & Isaac Company (FICO). Lo anterior quiere decir que para el desarrollo de las scorecards predictivas, las instituciones financieras proporcionaban sus datos a los proveedores y ellos generaban los modelos. No obstante, la tendencia de la industria ha sido moverse hacia el desarrollo interno de scorecards. Siddiqi [13] menciona que un factor importante en dicha evolución es el desarrollo tecnológico:

- Disponibilidad de software que permitía a los usuarios desarrollar scorecards sin grandes inversiones en capital humano capacitado y en infraestructura avanzada. La existencia de dicho software provocó que algunas funciones complejas de minería de datos estuvieran disponibles al alcance de un clic.

- Avances en el almacenamiento de datos inteligente y de fácil acceso han eliminado gran parte de la carga de recopilar los datos necesarios y ponerlos en una forma que sea adecuada para el análisis.

Debido a que todas estas herramientas se volvieron accesibles, el desarrollo interno se convirtió en una opción viable para muchas instituciones pequeñas y medianas. Las instituciones financieras comenzaron a visualizar las ventajas de desarrollar scorecards internas:

- Recompensas significativas (medidas con el retorno de la inversión, Return on Investment ROI) que el desarrollo interno de las scorecards podría ofrecer a aquellos que pudieran crearlas adecuadamente.

- Desarrollo más rápido, más barato y con mucha más flexibilidad que antes.
- Mantenimiento más económico, ya que el costo de mantener una capacidad de calificación crediticia interna es menor en comparación al costo con un proveedor externo.
- Posibilidad de desarrollar muchas más scorecards (con segmentación mejorada) pero sin necesidad de gastar más.
- Además, las empresas se dieron cuenta de que al desarrollar sus propias scorecards de manera interna, los involucrados conocían mucho mejor los datos al ser propios de la institución. En particular los conocimientos de negocio eran superiores lo cual llevó a desarrollar scorecards de mejor rendimiento.
- Otro aspecto que también se observó fue que dar las definiciones de desempeño de la población es una parte crítica de la construcción del sistema de puntuación, y la capacidad de variar las definiciones para diferentes propósitos es un punto crucial.

Un ejemplo de este último punto, un puntaje de probabilidad de incumplimiento diseñado para propósitos de planificación de capital puede excluir cuentas moderadamente morosas (60 días vencidas dos veces durante los últimos 24 meses) que normalmente se incluyen en "mal comportamiento" y se rigen por la definición de Basilea para préstamos considerados con probabilidad de incumplimiento.

3.2. El proceso de crédito

Esta sección está dedicada a la descripción de los componentes del proceso de originación y monitoreo del crédito. Cada uno de los componentes se discutirá con cierto detalle con el propósito de documentar la posición de la calificación crediticia en todo este proceso.

Antes de comenzar con la descripción del proceso se definen a los involucrados en el proceso:

- El prestamista (lender) es la persona o entidad que facilita una determinada cantidad de dinero, en forma de crédito o préstamo, con el compromiso de la otra parte (o prestatario) de que este será devuelto, junto a los intereses, según las condiciones acordadas por contrato.
- El prestatario (borrower) es el solicitante, y por tanto receptor de la cantidad de dinero entregada por el prestamista.

- Los oficiales de crédito (credit officer) evalúan los datos crediticios para determinar todos los riesgos potenciales involucrados en la otorgación de crédito.
- Los examinadores (examiners) son auditores que revisan y monitorean las operaciones de las instituciones financieras. Trabajan en ubicaciones financieras que incluyen un banco, una asociación de ahorro y préstamo o una cooperativa de crédito. Revisan los procedimientos y las actividades de la institución financiera para garantizar que funcione de conformidad con todas las normas y reglamentaciones gubernamentales pertinentes y necesarias

3.2.1. El Proceso de Iniciación y Análisis de Crédito

Previo a que llegue una solicitud de crédito, es indispensable que la institución en consideración defina la política de negocio, o más exactamente, la política de crédito. Esta debe contener una colección de lineamientos que defina: (a) tipos de préstamos aceptables para la institución, (b) propósitos de los préstamos, (c) el plazo, (d) la garantía, (e) la estructura y (f) las garantías aceptables para respaldar un préstamo.

Partiendo de que la política de crédito ya ha sido definida, el proceso de crédito comienza con un análisis exhaustivo de la solvencia del prestatario. En [14] se comenta que se debe realizar una evaluación que incorpore los siguientes elementos:

- La condición financiera actual y esperada del prestatario.
- La capacidad del prestatario para soportar condiciones adversas o "estrés".
- El historial crediticio del prestatario y una correlación positiva entre la capacidad de pago histórica y proyectada.
- La estructura óptima del préstamo, incluida la amortización del préstamo, convenios, y requisitos de información.
- Garantías prendadas por el prestatario: cantidad, calidad y liquidez.
- Factores cualitativos, como la gestión, la industria y el estado de la economía.

Los anteriores puntos son formulados en [14] para créditos empresariales. No obstante, todos los puntos anteriores se pueden traducir al contexto de créditos a las personas. Por ejemplo, con respecto a los primeros tres puntos, generalmente las instituciones que otorgan un crédito a las personas requieren conocer la situación

laboral del individuo que solicita un crédito. Particularmente se requiere un comprobante de ingresos, en donde ya se obtiene varia información del solicitante: profesión, nivel de ingreso, y lugar de trabajo. Permite también estimar la estabilidad del trabajo y así hacer alguna apreciación con respecto al segundo punto. Con respecto al tercer punto es usual que la institución que gestiona un préstamo solicite información de historial crediticia a una oficina especializada, por ejemplo a la empresa conocida como Buró de Crédito. Para el cuarto punto con respecto a la estructura del préstamo, dependerá mucho del tipo de crédito. Por ejemplo, si es un crédito hipotecario, hay opciones a definir como son el enganche y plazo. Estos vendrán determinados por parámetros que define la institución financiera dejando algunas elecciones al solicitante.

De acuerdo a lo expuesto en [14], el análisis de la solvencia comienza con la recopilación, el análisis y la evaluación de la información requerida. Continua dicho análisis con la decisión de etiquetar el tipo de riesgo de la solicitud. En el caso de un riesgo aceptable, el oficial de crédito debe proponer una estructura de préstamo diseñada de tal forma que toma en cuenta las fortalezas del cliente además de proteger a la institución contra las debilidades identificadas del prestatario. El análisis finaliza con la **determinación de una calificación de riesgo** para la aprobación (o rechazo) del crédito y préstamo. Aparentemente la narrativa de [14] sugiere un proceso manual en el análisis. Sin embargo, es precisamente este proceso de originación el que interesa automatizar mediante un modelo matemático que genere objetivamente y expeditamente una calificación de riesgo. Hay varias consideraciones en este punto que generan matices:

- El monto del crédito y el volumen de la cartera. Cuando los montos son significativos en una cartera con pocos créditos, es factible y deseable un proceso manual de evaluación. En esta situación se puede prescindir de una scorecard, o bien tiene un papel complementario.

- En una cartera granular con muchos créditos de muy bajo monto, la automatización es deseable y la scorecard puede ser la herramienta definitiva de decisión.

- Puede darse una situación híbrida, en la que una solicitud recibe una calificación por una scorecard en una banda indeterminada: No es un “buen riesgo” pero tampoco es decididamente malo. Entonces se puede redirigir a un experto que haga una evaluación principalmente manual.

3.2.2. Estructuración del crédito

De acuerdo a [14], la estructuración del crédito (credit underwritting) es el proceso que emprenden las instituciones crediticias para estructurar una línea de crédito para minimizar los riesgos y generar el mejor rendimiento, dados los riesgos que asumen las instituciones. La estructura crediticia incluye la definición de: (a) el plazo del préstamo, (b) la garantía requerida, (c) el requisito de amortización, (d) programación de los pagos de interés y (e) los requisitos de información.

La estructuración debe incluir protecciones que ayuden a mitigar los riesgos y ayuden a aumentar la probabilidad de pago del préstamo. En los créditos empresariales, dichas protecciones incluyen la verificación del flujo de efectivo para cumplir con los requisitos de servicio del préstamo, clausulas adecuadas (covenants) para preservar las fortalezas del prestatario y limitar las debilidades, y el requisito de garantías suficientes y verificadas para proporcionar una fuente de pago alternativo. En el caso de crédito a las personas, como es el tipo de créditos considerado en la tesis, solo en algunos propósitos muy particulares, objetivo del crédito, se cuentan con algunos de estos elementos. Por ejemplo, en el caso de un crédito hipotecario, o de un crédito para la adquisición de un automovil, es el mismo objeto que queda como garantía. En ambos casos, el solicitante debe solventar el costo de una aseguranza que cubra varios conceptos: desempleo, siniestro, liquidación en caso de declararse la bancarota.

La estructuración también incluye especificar los informes requeridos del prestatario durante la vida del préstamo. Cuanto mayor sea el riesgo identificado en el crédito, más información se requerirá y mayor será la frecuencia de la información. Por ejemplo, es posible que se requieran estados financieros mensuales en lugar de trimestrales.

Este informe constituye la base del seguimiento del préstamo posterior al desembolso. En el caso de créditos a las personas no existe este requerimiento. Sin embargo, hay algunos casos en que la institución financiera podrá acceder a esta clase de información. Particularmente cuando la institución financiera es un banco que otorga el crédito y ofrece (incluso de manera vinculante) el adquirir otros servicios como es la administración de cuenta de nomina o una tarjeta de crédito.

El desembolso del préstamo debe ocurrir una vez que todos los documentos requeridos hayan sido firmados y entregados a la institución. Los documentos del préstamo constituyen la protección primaria de la institución una vez que se ha desembolsado el préstamo. El contrato de préstamo, un documento legal que vincula a ambas partes, es el documento clave para el prestamista. Debe estar diseñado para “controlar” al prestatario y contener protecciones para la institución.

El incumplimiento de las clausulas en un convenio de préstamo generalmente desencadena un incumplimiento del préstamo. Sin embargo, antes de llegar a una cancelación (charge-off) la mayoría de las instituciones prefieren negociar una exención del incumplimiento basada en una tarifa siempre que el reembolso final del préstamo esté relativamente asegurado.

3.2.3. Administración de la cartera de crédito

Una vez que el crédito ha sito otorgado, debe darse seguimiento al mismo. El propósito del seguimiento de préstamos es identificar lo antes posible cualquier cambio en la condición financiera o el desempeño del prestatario que afecte, o pueda afectar, la capacidad del prestatario para pagar los préstamos pendientes a la institución según lo acordado. Como se señaló anteriormente, el contrato de préstamo estipula las herramientas de seguimiento; [14].

En los creditos empresariales, el prestamista debe monitorear regularmente y activamente las fortalezas del prestatario y, en particular, las debilidades identificadas durante el proceso de estructuración. La institución/prestamista debe prestar especial atención a la preservación de las dos fuentes de reembolso del préstamo: el flujo de efectivo y la garantía. Cuanto mayores sean las debilidades identificadas, más frecuente será el seguimiento. Por ejemplo un deterioro en la calidad del crédito por la identificación de debilidades nuevas requieren un monitoreo más frecuente. En general, cuanto mayor es el riesgo crediticio, con más frecuencia la institución financiera debe requerir información. Además de este flujo regular de información, el prestamista debe estar en comunicación con el prestatario para realizar un seguimiento del desempeño del prestatario. El archivo de crédito debe contener evidencia de monitoreo por parte del oficial de crédito como por ejemplo llamadas telefónicas, informes financieros intermedios y estados financieros anuales. Lo anterior en el caso de créditos personales se limitará, como ya se mencionó, a monitorear otras cuentas del solicitante dentro de la institución y también al comportamiento histórico en el pago del crédito, lo que llevá al concepto de scorecard de comportamiento (behavioral scoring).

La principal herramienta de seguimiento, el contrato de préstamo, debe estipular los requisitos de información que debe proporcionar el prestatario: tipo de información y frecuencia de presentación. El acuerdo de préstamo también debe contener convenios que el prestatario debe observar durante la vida del préstamo. Dichos convenios pueden incluir requisitos para observar ciertos índices, como apalancamiento y liquidez, en todo momento. También pueden incluir ciertas prohibiciones, como préstamos realizados por parte del prestatario o compra de equipos sin el consentimiento por escrito del prestamista. Tales requerimientos no aplican por lo general al crédito personal.

En el caso de que se identifique algún deterioro o cambio negativo en el prestatario, la institución/prestamista debe determinar si estos cambios son lo suficientemente importantes como para afectar la capacidad de pago y, por lo tanto, efectuar un cambio en la calificación de riesgo. Un cambio importante puede incluso representar un incumplimiento del préstamo, una violación del contrato de crédito (denominada cláusula MAC o Material Adverse Change), lo que le daría a la institución otorgante la oportunidad de reestructurar el crédito.

Un cambio en la calificación de riesgo puede desencadenar otra acciones en el interior de la institución, tales como un aumento en la reserva para pérdidas crediticias. Otro tipo de acciones que pude causar es la inclusión de demandas de garantías adicionales, un aumento en la tasa de interés e incluso la demanda de reembolso inmediato del préstamo, "llamar al préstamo" (calling the loan).

3.2.4. Reestructuración de crédito (créditos problemáticos)

De acuerdo a [14], la reestructuración del crédito cuando el comportamiento de pago o la calidad del crédito comienza a deteriorarse debe ser estudiado de antemano y el proceso a seguir debe ser especificado en la política de crédito de la institución.

Debe existir un mecanismo para la identificación formal de un préstamo problemático, como una "lista de vigilancia" y un comité de "lista de vigilancia" de créditos problemáticos, ver [14]. Este mecanismo es parte del proceso de informar a la alta dirección sobre el problema, para que se pueda tomar una decisión sobre la acción de la institución lo antes posible. Generalmente, cuando un crédito se coloca en la "lista de vigilancia", recibe una reducción en la calificación de riesgo, además de que aumenta la reserva porque la capacidad de pago se ha visto afectada o está a punto de verse afectada. Un comité de "lista de vigilancia" se debe de reunir regularmente para monitorear el progreso en la gestión de los préstamos de la "lista de vigilancia".

Cuando se coloca un préstamo en la lista de vigilancia, se debe acordar un plan de acción internamente y con el prestatario lo antes posible. Tomando en cuenta que tan grave es el deterioro, este plan deberá incluir una estrategia de rehabilitación o una estrategia de salida, según la determinación de la institución de varios factores, tales como ([14]):

- La probabilidad de éxito de la estrategia elegida,
- El nivel de cooperación que se espera del prestatario,

- Gastos que probablemente se incurrirán en la implementación de la estrategia
- El valor presente de la recuperación esperada (si la recuperación requerirá una cantidad significativa de tiempo).

Los préstamos problemáticos se desarrollan cuando aparecen debilidades crediticias que afectan, o afectarán pronto, la capacidad de pago del prestatario. Las renegociaciones de préstamos exitosas dependen de la identificación temprana de debilidades crediticias y tendencias crediticias adversas. Es por ello que se requiere tener un seguimiento constante de los préstamos con la finalidad de identificar cualquier deterioro en la solvencia del prestatario que pueda causar una reducción en la calificación de riesgo. Es importante que la alta dirección sea informada con prontitud del deterioro ya que la acción inmediata es crucial para una gestión y una recuperación exitosas.

Capítulo 4

Preparación de los datos para modelos de calificación crediticia

La calificación de riesgo es una herramienta utilizada para evaluar el nivel de riesgo asociado con los solicitantes de un crédito, o incluso de clientes que ya forman parte de la cartera; ver [13]. Por ejemplo, en el modelo logístico se estima la probabilidad de que un solicitante con una puntuación dada sea buen pagador. La definición de que un cliente sea "bueno" o "malo", requiere por si mismo un estudio minucioso para que sean compatibles por un lado, las necesidades de negocio de la empresa, y por el otro, la información disponible.

En este capítulo nos concentramos en la tarea de presentar los principios para procesar la información disponible. El texto del capítulo se basa en la referencia Siddiqi [13]. El propósito es presentar la preparación necesaria que debe hacerse en los "datos brutos" para obtener datos limpios que puedan ser utilizados en la estimación de los modelos enlistados en la introducción, a saber, modelo logístico, LDA, SVM, y bosques aleatorios. Para hablar de manera genérica de un modelo de calificación crediticia utilizaremos el término **tarjeta de calificación** (en inglés scorecard). No debe causar confusión a pesar de que en muchas referencias se utilice para un modelo en particular, como es el caso de la referencia [13].

El capítulo está organizado de la siguiente forma. En la Sección 4.1 de consideraciones generales se habla de: Las variables, la comparación entre modelos para la originación y del comportamiento, así como el uso a los modelos de calificación.

La Sección 4.2 presenta consideraciones preliminares para la creación de un sistema de calificación de crédito, particularmente señala la importancia de la planificación inicial y del aseguramiento de datos de entrada. La Sección 4.3 es la más importante del capítulo y como su título lo indica, presenta consideraciones cruciales para preparar los datos. En nuestro caso, es particularmente importante el así llamado análisis de tasa de balanceo (roll rate) ya que será la guía para una

correcta definición de la variable a explicar.

4.1. Consideraciones generales en la calificación crediticia

En su forma más simple, una tarjeta de calificación consta de un grupo de variables (el término que se acostumbra a usar en la literatura de credit scoring es característica, aquí las usaremos como sinónimos privilegiando la primera sobre la segunda), las cuales se usan para predecir y separar las solicitudes buenas de las malas.

Las variables que se usan en los **modelos de calificación crediticia de originación** se deben seleccionar de las fuentes de datos disponibles para el prestamista en el momento de la solicitud. Ejemplos de tales variables son:

- datos demográficos (edad, tiempo de residencia, tiempo en el trabajo, código postal),
- relación existente (tiempo en el banco, número de productos, rendimiento de pago, reclamos anteriores),
- buró de crédito (consultas, oficios, morosidad, registros públicos),
- datos inmobiliarios, etc.

Por ejemplo, en las tarjetas de calificación a cada valor (también llamada atributo) que toma la variable (por ejemplo en la variable "Edad" el valor "23–25" es un atributo) se le asignan puntos en función de análisis estadísticos, teniendo en cuenta varios factores, como la correlación entre las características y los factores operativos. La puntuación total de un solicitante es la suma de las puntuaciones de cada atributo presente en la tarjeta de calificación de ese solicitante; ver [13]. Lo anterior requiere la estimación de un modelo.

La puntuación de las solicitudes sirven para la toma de decisión. En particular, pueden ser una herramienta para establecer políticas de "auditoría" (due diligence). Por ejemplo, un solicitante con una puntuación muy alta puede ser aprobado directamente sin obtener más información sobre bienes inmuebles, verificación de ingresos o análisis del valor subyacente. Recíprocamente, cuando la puntuación es muy baja puede rechazarse sin mayor tramite.

La calificación de riesgo se usa de manera similar con los clientes existentes en la cartera de crédito. En este caso, los datos de comportamiento del cliente se utilizan para predecir la probabilidad de comportamiento negativo a través de

la generación de un **modelo de calificación crediticia de comportamiento** (tarjeta de calificación de comportamiento). Diferentes consideraciones de negocio, tales como niveles de riesgo y rentabilidad esperados, pueden determinar diferentes tratamientos a las cuentas. De acuerdo a Siddiqi [13] estas pueden ser:

- Ofrecer actualizaciones de productos o productos adicionales
- Aumento de los límites de crédito en tarjetas de crédito y líneas de crédito
- Permitir que algunos clientes de crédito rotativo vayan más allá de sus límites de crédito
- Marcar transacciones potencialmente fraudulentas
- Ofrecer mejores precios en las renovaciones de pólizas de seguros/préstamos
- Decidir si volver a emitir o no una tarjeta de crédito vencida
- Precalificación de listas de marketing directo para venta cruzada
- Dirigir las cuentas morosas a métodos de cobro más estrictos o subcontratar a una agencia de cobro
- Suspender o revocar servicios telefónicos o facilidades de crédito
- Poner una cuenta en una "lista de vigilancia" para posibles actividades fraudulentas

De acuerdo a [13], la calificación de riesgo al ser una herramienta para evaluar los niveles de riesgo, también se aplica en otras áreas operativas tales como:

- Simplificar el proceso de toma de decisiones, es decir, las solicitudes de mayor riesgo y límite se entregan al personal con más experiencia para un mayor escrutinio, mientras que las solicitudes de bajo riesgo se asignan al personal subalterno. Esto se puede hacer en sucursales, centros de adjudicación de crédito y departamentos de cobranza.
- Reducción del tiempo de respuesta para el procesamiento de solicitudes a través de la toma de decisiones automatizada.
- Fijación de la asignación de capital económico y regulatorio.
- Fijación de precios para la bursatilización de carteras de cuentas por cobrar.
- Comparar la calidad del negocio de diferentes canales/regiones/proveedores.

4.2. La creación de tarjetas de calificación de riesgo crediticio

El proceso de desarrollo de una tarjeta de calificación involucra diferentes departamentos de la empresa, tales como: Análisis de riesgo, tecnología de la información (TI) incluyendo dataware house, el área de negocio. Esto crea mejores tarjetas de calificación desde el punto de vista del pronóstico, incorpora necesidades de negocio, y además da viabilidad desde el punto de vista de la información presente y sobre todo, de la disponibilidad de la información futura. Además ayuda a la transferencia de conocimiento para el propio uso de la herramienta. De acuerdo a [13], se ha observado que el desarrollo de tarjetas de calificación de forma aislada puede generar problemas como la inclusión de variables cuyos datos ya no se recopilan, que son legalmente sospechosas (por ejemplo en algunas regulaciones es ilegal el uso de la variable "raza") o difíciles de recopilar operativamente (por ejemplo la variable "ingreso" siempre es problemática).

En el proceso de creación de una tarjeta de calificación, Siddiqi [13] recomienda considerar entre otros, los siguientes pasos:

- Planificación
- Revisión de los datos y de los parámetros del proyecto
- Creación de la base de datos
- Desarrollo de la tarjeta de calificación
- Informes de gestión de la tarjeta de calificación.

A continuación en los siguientes apartados exponemos un resumen de cada uno de los anteriores puntos.

4.2.1. Planificación

El desarrollo de una tarjeta de calificación requiere una planificación adecuada antes de que pueda comenzar cualquier trabajo de modelación. Esto incluye identificar el objetivo del proyecto, identificar a los participantes clave en el desarrollo e implementación de la tarjeta de calificación y asignarles tareas.

Entre las principales tareas que se tienen en esta etapa son:

- Crear un plan de negocios

- Identificar los objetivos organizacionales y el rol de la tarjeta de calificación
- Determinar el desarrollo interno versus externo y el tipo de tarjeta de calificación
- Crear un plan de proyecto
- Identificar los riesgos del proyecto
- Identificar el equipo y las responsabilidades del proyecto

En el presente trabajo de tesis, tomamos en cuenta los anteriores puntos en tanto que el objetivo estaba bien definido: crear un modelo de calificación que permitiera entender los datos reales a los cuales se tiene acceso, y que en definitiva podríamos decir que este es un desarrollo interno. También fue relevante s la identificación de riesgos del proyecto. Específicamente, como veremos en el capítulo 5, en la descripción de datos, nuestra base de datos da acceso muy limitado a variables sociodemograficas lo que hizo inviable la creación de un modelo de originación y optamos por un modelo de comportamiento. También el último punto es de consideración en tanto que el presente trabajo, al ser un proyecto de tesis, tiene un único integrante que realizará todo el desarrollo.

4.2.2. Revisión de los datos y de los parámetros del proyecto

Entre las principales tareas que se tiene en esta etapa de revisión son:

- Disponibilidad y calidad de los datos
- Recopilación de datos para la definición de los parámetros del proyecto
- Definición de parámetros del proyecto
- Ventana de rendimiento y ventana de muestra
- Exclusiones
- Segmentación
- Metodología
- Revisión del plan de implementación.

Esta etapa es reportada en [13] como la más larga, y efectivamente, en el desarrollo de la tesis podemos confirmar tal afirmación. Aquí se invirtió mucho tiempo y requirió varias lineas de código R como evidencia el Apéndice A en donde se le hace disponible. Esta etapa es critica ya que determina parámetros de alto nivel para el proyecto, entre otros: exclusiones, ventanas de muestra/rendimiento. Los anteriores parámetros, o conceptos, se revisan con mayor detalle más adelante en la Sección 4.3. Como consecuencia de los resultados obtenidos en esta etapa se ve la conveniencia de definir de una manera dada al evento crediticio, a.k.a., la variable a explicar. Posiblemente el último punto, "Revisión del plan de implementación" sea el que es menos relevante para la tesis ya que no se tiene contemplado una implementación en producción.

4.2.3. Creación de la base de datos

Respetando la especificación de los parámetros referidos en la Sección 4.2.2 se genera una base de datos para el desarrollo de la tarjeta de calificación. Esta base de datos contendrá un conjunto de variables más una variable objetivo que en su conjunto serán la información de entrada para el desarrollo de la tarjeta de calificación.

Entre las principales tareas que se tiene en esta etapa son:

- Muestreo
- Recopilación y construcción de datos de desarrollo
- En su caso, ajuste por probabilidades previas (factoring).

4.2.4. Desarrollo de las tarjetas de calificación

Una vez llegando al punto en que se cuenta con una base de datos depurada en la que se incluye un conjunto de variables (explicativas) y una variable objetivo se puede pasar a la etapa de modelación. Hay varios métodos que se pueden utilizar para desarrollar las tarjetas de calificación a partir de estos datos. Todos ellos implican establecer y cuantificar la relación entre las características y el buen/mal desempeño (objetivo). En nuestro caso el modelo de comparación (baseline) es el modelo logístico y se comparará con los modelos que ya se han mencionado anteriormente: LDA, SVM, bosques aleatorios.

De acuerdo a [13], entre otras, las principales tareas que se tiene en esta etapa son:

- Exploración de datos

- Identificación de valores perdidos y atípicos
- Correlación
- Análisis inicial de variables
- Generación de tarjetas de calificación preliminar con datos de entrenamiento
- Elaboración de tarjetas de calificación final
- Validación.

Como puede verse, en una buena medida las tareas involucradas son como en cualquier proceso de modelación de datos. En nuestro caso, consideramos cada uno de los anteriores puntos. La exploración de datos se obtendrá a través de estadísticas descriptivas de las columnas que conforman las diferentes tablas de información. Seguimos la práctica común de dividir en dos partes los datos para tener datos de entrenamiento y datos de prueba. Lo anterior será la base de nuestra validación del modelo.

4.2.5. Informes de gestión de tarjetas de calificación

Una vez que se llega a una tarjeta de calificación en versión final, se debe considerar la generación de informes de gestión. Por ejemplo, hoy en día es muy común el concepto de tablero de mando (dashboard) que en una sola pantalla brinda un vistazo a métricas que consolidan la información, dan una idea del estado del sistema. Estos informes son herramientas valiosas que apoyan en la toma de decisiones de alta jerarquía, tales como el diseño de nuevas estrategias de adquisición. También ayudan a supervisar el desempeño de pronóstico de la tarjeta de calificación. De acuerdo a [13], estos informes deben diseñarse y producirse para ayudar al usuario comercial a responder preguntas como: "¿Dónde debo establecer mi límite para cumplir mis objetivos?" y "¿Qué impacto tendrá esto en mi cartera?". Por lo tanto, una buena práctica es obtener la opinión de los usuarios finales sobre qué informes les resultarían útiles para tomar decisiones y usarlos como guía para producir informes.

Por lo general, estos incluyen histogramas de las variables características de la tarjeta de calificación, gráficos de tasa de aprobación/rechazo esperado y los efectos de la tarjeta de calificación en subpoblaciones clave.

Además de estos informes de gestión, se debe producir como en todo proyecto de ingeniería de software, la debida documentación de la tarjeta de calificación que detalle los análisis realizados en cada fase clave del proyecto (es decir, desarrollo de casos de negocios, definiciones de bueno/malo/indeterminado, exclusiones,

segmentación, muestreo y recopilación de datos, análisis inicial de características , desarrollo de modelos, inferencia de rechazo, estadísticas de rendimiento de la tarjeta de calificación y validación) y el resultado generado. Entre las principales tareas que se tiene en esta etapa son:

- Generar tablas de ganancias
- Informes de variables.

Aunque muy importantes como son las anteriores consideraciones, estas no aplican en el presente trabajo ya que no está contemplada la implementación en un ambiente de producción.

4.3. La preparación de los datos

Ahora revisaremos mas a detalle una de las etapas que más nos interesan en el desarrollo de las tarjetas de calificación, la cual es la etapa de preparación de datos. La idea es definir una serie de pasos a seguir para preparar nuestros datos y así procurar tener un mejor desempeño de nuestro modelo. De acuerdo a [13] se deben considerar los siguiente pasos (o subetapas) que iremos presentando a continuación:

- Disponibilidad y calidad de los datos, Sección 4.3.1.
- Definición de los parámetros del proyecto, Sección 4.3.2.
- Segmentación, Sección 4.3.3
- Metodología, Sección 4.3.4.

En la anterior lista, el paso más extenso es el segundo y lo veremos con mucho detalle. El tercer paso de segmentación es importante y en general ayuda a mejorar la calidad de predicción de un modelo. En nuestro caso no implementamos este paso por falta de tiempo y porque tuvimos un buen desempeño en los modelos ajustados. El último paso de metodología se refiere a que existe una variedad de modelos que se pueden utilizar para la predicción y se dan una serie de consideraciones para seleccionar uno en particular.

4.3.1. Disponibilidad y calidad de los datos

Es necesario tener datos confiables y limpios para el desarrollo de la tarjeta de calificación, con un número mínimo aceptable de “buenos” y “malos”. La

cantidad de datos necesarios puede variar, pero en general, debe cumplir con los requisitos de significancia estadística y aleatoriedad. Siddiqi [13] menciona como regla del dedo pulgar que para el desarrollo de la tarjeta de calificación debe haber aproximadamente 2000 cuentas "malas" y 2000 cuentas "buenas" que se pueden seleccionar aleatoriamente para cada tarjeta de calificación propuesta dentro de un grupo de cuentas aprobadas abiertas en un marco de tiempo definido. En nuestro caso, dichos números serán significativamente rebasados, así es que no tendremos dificultad por datos y sus significancias. Al menos, de acuerdo a la regla del dedo.

En esta fase debe revisarse la confiabilidad, y en su caso seleccionar/filtrar variables de los datos disponibles. También debe decidirse si el modelo a producir se alimentará únicamente con datos internos de la empresa, o en su caso, se requiere contratar servicios de información especializados, por ejemplo de alguna oficina de información de crédito como pueden ser el Buro de crédito, Círculo de Crédito, Dun & Bradstreet y Trans-Union de México, por ejemplo ver la nota periodística www.elfinanciero.ysivienenmasburos.

En cuanto a la calidad de los datos, se debe investigar cada una de las variables para ver si no han sido alteradas y si son confiables. Los datos demográficos y otros datos que no se verifican, como los ingresos, son más susceptibles de tergiversarse. Por otro lado, los datos externos de agencias de crédito son por lo general más sólidos y se pueden usar.

Una vez que se determina que hay suficientes datos internos de buena calidad para proceder, se deben evaluar, cuantificar y definir las necesidades de datos externos. La organización puede decidir desarrollar una tarjeta de calificación basados únicamente en datos internos, o puede optar por complementar estos datos de fuentes externas, como oficinas de crédito, repositorios centrales de reclamaciones, proveedores de datos geodemográficos, etc.

Como veremos más adelante, la información con que se cuenta en el presente trabajo de tesis tiene una procedencia mixta. Principalmente es información interna, pero también se tiene información externa de una agencia de información crediticia.

4.3.2. Definición de los parámetros del proyecto

En esta sección se presenta los pasos a seguir para preparar los datos en lo referente a definir unos parámetros del proyecto que reflejan la lógica de negocio de la cartera. Como se menciono anteriormente, esta es la etapa de mayor volumen y con mayor número de divisiones. Estas incluyen:

- Exclusiones, Sección 4.3.2.1.

- Ventanas de rendimiento y muestra, Sección 4.3.2.2
- Efectos de la estacionalidad, Sección 4.3.2.3.
- Definición del evento crediticio, Sección 4.3.2.4
- Confirmación de la definición del evento crediticio, Sección 4.3.2.5
- Indeterminadas, Sección 4.3.2.6

4.3.2.1. Exclusiones

El primer parámetro de proyecto en consideración está determinado por el concepto de exclusiones. Se refiere a que en la estimación de una tarjeta de calificación se consideran registros de cuentas que sean representativas de las nuevas solicitudes que llegarán en un futuro y las cuales se deben de calificar para decidir si se otorga el crédito o no. Esto en el término especializado y en inglés se le conoce como clientes "through-the-door". Lo anterior aplica a una tarjeta de calificación de originación pero también es relevante para las tarjetas de calificación de comportamiento. En efecto, para este segundo estilo de modelos también se requiere una consistencia entre los datos de "entrenamiento" y los registros en la cartera que se calificarán.

Ciertos tipos de cuentas deben excluirse de la muestra de desarrollo. Entre otras:

- Las cuentas que tienen un rendimiento anormal (por ejemplo, fraudes)
- Aquellas que se otorgan utilizando algún criterio que no depende de la puntuación. Por ejemplo créditos: VIP, fuera del país, o preaprobados.
- Si hay áreas geográficas o mercados en los que la empresa ya no opera, los datos de estos mercados también deben excluirse.
- Relacionado al anterior punto, en el caso de una tarjeta de calificación para aplicar a los residentes de una ciudad, no incluiría en la muestra de desarrollo a aquellos que viven en áreas rurales.

4.3.2.2. Ventanas de rendimiento y de muestra

La definición del evento crediticio de interés y por tanto la definición de crédito "malo" dependen de la necesidad de negocio en particular. Más precisamente,

cuando el evento de interés es el de fraude (un crédito que nunca recibió un pago programado), bancarrota (incapacidad de pago), o cuentas canceladas "chargeoff", la definición del evento es de antemano determinada. Cuando el evento crediticio tiene que ver con la morosidad (pagos impuntuales) que en el inglés le llaman delinquency, la definición no es única y debe determinarse siguiendo algún criterio. En todos estos casos es necesario fijar el período durante el cual se observará al crédito para declarar la presencia del evento y levantar la bandera de crédito malo. Sobretodo cuando se trata de un crédito revolvente o una línea de crédito. Para este fin se introducen las definiciones de dos conceptos:

- Ventana de muestreo: es el período de tiempo en el que se incluyen las cuentas de crédito a analizar. Por ejemplo, todas las cuentas que se aperturaron en el mes de enero del año pasado.
- Ventana de desempeño: es el período de tiempo en el que se observan las cuentas seleccionadas en el paso anterior. Por ejemplo, se puede observar el comportamiento durante doce meses de todas las cuentas aperturadas el mes de enero del año pasado.

La idea más importante en el concepto de ventana de desempeño es dar el suficiente tiempo para que una cuenta que en retrospectiva fue mala, efectivamente se calificó como tal en el período de observación. Llegamos así a la siguiente pregunta

> Como determinar el tamaño del período de desempeño?

Diferentes tipos de crédito requieren diferentes tamaños, de acuerdo a Siddiqi [13] tenemos que

- Las tarjetas de crédito requieren por lo general de un período de desempeño de entre 18 y 24 meses.
- Los créditos hipotecarios requieren un período de entre tres a cinco años.
- Una tarjeta de calificación de comportamiento generalmente utiliza de entre seis a doce meses.

El lector se preguntará por la forma en que se determinan dichos períodos. Hay dos métodos íntimamente relacionados, el primero **el análisis de cohortes o el análisis vintage**, y el segundo, **la tasa de evolución en la aparición del evento crediticio** (o simplemente la tasa de malos). Dado que están relacionados estos dos métodos, concentrémonos en el segundo.

La Figura 4.1 presenta un gráfico de la tasa de malos calculada para las cuentas que llevan x meses originadas y se repite (arbitrariamente) durante un período de 14 meses.

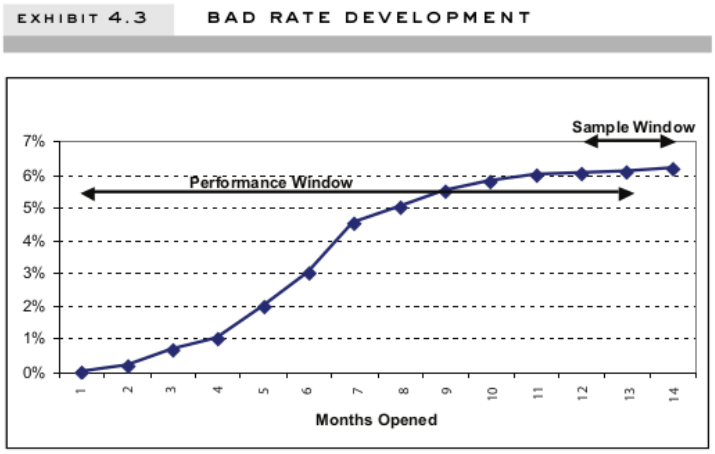

Figura 4.1: Desarrollo de tasa de malos. Fuente: [13, Figura 4.3]

Esta figura muestra un ejemplo de una cartera en la que se ha trazado la tasa de malos para cuentas abiertas en un período de 14 meses. Puede observarse en la gráfica que la tasa de malos tiene una forma convexa en una primera parte del intervalo de tiempo y en la segunda parte terminal es concava, además en su concavidad, hay un punto en el cual el cambio marginal ("la derivada") es prácticamente cero. Este comportamiento es lo que permite mediante la observación de la curva el definir el período de desempeño:

> Se elige un período en el que el cambio marginal es prácticamente (subjetivamente) cero. En este sentido, se elige el período cuando la tasa de malos se "comienza a estabilizar".

En el ejemplo anterior, una buena ventana de muestra sería entre 12 y 14 meses en el pasado con una ventana de rendimiento de aproximadamente 12 meses.

En el caso de una tarjeta de calificación de morosidad, es recomendable graficar la tasa de malos para varias definiciones de morosidad relevantes. Esto se hace porque las diferentes definiciones producirán diferentes recuentos de En cambio, para una tarjeta de calificación de bancarrota o canceladas, solo un análisis es suficiente ya que solo hay una definición posible de "malo".

4.3.2.3. Efectos de la estacionalidad

Los datos seleccionados en la ventana de muestreo deberán ser representativos de los clientes esperados futuros. Es necesario que en la etapa de exclusiones se dejen fuera a aquellos registros de cuentas anormales para que únicamente se mantengan a cuentas que represente a la "población normal" (población

"through-the-door"). Ciertamente la anterior consideración tiene un grado de subjetividad y como menciona [13], lo más importante es excluir casos extremos. Un caso frecuente que puede introducir sesgos ("shifts") puede ser una campaña de mercadeo destinada a un segmento específico de la población. Por ejemplo, el objetivo de captar más clientes en un intervalo de edad para una cartera de tarjetas de crédito claramente introduce una modificación a la distribución de la cartera.

Además de las exclusiones, otra forma de contrarestar los efectos de la estacionalidad es construir varias ventanas de muestra todas ellas con la misma ventana de rendimiento. Por ejemplo, se pueden tomar tres muestras de cada uno de los meses de enero, febrero y marzo, con ventanas de rendimiento de 12 meses cada una.

4.3.2.4. Definición del evento crediticio

La definición del evento crediticio dependerá del objetivo de la tarjeta de calificación (por ejemplo, si es una tarjeta de calificación para cuentas canceladas, fraude, morosidad, etc.). En todo caso, [13] recomienda mantener en mente las siguientes observaciones:

- La definición debe ser consistente con los objetivos organizacionales. Si el objetivo es aumentar la rentabilidad, entonces la definición debe establecerse en un punto de morosidad en el que la cuenta deja de ser rentable, para lo cual deberán hacerse las debidas estimaciones. Si el objetivo es la detección de morosidad, la definición será más simple (por ejemplo 90 días).
- La definición debe ser consistente con el producto o propósito para el cual se crea la tarjeta de calificación, por ejemplo, quiebra, fraude, reclamos (este último en el caso de seguros).
- Una definición "más estricta", por ejemplo, "120 días de mora", brinda una diferenciación más extrema (y precisa), pero en algunos casos puede generar tamaños de muestra bajos.
- Una definición "más flexible" (p. ej., 30 días de morosidad) generará un mayor número de cuentas para la muestra, pero puede no ser un diferenciador lo suficientemente bueno entre cuentas buenas y malas y, por lo tanto, producirá una tarjeta de calificación débil. A este respecto ver la Sección 4.3.2.5 de confirmación de la definición del evento crediticio, en donde se introduce el análisis de balanceo.
- La definición debe ser fácilmente interpretable (p. ej., c 90 días de morosidad, etc.). Definiciones como "tres veces 30 días de morosidad o dos veces 60

días de morosidad, o una vez 90 días", son mucho más difíciles de rastrear y pueden no ser apropiadas para todas las empresas.

- En algunos casos, puede ser beneficioso tener definiciones coherentes de "malo" en varios segmentos y otras tarjetas de calificación en uso dentro de la empresa. Esto facilita la gestión y la toma de decisiones, especialmente en entornos donde se utilizan muchas tarjetas de calificación.
- Puede haber requisitos normativos u otros requisitos externos que rijan cómo se define la morosidad. Por ejemplo el Acuerdo de Capital de Basilea II impone requisitos de presentación de informes.

4.3.2.5. Confirmación de la definición del evento crediticio

Una vez que se identifica una definición de 'cliente malo" utilizando el análisis descrito anteriormente, se debe realizar un análisis adicional para confirmarla. Lo anterior con el propósito de confirmar que las cuentas identificadas como malas realmente lo sean. Dicha confirmación será más relevante cuando la asignación de clase "mala" se basa en algún nivel de morosidad, como es nuestro caso. Aunque hay varias alternativas, nosotros nos enfocaremos en un método analítico conocido como análisis de tasa de balanceo (roll rate) cuyo resultado con nuestro datos será reportado en el Capítulo 5. El **Análisis de tasa de balanceo (roll rate)** consiste en determinar un punto en el tiempo t que denominaremos "presente" y comparar la peor morosidad en un período de "x" meses anteriores a t con los "x próximos" meses posteriores a t, y luego calcular el porcentaje de cuentas que mantienen su peor morosidad, mejoran o "avanzan" durante los siguientes meses.

El propósito de este ejercicio es identificar un '**punto de no regreso** (point of no return), es decir, el nivel de morosidad en el que la mayoría de las cuentas se vuelven incurables. Por lo general, la gran mayoría de las cuentas que alcanzan los 90 días de morosidad no se curan: empeoran (avanzan), lo que confirma que esta definición de "malo" es adecuada.

Cabe señalar que el método descrito para determinar y confirmar las definiciones "malas" se puede realizar tanto para una tarjeta de calificación de originación como para las de comportamiento.

4.3.2.6. Indeterminadas

Una vez que se definen las cuentas "malas" también es necesario definir las cuentas "buenas". Hay una sutileza ya que la definición de cuenta buena no necesariamente debe ser el complemento lógico de cuenta mala. Lo anterior es consecuencia al hecho de que la definición de buena debe incorporar los objetivos

de la empresa. Como consecuencia al hecho de que las cuentas buenas no son el complemento lógico de las cuentas malas, entonces aparece una tercera categoría que dará origen a las cuentas indeterminadas.

Las cuentas indeterminadas son aquellas que al analizarlas no pueden clasificarse como "buenas" o "malas". Una forma en que surgen tales cuentas es debido a que no se cuenta con el suficiente historial. También puede ser, en el caso de una tarjeta de calificación de morosidad, a que una cuenta presente una morosidad leve y por lo tanto sea insuficiente para clasificarla como buena o mala. De acuerdo a Siddiqi [13], las cuentas indeterminadas pueden incluir entre otras:

- Cuentas que alcanzan una morosidad de 30 o 60 días pero que no empeoran.
- Cuentas inactivas y canceladas voluntariamente. Cuentas de "oferta rechazada", solicitudes que fueron aprobadas pero no reservadas y otras que fueron aprobadas pero tienen un historial de rendimiento insuficiente para la clasificación
- Cuentas con uso insuficiente.
- Cuentas de seguros con reclamos por debajo de un valor en dólares específico.

Otras consideraciones:

- Las cuentas que se cancelan voluntariamente deben considerarse como indeterminadas. Lo anterior debido a que si estos clientes volvieran a presentar una solicitud, serían nuevamente calificados y probablemente aprobados nuevamente.
- Las cuentas indeterminadas solo se usan cuando la definición de "malo" se puede establecer de varias maneras y, por lo general, no se requieren cuando la definición es clara (por ejemplo, en tarjeta de calificación de fraude o quiebra).
- Como regla general, los indeterminados no deben exceder más del 10 % al 15 % de la cartera.
- En caso que dicha proporción sea mayor, es necesario un análisis para abordar las causas fundamentales de dicha indeterminación. Por ejemplo si la indeterminación es por bajo uso del producto, determinar la presencia de otros productos internos que compiten entre si, ya bien sea por mejores programas de lealtad o mayores límites de crédito, etc.

En nuestro caso, para el ejercicio de estimación de esta tesis, hemos decidido no incluir el estado de indeterminado y considerar únicamente créditos que han concluido con anterioridad a la fecha de la generación de la base de datos. Nuestra decisión también está fundamentada en el hecho de que no contamos con información de calidad sobre aquellas solicitudes de crédito aceptadas pero que se cancelaron voluntariamente.

4.3.3. Segmentación

En esta sección contestaremos las siguientes preguntas:

- Que es la segmentación y cuales son los métodos para realizarla?
- Porque segmentar en calificación crediticia?
- Porque es importante diferenciar entre segmentos por perfil del cliente y por perfil de riesgo?
- Cuales son los casos en que no es recomendable segmentar?

Con respecto a la primera pregunta, Que es la segmentación y cuales son los métodos para realizarla?, tenemos la siguiente respuesta. Sucede con frecuencia que en una población en observación hay de manera natural una estructura en el sentido de que dicha población se forma como la unión de una colección de subpoblaciones. La solución al problema inverso de identificar cuales son dichas subpoblaciones es de mucho interés, y las técnicas involucradas se denominan técnicas de segmentación. Se pueden reconocer dos formas de realizar la segmentación:

- Generar segmentos en base a la experiencia y el conocimiento de la industria.
- Generar segmentos usando técnicas estadísticas como clustering o árboles de decisión.

En las siguientes dos Subsecciones 4.3.3.1 y 4.3.3.2 veremos mayor detalle de estas dos alternativas.

La respuesta a la segunda pregunta, Porque segmentar en calificación crediticia? es la siguiente. De manera natural, frecuentemente una cartera de crédito se compondrá de distintas subpoblaciones con distintos pérfiles de riesgo. Luego entonces, el uso de varias tarjetas de calificación para cada uno de los segmentos de una cartera proporciona una mejor diferenciación de riesgos que el uso de una única tarjeta de calificación para todos. Es entonces también de esperarse que una única tarjeta de calificación no funcionará de manera eficiente para todos los

segmentos, o subpoblaciones. En la Figura 4.2 se ve un caso extremo en que las variables "Numero de transacciones" y "Residencia" (que puede tener los valores propia, rentada, padres) tienen muy diferentes relaciones con el evento crediticio dependiendo del segmento definido por la variable "Edad". El fenómeno es espectacular en tanto que una variable pasa de un efecto positivo a uno negativo dependiendo del segmento.

EXHIBIT 4.8 BAD RATES BY ATTRIBUTES FOR AGE-BASED SEGMENTATION

Bad Rate			
	Age > 30	*Age < 30*	*Unseg*
Res Status			
Rent	2.1%	4.8%	2.9%
Own	1.3%	1.8%	1.4%
Parents	3.8%	2.0%	3.2%
Trades			
0	5.0%	2.0%	4.0%
1–3	2.0%	3.4%	2.5%
4+	1.4%	5.8%	2.3%

Figura 4.2: Segmentos por edad. Fuente: Siddiqi [13].

Con respecto a la tercera pregunta, Porque es importante diferenciar entre segmentos por perfil del cliente y por perfil de riesgo?, se tiene la siguiente consideración. Se debe tener en cuenta que en el desarrollo de la tarjeta de calificación de riesgo, una población "distinta" no se reconoce como tal basándose únicamente en sus características que la definen (como la demografía), sino más bien en su desempeño con respecto al evento crediticio en observación. El objetivo es definir segmentos basados en el desempeño basado en el riesgo.

Ahora bien, para que la detección de un comportamiento diferenciado sea de interés, debe llevar a efectos positivos de negocio (por ejemplo, menores pérdidas, mayores tasas de aprobación para ese segmento, etc). Contestando la última pregunta, Cuales son los casos en que no es recomendable segmentar?, tenemos de acuerdo a Siddiqi [13]:

- Costo de Desarrollo. Esto incluye el esfuerzo interno y externo involucrado para producir tarjetas de calificación con documentación completa.
- Costo de Implementación. Las tarjetas de calificación adicionales cuestan recursos del sistema para implementarlos.
- Procesamiento. Hay costos de procesamiento adicionales asociados con más tarjetas de calificación.

- Desarrollo y Seguimiento de la Estrategia. Cada tarjeta de calificación requiere un conjunto de estrategias asociadas, reglas de política e informes de seguimiento. Crearlos, administrarlos y mantenerlos requieren recursos.

4.3.3.1. Segmentación basada en la experiencia

La segmentación basada en la experiencia incluye ideas generadas a partir del conocimiento y la experiencia del negocio, consideraciones operativas y prácticas de la industria. Las áreas de segmentación típicas utilizadas en la industria incluyen:

- Demografía. Regional (provincia/estado, definición interna, urbano/rural, basado en código postal, vecindario), edad, código de estilo de vida, tiempo en la oficina, permanencia en el banco
- Tipo de producto. Tarjetas oro/platino, duración de la hipoteca, tipo de seguro, garantizado/no garantizado, arrendamiento de automóviles nuevos versus usados, monto del préstamo.
- Fuentes de Negocio. Tienda, sucursal, Internet, distribuidores, etc.
- Tipo de solicitante. Cliente existente/nuevo, comprador de vivienda por primera vez/renovación de hipoteca, grupos comerciales profesionales, etc.
- Propiedad del producto. Por ejemplo, hipotecarios que solicitan tarjetas de crédito en el mismo banco.

Un método simple para confirmar la segmentación es analizar el comportamiento de riesgo de una misma característica en diferentes segmentos predefinidos. Si la misma característica (por ejemplo, "Residencia") predice de manera diferente en segmentos únicos, esto puede presentar un caso para tarjetas de calificación segmentadas; ver la Figura 4.2.

4.3.3.2. Segmentación basada en técnicas estadísticas

Existen diferentes técnicas estadísticas para la segmentación como lo son el agrupamiento o formación de cúmulos por K-medias, y los árboles de decisión. Siendo un tema importante y recurrente existen diferentes referencias donde puede consultarse a detalle; ver por ejemplo [7, Capítulos 9 y 14].

En el clustering, la agrupación en clústeres, coloca objetos en grupos o "clústeres" sugeridos por los datos. Los objetos de cada grupo tienden a ser similares entre sí con respecto a un criterio bien específico, y los objetos de diferentes grupos

tienden a ser distintos. Es importante señalar que el clustering identifica grupos que son similares en función de sus características, no de su desempeño de riesgo crediticio. [1] Por lo tanto, los clústeres diferencian en las variables consideradas y debe verificarse que también diferencian con respecto al desempeño de riesgo. Por ello es necesario analizar cada segmento con respecto a su riesgo, por ejemplo a través de un análisis del desarrollo de la tasa de malos conforme a lo expuesto en la Sección 4.3.2.2.

Tanto los análisis basados en la experiencia como los estadísticos definen pasos a seguir para una segmentación. Es una tarea aparte el evaluar los beneficios de tal segmentación. Por razones de negocio se deben considerar los puntos al final de la Sección 4.3.3. En el campo puramente de la estadística, una forma de determinar la conveniencia de la segmentación es medir la mejora en el poder predictivo. Esto se puede hacer utilizando una serie de estadísticas, como Kolmogorov-Smirnov (KS), o la curva ROC. Está última se expuso con detalle en el Capítulo 2, Sección 2.5.

Siempre será el usuario final quien deba establecer los criterios bajo los cuales se justifica el esfuerzo adicional de desarrollo e implementación.

4.3.4. Metodología

Hay varias técnicas matemáticas disponibles para crear tarjetas de calificación de predicción de riesgos, por ejemplo, regresión logística, redes neuronales, árboles de decisión, etc. La técnica más adecuada a utilizar puede depender de varias situaciones como lo señala Siddiqi [13]:

- La calidad de los datos disponibles. Un árbol de decisión puede ser más apropiado para los casos en los que faltan muchos datos o en los que la relación entre las características y la variable objetivo no es lineal.
- Tipo de variable objetivo, es decir, binario (bueno/malo) o continuo (ganancia/pérdida).
- Tamaños de muestra disponibles. Por ejemplo las redes neuronales profundas requieren estimar un número considerable de parámetros que en algunos casos puede rebasar considerablemente el número de datos disponibles.
- Interpretabilidad de los resultados, como en el caso de la facilidad con la que se pueden interpretar las tarjetas de calificación basadas en puntos

[1] Si existiera a priori una variable que mida tal riesgo entonces los grupos podrían reflejarlo. Los resultados que se reportarán en el Capítulo 5 corresponden a construir tal variable.

desarrollados por regresión logística con variables agrupadas y expresadas con el peso de la evidencia.

- Cumplimiento legal de la metodología, generalmente requerida por los reguladores locales para que sea transparente y explicable.
- Capacidad para realizar un seguimiento y diagnosticar el rendimiento de la tarjeta de calificación. Es decir, el mantenimiento del modelo.

La técnica y el formato previsto de la tarjeta de calificación deben comunicarse al interior de la empresa, particularmente a los gerentes de riesgos y departamento de tecnologías de la información, para asegurarse de que se comprendan los datos y los problemas teóricos sobre las técnicas identificadas.

Capítulo 5

Reporte de Resultados

En este capítulo se reportan resultados de modelación de datos reales de una cartera de crédito al consumo. El objetivo es el pronóstico de la probabilidad de un evento crediticio en una segunda mitad de la vida del crédito dado que el evento se observó en la primera mitad. Investigamos la siguiente definición:

- El cliente falla en pagar puntualmente en al menos 3 pagos programados de manera consecutiva.

La cartera de crédito consiste de datos reales de una empresa especializada en el otorgamiento de crédito a las personas en la que los créditos están destinados a la adquisición de dispositivos electrónicos. Por confidencialidad no se menciona el nombre de la empresa. Para la modelación se utilizarán las siguientes alternativas:

- Regresión logística
- Análisis Discriminante Lineal (LDA)
- Maquinas de Soporte Vectorial (SVM)
- Bosques aleatorios

Para medir el desempeño del modelo se observará principalmente el indicador AUC, el área bajo la curva ROC (Receiver Operatoring Characteristic Curve). Nos enfocamos principalmente en esta métrica ya que es práctica comun en la literatura especializada el enfocarse en ella. En dicha literatura complementan el análisis con otras métricas aparte del AUC principalmente en métricas basadas en la matriz de confusión. Por ello es que nosotros reportaremos el AUC principalmente, y complementaremos reportando la matriz de confusión.

Aparte de esta breve introducción el capítulo está organizado de la siguiente forma. En la Sección 5.1 se hace una descripción detallada de los datos. En la Sección 5.2 se presentan resultados de análisis del evento crediticio con un modelo

logístico. En la Sección 5.3 se presentan resultados con el modelo LDA. En la Sección 5.4 se presentan resultados con el modelo SVM. En la Sección 5.5 se presentan resultados con el modelo de bosques aleatorios.

5.1. Descripción de datos

La información que define la cartera de crédito está estructurada en varias tablas. Previo a describir los datos con los cuales se cuenta, es importante señalar que los créditos considerados son estructurados: Para cada solicitud se determina un pago inicial (un enganche determinado por la empresa) que representa un porcentaje del costo de venta del dispositivo y se programan n pagos (seleccionado por el cliente), todos ellos iguales entre si, típicamente n puede ser 13, 26, 39 o 52 en una periodicidad semanal. Cada pago programado tiene asociado una fecha limite en la que se debe liquidar el monto total. En caso contrario, se levanta una bandera de morosidad para dicha semana.

A continuación hacemos una descripción de cada una de las tablas que conforman nuestra información.

Existe una primera tabla que contiene información de contacto asociada a la solicitud. Esta contiene las siguientes columnas

```
"clave_solicitud"      "edad"          "codigo_postal"
"estado"               "municipio"     "ciudad"
"tipo_asentamiento"    "oper_alta"     "fecha_alta"
```

En una segunda tabla se tienen datos de venta. Aquí cada registro es una solicitud de crédito aprobada para la adquisición de un dispositivo, además de una variable de un perfil de riesgo que la empresa asocia a la la solicitud. Esta tabla contiene las siguientes columnas

```
"clave_solicitud"  "clave_cliente"   "fecha_venta"
"precio_venta"     "enganche"        "plazo"
"perfil_riesgo"
```

La llave de entrada para ambas tablas es “clave_solicitud” y puede suceder que una misma “clave_cliente” tenga asociada una o mas solicitudes. De hecho, adelantándonos a la presentación del modelo que se verá mas adelante, diremos que la variable que cuenta el número de solicitudes asociadas a un mismo cliente es

significativa. El cociente "enganche"/"precio_venta" nos permite crear una nueva variable que llamaremos **porcentaje**.

Una tercera tabla contiene información de como fue el comportamiento histórico en cada solicitud. Particularmente, para cada solicitud se registraba información cada uno de los pagos programados. Particularmente los siguientes campos:

`"clave_solicitud"`

`"numero_pago"`

`"fecha_limite"`

`"pago_realizado"`

`"pago_restante"`

`"pago_fecha"`

`"fecha_alta"`

Esta es la información con la que contamos para realizar el modelo. El lector atento observará que no contamos con información sociodemográfica del cliente. Aún así, con agradable sorpresa, es posible generar un modelo de comportamiento con un buen desempeño predictivo.

Antes de presentar los modelos, en las Tablas 5.2 y 5.1 se muestra algunas estadísticas descriptivas de algunas variables:

Perfil de riesgo	Porcentaje
A	22.74
B	25.49
C	51.75

Tabla 5.1: Porcentajes por tipo de perfil de riesgo.

Con la información disponible en las tablas de la base de datos, consideraremos las siguientes variables para generar un modelo:

- La variable "numero_plazo" especifica el numero de semanas para el credito, pueden ser los valores 13, 26, 39 o 52.
- La variable "solic_prev" da el numero total de solicitudes previas realizadas por el mismo cliente.

- La variable “porcentaje” indica el porcentaje que representa el enganche sobre el valor total de la compra.
- La variable “edad” es del cliente asociado a la solicitud.
- La variable “rural” es una bandera que indica cuando el asentamiento en donde vive el cliente es catalogado como rural.
- La variable “perfil_riesgo” es una variable asociada al riesgo del crédito a la cual se le asigna una letra de la “A” a la “C”. El riesgo mayor se representa por “C”, el intermedio por “B” y el menor por “A”.
- La variable “EVM1” toma los valores cero y uno. Solamente es uno cuando en la **primera mitad** de la vida del crédito hay tres o mas retrasos de manera consecutiva en los pagos programados. Esta es una variable explicativa.
- La variable “EVM2” toma los valores cero y uno. Solamente es uno cuando en la **segunda mitad** de la vida del crédito hay tres o mas retrasos de manera consecutiva en los pagos programados. Esta es la variable a explicar.

	EVM1	EVM2	rural	edad	numero_plazos	solic_prev	porcentaje
min	0.00	0.00	0.00	18.00	13.00	0.00	0.13
max	1.00	1.00	1.00	88.00	52.00	2.00	0.94
range	1.00	1.00	1.00	70.00	39.00	2.00	0.81
median	0.00	0.00	0.00	34.00	26.00	0.00	0.28
mean	0.26	0.43	0.19	35.08	22.94	0.02	0.27

Tabla 5.2: Estadísticas descriptivas

También es importante mencionar el análisis previo realizado a los datos, en la Figura 5.1 se muestra el análisis de tasa de balanceo (ver Capitulo 4, Sección 4.3.2.2) para las cuentas a un plazo de 26 semanas, tomando $x = 13$ semanas, es decir, el punto medio de vida del crédito.

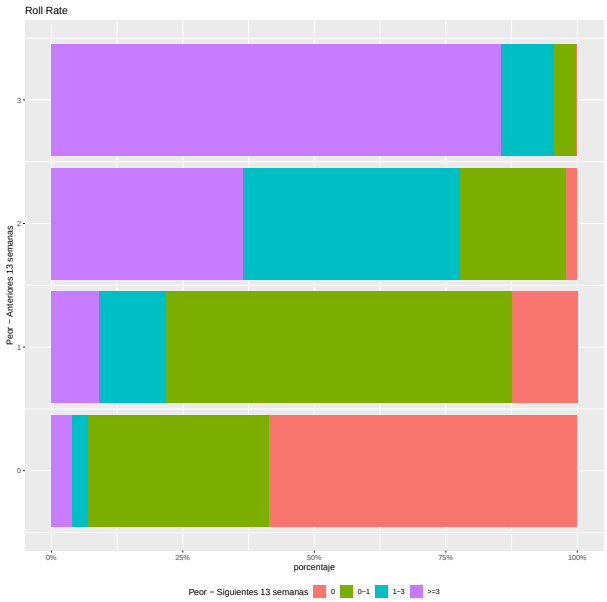

Figura 5.1: Análisis de tasa de balanceo (26 semanas) para tres semanas de morosidad. Elaboración propia.

Podemos observar en la Figura 5.1 como la mayoría de créditos que se atrasan tres semanas en realizar algún pago durante la primera mitad de vida del crédito, en la segunda mitad no "mejoran", por lo que siguen incurriendo en morosidad, por su parte un buen porcentaje de los clientes que durante la primera mitad nunca se atrasaron permanecieron pagando a tiempo sus créditos. En el apéndice B se presentan las gráficas para definiciones con mas semanas y para diferentes plazos. En las Figuras B.1 y B.2 se encuentra el análisis de tasa de balanceo para las cuentas de 26 semanas para definiciones de malo de 4 y 5 semanas respectivamente, el comportamiento es similar al análisis con definición de malo de tres semanas. Realizando el mismo análisis de tasa de balanceo pero para cuentas de 52 semanas se notaron resultados muy similares a los obtenidos para los créditos de 26 semanas. En las Figuras B.3, B.4 y B.5 se encuentra el análisis de tasa de balanceo para las cuentas de 52 semanas para definiciones de malo de 3, 4 y 5 semanas respectivamente.

Para la definición del evento crediticio incorporado en la variable EVM2 se utilizan los campos "fecha_limite" y "pago_fecha". El primero nos permite saber cual es la fecha limite en que se debe liquidar el saldo del pago cuyo numero esta registrado en "numero_pago" y el campo "pago_fecha" nos permite conocer la fecha exacta en que se realizó.

Solamente se considerarán para la modelación aquellas solicitudes cuya fecha límite del último pago programado sea antes de la fecha de corte de la base de

datos.

5.2. Modelo logístico

En esta sección reportaremos resultados que hemos obtenido al ajustar un modelo logístico a la variable a pronosticar EVM2. Los datos se partieron en dos grupos, uno primero con el 70 % de las observaciones en el conjunto de datos de entrenamiento y el 30 % restante para una prueba de validación. Como mencionamos anteriormente, únicamente se consideran créditos que ya hayan concluido y en total fueron 63,645 registros.

Recordamos las variables del modelo:

- La variable "numero_plazo" especifica el numero de semanas para el credito, pueden ser los valores 13, 26, 39 o 52.
- La variable "solic_prev" da el numero total de solicitudes previas realizadas por el mismo cliente.
- La variable "porcentaje" indica el porcentaje que representa el enganche sobre el valor total de la compra.
- La variable "edad" es del cliente asociado a la solicitud.
- La variable "rural" es una bandera que indica cuando el asentamiento en donde vive el cliente es catalogado como rural.
- La variable "perfil_riesgo" es una variable asociada al riesgo de crédito la cual se le asigna una letra de la "A" a la "C", siendo "C" el cliente con más riesgo. En la regresión logística está variable categórica dará lugar a las variables dicotómicas "perfil_riesgoB" y "perfil_riesgoC".
- La variable "EVM1" toma los valores cero y uno. Solamente es uno cuando en la **primera mitad** de la vida del crédito hay tres o mas retrasos de manera consecutiva en los pagos programados. Esta es una variable explicativa.
- La variable "EVM2" toma los valores cero y uno. Solamente es uno cuando en la **segunda mitad** de la vida del crédito hay tres o mas retrasos de manera consecutiva en los pagos programados. Esta es la variable a explicar.

En la Tabla 5.3 se muestran los resultados de la regresión logística. En la Tabla 5.4 se muestran prueba ANOVA para dicho modelo.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.78	0.09	-9.05	0.00
EVM1	4.75	0.06	78.39	0.00
rural	0.03	0.03	1.06	0.29
edad	-0.02	0.00	-15.00	0.00
perfil_riesgoB	0.24	0.04	5.81	0.00
perfil_riesgoC	0.52	0.05	9.79	0.00
numero_plazos	0.02	0.00	13.48	0.00
solic_prev	-0.28	0.10	-2.88	0.004
porcentaje	-1.87	0.32	-5.87	0.00

Tabla 5.3: Resultados de la regresión logística.

	Df	Deviance	AIC	LRT	Pr(>Chi)	
<none>		37665.21	37683.21			
EVM1	1	57698.60	57714.60	20033.39	0.0000	***
rural	1	37666.33	37682.33	1.12	0.2907	
edad	1	37897.16	37913.16	231.94	0.0000	***
perfil_riesgo	2	37773.08	37787.08	107.87	0.0000	***
numero_plazos	1	37845.57	37861.57	180.35	0.0000	***
solic_prev	1	37673.97	37689.97	8.75	0.0031	**
porcentaje	1	37703.55	37719.55	38.33	0.0000	***

Tabla 5.4: Prueba ANOVA de la regresión logística.

Las variable rural no fue significativa. Las variables que fueron significativas tienen un signo que nos hace sentido. El coeficiente de la variable numero_plazos es positivo y significa que entre mayor es el plazo del crédito mayor es la probabilidad de observar el evento crediticio. La variable solic_prev tiene signo negativo y nos indican que entre mas solicitudes tiene el cliente, menor la probabilidad de

observar el evento crediticio. La variable edad tuvo un coeficiente negativo lo que indica que entre mas maduro es el cliente, menor es la probabilidad de observar el evento crediticio. También la variable porcentaje tiene signo negativo por lo que entre más alto sea el porcentaje de enganche, menor será la probabilidad del evento crediticio. El signo positivo de la variable EVM1 nos hace sentido y significa que observar tres demoras consecutivas en el comienzo del crédito tiene un impacto positivo en la probabilidad de observar al menos otras tres demoras consecutivas hacia el final, además cabe resaltar que esta variable es la que tiene el coeficiente más alto del modelo.

El área bajo la curva ROC del modelo para los datos de entrenamiento fue de

$$AUC = 0.8292$$

Para los datos de prueba se puede visualizar la curva en la Figura 5.2 y obtuvo un área

$$AUC = 0.8269.$$

La matriz de confusión con los datos de prueba para el modelo LOGIT se visualiza en la Tabla 5.5.

	No evento	Evento	Exactitud
Pronosticado: No evento	0.99	0.40	
Pronosticado:Evento	0.01	0.60	
Exactitud			0.82

Tabla 5.5: Matriz de confusión modelo LOGIT y la tasa global de pronóstico.

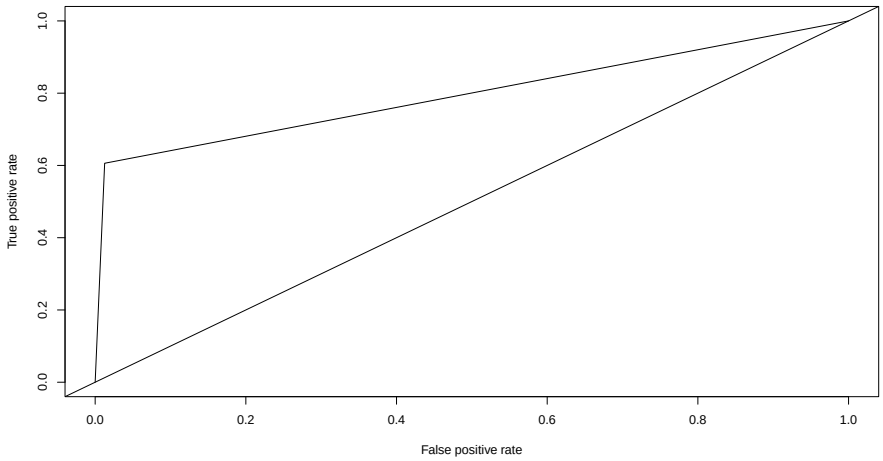

Figura 5.2: Curva ROC del modelo logístico con los datos de prueba.

5.2.1. Interacción de variables

En esta sección se presenta el ejercicio de considerar un conjunto de variables explicativas extendido que se obtiene al considerar las interacciones entre las variables existentes. Los resultados se reportan en las Tablas 5.6 y 5.7. En la Tabla 5.8 se reporta prueba ANOVA del modelo. Como puede verse, muy pocas de las nuevas variables son significativas. En términos de desempeño no hay una mejoría substancial. Con mayor detalle, el AUC obtenido es de 0.8269 y en la matriz de confusión la tasa de éxito en el pronóstico de no evento es 0.988 y para el pronostico del evento es de 0.596 que son similares al modelo sin interacciones.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.34	0.32	-1.06	0.29
EVM1	5.68	0.43	13.09	0.00
rural	0.04	0.21	0.22	0.83
edad	-0.03	0.01	-3.49	0.00
perfil_riesgoB	0.12	0.25	0.49	0.62
perfil_riesgoC	0.51	0.29	1.73	0.08
numero_plazos	0.00	0.01	0.44	0.66
solic_prev	-0.15	0.67	-0.23	0.82
porcentaje	-3.61	1.47	-2.45	0.01
EVM1:rural	-0.08	0.15	-0.54	0.59
EVM1:edad	0.00	0.01	0.25	0.80
EVM1:perfil_riesgoB	0.29	0.19	1.49	0.14
EVM1:perfil_riesgoC	0.35	0.25	1.40	0.16
EVM1:numero_plazos	-0.02	0.01	-2.51	0.01
EVM1:solic_prev	0.38	0.72	0.53	0.60
EVM1:porcentaje	-2.83	1.56	-1.81	0.07

Tabla 5.6: Resultados de la regresión logística con interacciones 1/2.

	Estimate	Std. Error	z value	Pr(>|z|)
rural:edad	-0.00	0.00	-1.05	0.30
rural:perfil_riesgoB	0.17	0.10	1.74	0.08
rural:perfil_riesgoC	0.23	0.12	1.92	0.05
rural:numero_plazos	0.00	0.00	0.65	0.51
rural:solic_prev	-0.39	0.30	-1.32	0.19
rural:porcentaje	-0.35	0.73	-0.48	0.63
edad:perfil_riesgoB	0.00	0.00	0.26	0.79
edad:perfil_riesgoC	-0.00	0.01	-0.74	0.46
edad:numero_plazos	0.00	0.00	1.02	0.31
edad:solic_prev	0.01	0.01	0.83	0.41
edad:porcentaje	0.02	0.03	0.56	0.58
perfil_riesgoB:numero_plazos	0.00	0.00	0.26	0.79
perfil_riesgoC:numero_plazos	-0.01	0.01	-1.47	0.14
perfil_riesgoB:solic_prev	-0.12	0.32	-0.37	0.71
perfil_riesgoC:solic_prev	-0.19	0.41	-0.47	0.64
perfil_riesgoB:porcentaje	0.02	0.83	0.02	0.98
perfil_riesgoC:porcentaje	0.94	0.76	1.23	0.22
numero_plazos:solic_prev	-0.03	0.02	-1.15	0.25
numero_plazos:porcentaje	0.06	0.04	1.36	0.17
solic_prev:porcentaje	0.54	2.36	0.23	0.82

Tabla 5.7: Resultados de la regresión logística con interacciones 2/2.

	Df	Deviance	AIC	LRT	Pr(>Chi)
<none>		37631.23	37703.23		
EVM1:rural	1.00	37631.51	37701.51	0.29	0.59
EVM1:edad	1.00	37631.29	37701.29	0.06	0.80
EVM1:perfil_riesgo	2.00	37633.68	37701.68	2.45	0.29
EVM1:numero_plazos	1.00	37637.50	37707.50	6.27	0.01
EVM1:solic_prev	1.00	37631.54	37701.54	0.32	0.57
EVM1:porcentaje	1.00	37634.16	37704.16	2.94	0.09
rural:edad	1.00	37632.33	37702.33	1.10	0.29
rural:perfil_riesgo	2.00	37635.39	37703.39	4.16	0.12
rural:numero_plazos	1.00	37631.65	37701.65	0.43	0.51
rural:solic_prev	1.00	37633.10	37703.10	1.88	0.17
rural:porcentaje	1.00	37631.46	37701.46	0.24	0.63
edad:perfil_riesgo	2.00	37632.84	37700.84	1.61	0.45
edad:numero_plazos	1.00	37632.27	37702.27	1.04	0.31
edad:solic_prev	1.00	37631.91	37701.91	0.68	0.41
edad:porcentaje	1.00	37631.53	37701.53	0.31	0.58
perfil_riesgo:numero_plazos	2.00	37636.65	37704.65	5.42	0.07
perfil_riesgo:solic_prev	2.00	37631.44	37699.44	0.22	0.90
perfil_riesgo:porcentaje	2.00	37633.14	37701.14	1.92	0.38
numero_plazos:solic_prev	1.00	37632.61	37702.61	1.38	0.24
numero_plazos:porcentaje	1.00	37633.07	37703.07	1.85	0.17
solic_prev:porcentaje	1.00	37631.28	37701.28	0.05	0.82

Tabla 5.8: Resultados de prueba ANOVA de la regresión logística con interacciones.

5.2.2. AUCs para diferentes particiones entrenamiento-prueba

Adicionalmente se calculó el AUC para diferentes particiones en train-test de la base para observar si el comportamiento del AUC pudiera depender de la elección de las muestras, se realizaron 100 particiones distintas de la base de las cuales se puede observar un resúmen de dichos AUCs en la tabla 5.9.

	Valor
Min.	0.82
1st Qu.	0.82
Median	0.82
Mean	0.82
3st Qu.	0.82
Max.	0.83
Sd.	0.003

Tabla 5.9: Resúmen AUCs para diferentes particiones entrenamiento-prueba en el modelo Logit.

Como se puede observar el AUC solo tiene una variación muy pequeña ya que la diferencia entre el valor máximo y el mínimo es pequeña, además de que la desviación estandar también es un valor pequeño.

Este mismo procedimiento, con las mismas particiones se realizó para el modelo con interacciones obteniendo lo que se muestra en la tabla 5.10

	Valor
Min.	0.82
1st Qu.	0.82
Median	0.82
Mean	0.82
3st Qu.	0.82
Max.	0.83
Sd.	0.0028

Tabla 5.10: Resúmen AUCs para diferentes particiones entrenamiento-prueba en el modelo logit con interacciones.

Como se puede observar en la tablas 5.9 y 5.10 la direncia entre el uso de interacciones de variables en el modelo logit no cambia los resultados del AUC de menera significativa ya que la diferencia entre dichas tablas es mínima.

5.3. Modelo LDA

Para este modelo se utilizaron los mismos datos ya comentados para el modelo logístico, la intención de usar este modelo es observar la capacidad predictiva de un modelo más simple para nuestros datos de comportamiento crediticio, por esta razón solo se utilizara en su forma más simple únicamente como referencia.

```
Prior probabilities of groups:
   0    1
0.57 0.43


Group means:
   EVM1 rural edad perfil_riesgoB perfil_riesgoC numero_plazos
0 0.012  0.19   36           0.26           0.50            21
1 0.604  0.20   34           0.25           0.54            25
```

```
  solic_prev porcentaje
0      0.025       0.27
1      0.011       0.27


Coefficients of linear discriminants:
                   LD1
EVM1             2.921
rural            0.018
edad            -0.011
perfil_riesgoB   0.132
perfil_riesgoC   0.283
numero_plazos    0.011
solic_prev      -0.166
porcentaje      -1.028
```

Como podemos observar el coeficiente mayor es el de la variable EVM1, además de ser positivo, lo cual indica una relación positiva (aunque no necesariamente hay causalidad) entre observar tres demoras consecutivas hacia el final dado que se observan al menos tres demoras consecutivas al comienzo del crédito. El área bajo la curva ROC del modelo para los datos de entrenamiento fue de

$$AUC = 0.794$$

Para los datos de prueba se puede visualizar la curva en la Figura 5.3 y obtuvo un área

$$AUC = 0.796.$$

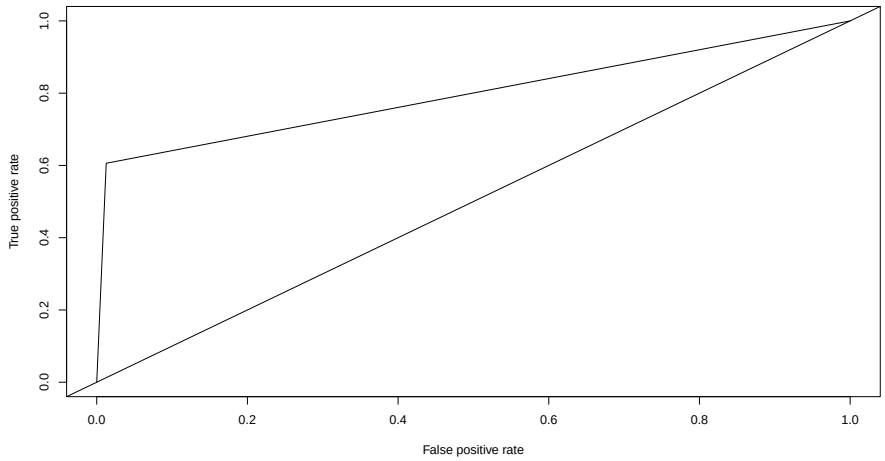

Figura 5.3: Curva ROC del modelo LDA con los datos de prueba.

5.4. Modelo SVM

El área bajo la curva ROC del modelo con maquinas de soporte vectorial para los datos de entrenamiento fue de

$$AUC = 0.794.$$

Para los datos de prueba se puede visualizar la curva en la Figura 5.4 y obtuvo un área bajo la curva ROC igual a

$$AUC = 0.796.$$

La matriz de confusión con los datos de prueba para el modelo SVM fue similar al modelo LOGIT reportada en la Tabla 5.5.

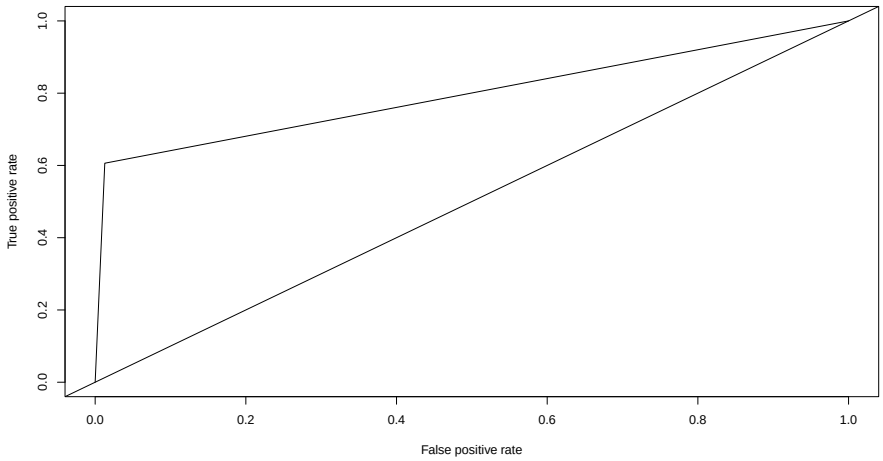

Figura 5.4: Curva ROC del modelo SVM con los datos de prueba.

Cabe destacar que se realizo una comparación con diferentes kernels del modelo SVM (radial, lineal, polinomial y sigmoide) de los cuales los resultados reportados son los correspondientes al kernel radial, ya que son los que mejores resultados producen, seguidos por los del lineal, el polinomial y siendo los del sigmoide los de peor comportamiento con un AUC de 0.7108.

5.4.1. AUCs para diferentes particiones entrenamiento-prueba

También, como en el modelo logit, se calculó el AUC para diferentes particiones en train-test para el kernel radial, esto para 20 diferentes elecciones datos, es importante mencionar que se decidió realizarlo únicamente 20 debido a las limitaciones computacionales. Los resultados obtenidos se presentan en la tabla 5.11.

	Valor
Min.	0.7885
1st Qu.	0.7929
Median	0.7941
Mean	0.7943
3st Qu.	0.7954
Max.	0.7991
Sd.	0.0026

Tabla 5.11: Resúmen AUCs para diferentes particiones entrenamiento-prueba modelo SVM.

5.5. Modelo de Bosques Aleatorios

En la Tabla 5.12 podemos apreciar la importancia de las variables para el modelo de bosques aleatorios[1]:

	IncNodePurity
EVM1	4310.61
numero_plazos	270.98
edad	165.55
perfil_riesgo	20
rural	11.9
solic_prev	11.4

Tabla 5.12: Importancia de las variables en el modelo de bosques aleatorios.

[1]Se calcula utilizando el comando varImpPlot del paquete mtcars de R. Es una métrica basada en medir la precisión del modelo al excluir la variable en turno.

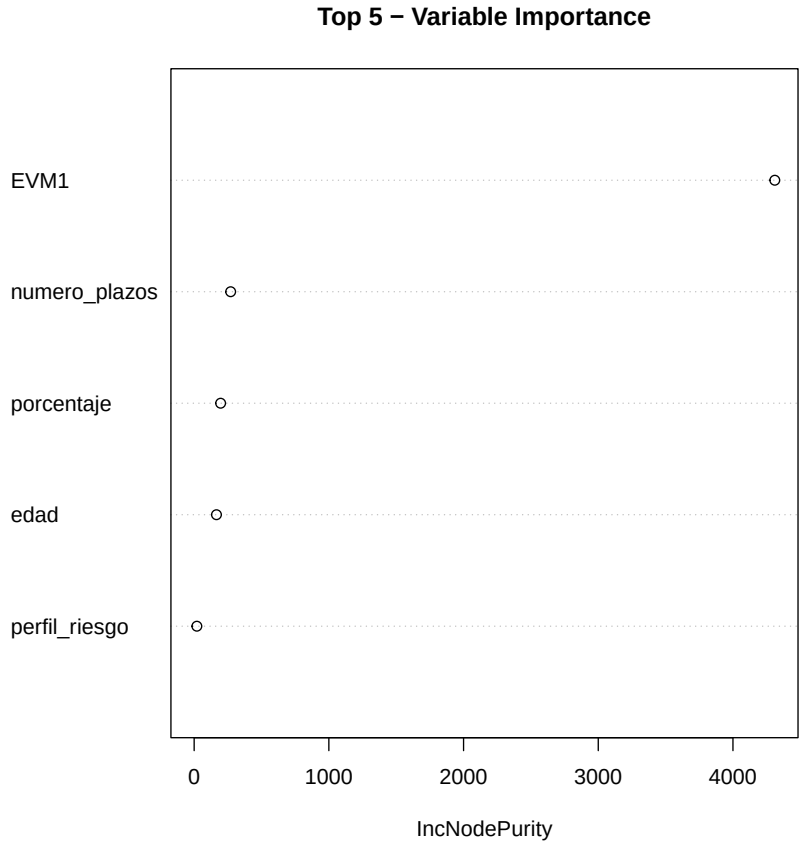

Figura 5.5: Importancia de las variables en el modelo de bosques aleatorios.

Como podemos observar en la Figura 5.5 la variable más importante es EVM1 lo cual muestra un cierto grado de consistencia con los otros modelos, en segundo lugar viene "numero_plazos", aunque en mucho menos medida que EVM1.

La Figura 5.6 muestra como el error de modelo de bosques aleatorios se va estabilizando con forme aumenta el número de arboles.

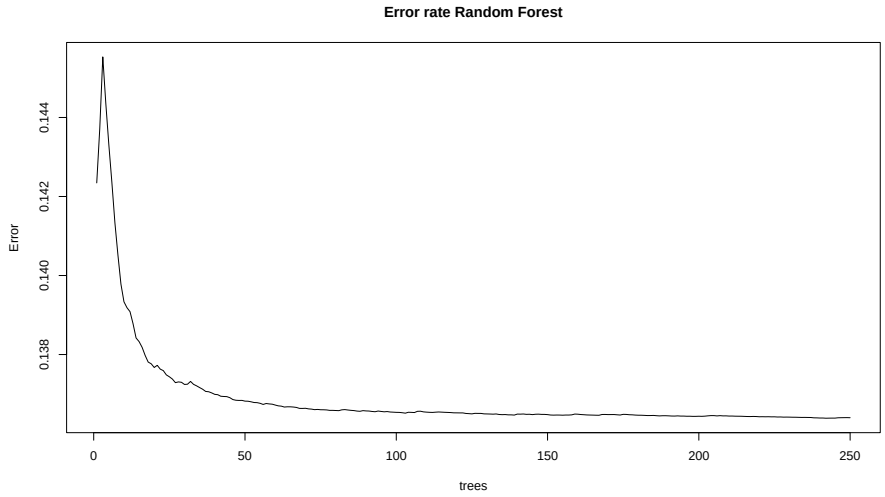

Figura 5.6: Evolución del error en el modelo de bosques aleatorios.

El área bajo la curva ROC del modelo de bosques aleatorios para los datos de entrenamiento fue de

$$AUC = 0.8759.$$

Para los datos de prueba se puede visualizar la curva en la Figura 5.7 y obtuvó un área

$$AUC = 0.8291.$$

La matriz de confusión con los datos de prueba para el modelo de bosques aleatorios se visualiza en la Tabla 5.13.

	No-Evento	Evento	
Pronostico:No-evento	0.987	0.402	
Pronostico:Evento	0.012	0.597	
Exactitud (Accuracy)			0.8207

Tabla 5.13: Matriz de confusión modelo de bosques aleatorios.

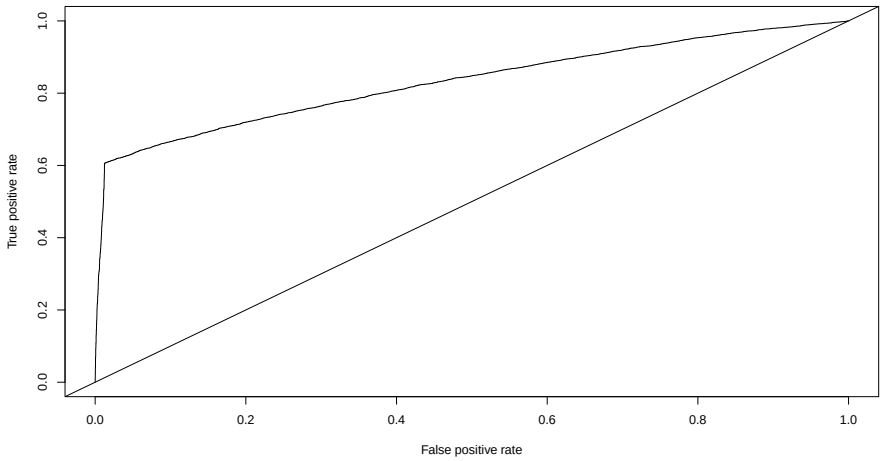

Figura 5.7: Curva ROC del modelo de bosques aleatorios con los datos de prueba.

5.6. Malla en el número de árboles

Para explorar el efecto que tiene el número de árboles en el poder de ajuste del modelo de bosques aleatorios hicimos el experimento de recorrer una malla de diez en diez comenzando en 200 y terminando en 500 árboles. El comportamiento del AUC se visualiza en la Figura 5.8. Puede verse una estabilidad con respecto a este parámetro. Comportamientos análogos presentan las tasas de éxito de pronóstico de la presencia/ausencia del evento crediticio; ver Figuras 5.9 y 5.10.

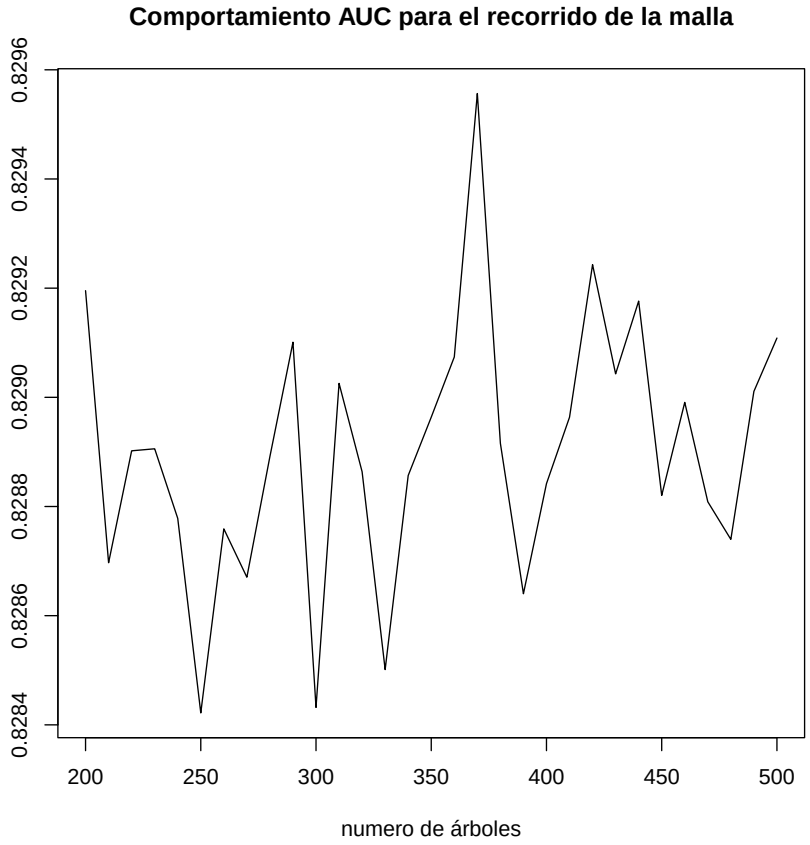

Figura 5.8: Comportamiento del AUC en función del número de árboles.

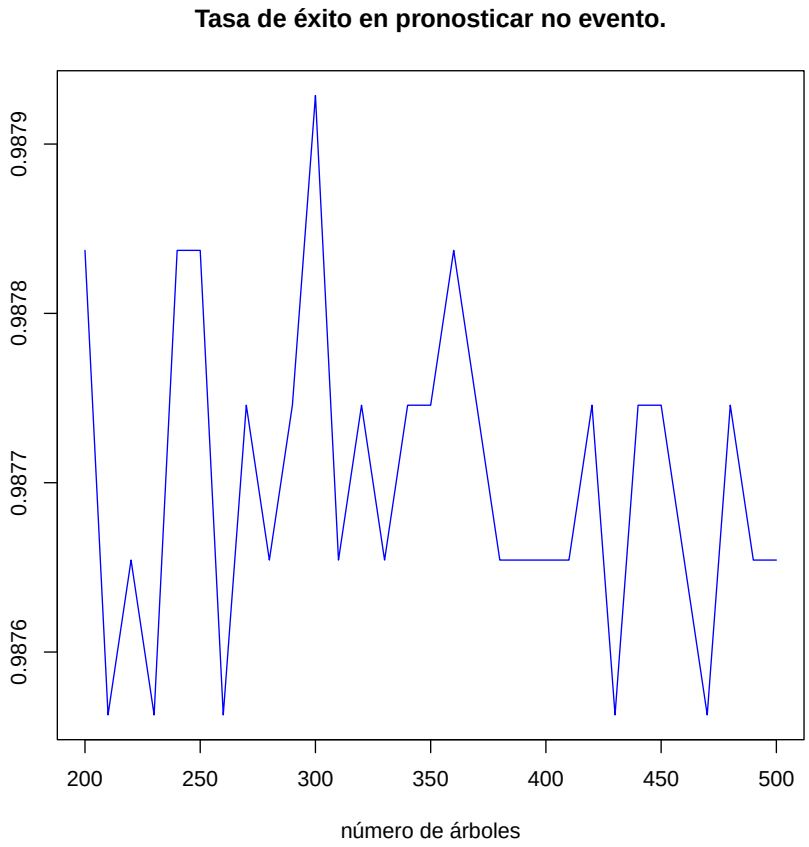

Figura 5.9: Comportamiento de la tasa de éxito de pronóstico de la ausencia del evento crediticio como función del número de árboles.

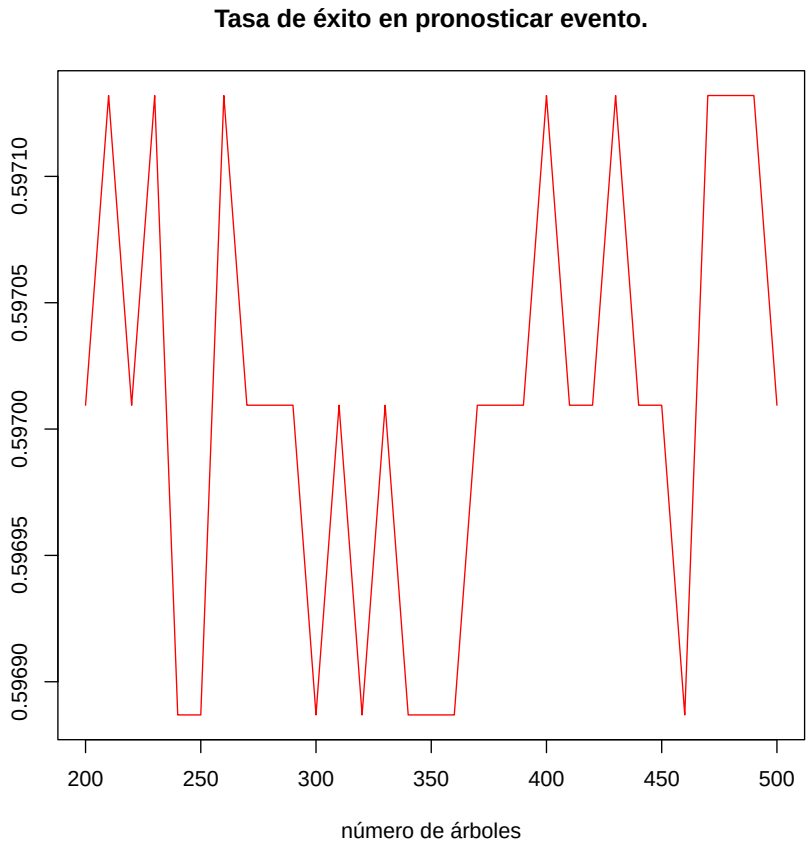

Figura 5.10: Comportamiento de la tasa de éxito de pronóstico de la presencia del evento crediticio como función del número de árboles.

5.6.1. AUCs para diferentes particiones entrenamiento-prueba

Para el modelo de Random Forest también se obtuvieron los AUC para diferentes particiones entrenamiento-prueba de la base, se realizaron 20 particiones distintas de las cuales se puede observar un resúmen de dichos AUCs en la tabla 5.14.

	Valor
Min.	0.8237
1st Qu.	0.8267
Median	0.8283
Mean	0.8287
3st Qu.	0.8310
Max.	0.8332
Sd.	0.0029

Tabla 5.14: Resúmen AUCs para diferentes particiones entrenamiento-prueba en el modelo de bosques aleatorios.

Como se puede observar el AUC solo tiene una variación muy leve ya que la diferencia entre el valor máximo y el mínimo es pequeña, además de que la desviación estándar también es un valor no muy grande.

Capítulo 6

Conclusiones

Con base en los resultados del Capítulo 5 generamos la Tabla 6.1 con un comparativo del desempeño de los modelos ajustados en datos de prueba.

Modelo	AUC
Modelo logístico	0.827
Media modelo logístico (100)	0.828
Media modelo logístico interacciones (100)	0.827
LDA	0.797
SVM	0.798
SVM media (20)	0.794
Bosques Aleatorios	0.83
Media Bosques Aleatorios (20)	0.829

Tabla 6.1: Comparativo del desempeño de diversos modelos.

El mejor modelo para nuestros datos resultó ser el de bosques aleatorios con un $AUC = 0.83$ para los datos de prueba, sin embargo la regresión logística fue competitiva con una $AUC = 0.82$. Es decir, la diferencia en el desempeño es muy baja y se podría valorar el uso de regresión logística considerando que la especificación de este modelo tiene una fácil interpretación. Se debe valorar si el aumento en la complejidad del modelo de bosques aleatorios no acota la importancia de su desempeño ligeramente superior.

Por otro lado, es de observarse que incluso los modelos peor calificados (LDA y SVM) tuvieron un desempeño competitivo. Consideramos que lo anterior se

consiguió debido a la definición de la variable a explicar y el hecho de que se haya ajustado un modelo de calificación de comportamiento y no así de originación de crédito.

En resumen tenemos los siguientes hallazgos:

- Se desarrollaron modelos de comportamiento y no así de originación debido a la limitante de información.
- En general, todos los modelos fueron competitivos, especialmente el modelo logístico que usualmente es el modelo de referencia. El modelo que tuvó mejor desempeño fue el de bosques aleatorios.
- Es interesante que en nuestro caso el modelo LDA tuvo un desempeño aceptable a pesar que en la literatura se reporta una pobre capacidad explicativa.
- El modelo logit es el de más fácil interpretación y todas la variables tuvieron coeficientes cuyos signos hacen sentido; ver el párrafo siguiendo la Tabla 5.3.
- Los modelos que se desarrollaron son estables con respecto a la métrica AUC al hacer repetidos muestreos y ajuste de modelos.
- Las interacciones de variables no aportaron poder predictivo substancial en comparación a las variables originales.
- En las matrices de confusión de los modelos se observa que en general la tasa de éxito del pronóstico de la no aparición del evento crediticio es casi perfecta (mayor a 98 % en el caso del modelo logístico y bosques aleatorios). La tasa de un pronóstico correcto de la aparición del evento fue cercano al 60 % en el caso del modelo logístico y bosques aleatorios.

Apéndice A

Código R

En este capítulo de apéndice reportamos el código R utilizado en la generación de modelos.

A.1. Preliminares y generación de modelos

```
#------------------------------------------------------------------
#PRELIMINARES
#------------------------------------------------------------------
library(ROCR)
library(caret)
library(MASS)
library(e1071)
library(randomForest)

setwd("mydir/")
#------------------------------------------------------------------
#CARGAR DATOS
#------------------------------------------------------------------
#Bases de datos
```

```
BMod<-read.csv("Datos.csv")
#BMod$EVM1=as.factor(BMod$EVM1)
#BMod$EVM2=as.factor(BMod$EVM2)

set.seed(12345)
row_index<-sample(1:nrow(BMod),0.7*nrow(BMod))
train<-BMod[row_index,]
test<-BMod[-row_index,]

#-----------------------------------------------------------------
#SECCION FUNCIONES
#-----------------------------------------------------------------
f.output= function(modelo,name,data=test)
{
    pred_modelo<-predict(modelo,newdata=data,type="response")

    #matriz de confusion
    if(!class(pred_modelo)=="factor")
    {
        cm=confusionMatrix(as.factor(ifelse(pred_modelo>0.5,1,0)),
                           as.factor(data$EVM2))
    }else
    {
        cm=confusionMatrix(pred_modelo,as.factor(data$EVM2))
    }

    x=cm$table
```

```
    x0=cbind(x[,1]/sum(x[,1]), x[,2]/sum(x[,2]))
    colnames(x0)=c("No evento","Evento")
    rownames(x0)=c("Pronosticado:No evento","Pronosticado:Evento")

    #curva roc
    pred1<-prediction(as.numeric(pred_modelo),data$EVM2)
    perf1<-performance(pred1,"tpr","fpr")
    pdf(paste0(name,".pdf"))
    plot(perf1,main=paste0("curva ROC del modelo ", name))
    abline(a=0,b=1)
    dev.off()
    #AUC
    auc_ROC<-performance(pred1,measure = "auc")
    auc=auc_ROC@y.values[[1]]
    #guardar resumen modelo
    sink(paste0(name,"summary.txt"))
    print("Resumen modelo ")
    print(summary(modelo))
    print("Matriz de confusion")
    print(x0)
    print("AUC")
    print(paste0("AUC = ",auc))
    sink()
    return(list(x0,auc))
}
#------------------------------------------------------------------
#SECCION MODELOS
```

```
#-----------------------------------------------------------------
#MODELO LOGIT
logitmod<-glm(formula=EVM2~.,family = "binomial",data = train)
summary(logitmod)
f.output(logitmod,"logit")


#MODELO LOGIT CON INTERACCIONES
logitmod2<-glm(formula=EVM2~.^2,family = "binomial",data = train)
summary(logitmod2)
f.output(logitmod2,"logitInteracciones")



#MODELO LDA
modlda<-lda(EVM2~.,train)
summary(modlda)
pred_lda<-predict(modlda,newdata = test)
pred2<-prediction(as.numeric(pred_lda$class),test$EVM2)
perf2<-performance(pred2,"tpr","fpr")
par(mfrow=c(1,1))
plot(perf2)
abline(a=0,b=1)
auc_ROC<-performance(pred2,measure = "auc")
auc_ROC@y.values[[1]]

#MODELO SVM
```

```
A1=Sys.time()
modsvm<-svm(EVM2~.,data=train,
            type="C-classification",kernel='radial',scale = T)
summary(modsvm)
f.output(modsvm,"SVM")
B1=Sys.time()
B1-A1

A2=Sys.time()
#MODELO SVM CON INTERACCIONES
modsvm2<-svm(EVM2~.^2,data=train,
            type="C-classification",kernel='radial',scale = T)
summary(modsvm2)
f.output(modsvm2,"SVMInteracciones")
B2=Sys.time()
B2-A2

#-----------------------------------------------------------------
#SECCION MODELOS RANDOM FOREST
#-----------------------------------------------------------------
#MODELO RANDOM FOREST
(A3=Sys.time())
modRF<-randomForest(EVM2~.,data=train,ntree=250)
summary(modRF)
f.output(modRF,"bosquesAleatorios",data=test)
#Importancia de las variables
sink("ImportancemodRF.txt")
```

```
importance(modRF)
sink()
pdf("varImpPlot.pdf")
varImpPlot(modRF,sort=T,n.var = 5,
           main="Top 5 - Variable Importance")
dev.off()
B3=Sys.time()
B3-A3

#MODELO RANDOM FOREST
(A4=Sys.time())
modRF2<-randomForest(EVM2~.^2,data=train,ntree=250)
summary(modRF2)
f.output(modRF2,"bosquesAleatoriosInteracciones")
#Importancia de las variables
importance(modRF2)
varImpPlot(modRF2,sort=T,n.var = 5,
           main="Top 5 - Variable Importance")
B4=Sys.time()
B4-A4



#-----------------------------------------------------------------
#SECCION MODELOS RANDOM FOREST MALLA
#-----------------------------------------------------------------
#MODELO RANDOM FOREST
(A5=Sys.time())
```

```
modelos=list()
auc=c()
matrices=list()
coun=0
malla =seq(200,500, by=10)
for(numt in malla)
{
    coun=coun+1
    modRF3<-randomForest(EVM2~.,data=train,ntree=numt)
    modelos[[coun]]=modRF3
    summary(modRF3)
    x=f.output(modRF3,paste0("MallabosquesAleatorios",numt))
    auc[coun]=x[[2]]
    matrices[[coun]]=x[[1]]
}
B5=Sys.time()
B5-A5
```

A.2. Para repeticiones

A continuación se muestra el código usado para generar repeticiones de los modelos para diferentes particiones de entrenamiento-prueba.

```
modLogit<-list()
perf<-list()
AUC<-list()

MMRL<-function(i,BMod,c=1){
  set.seed(i*197)
```

```
  row_index<-sample(1:nrow(BMod),0.7*nrow(BMod))
  train<-BMod[row_index,]
  test<-BMod[-row_index,]
  if(c==1){
    logitmod<-glm(formula=EVM2~.,family = "binomial",data = train)
  }else{
    logitmod<-glm(formula=EVM2~.^2,family="binomial",data = train)
  }
  pred_logit<-predict(logitmod,newdata=test)
  table(ifelse(pred_logit>0.5,1,0),test$EVM2)
  pred1<-prediction(as.numeric(pred_logit),test$EVM2)
  perf1<-performance(pred1,"tpr","fpr")
  auc_ROC<-performance(pred1,measure = "auc")
  return(list(logitmod,table(ifelse(pred_logit>0.5,1,0),
                             test$EVM2),
              perf1,auc_ROC@y.values[[1]]))
}

RR<-lapply(as.list(1:1000),function(i) MMRL(i,BMod))
RRI<-lapply(as.list(1:100),function(i) MMRL(i,BMod,2))
AUC<-sapply(1:100, function(i) RR[[i]][[4]])
AUCI<-sapply(1:100, function(i) RRI[[i]][[4]])

MMSVM<-function(i,BMod,ker='radial'){
  set.seed(i*197)
  row_index<-sample(1:nrow(BMod),0.7*nrow(BMod))
  train<-BMod[row_index,]
```

```
  test<-BMod[-row_index,]
  modsvm<-svm(EVM2~.,data=train,
              type="C-classification",kernel=ker,scale = T)
  summary(modsvm)
  pred_svm<-predict(modsvm,newdata=test)
  pred3<-prediction(as.numeric(pred_svm),test$EVM2)
  perf3<-performance(pred3,"tpr","fpr")
  auc_ROC<-performance(pred3,measure = "auc")
  return(list(modsvm,table(pred_svm,test$EVM2),perf3,
              auc_ROC@y.values[[1]]))
}


clsvm<-apply(data.frame(a=as.numeric(gl(5,4)),
                        b=rep(c('radial','linear',
                                'polynomial','sigmoid'),5)),1,
                        function(i) MMSVM(as.numeric(i[1]),
                                          BMod,i[2]))
clsvm2<-apply(data.frame(a=seq(20),
                        b=rep('radial',20)),1,
             function(i) MMSVM(as.numeric(i[1]),BMod,i[2]))
clsvm<-mapply(MMSVM())


MMRF<-function(i,BMod){
  set.seed(i*197)
  row_index<-sample(1:nrow(BMod),0.7*nrow(BMod))
```

```
  train<-BMod[row_index,]
  test<-BMod[-row_index,]
  modRF<-randomForest(EVM2~.,data=train,ntree=250)
  pred_RF<-predict(modRF,newdata=test)
  pred4<-prediction(as.numeric(pred_RF),test$EVM2)
  perf4<-performance(pred4,"tpr","fpr")
  auc_ROC<-performance(pred4,measure = "auc")
  auc_ROC@y.values[[1]]
  return(list(modRF,table(pred_RF,test$EVM2),perf4,
              auc_ROC@y.values[[1]]))
}

clRF<-lapply(as.list(1:20),function(i) MMRF(i,BMod))
```

Apéndice B

Figuras Adicionales

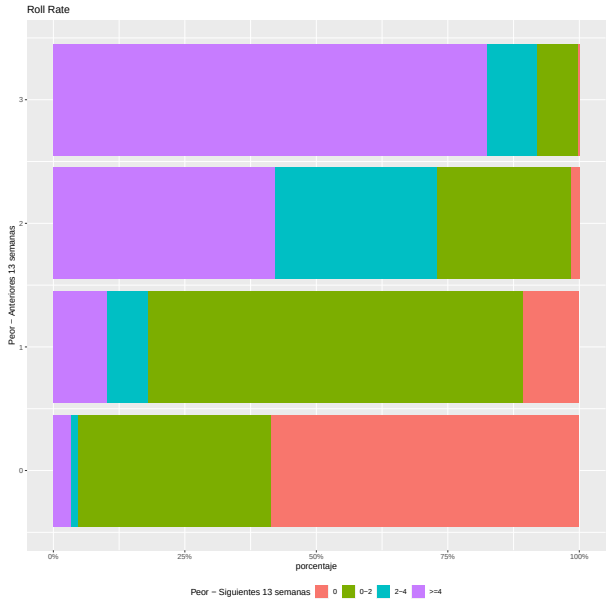

Figura B.1: Análisis de tasa de balanceo (26 semanas) para cuatro semanas de morosidad. Elaboración propia.

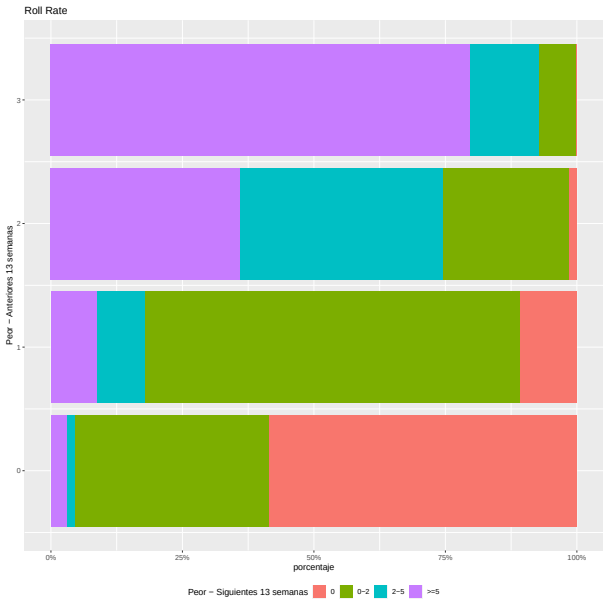

Figura B.2: Análisis de tasa de balanceo (26 semanas) para cinco semanas de morosidad. Elaboración propia.

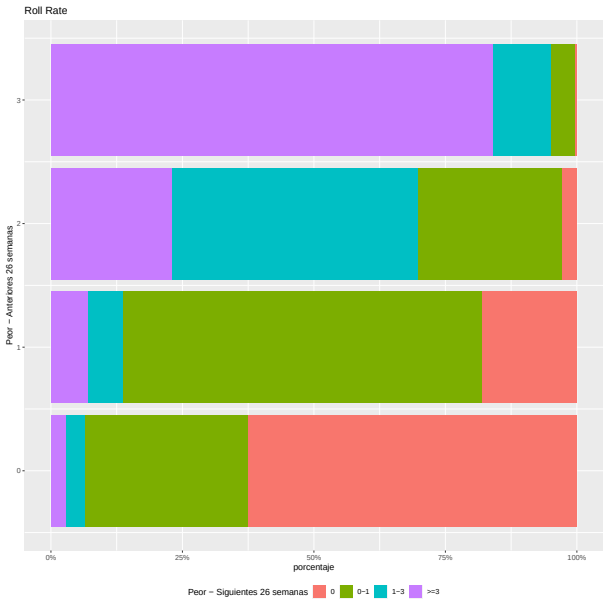

Figura B.3: Análisis de tasa de balanceo (52 semanas) para tres semanas de morosidad. Elaboración propia.

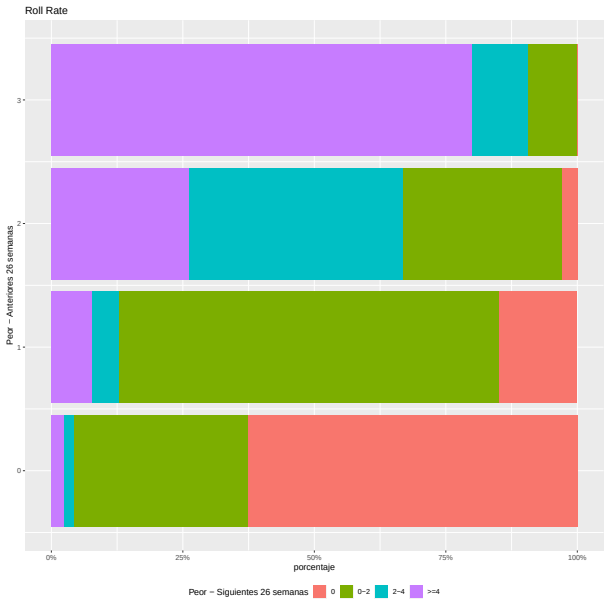

Figura B.4: Análisis de tasa de balanceo (52 semanas) para cuatro semanas de morosidad. Elaboración propia.

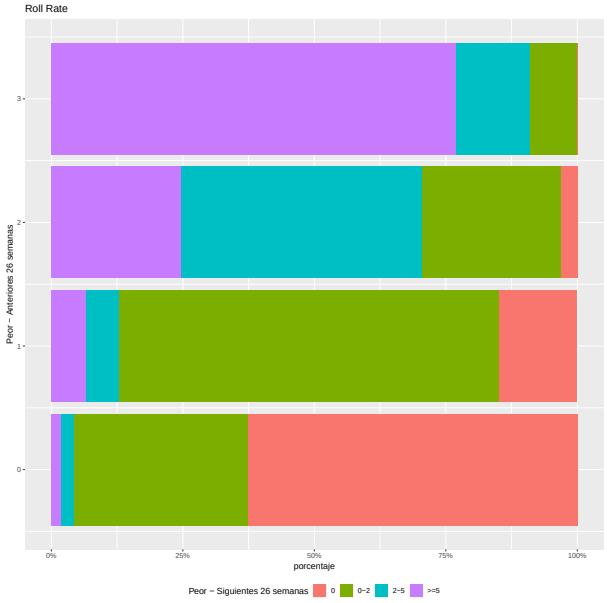

Figura B.5: Análisis de tasa de balanceo (52 semanas) para cinco semanas de morosidad. Elaboración propia.

Bibliografía

[1] Baesens, B., Lessmann, S., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

[2] Breiman, L. (2001). Random forest. *Machine Learning*, 45(1):5–32.

[3] Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. *Conf. on Knowledge Discovery and Data Mining*, page 155–164.

[4] Egan, J. P. (1975). Signal detection theory and roc analysis. series in cognitition and perception. *Academic Press*.

[5] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

[6] Gunnarsson, B. R., van den Broucke, S., Baesens, B., Óskarsdóttir, M., and Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, pages 292–305.

[7] Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, second edition.

[8] Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1):489–501.

[9] Johnson, R. A. and Wichern, D. W. (2019). *Applied multivariate statistical analysis*. Pearson, New Jersey.

[10] Louzada, F., Ara, A., and Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2):117–134.

[11] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Champman and Hall/CRC, second edition.

[12] Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Kluwer Academic Publishers*.

[13] Siddiqi, N. (2006). *Credit risk scorecards developing and Implementing Intelligent Credit scoring*. Wiley.

[14] USAID (2006). The credit process - united states agency for international development. https://pdf.usaid.gov/pdf_docs/pnadq084.pdf.

[15] Weinstein, M. C. and Fineberg, H. V. (1980). Clinical decision analysis. *PA: W. B. Saunders Company*.

[16] Ye, J., Janardan, R., and Li, Q. (2004). Two-dimensional linear discriminant analysis. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press.