Modelos de mezclas Bernoulli con regresión Logística: Una aplicación en la valoración de carteras de crédito.

Trabajo final de Maestría para optar al título de:
Maestría en Ciencias - Estadística

Presentada por:
Esteban Tabares Alzate
estabaresal@unal.edu.co

Asesor:
Norman Diego Giraldo, MS Matemáticas.
ndgirald@unal.edu.co

UNIVERSIDAD NACIONAL DE COLOMBIA
Sede Medellín
Escuela de Estadística
Facultad de Ciencias

Octubre, 2019

Contenidos

Lista de figuras

Lista de tablas

Resumen

Este trabajo final de maestría, modalidad de profundización, consiste en la elaboración de un problema de modelacióon estadística aplicada al sector crediticio. El objetivo es aplicar un modelo de regresión logística o un modelo de aprendizaje de máquina para calcular la probabilidad de default e incorporarla en la fórmula para hallar la distribución de costos totales en el modelo de Mezclas Bernoulli, con el fin de estimar valores de cuartiles superiores de la distribución de los costos totales, denominados la provisión. Para cumplir lo anterior se deben calcular las probabilidades de incumplimiento (o de default) después de realizar una competencia entre modelos vía mejor medida de ajuste (AUC), precisión y rango del intervalo de precisión (IC_precisión); si bien el título solo menciona regresión logística, este modelo competirá con modelos de aprendizaje de máquina como árboles aleatorios, bosques aleatorios, Knn y máquinas de soporte vectorial y con el de mejor AUC, precisión y IC_precisión se calcularán dichas probabilidades de default.

Además, se calcula la distribución aproximada del monto total de las pérdidas por incumplimiento para créditos originados entre 2014 y 2018. Tales costos totales se modelan mediante ciertos tipos de sumas de variables aleatorias que se denominan Mezclas Bernoulli para, finalmente, evaluar el capital expuesto de una cartera de créditos y así entender el grado de deterioro de esta cartera para créditos originados entre 2014 y 2018.

Palabras claves Riesgo de crédito, distribuciones de Mezclas Bernoulli, regresión logística, modelos de aprendizaje de máquina, distribución de pérdidas, VaR y TVaR.

Abstract

This final master's work, deepening modality, consists in the elaboration of a statistical modeling problem applied to the credit sector. The objective is to apply a logistic regression model or a machine learning model to calculate the probability of default and incorporate it in the formula to find the distribution of total costs in the Bernoulli Blends model, in order to estimate values of higher quartiles of the distribution of total costs, called the provision. To accomplish the above, the probabilities of default must be calculated after competition between models via best fit measurement (AUC), precision and precision interval range (IC_precision); although the title only mentions logistic regression, this model will compete with machine learning models such as random trees, random forests, Knn and vector support machines and with the best AUC, precision and IC_precision these default probabilities will be calculated.

In addition, the approximate distribution of the total amount of default losses is calculated for credits originated between 2014 and 2018. These total costs are modeled using certain types of random variable sums called Bernoulli Blends to finally evaluate the exposed capital of a loan portfolio and thus understand the degree of impairment of this portfolio for loans originated between 2014 and 2018.

Keywords Credit risk, Bernoulli Blend distributions, logistic regression, machine learning models, loss distribution, VaR and TVaR.

Capítulo 1

Introducción

Este trabajo final de maestría, modalidad de profundización, consiste en la elaboración de un problema de modelación estadística aplicada al sector crediticio. La ocurrencia periódica de incumplimientos de créditos, con consecuencia económica apreciable, motiva el problema de cuantificar los costos de tales contingencias. Parte del análisis consiste en calcular las probabilidades de incumplimiento (o de default) después de realizar una competencia entre modelos vía mejor medida de ajuste (AUC) [1]; si bien el título solo menciona regresión logística, este modelo competirá con modelos de aprendizaje de máquina como árboles aleatorios, bosques aleatorios, Knn y máquinas de soporte vectorial y con el de mejor AUC calcular dichas probabilidades de default. También se calculará la distribución aproximada del monto total de las pérdidas por incumplimiento desde el año 2014 hasta 2018. Tales costos totales se modelan mediante ciertos tipos de sumas de variables aleatorias que se denominan Mezclas Bernoulli probando varios enfoques para la calibración de modelos basados en la representación de la mezcla de Bernoulli.

Las probabilidades asociadas con estas variables Bernoulli dependen de variables exógenas que se asumen contienen información individual (efectos fijos). Entonces, el objetivo de este trabajo final de mestría es aplicar un modelo de regresión logística o un modelo de aprendizaje de máquina con efectos fijos para calcular la probabilidad de default e incorporarla en la fórmula para la distribución de costos totales en el modelo de Mezclas Bernoulli, con el fin de estimar valores de cuartiles superiores de la distribución de los costos totales, denominados la provisión. El modelo utilizado, sin embargo, introduce variables condicionales con el fin de incorporar correlaciones entre los créditos que incumplen.

[1] A mayor AUC, mejor modelo.

La revisión de literatura realizada sobre riesgo de crédito ha sido McNeil y Wendin (2003) donde se analizan modelos de riesgo de cartera de crédito. En Bluhm, Overbeck, y Wagner (2016) se explica para el sector financiero, decisiones de la gestión de carteras y una financiación estructurada, con base a sólidos conocimientos cuantitativos. En Giraldo (2014) se plantean y desarrollan modelos cuantitativos utilizados en el campo de gestión del riesgo financiero. En Huang y Oosterlee (2011) se describe el planteamiento y desarrollo del modelo de riesgo de crédito CreditRisk. Finalmente se menciona Crouhy, Galai, y Mark (2000) en el que se presenta una revisión de los principales modelos de riesgo de crédito.

Capítulo 2

Planteamiento del problema

El problema consiste en realizar un análisis de la probabilidad de incumplimiento utilizando un modelo logístico con efectos fijos, e incorporar esta probabilidad en las fórmulas para calcular medidas de riesgo de carteras de crédito. Estas fórmulas se basan en expresiones que buscan aproximar de manera eficiente cuantiles altos de la distribución de pérdidas, que es una variable aleatoria definida como una suma ponderada de variables dicótomas Bernoulli. El plan de análisis se realizará con base en información (reservada) de una entidad bancaria, que consiste en datos de la forma:

$$\begin{aligned} D_{i,j}, Y_{i,j}, X_i, F_j, w_j \\ j = 1, 2, \ldots, N, \\ i = 1, 2, \ldots, n_i \\ N = n_1 + \ldots + n_k. \end{aligned}$$

Donde $Y_{i,j}$ son variables cualitativas-cuantitativas con información relevante de cada crédito. Las variables X_i son efectos fijos, que representan criterios de clasificación ó efectos en el tiempo de variables macroeconómicas, $D_{i,j}$ son variables dicótomas 0-1, que representan si un crédito incumple en un período y genera una pérdida económica. Las variables F_j son los saldos de las deudas, y w_j son las tasas de recuperación de la deuda después de subastar las garantías.

Capítulo 3

Marco teórico del problema propuesto

El trabajo busca dar solución a un problema cuantitativo propio del sector bancario. Consiste en estimar el capital que una entidad bancaria puede perder por concepto de incumplimiento de créditos, en un período de tiempo determinado. Este capital se estima hoy en día de manera obligatoria en todas la entidades, como una manera de ejercer vigilancia sobre el riesgo al cual se exponen éstas. Desde un punto de vista estadístico este capital se modela como una suma ponderada de variables dicótomas, 0,1, ó variables Bernoulli. El problema estadístico consiste en determinar los cuantiles altos de este tipo de variables aleatorias, y otras cantidades asociadas.

3.1. La variable de costo total L

El costo total de los incumplimientos (defaults) en el período $[0, T]$, se define como la variable aleatoria

$$L = \sum_{j=1}^{N} F_j(1 - w_j)D_j, \tag{3.1}$$

La cantidad $w_j \in [0, 1]$ es la tasa de recuperación, que puede ser aleatoria y que es el valor que la Entidad puede recuperar mediante el proceso de remate de la garantía (en caso de existir ésta). Muchas entidades bancarias utilizan garantías para los créditos mediante terceros, que son empresas que venden garantías a los usuarios (Compañías de Fianzas). En estos casos $w_j = 1$, y no interesan para este análisis. Algunos créditos utilizan un bien que se desea adquirir mediante el crédito. En este caso el bien es una garantía mediante un contrato de hipoteca.

Pero no se garantiza que al rematarlo se genere un valor igual al saldo adeudado, por lo que se asume $0 < w_j < 1$. La variable $LGD_j = F_j(1 - w_j)$ se define como la "pérdida dado el default".

El objetivo principal del modelo de mezclas Bernoulli consiste en poder evaluar aproximadamente la distribución $F_L(x) = \mathbb{P}(L \leq x)$, asumida concentrada en un retículo discreto finito (un conjunto de valores discreto). En particular, encontrar el Valor en Riesgo para un nivel de $100(1 - q)\,\%$, definido como el percentil de $1 - q$ de la distribución de L, $VaR_L(1 - q) = F_L^{-1}(1 - q)$. Por ejemplo, el $VaR(0.99)$ representa un valor de una pérdida que se supera solamente el $1\,\%$ de la veces en períodos de la misma duración. Estos valores corresponden a eventos de riesgo de crédito extremos, y serían atribuíbles al efecto de un riesgo sistemático sobre la economía.

Hay varios modelos en la literatura sobre riesgo de crédito para la variable L en:

1. El modelo de mezclas Bernoulli homogéneo con un factor.
2. El modelo Credit-Risk+

En este trabajo se utiliza el primero.

3.2. Modelos de mezclas Bernoulli

El modelo de mezclas Bernoulli se define como sigue. Se asume un portafolio de N créditos. A cada crédito se le asocia una variable aleatoria Bernoulli, denominada default,

$$D_j \in \{0, 1\} \sim Ber(p_j), \tag{3.2}$$

donde $p_j \in (0, 1)$ es la probabilidad de incumplimiento del crédito $\mathbb{P}(D_j = 1) = p_j$. El evento $(D_j = 1)$ se interpreta como que el crédito j-ésimo entró en default en el período en cuestión, que puede tratarse por ejemplo, de un año. Se define además una cantidad no aleatoria $F_j \geq 0$, el saldo del crédito j-ésimo, a la fecha del cálculo, denominado "exposición".

Se asume dado un vector de m variables aleatorias $\underline{X} = (X_1, \ldots, X_m)' \in R^m$, $N > m$. Este vector se denomina el riesgo sistemático. Las X_j se pueden interpretar como variables macroeconómicas. En McNeil y Wendin (2003) se hace enfásis en la importancia de incluir el efecto de riesgo sistemático en portafolios de crédito. Se extiende este modelo incorporando información sobre cada crédito, ó un riesgo idiosincrático, asumiendo un vector de r variables aleatorias

$\underline{Y}_j \in \mathbb{R}^r, j = 1, 2, \ldots, N.$

El supuesto básico es que cada variable default, D_j, dadas $\underline{X}, Y_j$, es una variable Bernoulli condicional. Las probabilidades de default dependen de la información proporcionada por $\underline{X}, \underline{Y}_j$. Se asume que existen N funciones continuas no decrecientes,

$$p_j : \mathbb{R}^{m+r} \to [0, 1], \ j = 1, 2, \ldots, N, \tag{3.3}$$

tales que $\underline{D}|\underline{X}, \underline{Y}_j$ es un vector de N variables Bernoulli condicionales, independientes, distribuídas

$$D_j|\underline{X}, \underline{Y}_j \sim Ber(p_j(\underline{X}, \underline{Y}_j)). \tag{3.4}$$

El modelo (3.3)-(3.4) se denomina modelo de mezclas Bernoulli.

Si en el modelo inicial (3.4) se toma $m = 1$ entonces $\underline{X} = X$ y solamente hay un factor que define el riesgo sistemático. Adicionalmente, si todas las funciones $p_j(.)$ son idénticas, $p_j(x) \equiv p(x)$, el modelo (3.4) se denomina modelo de mezclas Bernoulli homogéneo de un factor ó también modelo intercambiable (exchangeable Bernoulli mixture model), ver McNeil, Frey, Embrechts, et al. (2005). Es el modelo escogido en este trabajo.

$$D_j|X, \underline{Y}_j \sim Ber(p(X, \underline{Y}_j)). \tag{3.5}$$

3.3. Modelo para estimar la probabilidad de castigo: Regresión logística con efectos fijos

El objetivo es estimar la función $p(X_i, \underline{Y}_{i,j})$. La siguiente definición de regresión Logística con efectos fijos está tomada del reporte técnico Frey y McNeil (2003).

Se asumen r covariables que contienen información sobre cada crédito, conformadas como las columnas de la matriz $Y \in \mathbb{R}^{N \times r}$, $\underline{Y}_j \in \mathbb{R}^r, j = 1, 2, \ldots, N$

La variable aleatoria $X \sim N(0, \sigma^2)$ contiene información sobre el efecto aleatorio de un riesgo sistemático. Es supuesto es que esta variable puede ser un efecto macroeconómico en el tiempo, ó puede ser una variable de clasificación en categorías, que afecta por igual a todos los elementos de cada subgrupo de observaciones. Si es tiempo, cada subgrupo puede ser un año.

Se asumen n_i observaciones para cada subgrupo, $i = 1, 2, \ldots, k$, tales que $k_1 + k_2 + \ldots + k_n = N$. Las observaciones para el i-ésimo sector ó cluster son $D_{i,j} \in \{0, 1\}$, $j = 1, \ldots, n_i$. Es una variable dicótoma que representa el evento default del crédito j-ésimo. El modelo logístico con efectos fijos se define como (ver McNeil y Wendin (2003))

$$\mathbb{P}(D_{j,i} = 1 | \underline{Y}_{j,i}, X_i) = \frac{1}{1 + e^{-(\beta_0 + \alpha_i + \underline{Y}_{j,i}\underline{\beta} + X_i)}}. \tag{3.6}$$

donde $\underline{\beta} \in \mathbb{R}^p$, es el vector de parámetros de efectos fijos a estimar, β_0 es un intercepto promedio y α_i es un intercepto para cada grupo de n_i observaciones, y $\underline{X}_{i,j}$ es la celda i,j-ésima de X, del sujeto j-ésimo, dentro del grupo de n_i observaciones. En Demidenko (2013) se propone que el modelo (3.6) se puede dividir en dos modelos diferentes.

1) Pueden considerarse interceptos dependientes de i

$$\mathbb{P}(D_{j,i} = 1 | \underline{Y}_{j,i}, X_i) = \frac{1}{1 + e^{-\beta_{i,0} - \underline{Y}_{j,i}\underline{\beta}}}. \tag{3.7}$$

Los parámetros $\beta_{i,0}$ son fijos. Se podrían interpretar como $\beta_0 + \alpha_i = \beta_{i,0}$ en el modelo inicial en McNeil y Wendin (2003). El modelo es Logístico con efectos fijos, con interceptos fijos.

2) Se puede extender (3.6) colocando de nuevo

$$\mathbb{P}(D_{j,i} = 1 | \underline{Y}_{j,i}, X_i) = \frac{1}{1 + e^{-\beta_{i,0} - \underline{Y}_{j,i}\underline{\beta}}}. \tag{3.8}$$

asumiendo que los $\beta_{i,0}$ son aleatorios de la forma $\beta_{i,0} = \beta_0 + \alpha_i + X_i$, donde β_0 es el intercepto poblacional. Se asume que $D_{j,i}$ son independientes dadas las $\underline{Y}_{j,i}$. Los vectores fila $\underline{Y}_{i,\cdot}$ volverían a considerarse no aleatorios debido a que las $\beta_{i,0}$ aportan la aleatoriedad del modelo. El modelo es Logístico con efectos fijos, con interceptos aleatorios.

Como señala Demidenko (2013), no es posible incorporar en el modelo logístico otros factores de efectos aleatorios diferentes del intercepto porque toda la variabilidad de aquellos queda capturada en éste, y por tanto, si se incluyeran, el modelo no sería identificable. Por tanto, se excluyen modelos con factores aleatorios multivariados para el caso logístico.

3.4. Modelo para estimar la probabilidad de castigo: Modelos de aprendizaje de máquina

En el presente trabajo se utilizarán, además de la regresión logística, modelos de aprendizaje de máquina o machine learning que es la ciencia que basa sus estudios en inteligencia artificial y sistemas de aprendizaje automático tomando como datos de entrada bases de datos de gran dimensión. Algunos modelos de aprendizaje de máquina son bosques aleatorios, árboles aleatorios, k-nn o vecinos más cercanos y máquinas de soporte vectorial. En la actualidad es común hacer inferencia y modelación estadística con bases de datos de grandes dimensiones por lo que se han venido creando herramientas y métodos de aprendizaje de máquina que tienen como objetivo desarrollar técnicas y algoritmos para que las máquinas aprendan; éstos métodos trabajan cada vez más en dar la mejor respuesta posible a las siguientes necesidades:

a) Procesamiento de máquina: Consumo de recursos de máquina y tiempo que se demora en realizar las estimaciones el modelo.

b) Interpretabilidad de los resultados: Facilidad para explicar y llevar a la práctica los resultados obtenidos.

c) Precisión en las estimaciones de los modelos: Certeza con la que el modelo estima las probabilidades.

3.4.1. Modelos de bosques aleatorios

Los métodos basados en árboles dividen las observaciones de la muestra analizada en rectángulos y luego, después de correr un modelo simple en cada árbol, arroja el mejor resultado encontrado en cada bosque. Los modelos con base a árboles han servido para cubrir la limitación de los modelos aditivos que proporcionan una útil extensión de los modelos lineales, haciéndolos más flexible al tiempo que conserva gran parte de su interpretabilidad. Sin embargo, los modelos aditivos pueden tener limitaciones para el uso prolongado de algoritmos de ajuste. Se ajusta a todos los predictores, lo que no es factible o deseable cuando un gran número está disponible lo que sí se puede hacer con modelos de árboles; lo modelos no aditivos no tienen la característica de aprendizaje por medio de algoritmos que sí tienen los modelos basados en árboles.

Finalmente, se divide la muestra en n número de árboles y el resultado es la clase con mayor número de votos en todo el bosque. Los modelos de bosques

aleatorios utilizan el algoritmo de bosque aleatorio de Breiman; además se utiliza como técnica de reducción de la dimensionalidad.

Bosques aleatorios o Random Forests Breiman (2001) hacen parte de un grupo de métodos conocidos como métodos de ensamble, los cuales según Zhou (2012) "son atractivos principalmente porque son capaces de impulsar métodos débiles y convertirlos en métodos fuertes, con los cuales pueden hacerse predicciones muy precisas".

El desempeño de bosques aleatorios puede ser superior al de los métodos paramétricos clásicos cuando la forma de los modelos es no lineal. Además permiten identificar variables predictoras informativas por medio de medidas de importancia para las variables explicativas, lo que hace atractivo el método.

Ejemplo de bosques aleatorios en R.

```
# Función del modelo control <- trainControl(method = "none")

# Modelo random forest
rf.model <- train (Castigos~.,
                   data=datos.aprendizaje,
                   method = "rf",
                   ntree = 150,
                   trControl = control)

# Probabilidades de castigo
prediction <- predict(rf.model,datos.testing[,-188]),
type="prob")

# Obtenemos un resumen de desempeño con la matriz de confusión
confusionMatrix(prediction,datos.testing\$Castigos)

# Performance del modelo
pred_1 <- prediction(as.numeric(prediction),
as.numeric(datos.testing[,ncol(datos.testing)]))
pred_1
perf_1 <- performance(pred_1, "tpr", "fpr")
plot(perf_1, main="ROCR de bosques aletarorios", type="l",
lty=4,
pch=5, col=4)
abline(h=1,v=1)
```

```
perf_1

# auc
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc <- performance(pred_1, "auc")
```

3.4.2. Modelos de árboles aleatorios

Los modelos de árboles aleatorios siguen el mismo método de bosques aleatorios con la diferencia que el de árboles no promedia la decisión de diferentes árboles si no que toma la muestra completa como su único árbol y con base a éste toma la decisión final. En modelos de árboles aleatorios se identifican las variables más significativas por medio de metodologías que generan medidas de importancia para las variables como se verá en la siguiente sección.

Dos de las ventajas de los modelos con árboles aleatorios son:

1. Manejo implícito o recursivo de datos faltantes.
2. Procesamiento de variables cualitativas y cuantitativas sin la necesidad de crear variables indicadoras.

Ejemplo de árboles aleatorios en R.

```
# Modelo
tree.model <- train(Castigos~.,
                    data = datos.aprendizaje,
                    method = "rpart",
                    trControl = control)

# Probabilidades de castigo
predictiontree <- predict(tree.model,datos.testing[,-188]),
type = "prob")

# Obtenemos un resumen de desempeño con la matriz de confusión
confusionMatrix(predictiontree,datos.testing$Castigos)

# Performance del modelo
pred_2 <- prediction(as.numeric(predictiontree),
as.numeric(datos.testing[,ncol(datos.testing)]))
hist(pred_2)
```

```
pred_2
perf_2 <- performance(pred_2, "tpr", "fpr")
plot(perf_2)
abline(h=1,v=1)

# auc
auc <- performance(pred_2, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc
```

3.4.3. Medida de importancia Gini

En este trabajo se medirá la importancia o significancia de las variables explicativas con base a la medida de importancia de Gini que cuantifica la reducción que genera una variable predictora de interés en la impureza de la variable respuesta. La cuantificación es sobre cada uno de los árboles del bosque y se obtiene el promedio o la suma de todas las medidas de importancia que se generan para la variable explicativa que finalmente da una medida global de la importancia de Gini.

Esta medida de importancia tiene como no positivo el hecho que la generación de particiones de los árboles es sesgado cuando alguna de las variables predictoras es categórica y/o tiene valores faltantes Strobl, Boulesteix, Zeileis, y Hothorn (2007), por lo tanto las medidas de importancia calculadas sobre árboles sesgados también sean sesgadas. Para lo anterior Sandri y Zuccolotto (2008) proponen un algoritmo de corrección de sesgo para la medida de importancia de las variables en la explicación de la variable dependiente en modelos de árboles de clasificación, algoritmo que tiene como objetivo principal la reducción de la heterogeneidad total producida por una covariable dada en la variable de respuesta cuando el espacio muestral se divide en entrenamiento y validación.

3.4.4. Modelo de máquinas de soporte vectorial

Las máquinas de soporte vectorial (Support Vector Machines) son un conjunto de algoritmos de aprendizaje supervisado para solución de problemas de clasificación y regresión que con base a un conjunto de muestras de datos de entrenamiento se predice la probabilidad de que ocurra un evento.

Como se explica en Hastie, Tibshirani, y Friedman (2009), el objetivo del algoritmo SVR es hallar, vía optimización, una función $f(x)$ cuya distancia a los valores de respuesta observados, Y_i, sea lo más plana posible y que como máximo esta distancia sea el error, $\epsilon > 0$. Por ejemplo, dados unos indicadores económicos, X_i, y marcaciones de créditos con proceso de castigo ya observados, Y_i, este método encontrará una función lo más plana posible, que prediga la probabilidad de castigo y la pérdida de dinero sea como máximo ϵ con respecto a la probabilidad real de default.

Ejemplo modelo de máquina de soporte vectorial en R.

```
# Función del modelo
svmR.model <- train(Castigos~.,
                    data = datos.aprendizaje,
                    method = "svmRadial",
                    trControl = control)

# Probabilidades de castigo
predictionsvmR <- predict(svmR.model,datos.testing[,-188])

# Obtenemos un resumen de desempeño con la matriz de confusión
confusionMatrix(predictionsvmR,datos.testing$Castigos)

# Performance del modelo
pred_4 <- prediction(as.numeric(predictionsvmR),
as.numeric(datos.testing[,ncol(datos.testing)]))
pred_4
perf_4 <- performance(pred_4, "tpr", "fpr")
plot(perf_4)
abline(h=1,v=1)
perf_4

# auc
auc <- performance(pred_4, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc
```

3.4.5. Modelo k-nn o vecino más cercano

El modelo k-nn o regla del vecino más cercano es un método de aproximación simple no paramétrica que consiste en estimar el valor de un dato desconocido a partir de las características del dato más próximo con base a distancias estadísticas. Según Friedman, Hastie, y Tibshirani (2001), entre más valores vecinos se consideren en el modelo, mayor probabilidad hay de ruido o de independencia entre las variables explicativas y así se suaviza la curva de estimación.

Ejemplo k-nn en R.

```
# Función del modelo
knn.model <- train(Castigos~.,
                   data = datos.aprendizaje,
                   method = "kknn",
                   trControl = control,
                   metric = "Sensitivity",
                   verbose = FALSE)

# Probabilidades de castigo
predictionknn <- predict(knn.model,datos.testing[,-188]),
type="prob")

# Obtenemos un resumen de desempeño con la matriz de confusión
confusionMatrix(predictionknn,datos.testing\$Castigos)

# Performance del modelo
pred_3 <- prediction(as.numeric(predictionknn),
as.numeric(datos.testing[,ncol(datos.testing)]))
pred_3
perf_3 <- performance(pred_3, "tpr", "fpr")
plot(perf_3)
abline(h=1,v=1)

# auc
auc <- performance(pred\_3, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc
```

3.5. Las Provisiones VaR, TVaR y ES

Las provisiones son cantidades que se calculan a partir de la distribución de la variable L, asumiendo el modelo de mezclas Bernoulli. Entre varias propuestas de provisiones las más utilizadas son VaR, se revisó Lien, Stroud, y Ye (2016) que presenta comparaciones entre cuatro métodos de cálculo de VaR y también en Crouhy et al. (2000) se revisan las actuales metodologías de Valor en Riesgo de crédito; TVaR revisándose Bakar, Hamzah, Maghsoudi, y Nadarajah (2015) que explica el planteamiento y desarrollo de modelos compuestos para datos de pérdidas de seguros de cola gruesa y ES. La definición de estas alternativas de cálculo de provisión se hace a continuación.

Definición 3.5.1. *Para una variable aleatoria X con fda $F_X(x)$, dado $u \in (0,1)$, el Valor en Riesgo de X al nivel u es una medida de riesgo que se define como*

$$VaR_X(u) = F_X^{-1}(u). \tag{3.9}$$

Definición 3.5.2. *Se define la medida de riesgo "valor en riesgo en la cola " TVaR (Tail Value at Risk) ó CTE (Conditional Tail Espectation)*

$$TVaR_X(1-q) = \mathbb{E}(X|X > VaR_X(1-q)),\ 0 < q < 1. \tag{3.10}$$

Definición 3.5.3. *Se define la medida de riesgo ES (expected shortfall) para un riesgo X, a un nivel de probabilidad $1-q$ como*

$$ES_X(1-q) = \mathbb{E}[(X - VaR_X(1-q))_+]. \tag{3.11}$$

Donde $x_+ = max(x,0)$ es la parte positiva de x.

Ver los anexos I y II para complementar información sobre aproximaciones de provisiones VaR, TVaR y ES con el modelo de Mezclas Bernoulli.

Capítulo 4

Descripción de los Datos

Los datos son simulados del desempeño de una compañía de crédito y son con corte de tiempo a mayo de 2018. Las variables a analizar son continuas y describen el comportamiento financiero histórico y actual de los créditos, por medio de variables autorregresivas y variables con valores a mayo de 2018, respectivamente.

Las dos bases de datos insumos para trabajar los modelos se llaman "entrenamiento" y "prueba". Los datos disponibles en la base "entrenamiento" son 89.584 observaciones individuales identificadas para cada cliente mediante un consecutivo que no tiene relación con datos de identificación de los clientes y la variable dependiente "castigos" que toma valor 1 si fue castigado y 0 si no fue castigado. La base de datos "entrenamiento" se divide en 70 % (62.708) como datos de aprendizaje y 30 % (26.876) como datos para validación. En cuanto al número de variables, se tienen 187 con potencial poder explicativo sobre la probabilidad de castigo de un crédito.

Finalmente se espera tener el resultado de probabilidad de castigo en la base "prueba" que tiene 6.298 observaciones por evaluar y el mismo número de variables que la de base de "entrenamiento", de ésta se extraerán los créditos que hayan sido desembolsados en los años 2014, 2015, 2016, 2017 y 2018.

Antes de iniciar cualquier análisis, se considera necesario revisar la estructura, tipos y potenciales sesgos de las variables disponibles que son 187. Entonces, las tranformaciones realizadas sobre la base de datos orginal fueron:

- Estandarización de variables: Las variables numéricas se vuelven comparables por medio de la normalización debido a que en la base original tienen diferente escala; las variables dummy o binarias se conservan.
- Transformación de variables: La variable de tipo texto se transforma en

dummy lo que aporta al modelo tantas variables como número de categorías menos una tiene la variable.

- Eliminación de variables: Se eliminan variables que posiblemente generan sesgo en la estimación de la probabilidad de default debido a que su quantil 75 o mejor, el 75 % de los datos son cero.
- No eliminar outliers: Si bien esto impacta negativamente el resultado actual del AUC de los modelos, es un primer resultado, con el paso del tiempo se recoge información y el modelo aprenderá a capturar las características de los outliers o datos extremos e ir aumentando la precisión en la predicción de los castigos.
- Análisis de componentes principales: Se hace para las variables transaccionales que muestran alto grado de correlación entre ellas pasando de un número de 9 variables explicativas a 2 grupos de componentes que capturan significativamente la varianza en las demás variables, por ejemplo, variables de comportamiento del cliente, el resultado de componentes principales no fue significativo y entran uno a uno porque su correlación no es significativa entre ellas lo que no muestra multicolinealidad.

$Nota$: El código de R para lograr lo descrito anteriormente se encuentra en el Apendice B.

Como se ve en la siguiente gráfica, a la base de entrenamiento se le hacen pruebas de calidad para garantizar que la estructura y calidad de los valores de las observaciones cumplen las particularidades que exige cada modelo que en este caso son variables continuas y no valores nulos o vacíos.

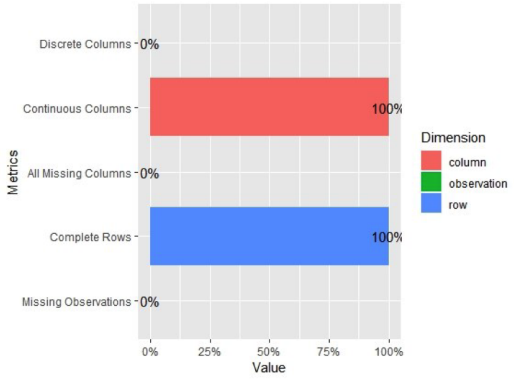

Figura 4.1: En la gráfica se observa que todas las variables son continuas y no poseen datos faltantes o datos missing.

Con base a lo observado en la gráfica anterior, no es necesario realizar imputación de datos, tampoco se eliminarán outliers que aunque en los actuales modelos se sacrifica un poco el resultado de AUC, se entiende que son modelos base o primeros modelos para dar solución al problema actual y que con el paso del tiempo los modelos aprenderán y capturarán en mayor medida los comportamientos de estos outliers y el resultado del AUC cada vez será más robusto y estimará probabilidades de default más aproxiamdas a la realidad.

Ahora, la base de "entrenamiento" contiene la variable "Castigos" y se distribuye de la siguiente manera:

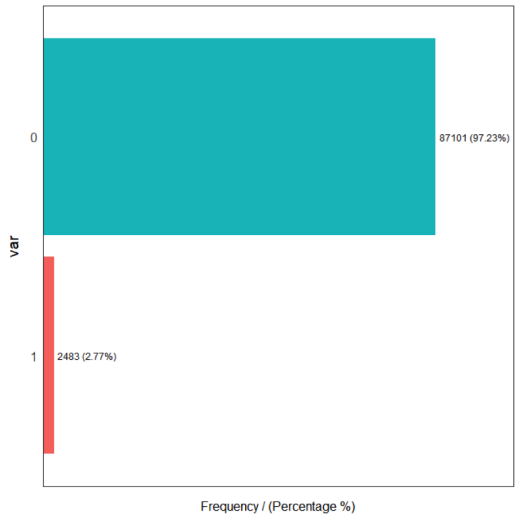

Figura 4.2: En la gráfica se observa el porcentaje de créditos de la base de la base de entrenamiento que han sido castigados y los que no

En la gráfica 2 se puede ver que el modelo se entrena con una base de datos que tiene el 97.23 % créditos que no han sido castigados respecto a 2.77 % que sí fueron castigados.

A continuación se hace análisis exploratorio de 10 variables explicativas entandarizadas de la base de entramiento potencialmetne significativas en la explicación de si un crédito se castiga o no.

4.1. Análisis univariado de frecuencias para variables independientes

El análisis univariado permite visualizar asimetrías[1] y curtosis[2] de los valores de las variables independientes comparado con su frecuencia.

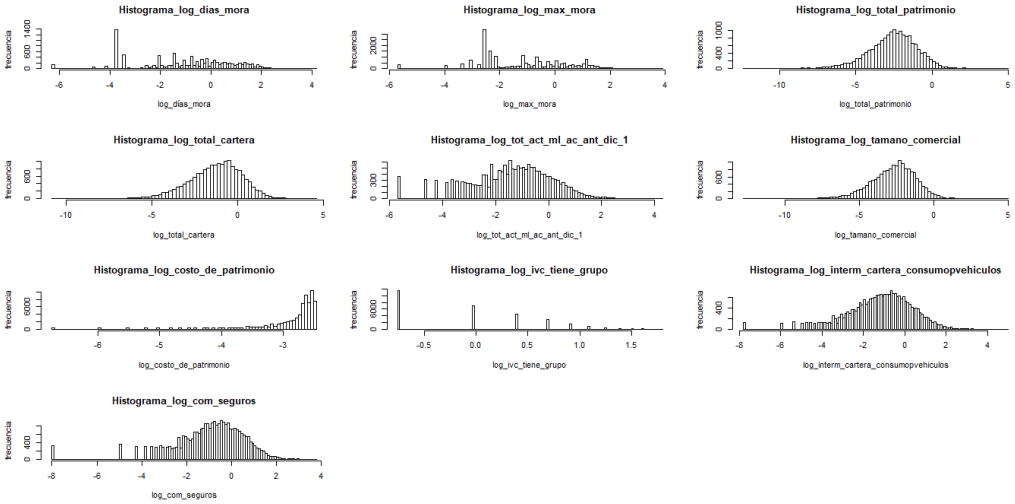

Figura 4.3: Distribución de las variables días_mora, max_mora, total_patrimonio, total_cartera, tot_act_ml_ac_ant_dic_1, tamano_comercial, costo_de_patrimonio, ivc_tiene_grupo, interm_cartera_consumopvehicuclos y com_seguros todas en escala logarítmica.

La distribución de la variable días_mora presenta asimetría izquierda con curtosis alta en el nivel logarítmico -3 lo que indica que la mayoría de créditos de la organización están con moras tolerables al 31 de mayo_2018. La distribución de la variable max_mora presenta un comportamiento similar a la distribución de dias_mora. La distribución de la variable total_patrimonio presenta asimetría izquierda con curtosis alta. La distribución de la variable total_cartera presenta asimetría izquierda con curtosis alta en su moda que es -1. La distribución de la variable tot_act_ml_ac_ant_dic_1 presenta asimetría izquierda con curtosis alta en el nivel logarítmico -3. La distribución de la variable tamano_comercial presenta asimetría izquierda y curtosis en -3. La distribución de la variable costo_de_patrimonio presenta alta asimetría izquierda con curtosis media respecto a su moda. La distribución de la ivc_tiene_grupo presenta asimetría derecha, también clasifica por niveles de ivc_tiene_grupo y se ve que la mayor concentración en ivc_tiene_grupo está en los primeros valores positivos de este índice. La

[1] Tiene relación con el ancho de la gráfica, con frecuencias y colas de la distribución.
[2] Tiene relación con la altura o moda de la distribución

distribución de la variable interm_cartera_consumopvehicuclos presenta asimetría izquierda con curtosis alta en el nivel logarítmico -3 lo que significa que la variable tiene una concentración en los que la interm_cartera_consumopvehicuclos es baja. La distribución de la variable com_seguros presenta asimtería izquierda y curtosis en -1 lo que quiere decir que hay un considerable número de créditos concentrados en 0 o bajo número de seguros para los crédtos adquiridos.

4.2. Análisis bivariado. Correlación y distribución de castigos respecto a variables independientes

El análisis bivariado permite visualizar la concentración o distribución de la variable dependiente, en este caso "castigos de créditos", respecto a las variables independientes estando éstas estandarizadas.

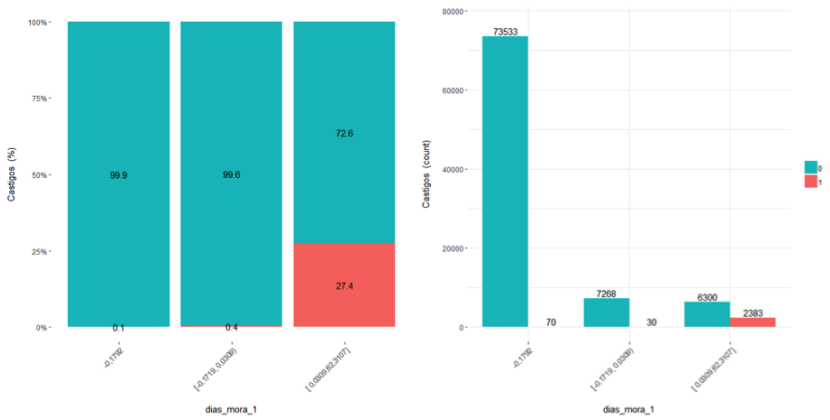

Figura 4.4: Concentración o distribución de "castigos de créditos" en la variable días_mora.

La gráfica muestra la relación entre castigos y días_mora observándose una gran concentración de castigos en el último rango de días de mora lo que da una señal, sin hablar aún de significancia en la explicación de la variable respuesta, que a mayor número de días de mora, es más probable que hayan castigos. También se observa que el 93 % de los clientes tienen menos de 120 días de mora lo que indica implícitamente que se tiene un bajo o controlado ICV.

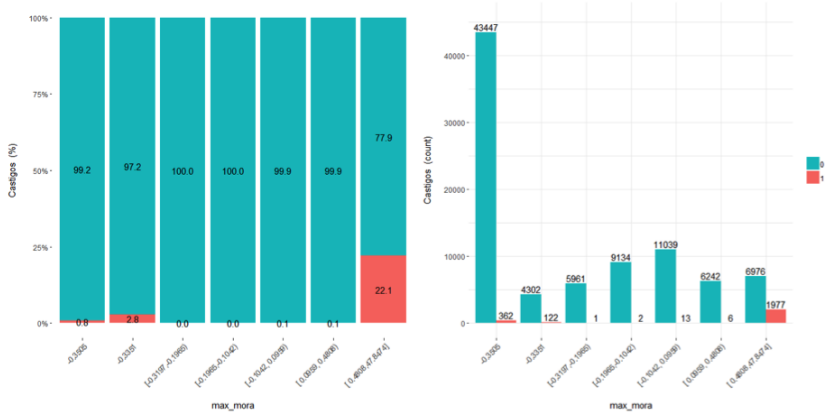

Figura 4.5: Concentración o distribución de "castigos de créditos" en la variable max_mora.

La gráfica muestra la relación entre castigos y max_mora observándose una gran concentración de castigos en el último rango max_mora entendiéndose este último rango como el grupo de clientes que han alcanzado una altura mayor máxima considerable en el periodo de tiempo estudiado lo que da una señal, sin hablar aún de significancia en la explicación de la variable respuesta, que entre más alta haya sido la max_mora, es más probable que hayan castigos.

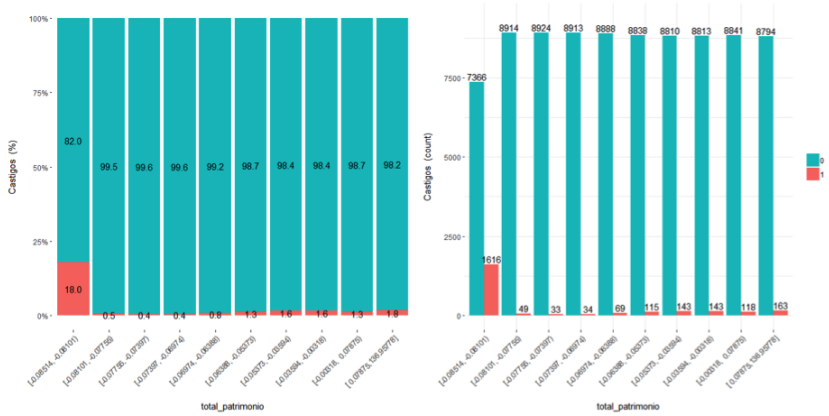

Figura 4.6: Concentración o distribución de "castigos de créditos" en la variable total_patrimonio.

La gráfica muestra la relación entre castigos y total_patrimonio observándose que el 18 % del primer rango de total_patrimonio ha sido castigado en el periodo de tiempo estudiado lo que da una señal, sin hablar aún de significancia en la explicación de la variable respuesta, a menor patrimonio, mayor concentración de castigos.

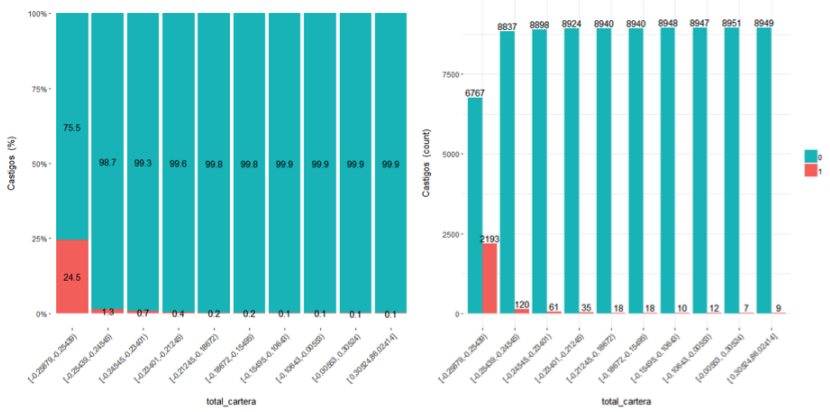

Figura 4.7: Concentración o distribución de "castigos de créditos" en la variable total_cartera.

La gráfica muestra la relación entre castigos y total_cartera observándose que el 24.5 % del primer rango de total_cartera ha sido castigado en el periodo de tiempo estudiado lo que da una señal, sin hablar aún de significancia en la explicación de la variable respuesta, a menos dinero en total_cartera, mayor concentración de castigos.

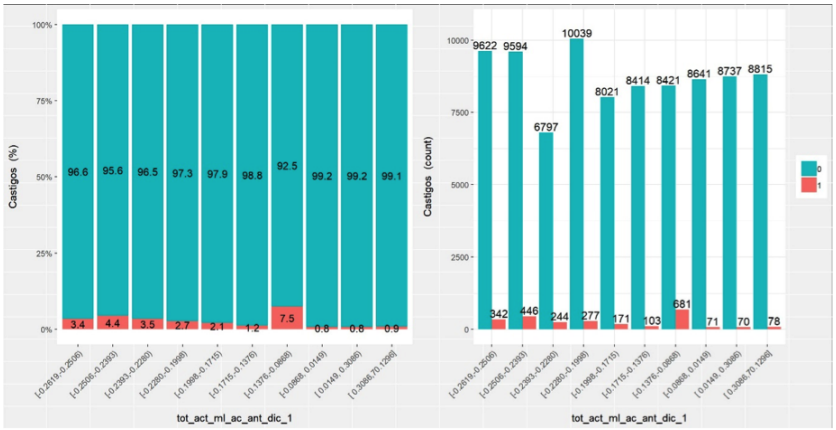

Figura 4.8: Concentración o distribución de "castigos de créditos" en la variable tot_act_ml_ac_ant_dic_1.

La gráfica muestra la relación entre castigos y tot_act_ml_ac_ant_dic_1 observándose que un máximo créditos castigados en el rango 7 de tot_act_ml_ac_ant_dic_1 en el periodo de tiempo estudiado lo que da una señal, sin hablar aún de significancia en la explicación de la variable respuesta, a menor valor de tot_act_ml_ac_ant_dic_1, mayor concentración de castigos.

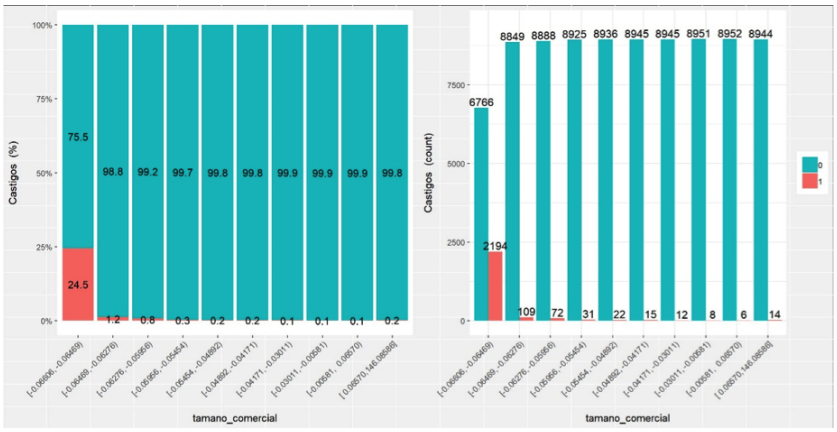

Figura 4.9: Concentración o distribución de "castigos de créditos" en la variable tamano_comercial.

La gráfica muestra la relación entre castigos y tamano_comercial observándose que el 24.5 % del primer rango de tamano_comercial ha sido castigado en el periodo de tiempo estudiado lo que da una señal, sin hablar aún de significancia en la explicación de la variable respuesta, a menor tamano_comercial, mayor concentración de castigos.

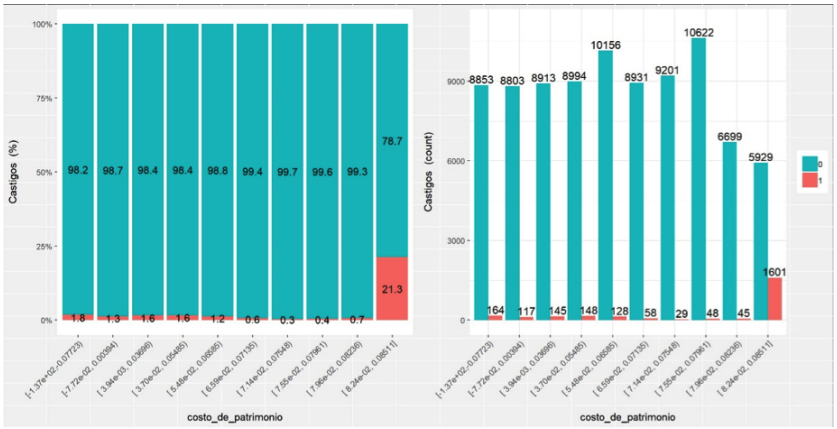

Figura 4.10: Concentración o distribución de "castigos de créditos" en la variable costo_de_patrimonio.

La gráfica muestra la relación entre castigos y costo_de_patrimonio observándose que el 21.3 %, o sea, 1601 créditos, del último rango de costo_de_patrimonio ha sido castigado en el periodo de tiempo estudiado lo que da una señal, sin hablar aún de significancia en la explicación de la variable respuesta, a mayor costo_de_patrimonio, mayor concentración de castigos.

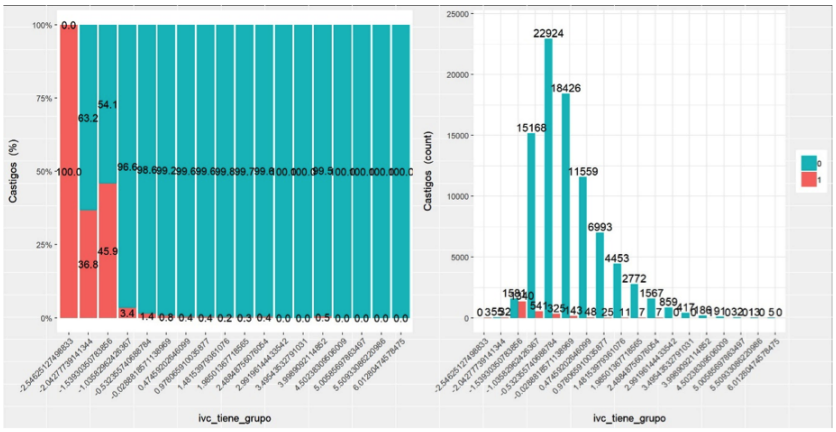

Figura 4.11: Concentración o distribución de "castigos de créditos" en la variable ivc_tiene_grupo.

La gráfica muestra que a menor ivc_tiene_grupo, mayor concentración de castigos. A pesar de que no es lógico desde un contexto de negocio el sentido que toma una alta concentración de castigos en bajos niveles de ivc_tiene_grupo, el argumento que se da para que esta variable permanezca es que los créditos castigados no tienen o presentan niveles muy bajos de ivc_tiene_grupo porque éstos ya entraron en la provisión contable, o sea, en el estado de pérdidas y gastos ya se declaró dinero perdido. De hecho, se tienen metas o promedios de castigos mensuales para sanear el índice de cartera vencida de la empresa.

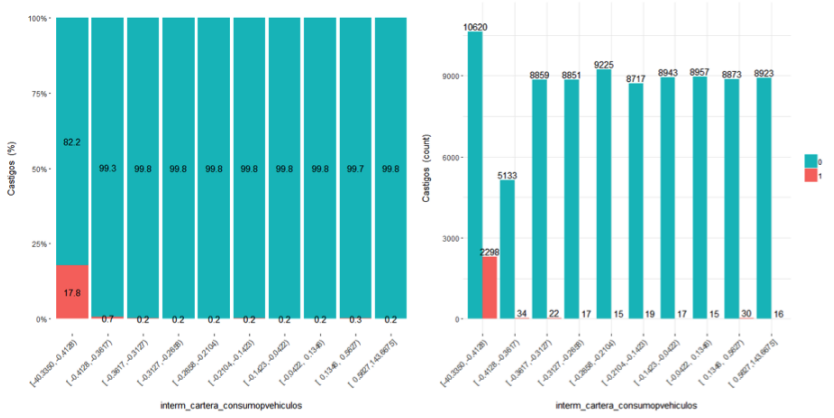

Figura 4.12: Concentración o distribución de "castigos de créditos" en la variable interm_cartera_consumopvehicuclos.

La gráfica muestra la relación entre castigos y interm_cartera_consumopvehicuclos observándose que el 17.8 %, o sea, 2298 créditos, del primer rango de costo_de_patrimonio ha sido castigado en el periodo de tiempo estudiado lo que da

una señal, sin hablar aún de significancia en la explicación de la variable respuesta, a menor interm_cartera_consumopvehiculos, mayor concentración de castigos.

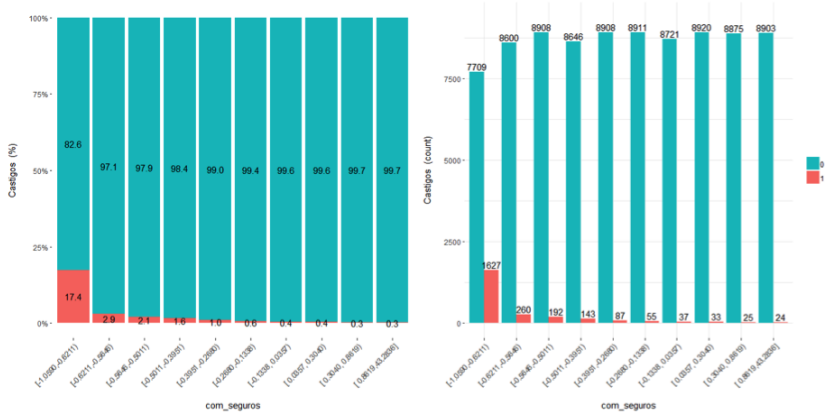

Figura 4.13: Concentración o distribución de "castigos de créditos" en la variable com_seguros.

La gráfica muestra la relación entre castigos y com_seguros observándose que el 17.4 %, o sea, 1627 créditos, del primer rango de com_seguros ha sido castigado en el periodo de tiempo estudiado lo que da una señal, sin hablar aún de significancia en la explicación de la variable respuesta, a menor com_seguros, mayor concentración de castigos.

Capítulo 5

Resultados de las Estimaciones

5.1. Competencia de modelos, resultados y selección del mejor modelo para la estimación de la probabilidad de castigo de los créditos

Se genera una competencia entre los modelos de aprendizaje de máquina[1] como lo son bosques aleatorios, árboles aleatorios, knn o vecinos más próximos, máquinas de soporte vectorial y el modelo de regresión logística. El modelo que genera el mayor AUC sobre la base de datos "entrenamiento", mayor precisión y el más aceptable rango de precisión con IC del 95 % será el seleccionado para estimar las probabilidades finales sobre la base "prueba".

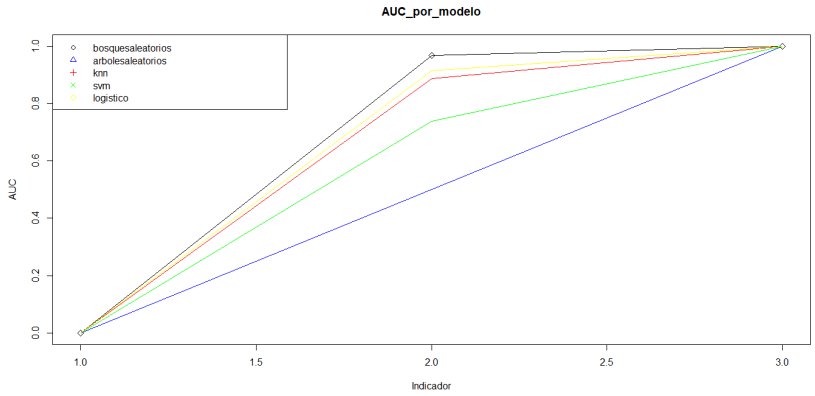

Figura 5.1: Curvas AUC para los modelos bosques aleatorios, árboles aleatorios, knn, svm y logístico.

[1] En Demidenko (2013) se precisa el impacto de la tecnología en la estadística y explica teoría de la modelos mixtos y métodos de diagnóstico de modelos y análisis influyentes.

El modelo de bosques aleatorios muestra una muy buena curva AUC. En la gráfica se ve que es el modelo que mayor área cubre sobre la curva, su capacidad discriminante es mayor a la de los demás modelos lo que da señal de que es el que mejor pronóstico de probabilidades de default estima.

El modelo de árboles aleatorios muestra una mala curva AUC respecto a bosques aleatorios y aporta poco en la estimación de probabilidades de ser o no castigado un crédito. En la gráfica se ve que el modelo es el que menor área bajo la curva captura o sea que su capacidad discriminante es baja lo que da señal de que es el modelo que menor confianza genera para estimar probabilidades de default.

Los modelos knn y svm muestran una aceptable curva AUC. En la gráfica se ve que los modelos capturan un área bajo la curva aceptable o sea que su capacidad discriminante es normal y podrían ser una alternativa para estimar probabilidades de default.

El modelo logístico muestra una muy buena curva AUC ; compite con el modelo de bosques aleatorios. En la gráfica se ve que es el segundo modelo que mayor área cubre sobre la curva y su capacidad discriminante es buena, lo que da señal que es el segundo modelo que genera mayor confianza para la estimación de probabilidades de default.

Ahora, después de revisar el AUC o área bajo la curva que captura cada modelo, la siguiente tabla complementa el análisis con precisión e IC_precisión que son la 3 medidas que se definieron para seleccionar el mejor modelo. Entonces:

Precisión, IC_precisión y AUC

Modelo	Precisión	IC precisión	AUC
Bosques aleatorios	0.996	(0.9952, 0.9967)	0.9536
Árboles de decisión	0.9723	(0.9703, 0.9742)	0.500
KNN	0.9912	(0.99, 0.9923)	0.8884
SVM	0.9916	(0.9904, 0.9926)	0.8683
Logit	0.9929	(0.9918, 0.9938)	0.9141

Tabla 5.1: La tabla muestra el porcentaje de precisión, el intervalo de confianza en el que se encuentra esa precisión y el AUC que son las métricas con las que se evalúan los modelos en competencia.

En la tabla se observa que el modelo de bosques aleatorios es el que arroja mejores resultados en las medidas de desempeño precisión y AUC. A pesar que los demás modelos poseen un rango similar para el intervalo de confianza, el rango

con valores superiores para la precisión es el de bosques aleatorios.

El mejor modelo vía mejor desempeño en AUC, precisión y rango IC_precisiónbosques es el de bosques aleatorios. Con bosques aleatorios se estima sobre base "prueba" la probabilidad de que un crédito sea o no castigado.

Después de correr el modelo de bosques aleatorios y estimar las probabilidades de default, se encuentra por medio de Gini, explicado en la sección 4.4.3, que las variables independientes que mayor importancia tinene en la explicación de los castigos son:

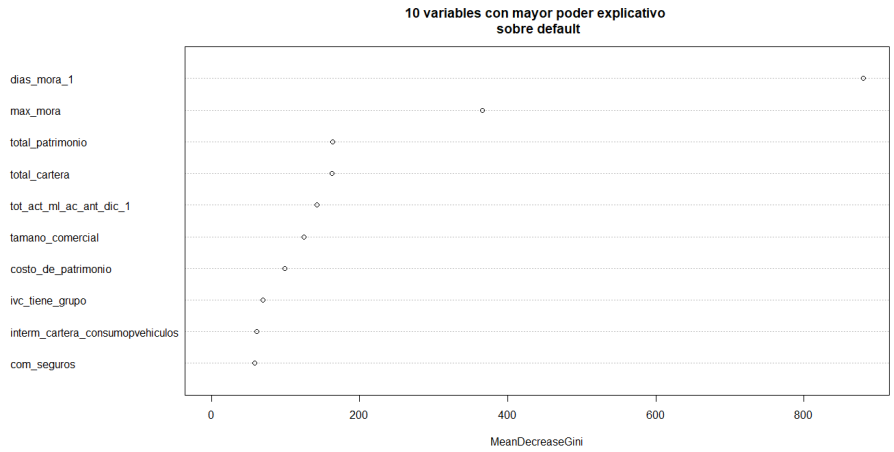

Figura 5.2: Las 10 variables más signigficativas en la explicación en la probabilidad de castigo de un crédito.

En la gráfica se observa que las variables más significativas son precisamente las analizadas y explicadas en la sección anterior; éstas variables son: días_mora, max_mora, total_patrimonio, total_cartera, tot_act_ml_ac_ant_dic1, tamano_comercial, costo_de_patrimonio, ivc_tiene_grupo, interm_cartera_consumopvehiculos, com_seguros.

Finalmente, después de analizar los resultados de cada modelo se observan las distribuciones de sus funciones de probabilidad:

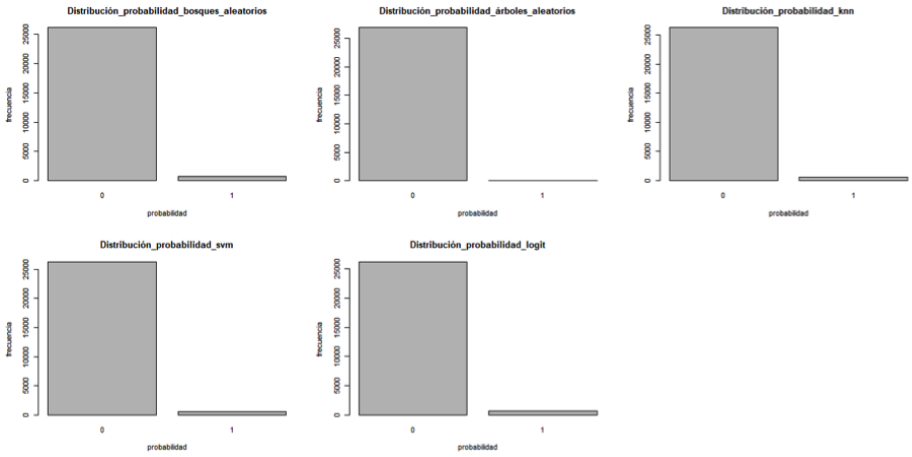

Figura 5.3: Gráficas de distribución de probabilidad para los modelos bosques aleatorios, árboles aleatorios, knn, svm y logístico.

En la gráfica se ve que la distribución de probabilidades del modelo bosques aleatorios tiene un alto porcentaje en 0, osea que son muchos más los créditos que son probables que no sean castigados o que paguen hasta finalizar la deuda respecto a la proporción de créditos que sí serán castigados. Ahora, se ve que la distribución de probabilidades del modelo árboles aleatorios tiene un 100 % con 0, osea que el modelo de árboles aleatorios estima que ningún crédito será castigado; este modelo, como ya se mencionó, tiene poca capacidad discriminante y el AUC en la competencia de modelos de este trabajo lo corrobora.

La distribución de probabilidades de los modelos knn y svm tienen un alto porcentaje en 0, osea que son muchos más los créditos con probabilidad de que no sean castigados o que paguen hasta finalizar la deuda. Por último, se tiene que la distribución de probabilidades del modelo logístico muesta un alto porcentaje en 0, osea que son muchos más los créditos que son probables que no sean castigados o que paguen hasta finalizar la deuda.

$Nota$: El código de R para lograr lo descrito anteriormente se encuentra en el Apendice B.

5.2. Cálculo de la distribución de costos para los años 2014, 2015, 2016, 2017 y 2018.

Se realizará la estimación del capital expuesto con intervalos de confianza del 90 %, 95 % y 99 % para créditos con años de originación de 2014, 2015, 2016,

2017 y 2018 revisandose:

a) Var para el métodos NP: En la gráficas se ve con puntos rojos

b) TVar para el método NP: En la gráficas se ve con puntos azules

c) Función de distribución de costos

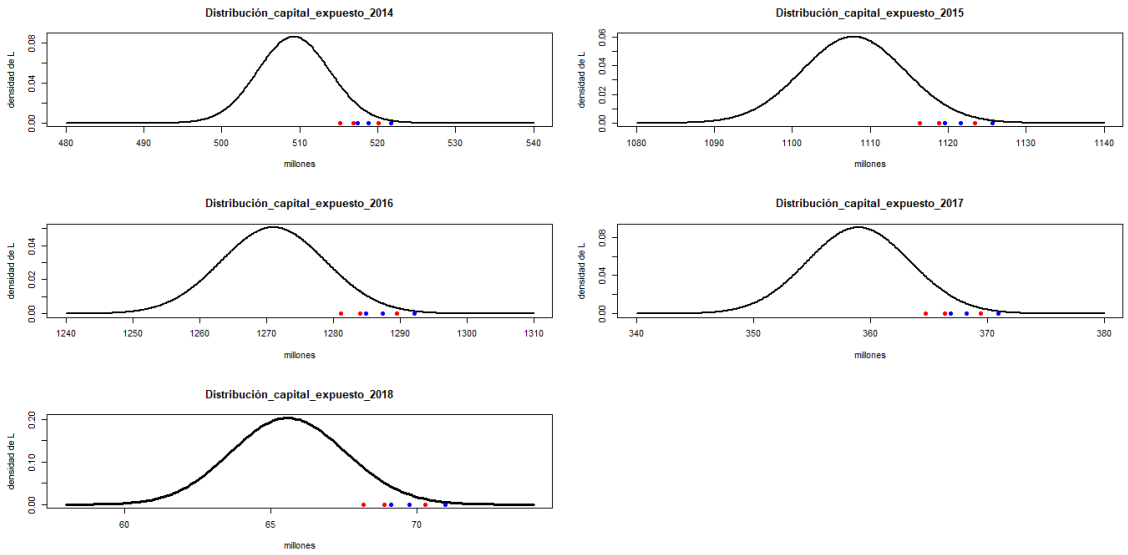

Figura 5.4: Distribución de la función L para créditos originados en los años 2014, 2015, 2016, 2017 y 2018.

En la gráfica se observa que para los créditos originados en los años 2014 y 2015, con una confianza del 99 %, el capital expuesto sería menor o igual $520'000.000$ y $1.123'000.000$;respectivamente. Estos serían los valores de la provisón contable que garantizaría cubrir todas las posibles pérdidas; en Dutang, Goulet, Pigeon, et al. (2008) se presentan las facilidades de trabajar con la versión actual del paquete "actuar" de R porque contiene funciones de modelación de distribuciones de pérdidas y teoría de riesgo de crédito. Para los crédtios de estos dos años de originación, se observa que hay una gran concentración de probabilidades altas de castigos en saldos de monto alto lo que se traduce en alto riesgo de pérdida sobre el capital expuesto.

También se puede ver que para el año 2016, con una confianza del 99 %, el capital expuesto sería menor o igual $1.289'000.000$ y este sería el valor de la provisón

contable que garantizaría cubrir todas las posibles pérdidas. El 2016 muestra que fue el año de desembolsos en el que menos ajustado se tenía el perfil de riesgos para hacer efectivo o no el préstamo. Para los créditos originado en el año 2017 se observa, con una confianza del 99 %, que el capital expuesto sería menor o igual $369'000.000$ y este sería el valor de la provisón contable que garantizaría cubrir todas las posibles pérdidas. Comparado con el capital expuesto por los créditos orginados en 2016, el capital expuesto para créditos originados en 2017 es mucho menor.

Finalmente, se puede ver que para los primeros 5 meses del 2018, con una confianza del 99 %, el capital expuesto sería menor o igual $70'000.000$ y este sería el valor de la provisón contable que garantizaría cubrir todas las posibles pérdidas. A pesar de que a mayo de 2018 ya se había castigado un número de créditos similar al de 2017 en diciembre, el capital expuesto en 2018 es mucho menor. Las probabilidades altas de castigos están distribuidas por lo diferentes niveles de saldos sin mostrar un concentración sobre algún rando de saldos lo que permite concluir que en este año tenemos un menor capital expuesto y menos riesgoso.

$Nota$: El código de R para lograr lo descrito anteriormente se encuentra en el Apendice B.

Capítulo 6

Conclusiones y sugerencias próximos trabajos

6.1. Conclusiones

- Los créditos desembolsados los años 2015 y 2016 son los que mayor capital expuesto generan. Hay una gran concentración de probabilidades altas de castigos en saldos de monto alto.
- El perfil de apetito de riesgo se viene ajustando en los últimos años y muestra una cartera menos expuesta. Los desembolsos de años recientes muestran una menor concentración en saldos altos de la probabildad de castigo respecto a desembolsos de años anteriores.
- El mejor modelo para estimar probabilides de castigo fue el de bosques aleatorios con AUC del 95 %. Las variables que muestran mayor importancia vía Gini en la explicación de la variable respuesta son días_mora, max_mora, total_patrimonio, total_cartera, tot_act_ml_ac_ant_dic1, tamano_comercial, costo_de_patrimonio, ivc_tiene_grupo, interm_cartera_consumopvehiculos y com_seguros.

6.2. Sugerencias próximos trabajos

Para próximos trabajos de modelo de riesgos de crédito se sugiere incluir variables de gestión o negociaciones de conciliación de la empresa con el cliente en mora; también incluir serie de tiempo de pagos con variables como la fecha, valor del pago, hora del pago y canal por el que realizó el pago.

Referencias

Albrecher, H., Beirlant, J., y Teugels, J. L. (2017). *Reinsurance: actuarial and statistical aspects*. John Wiley & Sons.

Bakar, S. A., Hamzah, N., Maghsoudi, M., y Nadarajah, S. (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics*, *61*, 146–154.

Bluhm, C., Overbeck, L., y Wagner, C. (2016). *Introduction to credit risk modeling*. Chapman and Hall/CRC.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Castaner, A., Claramunt, M., y Marmol, M. (2013). Tail value at risk. an analysis with the normal-power approximation. *Statistical and Soft Computing Approaches in Insurance Problems; Nova Science Publishers: Hauppauge, NY, USA*, 87–112.

Crouhy, M., Galai, D., y Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking & Finance*, *24*(1-2), 59–117.

Demidenko, E. (2013). *Mixed models: theory and applications with r*. John Wiley & Sons.

Dutang, C., Goulet, V., Pigeon, M., y cols. (2008). actuar: An r package for actuarial science. *Journal of Statistical software*, *25*(7), 1–37.

Fisher, S. R. A., y Cornish, E. (1960). The percentile points of distributions having known cumulants. *Technometrics*, *2*(2), 209–225.

Frey, R., y McNeil, A. J. (2003). Dependent defaults in models of portfolio credit risk. *Journal of Risk*, *6*, 59–92.

Friedman, J., Hastie, T., y Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (n.º 10). Springer series in statistics New York.

Giraldo, N. (2014). Gestión cuantitativa de riesgo. aplicaciones con r. notas de clase (sin publicar). *Escuela de Estadística. Universidad Nacional de Colombia. Medellin*.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning* (2.ª ed.). Springer-Verlag New York. doi: 10.1007/978-0-387-84858-7

Huang, X., y Oosterlee, C. W. (2011). Saddlepoint approximations for ex-

pectations and an application to cdo pricing. *SIAM Journal on Financial Mathematics*, *2*(1), 692–714.

Kaas, R., Goovaerts, M., Dhaene, J., y Denuit, M. (2008). *Modern actuarial risk theory: using r* (Vol. 128). Springer Science & Business Media.

Lieberman, O. (1994). Saddlepoint approximation for the distribution of a ratio of quadratic forms in normal variables. *Journal of the American Statistical Association*, *89*(427), 924–928.

Lien, D., Stroud, C., y Ye, K. (2016). Comparing var approximation methods that use the first four moments as inputs. *Communications in Statistics-Simulation and Computation*, *45*(2), 491–503.

Maillard, D. (2018). A user's guide to the cornish fisher expansion. *Available at SSRN 1997178*.

McNeil, A. J., Frey, R., Embrechts, P., y cols. (2005). *Quantitative risk management: Concepts, techniques and tools* (Vol. 3). Princeton university press Princeton.

McNeil, A. J., y Wendin, J. (2003). *Generalized linear mixed models in portfolio credit risk modelling* (Inf. Téc.). ETH Zurich.

Merino, S., y Nyfeler, M. (2002). Credit portfolio modelling calculating portfolio loss. *RISK-LONDON-RISK MAGAZINE LIMITED-*, *15*(8), 82–86.

Pav, S. E. (2017). The sadists package.

Peters, G., Targino, R., y Shevchenko, P. V. (2013). Understanding operational risk capital approximations: first and second orders. *Available at SSRN 2980465*.

Sandri, M., y Zuccolotto, P. (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*(3), 611-628. Descargado de https://doi.org/10.1198/106186008X344522 doi: 10.1198/106186008X344522

Strobl, C., Boulesteix, A.-L., Zeileis, A., y Hothorn, T. (2007, 25 de Jan). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. Descargado de https://doi.org/10.1186/1471-2105-8-25 doi: 10.1186/1471-2105-8-25

Sundt, B. (1999). *An introduction to non-life insurance mathematics* (Vol. 28). VVW GmbH.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

Apéndice A

A.1. Aproximaciones para VaR, TVaR y ES

En Merino y Nyfeler (2002) se desarrolla una metodología para el cálculo de la distribución de la pérdida L y las provisiones VaR, TVaR y ES en el modelo de mezclas Bernoulli. Se basa en un aproximación de la distribución de L mediante una suma de distribuciones Poisson, usando simulación intensiva y aplicando la Ley Fuerte de Grandes Números. La metodología Merino-Nyfeler, que se utilizará en este trabajo, se presenta en Giraldo (2014). Es una modificación de la propuesta original en Merino y Nyfeler (2002), basada en la transformada rápida de Fourier, que consiste en calcular directamente los cumulantes a partir de la distribución Poisson y calcular las medidas de riesgo VaR, TVaR y ES.

A.2. Aproximaciones para VaR, ES y TVaR con el modelo de Mezclas Bernoulli

Los métodos empleados para aproximar la distribución de la pérdida L están en Kaas, Goovaerts, Dhaene, y Denuit (2008, sec. 2.5), ver también Peters, Targino, y Shevchenko (2013) y Bakar et al. (2015). Se basan en el supuesto de que los primeros cuatro momentos de L son finitos, y por tanto, los respectivos cumulantes también. En Peters et al. (2013) se encuentra una revisión del estado del arte de las aproximaciones que utilizan momentos, las aproximaciones consideradas en este trabajo son las siguientes.

A.2.1. VaR con la aproximación NP

El Método de Aproximación NP ó "Normal Power", permite una evaluación del percentil de F_S, S_q correspondiente a la probabilidad $(1-q)\,\%$.

$$VaR_S(q) = \mu_S + \sigma_S(z_q + \frac{\gamma_{1,s}}{6}(z_q^2 - 1)). \tag{A.1}$$

donde z_q es el q-percentil de una Normal estándar, con $\Phi(z_q) = q$, por ejemplo, $q = 0.99$.

A.2.2. TVaR con la aproximación NP

Para una variable $X \sim N(\mu, \sigma^2)$ se cumple la identidad:

$$\mathbb{E}(X|X > a) = \mu + \sigma \frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{A.2}$$

luego, con $a = VaR(u)$

$$\mathbb{E}(X|X > VaR(u)) = \mu + \sigma \frac{\phi(z_u)}{1-u} \tag{A.3}$$

En Castaner, Claramunt, y Marmol (2013) se modificó esta expresión para calcular TVaR a partir de la aproximación Normal Power

$$TVaR_X(u) = \mathbb{E}(X|X > VaR(u)) = \mu + \sigma \frac{\phi(z_u)}{1-u}\left(1 + \frac{\gamma_{1,X}}{6}\right) \tag{A.4}$$

A.2.3. VaR y TVaR con aproximación Cornish-Fisher

El desarrollo de Cornish-Fisher, orignal de Fisher y Cornish (1960), es una fórmula para aproximar los cuantiles de una distribución, utilizando los primeros cuatro cumulantes. Su definición en Lieberman (1994) es:

"The Cornish Fisher approximation is the Legendre inversion of the Edgeworth expansion of a distribution, but ordered in a way that is convenient when used on the mean of a number of independent draws of a random variable".

$$\frac{S - \mu_S}{\sigma_S} \approx Z + \frac{\gamma_{1,s}}{6}(Z^2 - 1) + \frac{1}{24}\gamma_{2,s}(Z^3 - 3Z) - \frac{1}{36}\gamma_{1,s}^2(2Z^3 - 5Z). \tag{A.5}$$

Permite también obtener una expresión del percentil S_q de F_S. Reemplazando S por S_q y z_p por Z en la aproximación CF (A.5), como se ve en Sundt (1999), se obtiene :

$$\frac{S_q - \mu_S}{\sigma_S} \approx z_q + \frac{\gamma_{1,s}}{6}(z_{2-1}) + \frac{1}{24}\gamma_{2,s}(z_q^3 - 3z_q) - \frac{1}{36}\gamma_{1,s}^2(2z_q^3 - 5z_q).$$

Y despejando S_q, se tiene

$$VaR_S(q) = \mu_S + \sigma_S(z_q + \nu). \tag{A.6}$$

donde

$$\nu = \frac{\gamma_{1,s}}{6}(z_q^2 - 1) + \frac{1}{24}\gamma_{2,s}(z_q^3 - 3z_q) - \frac{1}{36}\gamma_{1,s}^2(2z_q^3 - 5z_q), \tag{A.7}$$

es un factor de corrección de la aproximación Normal por asimetría y curtosis.

La implementación en R del VaR está el la librería de Pav (2017). La implementación de TVaR con base en Cornish-Fisher no está implementada en esta librería. Se implementará la versión que aparece en Maillard (2018).

Las Librerías en R que implementan esto desarrollos son:

- PDQutils de Pav (2017)
- RelIns de Albrecher, Beirlant, y Teugels (2017)

A.2.4. VaR, TVaR y ES con aproximación con una Gamma transladada

Utilizando una Gamma transladada $X \sim Gamma(\alpha, \beta, k)$, definida como una distribución Gamma con un rango $[k, \infty)$. Primero se estiman los parámetros (k, α, β) a partir de $\kappa_1 = \mathbb{E}(S), \kappa_2 = Var(S), \gamma_{1,s}$, con las ecuaciones:

$$\begin{aligned} \alpha &= 4/\gamma_{1,s}, \\ \beta &= \sqrt{\kappa_2/\alpha}, \\ k &= \kappa_1 - \alpha\beta. \end{aligned}$$

Luego se define

$$VaR_S(p) = k + F^{-1}_{G(\alpha,\beta)}(p).$$

La librería `VaRES` tiene programados el VaR y el TVaR para distribuciones $Gamma(\alpha, \beta)$. De la página de ayuda se extraen las fórmulas utilizadas, y se hacen algunos co-

mentarios.

$$f(x) = \frac{b^a x^{a-1} \exp(-bx)}{\Gamma(a)}, \tag{A.8a}$$

$$1 - F_{G(\alpha,\beta)}(x) = Q(\alpha, \beta x), \tag{A.8b}$$

$$VaR_p(X) = Q^{-1}(\alpha, \beta(1-p)), \tag{A.8c}$$

$$ES_p(X) = \frac{1}{p}\int_0^p Q^{-1}(\alpha, \beta(1-v))dv \tag{A.8d}$$

Donde $Q(\alpha, x)$ se define como la función gamma incompleta normalizada

$$Q(\alpha, x) = \int_x^\infty t^{\alpha-1}e^{-t}dt/\Gamma(\alpha) \tag{A.9}$$

luego $1 - F_{G(\alpha,\beta)}(x) = Q(\alpha, \beta x)$. Y el VaR Gamma es $F^{-1}_{G(\alpha,\beta)}(p)$, dado por

$$VaR_{G(\alpha,\beta)}(p) = F^{-1}_{G(\alpha,\beta)}(p) = Q^{-1}(\alpha, \beta(1-p)). \tag{A.10}$$

Utilizando la identidad

$$TVaR_X(p) = \frac{1}{p}\int_p^1 VaR_X(v)dv$$

se obtiene, según las definiciones en las ayudas con la librería VaRES, que para la Gamma transladada las fórmulas son las anteriores añadiendo el valor de k en cada caso.

A.2.5. VaR, TVaR y ES con aproximación de punto de silla

Una descripción de la metodología de punto de silla está en Huang y Oosterlee (2011).

Apéndice B

Implementación en R del cálculo de la probabilidad de default y del capital expuesto

```
#### Instalar_CargarLibrerias ####
library(PDQutils), library(VaRES), library(actuar),
library(missing), library(kernlab), library(readxl),
library(e1071),library(kknn), library(MASS), library(class),
library(rpart), library(ada),library(nnet),library(lattice),
library(rpart.plot), library(randomForest), library(caret),
library(fBasics), library(imputeTS), library(Hmisc),
library(funModeling), library(tabplot),library(plyr),
library(DataExplorer), library(data.table), library (ggplot2),
library(lme4), library(languageR), library(rms), library(aod),
library(MKmisc), library(MLDS),library(dummies), library(reshape),
library(ROCR)

#### Bases de datos ####
setwd ("D:/ETALZAT/Tesis_riesgo_credito_2018")

# Base prueba
Cast <- read.csv2(file.choose(),header = T)

# Variables que no incluiremos en el análisis debido a que son
# categóricas y otras causan sesgos.
which(colnames(Cast)== "anho_castigo_1")
which(colnames(Cast)== "anho_desembolso_0001_1") # Continúa
```

```
# Eliminar las variables que no se incluirán de la base prueba
Cast1 <- Cast[,-c(1,2,3,4,5,6,7,9,13,14,16,20,22,42,46,47,48,50,
51,52,54,56,58,59,60,61,63,64,74,86,92,95,104,118,119,120,121,122,
123,124,138,141,156,159,185,210,232,235)]

# Transformación de la base de datos: Estandarización
Cast2 <- as.data.frame(scale(Cast1,center = TRUE, scale = TRUE))

# Agregamos la variable nit (primary key)
Cast3 <- as.data.frame(cbind(Cast2,Cast[,235]))

names(Cast3)[188] = "Castigos"
Cast3$Castigos <- as.factor(Cast3$Castigos)
Cast3$Castigos

# Revisar que no se generen cambios de formato o que creen NA's
plot\_intro(Cast3)
names(Cast3)
View(Cast3)
str(Cast3)
summary(Cast3)

# Identificar en que posicion de la base Cast3 estan las
variables significativas ####
which(colnames(Cast3)== "dias_mora_1")
which(colnames(Cast3)== "max_mora") # Continúa

# En orden de significancia
Cast4 <- as.data.frame(Cast3[,c(3,185,182,139,12,143,183,128,
156,170)])

#### Descriptiva de las 10 variables mas significativas ####
par(mfrow=c(2,2))
hist(log(Cast3$dias_mora_1),100,main="Histograma_log_días_mora",
xlab = 'log_días_mora', ylab='frecuencia')

hist(log(Cast3$max_mora),100,main="Histograma_log_max_mora",
xlab = 'log_max_mora', ylab='frecuencia')

par(mfrow=c(2,1))
hist(log(Cast3$interm_cartera_consumopvehiculos),100,main="Histog
```

```
rama_log_interm_cartera_consumopvehiculos", xlab =
'log_interm_cartera_consumopvehiculos', ylab='frecuencia')

hist(log(Cast3$com_seguros),100,main="Histograma_log_com_segu
os", xlab = 'log_com_seguros', ylab='frecuencia') # Continúa

#### Correlaciones y distribucion de las 10 variables
mas significativas ####
cross_plot(Cast3,target = "Castigos",input =
c("dias_mora_1"),path_out = 'my_folder')

cross_plot(Cast3,target = "Castigos",input =
c("max_mora"),path_out = 'my_folder') # Continúa

#### Particionamos la base de datos en 70% entrenamiento
y 30% prueba
p <- sample(1:nrow(Cast3),nrow(Cast3)*0.7)
datos.aprendizaje <- Cast3[p,]
datos.testing <- Cast3[-p,]

# Creamos la partición de la base de datos para las estimaciones
de 5 modelos de machine learning
datos <- Cast3
set.seed(42)
p <- createDataPartition(datos$Castigos, p = 0.7, list = F)
datos.aprendizaje <- datos[p,]
datos.testing <- datos[-p,]

#### Modelo de bosques aleatorios ####
control <- trainControl(method = "none")
rf.model <- train (Castigos~.,
                   data=datos.aprendizaje,
                   method = "rf",
                   ntree = 150,
                   trControl = control)

# Probabilidades de castigo o default
prediction <- predict(rf.model,datos.testing[,-188]),
type="prob")

# Obtenemos un resumen de desempeño con la matríz de confusión
```

```
confusionMatrix(prediction,datos.testing$Castigos)

# Performance del modelo - ROCR & AUC
pred_1 <- prediction(as.numeric(prediction),
as.numeric(datos.testing[,ncol(datos.testing)]))
pred_1
perf_1 <- performance(pred_1, "tpr", "fpr")
plot(perf_1, main="ROCR de bosques aletarorios",
type="l", lty=4,
pch=5, col=4)
abline(h=1,v=1)
perf_1

# auc
auc <- performance(pred\_1, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc

#### Modelo de árboles de decisión ####
tree.model <- train(Castigos~.,
                    data = datos.aprendizaje,
                    method = "rpart",
                    trControl = control)

# Probabilidades de castigo o default
predictiontree <- predict(tree.model,datos.testing[,-188]),
type ="prob")

# Obtenemos un resumen de desempeño con la matríz de confusión
confusionMatrix(predictiontree,datos.testing$Castigos)

# Performance del modelo - ROCR & AUC
pred_2 <- prediction(as.numeric(predictiontree),
as.numeric(datos.testing[,ncol(datos.testing)]))
hist(pred_2)
pred_2
perf_2 <- performance(pred\_2, "tpr", "fpr")
plot(perf_2)
abline(h=1,v=1)
```

```
# auc
auc <- performance(pred\_2, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc

#### Modelo Knn ####
knn.model <- train(Castigos~.,
                   data = datos.aprendizaje,
                   method = "kknn",
                   trControl = control,
                   metric = "Sensitivity",
                   verbose = FALSE)

# Probabilidades de castigo o default
predictionknn <- predict(knn.model,datos.testing[,-188])
,type="prob")

# Obtenemos un resumen de desempeño con la matríz de confusión
confusionMatrix(predictionknn,datos.testing$Castigos)

# Performance del modelo - ROCR & AUC
pred_3 <- prediction(as.numeric(predictionknn),
as.numeric(datos.testing[,ncol(datos.testing)]))
pred_3
perf_3 <- performance(pred_3, "tpr", "fpr")
plot(perf_3)
abline(h=1,v=1)

# auc
auc <- performance(pred_3, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc

#### Modelo SVM ####
svmR.model <- train(Castigos~.,
                    data = datos.aprendizaje,
                    method = "svmRadial",
                    trControl = control)
```

```
# Probabilidades de castigo o default
predictionsvmR <- predict(svmR.model,datos.testing[,-188])

# Obtenemos un resumen de desempeño con la matríz de confusión
confusionMatrix(predictionsvmR,datos.testing$Castigos)

# Performance del modelo - ROCR & AUC
pred_4 <- prediction(as.numeric(predictionsvmR),
as.numeric(datos.testing[,ncol(datos.testing)]))
pred_4
perf_4 <- performance(pred\_4, "tpr", "fpr")
plot(perf_4)
abline(h=1,v=1)

# auc
auc <- performance(pred\_4, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc

#### Modelo Logit ####
logit.model <- train(Castigos~.,
                     data = datos.aprendizaje,
                     method = "LogitBoost",
                     trControl = control)

# Probabilidades de castigo o default
predictionlogit <- predict(logit.model,datos.testing[,-188])
\#, type = "prob")

# Obtenemos un resumen de desempeño con la matríz de confusión
confusionMatrix(predictionlogit,datos.testing$Castigos)

# Performance del modelo - ROCR & AUC
pred_5 <- prediction(as.numeric(predictionlogit),
as.numeric(datos.testing[,ncol(datos.testing)]))
pred_5
perf_5 <- performance(pred_5, "tpr", "fpr")
plot(perf_5)
abline(h=1,v=1)
```

```
perf5_num <- as.numeric(perf_5$y.values)
perf5_num <- perf_5$x.values
perf5_num <- perf_5$alpha.values
perf5_num <- perf\_5$x.name

# auc
auc <- performance(pred\_5, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,4)
auc

#### Graficas AUC ####
plot(perf1_num, pch=5, col= 1
     ,main = "AUC_por_modelo", xlab='Indicador', ylab='AUC')
lines(perf1_num,col=c("black"))
lines(perf2_num,col=c("blue"))
lines(perf3_num,col=c("red"))
lines(perf4_num,col=c("green"))
lines(perf5_num,col=c("yellow"))
legend("topleft",legend=c("bosquesaleatorios","arbolesaleatorios",
                          "knn","svm","logistico"),
       pch=c(1,2,3,4,5),col=c("black","blue","red","green","yellow"))

#### Graficas distribucion probabilidades todos los
modelos ####
par(mfrow=c(2,3))
plot(prediction, main =
"Distribución_probabilidad_bosques_aleatorios",
     xlab='probabilidad', ylab='frecuencia')
plot(predictiontree, main =
"Distribución_probabilidad_árboles_aleatorios",
     xlab='probabilidad', ylab='frecuencia')
plot(predictionknn, main = "Distribución_probabilidad_knn",
     xlab='probabilidad', ylab='frecuencia')
plot(predictionsvmR, main = "Distribución_probabilidad_svm",
     xlab='probabilidad', ylab='frecuencia')
plot(predictionlogit, main = "Distribución_probabilidad_logit",
     xlab='probabilidad', ylab='frecuencia')

#### Calcular la distribucion perdidas L con base a p
(probabilidades) y a F(saldos) ####
```

```
setwd ("D:/ETALZAT/Tesis\_riesgo\_credito\_2018")
Cast\_final2 <- read.csv2(file.choose(),
                         header = T)

# Distribucion de capital expuesto con  base a
(probabilidades) y a F(saldos).
setwd ("D:/ETALZAT/Tesis\_riesgo\_credito\_2018")
Cast\_final2 <- read.csv2(file.choose(),
                         header = T)

names(Cast_final2)[1] = "lj"
names(Cast_final2)[2] = "Fj"
Cast_final2$Fj <- (Cast_final2$Fj/1.0e+09)
Cast_final2$lj <- as.numeric(Cast_final2$lj)
summary(Cast_final2)

# Cifras en millones de pesos
LPcumul_2018 = function(lj,Fj){
  k1 = sum(lj*Fj)
  k2 = sum(lj*Fj^2)
  k3 = sum(lj*Fj^3)
  k4 = sum(lj*Fj^4)
  g1 = k3/k2^(3/2)
  g2 = k4/k2^2
  kS = c(k1,k2,k3,k4,g1,g2)
  names(kS)=c("k1","k2","k3","k4","g1","g2")
  return(kS)}

LPcumul_2018
attach(Cast_final2)

#### Probabilidades vs saldos en escala log ####
plot(log(Fj),log(lj))

#### Cálculo de VaR y TVaR con metodología NP por IC del
90%, 95% y 99% ####
kS = LPcumul\_2018(lj,Fj)
Fs = aggregateDist("npower",  moments = kS[-c(3,4)])
VaR.L.np = actuar:::VaR(Fs)
VaR.L.np
TVaR.L.np =  CTE(Fs)
```

```
TVaR.L.np

#### Tabla VaR y TVaR de L por lo métodos NP ####
M_2018=cbind(VaR.L.np,VaR.g,TVaR.L.np,TVaR.p,VaR.L.cf)
M_2018

# Densidad Edgeworth con PDQutils\_Calculo de la distribucion
de costos o capital expuesto
xvals = seq(58,74,0.1)
p1e = dapx_edgeworth(xvals, kS[1:4])

# Nombre de la grafica: Distribucion de perdidas agregadas por
# default
par(mfrow=c(1,1))
plot(xvals,p1e,type='l',lwd=3, main =
"Distribución_capital_expuesto_2018",
xlab='millones', ylab='densidad de L')
points(VaR.L.np,rep(0,3),pch=19,col='red')
points(TVaR.L.np,rep(0,3),pch=19,col='blue')
```