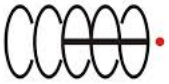

Facultad de Ciencias Económicas y de Administración
Universidad de la República

UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRACIÓN.

Pasantía para obtener el Título de Licenciado en Estadística.

Informe de Pasantía.
Modelo de Scoring Crediticio en una empresa financiera.

Leticia Colombo, Camila Cosentino.

Montevideo,
URUGUAY
27 de octubre de 2015

UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN.

El tribunal integrado por los abajo firmantes aprueba el trabajo de Pasantía:

Modelo de Scoring Crediticio en una empresa financiera.

Leticia Colombo, Camila Cosentino.

Tutores académicos: Ramón Álvarez, Andrés Castrillejo.

Tutor empresarial: Martín Rivero.

Cátedra:

Puntaje:

Tribunal:
Profesor: Alvarez, Ramón.
Profesor: Mesa, Andrea.
Profesor: Nalbarte, Laura.

Fecha:

Agradecimiento

A la Universidad de la República, por darnos la oportunidad de estudiar y ser profesionales. A los profesores durante toda la carrera profesional porque todos han aportado con un granito de arena a nuestra formación y han compartido sus conocimientos con nosotros.

De igual manera agradecer a nuestros profesores de Investigación, Ramón Álvarez y Andrés Castrillejo por sus visiones críticas de muchos aspectos cotidianos de la vida, por su rectitud en sus profesiones como docentes, por sus consejos, que ayudan a formarte como persona e investigador.

Un especial agradecimiento a la Empresa que colaboró con la información pertinente para la realización de esta Pasantía. Por brindarnos un lugar en su empresa para de esta manera desempeñar una labor de trabajo en la misma, y a su vez permitir la culminación de una etapa importante en nuestras vidas.

A Martín Rivero, tutor empresarial por compartir todos sus conocimientos, por estar siempre pendiente y darnos apoyo en todo momento. A nuestros compañeros de trabajo gracias por el apoyo y amistad brindados durante el período de la pasantía.

A nuestras familias y amigos por su comprensión, dedicación y apoyo. Y a todas las demás personas que no fueron citadas, pero que de alguna manera directa o indirecta contribuyeron a la realización de este trabajo.

Resumen del Informe

En este trabajo se realizaron modelos de Credit Scoring, desarrollados utilizando información de una financiera real. La población objetivo fue toda persona física que haya solicitado un crédito al consumo y cuyo crédito fue aprobado por los analistas, durante el período del segundo semestre de 2011 al primer semestre de 2014.

Los Credit Scoring son procedimientos estadísticos que se usan para clasificar a los solicitantes del crédito, inclusive a los que ya son clientes de la entidad crediticia, en los tipos de riesgo *Bueno* y *Malo*.

Mediante una puntuación se mide el riesgo de un prestatario y/o de la operación en el momento en el que se está llevando a cabo la solicitud, es decir, se estima cuál será el comportamiento del crédito hasta su vencimiento atendiendo al riesgo del cliente.

En la actualidad el riesgo se ve como una oportunidad más que una amenaza, debe ser considerado como una inversión en la organización, por lo que debería brindar a las instituciones ventajas competitivas y mejoras de gestión. El riesgo se mide, se evalúa y se cuantifica. De los distintos tipos, puede considerarse al riesgo de Crédito como el más importante al que deben hacer frente las entidades financieras.

Para evaluar el riesgo crediticio o la conveniencia de otorgar un crédito, se utilizan una gran variedad de metodologías. De todas ellas se estudiaron la Regresión Logística y los modelos CART.

Modelo de Regresión Logística

Los modelos de regresión logística permiten estudiar las diferencias entre dos o más grupos de individuos definidos a priori, con respecto a una serie de variables. Modelan la probabilidad de pertenencia a una categoría en particular.

La variable dependiente es definida como la ocurrencia o no de un acontecimiento, en este caso de ser *Malo* o *Bueno* en relación al comportamiento en los pagos. El principal objetivo fue encontrar el modelo que ajuste mejor a los datos y que sea lo más parsimonioso posible.

Luego de estimar los parámetros del modelo, se debía poder predecir el

valor de la variable en función de las variables explicativas. Para realizar dicho procedimiento se debió determinar cuál es el valor crítico a partir del cual las estimaciones implican un valor de 1 para la variable de respuesta. El problema fue determinar cuándo un valor es chico o grande.

Se realizan test de significación de los modelos para ver si estos eran adecuados y test de significación de los parámetros para ver cuales debían incluirse o no en los modelos.

La bondad de ajuste fue utilizada para resumir la discrepancia entre los valores observados y los valores esperados en el modelo de estudio.

Una vez que se obtuvo un modelo en donde tanto los parámetros como el modelo en su conjunto eran significativos, se procedió a elegir el punto de corte más apropiado y a comprobar cuán bueno fue el ajuste de los valores predichos por el modelo, utilizando otras herramientas.

Para la elección del punto de corte se utilizó el estadístico de *Kolmogorov-Smirnov* junto con la *curva ROC* y el área debajo de la curva. Para ello se utilizaron las tablas de clasificación en dónde se cruzaron el número de observaciones que tenía cada grupo a priori con las perdiciones realizadas. Con estas tablas se puede calcular la tasa de predicción positiva y la tasa de predicciones negativas.

En un principio se realizan las estimaciones en base a una muestra del 90 % de la población y luego con el 50 %, con el fin de contar con más datos de prueba para evaluar el desempeño del modelo .

Cómo los clientes calificados como $Malo$ no llegan a ser el 10 % del total de la población se decidió, para explorar la técnica, tomar una muestra en la que la proporción de clientes $Malo$ fuese igual a la de *Bueno*.

En la búsqueda de mejorar los resultados, como se observaba que los clientes con categoría ocupacional *Activos* tenían un perfil diferente a los *Pasivos* se decidió considerar un modelo diferente para cada uno de ellos tomando las respectivas muestras al 50 %.

Lo mismo se realizó con los clientes que ya habían operado en la empresa más de una vez y con los que era su primera operación, se estimó un modelo para cada perfil.

Luego de probar varias alternativas se decidió que el más adecuado era el que se estimó con una muestra del 50 % incluyendo las siguientes variables: *Cantidad de veces que operó, Edad, Sexo, Antigüedad Laboral, Clearing, Ocupación, Cuotas totales, Valor cuota / Total de Ingresos.*

Luego de estimado el modelo de regresión logística se procede a implementar la técnica CART.

Árboles de regresión y clasificación - CART -

Los arboles de regresión y clasificación fueron propuestos para separar las observaciones que componen la muestra asignándolas a grupos establecidos a priori. Pueden verse como la estructura resultante de la partición recursiva del espacio de representación, esta partición se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada nodo hoja se refiere a una decisión (clasificación).

En el proceso de construcción de todos los arboles de clasificación estimados resultan ser "significativas" las mismas variables que fueron consideradas en el modelo final de regresión logística. Por este motivo se decide considerar, en la construcción de los árboles, las variablese: Cantidad de veces que operó, Sexo, Edad, Antigüedad Laboral, Clearing, Ocupación, Cuotas totales y Valor cuota sobre Total de Ingresos.

Para llevar a cabo el procedimiento se realiza, en primera instancia, un árbol de clasificación considerando una muestra aleatoria simple del 50 % de las observaciones. Luego se realiza otro con una muestra que tuviese igual proporción de clientes clasificados como *Bueno* y clientes clasificados como *Malo*, a modo de ejemplo.

En ambos casos se obtuvieron los arboles completos, luego se evaluó cuál era la poda más adecuada utilizando la medida CP y el error de validación cruzada.

Dejando de lado el árbol con la muestra equilibrada, cabe destacar que esta metodología proporciona estimaciones similares al modelo de regresión logística. Esto es importante ya que reafirma los resultados obtenidos en la metodología anterior, pude ser un buen complemento debido a su fácil interpretación.

Sucede lo mismo con el árbol de clasificación realizado con la muestra con igual proporción de *Bueno* y *Malo*, sin embargo esta muestra no es representativa de la realidad.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Hoy en día el consumo de créditos es muy utilizado tanto en Uruguay como en todo el mundo. Para las empresas financieras la eficiencia a la hora de tomar la decisión de otorgar un crédito es primordial por lo que, entre otros recursos, es habitual el uso de herramientas estadísticas para disminuir los riesgos.

En este informe se desarrollará un modelo de Scoring Crediticio que logre predecir el comportamiento de los clientes que solicitan créditos al consumo en una financiera del mercado uruguayo. La información con la que se cuenta permite mediante técnicas estadísticas determinar un puntaje a cada cliente para de esa forma tener un control subjetivo del riesgo, complementario al realizado por los analistas.

Este informe muestra, en el actual capítulo la justificación de la pasantía y sus objetivos; en el capítulo 2 se muestra la fundamentación teórica que se necesita para llevar a cabo el mismo y en el capítulo 3 se describe el proceso de creación y análisis de los modelos de Credit Scoring tanto para la regresión logística como para el análisis de árboles de clasificación, CART.

1.1. Justificación

Este estudio se realiza debido a la necesidad de una empresa de mediano tamaño del mercado uruguayo de poder contar con un modelo que ayude a los analistas en la toma de decisiones al momento de otorgar un crédito. El objetivo era contar con una herramienta objetiva para poder caracterizar el perfil de sus clientes y discriminarlos entre “buenos” y “malos”.

Las empresas deben asumir riesgos en su toma de decisiones en busca de la máxima rentabilidad en relación al riesgo-rentabilidad como algo inseparable de la gestión de las mismas, por lo que el estudio de este último se ha convertido en algo esencial para el desarrollo de sus actividades.

Las técnicas de Credit Scoring son muy utilizadas y además rentables, dado que una pequeña mejora en el desempeño puede significar un incremento en las ganancias. Si bien estas no sustituyen a los analistas, sí tienen en general suficiente capacidad predictiva como para introducir mejoras importantes en la evaluación de los créditos.

Estas técnicas tienen muchas ventajas sobre todo cuando se compara con el análisis subjetivo pero también tiene algunas desventajas [Schreiner, 2002.].

La ventaja principal es que cuantifica al riesgo de morosidad como una probabilidad, a través de un puntaje que penaliza varios factores.

Es consistente, ya que dos personas con las mismas características serán calificadas del mismo modo, sin embargo la sentencia de un analista podría verse influenciada por factores externos.

El scoring estadístico considera una amplia gama de factores, las normas para la evaluación subjetiva de solicitudes pueden especificar que una solicitud debe cumplir ciertas disposiciones, pero, a diferencia del scoring estadístico, el scoring subjetivo no puede considerar tantas características simultáneamente.

El scoring estadístico puede probarse antes de usarlo para ver cómo funciona y si es necesario hacerle ajustes previos a su implementación.

Revela las relaciones entre el riesgo y las características del cliente que solicita el crédito. No sólo es posible obtener la probabilidad de mora teniendo en cuenta todas las características implícitas sino que es posible analizar la relación con una característica en particular. El scoring estadístico indica precisamente qué tan fuertes son las relaciones a diferencia del análisis subjetivo.

Si bien tiene todas estas ventajas también tiene alguna desventaja que deberían ser tenidas en cuenta [Schreiner, 2002.].

Se requiere de una base de datos extensa para poder elaborar el modelo,

no todas las empresas cuentan con esa información o en muchos casos no han sido respaldados de manera adecuada.

No sólo la base debe ser extensa sino que se requiere de muchos datos de cada préstamo, para evaluar las diferentes características.

En muchos casos la información es imprecisa o aleatoria, o los datos son erróneos, es un punto que hay que tener en cuenta. Mientras esas perturbaciones no sean demasiadas no habría demasiados problemas.

El scoring estadístico supone que una buena parte del riesgo está vinculada con características cuantificadas. Supone, por ejemplo, que el riesgo está vinculado, por ejemplo, con el género, la edad, los atrasos en créditos anteriores, la actividad laboral, etc. Pero la cuestión es qué proporción del riesgo está asociada con esos factores y qué proporción está asociada con los factores.

Este tipo de modelo estadístico supone que el futuro será como el pasado, no prevé riesgos externos al solicitante del crédito como catástrofes naturales o cambios en la economía, por ejemplo. Esto hace necesaria su actualización periódica.

El scoring estadístico es susceptible al mal uso sobre todo si se ignora el pronóstico y se continúan haciendo lo que siempre se ha hecho. La solución para este punto sería la capacitación y seguimiento de los analistas.

Habiendo establecido la justificación del trabajo, los riesgos, y las ventajas y desventajas de la implementación de un modelo de Scoring Crditicio estamos en condiciones de empezar a realizarlo.

1.2. Objetivos

Obtener un modelo de “Credit Scoring” alternativo al utilizado actualmente por la empresa, que logre predecir de la mejor manera posible las futuras solicitudes.

Capítulo 2

Marco Teórico

2.1. ¿Qué es el "Credit Scoring"?

Las técnicas de Credit Scoring se han utilizado para otorgar créditos en la industria crediticia por más de 40 años, permitiendo el crecimiento del número de consumidores de crédito, crecimiento que ha sido propiciado por el uso de la informática lo que permitió el avance de las técnicas estadísticas por el manejo de grandes cantidades de datos.

Según Hand y Henley [Hand et al., 1997], los Credit Scoring son procedimientos estadísticos que se usan para clasificar a aquellos que solicitan crédito en los tipos de riesgo bueno y malo. La construcción de toda aplicación del Credit Scoring se realiza tomando la información del cliente contenida en las solicitudes del crédito, de fuentes internas e, incluso, de fuentes externas de información.

El Credit Scoring estima, en el momento en el que se está llevando a cabo la solicitud, cuál será el comportamiento del cliente atendiendo al riesgo. Se evalúa a través de un modelo predictivo de comportamiento de pago o reembolso mediante una puntuación que mide el riesgo del prestatario y/o de la operación. En general, estos métodos de calificación de créditos se aplican para obtener un conocimiento sobre distintos aspectos tales como [Hand et al., 1997]:

- el comportamiento financiero en cuanto a los productos solicitados y a la morosidad;
- la relación entre el riesgo y rentabilidad. El Credit Scoring aporta información sobre el precio o prima por riesgo, volatilidad, diversificación, etc.;
- el costo de la operación. La agilización general de procesos que se consigue con el Credit Scoring permite la reducción del costo en el proceso de concesión de un crédito.

2.2. Riesgo en las entidades financieras

2.2.1. Concepto y tipos de riesgo.

Los conceptos de riesgo fueron extraídos de las notas internas del Contador Martín Rivero [Rivero, 2012].

Según estas notas, "...la actividad económica se desarrolla en un ambiente de incertidumbre, convirtiendo el riesgo en un factor inherente a la misma. De este modo, surge aquél como la contingencia, probabilidad o proximidad de un daño o peligro, en concreto, de sufrir una pérdida. La incertidumbre, junto con la aleatoriedad, constituyen las características principales del riesgo, añadiéndose como tal el conflicto, ya que el riesgo se presenta ante situaciones diferenciadas entre las que elegir."

Así pues, el riesgo se ha convertido en uno de los rasgos básicos del entorno económico actual al que se enfrentan las empresas, que deben asumir riesgos en su toma de decisiones en busca de la máxima rentabilidad en relación al binomio riesgo-rentabilidad como algo inseparable de la gestión de las mismas.

En el caso de las entidades financieras, esta característica es esencial a la actividad que desarrollan, consistente en la concesión de créditos, asumiendo un riesgo cuando prestan unos recursos financieros que otros clientes les han cedido, sin controlar posteriormente el destino y utilización de los mismos. Desde un punto de vista tradicional, "...el riesgo es todo aquello que podrá impedir u obstaculizar el cumplimiento de los objetivos. Se define como la eventualidad de que el patrimonio institucional se vea afectado negativamente por la probable ocurrencia de un evento. Si sólo se visualiza el riesgo de

esta manera, se está limitando o restringiendo el concepto al término de amenaza" [Rivero, 2012].

En la nueva concepción, el riesgo es una oportunidad más que una amenaza. De este modo, debe brindar a la institución ventajas competitivas y mejoras de gestión. Para poder llegar a visualizar al riesgo como una oportunidad, el mismo debe ser considerado como una inversión en la organización.

Entonces, el riesgo hoy en día, es la incerteza en los objetivos. Es tanto una amenaza al cumplimiento de los objetivos, como una oportunidad a que los mismos se cumplan [Rivero, 2012].

"Desde el punto de vista matemático financiero, se puede definir como una medida cuantitativa que expresa, tanto el grado en que un resultado tiene el potencial de ser diferente al esperado como, el impacto asociado a dicha variación" [Rivero, 2012].

El riesgo se mide, se evalúa y se cuantifica. La medición del riesgo es semejante a una regla de estimación del nivel de incertidumbre sobre la ocurrencia de este tipo de eventos. La herramienta para medir el riesgo es la probabilidad.

El riesgo global en la actividad financiera resulta de la suma de distintos tipos de eventos de riesgo, los cuales se describen de forma breve a continuación. Esta información fue extraída de las capacitaciones brindadas en la empresa.

1 Eventos Accidentales: son los asociados a los eventos súbitos e imprevistos no predecibles a los cuales está expuesto una organización. Puede haber eventos accidentales a las propiedades, personas, responsabilidad civil, beneficio bruto, etc.

2 Eventos de Fraude: son los eventos que derivan de una acción que resulta contraria a la verdad y a la rectitud. Se procede de manera ilegal o incorrecta según los parámetros establecidos con el objetivo de obtener algún beneficio.

3 Eventos Operacionales: este tipo de eventos son los asociados a eventos no accidentales originados por el no funcionamiento o el funcionamiento inadecuado de los procesos internos (incluye los informáticos); los sistemas de información y el personal de una organización.

4 Eventos Financieros: son aquellos tipos de eventos que causan impacto en los resultados financieros de una organización debido a cambios en las condiciones de mercado; el no cumplimiento por parte de un tercero de las obligaciones financieras para con la misma; la responsabilidad de cumplir con las obligaciones financieras por parte de la institución. Estos pueden ser: de mercado, de crédito o de liquidez.

5 Eventos de Reputación: eventos causados por una opinión pública negativa, afectando con esto la habilidad de la organización de mantener las actuales y/o establecer nuevas relaciones o servicios.

6 Eventos Estratégicos: eventos asociados tanto con la toma de decisiones que sobre el negocio hacen las organizaciones como con el entorno en que el negocio se desenvuelve.

7 Evento de Cumplimiento: es la exposición al riesgo derivado de omisiones o actuaciones del Banco en sus obligaciones regulatorias, administrativas, tributarias, de seguridad social y de prevención contra el lavado de activos.

8 Evento de Lavado de Activos: es el riesgo a que se utilice la estructura de la institución para que bienes de origen delictivo integren el sistema financiero, aparentando haber sido obtenidos en forma lícita.

2.2.2. Riesgo de crédito.

De los distintos tipos de riesgo, puede considerarse al riesgo de crédito como el más importante al que deben hacer frente las entidades financieras, por ser intrínseco a la actividad que desarrollan, y porque es la principal incertidumbre a la que estas entidades se enfrentan en las operaciones de activos que les vinculan a sus clientes.

El riesgo de crédito se define como la eventualidad de que el patrimonio institucional se vea afectado debido a la incapacidad del cliente o contraparte de cumplir en tiempo y forma con los acuerdos contractuales pactados con la institución [Rivero, 2012].

Se determina que existen dos tipos de riesgo de crédito: el riesgo de incertidumbre, que se refiere a la pérdida potencial derivada de que la contraparte no pueda cumplir con sus obligaciones financiera en las condiciones definidas contractualmente; y el riesgo de mercado, que se define como la pérdida

potencial que podría sufrir una institución financiera o derivados, como consecuencia de que el valor de mercado de éstos disminuya. El segundo tipo, plantea exposición al riesgo de crédito aún en el caso de que la contraparte no sufra quebranto alguno.

Importancia del Riesgo.

Hasta hace poco tiempo, los altos tipos de intereses existentes y la escasa competencia permitían a las entidades financieras mantener elevados márgenes con los que se cubría el riesgo de crédito. Los bancos conocían relativamente bien el riesgo asumido y la rentabilidad que las operaciones producían. En la última década, sin embargo, tanto los avances tecnológicos y financieros como la globalización de los mercados, han hecho que los márgenes disminuyan y la competencia alcance cuotas antes impensables, lo que ha llevado a los bancos a replantearse la rentabilidad que obtienen con sus operaciones, y sobre todo, el riesgo que asumen.

Uno de los factores necesarios para medir el riesgo de crédito es la *probabilidad de incumplimiento*, esta es la "probabilidad de que la contraparte no haga frente a sus obligaciones contractuales" [Reyes, 2007].

A la hora de estimar la probabilidad de impago no hay que olvidar la fuerte correlación existente entre el grado de incumplimiento y los ciclos económicos. El problema reside en que, a la hora de medir el nivel de riesgo, se está ignorando uno de sus elementos claves, la existencia de ciclos económicos, disminuyéndose el riesgo en épocas de bonanzas y sobrevalorándose en épocas de crisis.

Por lo que el incumplimiento no es una variable aislada, sino que su valor afectará el resto de los factores que determinan el riesgo de crédito.

La variable incumplimiento (o default) depende, a su vez, de los siguientes factores: definición del incumplimiento, calidad crediticia de la contraparte, ciclo económico, y condiciones del mercado (tipo de interés).

Hasta ahora, cuando se ha hablado de probabilidad de impago, sólo se ha considerado dos posibilidades: cumplimiento e incumplimiento, en este caso resultado "Bueno" o "Malo", pero en muchos casos se establece más de dos estados, tantos como niveles de score existan.

2.3. Metodología

Una vez presentados los riesgos a los que deben enfrentarse las entidades financieras y, en especial, el de crédito, el presente trabajo se completa con un estudio empírico, cuyo objetivo es analizar y valorar la morosidad, como forma de manifestación de dicho riesgo, en las entidades financieras.

Se tratará de determinar los factores de mayor influencia en el comportamiento de pago de los clientes de las entidades financieras, y que permitan distinguir los clientes solventes que cumplen con sus obligaciones, de los morosos que las incumplen o se retrasan en su cumplimiento.

Por tanto, el objetivo del trabajo se centra en explicar el comportamiento de una variable categórica con dos modalidades: ser un cliente moroso, $Malo$, o bien un cliente no moroso, $Bueno$.

Para evaluar el riesgo crediticio o la conveniencia de otorgar un crédito, se puede utilizar una gran variedad de metodologías: análisis discriminante, regresión lineal, regresión logística, algoritmos de particiones recursivas (árboles o modelos CART), redes neuronales, etc.; y por otra parte la decisión de un analista acerca de si otorgar un crédito o no. Este último se considera que es esencial ya que complementa la herramienta estadística utilizada.

Entre todas las metodologías disponibles, se estudiaran e implementarán la Regresión Logística y también, para complementar el estudio, los Árboles de Regresión y Clasificación .

Estos proveen para cada deudor una probabilidad de default y clasifica a los deudores en uno de los dos grupos, Bueno o Malo.

2.3.1. Modelo de Regresión Logística

Para establecer la relación existente entre una variable dependiente Y no métrica, en particular dicotómica, y un conjunto de variables independientes $(x_1, x_2, ..., x_k)$ que pueden ser tanto cualitativas como cuantitativas se podrá ajustar un ecuación de tal forma de que a través de la estimación los parámetros $(\beta_0, \beta_1, \beta_2, ..., \beta_k)$ se prediga el comportamietno de Y.

Cuando la respuesta a un problema, Y, está dentro de dos categorías. En lugar de modelarla directamente, la regresión logística modela la probabilidad de que ésta pertenezca a una categoría en particular. Y está dada por:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}} \tag{2.1}$$

El modelo de regresión logística permite estudiar las diferencias entre dos o más grupos de individuos definidos a priori, con respecto a una serie de variables.

Tiene como objetivo analizar la relación entre una variable dependiente categórica con modalidades que se corresponden con los grupos analizados, y un conjunto de variables independientes.

Una vez definida la variable dependiente como la ocurrencia o no de un acontecimiento, en este caso de ser $Malo$ o $Bueno$ en relación al comportamiento en los pagos, el modelo de regresión logística la expresa en términos de probabilidad. Se utiliza la función logística para estimar la probabilidad de que ocurra el acontecimiento dados determinados valores de las variables explicativas.

Puesto que el modelo no es lineal, para lograrlo se considera una transformación de la función logística, $logit$ o logaritmo de los odds, $logit(\pi) = log(\frac{\pi}{1-\pi})$. Si π es la probabilidad de que un suceso ocurra, al cociente de probabilidades $\pi/(1-\pi)$ es llamado $odds$.

La formulación anterior facilita la interpretación del modelo y de sus parámetros. Para la interpretación de los parámetros se debe calcular el cociente llamado $odds$ que más adelante se detallará como realizarlo.

En este caso la población se encuentra dividida en dos grupos : P los que a priori fueron clasificados como $Malo$ y N los que fueron clasificados como $Bueno$, asociado a cada individuo se conoce un vector de características.

Estimado el modelo, su capacidad predictiva se evalúa mediante el establecimiento de un punto de corte óptimo, que permite asignar los casos a cada uno de los grupos definidos por la variable dependiente.

El modelo, a su vez, tendrá un poder predictivo pues se considera que los criterios utilizados para clasificar a la población actual, podrían ser utilizados para los nuevos elementos que se incorporen en ella.

2.3.1.1. Estimación del Modelo de Regresión Logística.

Los modelos de regresión se han convertido en una herramienta fundamental en el análisis de datos en donde se describe la relación entre una variable de respuesta y una o más variables explicativas [Hosmer y Lemeshow, 2013].

Es necesario destacar que el principal objetivo en toda regresión es encontrar el modelo que ajuste mejor a los datos y que sea lo más parsimonioso posible.

Se tienen observaciones independientes de (X_i, y_i), $i = 1, 2, ..., n$, donde y_i es el valor de una variable dicotómica y X_i es un vector con los valores de las diferentes variables para las i observaciones. La variable de respuesta se asume que toma los valores 0 y 1, representando la ausencia o presencia de determinada característica.

En un modelo de regresión lineal dado por $y = E(Y|x) + \varepsilon$, el ajuste se da a través de la siguiente ecuación [Hosmer y Lemeshow, 2013]:

$$E(Y|x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2.2}$$

Esta expresión implica que los valores posibles de $E(Y|x)$ van de $-\infty$ a $+\infty$. Asumiendo que ε es el error y tiene una distribución normal con media cero y varianza constante.

Cuando se tiene una variable de respuesta dicotómica no es posible utilizarlo, a no ser que se haga una transformación para que el rango de valores esté entre $(0, 1)$. Una de las opciones es la regresión logística ya que no sólo desde el punto de vista matemático es una función extremadamente flexible sino que además es de fácil interpretación.

Simplificando la notación se utiliza $\pi(x) = E(Y|x)$ para representar la esperanza condicional de Y dado x cuando se utiliza la regresión logística. En este caso, el modelo está dado por $y = \pi(x) + \varepsilon$, asumiendo que el error puede tomar dos valores: $\varepsilon = 1 - \pi(x)$ cuando $y = 1$ o $\varepsilon = -\pi(x)$ cuando $y = 0$. Por lo que ε tiene una distribución con media cero y varianza $\pi(x)[1 - \pi(x)]$ [Hosmer y Lemeshow, 2013]. Esta es la distribución de una variable binomial con probabilidad dada por la esperanza condicional $\pi(x)$ [Hosmer y Lemeshow, 2013].

La expresión para el modelo de regresión logística está dada por:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}} \tag{2.3}$$

La transformación de $\pi(x)$, $g(x) = log[\frac{\pi(x)}{1-\pi(x)}] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ es llamada transformación *logit*.

Ajuste del Modelo

El ajuste del modelo de regresión logística se realiza a través del método de máxima verosimilitud que estima valores para los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto de datos observados. En primera instancia se debe construir la función de verosimilitud, que expresa la probabilidad de las observaciones como una función de parámetros desconocidos. El estimador de máxima verosimilitud de los parámetros es elegido de tal forma que maximice esta función, por lo que va a ser el que se ajuste mejor a los datos.

Si Y es codificada como 0 y 1 entonces la expresión para $\pi(x)$ dada en la ecuación (2.3) provee la probabilidad condicional de que Y sea igual a 1 dado x, $P(Y = 1|x)$. Por lo que $1 - \pi(x)$ es la probabilidad de que Y sea 0 dado x, $P(Y = 0|x)$. Entonces, para (x_i, y_i), donde $y_i = 1$, la contribución a la función de verosimilitud es $\pi(x_i)$, y cuando $y_i = 0$, la contribución a la función de verosimilitud es $1-\pi(x_i)$ [Hosmer y Lemeshow, 2013]. Una forma conveniente de expresar la contribución a la función de verosimilitud para (x_i, y_i) es a través de la expresión [Hosmer y Lemeshow, 2013]:

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{2.4}$$

Como las observaciones son independientes, la función de verosimilitud, a la que llamaremos $L(\beta)$, se obtiene como la productoria de la expresión dada en la ecuación anterior,[Hosmer y Lemeshow, 2013]

$$L(\beta) = \prod_{i=1}^{N}[\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}] \tag{2.5}$$

El principio de máxima verosimilitud establece que el estimador de β, es el que maximiza la expresión anterior. Para trabajar con mayor facilidad se trabajara con el logaritmo de la expresión anterior [Hosmer y Lemeshow, 2013]:

$$L(\beta) = ln(L(\beta)) = \sum_{i=1}^{N}[y_i ln[\pi(x_i)] + (1-y_i)ln[1-\pi(x_i)]] \quad (2.6)$$

$$= \sum_{i=1}^{N}[y_i \beta' x_i - ln(1+e^{\beta' x_i})] \quad (2.7)$$

Para encontrar los valores de β que maximizan $L(\beta)$ se deriva con respecto a cada β y se iguala a 0.

Las ecuaciones de verosimilitud, de acuerdo a [Hastie et al., 2009] son:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i(y_i - \pi(x_i)) = 0 \quad (2.8)$$

Son $p+1$ ecuaciones no lineales en β. El primer componente de x_i es 1, entonces la primera ecuación de verosimilitud es $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \pi(x_i)$, que es el número esperado de clases que coinciden con el valor observado.

Para resolver las ecuaciones se utiliza el algoritmo del Newton-Raphson.

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} x_i x_i' \pi(x_i)(1-\pi(x_i)) \quad (2.9)$$

De forma iterativa se empieza con un β^0, y luego se actualiza con

$\beta^{s+1} = \beta^s - (\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'}^{-1}) \frac{\partial L(\beta)}{\partial \beta}$, evaluando las derivadas en β^s.

Para facilitar la interpretación se escribirán las ecuaciones en notación matricial. Al vector de los valores y_i se lo notará como y; a los valores x_i como $X_{N\times(k+1)}$ siendo k el número de variables; p al vector de probabilidades $\pi(x_i)$; y $W_{N\times N}$ será una matriz diagonal de pesos, con elementos $\pi(x_i)(1-\pi(x_i))$ evaluados en un β^s. Por lo tanto las ecuaciones anteriores para esta notación, de acuerdo a [Hastie et al., 2009], son:

$$\frac{\partial L(\beta)}{\partial \beta} = X'(y-p) \quad (2.10)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} = -X'WX \quad (2.11)$$

En el paso s+1 del algoritmo de Newton-Rapshon se tiene:

$$\begin{aligned} \beta^{s+1} &= \beta^s + (X'WX)^{-1}X'(y-p) && (2.12)\\ &= (X'WX)^{-1}X'W(X\beta^s + W^{-1}(y-p)) && (2.13)\\ &= (X'WX)^{-1}X'Wz. && (2.14) \end{aligned}$$

En la segunda y tercera línea se reescriben las ecuaciones como un paso del algoritmo de mínimos cuadrados ponderados, con $z = X\beta^s + W^{-1}(y-p)$.

Estas ecuaciones se resuelven ya que, en cada iteración, p cambia y por lo tanto lo hacen W y z. El problema de mínimos cuadrados ponderados está dado por [Hastie et al., 2009]:

$$\beta^{s+1} \leftarrow \arg\min_{\beta}(z - X\beta)'W(z - X\beta). \quad (2.15)$$

En muchos casos empezar con $\beta = 0$ es una buena elección aunque no garantiza la convergencia pero generalmente el algoritmo sí converge [Hastie et al., 2009].

Punto de Corte.

Luego de estimar los parámetros del modelo, se podrá predecir el valor de la variable en función de las variables explicativas. Para realizar dicho procedimiento se debe determinar cuál es el valor crítico a partir del cual las estimaciones implican un valor de 1 ($Malo$) para la variable de respuesta. Valores grandes de π_i implicarán $y_i = 1$, mientras que los valores chicos implicarán $y_i = 0$. El problema está en determinar cuándo un valor es chico o grande [Blanco, 2006].

Si el punto de corte es 0.5, la regla de decisión será que si $\pi_i > 0{,}5$ entonces $y_i = 1$. Sin embargo esta aproximación es válida si es igualmente probable que ocurra 0 o 1 o si los costos de predecir uno u otro son los mismos.

Encontrar el mejor punto de corte para los datos, implica calcularlos y evaluar en cada caso cómo son pronosticadas las n observaciones.

Cuando el costo de predecir incorrectamente 1 no es el mismo que predecir incorrectamente 0 se pueden utilizar las probabilidades a priori [Blanco, 2006].

En muchos casos en la elección del punto de corte se busca optimizar la sensibilidad y la especificidad del modelo. La sensibilidad se define como el

cociente entre los éxitos observados clasificados como éxitos y el total de éxitos observados, mientras que la especificidad se define como el cociente entre los fracasos observados clasificados como fracasos y el total de los fracasos observados.

Los dos conceptos se basan en un punto de corte óptimo a partir del cual se clasifican observaciones como éxito o fracaso. Este punto óptimo se puede encontrar a partir de la curva ROC. En la siguiente sección se detallará el procedimiento para la elección del punto de corte óptimo y se estudiará la bondad de ajuste del modelo [Blanco, 2006].

2.3.2. Validación de los Modelos y Elección del Punto de corte.

La significación del modelo sirve para testear si el modelo es adecuado o no y la significación de los parámetros se utiliza para testear cuales variables deben ser incluidas en el modelo.

La bondad de ajuste de un modelo estadístico describe lo bien que se ajusta un conjunto de observaciones. Las medidas de bondad en general resumen la discrepancia entre los valores observados y los valores esperados en el modelo de estudio.

2.3.2.1. Test de Razón de Verosimilitud

La razón de verosimilitud del modelo es una prueba para testear la significación del modelo. Se define $\lambda = \frac{L_R}{L_M}$ donde L_M es la verosimilitud del modelo completo, $\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_q+\ldots+\beta_k x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_q+\ldots+\beta_k x_p}}$ y L_R la del modelo reducido, $\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_q}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_q}}$, $q < p$.

$H_0)\beta_1 = \beta_2 = \ldots = \beta_p = 0$

$H_1)$ algún $\beta_k \neq 0$, $k = 1, 2, ..., p$

$-2ln(\lambda)$ se distribuye $\chi^2_{(p+1-q),\alpha}$, siendo $(p+1)$ y q la cantidad de parámetros incluidos en el modelo completo y el modelo reducido respectivamente .

Test de razón de verosimilitud:

$$-2ln(\lambda) = -2ln(\frac{L_R}{L_M}) = -2(lnL_R - lnL_M) \tag{2.16}$$

La hipótesis nula será rechazada para el nivel de significacion α cuando $-2ln(\lambda) > \chi^2_{(p+1-q),\alpha}$. Esto es equivalente a que el p valor del contraste sea menor que el nivel de significación fijado [Blanco, 2006].

2.3.2.2. Estadístico de Wald

Significación de un parámetro en particular.

$H_0)\beta_k = 0$

$H_1)\beta_k \neq 0$

$$W = \frac{\widehat{\beta_k}}{sd(\widehat{\beta_k})}, \tag{2.17}$$

se distribuye aproximadamente normal.

El nivel de significación de un test es un concepto estadístico asociado a la verificación de una hipótesis. Se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula (H_0) cuando esta es verdadera (decisión conocida como "Error de tipo I", o "falsos positivos"). La decisión se toma a menudo utilizando el *p-valor*: si el valor p es inferior a nivel de significación, entonces la hipótesis nula es rechazada [Blanco, 2006].

Una vez que se obtiene un modelo en donde tanto los parámetros como el modelo en su conjunto son significativos, se procede a elegir el punto de corte más apropiado y a comprobar cuán bueno fue el ajuste de los valores predichos por el modelo utilizando otras herramientas.

2.3.2.3. Curva ROC

Una forma de evaluar la calidad de ajuste de un modelo es utilizando la curva ROC (Receiver Operating Characteristic), que capta las características del funcionamiento del modelo a través de la variación de su comportamiento.

Su primera utilización fue durante la Segunda Guerra Mundial para el análisis de las señales de radar, y en consecuencia, entró en la literatura científica en la década de 1950 en el marco de la teoría de detección de señales y la psicofísica. Más tarde, en los años 1970 y 1980, se hizo evidente la importancia de la técnica para la evaluación médica de pruebas y toma de decisiones, y desde entonces se ha visto mucho el desarrollo y uso de la técnica en áreas tales como radiología, cardiología, química clínica y la epidemiología [Krzanowski y Hand, 2009.].

La curva ROC es utilizada para evaluar situaciones en las que el objetivo del modelo es asignar las observaciones a una o más clases. Desafortunadamente, los procedimientos no son perfectos, se cometen errores asignando observaciones a la clase incorrecta por lo que se hace necesario evaluar no sólo el comportamiento del modelo y sus variaciones sino también, si es necesario reemplazarlo por otro [Krzanowski y Hand, 2009.].

Función de clasificación

En este estudio, el objetivo es determinar si el comportamiento del cliente al que se le otorgará un crédito será *Bueno* o *Malo* a través de la regresión logística. Para poder inferir cuál será el comportamiento del cliente se utilizan ciertas variables cuantitativas y cualitativas, X, que producen un score $S(X)$, continuo, como resultado de aplicar dicha función S. La asignación a cada clase se realiza luego comparando el score con un umbral T (perteneciente al recorrido de S(X)), si está por encima de dicho umbral se clasificará en *Malo* sino será *Bueno* [Krzanowski y Hand, 2009.].

Se denota como P a los clientes que a priori fueron clasificados como *Malo* y N a los que fueron calificados como *Bueno*. Se tratará de encontrar una

función score $S(X)$ que produzca puntajes tales que se puedan diferenciar claramente las dos clases, y un umbral que separe por encima aquellos que a priori fueron clasificados como P y por debajo los que se clasificaron como N.

La muestra de entrenamiento servirá para construir la regla de clasificación con la que se verá luego cuán efectivo va a ser el modelo asignando las nuevas observaciones a las diferentes clases.

Teniendo entonces el score $S(X)$, las observaciones provenientes del grupo P, resultarán en la probabilidad condicional $p(s|P)$ y las del grupo N en la probabilidad condicional $p(s|N)$. Las clasificaciones surgen de comparar los scores con el umbral T. Si se puede encontrar un umbral $T = t$ tal que todas las observaciones del grupo P tengan puntajes mayores a t y todos los clasificados en N tengan puntajes menores al umbral entonces se logrará la clasificación perfecta. Sin embargo esto es casi imposible, sucederá que las observaciones en el grupo P tenderán a tomar valores más altos mientras que las del grupo N tomarán valores más pequeños.

En las tablas de clasificación se cruza el número de observaciones que tenía cada grupo a priori con las predicciones. Para construirlas se necesitan las probabilidades conjuntas $p(s > t, P)$, $p(s > t, N)$, $p(s < t, P)$ y $p(s < t, N)$.

Una de las medidas más utilizadas es el error de clasificación, es decir utilizan como medida la probabilidad de que las observaciones del grupo N tengan puntajes mayores a t o que las observaciones del grupo P tengan puntajes menores a t. Sin embargo, el error de clasificación considera ambos errores con igualdad de importancia, pero el costo de clasificar mal en un grupo no es el mismo que en el otro [Krzanowski y Hand, 2009.].

Resumiendo las cuatro probabilidades conjuntas nombradas anteriormente, se tiene que :

- la probabilidad de que una observación de la clase N produzca un valor mayor a t, $p(s > t|N)$, es llamada falsos positivos y se denotará por f_p.
- la probabilidad de que una observación del grupo P produzca valores mayores a t, $p(s > t|P)$, se le llama verdaderos positivos y se denota por t_p. (*Sensibilidad*)
- la probabilidad de que una observación pertenezca a la clase P es $p(P)$.

También existen otras probabilidades complementarias :

- el verdaderos negativos, $p(s < t|N)$, es la proporción de observaciones del grupo N que son clasificados correctamente, es igual a $1 - f_p$ y se denota por t_n. (*Especificidad*).

- falsos negativos, $p(s < t|P)$, que es la proporción de observaciones del grupo P que son mal clasificados en la clase N, es igual a $1 - t_p$ y se denota por f_n.

- la probabilidad de que una observación pertenezca a la clase N es $p(N) = 1 - p(P)$.

Utilizando las probabilidades antes descritas se puede calcular la tasa de predicción positiva, que es la proporción de observaciones que realmente pertenecen al grupo P sobre el total de observaciones que la regla asignó a dicho grupo, $p(N|s > t)$. La tasa de predicciones negativas es la proporción de observaciones que realmente pertenecen al grupo N en relación a las que la regla clasifica como N, $p(P|s < t)$.

Estas probabilidades pueden calcularse utilizando el teorema de Bayes,

$$P(P|s > t) = \frac{P(s > t|P)p(P)}{P(s > t|P)p(P) + P(s > t|N)p(N)} \tag{2.18}$$

Todas estas medidas se basan en la comparación entre las distribuciones de los scores de uno u otro grupo. Una buena regla, tiende a producir valores altos para las observaciones P y bajos para los valores de N.

La curva ROC es una forma de mostrar en forma conjunta estas dos probabilidades. La interpretación adecuada puede mostrar como es el funcionamiento del modelo, el área debajo de la curva se puede utilizar como medida global de cuán separados están los score de un grupo y otro. Esto no exige elegir un sólo valor para el umbral pero resume los resultados de las posibles opciones, por lo que perimirá elegir cuál es el corte óptimo [Krzanowski y Hand, 2009.].

El gráfico de la curva ROC muestra la tasa de verdaderos positivos en el eje vertical y la tasa de falsos positivos en el eje horizontal cuando el umbral de clasificación t varía en el rango $(0, 1)$. Es una curva que resume la información en una función de distribución acumulada de los puntajes de ambos grupos. Se puede pensar como una completa representación del funcionamiento de la función de clasificación [Krzanowski y Hand, 2009.].

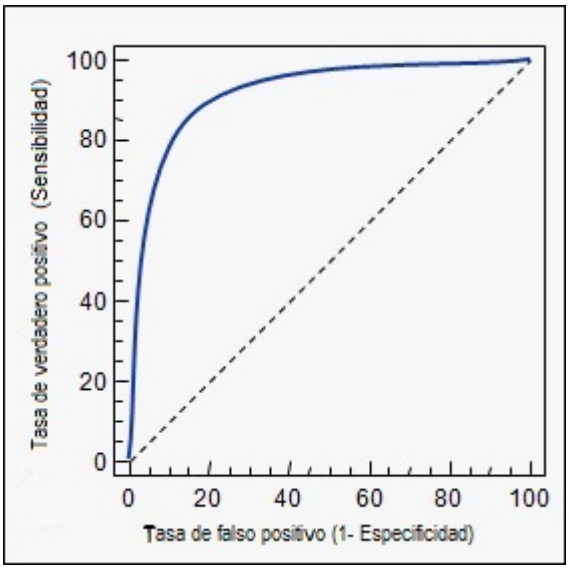

Figura 2.1: Curva ROC

A continuación se describirá la interpretación, y cómo además se pueden obtener otros resultados a partir de ella.

La función de clasificación $S(X)$ es el componente determinante en el análisis, por lo que llamaremos $p(s|P)$ y $p(s|N)$ a las funciones de densidad del score de las observaciones que pertenecen al grupo P y N respectivamente. Sea t un valor del umbral T para una regla de clasificación particular, para evaluar la eficacia del estimador es necesario calcular la probabilidad de hacer una clasificación incorrecta. Dicha probabilidad nos puede dar una idea de cómo las nuevas observaciones van a ser clasificadas.

Dadas las densidades $p(s|P)$ y $p(s|N)$, y el valor t, se pueden obtener los valores de las tasas definidas en la página 31, t_p, f_p, t_n y f_n, para un valor particular de t. Como no siempre es posible determinar el valor de t más adecuado, éste debe ser determinado como parte de la construcción de la función de clasificación. Por lo tanto, variando este valor y con los diferente valores de las tasas se puede obtener la información suficiente para evaluar el desempeño del estimador.

La curva ROC se obtiene entonces, variando t pero utilizando solamente las tasas de falsos y verdaderos positivos, (f_p, t_p), graficándolas sobre ejes ortogonales. En este caso se necesitará las tasas de clasificación sólo para

la proporción de los clientes clasificados por el modelo como $Malo$ y que a priori no lo eran y aquellos que fueron clasificados como $Malo$ siendo que realmente lo eran.

El objetivo de la curva ROC es mostrar el comportamiento del estimador sobre todos los valores posibles de t y no sólo de uno. Se observará cuánto difieren las distribuciones de los scores de $p(s|P)$ y $p(s|N)$. Cuanta mayor diferencia haya, menos solapamiento habrá, por lo que será menos probable que las asignaciones a uno u otro grupo sean incorrectas y por lo tanto más exitosa será nuestra función de clasificación. Por el contrario, cuanto más parecidos son las dos distribuciones, más solapamiento existe entre ellas y por lo tanto más posibilidades existen de que hayan asignaciones incorrectas [Krzanowski y Hand, 2009.].

Considerando los extremos, el resultado menos exitoso sería aquél en el que $p(s|P) = p(s|N) = p(s)$. En este caso clasificar una observación en P es lo mismo que en N cualquiera sea el valor de t. Los valores de las tasas también serán iguales por lo que la curva ROC quedará determinada por la unión de los puntos $(0, 0)$ con $(1, 1)$, es decir la diagonal $x = y$.

En el otro extremo está la separación completa de $p(s|P)$ y $p(s|N)$ en el cual habrá por lo menos un valor de t en el que la asignación a cada grupo es perfecta, en ese caso $t_p = 1$ y $f_p = 0$. Pero como la curva ROC se centra sólo en las probabilidades en que $s > t$ entonces para todos los valores más pequeños de t, $t_p = 1$, mientras que f_p varía de 0 a 1 y para todos los valores más grandes de t debemos tener $f_p = 0$, mientras que t_p varía de 1 a 0. Así que la curva se encontrará a lo largo de los bordes superiores de la gráfica: una línea recta a partir de $(0, 0)$ a $(0, 1)$, seguido por una línea recta a partir de $(0, 1)$ a $(1, 1)$ [Krzanowski y Hand, 2009.].

En la práctica esto no sucede, se consiguen curvas situadas en el extremo superior de la gráfica. Cuanto más cercana esté al extremo superior izquierdo más cerca se estará de la situación de completa separación y por lo tanto mejor será el desempeño de la función de clasificación.

Estimación de la curva ROC

Para estimar la curva ROC en el caso que se esté trabajando con una función de score S continua, se utiliza [Krzanowski y Hand, 2009.]:

$$y = 1 - G[F^{-1}(1-x)], \quad (0 \leq x \leq 1) \tag{2.19}$$

donde f es la función de densidad y F la función de distribución de S en el grupo N, y g, G las funciones de S para el grupo P. El problema está en la estimación de la curva a partir de los datos.

Para obtener el estimador empírico se aplicarán las definiciones dadas en la página 31, t_p, f_p, t_n, f_n. Si n_p y n_N es el número de individuos del grupo P y N respectivamente, $n_{PP(t)}$ denota el número de individuos en la muestra de la población P cuyos scores son mayores que t y $n_{NP(t)}$ denota el número de individuos en la muestra de la población N cuyos scores son mayores que t, entonces el estimador empírico para la tasa de verdaderos positivos, $t_p = p(S > t|P)$, y tasa de falsos positivos, $f_p = p(S > t|N)$, para el umbral t es [Krzanowski y Hand, 2009.]:

$$\hat{t_p} = \frac{n_{PP_{(t)}}}{n_P} \tag{2.20}$$

y

$$\hat{f}_p = \frac{n_{NP_{(t)}}}{n_N} \tag{2.21}$$

Por lo tanto el trazado de los valores $1 - \hat{f}_P$ contra t nos lleva a la distribución empírica de $\hat{F}(t)$, y de la misma manera $1 - \hat{t_p}$ nos lleva a la distribución empírica $\hat{G}(t)$.

La curva ROC está dada simplemente por el gráfico de $(\hat{f}_p, \hat{t_p})$ obtenidas de variar t, por lo que la curva está dada por [Krzanowski y Hand, 2009.]:

$$y = 1 - \hat{G}[\hat{F^{-1}}(1-x)], \quad (0 \leq x \leq 1). \tag{2.22}$$

Aunque técnicamente se deben considerar todos los valores posibles de t, en la práctica $\hat{f}_p$ cambiará solamente cuando t cruza el valor del score de las n_N observaciones y $\hat{t}_P$ sólo va a cambiar cuando t cruce el valor del score de los n_p individuos, por lo que habrá como mucho $n_N + n_P + 1$ puntos en el gráfico. Los puntos son unidos por líneas que producen un aspecto irregular

ya que el cambio en la dirección está dado por el cambio en $\hat{f}_p$ o $\hat{t}_p$.

Como se dijo anteriormente, la curva ROC provee una descripción de la separación de la distribución de la función de clasificación S en los dos grupos, y la línea que une los puntos (0,0) y (1,1) es aquella en donde la probabilidad de clasificar a un individuo en el grupo P es igual a la de clasificarla en N. Por lo que, para medir la diferencia en el score de diferentes poblaciones se necesita medir la diferencia entre la curva ROC y la diagonal. Una forma de cuantificarlo es medir directamente la mayor separación entre la curva y la diagonal, y la otra es utilizando la diferencia entre el área de las curvas.

Medidas de la curva ROC.

Existen ciertas medidas que se obtienen a través de la curva ROC, que de manera complementaria capturan y resumen la esencia de los datos.

- Uno de ellos es el área debajo de la curva, comúnmente denotado AUC [Krzanowski y Hand, 2009.].

$$AUC = \int_0^1 y(x)dx \tag{2.23}$$

  El AUC es la verdadera tasa positiva promedio, tomada de manera uniforme sobre todas las posibles tasas de falsos positivos en el rango (0, 1). Es decir, el área debajo de la curva proporciona una medida de la habilidad del modelo para discriminar entre las observaciones que presentan el suceso de interés.

  Una interpretación menos obvia pero usada frecuentemente es que sea la probabilidad de que el clasificador asigne una puntuación más alta a un individuo de la población P elegida al azar de lo que le asignará a un individuo de la población N elegido al azar y de manera independiente.

Una regla empírica sobe el AUC [Blanco, 2006] establece que:

- Si es menor a $0,7$ no es bueno el modelo.
- Si está entre $0,7$ y $0,8$ su ajuste y poder predictivo son aceptables.
- Si está entre $0,8$ y $0,9$ su ajuste y poder predictivo son muy buenos.
- Un valor mayor a $0,9$ es poco probable que suceda.

En algunos casos podría llegar a utilizarse este valor para comparar dos funciones de clasificación pero se debe tener precaución ya que existe la posibilidad de que las curvas se crucen.

- Otra medida de resumen es el índice que mide la máxima distancia vertical [Krzanowski y Hand, 2009.], MVD, entre la diagonal y la curva ROC,

$$MVD = \text{máx}\,|y(x) - x| \tag{2.24}$$

Utilizando las funciones en términos de la variación de t y las probabilidades definidas anteriormente, se obtiene:

$$MVD = \max_t |p(S > t|P) - p(S > t|N)| = \max_t |t_p - f_p| \tag{2.25}$$

Por lo tanto MVD es la máxima distancia, en un rango de 0 a 1, entre la distribución acumulada de S en P y N. Este índice es equivalente al estadístico de *Kolmogorov-Smirnov*.

- Estadístico de Kolmogorov-Smirnov [Krzanowski y Hand, 2009.]

El indicador MVD, es una medida simple para medir la diferencia entre la curva ROC y la diagonal, es la máxima distancia vertical. Este indicador mide por lo tanto hasta que punto se desvía de la "aleatoriedad", y varía desde 0, para una curva poco informativa, a 1 para un discriminador perfecto.

La ecuación de la curva ROC es [Krzanowski y Hand, 2009.]:

$$Y = 1 - G[F^{-1}(1 - x)] \;\; (0 \leq x \leq 1) \tag{2.26}$$

Siendo F la distribución de la función de clasificación S en el grupo N y G la distribución de la función de clasificación en el grupo P.

Tomando $x = y$ como la ecuación de la diagonal y escribiendo [Krzanowski y Hand, 2009.]:

$$MVD = \max_x |1 - G[F^{-1}(1-x)] - x| \quad (2.27)$$

$$= \max_x |(1-x) - G[F^{-1}(1-x)]| \quad (2.28)$$

Si $t = F^{-1}(1-x)$ entonces $(1-x) = F(t)$, y como $F(.)$ es una función de distribución, entonces el rango de t es $\mathbf{R}$. Sustituyendo en la ecuación anterior obtenemos [Krzanowski y Hand, 2009.],

$$MVD = \max_t |F(t) - G(t)| = \max_{t \in \mathbf{R}} |F(t) - G(t)| \quad (2.29)$$

Utilizando los datos de la muestra se tiene,

$$Y = 1 - \hat{G}[\hat{F}^{-1}(1-x)] \ \ (0 <= x <= 1) \quad (2.30)$$

El estimador de la máxima distancia vertical $M\hat{V}D$ entre la curva y la diagonal es [Krzanowski y Hand, 2009.],

$$M\hat{V}D = \max_{t \in (-\infty, \infty)} |\hat{F}(t) - \hat{G}(t)| \quad (2.31)$$

Este indicador $M\hat{V}D$ es conocido como el estadístico de *Kolmogorov-Smirnov* (K-S) utilizado para testear la igualdad de dos distribuciones de probabilidad F y G [Krzanowski y Hand, 2009.].

Elección del umbral óptimo

La curva ROC muestra el valor de la función de clasificación a través de todas los posibles valores del umbral, pero si debe ser utilizado luego para clasificar nuevas observaciones es necesario establecer un único valor de t [Krzanowski y Hand, 2009.].

Si los costos de clasificar en uno u otro grupo son diferentes entonces se debe proceder de tal forma de minimizar los costos esperados de la clasificación errónea. Sin embargo, si no se tiene información al respecto o no difieren

entonces es necesario adoptar algún procedimiento para la determinación del umbral óptimo.

Uno de los procedimientos utilizados es el de localizar en la curva ROC el punto más cercano a la esquina superior izquierda y utilizar dicho valor de t. Sin embargo, esto no se ha verificado y, además, algunos autores han advertido que este procedimiento puede llegar a introducir una mayor tasa de error de clasificación [Krzanowski y Hand, 2009.].

Un criterio a considerar es el Índice de Youden,

$$YI = \max_t |t_p - f_p| = \max_t |t_p - (1 - t_n)|, \tag{2.32}$$

este criterio es utilizado con el fin de determinar un umbral óptimo para el uso de un solo clasificador. De hecho, en [Fluss et al., 2005] se señala que:

$$\begin{aligned} YI &= \max_t |t_p - f_p| = \max_t |t_p + t_n - 1| && (2.33)\\ &= \max_t |F(t) - G(t)| && (2.34)\\ &= K\text{-}S, && (2.35) \end{aligned}$$

de modo que YI se puede estimar utilizando cualquiera de los estimadores para los F y G ya descritos. Por lo que el umbral óptimo t^* sería el valor de t que maximiza $F(t)-G(t)$, que es el valor que maximiza el estadístico K-S.

Se tienen dos grupos, los que fueron calificados como $Malo$ (P) y los que fueron clasificados como $Bueno$ (N). Se tiene la matriz X con los datos de cada cliente y la función de calificación del comportamiento del cliente $S(X)$. Esta función, a través de la regresión logística, convierte la matriz de valores en una sola puntuación de tal manera que la solicitud sea rechazada si la puntuación supera un determinado umbral t y aceptada si sucede lo contrario.

La convención habitual, para este tipo de procedimientos, es que una alta puntuación implica que el cliente tiene un perfil más moroso.

Entonces, la principal tarea es determinar un valor umbral t adecuado para el modelo elegido. En [Blöchlinger y Leippold, 2006] se señala que en general se trata de una elección arbitraria, en base a argumentos cualitativos tales como restricciones comerciales, y por lo general será subóptima. También argumentan que un criterio más riguroso puede derivarse de conocimiento de la probabilidad a priori de forma predeterminada junto con los costos y los ingresos asociados. En otros casos cuando no se penalizan los

costos se considera entonces el umbral donde el MVD se maximiza, el K-S.

En primera instancia, como medida de calidad de ajuste, se utilizarán la representación gráfica de la curva ROC y el estadístico *Kolmogorov-Smirnov* (K-S).

En la curva ROC se representan los resultados para diferentes puntos de corte teniendo en cuenta el estadístico K-S, cuyo máximo genera un umbral óptimo, según dicho criterio.

Para cada punto de corte t se deberá calcular la llamada matriz de confusión, indicando la cantidad de individuos que fueron clasificados en cada grupo teniendo en cuenta el grupo al que pertenecían a priori.

	Estimación		
	Bueno (N)	Malo (P)	Total
Bueno (N)	$n_{NN(t)}$	$n_{NP(t)}$	n_N
Malo (P)	$n_{PN(t)}$	$n_{PP(t)}$	n_P
	$n_{N(t)}$	$n_{P(t)}$	n

Cuadro 2.1: Matriz de confusión.

Según van variando los puntos de corte, se van obteniendo las tasas de falsos y verdaderos positivos y los puntos que conforman la curva, que a su vez permite identificar el corte que maximiza el estadístico K-S.

En términos de la matriz anterior, la sensibilidad es el cociente entre los éxitos observados clasificados como éxitos y el total de éxitos observados, es decir $\frac{n_{PP(t)}}{n_P}$.

La especificidad se define como el cociente entre los fracasos observados clasificados como fracasos y el total de los fracasos observados, $\frac{n_{NN(t)}}{n_N}$. En muchos casos la curva ROC se realiza utilizando estos datos, se grafica la *Sensibilidad* contra *1 – Especificidad*.

Para calcular el K-S se debe obtener para cada t la diferencia entre la tasas de verdaderos positivos y la tasa de falsos positivos y el punto de corte óptimo será aquel cuya diferencia sea máxima [Krzanowski y Hand, 2009.].

$$M\hat{V}D = \max_t |\hat{t_p} - \hat{f_p}| = \max_t |\frac{n_{PP_{(t)}}}{n_P} - \frac{n_{NP_{(t)}}}{n_N}| \tag{2.36}$$

Mientras que el AUC es únicamente un medida global de la calidad del modelo, el estadístico K-S además de medir la calidad de ajuste cuando el valor es máximo, permite identificar el punto de corte "óptimo".

2.3.3. Árboles de regresión y clasificación - CART -

Originariamente fueron propuestos para separar las observaciones que componen la muestra asignándolas a grupos establecidos a priori, de forma que se minimizara el costo esperado de los errores cometidos.

Esta técnica fue presentada por Friedman en 1977, pero originariamente sus aplicaciones a las finanzas no fueron muy numerosas, si bien corresponde destacar dos estudios pioneros: *Friedman y otros* [Altman et al., 1985] en el que utilizan el modelo para clasificar empresas, comparando su capacidad predictiva con el Análisis Discriminante, y *Marais y otros* [Marais et al., 1984] que, por el contrario, lo aplican a préstamos bancarios. En ambos trabajos se ha llegado a demostrar la gran potencia que presenta este algoritmo como técnica de clasificación.

Un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto (numeroso) de ejemplares. Esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada nodo hoja se refiere a una decisión (clasificación).

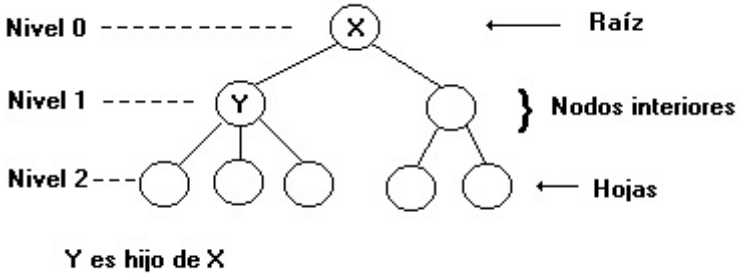

Figura 2.2: Árboles de regresión y clasificación.

La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezado por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja (o nodo hijo). La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.

Los métodos basados en los árboles son simples y útiles para la interpretación. Muchas veces esta técnica va acompañada de procedimientos de agregación. Cada uno de estos enfoques implica producir múltiples árboles que después se combinan para producir un sólo consenso de predicción. Y en varios casos la combinación de un gran número de árboles a menudo puede resultar en grandes mejoras en la precisión de la predicción, a expensas de alguna pérdida en la interpretación.

Entre los clasificadores basados en árboles descritos en la literatura (ID3, C4, C4.5, Árboles Bayesianos, etc.) se estudiará CART, acrónimo de Classification And Regression Trees o Árboles de Clasificación y Regresión, propuesto por Breiman [Breiman, 1994]. Las diferencias principales entre los distintos algoritmos de construcción de los árboles de decisión radican en las estrategias de poda y en la regla adoptada para particionar nodos. Así, CART se caracteriza, fundamentalmente, por realizar particiones binarias y por utilizar una estrategia de poda basada en el criterio de costo-complejidad.

Los árboles de decisión se pueden aplicar tanto a problemas de regresión como de clasificación.

Dado un conjunto de datos de entrenamiento $L(X, Y)$, donde Y es la variable a explicar y $X = (X_1, ..., X_k)$ es un conjunto de k características

que describe a los individuos, el objetivo de CART es predecir los valores de Y a partir de los valores observados de las variables X. Tanto la variable dependiente Y, como cada una de las variables explicativas X_i puede ser cuantitativa o cualitativa, esto hace de CART una técnica de gran flexibilidad pues se puede aplicar en muchos contextos distintos.

En el caso en que la variable dependiente sea cualitativa se dice que CART es un árbol de clasificación y lo que se busca es clasificar a los individuos objeto de estudio en alguno de los grupos predeterminados usando k características $(X_1, ..., X_k)$. Por otro lado si Y es una variable continua entonces CART es llamado árbol de regresión y su objetivo es obtener una estimación del valor de Y.

2.3.3.1. Árboles de Clasificación

Para un árbol de clasificación, a diferencia de un árbol de regresión, en el cual, la respuesta pronosticada para una observación es dada por la respuesta media de las observaciones de entrenamiento que pertenecen al mismo nodo terminal.

En la interpretación de los resultados de un árbol de clasificación, se está a menudo interesado no sólo en la predicción de la clase correspondiente para un nodo terminal en particular, sino también en las proporciones de los grupos que caen en esa región.

Reglas de división y criterio de mejor división.

Cada partición tiene asociada una medida de impureza, de forma genérica $i(t)$ es la medida de impureza del nodo t. Y se tratará de incrementar la homogeneidad de los subconjuntos resultantes de la partición, esto es, que sean más puros que el conjunto original.

Entonces, ¿cómo medir si un nodo es puro o impuro? Pueden utilizarse distintos criterios como: error de clasificación, índice de Gini y entropía.

Sea $j = 1, \ldots, k$ siendo k el número de clases de la variable dependiente, definiendo $p(j|t)$ como la distribución de probabilidad de la clase de la variable dependiente para el nodo t (la probabilidad de pertenecer a la clase j estando en el nodo t), entonces $p(1|t) + p(2|t) + p(3|t) + \ldots + p(k|t) = 1$.

Criterios de impureza

Para la creación de un árbol de clasificación utilizamos una división binaria recursiva. Sin embargo, en el ajuste de la clasificación la suma de los cuadrados de los residuos (RSS) no puede ser utilizado como criterio para la clasificación de las particiones binarias, como se hace para los árboles de regresión. Una alternativa al RSS es la tasa de error de clasificación, ésta es simplemente la fracción de las observaciones de formación en esa región que lo hacen pertenecer a la clase más común:

$$E = 1 - \max_k(\widehat{p}_{km}) \tag{2.37}$$

Donde, $\widehat{p}_{km}$ representa la proporción de observaciones en la m-ésima región de la clase k. Sin embargo, resulta que el error de clasificación no es suficientemente sensible para la elaboración de árboles, y en la práctica otras dos medidas son preferibles.

Una de ellas, la medida de impureza de Gini para un nodo t, es definida como $i(t) = 1 - S$, donde S (la función de impureza) es:

$$S = \sum_j p^2(j|t), \tag{2.38}$$

para $j = 1, 2, \ldots, k$. [Hastie et al., 2009]

La función de impureza alcanza el máximo si cada clase en la población se encuentra con igual probabilidad. Esto es, $p(1|t) = p(2|t) = p(3|t) = \ldots = p(k|t)$ para $j = 1, 2, \ldots, k$. Sin embargo, la función de impureza alcanza éste máximo si todos los casos del nodo pertenecen a una sola clase. Esto es, si un nodo t es puro con una tasa de error de clasificación igual a cero, $i(t) = 0$. Un nodo de valor de $i(t)$ pequeño indica que contiene predominantemente observaciones de una sola clase.

Una alternativa para el *Índice de Gini* es la *Entropía Cruzada*, dada por,

$$D = -\sum_{k=1}^{K} \widehat{p}_{km} log(\widehat{p}_{km}) \tag{2.39}$$

donde $0 \leq \widehat{p}_{km} \leq 1$, se deduce que $0 \leq -\widehat{p}_{km} log(\widehat{p}_{km})$. Se puede demostrar que la *Entropía Cruzada* presentará a un valor cercano a cero si $\widehat{p}_{km}$ están todos cerca de cero o cerca de uno. Por lo tanto, al igual que el *Índice de Gini*, en la *entropía cruzada* se dará un valor pequeño si el nodo m-ésimo es puro. De hecho, resulta que el coeficiente del *Índice de Gini* y la *Entropía Cruzada* son bastante similares numéricamente.

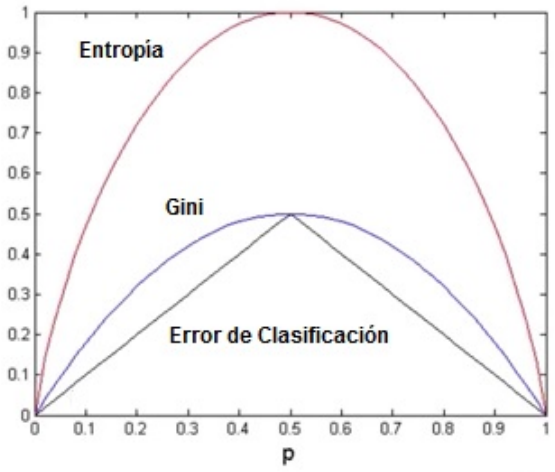

Figura 2.3: Ajustes de clasificación, CART.

Cuando se construye un árbol de clasificación, ya sea el *Índice de Gini* o la *Entropía Cruzada* se utilizan normalmente para evaluar la calidad de una división en particular, ya que estos dos enfoques son más sensibles a la pureza del nodo que la tasa de error de clasificación. Cualquiera de estos tres enfoques podría utilizarse cuando se poda el árbol, pero la tasa de error de clasificación es preferible cuando el objetivo es la precisión de la predicción del árbol final podado.

Poda de Árboles.

Los árboles grandes pueden tener dos problemas:

1. Sobre-ajuste: aunque son de gran precisión, con errores bajos o nulos, proporcionan resultados pobres cuando se aplica a nuevos conjuntos de datos.

2. Complejidad: la comprensión e interpretación de los árboles con un gran número de nodos terminales es un proceso complicado. La complejidad de un árbol se mide por el número de sus nodos terminales.

La situación ideal de un error de clasificación bajo o nulo implica un compromiso entre la precisión y la complejidad del árbol. La relación entre la complejidad y precisión del árbol puede ser entendido con la medida de costo de complejidad asociada al árbol T, $R_\beta(T)$ que se define como:

$$R_\beta(T) = R(T) + \beta * O \tag{2.40}$$

- R(T) es el error de clasificación asociado al árbol T.

- β, ($\beta \geq 0$, parámetros de complejidad) se interpreta como el costo de complejidad por nodo terminal.

- O es el número de nodos terminales.

  Si $\beta = 0$, el costo de complejidad alcanza su máximo para el árbol más largo posible. Cuando los valores de β decrecen y se aproximan a cero, los árboles minimizan el costo de complejidad.

Criterio de mejor división

Sea s una división del nodo t, la mejor división s es definida como la disminución de la medida de impureza:

$$\Delta_i(s,t) = i_t - p_L[i(t_L)] - p_R[i(t_R)] \tag{2.41}$$

Donde,

- s = división particular
- p_L=la proporción de casos del nodo t que van en el nodo hijo izquierdo, t_L.
- p_R=la proporción de casos del nodo t que van en el nodo hijo derecho, t_R.
- $i(t_L)$ = impureza del nodo hijo izquierdo.
- $i(t_R)$ = impureza del nodo hijo derecho.

Regla de asignación de clases

Hay dos reglas de asignación de clases para los nodos.

1 La regla de mayoría relativa (The pluralty rule), asigna el nodo terminal t a la clase con mayor $p(j|t)$. Si la mayoría de los casos en un nodo terminal pertenecen a una clase específica, el nodo es asignado a esa clase. La regla asume mismo costo de error de clasificación para cada clase. Esto no toma en cuenta la gravedad del costo de cometer un error (caso particular de la segunda).

2 Una segunda regla asigna el nodo terminal t a la clase con el mínimo costo de error de clasificación esperado. La aplicación de esta norma tiene en cuenta la gravedad de los costos de error de clasificación de casos u observaciones en una cierta clase, e incorpora la variabilidad del costo en la regla de partición de Gini.

Sea $c(i|j)$ el costo de clasificar una clase j como una clase i:

$c(i|j) \geq 0$ si $i \neq j$, $c(i|j) = 0$ si $i = j$

Asumiendo un problema con dos clases, se tiene:

$\pi_t(1)$ = probabilidad a priori de la clase 1 en el nodo t

$\pi_t(2)$ = probabilidad a priori de la clase 2 en el nodo t

$r_1(t)$ = el costo de asignar el nodo t para la clase 1

$r_2(t)$ = el costo de asignar el nodo t para la clase 2

Dadas las distribuciones a priori y el costo variable del error de clasificación, $r_1(t)$ y $r_2(t)$ son estimados como:

$$r_1(t) = \pi(1)c(2|1) \quad y \quad r_2(t) = \pi(2)c(1|2) \tag{2.42}$$

De acuerdo con la regla 2, si en el nodo t, $r_1(t) < r_2(t)$, el nodo t es asignado a la clase 1. Si $c(1|2) = c(2|1)$, entonces aplicando la regla 1, el nodo es asignado a la clase donde la probabilidad a priori es la mayor.

Pasos para la construcción del Árbol de Clasificación.

El proceso de construcción de los árboles comienza dividiendo una muestra o el nodo raíz en nodos binarios basado en la pregunta de si $x \leq d$. Donde x es una variable del conjunto de datos y d es una constante.

Inicialmente todas las observaciones son colocadas en el nodo raíz. Este nodo es impuro o heterogéneo porque contiene observaciones de diferentes clases. El objetivo es diseñar una regla que divida estas observaciones y cree grupos o nodos binarios que sean internamente más homogéneos que el nodo raíz. Se utilizan algoritmos iterativos computacionales que buscan la mejor partición dentro de todas las posibles para cada variable.

La metodología que se utiliza para la creación de árboles técnicamente se conoce como partición recursiva binaria [Hastie et al., 2009]. Comienza del nodo raíz y usando, por ejemplo, el *Índice de Gini* como regla de partición, el proceso es el siguiente:

1. Se divide la primera variable en todos sus posibles puntos de división (en todos los valores que la variable asume en la muestra). En cada posible punto de partición de la variable, la muestra se divide en nodos secundarios binarios o dos nodos hijos. Los casos con "sí" como respuesta a la pregunta formulada, se envían al nodo izquierdo y aquellos con respuestas "no", se envían al nodo derecho.

2. Luego, aplica sus criterios de "bondad de división" para cada punto y evalúa la reducción de la impureza que se logra mediante la fórmula:

$$\Delta_i(s,t) = i(t) - p_L[i(t_L)] - p_K[i(t_K)] \tag{2.43}$$

   como fue descrito más arriba.

3. Selecciona la mejor división de la variable como aquella donde la reducción de la impureza es la mayor.

4. Los pasos 1, 2 y 3 son repetidos para cada una de las variables del nodo raíz.

5. Luego clasifica todas las mejores divisiones de cada una de las variables acorde con la reducción de la impureza alcanzada por cada división.

6. Selecciona la variable y su punto de división que más reduce la impureza del nodo raíz o padre.

7. Asigna clases a estos nodos de acuerdo a la regla que minimiza el costo de error de clasificación.

8. Debido a que el procedimiento es recursivo, los pasos 1 a 7 se aplican varias veces para cada nodo hijo no terminal en cada etapa sucesiva.

9. Continúa el proceso de división y se construyen árboles largos. El árbol más largo es construido si el proceso de división continúa hasta que todas las observaciones constituyan un nodo terminal.

Outliers

Los outliers de las variables independientes raramente afectan el análisis de CART, porque las divisiones generalmente son determinadas por los valores que no son atípicos. Si existen valores atípicos en la variable dependiente, están aislados en pequeños nodos, en los que no afectan al resto del árbol [Hastie et al., 2009].

Ventajas y desventajas de los árboles de clasificación

Ventajas:

- Los árboles se pueden visualizar gráficamente y son muy fáciles de explicar a cualquier tipo de persona (dentro y fuera del área estadística). De hecho, son incluso más fácil de explicar que la regresión lineal.

- Los árboles pueden manejar fácilmente predictores cualitativos sin la necesidad de crear variables ficticias.

Desventajas:

- Inestabilidad: desafortunadamente, los árboles generalmente no tienen el mismo nivel de predicción y exactitud como algunos de los otros métodos de regresión y clasificación. Sin embargo, mediante la agregación de muchos árboles de decisión el rendimiento predictivo de árboles puede ser mejorado sustancialmente.

Capítulo 3

Aplicación

3.1. Resumen del procedimiento a realizar

Considerando que la base de datos fue debidamente depurada, finalmente sólo se incluyen aquellas observaciones que se consideran pertinentes estudiar por su historial y condiciones de operación en la empresa.

Estas observaciones se clasificaron en *Bueno* o $Malo$, variable de respuesta, la que luego se va a querer predecir. Es decir, en este caso se cuentan con dos grupos, los clasificados como *Bueno* y los clasificados como $Malo$.

De cada una de estas observaciones se cuenta con información de 35 variables en relación a los datos personales del cliente, del crédito otorgado y del comportamiento frente a este. Se realiza el análisis estadístico de cada una de ellas. Se discrimina por la variable de referencia, para investigar cuál era la relación con la variable de interés según el comportamiento del cliente ya sea $Malo$ o *Bueno*.

Teniendo entonces la base de datos y los grupos que se tienen a priori, se procede a la estimación de los modelos. El programa elegido para realizarla es el R-project [R Core Team, 2014].

Mediante este software es posible considerar diferentes modelos de tal forma de poder elegir no sólo cuáles variables se incluirán en el modelo sino cuál es el modelo que provee un mejor ajuste a los datos.

Se estiman los coeficientes de regresión, los $\hat{\beta}'s$. Se testea su significación y la significación de los modelos, para comprobar que el modelo prediga de

la mejor forma posible. También se verifica que las predicciones sean aceptables, con los datos que se utilizaron para estimar el modelo y con aquellos que se reservaron para utilizarlos como control. Se estudian los errores de clasificación, la curva ROC, el área debajo de la curva, el estadístico *K-S*, la sensibilidad y la especificidad del modelo.

Una vez que se elige el modelo que se considera que provee un "mejor" ajuste con la estimación de los $\beta's$, se interpreta el significado de cada una de ellas en relación al incremento o no de las probabilidades de ser *Bueno* o *Malo*. Finalmente se determina cual va a ser el modelo que se utilizará para predecir los nuevos casos, modelo que se incorporará en el sistema de la empresa.

Más adelante se detallará con mayor precisión las estimaciones realizadas.

3.2. Consideraciones Generales

La población objetivo es toda persona física que haya solicitado un crédito al consumo en la financiera y cuyo crédito fue aprobado por los analistas, durante el período transcurrido entre el segundo semestre de 2011 y el primer semestre de 2014.

De este modo, se puede obtener la información del comportamiento del cliente durante el transcurso del Crédito, dato que interesa para clasificar a cada uno como *Bueno* o *Malo* según su comportamiento en los pagos.

Se utilizan tres años de contratación para obtener una muestra de mayor tamaño, ya que la muestra se reduce en el proceso de limpieza.

Se crean las variables *Cociente cuotas pagas/cuotas totales*, *Valor cuota/total ingreso líquido*, *Cantidad de veces que operó*, *Total de ingresos* y *Contactabilidad*. (Ver Anexo B)

Las observaciones de la base se las denomina Instancias, estas son un número identificador de cada acción de un cliente en la empresa. Es decir, un mismo cliente puede tener varias instancias, tantas como préstamos tenga.

Para la estimación del modelo no se consideran aquellas observaciones clasificadas como $Indiferentes$ pues se desean perfiles más marcados de comportamiento.

Se quitan 23 instancias identificados como fraudulentas; por último se quitan aquellas instancias que habían tenido sólo una operación como cliente en la empresa solicitado hace menos de 12 meses. Esto último se realiza para obtener una tabla de datos con cierta "historia" crediticia, para poder analizar mejor su comportamiento.

Las consideraciones anteriores se realizan a pedido de la empresa, conociendo éstos el comportamiento de sus clientes.

3.3. Análisis de las Variables

A continuación se procede a realizar un análisis de aquellas variables que podrían llegar a incluirse en el modelo, el estudio de las otras variables se encuentra en el Anexo B.

1. **Antigüedad Laboral**

   Antigüedad laboral que tiene una persona, variable cuantitativa, expresada en meses.

   Se decide trabajar con rangos de meses laborales quedando cuatro grupos: menos de 24 meses, entre 25-48, de 48-60 meses y por último mayor a 60 con los Jubilados y Pensionistas. Debido a que las personas jubiladas algunas veces registraban números muy altos en esta variable o no registran dato y a su vez presentan un comportamiento similar a los mayores a 60 meses, se decide agruparlos de esa forma.

Antigüedad laboral	**Frecuencia Relativa** (%)	**Frecuencia Relativa** *Bueno* (%)	**Frecuencia Relativa** *Malo* (%)
<=24	9	8	21
25-48	9	8	16
49-60	3	3	5
>60, Jubilado, Pensionista	79	81	58

Cuadro 3.1: Frecuencia relativa de la variable Antigüedad Laboral.

Esta variable parece ser muy importante, ya que al diferenciar según *Bueno* o *Malo*, se observan grandes cambios. Dentro de los clientes malos se aprecia un aumento considerable de la proporción de aquellos con menor antigüedad laboral.

2. **Cantidad de veces que operó.**

   Variable cuantitativa que hace referencia a la cantidad de veces que el cliente ha operado en la empresa entre el 2011 y 2014, es decir la cantidad de veces que se le otorgó un crédito. Aquellos clientes nuevos van a tomar el valor cero.

Cantidad de de veces que operó	**Frecuencia Relativa** (%)	**Frecuencia Relativa** *Bueno* (%)	**Frecuencia Relativa** *Malo* (%)
0	0,4	0,4	1,3
1	96	96,9	97,7
2	1,1	1,2	0,5
3	0,9	1,0	0,2
4	0,6	0,6	0,2
⋮	⋮	⋮	
27	0,0	0,0	0,0

Cuadro 3.2: Frecuencia relativa de la variable Cantidad de veces que operó.

Se observa que la cantidad de veces promedio que ya ha operado una persona en la empresa es una, siendo éste un 96 % de todos los casos. A su vez se puede apreciar que si operan una vez es muy probable que lo vuelvan a hacer una vez más.

Si bien el valor de la media es la misma para ambos grupos, el porcentaje (dentro de cada grupo) de personas calificadas como malos pagadores que hasta la fecha no habían operado, es casi el triple que los buenos.

3. **Clearing**

   Variable creada en base a las reglas de medición de la información que provee el Bureau de crédito "Clearing de Informes" y el seguimiento comportamental de su cartera.

   Las codificaciones son las siguientes:

Nombre	**Descripción**
ROJO	Tiene incumplimientos vigentes diferentes a incumplimientos de intendencias municipales o tienen incumplimientos de intendencias municipales anteriores al 30/06/2011 o cheques devueltos por falta de fondos o refinanciación atrasada o deuda actualizada atrasada. Tienen cuenta clausurada en los últimos veinticuatro meses Tiene más de tres cancelaciones con atrasos con empresas distintas No tiene consultas en financieras con más de nueve meses de antigüedad, sí tiene consultas financieras en los últimos nueve meses y tienen un atraso cancelado No tiene consultas en financieras con más de nueve meses de antigüedad y tiene más de tres consultas en financieras en los últimos nueve meses, ignorando consultas realizadas por la misma empresa el mismo mes.

AMARILLO A2	Tiene más de seis consultas con sector financiero en los últimos doce meses sin contar e ignorando consultas realizadas por misma empresa en mismo mes
AMARILLO M	Tiene dos o más cancelaciones con atraso Tiene incumplimientos de intendencias municipales posteriores al 01/07/2011 Tiene cheques devueltos por falta de fondos cancelados Tiene refinanciación al día o deuda actualizada al día Tiene sólo una cancelación y está dentro de los últimos seis meses Tiene más de una consulta en financieras en los últimos tres meses e ignorando consultas realizadas por misma empresa el mismo mes No tiene consultas en financieras con más de nueve meses de antigüedad y si tiene consultas financieras en los últimos nueve meses
AMARILLO A3	Tienen una cancelación con atraso (por filtros anteriores ésta será anterior a los seis meses)
AMARILLO A1	No tiene consultas (no se consideran las consultas de la empresa) No tiene consultas con financieras y no tiene consultas con la empresa (si no tiene consultas en financieras pero existe al menos una con nosotros ya no es A1) Sin antecedentes en Clearing de informes
2 VERDE y LC	Si no cumple ninguna de las condiciones anteriores o tiene Línea de Crédito. La Línea de crédito es un cupo contingente de capital a riesgo pre-definido y establecido por la compañía para que un beneficiario (cliente), lo pueda utilizar a su discreción para el servicio de crédito, en las condiciones establecidas por la entidad.

Cuadro 3.3: Codificación de la variable Clearing.

Clearing	**Frecuencia Relativa** (%)	**Frecuencia Relativa** *Bueno* (%)	**Frecuencia Relativa** *Malo* (%)
AMARILLO_A1	1	1	1
AMARILLO_A2 y ROJO	1	1	5
AMARILLO_A3	3	3	6
AMARILLO_M	9	8	20
VERDE y LC	86	87	68

Cuadro 3.4: Frecuencia relativa de la variable Clearing.

En más del 80 % de los créditos otorgados los clientes tenían una calificación *Verde o LC* en el Clearing, dicha calificación fue realizada por la empresa teniendo en cuenta los datos extraídos del Clearing de Informes.

Como es de esperar, al discriminar según *Bueno* o *Malo*, se observan grandes diferencias. Ahora, dentro del grupo de los malos pagadores, el porcentaje con calificación *Verde* o con *LC* según la variable Clearing, disminuye considerablemente aumentando los calificados con *Amarillo M*. Dentro de la categoría *Bueno*, en proporción, el comportamiento en el Clearing es muy similar.

4. **Contactabilidad**

Esta variable fue creada en base a los datos que se tenían de los teléfonos de contacto que fueron brindados por el cliente, y permite resumir la cantidad de teléfonos y/o celulares que brinda el cliente.

Código	**Teléfono Fijo**	**Teléfono Alt.**	**Celular**	**Teléfono laboral**	**Total**
1	si	si	si	si	4
2	si	si		si	3
3	si	si	si		3
4	si		si	si	3
5	si		si		2
6	si	si			2
7	si			si	2
8	si				1

Cuadro 3.5: Codificación de la variable Contactabilidad.

Contactabilidad	**Frecuencia Relativa** *Bueno* (%)	**Frecuencia Relativa** *Malo* (%)
1	2	3
2	3	2
4	59	69
5	24	17
6	9	6
8	3	3

Cuadro 3.6: Frecuencia relativa de la variable Contactabilidad según la categoría *Bueno* y *Malo*.

Se observan algunas diferencias en los porcentajes por categoría de contactabilidad, pero no parece ser una estructura muy clara, posiblemente no aporte información clara para el modelo.

5. **Cuotas totales**

   Cantidad de cuotas con que el cliente solicita el crédito.

   Si bien en la base de datos esta variable no es categórica, a modo de visualizar mejor los resultados, se muestra a continuación la frecuencia agrupada en distintos rangos.

Cuotas Totales	**Frecuencia Relativa** (%)	**Frecuencia Relativa** *Bueno* (%)	**Frecuencia Relativa** *Malo* (%)
$\leq$6	7	7	3
7-12	54	55	43
13-18	29	28	37
19-24	10	9	17
>24	0	1	0

Cuadro 3.7: Frecuencia relativa de la variable Cuotas Totales.

Por lo que se puede observar parece haber cierta relación entre mayor cantidad de cuotas y la categoría *Malo*. En proporción casi el doble solicitaron el préstamo en más de 18 cuotas comparando con los calificados como buenos pagadores.

6. **Edad**

   Edad del cliente al momento de solicitar el crédito. Fue calculada restando las variables fecha valor (fecha en que se le otorgó el crédito) menos la fecha de nacimiento.

Medidas de Resumen	**Edad**	**Edad** *Bueno*	**Edad** *Malo*
Mínimo	18	18	18
1^{er} Cuartil	36	37	29
Mediana	51	52	40
Media	51	51	43
3^{er} Cuartil	65	65	56
Máximo	83	83	81
Desvío	17	17	17

Cuadro 3.8: Medidas de resumen de la variable Edad.

Al discriminar la variable Edad entre *Bueno* y *Malo*, se pueden observar diferencias importantes. Los clientes calificados como malos pagadores parecen ser de edades menores que los buenos pagadores, estos tienen una edad promedio menor.

Parece ser que las personas con mayor edad son mejores pagadoras. Todo esto da indicios de que sería buena opción incluirla en el modelo.

7. **Estado Civil**

   La variable Estado Civil del solicitante del crédito se decide reagrupar en menos niveles ya que algunas modalidades presentaban pocas observaciones:

Código	**Estado civil**
1	Soltero
2	Casado / Concubino
3	Separado / Divorciado
4	Viudo

Cuadro 3.9: Recodificación de la variable Estado Civil.

Estado Civil	**Frecuencia Relativa** (%)	**Frecuencia Relativa** *Bueno* (%)	**Frecuencia Relativa** *Malo* (%)
Soltero	29	28	40
Casado/ Concubino	46	47	40
Separado/ Divorciado	12	12	11
Viudo	13	13	9

Cuadro 3.10: Frecuencia relativa de la variable Estado Civil.

La diferencia más apreciable es en los solteros, aumenta la proporción dentro de los que fueron calificados como malos pagadores disminuyendo los Casados o con Concubino.

8. **Importe**

Importe del capital del crédito que se le otorga al cliente.

Mínimo	**Primer Cuartil**	**Mediana**	**Media**	**Tercer Cuartil**	**Máximo**	**Desvío**
1417	10360	15620	19050	23440	386500	7114

Cuadro 3.11: Medidas de resumen de la variable Importe.

9. **Ocupación**

La variable Ocupación del solicitante del crédito se recodifica en tres categorías estudiando previamente el comportamiento de cada una.

Código	**Ocupación**
R	Profesionales, Trabajador Temporal Privado, Domesticas/Rentas, Trabajadores Independientes, Contratado temporal Público u Otros.
A	Empleado Fijo Privado
V	Jubilados, Pensionistas o Empleado Fijo Público

Cuadro 3.12: Recodificación de la variable Ocupación.

Ocupación	**Frecuencia**	**Frecuencia**	**Frecuencia**

	Relativa (%)	**Relativa** *Bueno* (%)	**Relativa** *Malo* (%)
V	54	56	42
A	43	41	55
R	3	3	3

Cuadro 3.13: Frecuencia relativa de la variable Ocupación.

Más de la mitad de los créditos son otorgados a los jubilados, pensionistas y empleados fijos públicos, quizás porque es la categoría más estable en este sentido.

Al discriminar la variable Ocupación según *Bueno* o *Malo* se observan algunos cambios. Dentro de la categoría *Bueno*, más del 50 % pertenecen a la ocupación etiquetada como "V", mientras que en la otra pertenecen a la categoría "A". Esta variable entonces posiblemente sea importante incluirla en el modelo.

10. **Sexo**

Sexo de la persona solicitante del crédito.

Código	**Sexo**
0	Femenino
1	Masculino

Cuadro 3.14: Codificación de la variable Sexo.

Sexo	**Frecuencia Relativa** (%)	**Frecuencia Relativa** *Bueno* (%)	**Frecuencia Relativa** *Malo* (%)
Femenino	54	55	49
Masculino	46	45	51

Cuadro 3.15: Frecuencia relativa de la variable Sexo.

Dentro de la categoría *Bueno* mayoritariamente son mujeres. Sucede lo contrario en la otra categoría, igualmente no es mucha la diferencia.

11. **Total de Ingresos**

De todas las variables referentes al ingreso del cliente se decide utilizar

el Total de Ingresos debido a que, con los respectivos descuentos, permite apreciar la situación real de la persona cuando se enfrenta al pago de un crédito.

Suma del ingreso líquido, otros ingresos y los anticipos.

Medidas de Resumen	**Total de Ingresos**	**Total de** Ingresos *Bueno*	**Total de Ingresos** *Malo*
Mínimo	1942	1942	2730
1^{er} Cuartil	7266	7289	7049
Mediana	10490	10580	9614
Media	12560	12660	11390
3^{er} Cuartil	15460	15620	13650
Máximo	374500	374500	200100
Desvío	8700	8792	7482

Cuadro 3.16: Medidas de resumen de la variable Total de Ingresos.

Se observa que, en promedio, los sueldos de las personas calificadas como *Malo* se concentran en un rango menor, y también su media es menor comparado con los clientes calificados *Bueno*.

Este resultado nos permite concluir que la variable Total de Ingresos discrimina a las personas buenas y malas pagadoras; y en promedio se podría pensar en términos generales, que las personas con un sueldo mayor, son mejores pagadores.

12. **Valor Cuota**

Valor mensual de la cuota del crédito solicitado incluyendo intereses.

Mínimo	**Primer Cuartil**	**Mediana**	**Media**	**Tercer Cuartil**	**Máximo**	**Desvío**
227	1040	1286	1414	1633	22830	646

Cuadro 3.17: Medidas de resumen de la variable Valor Cuota.

Al discriminar según *Bueno* o *Malo* no se observan grandes diferencia.

13. **Valor Cuota / Total de Ingresos.**

Ratio calculado a partir del Valor Cuota y el Total de Ingresos para calcular el nivel de endeudamiento en base al ingreso del cliente.

Mínimo	**Primer Cuartil**	**Mediana**	**Media**	**Tercer Cuartil**	**Máximo**	**Desvío**
0,01	0,09	0,13	0,14	0,18	0,43	0,063

Cuadro 3.18: Medidas de resumen de la variable Valor Cuota/Total de Ingresos.

Se puede decir que el ratio promedio, 0,14, hace accesible el pago del crédito ya que el valor de la cuota es un 14 % del sueldo de la persona. Casi el 95 % de los casos no superan un ratio de 0,25 siendo el límite aceptable para la empresas 0,40 (éste último no llega al 1 % de los casos). Al discriminar según *Bueno* o *Malo* no se observan grandes diferencias.

14. **Bueno y Malo**

Esta variable es la variable dependiente del modelo. Fue calculada a partir de los días de atrasos actuales (MORA) y los atrasos que el cliente tuvo en cada cuota durante el crédito (estos fueron medidos por las variables tramo 1, tramo 2, tramo 3, tramo 4, tramo 5 y tramo 6).

Mora: días de atraso en la cuota actual.

Tramo 1: cantidad de veces que la persona cayó en mora en el tramo 1 (menos de 6 días de atraso).

Tramo 2: cantidad de veces que la persona cayó en mora en el tramo 2 (entre 6 y 29 días de atraso).

Tramo 3: cantidad de veces que la persona cayó en mora en el tramo 3 (entre 30 y 59 días de atraso).

Tramo 4: cantidad de veces que la persona cayó en mora en el tramo 4 (entre 60 y 89 días de atraso).

Tramo 5: cantidad de veces que la persona cayó en mora en el tramo 5 (entre 90 y 119 días de atraso).

Tramo 6: cantidad de veces que la persona cayó en mora en el tramo 6 (más de 120 días de atraso o venta de cartera).

Estas variables se utilizaron solamente para calcular la variable BYM.

Observaciones

Atraso actual	Días de atraso desde el último pago de la cuota.	Todo dentro de un mismo préstamo, ya que sólo se tiene en cuenta los registros dentro de la misma instancia.
Cantidad de caídas	Cantidad de veces que cayó en mora en las cuotas pagas anteriores.	

Cuadro 3.19: Definiciones para la clasificación de *Bueno* y *Malo*.

La variable BYM califica el comportamiento del cliente en *Bueno*, *indiferente* o *Malo*, esta clasificación se realiza según criterios de la empresa de la siguiente manera:

Cuadro 3.20: Clasificación de Bueno (B), Indiferente (I) y Malo (M)

		<6	6-29	30-59		60-89		90-119		>=120	
Atraso actual/ Nº Caídas		Tramo 1: Sin tope	Tramo 2: Sin tope	Tramo 3: =<4	>4	Tramo 4: =<2	>2	Tramo 5: =<1	>1	Tramo 6: =<1	>1
<6	Tramo 1	B	B	B	I	B	I	B	I	I	M
6-29	Tramo 2	B	B	B	I	I	I	I	I	M	M
30-59	Tramo 3	B	I	I	M	I	M	M	M	M	M
60-89	Tramo 4	I	I	M	M	M	M	M	M	M	M
90-119	Tramo 5	M	M	M	M	M	M	M	M	M	M
>=120	Tramo 6	M	M	M	M	M	M	M	M	M	M

Como se dijo anteriormente aquellas instancias clasificadas como "Indiferentes" no se van a considerar para la estimación del modelo ya que la empresa quería contar con perfiles de clientes más marcados.

3.4. Modelo de Regresión Logística

Una vez realizado el análisis exploratiorio de las variables y cada una de ellas respecto a la variable dependiente se procede a realizar las pruebas correspondientes con diferentes modelos y diferentes muestras, con el objetivo de encontrar el más adecuado.

En un principio se realizan las estimaciones en base a una muestra del 90 % de la población (Modelos 1.a, 1.b y 1.c) y luego con el 50 % (Modelo 2), con el fin de contar con más datos donde evaluar el desempeño del modelo.

Cómo los clientes calificados como *Malo* no llegan a ser el 10 % del total de la población, para explorar la técnica, se decide tomar una muestra en la que la proporción de clientes *Malo* fuese igual a la de *Bueno* (Modelo 3).

En la búsqueda de mejorar los resultados, como se observaba que los clientes con categoría ocupacional *Activos* tenían un perfil diferente a los *Pasivos* se decide considerar un modelo diferente para cada uno de ellos tomando las respectivas muestras al 50 % (Modelos 4 y 5).

Lo mismo se realizó con los clientes que ya habían operado en la empresa más de una vez y con los que era su primera operación, se estimó un modelo para cada perfil (Modelos 6 y 7).

Para realizar estos cálculos se utiliza el software R versión 3.0.1.[R Core Team, 2014].

A continuación se brinda una breve descripción de cada modelo estimado para lograr obtener un modelo que sea parsimonioso, que logre buenas predicciones y que a su vez se adapte a las necesidades de la empresa.

Para cada uno de ellos se estudió la significación de los modelos a través del *test de razón de verosimilitud*. Se comparó el modelo nulo con el modelo completo, éste tendrá una mayor probabilidad logarítmica o al menos la misma que el modelo nulo.

El test de *Razón de Verosimilitud* se utilizó como medida global para evaluar el ajuste del modelo a los datos, con un nivel de significación del 5 %.

La hipótesis nula plantea que los coeficientes estimados del modelo son todos cero contra la alternativa de que alguno es diferente de cero.

También se evaluó la significación de cada uno de los parámetros del modelo a través del estadístico de $Wald$, test que evalúa el nivel de significación de cada parámetro. En la estimación del modelo éste nos devuelve además el valor estimado del parámetro, el desvío estándar, el error y por último el nivel de significación.

Para comparar y analizar los diferentes modelos, las herramientas que se utilizaron fueron la curva ROC, el área debajo de la curva y el estadístico K-S, para estos procedimientos se utilizó el paquete "proc" [Robin et al., 2011.] del software R.

Para poder graficar la curva se utilizan los valores de las tasas de verdadero positivo y falso positivo para algunos puntos de corte.

Como se dijo en el marco teórico, el punto de corte que maximiza el estadístico $K - S$ se corresponde con el punto en la curva ROC cuya distancia vertical al eje es máxima, este punto es el utilizado para realizar las tablas con los errores de predicción.

En cuanto al estadístico K-S, cabe aclarar que durante la pasantía se realizó una indagatoria de campo con profesionales que aplicaron modelos de scoring crediticio y que con base a juicio experto dieron su opinión en este tipo de diseños. Estos establecieron que un buen valor de la medida del estadístico K-S, para un Scoring de aprobación crediticia, debe estar entre 30 y 45 puntos. Esta fue una de las herramientas utilizadas para la interpretación de los resultados.

3.4.1. Calibración del Modelo

A continuación se presentan los resultados de algunos de los modelos estimados, considerando distintas variables y a su vez diferentes muestras.

3.4.1.1. Estimación de los diferentes modelos.

1 **Modelos con muestra del 90 % de la población**.

Para estos modelos se saca una muestra por muestreo aleatorio simple sin reposición del 90 % de la población.

Figura 3.1: Muestra del 90 % de la población.

Con $N = 236418$, $n = 212794$, la proporción clientes calificados como $Malo$: $8,16\,\%$.

Dado $P(Y = 1|x) = \pi = \frac{\exp(\sum X_i \widehat{\beta}_i)}{(1+\exp(\sum X_i \widehat{\beta}_i))}$

Se muestran los resultados de los siguientes modelos:

a) X =(Edad , Clearing , Sexo , Ocupación , Actividad Económica, Total de Haberes , Antigüedad , Estado Civil , LC , Total de Ingresos , Cuotas totales , Valor cuota / Total de Ingresos , Valor cuota , Cantidad de veces que operó , Contactabilidad , Tiene RUT)

   Estimado el modelo se comprueba que es significativo con una confianza del 95 %. También se testea la significación de cada parámetro estimado a través del estadístico de *Wald*.

Para evaluar el modelo, y los errores de predicción se calculan las tasas de falso positivo y las tasas de verdadero negativo, la curva ROC y el área debajo de la curva, así como también el estadístico *K-S*. Con este último se determina cuál es el umbral óptimo para calcular los errores de predicción.

Dónde se maximiza la distancia, es en el punto de corte: 0,08, que además coincide con el criterio utilizado para la elección del punto de corte cuando los costos de clasificar en uno u otro grupo son iguales. La distancia máxima es: $0,32$.

Los resultados de las predicciones, utilizando el punto de corte óptimo son:

		Predicción	
		Bueno	*Malo*
Observado	*Bueno*	68 %	32 %
	Malo	35 %	65 %

Cuadro 3.21: Errores de clasificación modelo a, muestra del 90 % de la población.

b) X = (Edad , Clearing , Sexo , Línea de Crédito , Total de Ingresos, Cuotas totales , Valor cuota / Total de Ingresos , Valor cuota, Cantidad de veces que operó , Contactabilidad , Tiene RUT)

En el caso anterior, el modelo es significativo con un 95 % de confianza según el *Test de Razón de Verosimilitud*. Así como también todos los parámetros son significativos con un 95 % de confianza excepto el de *Cantidad de veces que operó* y algunas modalidades de *Contactabilidad*.

Se evalúa el modelo utilizando el mismo procedimiento que en el anterior. Se calcula dónde es que se maximiza la distancia *K-S*, la curva ROC y el AUC. El punto de corte óptimo es $0,08$. La distancia máxima es $0,56$.

Los resultados de las predicciones, utilizando el punto de corte óptimo son:

		Predicción	
		Bueno	*Malo*
Observado	*Bueno*	73 %	27 %
	Malo	17 %	83 %

Cuadro 3.22: Errores de clasificación modelo b, muestra del 90 % de la población.

c) X= (Cantidad de veces operó, Edad, Sexo, Antigüedad, Clearing, Cuotas totales, Valor cuota / Total de Ingresos)

Tras realizar diferentes análisis se concluye que la variable Línea de Crédito es la más discriminante en la mayoría de los modelos, por lo que se decide integrar esta variable dentro de la categoría $Verde$ del Clearing, correspondiente al mejor comportamiento. Como la línea de crédito se les otorga a los clientes por su buen comportamiento, todos los casos con línea tenían la misma calificación en el Clearing.

En este caso, no sólo el modelo es significativo con un 95 % de confianza sino que también lo son todas las variables que se incluyeron en este modelo, para testear la significación de los parámetros se utiliza el estadístico de $Wald$.

Para este modelo a continuación se presenta la tabla con valores de las tasas de verdaderos negativos, falsos positivos, falsos negativos y verdaderos positivos para algunos puntos de corte:

Punto de Corte: t	$\hat{t_N}$	$\hat{f_P}$	$\hat{f_N}$	$\hat{t_P}$	$\lvert\hat{t_P} - \hat{f_P}\rvert$
0,00	0,04	0,96	0,01	0,99	0,03
0,02	0,25	0,75	0,09	0,91	0,16
0,04	0,53	0,47	0,22	0,78	0,31
0,06	0,57	0,43	0,20	0,80	0,37
0,08	**0,68**	**0,32**	**0,30**	**0,70**	**0,38**
⋮	⋮	⋮	⋮	⋮	⋮
1,00	0,00	0,00	0,00	0,00	0,00

Cuadro 3.23: Punto de corte óptimo según el estadístico $K-S$ modelo c, muestra del 90 % de la población.

Dónde se maximiza la distancia, es en el punto de corte: $0,08$, que además coincide con el criterio utilizado para la elección del punto de corte cuando los costos de clasificar en uno u otro grupo son iguales. La distancia máxima es: $0,38$.

		Predicción	
		Bueno	*Malo*
Observado	*Bueno*	68 %	32 %
	Malo	30 %	70 %

Cuadro 3.24: Errores de clasificación modelo c, muestra del 90 % de la población.

A continuación se presenta el gráfico de la curva ROC:

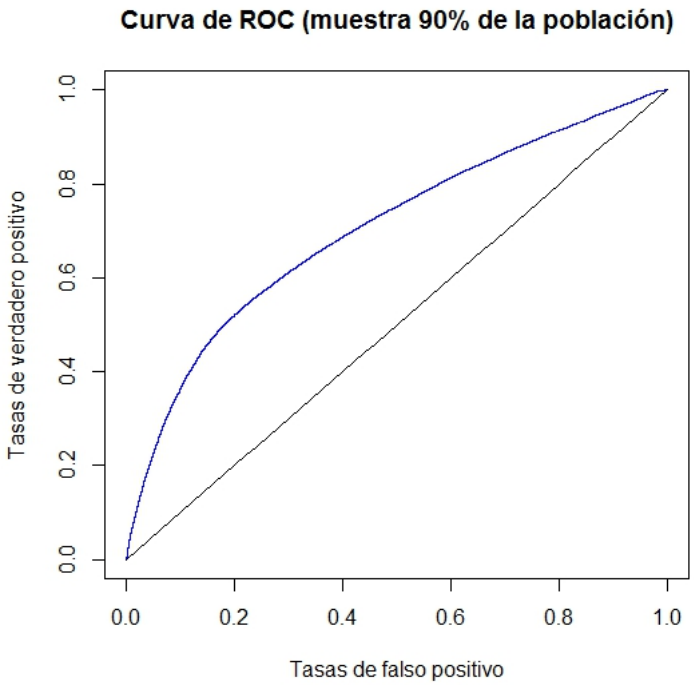

Figura 3.2: Curva ROC modelo c, muestra 90 % de la población.

Este modelo tiene una especificidad de $0,68$ y una sensibilidad de $0,70$. El área debajo de la curva, utilizada como medida global, es de: $0,75$. Por lo que el ajuste y poder predictivo de este modelo es aceptable.

2 **Modelo con muestra del 50 % de la población**.

Para este modelo se saca una muestra por muestreo aleatorio simple sin reposición del 50 % de la población.

Figura 3.3: Muestra del 50 % de la población.

Dónde $N = 236418$, $n = 118306$, la proporción de clientes calificados como $Malo$: $8,13\,\%$.

Dadas las siguientes variables:

X= (Cantidad de veces operó, Edad, Sexo, Antigüedad, Clearing, Cuotas totales, Valor cuota / Total de Ingresos)

Estimado el modelo se comprueba que es significativo con una confianza del 95 % según el *Test de razón de verosimilitud*. También se testea la significación de cada uno de los parámetros a través del estadístico de *Wald*.

Para evaluar el modelo, y los errores de predicción como en los modelos anteriores se calculan las tasas de falso positivo y verdadero negativo, la curva ROC y el área debajo de la curva, así como también el estadístico *K-S*. Con este último se determina cuál es el umbral óptimo para calcular los errores de predicción.

Punto de Corte: t	$\hat{t_N}$	$\hat{f_P}$	$\hat{f_N}$	$\hat{t_P}$	$\lvert\hat{t_P} - \hat{f_P}\rvert$
0,00	0,04	0,96	0,01	0,99	0,03
0,02	0,25	0,75	0,09	0,91	0,16
0,04	0,53	0,47	0,22	0,78	0,31
0,06	0,59	0,41	0,23	0,77	0,36
0,08	**0,68**	**0,32**	**0,30**	**0,70**	**0,38**
0,10	0,77	0,23	0,37	0,63	0,40
0,12	0,88	0,12	0,49	0,51	0,39
⋮	⋮	⋮	⋮	⋮	⋮
1,00	0,00	0,00	0,00	0,00	0,00

Cuadro 3.25: Punto de corte óptimo según el estadístico $K-S$ modelo c, muestra del 50 % de la población.

Dónde se maximiza la distancia, es en el punto de corte: $0,08$ y la distancia máxima, el $K - S$ es: $0,38$.

Utilizando el punto de corte óptimo, los resultados de predicción son:

		Predicciones	
		Bueno	*Malo*
Observados	*Bueno*	68 %	32 %
	Malo	30 %	70 %

Cuadro 3.26: Errores de clasificación modelo c, muestra del 50 % de la población.

El gráfico de la curva ROC presentado a continuación muestra la variación de los errores de clasificación a medida que varía el punto de corte:

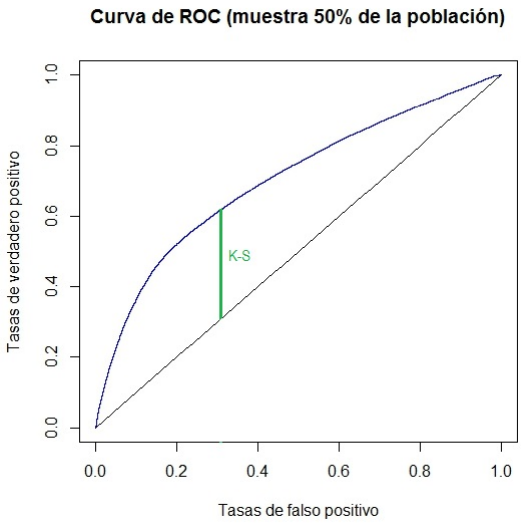

Figura 3.4: Curva ROC modelo c, muestra 50 % de la población.

El modelo también tiene una especificidad de $0,68$ y una sensibilidad de $0,70$. El área debajo de la curva, que se puede utilizar como medida global, es de $0,75$. Por lo que el ajuste y poder predictivo de este modelo es aceptable.

3 **Modelo para muestra con igual proporción de** *Bueno* **y** *Malo*.

Para trabajar el siguiente modelo se realiza una muestra con igual proporción de clientes *Bueno* y *Malo*, de forma de poder observar si hay algún cambio cuando el peso de los dos perfiles de clientes es el mismo. Cabe destacar que dicho escenario no se ajusta con la realidad de la empresa.

Para estos modelos, primero se obtiene una muestra (de tamaño n) por muestreo aleatorio simple sin reposición, del 90 % de la población calificada como *Malo*. Luego, a partir de la población de clientes *Bueno*, se obtiene una muestra del mismo tamaño (n). Estas dos submuestras formarán la llamada "Muestra con igual proporción de *Bueno* y *Malo*".

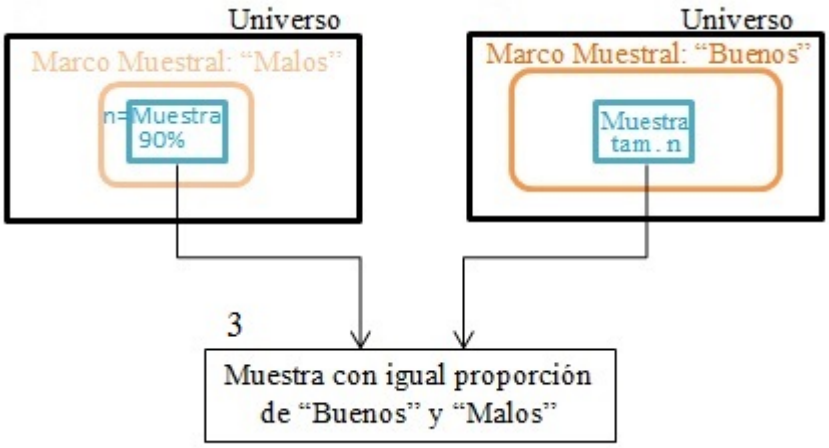

Figura 3.5: Muestra igual proporción de *Bueno* y *Malo*.

Donde, $N_M = 19208$, $n_M = 17240$, $N_B = 217210$, $n_B = 17240$.

Dadas las siguientes variables:

X= (Cantidad de veces operó, Edad, Sexo, Antigüedad, Clearing, Ocupación, Cuotas totales, Valor cuota / Total de Ingresos)

En este modelo todos los parámetros son significativas con un 95 % de confianza, al igual que el modelo en su conjunto.

A continuación se presentará la tabla con las tasas de falso positivo y verdadero negativo para determinar cuál será el umbral óptimo y el valor del estadístico $K - S$.

Punto de Corte: t	$\hat{t_N}$	$\hat{f_P}$	$\hat{f_N}$	$\hat{t_P}$	$\lvert\hat{t_P} - \hat{f_P}\rvert$
0,00	0,02	0,98	0,00	1,00	0,01
⋮	⋮	⋮	⋮	⋮	⋮
0,46	0,72	0,28	0,33	0,67	0,39
0,48	0,75	0,25	0,35	0,65	0,39
0,50	0,78	0,22	0,37	0,63	0,41
0,52	**0,83**	**0,17**	**0,42**	**0,58**	**0,41**
0,54	0,85	0,15	0,45	0,55	0,40
⋮	⋮	⋮	⋮	⋮	⋮
1,00	0,00	0,00	0,00	0,00	0,00

Cuadro 3.27: Punto de corte óptimo según el estadístico $K - S$, muestra igual proporción de *Bueno* y *Malo*.

Utilizando el punto de corte óptimo, en este caso es $0,50$ con un $K - S$ de 0,41, los resultados de predicción son:

		Predicciones	
		Bueno	*Malo*
Observados	*Bueno*	78 %	22 %
	Malo	37 %	63 %

Cuadro 3.28: Errores de clasificación, muestra igual proporción de *Bueno* y *Malo*.

La curva ROC para este modelo es:

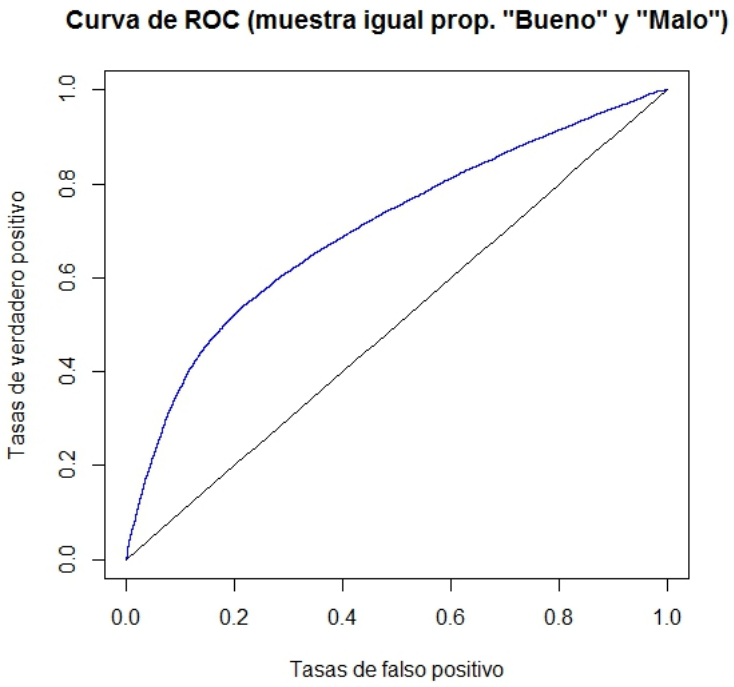

Figura 3.6: Curva ROC, muestra igual proporción de *Bueno* y *Malo*.

Este modelo tiene una especificidad de $0,78$ y una sensibilidad de $0,63$. El área debajo de la curva es de $0,75$. Por lo que el ajuste y poder predictivo de este modelo es aceptable.

Este modelo se realiza para observar el comportamiento del modelo cuando se tiene un escenario dónde la proporción de clientes morosos es igual a la de no morosos. **Se entiende que no es representativa de la realidad por lo que no se podría tomar como modelo para realizar las futuras predicciones.**

4 **Modelo para muestra con 50 % de clientes** *Activos*.

Considerando todos aquellos clientes cuya ocupación no está dentro de la categoría Jubilado o Pensionista, se saca una muestra por muestreo aleatorio simple sin reposición del 50 % de dicha población.

Figura 3.7: Muestra 50 % de clientes *Activos*.

Donde, $N = 155262$, $n = 77986$, la proporción de clientes clasificados como $Malo$ en la muestra es de: $9,6\,\%$.

Dadas las siguientes variables:

X= (Cantidad de veces operó, Edad, Sexo, Antigüedad, Clearing, Ocupación, Cuotas totales, Valor cuota / Total de Ingresos)

Este modelo es significativo con un 95 % de confianza, todos los parámetros son significativos con ese nivel de confianza.

Para evaluar el modelo, y los errores de predicción como en los modelos anteriores se calculan las tasas de falso positivo y verdadero negativo, la curva ROC y el área debajo de la curva, así como también el estadístico *K-S*.

Punto de Corte: t	$\hat{t_N}$	$\hat{f_P}$	$\hat{f_N}$	$\hat{t_P}$	$\lvert\hat{t_P} - \hat{f_P}\rvert$
0,00	0,04	0,96	0,01	0,99	0,03
0,02	0,20	0,80	0,05	0,95	0,15
0,04	0,43	0,57	0,14	0,86	0,29
0,06	0,57	0,43	0,21	0,79	0,36
0,08	0,68	0,32	0,28	0,72	0,40
0,10	**0,76**	**0,24**	**0,34**	**0,66**	**0,42**
0,12	0,82	0,18	0,41	0,59	0,41
0,14	0,86	0,14	0,47	0,53	0,39
0,16	0,89	0,11	0,53	0,47	0,36
⋮	⋮	⋮	⋮	⋮	⋮
1,00	0,00	0,00	0,00	0,00	0,00

Cuadro 3.29: Punto de corte óptimo según el estadístico $K - S$, muestra 50 % de la población *Activos*.

Dónde se maximiza la distancia, es en el punto de corte: $0,10$ y la distancia máxima, el $K - S$ es: $0,42$.

Utilizando el punto de corte óptimo, los resultados de predicción son:

		Predicciones	
		Bueno	*Malo*
Observados	*Bueno*	76 %	24 %
	Malo	34 %	66 %

Cuadro 3.30: Errores de clasificación, muestra 50 % de la población *Activos*.

Como se puede observar, al realizar la primer división de la población (Activos-Pasivos), se predicen bien los clientes calificados como *Bueno* que son pasivos, pero no tanto los clientes calificados como *Malo*.

El gráfico de la curva ROC que muestra la variación de los errores de clasificación a medida que varía el punto de corte, es:

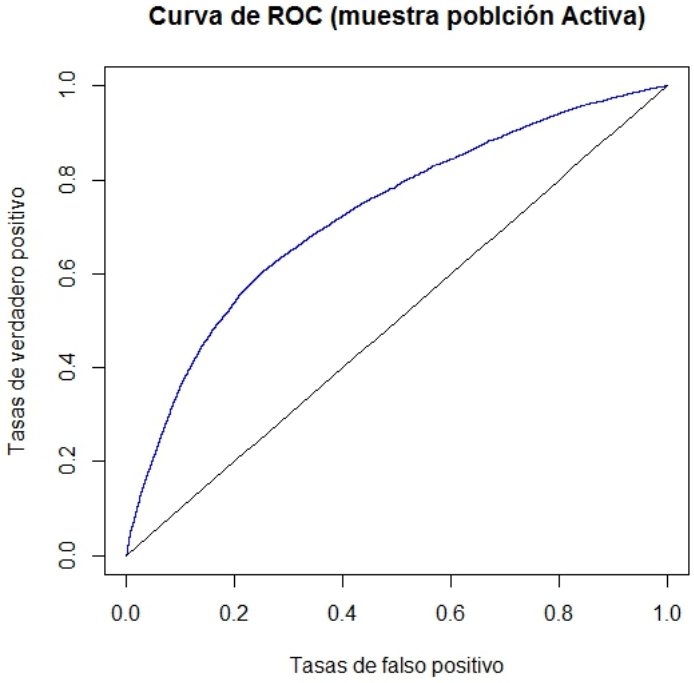

Figura 3.8: Curva ROC, muestra 50 % de la población *Activos*.

Este modelo tiene una especificidad de $0,76$ y una sensibilidad de $0,66$. El área debajo de la curva es de $0,77$. Por lo que el ajuste y poder predictivo de este modelo es aceptable.

5 **Modelo para muestra con 50 % de clientes** *Pasivos*.

Considerando todos aquellos clientes pasivos, se saca una muestra por muestreo aleatorio simple sin reposición del 50 % de dicha población.

Figura 3.9: Muestra 50 % de clientes *Pasivos*.

Donde, $N = 81156$, $n = 40632$, la proporción de clientes calificados como $Malo$ en la muestra es de: $5,18\,\%$

Dadas las siguientes variables:

X= (Cantidad de veces operó, Edad, Sexo, Antigüedad, Clearing, Cuotas totales, Valor cuota / Total de Ingresos)

Los parámetros son significativos con un 95 % de confianza, al igual que el modelo en su conjunto según el test realizado.

A continuación se presenta la tabla con las tasas de falso positivo y verdadero negativo para determinar cuál será el umbral óptimo y el valor del estadístico $K - S$.

Punto de Corte: t	$\hat{t_N}$	$\hat{f_P}$	$\hat{f_N}$	$\hat{t_P}$	$\lvert\hat{t_P} - \hat{f_P}\rvert$
0,00	0,04	0,96	0,01	0,99	0,03
0,02	0,17	0,83	0,11	0,89	0,06
0,04	0,76	0,24	0,53	0,47	0,23
0,06	**0,91**	**0,09**	**0,67**	**0,33**	**0,24**
0,08	0,93	0,07	0,71	0,29	0,23
0,10	0,95	0,05	0,73	0,27	0,22
⋮	⋮	⋮	⋮	⋮	⋮
1,00	0,00	0,00	0,00	0,00	0,00

Cuadro 3.31: Punto de corte óptimo según el estadístico $K - S$, muestra 50 % de la población *Pasivos*.

Utilizando el punto de corte óptimo, en este caso es $0,06$ con un $K - S$ de $0,24$, los resultados de predicción son:

		Predicciones	
		Bueno	*Malo*
Observados	*Bueno*	91 %	9 %
	Malo	67 %	33 %

Cuadro 3.32: Errores de clasificación, muestra 50 % de la población *Pasivos*.

En este caso no son buenas las predicciones, principalmente para los clientes con categoría *Malo*.

La curva ROC para este modelo es:

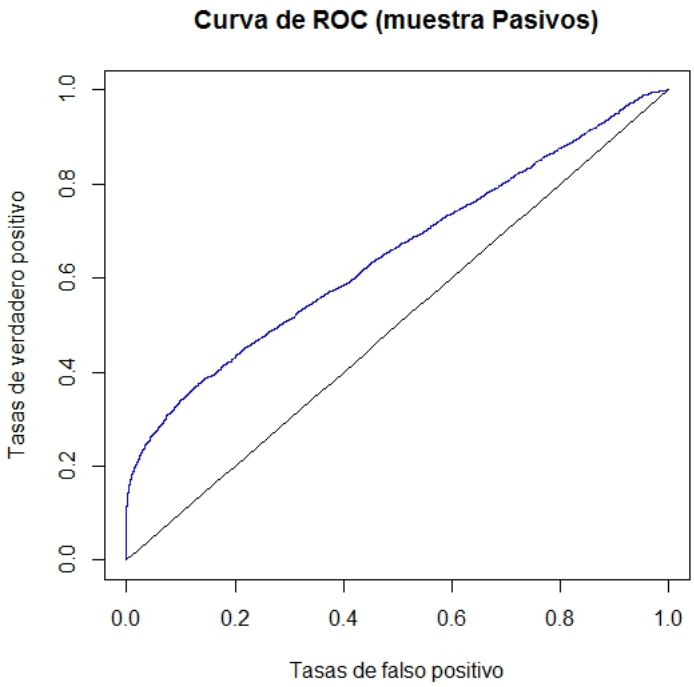

Figura 3.10: Curva ROC, muestra 50 % de la población *Pasivos*.

La especificidad es $0,91$ y la sensibilidad es $0,33$. El área debajo de la curva es de $0,65$. Por lo que el ajuste y poder predictivo de este modelo no es aceptable.

6 **Modelo con muestra 50 % de clientes que operaron por primera vez en la empresa.**

Considerando todos aquellos clientes que operan por primera vez en la empresa, se saca una muestra por muestreo aleatorio simple sin reposición del 50 % de dicha población.

Figura 3.11: Muestra 50 % de clientes que operaron por primera vez en la empresa.

Donde, $N = 1010$, $n = 504$, la proporción de clientes calificados como $Malo$ en la muestra es de: 80 %

Dadas las siguientes variables:

X= (Edad, Sexo, Antigüedad, Clearing, Ocupación, Cuotas totales, Valor cuota / Total de Ingresos)

El modelo es significativo con un 95 % de confianza. En el caso de clientes que operan por primera vez, sólo el parámetro de la variable $Edad$ es significativo con una confianza del 95 % y algunas modalidades de la variable $Clearing$.

Para evaluar el modelo, y los errores de predicción como en los modelos anteriores se calculan las tasas de falso positivo y verdadero negativo, la curva ROC y el área debajo de la curva, así como también el estadístico K-S.

Dónde se maximiza la distancia, es en el punto de corte: $0,8$ y la distancia máxima, el $K - S$ es: $0,69$. Utilizando el punto de corte óptimo, los resultados de predicción son:

		Predicciones	
		Bueno	*Malo*
Observados	*Bueno*	86 %	14 %
	Malo	17 %	83 %

Cuadro 3.33: Errores de clasificación, muestra 50 % de los clientes que operaron sólo una vez en la empresa.

Las predicciones para los clientes clasificados como *Bueno* son muy buenas. Para observar mejor el comportamiento del modelos se muestra a continuación la curva ROC y el AUC.

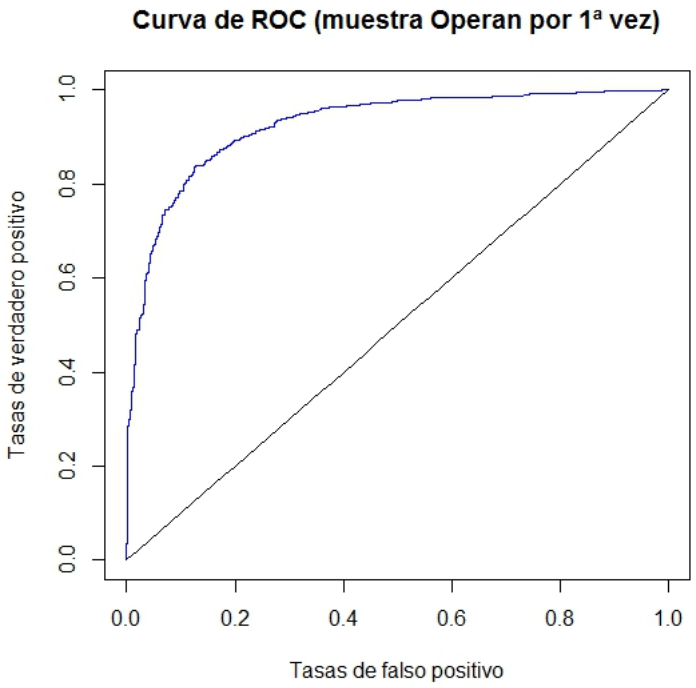

Figura 3.12: Curva ROC, muestra 50 % de los clientes que operaron sólo una vez en la empresa.

La especificidad es de $0,86$ y la sensibilidad es de $0,83$. El área debajo de la curva es de $0,91$. Por lo que el ajuste y poder predictivo de este modelo es más que aceptable.

7 **Modelo con muestra 50 % de clientes que han operado más de una vez en la empresa.**

Considerando todos aquellos clientes que ya han operado en la empresa, se saca una muestra por muestreo aleatorio simple sin reposición del 50 % de dicha población.

Figura 3.13: Muestra 50 % de clientes que han operado más de una vez en la empresa.

Con $N = 235408$, $n = 117967$, la proporción de clientes calificados como $Malo$ en la muestra es de: 2 %

Dadas las siguientes variables:

X= (Cantidad de veces operó, Edad, Sexo, Antigüedad, Clearing, Ocupación, Cuotas totales, Valor cuota / Total de Ingresos)

En el caso de los clientes que han operan más de una vez, todos los parámetros estimados son significativas con un 95 % de confianza.

Se estudia la tabla con las tasas de falso positivo y verdadero negativo para determinar cuál será el umbral óptimo y el valor del estadístico $K - S$.

Utilizando el punto de corte óptimo, en este caso es $0,02$ con un $K-S$ de $0,24$, los resultados de predicción son:

		Predicciones	
		Bueno	*Malo*
Observados	*Bueno*	63 %	37 %
	Malo	39 %	61 %

Cuadro 3.34: Errores de clasificación, muestra 50 % de los clientes que han operado más de una vez en la empresa.

La curva ROC para este modelo es:

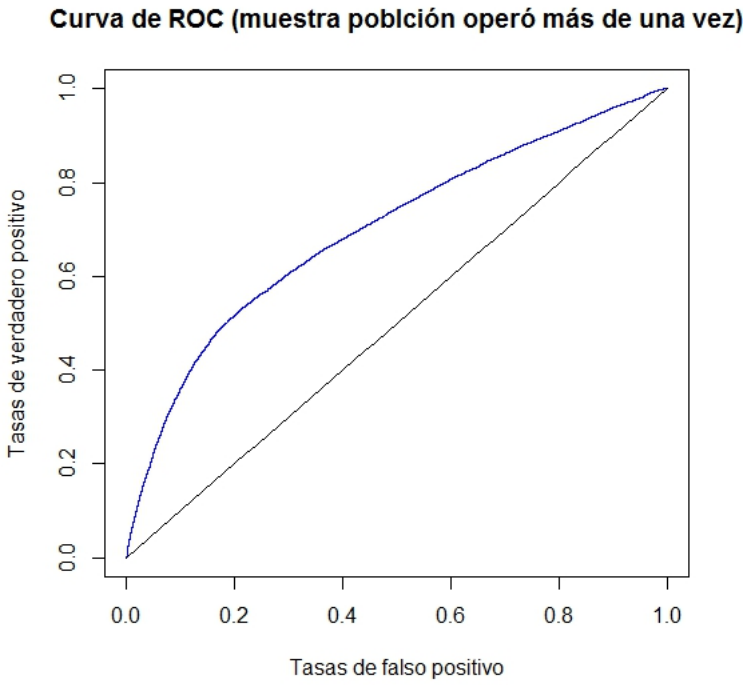

Figura 3.14: Curva ROC, muestra 50 % de clientes que han operado más de una vez en la empresa.

Para este modelo la especificidad es de $0,63$ y la sensibilidad es de $0,61$. El área debajo de la curva es de $0,67$. Por lo que el ajuste y poder predictivo de este modelo no es aceptable.

8 **Modelos con una muestra del 50 % de la población cambiando la definición de la variable *Bueno* y *Malo*.**

En este caso se realiza un muestreo aleatorio simple sin reposición del 50 % de la población, al igual que en el modelo 2. Pero con el fin de poder mejorar los resultados, se decide realizar cambios en la definición de *Bueno* y *Malo* (variable dependiente), ya que se cree que los resultados pueden mejorar si se es más estricto en el criterio de clasificación de esta variable.

Dado que no se observan mejoras en los resultados utilizando las diferentes definiciones, se decide trabajar con ésta última.

Por lo tanto, considerando todos los análisis anteriores y debido a que las distintas definiciones de *Bueno* y *Malo*, así como también los resultados con las distintas muestras no presenta grandes variaciones, se decide darle prioridad a los modelos que se describen a continuación:

3.4.1.2. Estimación del modelo elegido.

Considerando los análisis anteriores y debido a que las distintas definiciones de *Bueno* y *Malo*, así como también los resultados con las distintas muestras no presentan grandes variaciones, se decide darle prioridad a los modelos que se describen a continuación.

Para comparar y analizar los diferentes modelos, las herramientas que se utilizaron fueron la curva ROC, el área debajo de la curva y el estadístico $K-S$, para estos procedimientos se utilizó el paquete "proc" [Robin et al., 2011.] del software R.

El punto de corte que maximiza el estadístico $K-S$ se utiliza para determinar el umbral "óptimo", utilizado para obtener las tablas con los errores de clasificación.

- Modelo 2:

  X=(Cantidad de veces que operó, Edad, Sexo, Antigüedad, Clearing, Cuotas totales, Valor cuota / Total de Ingresos)

Coeficientes	**Estimación**	**Error std.**	**Valor z**	**P($>\|z\|$)**
Intercepción	-0,12	0,13	-0,94	0,346
Cant veces operó	-2,69	0,05	-50,03	< 2e-16 ***
Edad	-0,01	0,00	-16,95	< 2e-16 ***
Sexo M	0,22	0,02	10,33	< 2e-16 ***
Antigüedad > 60 o JP	-0,89	0,03	-27,79	< 2e-16 ***
Antigüedad 25-48	-0,30	0,04	-8,11	4,9e-16 ***
Antigüedad 49-60	-0,50	0,06	-9,00	< 2e-16 ***
Clearing AMARILLO A2	2,01	0,12	16,96	< 2e-16 ***
Clearing AMARILLO A3	0,98	0,12	8,36	< 2e-16 ***
Clearing AMARILLO M y ROJO	1,16	0,11	10,46	< 2e-16 ***
Clearing VERDE y LC	0,08	0,11	25,75	< 2e-16 ***
Cuotas totales	0,07	0,00	31,21	< 2e-16 ***
Valor cuota TotIng	2,43	0,18	13,30	< 2e-16 ***

Cuadro 3.35: Resumen del modelo 2, muestra del 50 % de la población.

En el cuadro anterior se presentan los resultados de las estimaciones realizadas por el programa utilizado. Se aprecia que todos los parámetros estimados son significativos con una confianza mayor al 95 %, según el estadístico de Wald.

Para el modelo en conjunto, el test de *Razón de Verosimilitud* indica que el modelo es significativo según el valor de tabla de una $\chi^2_{(p+1-q),0,05}$, el p-valor devuelto es muy pequeño. Esto se puede apreciar en la siguiente tabla:

	Resid. Df	**Resid. Dev**	**Df**	**Deviance**	**Pr(>Chi)**
Modelo nulo	120505	76448,32			
Modelo c	120493	63816,40	12	12631,92	<2,2e-16 ***

Cuadro 3.36: Test de razón de Verosimilitud modelo 2, muestra del 50 % de la población.

Predicciones:

		Predicción	
		Bueno	*Malo*
Observado	*Bueno*	68 %	32 %
	Malo	30 %	70 %

Cuadro 3.37: Errores de clasificación modelo 2, muestra del 50 % de la población.

Como se analizó al describir el modelo 2, se aprecia que proporciona un buen ajuste y un buen poder predictivo.

Tiene una especificidad de $0,68$ y una sensibilidad de $0,70$. El área debajo de la curva, que se puede utilizar como medida global, es de $0,75$. Por lo que el ajuste y poder predictivo de este modelo es aceptable.

Como a la empresa le interesaba contar con la variable *ocupación* y dado que las predicciones quedan casi invariantes, se decide estimar el modelo con el agregado de esta variable. Antes que nada se decide realizar un test de *Razón de Verosimilitud* para comparar ambos modelos.

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Modelo 2	120493	63816,40			
Modelo 2 más Ocupación	120491	63807,81	2	8,59	0,0136 *

Cuadro 3.38: Test de razón de Verosimilitud modelo 2 vs. modelo 2 más la variable *Ocupación*.

Este cuadro nos indica que agregar la variable *Ocupación* es una buena decisión aunque una de sus modalidades no sea significativa según el estadístico de $Wald$, con el nivel de confianza establecido.

- Variables consideradas en la estimación del modelo:

X=(Cantidad de veces que operó, Edad, Sexo, Antigüedad, Clearing, Ocupación, Cuotas totales, Valor cuota / Total de Ingresos)

Coeficientes	Estimación	Error std.	Valor z	P(> |z|)
Intercepción	-0,13	0,13	-1,01	0,31
Cant veces operó	-2,57	0,05	-49,90	< 2e-16 ***
Edad	-0,01	0,00	-15,91	< 2e-16 ***
SexoM	0,21	0,02	10,49	< 2e-16 ***
Antigüedad>60oJP	-0,87	0,03	-27,42	< 2e-16 ***
Antigüedad 25-48	-0,30	0,04	-8,20	2,5e-16 ***
Antigüedad 49-60	-0,44	0,06	-9,05	< 2e-16 ***
Clearing AMARILLO A2	1,84	0,12	16,94	< 2e-16 ***
Clearing AMARILLO A3	0,78	0,12	8,35	< 2e-16 ***
Clearing AMARILLO M y ROJO	1,05	0,11	10,46	< 2e-16***
Clearing VERDE y LC	-0,06	0,11	24,75	< 2e-16 ***
Ocu R	0,20	0,06	2,97	0,00297 **
Ocu V	0,001	0,03	0,40	0,69
Cuotas totales	0,07	0,00	31,17	< 2e-16 ***
Valor cuota Tot Ing	2,52	0,18	13,31	< 2e-16 ***

Cuadro 3.39: Resumen del modelo 2 incluyendo la variable *Ocupación*, muestra del 50 % de la población.

Al igual que en el caso anterior se presentan los resultados de las estimaciones realizadas. Como se puede apreciar el parámetros estimado de la modalidad V de la variable ocupación no es significativa con el nivel de confianza establecido.

El test de razón de verosimilitud, cuando se compara el modelo nulo con el modelo completo, indica que es significativo según el valor de tabla de una $\chi^2_{(p+1-q),0,05}$, el pvalor devuelto es muy pequeño. Esto se puede apreciar en la siguiente tabla:

	Resid. Df	**Resid. Dev**	**Df**	**Deviance**	**Pr(>Chi)**
Modelo nulo	120505	76448,32			
Modelo2 más Ocupación	120491	63807,81	14	12640,51	<2,2e-16***

Cuadro 3.40: Test de razón de Verosimilitud modelo 2 incluyendo la variable *Ocupación*, muestra del 50 % de la población.

		Predicción	
		Bueno	*Malo*
Observado	*Bueno*	68 %	32 %
	Malo	30 %	70 %

Cuadro 3.41: Errores de clasificación modelo 2 incluyendo la variable *Ocupación*, muestra del 50 % de la población.

La medida del estadístico $K - S$ en este modelo es de: 38 %. Al igual que el modelo anterior tiene una especificidad de $0,68$ y una sensibilidad de $0,70$. El área debajo de la curva, que se puede utilizar como medida global, es de $0,75$. Por lo que el ajuste y poder predictivo de este modelo es aceptable.

Las variables que se incluirán en el **Modelo Definitivo** son:

- Cantidad de veces que operó
- Edad
- Sexo
- Antigüedad laboral
- Clearing
- Ocupación
- Cuotas totales
- Ratio Valor de la cuota / Total de Ingresos líquidos.

Dicho modelo tiene una especificidad de un 68 %, nos indica la capacidad de nuestro estimador para predecir que el cliente es *Bueno* dado que realmente lo es. Y una sensibilidad de 70 % que indica la capacidad de nuestro estimador para clasificar a los clientes como *Malo* dado que realmente esa era su clasificación.

3.4.2. Parámetros del modelo e interpretación

A continuación se realizará una descripción e interpretación detallada de los parámetros del modelo elegido.

Para esta interpretación se necesita, en primer lugar, del cociente o razón de odds (Odd Ratio):

$$
\begin{aligned}
lO(x_1, \ldots, x_{j+1}, \ldots, x_k) &= \frac{P(Y=1)}{P(Y=0)} && (3.1)\\
&= \frac{\frac{1}{e^{-\beta_0-\beta_1 x_1-\ldots-\beta_k x_k}}}{1-\frac{1}{e^{-\beta_0-\beta_1 x_1-\ldots-\beta_k x_k}}} && (3.2)\\
&= e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j+\ldots+\beta_k x_k} && (3.3)
\end{aligned}
$$

Aumentando la variable x_j una unidad, manteniendo las demás constante, quedaría:

$$O(x_1, ..., x_j + 1, ..., x_k) = e^{\beta_0 + \beta_1 x_1 + ... + \beta_j (x_{j+1}) + ... + \beta_k x_k} \tag{3.4}$$

Si se dividen los dos cocientes:

$$\frac{O(x_1, ..., x_j + 1, ..., x_k)}{O(x_1, ..., x_j + 1, ..., x_k)} = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_j (x_j + 1) + ... + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + ... + \beta_j x_j + ... + \beta_k x_k}} \tag{3.5}$$

$$= e^{\beta_j} \tag{3.6}$$

Escrito de otra forma:

$$O(x_1, ..., x_j + 1, ..., x_k) = \exp^{\beta_j} \ O(x_1, ..., x_j + 1, ..., x_k) \tag{3.7}$$

En consecuencia, la razón de odds se multiplicará por $\exp^{\beta_j}$ cuando se aumenta una unidad el valor de x_j (manteniendo constantes todas las demás).

En cada caso se toma como referencia el aumento o disminución de la probabilidad de ser *Malo*.

Para la interpretación de los parámetros del modelo, se realizan los siguientes cálculos.

1. **Término independiente.**

   El término independiente en el modelo de regresión logística es el siguiente:

	$\widehat{\beta}_0$	**Exp($\widehat{\beta}_1$)**
	-0,14	0,87

Cuadro 3.42: Término independinete.

2. **Cantidad de veces que operó.**

   Para cada instancia, cuando ingresa la solicitud de un crédito ésta variable tomará el valor cero si la persona no ha operado anteriormente

en la empresa y sino será la cantidad de veces que ha operado.

El parámetro asignado a esta variable es:

	$\widehat{\beta}_1$	$\mathbf{Exp}(\widehat{\beta}_1)$
Cant. de veces que operó	-2,57	0,077

Cuadro 3.43: Estimación del parámetro Cantidad de veces que operó.

Al aumentar la cantidad de veces que operó, dejando el resto de los valores constantes, el odd ratio de ser $Malo$ disminuye.

3. **Edad.**

   Esta variable se considera al momento de la solicitud del préstamo.

   La estimación del parámetro de esta variable es:

	$\widehat{\beta}_2$	$\mathbf{Exp}(\widehat{\beta}_2)$
Edad	-0,01	0,99

Cuadro 3.44: Estimación del parámetro Edad.

Al aumentar la edad, dejando el resto de los valores constantes, el odd ratio de ser $Malo$ disminuye.

4. **Sexo.**

   La categoría de referencia es *Femenino*. Por lo tanto el peso asignado a la categoría *Masculino* es:

	$\widehat{\beta}_3$	**Exp**$(\widehat{\beta}_3)$
Sexo	0,21	1,23

Cuadro 3.45: Estimación del parámetro Sexo.

   El sexo masculino aumenta el odd ratio de ser *Malo* con respecto al femenino, dejando el resto de los valores constantes.

5. **Antigüedad laboral.**

   La variable Antigüedad está dividida en cuatro categorías, la modalidad de referencia es Antigüedad *menor que* 25 *meses*. Los coeficientes para el resto de las modalidades son los siguientes:

	$\widehat{\beta}_4$	**Exp**$(\widehat{\beta}_4)$
Antigüedad>60, Jubilado o Pensionista	-0,87	0,42

Cuadro 3.46: Estimación del parámetro de la categoría Antigüedad>60, Jubilado o Pensionista.

   El tener una antigüedad *mayor a 60 meses* o ser *Jubilado* o *Pensionista* disminuye el odd ratio de ser *Malo* con respecto a tener una antigüedad *menor a 25 meses*, dejando el resto de los valores constantes.

	$\widehat{\beta}_5$	$\mathbf{Exp}(\widehat{\beta}_5)$
Antigüedad 25-48	-0,30	0,74

Cuadro 3.47: Estimación del parámetro de la categoría Antigüedad 25-48 meses.

El tener una antigüedad laboral de *25-48 meses* disminuye el odd ratio ser $Malo$ con respecto a tener una antigüedad *menor a 25 meses*, dejando el resto de los valores constantes.

	$\widehat{\beta}_6$	$\mathbf{Exp}(\widehat{\beta}_6)$
Antigüedad 49-60	-0,44	0,64

Cuadro 3.48: Estimación del parámetro de la categoría Antigüedad Antigüedad 49-60 meses.

El tener una antigüedad laboral de *48-60 meses* disminuye el odd ratio de ser $Malo$ con respecto a tener una antigüedad *menor a 25 meses*, dejando el resto de los valores constantes.

6. **Resultado Experto Clearing.**

Dicha variable consta de 5 categorías, de estas, *AMARILLO A1* es la categoría de referencia.

La estimación de los parámetros para cada categoría son los siguientes:

	$\widehat{\beta}_7$	$\mathbf{Exp}(\widehat{\beta}_7)$
Clearing AMARILLO A2	1,84	6,29

Cuadro 3.49: Estimación del parámetro de la categoría Clearing AMARILLO A2.

El tener calificación *AMARILLO A2* en Clearing aumenta el odd ra-

tio de ser *Malo* con respecto a estar en la categoría *AMARILLO A1*, dejando el resto de las variables constantes.

	$\widehat{\beta}_8$	$\mathbf{Exp}(\widehat{\beta}_8)$
Clearing AMARILLO A3	0,78	2,18

Cuadro 3.50: Estimación del parámetro de la categoría Clearing AMARILLO A3.

El tener calificación *AMARILLO A3* en Clearing aumenta el odd ratio de ser *Malo* con respecto a estar en la categoría *AMARILLO A1*, dejando el resto de las variables constantes.

	$\widehat{\beta}_9$	$\mathbf{Exp}(\widehat{\beta}_9)$
Clearing AMARILLO M o ROJO	1,05	2,86

Cuadro 3.51: Estimación del parámetro de la categoría Clearing AMARILLO MANUAL o ROJO.

El tener calificación *AMARILLO M o ROJO* en Clearing aumenta el odd ratio de ser *Malo* con respecto a estar en la categoría *AMARILLO A1*, dejando el resto de las variables constantes.

	$\widehat{\beta}_{10}$	$\mathbf{Exp}(\widehat{\beta}_{10})$
Clearing VERDE o LC	-0,06	0,94

Cuadro 3.52: Estimación del parámetro de la categoría Clearing VERDE o LC.

El tener calificación en el Clearing *VERDE o LC* disminuye el odd ratio de ser *Malo* con respecto a estar en la categoría *AMARILLO A1*, dejando el resto de las variables constantes.

7. **Ocupación.**

   La variable Ocupación se reagrupa en 3 categorías llamadas: Ocupación Rojo, Ocupación Amarilla y Ocupación Verde.

   - Ocupación Rojo contiene: Profesionales, Trabajador Temporal Privado, Domesticas/Rentas, Trabajadores Independientes, Contratado temporal Público u Otros.
   - Ocupación Amarilla contiene: Empleado Fijo Privado.
   - Ocupación Verde contiene: Jubilados, Pensionistas o Empleado fijo Público.

   La categoría de referencia es *Ocupación Amarilla*. Las estimaciones de los parámetros para cada categoría son los siguientes:

	$\widehat{\beta}_{11}$	$\mathbf{Exp}(\widehat{\beta}_{11})$
Ocupacion R	0,20	1,22

Cuadro 3.53: Estimación del parámetro de la categoría Ocupacion R.

   Pertenecer al grupo de *Ocupación Rojo* aumenta el odd ratio de ser $Malo$ con respecto a ser empleado fijo privado, dejando el resto de las variables constantes.

	$\widehat{\beta}_{12}$	$\mathbf{Exp}(\widehat{\beta}_{12})$
Ocupacion V	-0,001	0,99

Cuadro 3.54: Estimación del parámetro de la categoría Ocupación V.

   Ser *Jubilado Pensionista* o *Empleado Fijo Público* disminuye el odd ratio de ser $Malo$ con respecto a ser *Empleado Fijo Privado*, dejando el resto de las variables constantes.

8. **Plazo del préstamo (Cuotas Totales)**

La estimación del parámetro de dicha variable es la siguiente:

	$\widehat{\beta}_{13}$	$\mathbf{Exp}(\widehat{\beta}_{13})$
Cuotas totales	0,07	1,07

Cuadro 3.55: Estimación del parámetro de la variable Cuotas Totales.

Dejando el resto de las variables constantes y al aumentar la cantidad de *cuotas totales*, aumenta el odd ratio de ser $Malo$.

9. **Ratio Valor Cuota/Ingresos Líquidos Totales**

Para dicho ratio la estimación es la siguiente:

	$\widehat{\beta}_{14}$	$\mathbf{Exp}(\widehat{\beta}_{14})$
Valor cuota/Tot Ing	2,52	12,43

Cuadro 3.56: Estimación del parámetro de la variable Valor cuota/Tot Ing.

Al aumentar el ratio *Valor Cuota/Ingresos Líquidos Totales* el odd ratio de ser $Malo$ aumenta, dejando el resto de las variables constantes.

3.4.3. Cálculo de la Probabilidad de incumplimiento.

En la sección anterior se describió cada uno de los parámetros del modelo. Obtenidos esos coeficientes, estamos en condiciones de poder calcular la probabilidad de mora.

La función para dicho cálculo, utilizando la regresión logística, es la siguiente:

$$P(Y = 1|x) = \pi = \frac{\exp\left(\sum_{i=1}^{N} X_i \widehat{\beta}_i\right)}{(1+\exp\left(\sum_{i=1}^{N} X_i \widehat{\beta}_i\right))}$$

donde los X_i van a ser los valores que toman las distintas variables y los $\widehat{\beta}_i$, $i = 1, 2, ..., 15$ son los pesos fijos de cada variables y/o categoría.

$\widehat{\beta} = (-0{,}14;\ -2{,}57;\ -0{,}01;\ 0{,}22;\ -0{,}87;\ -0{,}30;\ -0{,}44;\ 1{,}84;\ 0{,}78;\ 1{,}05;\ -0{,}06;\ 0{,}20;\ -0{,}001;\ 0{,}07;\ 2{,}52))$

X = (Cantidad de veces que operó, Edad, SexoM, Antigüedad$>$ 60 *o* JP, Antigüedad $25-48$, Antigüedad $49-60$, Clearing AMARIO A2, Clearing AMARILLO A3, Clearing AMARILLO M y ROJO, Clearing VERDE y LC, Ocupación R, Ocupación V, Cuotas totales, Valor cuota/TotIng)

Ejemplos prácticos.

A continuación se detalla cómo se debe proceder para calcular la probabilidad de mora, a través de unos ejemplos para facilitar su comprensión.

- El primer que se presentará es el de una persona que no ha operado, tiene 22 años, sexo masculino, con menos de 24 meses de antigüedad laboral (categoría que se toma como referencia), es empleado fijo privado por lo que pertenece a la categoría ocupación A (categoría de referencia), en el Clearing pertenece a la categoría Amarillo A2, el crédito fue solicitado en 15 cuotas y el ratio de valor cuota sobre total de ingreso es de $0,18$.
  Ver cuadro 3.57

Cuadro 3.57: Primer ejemplo práctico.

	$\widehat{\beta}_i$	x_i	$\widehat{\beta}_i * x_i$
	-0,14	1	-0,14
Cant. veces que operó	-2,57	0	0
Edad	-0,013	22	-0,28
SexoM	0,22	1	0,22
Antigüedad>60 o JP	-0,87	0	0
Antigüedad 25-48	-0,30	0	0
Antigüedad 49-60	-0,44	0	0
Clearing AMARILLO A2	1,84	1	1,84
Clearing AMARILLO A3	0,78	0	0
Clearing AMARILLO M y ROJO	1,05	0	0
Clearing VERDE y LC	-0,06	0	0
Ocu5R	0,20	0	0
Ocu5V	-0,001	0	0
Cuotas totales	0,07	15	1,06
Valor cuota/TotIng	2,52	0,18	0,45
$\sum_i x_i \widehat{\beta}_i$			3,15
$\exp\left(\sum_i x_i \widehat{\beta}_i\right)$			23,25
$\frac{\exp\left(\sum_i x_i \widehat{\beta}_i\right)}{\left(1+\exp\left(\sum_i X_i \widehat{\beta}_i\right)\right)}$			0,99

Este individuo tiene un puntaje de 0,99. En la siguiente sección se indicará como clasificar a cada puntaje, que va de 0 a 999.

- Luego se estudiará una persona que ya operó en la empresa una vez, tiene 59 años, sexo femenino, con una antigüedad laboral de entre 49-60 meses, la calificación en el Clearing es Verde, es empleado fijo privado, el crédito fue otorgado en 10 cuotas y tiene un ratio de $0,18$.
  Ver cuadro 3.58

	$\widehat{\beta}_i$	x_i	$\widehat{\beta}_i * x_i$
	-0,14	1	-0,14
Cant. veces que operó	-2,57	1	-2,57
Edad	-0,013	59	-0,76
SexoM	0,22	0	0
Antigüedad>60 o JP	-0,87	0	0
Antigüedad 25-48	-0,30	0	0
Antigüedad 49-60	-0,44	1	-0,44
Clearing AMARILLO A2	1,84	0	0
Clearing AMARILLO A3	0,78	0	0
Clearing AMARILLO M y ROJO	1,05	0	0
Clearing VERDE y LC	-0,06	1	-0,06
Ocu5R	0,20	0	0
Ocu5V	-0,001	0	0
Cuotas totales	0,07	10	0,71
Valor cuota/TotIng	2,52	0,18	0,45 8
$\sum_i x_i \widehat{\beta}_i$			-2,81
$\exp\left(\sum_i x_i \widehat{\beta}_i\right)$			0,06
$\frac{\exp\left(\sum_i x_i \widehat{\beta}_i\right)}{\left(1+\exp\left(\sum_i X_i \widehat{\beta}_i\right)\right)}$			0,057

Cuadro 3.58: Segundo ejemplo práctico.

Este individuo tiene un puntaje de 0,057. En la siguiente sección se indicará como clasificar a cada puntaje, que va de 0 a 999.

3.4.4. Dictamen del Score

El modelo de regresión, para cada individuo, estima un valor comprendido entre $(0, 1)$ de acuerdo a la probabilidad de ser moroso o no. Como el objetivo de la empresa es conseguir un score que provea un puntaje de 0 a 999 se realizará un cambio de escala. Además se especificarán los umbrales de tal forma de poder clasificar al cliente en Rojo, Amarillo o Verde.

Según como se especificó en secciones anteriores el umbral "óptimo" que determinaba si un cliente tenía perfil moroso o no, era de $0,08$. Si el resultado del score es mayor a $0,08$ entonces el perfil del cliente será $Malo$ o calificado

como *Rojo* y si es menor entonces el perfil será *Bueno* o $Verde$.

El objetivo de obtener una escala entre 000 y 999 es poder brindarle mayor claridad a los analistas a la hora de interpretar el resultado. Entonces, luego de obtener la probabilidad de ser moroso para cada individuo se deberá realizar la siguiente operación:

$$R(x) = (1 - \hat{S}(x)) * 1000 \tag{3.8}$$

Como la empresa quiere tener tres escalas de medición *Rojo* ($Malo$), *Amarillo* (*Dudoso*) y $Verde$ (*Bueno*), además de obtener el puntaje correspondiente, a continuación se procede a determinar cuáles serán los umbrales para cada caso.

Para la determinación de las franjas se contó con el apoyo del tutor de la empresa, el contador Martin Rivero, que nos brindó casos que consideraba que tenían un perfil marcado tanto por ser buenos, malos o dudosos en relación al comportamiento en el pago del crédito.

Luego del estudio de las probabilidades estimadas para cada caso, se propuso la siguiente partición: si el valor del score es mayor a $0,30$ se calificará a la instancia como *Rojo*, si está entre $0,08$ y $0,30$ será *Amarilla* y sino será $Verde$.

Entonces para la nueva escala de medición se tiene:

Score	Calificación
≥ 920	VERDE
700−920	AMARILLO
<700	ROJO

Cuadro 3.59: Dictamen del Score.

3.5. Árboles de Regresión y Clasificación, CART

Para continuar con el análisis de modo de poder complementar la investigación realizada durante la pasantía, se decide evaluar otra técnica de análisis: Árboles de regresión y clasificación, CART.

El estudio se realiza en primera instancia considerando aquellas variables que podían llegar a influir en el comportamiento del cliente. En el proceso de construcción de los árboles de clasificación resultan ser "significativas" las mismas variables que fueron consideradas en los modelos de regresión logística. Por este motivo se decide considerar, en la construcción de los árboles, las variables: Cantidad de veces que operó, Sexo, Edad, Antigüedad Laboral, Clearing, Ocupación, Cuotas totales y Valor cuota sobre Total de Ingresos.

Para llevar a cabo este procedimiento, se estima un modelo con una muestra del 50 % del total de la población, comparándolo luego con el modelo de regresión logística. Por otro lado, se estima también un árbol de clasificación considerando una muestra que tuviese igual proporción de clientes clasificados como *Bueno* y clientes clasificados como $Malo$. Esto se realiza para poder compararlo con el procedimiento realizado en el modelo de regresión logística y para evaluar su comportamiento.

Esta técnica realiza particiones recursivas del espacio de las variables a partir de ciertas reglas de decisión. Como son particiones encajadas se puede llegar a una partición total de tal forma que en cada nodo quede una sola observación. Sin embargo, ésta no sería una buena decisión porque se estaría sobre ajustando, por lo que se debe buscar el corte "óptimo".

Por otro lado, tampoco sería una buena decisión quedarse con un árbol muy pequeño ya que no sólo no captaría la estructura de los datos sino que seguramente los errores de predicción serían más elevados que los esperados.

Las reglas de clasificación se pueden basar en la tasa de errores de clasificación. Sin embargo, esta tasa siempre se reducirá (con cada división. Esto no significa sin embargo que la tasa de error de predicción final vaya a mejorar.

Una de las soluciones a este problema es la validación cruzada. Las estimaciones y los ajustes se realizan mediante los comandos *rpart* [Therneau et al., 2014], *printcp* [Kuehnapfel, 2014] y *plotcp* [Kuehnapfel, 2014], paquetes específicos del software *R* [R Core Team, 2014].

En conclusión, lo que se realizará en primera instancia es la construcción del árbol completo utilizando el comando *rpart*.

Luego se validará usando el *parámetro de complejidad* (cp) y el error de validación cruzada con el comando *printcp*.

La función *printcp* proporciona una tabla con los valores de *cp*, el número de divisiones, el error relativo, el error de validación cruzada y el error estándar.

El comando *plotcp* provee una representación gráfica del error de validación cruzada estándar, los valores de *cp* y el número de particiones.

Para realizar la poda adecuada, de tal forma de evitar cualquier sobreajuste, se utilizará el valor *cp* y se elegirá el que tiene menor valor de error de validación cruzada. El *parámetro de complejidad* no es el error en un nodo particular sino que representa el valor en la mejora del error relativo cuando se divide ese nodo. Una medida de consulta es $R_{cp}(T)$, el costo complejidad del árbol T [cp].

$$R_{\beta}(T) = R(T) + \beta * O \tag{3.9}$$

- R(T) es el error de clasificación asociado al árbol T.
- β, ($\beta \geq 0$, parámetros de complejidad) se interpreta como el costo de complejidad por nodo terminal.
- O es el número de nodos terminales.

Si $\beta = 0$, el costo de complejidad alcanza su máximo para el árbol más largo posible. Cuando los valores de β decrecen y se aproximan a cero, los árboles minimizan el costo de complejidad.

Un valor de $cp = 1$ es de un árbol sin particiones, siendo cero cuando se tiene un árbol completo. La interpretación es muy sencilla, cuando al realizar una partición el error de validación cruzada del modelo no aumenta entonces no vale la pena realizar esa partición, ya que aumentará la complejidad del árbol sin tener mejoras reales.

Se observa que $R_{cp}(T)$ es una combinación lineal entre el error o costo del árbol y su complejidad (tamaño). Donde *cp* es la penalidad por nodo

terminal adicional. Cuando los valores de *cp* decrecen y se aproximan a cero, los árboles minimizan el costo de complejidad.

Otra medida de presión de la clasificación es:

$$Precision = \frac{n_{NN} + n_{PP}}{n} = \frac{t_{NN} + t_{PP}}{2} \tag{3.10}$$

1. **Árbol de clasificación, muestra del 50 % de la población.**

   Para realizar la poda adecuada, de tal forma de evitar cualquier sobreajuste, se utilizará la $MedidadeCP$ y se elegirá el que tiene menor valor.

CP	**Nº de particiones**	**Error relativo**	**Error de validación cruzada**	**Error std.**	**Medida de CP**
0,1132	0	1,0000	1,0000	0,0089	1,0000
0,0053	1	0,8868	0,8868	0,0084	0,8921
0,0027	3	0,8761	0,8767	0,0083	0,8849
0,0005	**6**	**0,8707**	**0,8712**	**0,0083**	**0,8740**
0,0005	13	0,8655	0,8728	0,0083	0,8788
0,0004	16	0,8641	0,8724	0,0083	0,8793
⋮	⋮	⋮	⋮	⋮	⋮
0,0000	628	0,8001	0,9534	0,0086	0,9611
0,0000	649	0,7999	0,9539	0,0086	0,9604

Cuadro 3.60: Costo complejidad CART, muestra 50 % de la población.

Tomando en cuenta la menor medida de costo complejidad el árbol que se debería considerar es aquél con 6 particiones. Su representación es la siguiente:

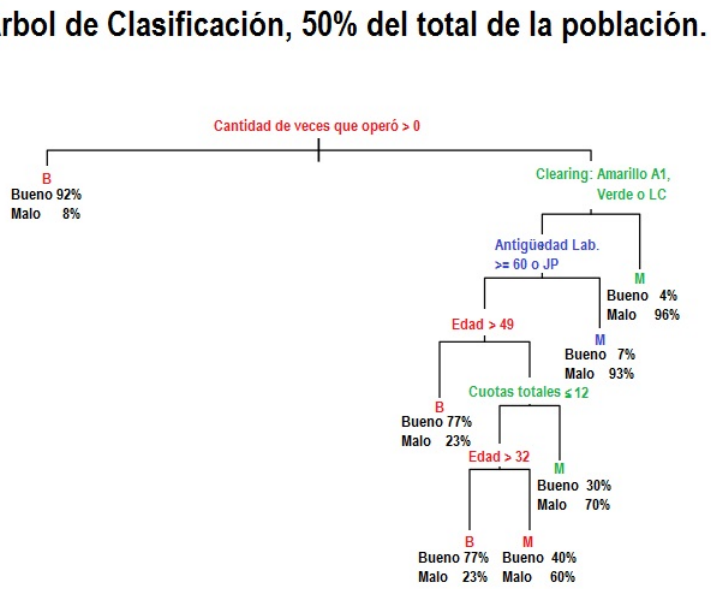

Figura 3.15: Árbol de clasificación podado, muestra 50 % de la población.

Como se puede observar en uno de los nodos terminales quedan agrupados todos aquellos clientes que ya han operado en la empresa más de una vez sin importar sus otras características, estos representean el 98 % del total de las observacioens de la muestra. Si bien, el 92 % de estos clientes son clasificados a priori como *Bueno* se busca poder obtener un árbol que capte con más detalle el comportamento de estos. Es decir se busca poder encontrar cómo es el perfil de aquellos clientes que habiendo operado más de una vez en la empresa a priori fueron clasificados como *Malo*.

Por esta razón se decide construir el árbol cuyo número de particiones es 13.

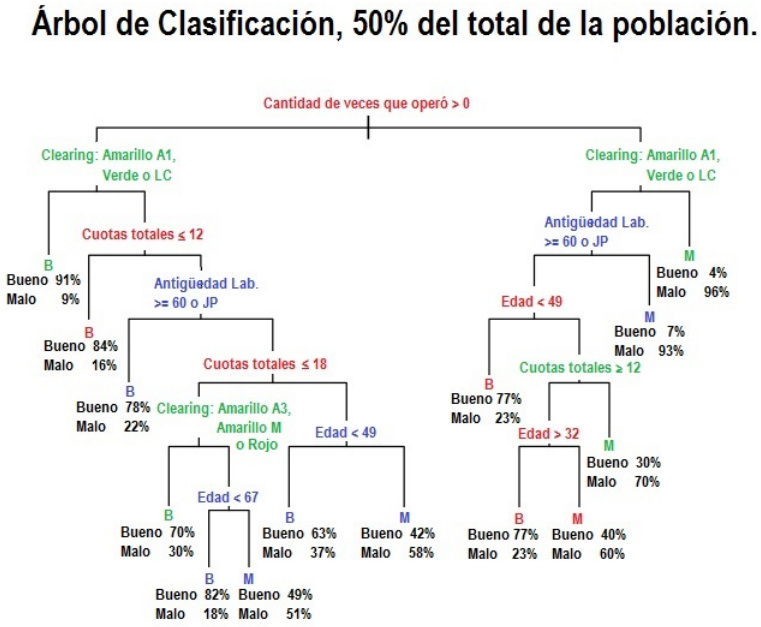

Figura 3.16: Árbol de clasificación podado, muestra 50 % de la población.

Habiendo logrado un árbol con más nodos terminales a continuación se los caracteriza siguiendo el camino de condiciones establecidas sobre los datos.

Primero que nada se debe tener en cuenta la cantidad de veces que el cliente operó. Si no ha operado en la empresa y tiene una clasificación en el Clearing *Amarillo A2*, *Amarillo A3*, *Amarillo M* o *Rojo*, es clasificado como *Malo* con una probabilidad de 96 %.

Si no ha operado en la empresa, tiene una calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito* y su *Antigüedad Laboral* es menor a 60 meses es calificado como *Malo* con un 93 % de probabilidad.

Si no ha operado en la empresa, tiene una calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito*, su *Antigüedad Laboral* es

mayor a 60 meses, es Jubilado o es Pensionista y tiene más de 49 años es calificado como *Bueno* con un 77 % de probabilidad.

Si no ha operado en la empresa, tiene una calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito*, su *Antigüedad Laboral* es mayor a 60 meses, es Jubilado o es Pensionista, es menor de 49 años y el préstamo fue solicitado en más de 12 cuotas es clasificado como *Malo* con un 70 % de probabilidad.

Si se da la misma situación anterior pero fue solicitado en menos de 12 cuotas, ésta será clasificada como *Bueno* con un 77 % de probabilidad si el cliente es mayor a 32 años y como *Malo* con un 60 % si es menor .

Luego nos fijamos en los clientes que ya han operado en la empresa.

Si tiene una calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito* es *Bueno* con una probabilidad de 91 %.

Si tiene una clasificación en el Clearing *Amarillo A2*, *Amarillo A3*, *Amarillo M* o *Rojo* y las cuotas totales son menores a 12 entonces es clasificado como *Bueno* con un 84 % de probabilidad.

Si tiene una clasificación en el Clearing *Amarillo A2*, *Amarillo A3*, *Amarillo M* o *Rojo*, las cuotas totales son mayores a 12 y su *Antigüedad Laboral* es mayor a 60 meses, es Jubilado o es Pensionista es *Bueno* con un 78 % de probabilidad.

Si tiene una clasificación en el Clearing *Amarillo A2*, *Amarillo A3*, *Amarillo M* o *Rojo*, las cuotas totales son mayores a 12 y su *Antigüedad Laboral* es menor a 60 meses entonces se deben tener en cuenta otros factores, tal como se puede apreciar en el gráfico anterior.

Luego de estimado el árbol se estudian los errores de predicción.

		Predicción	
		Bueno	*Malo*
Observados	*Bueno*	93 %	7 %
	Malo	34 %	66 %

Cuadro 3.61: Errores de clasificación CART, muestra 50 % de la población.

La medida de precisión de clasificación es de 93 %, es decir un 93 % de las observaciones están bien clasificadas, estos valores se daban también para el anterior árbol.

Para poder comparar con el modelo de regresión logística estimado, utilizando la misma muestra, a continuación se muestran los errores de predicción.

		Predicción	
		Bueno	*Malo*
Observados	*Bueno*	68 %	32 %
	Malo	30 %	70 %

Cuadro 3.62: Errores de clasificación regresión logística Modelo 2, muestra 50 % de la población.

En este árbol es posible captar el comportamiento de todos los perfiles de clientes, destacándose que la variable más importante es la cantidad de veces que operó y luego el clearing. Estas variables también eran de gran importancia en el modelo de regresión logística.

2. **Árbol de clsificación, muestra igual proporción de** *Bueno* **y** *Malo***.**

   Al igual que en el caso anterior para realizar la poda adecuada se utilizará la *Medida de CP* y se elegirá el que tiene menor valor.

CP	**Nº de particiones**	**Error relativo**	**Error de validación cruzada**	**Error std.**	**Medida de CP**
0,2655	0	1,0000	1,0100	0,0049	1,0100
0,0864	1	0,7345	0,7345	0,0047	0,8209
0,0171	2	0,6480	0,6480	0,0046	0,6822
0,0060	3	0,6309	0,6309	0,0045	0,6489
0,0010	**6**	**0,6087**	**0,6122**	**0,0045**	**0,6180**
0,0010	8	0,6067	0,6151	0,0045	0,6228
0,0009	12	0,6028	0,6154	0,0045	0,6265
⋮	⋮	⋮	⋮	⋮	⋮
0,0001	321	0,5324	0,6213	0,0045	0,6534

Cuadro 3.63: Errores de validación cruzada CART, muestra igual proporción de Bueno y Malo.

En este caso la menor medida de CP se da para el siguiente árbol con 6 particiones:

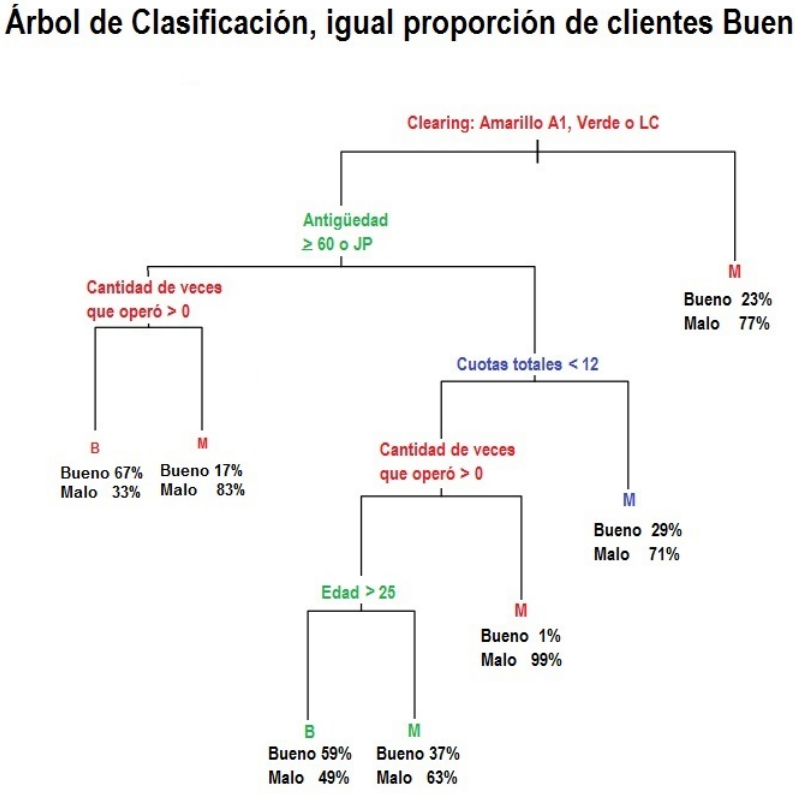

Figura 3.17: Árbol de clasificación podado, muestra igual proporción de *Bueno* y *Malo*.

Luego de obtener el árbol de clasificación se caracterizarán los nodos terminales siguiendo el camino de condiciones establecidas sobre los datos.

En un principio se debe tener en cuenta la clasificación del cliente en el Clearing, si es *Amarillo A2*, *Amarillo A3*, *Amarillo M* o *Rojo* entonces es clasificado como *Malo* con una probabilidad de 77 %.

De lo contrario si tiene calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito* entonces se debe tener en cuenta otras ca-

racterísticas. Si además, tiene una *Antigüedad laboral* mayor o igual a 60 meses, es Jubilado o es Pensionista y operó más de una vez en la empresa entonces es clasificado como *Bueno* con una probabilidad de 67 %. En caso contrario es clasificado como $Malo$ con un 83 % de probabilidad.

Si tiene calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito*, si su *Antigüedad laboral* es menor a 60 meses y las cuotas totales son mayores a 13, entonces es clsificado como $Malo$ con un 68 % de probabilidad.

Si tiene calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito*, si su *Antigüedad laboral* es menor a 60 meses y las cuotas totales son mayores a 12, entonces es clasificado como $Malo$ con un 71 % de probabilidad.

Si tiene calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito*, si su *Antigüedad laboral* es menor a 60 meses, el préstamo es solicitado en menos de 12 cuotas y no ha operado en la empresa, entonces es clasificado como $Malo$ con un 99 % de probabilidad.

Si tiene calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito*, si su *Antigüedad laboral* es menor a 60 meses, el préstamo es solicitado en menos de 12 cuotas y ya ha operado en la empresa y es mayor a 25 años, entonces es clasificado como *Bueno* con un 59 % de probabilidad.

Si tiene calificación en el Clearing *Amarillo A1*, *Verde* o tiene *Línea de Crédito*, si su *Antigüedad laboral* es menor a 60 meses, el préstamo es solicitado en menos de 12 cuotas y ya ha operado en la empresa y es menor a 25 años, entonces es clasificado como $Malo$ con un 63 % de probabilidad.

El resultado global de los errores de predicción es:

		Predicción	
		Bueno	*Malo*
Observados	*Bueno*	75 %	25 %
	Malo	24 %	76 %

Cuadro 3.64: Errores de clasificación CART, muestra igual proporción de *Bueno* y *Malo*.

En este caso la medida de precisión de clasificación es de 75 %, es decir un 75 % de las observaciones están bien clasificadas. Al compararlo con el modelo de regresión logística los resultados son parecidos, CART mejora un poco las predicciones de los clientes clasificados como *Malo*.

		Predicción	
		Bueno	*Malo*
Observados	*Bueno*	78 %	22 %
	Malo	37 %	63 %

Cuadro 3.65: Errores de clasificación regresión logística Modelo 3, muestra igual proporción de *Bueno* y *Malo*.

Dejando de lado el árbol con la muestra equilibrada, cabe destacar que esta metodología proporciona estimaciones similares al modelo estimado a través de la regresión logística. Esto es importante ya que reafirma los resultados obtenidos en la metodología anterior, puede servir como guía para la interpretación del comportamiento de los diferentes perfiles de clientes. Cuando se caracterizan los nodos siguiendo las condiciones establecidas en cada partición en ambos casos, las decisiones son coherentes.

Sucede lo mismo con el árbol de clasificación realizado con la muestra con igual proporción de *Bueno* y *Malo*, sin embargo esta muestra no es representativa de la realidad.

Capítulo 4

Conclusiones y Recomendaciones

4.1. Conclusiones

Durante el transcurso de la pasantía se ha logrado cumplir con los objetivos planteados en tiempo y forma. Se ha logrado construir un modelo de Scoring Crediticio alternativo al implementado en la empresa, tratando de que sea parsimonioso y prediga de la mejor manera. También se estudiaron los árboles de regresión y clasificación, una técnica nueva para las pasantes.

A su vez, se ha logrado experimentar las dificultades provenientes de enfrentarse con datos reales, sumando sus dificultades operacionales, adecuándose a los requerimientos de la empresa.

Regresión Logística

Para poder lograr obtener un modelo de regresión logística adecuado, se realizaron varias pruebas y análisis.

En un principio se realizan las estimaciones en base a una muestra del 90 % de la población y luego con el 50 %, con el fin de contar con más datos de prueba para evaluar el desempeño del modelo .

Cómo los clientes calificados como $Malo$ no llegan a ser el 10 % del total de la población se decidió, para explorar la técnica, tomar una muestra en la que la proporción de clientes $Malo$ fuese igual a la de $Bueno$.

En la búsqueda de mejorar los resultados, como se observaba que los clientes con categoría ocupacional *Activos* tenían un perfil diferente a los *Pasivos* se decidió considerar un modelo diferente para cada uno de ellos tomando las respectivas muestras al 50 %.

Lo mismo se realizó con los clientes que ya habían operado en la empresa más de una vez y con los que era su primera operación, se estimó un modelo para cada perfil.

Luego de probar varias alternativas se decidió que el más adecuado era el que se estimó con una muestra del 50 % incluyendo las siguientes variables: *Cantidad de veces que operó, Edad, Sexo, Antigüedad Laboral, Clearing, Ocupación, Cuotas totales, Valor cuota / Total de Ingresos.*

Alguno de los parámetros significativos son *cantidad de veces que operó* y *Ratio valor cuota / total de ingresos*, el primero tiene un impacto positivo, ya que al aumentar una unidad disminuye la probabilidad de ser malo; en cambio el segundo tiene un impacto negativo, ya que al aumentar su valor crece la probabilidad de ser malo.

En particular el modelo elegido tiene una especificidad de un 81 %, indica la capacidad del estimador para predecir que el cliente es *Bueno* dado que realmente lo es. Y una sensibilidad de 59 % que indica la capacidad del estimador para clasificar a los clientes como *Malo* dado que realmente esa era su clasificación.

Se observa que donde se maximiza la distancia, es en el punto de corte: $0,08$, que además coincide con el criterio utilizado para la elección del punto de corte cuando los costos de clasificar en uno u otro grupo son iguales. La distancia máxima del estadístico K-S es: $0,41$.
Para las predicciones futuras se decide tomar tres franjas para el dictamen del score.

Estas son:

- conceder con seguridad pues el puntaje es superior o igual a 920.
- rechazar con seguridad dado que el puntaje es inferior a 720.
- y el tercer estado que se ha establecido como dudoso, en el cual se aconseja su estudio más cauteloso por parte del analista.

En el análisis del Modelo de Regresión Logística no es tan sencillo determinar de forma directa un perfil particular de un cliente *Bueno* o *Malo*. Sin embargo se puede definir a modo de ejemplo qué características debe cumplir la persona con un perfil "idealmente" bueno.

Por ejemplo, haber *operado* la mayor cantidad de veces posible, tener la mayor edad permitida, sexo femenino, una *antigüedad laboral mayor a 60 meses*, ser *jubilado o pensionista*, tener una calificación en el Clearing *Verde* o tener una *Línea de Crédito*, ocupación *Jubilado, Pensionista o Empleado Fijo Público*, solicitar el crédito con la menor *cantidad de cuotas* posible y tener el menor ratio posible de *valor cuota / total de ingresos*. Por el contrario un ejemplo concreto de persona con un perfil malo sería: operar por primera vez, tener 18 años, sexo *masculino*, antigüedad laboral menor a 25 meses, tener calificación en el Clearing como *Amarillo A2*, ser un *trabajador independiente*, cantidad de *cuotas* igual a 24, ratio *valor cuota sobre total de ingresos* igual a $0,4$ (máximo aceptable).

Árboles de Clasificación

Los resultados obtenidos en los árboles de clasificación son consistentes con el Modelo de Regresión Logística. Esta metodología logró modelos con buenas predicciones.

En el proceso de construcción de todos los arboles de clasificación estimados resultan ser "significativas" las mismas variables que fueron consideradas en el modelo final de regresión logística.

Para llevar a cabo el procedimiento se realiza, en primera instancia, un árbol de clasificación considerando una muestra aleatoria simple del 50 % de las observaciones. Luego se realiza otro con una muestra que tuviese igual proporción de clientes clasificados como *Bueno* y clientes clasificados como *Malo*.

En ambos casos se obtuvieron los arboles completos, luego se evaluó cuál era la poda más adecuada utilizando la medida CP y el error de validación cruzada.

Al estimar el árbol de clasificación con la muestra del 50 % de la población las predicciones son buenas pues se tiene una precisión del 79 %, se logra captar el comportamiento de los diferentes perfiles de clientes.

Dejando de lado el árbol con la muestra equilibrada, cabe destacar que esta metodología proporciona estimaciones similares al modelo de regresión logística. Esto es importante ya que reafirma los resultados obtenidos en la metodología anterior, puede ser un buen complemento debido a su fácil interpretación.

Sucede lo mismo con el árbol de clasificación realizado con la muestra con igual proporción de *Bueno* y *Malo*, sin embargo esta muestra no es representativa de la realidad.

Hay que tener en cuenta que se está evaluando la probabilidad de que un cliente caiga en mora o no luego de haber solicitado el préstamo. El comportamiento de las personas es difícil de predecir, por lo que puede haber muchas excepciones. Por ejemplo, un cliente puede tener el peor perfil y sin embargo haber pagado todas sus cuotas en tiempo y forma, como puede suceder lo contrario. También existen factores externos que no se pueden controlar y que pueden incidir en el cumplimiento de sus derechos.

En este sentido, se considera que modelos con este nivel de error de predicción, son lógicos para ser utilizado en la práctica por una empresa crediticia.

En concreto, el trabajo se orientó a estimar la probabilidad de incumplimiento de pago de un cliente en función de una serie de características, utilizando la metodología del Credit Scoring (este método se emplea mayormente para evaluar individuos y, pequeñas y medianas empresas). Una buena aproximación de estas probabilidades resulta muy importante para que la empresa reduzca sus pérdidas por morosidad o que el proceso de análisis por parte de los analistas se vea facilitado.

4.2. Recomendaciones

En primera instancia, cabe destacar que para implementar un modelo estadístico es necesario definir con claridad la variable de respuesta. Para este caso en particular, a priori no deben haber dudas sobre cuándo un cliente es clasificado como *Bueno* o *Malo*, no pueden existir ambigüedades al respecto.

Luego de haber culminado el ajuste del modelo se entedió que otra alternativa pordría haber sido considerar, desde un principio, un modelo de regresión logística multinomial. Considerar la variable dependiente con sus

tres categorías ($Bueno$,$Indiferente$ o $Malo$), en vez de dicotómica. En ese caso se hubiese utlizado toda la base completa obteniedo quizás un resultado diferente para la franja "Amarilla", los $indiferentes$ en este caso.

Por otra parte también podría haber sido una solución utilizar CART como análisis preliminar para la selección de las variable a incluir en el modelo de regresión logística.

Dado que tanto las situaciones de las personas, como de la economía y del mundo en general, están en constante cambio, se recomienda evaluarlo periódicamente para determinar si es necesario realizar ajustes a los nuevos datos recabados para poder corregir los desvíos.

Bibliografía

[Altman et al., 1985] Altman, E; Kao, D.; Frydman, H. Introducing recursive partitioning for financial classification: The case of financial distress. The Journal of Finance, v. XL, n. 1, p. 269-291, 1985.

[Blanco, 2006] Blanco, Jorge. Introducción al Análisis Multivariado. Instituto de Estadística, Montevideo, Uruguay, 2006.

[Blöchlinger y Leippold, 2006] Blöchlinger, A. y Leippold, M. Economic benefit of powerful credit scoring. Journal of Banking and Finance, 30, pag.: 851-873, 2006.

[Breiman, 1994] Breiman, Leo. Bagging Predictors. Technical Report No. 421. Statistics Department. University of California.Berkeley, California 94720. September 1994.

[Castor et al., 2011] Castor Guisande González, Antonio; Vaamonde, Lise; Barreiro Felpeto, Aldo. Tratamiento de datos con R, Statistica y SPSS. Editorial Díaz de Santos, S.A., 2011.

[Costa et al., 2012] Teresa Costa Cor, Eva Boj del Val y Fortiana Gregori, José. Bondad de Ajuste y Elección del punto de corte en regresión logística basada en distancias. Anales del Instituto de Actuarios Españoles, 2012.

[cp] *http://cran.r−project.org/web/packages/rpart/vignettes/longintro.pdf*

[DeVeaux et al., 2014] DeVeaux, Richard; Fienberg, Stephen E.; Olkin, Ingram. Springer Texts in Statistics. Science and Busines Media New YOrk, 2014.

[Fluss et al., 2005] Fluss, R., Faraggi, D. y Reiser, B. Estimation of the Youden Index and its associated cutoff point. Biometrical Journal, 47, pag. 458-472,2005.

[Hand et al., 1997] Hand, D.J y W.E. "Statistical Classification Methods in Consumer Credit Scoring: A Review". Journal of the Royal Statistical Society; Serie A. Henley. 1997.

[Hastie et al., 2009] Hastie, Trevor; Tibshirani, Robert y Friedman, Jerome. The Element of Statistical Learing: Data Mining, Inference and Prediction. Second Edition, Springer. February 2009.

[Hosmer y Lemeshow, 2013] Hosmer W., David y Lemeshow, Stanley. Applied Logistic Regression. John Wiley & Sons, Inc. Second Edition, 2013.

[Iglesias, 2013.] Iglesias Cabo,Tania. Métodos de Bondad de Ajuste en Regresión Logística. Máster Oficial en Estadística Aplicada. Trabajo Fin de Máster. Curso académico 2012/2013.

[J. A. Nelder y R. W. M., 1972] J. A. Nelder y R. W. M. Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), Vol. 135, 1972.

[James et al., 2013] James, Gareth; Witten, Daniela; Hastie, Trevor y Tibshirani, Robert. An Introduction to Statistical Learning with Applications in R. Springer Science. Business Media New York, 2013 (Corrected at 4 printing 2014).

[Krzanowski y Hand, 2009.] Krzanowski, Wojtek J. y Hand, David J. ROC Curves for Continuous Data. Chapman & Hall/CRC by Taylor and Francis Group, LLC, 2009.

[Kuehnapfel, 2014] Kuehnapfel, Andreas. CP: Conditional Power Calculations. R package version 1.5. $http://CRAN.R-project.org/package=CP$. (2014)

[Marais et al., 1984] Marais, M.L; Patell, J; Wolfson, M. The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications. Journal of Accounting Research, v. 22, n. 1, p. 87-114, 1984.

[Mesa, 2014] Mesa, Andrea. Notas del Curso de Aprendizaje Automático. Facultad de Ciencias Económicas y de Administración, Montevideo, Uruguay, 2014.

[Nieto,2010] Nieto Murillo, Soraida. Proyecto de Tesis: Crédito al Consumo. La Estadística aplicada a un problema de Riesgo Crediticio. Universidad Autónoma Metropolitana, 2010.

[R Core Team, 2014] R Core Team . R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL *http*: *//wwwR − project.org/*. (2014)

[RAE, 2013] *Revista de Administração de Empresas*, Análisis del credit scoring. Print version ISSN 0034-7590. Rev. Adm. Empres. vol.53 no.3 São Paulo May/June 2013.

[Reyes, 2007] Reyes Samaniego, Medin. "El riesgo de crédito en el marco del acuerdo de Basilea II" Delta Publicaciones, 2007.

[Ripley, 2014] Ripley, Brian. tree: Classification and regression trees. R package version 1.0-35. *http*: *//CRAN.R − project.org/package = tree*. (2014)

[Rivero, 2012] Rivero, Martin. Notas internas del Contador Martín Rivero. Se realizaron en base a una Certificación Internacional en Riesgos y Seguros de la Asociación Latinoamericana de Administradores de Riesgo y Seguros, y de una Maestría en Gestión de Riesgos que realizó en la Facultad de Francisco de Vitoria de España, 2012.

[Robin et al., 2011.] Robin, Xavier; Turck, Natacha; Hainard, Alexandre; Tiberti, Natalia; Lisacek, Frédérique; Sanchez, Jean-Charles y Müller; Markus. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 *http*: *//www.biomedcentral.com/*1471 − 2105/12/77/. (2011)

[Schreiner, 2002.] Schreiner, Mark. Ventajas y Desventajas del Scoring Estadístico para las Microfinanzas. Microfinance Risk Management. 6970 Chippewa St. 1W, St. Louis, MO 63109-3060, U.S.A., 2002.

[Therneau et al., 2014] Terry Therneau, Beth Atkinson y Brian Ripley. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-8. *http*: *//CRAN.R − project.org/package = rpart*. (2014)

Capítulo 5

Anexo A

5.1. Descripción de las Actividades Realizadas.

SEMANA Nº1 (5/5-9/5).

Se conoce el ambiente laboral en donde se asigna un lugar de trabajo y una computadora a cada pasante. Se informa de manera resumida algunas de las funciones de la empresa.

Se realizan dos capacitaciones *Capacitación en Riesgos crediticios* y *Manejo del riesgo Crediticio y variables determinantes para el análisis*, a cargo del gerente del Sector.

SEMANA Nº2 (12/5-16/5).

Se realiza una tercera capacitación donde se enseña el funcionamiento del programa *BanTotal* desde el ingreso de datos en las sucursales, mostrando los diferentes análisis crediticios que se realizan en cada sector, a cargo de un integrante del sector Sistemas.

También se efectúan actividades prácticas en sector riesgo: manejo del sistema *BanTotal* en dicho sector y aprendizaje de su labor. Se culmina la semana con la realización de una evaluación escrita sobre la primera capacitación.

SEMANA Nº3 (19/5-23/5).

Se continúan las actividades en el sector riesgo y se realizan actividades prácticas con el sector recupero utilizando el sistema *BanTotal*, desde el

punto de vista de dicho sector, para entender su labor. Se realiza una matriz de riesgo sobre el trabajo que se realizará durante toda la pasantía, donde se evalúa y mide el riesgo de cada actividad.

El riesgo se mide teniendo en cuenta una evaluación de riesgo cualitativa de la empresa:

Cuadro 5.1: Clasificación del riesgo.

		Apetito de riesgo				
Impacto	Alto	M	A	A	A	A
	Medio alto	M	A	A	A	A
	Medio	B	M	M	A	A
	Medio bajo	B	B	M	M	M
	Bajo	B	B	B	M	M
		Bajo	Medio bajo	Medio	Medio alto	Alto
				Frecuencia		

SEMANA Nº 4 (26/5-30/5).

Se realiza una última Capacitación sobre *Central de Riesgos B.C.U.*, a cargo del Gerente del Sector y se lleva a cabo una prueba escrita sobre dicha capacitación.

Se solicita por escrito la base de datos al sector Sistemas, donde se especifica las variables que pueden llegar a ser útiles para el análisis. Se solicita la mayor cantidad de información posible sobre los créditos tanto otorgados como rechazados por la empresa.

POBLACIÓN OBJETIVO: personas físicas que hayan solicitado algún crédito al consumo, tanto rechazado como aceptado.

PERÍODO DE ESTUDIO: se considera pertinente la inclusión de los últimos 4 años. Cuanto más años se incluyan en el análisis preliminar se podrá observar mejor la evolución de los datos con el paso del tiempo y dependiendo de ésta la determinación de la cantidad de datos a utilizar.

VARIABLES: el estudio de todas las variables es primordial para determinar cuáles son significativas o no en el modelo, utilizando las herramientas estadísticas adecuadas.

A continuación se detallan algunas de ellas.

Persona		
Nº	**Variable**	**Observación**
1	TOTAL DE HABERES (nominal)	
2	INGRESO LÍQUIDO	
3	ANTICIPOS DE INGRESO	
4	OTROS INGRESOS	
5	CAPITAL SOLICITADO	
6	IMPORTE de CUOTA	
7	C.I.	
8	FECHA DE NACIMIENTO	
9	SEXO	
10	TELÉFONO FIJO	
11	CELULAR	
12	TELÉFONO LABORAL	
13	TELÉFONO ALTERNATIVO	
14	RUT	
15	FECHA INICIO LABORAL	
16	OCUPACIÓN	
17	DEPARTAMENTO PERSONA	
18	SITUACIÓN VIVIENDA	
19	CANTIDAD DE PERSONAS A CARGO	
20	ESTADO CIVIL	
21	CODEUDOR	Si la solicitud presentó codeudor o no.

Cuadro 5.2: Disponibilidad de Variables (1)

Sistema		
Nº	**Variable**	**Observación**
22	MÓDULO	
23	Nº CUENTA	
24	Nº OPERACIÓN	
25	TIPO DE RESOLUCIÓN	Si fue rechazado (incluye codeudor) o aceptado.

26	TIPO DE RECHAZO	Si el rechazo fue manual o automático.
27	MOTIVO DE RECHAZO	
28	TIPO DE FLUJO	Cual es el flujo (evaluación manual, producto básico, control, renovación o L.C.
29	ACTIVIDAD ECONÓMICA	Sector de actividad del solicitante
30	FECHA DE CONTABILIZADO	
31	FECHA DE DESEMBOLSO	
32	FECHA DE OTORGAMIENTO	
33	TASA DE INTERÉS	
34	FECHA ALTA LC	
35	MONTO OTOROGADO LC	
36	COMERCIALIZADORA	
37	DEPARTAMENTO DE LA AGENCIA	
38	SCORING	Resultado de franja
39	SCORE	puntaje
40	CAMPAÑA	
41	FECHA PRIMERA OPERACIÓN	Fecha
42	CASTIGO CORREGIDO	
43	ESTADO	
44	TRAMO 1	Cantidad de veces que pagó con menos de 6 días de atraso según Antecedentes Internos.
45	TRAMO 2	Cantidad de veces que pagó atrasado entre 6 y 29 días, según Antecedentes Internos.
46	TRAMO 3	Cantidad de veces que pagó atrasado entre 30 y 59 días, según Antecedentes Internos.
47	TRAMO 4	Cantidad de veces que pagó atrasado entre 60 y 89 días, según Antecedentes Internos.
48	TRAMO 5	Cantidad de veces que pagó atrasado entre 90 y 119 días, según Antecedentes Internos.

49	TRAMO 6	Cantidad de veces que pagó con más de 119 días de atraso según Antecedentes Internos.
50	SALDO ACTUAL	De aquellas personas con operaciones vivas
51	CUOTAS PAGAS	
52	DIAS DE MORA	De aquellas personas con operaciones vivas.
54	EXPERTO CORREGIDO	Resultado del Clearing.

Cuadro 5.3: Disponibilidad de Variables (2)

SEMANA Nº 5 (2/6-6/6).

Mientras se espera la base de datos para la realización del modelo, se analiza otra base pequeña de la empresa, con algunas de las variables para poder ir familiarizándonos con la misma. Se realizan algunos análisis estadísticos descriptivos básicos.

Al culminar la semana se realiza una prueba escrita de la capacitación *Manejo del riesgo Crediticio y variables determinantes para el análisis.*

SEMANA Nº 6 (9/6-13/6).

Se sigue estudiando la base de datos brindada por la empresa, construyendo un script en el programa R-project. Se realiza una lectura y análisis de la misma, de esta forma se adelanta trabajo, ya que éste va a servir cuando se tenga la base original.

Al culminar la semana se tiene una reunión con el tutor de la facultad Ramón Álvarez, el tutor de la empresa y la gerente de recursos humanos.

SEMANA Nº 7 (16/6-20/6).

Se nos entrega parte de la base (primer semestre de 2014), la cual se empieza analizar y se encuentran datos incoherentes y con errores. Esto conlleva a una reunión con los integrantes del sector Sistemas y se prosigue a corregir la misma por partes de éstos últimos.

SEMANA Nº 8 (23/6-27/6) A LA SEMANA Nº 13 (21/7-25/7).

La base de datos como herramienta dinámica ayuda en el desarrollo del modelo, resulta sólido en sí mismo. Cuanto mayor sea el volumen que componga la base de datos, mayor será la precisión del resultado final y por lo tanto a medida que se va disponiendo de más datos se van añadiendo a los actualmente disponibles.

Se entregó la base en 6 archivos, cada uno contiene un semestre, empezando desde el segundo semestre de 2011, hasta el primer semestre de 2014. Cada fila contiene una Instancia distinta (un número distinto para cada crédito) y las columnas contienen diferentes variables las cuales tienen diferentes codificaciones.

Como uno de los objetivos era calificar a cada *Instancia* en *Bueno*, $Indiferente$ o $Malo$(según los criterios brindados por la empresa) el primer filtro que se hace de la base es considerando todas aquellas instancias que tienen datos sobre sus días de mora y atraso (ya que es a partir de esas variables que se clasificará), las instancias que no tienen estos datos no sirven, por no poder clasificarlas.

Los ajustes a la base realizados para cada semestre son:

. Se calcula en el programa R-project la variable BYM (brinda clasificaciones: *Bueno*, $Indiferente$ y $Malo$).

. Se crea la variable *Edad*, esta es la edad del cliente al momento de realizar el préstamo (fecha valor - fecha de nacimiento).

. Se crea la variable $TotaldeIngresos$, calculada como: Ingresos Líquidos + Anticipos + Otros Ingresos.

. Se quitan algunas variables: *Saldo*, *Analista*, *Analista Control Documentario*, $Jefe$, *Gerente*, *Campaña*; debido a que se analizó junto con el jefe del área, que no iban a aportan información.

. Se agrupan los barrios por 18 zonas.

. Se recodifica la variable *Rut*, en vez de brindar el número de *Rut*, se la codifica como 1 o 0, si tienen o no Rut respectivamente.

. Se crea la variable LC (línea de crédito), 1 o 0, si tienen o no Línea de Crédito respectivamente.

. Se reagrupan en menos categorías las variables: Ocupación, Profesión y Actividad Económica. Esta agrupación se hace teniendo en cuenta criterios de reagrupamientos brindados por la empresa.

. Se crea la variable *Contactabilidad*, a partir de la información de teléfonos y celulares que se tienen en la base.

. Se recodifica la variable *Codeudor*, pasando a ser 1 o 0, si presenata o no codeudor.

. Se recodifican las variables *AntecedentesInternos*, *EstadoCivil* en variables con menos categorías.

Luego se unen todos los semestres en un mismo archivo y se realizan las siguientes modificaciones:

. Se crea la variable *Cantidad de Veces que Operó* como cliente (aparecerá una columna, en la cual tendrá para cada instancia, cuantas veces la persona correspondiente a esa instancia operó como cliente en la empresa)

. Se crea una variable ratio $CuotasPagas/CuotasTotales$

. Se crea la variable ratio $CuotasTotales/TotalIngresoLiquido$.

. Se eliminan las instancias que tienen *Importe* (valor del crédito) igual a 0 pues son errores de tipeo, corresponden únicamente a 29 casos.

. A la variable *GrupoFamiliar* (Cantidad de menores a cargo) se la recodifica, aquellos grupos familiares mayores o iguales a 20 se ponen en una sola categoría.

. En la variable *Antigüedad Laboral* los valores que son mayores o iguales a 600 meses (equivalente a 50 años) se los deja vacíos, algunos son incoherencias, y otros son jubilados o pensionistas (ya que se verifica filtrando la variable por ocupación).

. Valores que en algunas variables significan que no hay datos y están codificados con algún factor, se las deja vacías.

. Se quitan 23 individuos cuyos *CupoLC* (cupo a favor del cliente) era igual a 1, ya que se verificó con una base que tiene la empresa que estos habían cometido algún fraude, por lo tanto se prosigue a eliminarlos.

Luego se sigue filtrando la base, se sacan aquellas instancias clasificadas como $Indiferente$ ya que a la empresa le interesa sólo aquellos clasificados

como *bueno* o *malo*.

Por último se quitan aquellas instancias que habían tenido sólo una operación en su historia como cliente (Cantidad de Veces que Operó como cliente = 1) y a su vez que ésta sea menor a 12 meses (Cuotas Totales 12). Esto se hace para obtener una base con cierta "historia" crediticia y de este modo poder analizar mejor su comportamiento.

SEMANA Nº 14 (28/7-1/8).

Una vez lograda la base, se empieza a realizar el modelo de regresión logística en R-project incluyendo todas las variables. Surge un Warning message: "glm.fit: fitted probabilities numerically 0 or 1 occurred" y problemas con la significación de las variables. Debido a este último problema, se decide antes que nada realizar un estudio de las variables, así de éste modo poder descartar algunas que no sean significativas en el modelo.

SEMANA Nº 15 (4/8-8/8) y SEMANA Nº 16 (11/8-15/8).

Se prosigue a realizar un análisis a cada variable. El objetivo de estas semanas se centra en el estudio de las mismas, comenzando por realizar un análisis estadístico descriptivo, y los procedimientos que lo componen, como las medidas de tendencia central, medidas de distribución, frecuencias, las medidas de dispersión, gráficos, tablas etc.; para después analizar los datos con la finalidad de encontrar relaciones y tendencias que serían utilizadas en el planteamiento del modelo.

Se trata de averiguar de todas las variables implicadas en la determinación del score, cuáles tienen un mayor impacto en el no pago del crédito (variable dependiente), cuál es su comportamiento al variar sus valores, y cuáles en definitiva, tienen significación suficiente como para sustentar el modelo que se desea crear.

Al realizar estos análisis surge la necesidad de realizar pequeños arreglos:

. Se modifican algunos valores de Ingresos Líquidos y Total de Haberes que por ser demasiados grandes se los corrobora en el sistema $BanTotal$ y estaban mal ingresados en la base.

. Al modificar los ingresos líquidos antes mencionados, se tuvo que arreglar también Total de Ingresos y ratio Valor Cuota / Total Ingresos,

para esas instancias. De aquí, es que surgen algunas conclusiones de variables que quizás no sirven para el modelo porque no están discriminando bien la variable BYM.

SEMANA Nº 16 (11/8-15/8).

Se redacta parte del Informe Teórico entre otras cosas.

SEMANA Nº 17 (18/8-22/8).

Se realizó pruebas con diferentes modelos con el fin de encontrar el mejor.

SEMANA Nº 18 (26/8-29/8).

Se redactó todo lo hecho hasta el momento en el editor de texto Latex.

SEMANA Nº 19 (1/9-5/9).

Se decide no considerar la variable Antecedentes Internos, ya que no se tiene esta variable cuando se registra un nuevo crédito. Se podría considerar sólo cuando clasifiquemos a la población en "Préstamos por primera vez y Prestamos con historia".

Se realizó una reunión con Ramón Álvarez y Andres Castrillejo. Luego de la misma se realiza lo aconsejado por los docentes.

En cuento a Muestras se realiza:

- Una muestra considerando el 90 % de los Malos (el otro 10 % se usa para la muestra de prueba) y la misma cantidad de Buenos. Se corren los modelos antes realizados pero con esta nueva muestra, se analiza que el signo de los coeficientes de las variables sea coherente, se calculan los odds ratios y por último se verifican las predicciones con una muestra de prueba.
- Se realiza una muestra del 50 % de la población, y se corren los mismos modelos antes vistos. Se analizan los signos de los coeficientes de las variables y se calculan los odds ratios.

- Se realiza una muestra del 50 % de la población y se corren los mismos modelos antes vistos, se analiza que el signo de los coeficientes de las variables sean coherentes; se calculan los odds ratios y por último se verifican las predicciones con una muestra de prueba.

  Luego se considera la posibilidad de hacer más de un modelo para diferentes particiones de la población:

- Se prueba realizar distintos modelos clasificando a la población, por ejemplo en: Jubilados y Activos, Préstamo por primera vez y Préstamos con historia.

  Al realizar todas estos modelos con muestras distintas, las predicciones no mejoran comparando con las predicciones de los mismos modelos pero realizados con toda la población. A su vez se hicieron otros modelos distintos para estas clasificaciones pero las predicciones tampoco mejoran.

  Se considera realizar el modelo teniendo en cuenta la *Cuenta* (por cliente) en vez de la *Instancia* (por préstamo):

- Para poder resumir la información de cada préstamo de un mismo cliente se considera: Clasificar a las Cuentas como *Malo* si el cliente tuvo al menos un préstamo Malo. Y se toma toda la información del préstamo más actual.

  Al realizar esto, las predicciones son parecidas a lo anterior.

  Se realiza a su vez, otro análisis; se sacan varias muestras distintas del 10 % y se retienen los coeficientes de cada variable, luego se realizan las predicciones considerando las medias de las coeficientes obtenidos en las distintas muestras.

  El resultado obtenido también es el mismo.

  Dada la cantidad de análisis realizados y la obtención de las mismas predicciones, se sospecha que el problema está en las variables o combinación de variables consideradas en el modelo.

  Se decide reducir las modalidades de algunas variables como: Ocupación, Clearing.

De esta manera se vuelven a probar los distintos modelos.

SEMANA Nº 20 (8/9-12/9).

Reunión con el profesor Andrés Castrillejo: Se aconseja realizar nuevas definiciones de *Bueno* y *Malo* y probar los mejores modelos hasta el momento con esas nuevas definiciones.

También se aconsejó probar con otra técnica estadística: CART.

Durante el transcurso de esta semana, se realiza lo aconsejado por el Profesor Andrés, se prueban los modelos con 4 definiciones distintas. Los resultados de los modelos mejoran. Se empieza a estudiar e implementar CART.

SEMANA Nº 21 (15/9-19/9).

Reunión con Martín Rivero, se muestran los modelos realizados hasta el momento con las nuevas definiciones, se aconseja por parte del mismo sacar de los modelos la variable *Línea de Crédito* ya que estas no son analizadas por los Analistas, debido a que el sistema automáticamente las clasifica así por tener buen comportamiento.

Durante el transcurso de la semana, se prueban todos los modelos antes vistos sin considerar esta variable. Las predicciones empeoran ya que esta variable era la más significativa.

SEMANA Nº 22 (22/9-26/9).

Martín Rivero se encargó de averiguar, con otras empresas financieras que trabajan con modelos de Scoring, qué margen de error de predictibilidad se entiende aconsejable (o manejable) trabajar en los modelos de scoring.

Las respuestas fueron que, en general, se trabaja con el estadístico K-S que mide cuan bien discrimina el modelo.

Se considera que un modelo con un KS 20 no sirve, 35 es útil, 75 "estaría mintiendo".

Por lo tanto, dadas estas nuevas informaciones, se decide calcular el estadístico Kolmogórov-Smirnov a los modelos sin LC.

Se solicita la información de las instancias rechazadas por el sistema, instancias las cuales no se utilizaron para realizar el modelo por no tener ca-

lificación como *Bueno* o *Malo*, pero son importantes ya que estas son las que los analistas consideran que podrían llegar a tener un mal comportamiento.

SEMANA Nº 23 (29/9-3/10).

Se terminan los últimos arreglos del modelo que se considera más adecuado.

SEMANA Nº 25 (6/10-10/10).

Con la información de las instancias rechazadas, se decide correr el modelo creado, para ver cómo son las predicciones.

Se realiza una reunión con Martín y se escoge el mejor modelo.

Se realiza una reunión con los integrantes de Sistemas para explicar el modelo y que puedan de este modo aplicarlo.

SEMANA Nº 26 (13/10-17/10).

Se prueba, con más conocimiento del tema implementar el modelo CART, ya que se concurrió a clases de Análisis Multivariado II en el IESTA con Andrea Mesa como profesora titular.

SEMANA Nº 27 (20/10-24/10).

Se termina de realizar las últimas redacciones del Informe Final para la empresa y se sigue con la elaboración del informe que se presentará en el Instituto de Estadística.

SEMANA Nº 28 (27/10-31/10).

Se sigue con la tarea de la semana anterior.

Culmina la pasantía presentando el informe solicitado por la empresa.

Las semanas posteriores, se realizan reuniones con los tutores y se sigue con la elaboración del informe que se presentará en la facultad.

Capítulo 6

Anexo B

6.1. Análisis de las variables

En esta sección se muestra el análisis de aquellas variables que se consideran secundarias.

- **Antecedentes Internos.**

Antecedentes Internos	**Frecuencia Relativa**	**Frecuencia Acumulada**
sin dato	2	2
A	6	8
R	0	8
V	92	100

Cuadro 6.1: Frecuencia de la variable Antecedentes Internos

La mayoría de los clientes son clasificados como verdes según antecedentes internos. Esto es razonable, ya que a la hora de pedir un préstamo, se les miden los antecedentes internos, si estos son malos (rojos), no se les debería otorgar el préstamo, a no ser alguna excepción.

Antecedentes Internos según *Bueno* **o** *Malo*.

Antecedentes Internos	**Frecuencia Relativa** *Bueno*	**Frecuencia Relativa** *Malo*
Sin dato	1	17
A	6	11
R	0	0
V	93	72

Cuadro 6.2: Frecuencia de la variable Antecedentes Internos según las categorías *Bueno* y *Malo*

En el caso del grupo calificador como *Malo* disminuye el porcentaje de clientes con antecedentes internos verdes en relación al total de cada grupo. A su vez, hay un incremento tanto en los calificados amarillos como en los sin datos.

- **Actividad Económica.**

  La variable actividad económica tiene un comportamiento similar a la Profesión y la Ocupación.

- **Departamento de la Persona.**

  Departamento de la Persona según *Bueno* **o** *Malo***.**

Departamento de la Persona	**Frecuencia Relativa** *Bueno*	**Frecuencia Relativa** *Malo*
1	2	2
2	16	16
3	3	3
4	2	3
5	2	2
6	1	0
7	2	2
8	2	2
9	4	4
10	38	39
11	4	4
12	1	2
13	4	3
14	3	3
15	3	4
16	3	2
17	3	3
18	5	4
19	2	2

Cuadro 6.3: Frecuencia de la Variable Departamento de la Persona según las categorías *Bueno* y *Malo*.

- **Normativa**

Normativa	**Frecuencia Relativa**	**Frecuencia Acumulada**
N/A	61	61
ROJO	0	61
VERDE	399	100

Cuadro 6.4: Frecuencia de la Variable Normativa.

No se tiene información del 60 % de los datos de esta variable. De los que se tienen datos, tienen, en un gran porcentaje normativa color verde.

- **Grupo Familiar**

  Grupo Familiar según *Bueno* **o** *Malo*

Malo			
Grupo familiar	**Frec Absoluta**	**Frec Relativa**	**Frec Acumulada**
1	22,879	21,5	
2	35,379	31,1	
3	23,269	24,8	
4	13,054	14,7	
5	3,914	5,4	
6	1,017	1,6	
7	0,312	0,5	
8	0,099	0,2	
9	0,028	0,07	
10	0,019	0,06	
11	0,004	0,01	
12	0,004	0	
13	0	0	
14	0	0,01	
15	0,001	0	
N/A	0,021	0,01	

Cuadro 6.5: Frecuencia relativa de la variable Grupo Familiar según *Bueno* o *Malo*.

Se puede apreciar que tanto las personas clasificadas como *Buenas* y como *Malas*, tienen un promedio muy parecido de grupo familiar en todos los niveles. Esto nos da indicios de que la variable Grupo Familiar no discrimina bien según la variable BYM. Por estas razones no sería necesario incluir dicha variable en el modelo.

- **Profesión.**

Profesión	**Frecuencia Relativa**
Cargos de alta responsabilidad	2
Profesionales	0
Militar - policía - seguridad	5
Oficios	1
Otros independientes	0
Administrativo	6
Ama de casa	0
Desempleado	0
Docente	4
Empleado público	6
Jubilado	28
Obrero	6
Obrero calificado	4
Otros	29
Otros asalariados	5
Vendedor	3
Sin Datos	1

Cuadro 6.6: Frecuencia de la variable Profesión

Profesión según *Bueno* **o** *Malo***.**

Los grupos se comportan de manera similar al discriminar según *Bueno* o *Malo*, y similar a la variable Ocupación.

- **Total de Haberes**

  El Total de haberes hace referencia al ingreso nominal del sueldo principal que recibe el trabajador.

Mín.	1^{er} **Cuart.**	**Mediana**	**Media**	3^{er} **Cuart.**	**Máx.**	**Desvío**
1417	10360	15620	19050	23440	386500	13890

Cuadro 6.7: Medidas de resumen de la variable Total de Haberes.

Total de haberes según *Bueno* **o** *Malo*.

Mín.	1^{er} **Cuart.**	**Mediana**	**Media**	3^{er} **Cuart.**	**Máx.**	**Desvío**
1417	10360	15720	19200	23650	386500	14054

Cuadro 6.8: Medidas de resumen de la variable Total de Haberes según la categoría *Bueno*.

Mín.	1^{er} **Cuart.**	**Mediana**	**Media**	3^{er} **Cuart.**	**Máx.**	**Desvío**
3000	10330	14560	17430	20980	275600	11757

Cuadro 6.9: Medidas de resumen de la variable Total de Haberes según la categoría *Malo*.

Al discriminar según la variable de interés se pueden observar algunos cambios, los Haberes de los clientes calificados como *Bueno* llegan a ser más altos que el de los calificados como *Malo*. Por lo que podría ser un buen aporte incluirla en el modelo. Pero como el comportamiento y la información es parecida a la variable "Ingresos Líquidos", se decide agregar solo ésta última.

Capítulo 7

Anexo C

7.1. Scripts utilizados

1. **Script utilizado para la obtención de las diferentes muestras.**

```
rm(list=ls()) #borra todo lo cargado anteriormente


set.seed(123)
base<-read.csv2("BASE_log.csv", header = TRUE) ##2014-2011
base2<-read.csv2("base_AYR.csv", header = TRUE) ##2014-2011
base<-as.data.frame(base)
base2<-as.data.frame(base2)
prop.table(table(base$BYM))

base$muestra<-runif(nrow(base[,]))
base2$muestra2<-runif(nrow(base2[,]))

############## MUESTRA 90% ########################

muestra0<-base[base$muestra<=.9,]
valida0<-base[base$muestra>.9,]


############## MUESTRA 50% ########################

muestra1<-base[base$muestra<=.5,]
length(muestra1[,1])

valida1<-base[base$muestra>.5,]


########### base igual B y M ######################
```

```
#nos quedamos con el 90% de los malos

Malo=which(base$BYM=='M')
baseM<-base[Malo,]
baseM$muestra<-runif(nrow(baseM[,]))
muestraM<-baseM[baseM$muestra<=.9,]
no_muestraM<-baseM[baseM$muestra>.9,]
length(muestraM[,1])


#extraer una muestra equivalente a la prop de malos.

Bueno=which(base$BYM=='B')
baseB=base[Bueno,]
baseB$muestra<-runif(nrow(baseB[,]))
muestraB<-baseB[baseB$muestra<=0.095,]
no_muestraB<-baseB[baseB$muestra>0.095,]
length(muestraB[,1])

muestraBM=rbind(muestraM,muestraB)
no_muestraBM=rbind(no_muestraM,no_muestraB)
length(muestraBM[,1])


valida_BM<-no_muestraBM

############  muestra Activos  ######################
#nos quedamos con el 50% de los Activos

Act=which(base$Ocu2!=4)
baseA<-base[Act,]
baseA$muestra<-runif(nrow(baseA[,]))
muestraA<-baseA[baseA$muestra<=.5,]
valida_A<-baseA[baseA$muestra>.5,]
length(muestraA[,1])


############  muestra Pasivos ######################
#nos quedamos con el 50% de los Pasivos

Pas=which(base$Ocu2==4)
baseP<-base[Pas,]
baseP$muestra<-runif(nrow(baseP[,]))
muestraP<-baseP[baseP$muestra<=.5,]
Valida_P<-baseP[baseP$muestra>.5,]
length(muestraP[,1])

############  muestra 1a vez  ######################
```

```
primera=which(base$Cant_veces_opero2==0)
base_1a<-base[primera,]
base_1a$muestra<-runif(nrow(base_1a[,]))
muestra_1a<-base_1a[base_1a$muestra<=.9,]
valida_1a<-base_1a[base_1a$muestra>.9,]
length(muestra_1a[,1])

############  muestra 2a vez o mas ##################

no_primera=which(base$Cant_veces_opero2>1)
base_2a<-base[no_primera,]
base_2a$muestra<-runif(nrow(base_2a[,]))
muestra_2a<-base_2a[base_2a$muestra<=.5,]
valida_2a<-base_2a[base_2a$muestra>.5,]
length(muestra_2a[,1])
```

2. **Script utilizado para la regresión logística.**

```
### REGRESION LOGISTICA ###

########## Correr el archivo muestra.R

##########  MUESTRAS

#muestra1<-muestra0
#
muestra1<-muestra1
#muestra1<-muestraBM
#muestra1<-muestraA
#muestra1<-muestraP
#muestra1<-muestra_1a
#muestra1<-muestra_2a

##########  validacon

#valida1<-valida0
#
valida1<-valida1
#valida1<-valida_BM
#valida1<-valida_A
#valida1<-valida_P
#valida1<-valida_1a
#valida1<-valida_2a
```

```
############ ALGUNOS MODELOS

rlog_0=glm(BYM ~ 1
           ,family=binomial,data=muestra1)

summary(rlog_0)

#rlog_b=glm(BYM~ Cant_veces_opero2 + Edad + Sexo +
#                Antiguedad + Clearing2 + Cuotas.totales +
#                Valor.cuota.TotIng, family=binomial,
#           data=muestra1)

#summary(rlog_b)



rlog_c=glm(BYM ~  Cant_veces_opero2 + Edad + Sexo +  Ocu5 +
    Antiguedad +
                  Clearing2 + Cuotas.totales +
                  Valor.cuota.TotIng   ,family=binomial,
           data=muestra1)

summary(rlog_c)

####################    Test sig. del modelo

anova(rlog_c,test="LRT")

a=anova(rlog_0,rlog_b,test="LRT")

b=anova(rlog_0,rlog_c,test="LRT")

c=anova(rlog_b,rlog_c,test="LRT")


################################## K-S ##################

scores<-predict(rlog_c,type='response')

SCOR<-vector('character',length=(length(muestra1[,1])))
p=0
y=seq(0.0, 1, by=0.02 )
x=rep(0,length(y))
KS=cbind(y,x,x,x,x,x)
for (j in 1:length(y)){
p=p+0.002

for (i in 1:length(muestra1[,1]))
  if (scores[i]>p) SCOR[i]<-'M' else SCOR[i]<-'B'
```

```
#creo tabla con errores de clasificacion

M_o=table(muestra1$BYM,SCOR)
a=(M_oprop=(prop.table(M_o, m=1)))
KS[j,2]=a[1,1]
KS[j,3]=a[1,2]
KS[j,4]=a[2,1]
KS[j,5]=a[2,2]
KS[j,6]=a[2,2]-a[1,2]
}

max(KS[,6])

(kolmog=round(KS,2))

########### Errores de clasificacin  ################

prop.table(table(muestra1$BYM))
#creo columna que me clasifique en moroso o no
scores<-predict(rlog_c,type='response')
SCOR<-vector('character',length=(length(muestra1[,1])))

#creo columna que me clasifique en moroso o no
for (i in 1:length(muestra1[,1]))
  if (scores[i]>0.022) SCOR[i]<-'M' else SCOR[i]<-'B'


#tabla con errores de clasificacion

M_o=table(muestra1$BYM,SCOR)
M_o
a=round((M_oprop=(prop.table(M_o, m=1)))*100,0)
a
(MM=a[2,2])
(BM=a[1,2] )

########################## ROC ########################

#install.packges("pROC")
library(pROC)

roc1=roc(muestra1$BYM, scores, muestra1$BYM~scores,
          auc=TRUE, plot=FALSE)
roc1$auc

plot(1-roc1$specificities,roc1$sensitivities, type="l",
     ylim=c(0,1), main='Curva ROC, muestra 50% pob. que
         operaron ms de una vez',
```

```
      xlab='Tasas de falso positivo',
      ylab='Tasas de verdadero positivo', col="blue")

lines(KS[,1],KS[,1],type="l")

max(abs((1-roc1$specificities)-roc1$sensitivities))

a=which.max(abs((1-roc1$specificities)-
                  roc1$sensitivities))

roc1$thresholds[a]

roc1$specificities[a]
roc1$sensitivities[a]

k_s=abs(roc1$specificities[a]-(1-roc1$sensitivities[a]))
k_s

########## Errores de clasificacin  ################

prop.table(table(valida1$BYM))

#creo columna que me clasifique en moroso o no

scores_v<-predict(rlog_c,newdata= valida1, type='response')
SCOR_v<-vector('character',length=(length(valida1[,1])))

#creo columna que me clasifique en moroso o no
for (i in 1:length(valida1[,1]))
  if (scores_v[i]>0.08) SCOR_v[i]<-'M' else SCOR_v[i]<-'B'


#tabla con errores de clasificacion, validacin

M_o=table(valida1$BYM,SCOR_v)
M_o
a=round((M_oprop=(prop.table(M_o, m=1)))*100,0)
a
(MM=a[2,2])
(BM=a[1,2] )
```

3. **Script utilizado para la realización de CART.**

```
### CART ###
```

```
set.seed(567)

base<-read.csv2("base_AYR2.csv") ##2014-2011

############ muestra50%pob.

N=length(base[,1])
N
train=sample (1: nrow(base), N/2)

dim(base[train,])
table(base[train,12])
table(base$Cant_veces_opero2==0)

########

library("rpart")
lossmatrix=matrix(c(0,1,1,1,0,1,2,1,0), ncol=9, nrow=9)
cart_50= rpart( BYM~ Clearing2 + Cant_veces_opero2 + Edad +
                    Sexo + Antiguedad  +
                    Ocu5 + Cuotas.totales +
                    Valor.cuota.TotIng , base ,
                  subset =train , na.action=na.rpart,
                  method="class",
                  control=rpart.control(cp=0.0001))


plotcp(cart_50, main="")#para ver donde podar
printcp(cart_50)


pruned_50=prune(cart_50, cp=0.00232919)
plot(pruned_50, uniform=TRUE, margin=0.1,
     main='Arbol de Clasificacion')
text(pruned_50, use.n=TRUE, cex=0.85,
     splits=TRUE, pretty=0,col=c("red","green","blue"))

library(xtable)

xtable(printcp(cart_50))

## Prediccion 50

base.test=N_base[train ,]
BYM.test=N_base$BYM[train ]


tree.pred_50=predict(cart_50, base.test ,type ="class")
```

```
table(tree.pred_50  ,BYM.test)

(prop=(prop.table(table(tree.pred_50  ,BYM.test), m=1))*100)

#############  muestra 2a vez o mas

no_primera=which(base$Cant_veces_opero2>0)
base_2a<-base[no_primera,]

N_2a=length(base_2a[,1])
train_2a=sample  (1: nrow(base_2a), N_2a/2)

dim(base[train_2a,])
table(base[train_2a,11])



##########

lossmatrix=matrix(c(0,1,1,1,0,1,2,1,0), ncol=9, nrow=9)
cart_2a= rpart( BYM~ Clearing2 + Cant_veces_opero2 + Edad +
                    Sexo + Antiguedad  +
                    Ocu5 + Cuotas.totales +
                    Valor.cuota.TotIng  , base_2a  ,
                  subset =train_2a  ,  na.action=na.rpart,
                  method="class",
                  control=rpart.control(cp=0.0001))


plotcp(cart_2a, main="")#para ver donde podar
printcp(cart_2a)


pruned_2a=prune(cart_2a, cp=0.00078801)
plot(pruned_2a, uniform=TRUE, margin=0.1,
     main='Arbol de Clasificacion')
text(pruned_2a, use.n=TRUE, cex=0.85,
     splits=TRUE, pretty=0,col=c("red","green","blue"))



############base igual B y M

#  Malo
Malo=which(base$BYM=='M')
base_M<-base[Malo,]
table(base_M[,11])
```

```
N_M=length(base_M[,1])
N_M


# misma cantidad de Bueno

Bueno=which(base$BYM=='B')
base_Bueno=base[Bueno,]
table(base_Bueno[,11])
train_B=sample(1: nrow(base_Bueno), N_M)
base_B=base_Bueno[train_B,]
dim(base_B)
table(base_B[,11])


# union de Bueno y Malo
base_BM=rbind(base_M, base_B)
dim(base_BM)
table(base_BM[,11])


# muestra del 50%
N_BM=length(base_BM[,1])
N_BM
n_BM=round(N_BM*0.9)
n_BM

train_BM=sample(1: nrow(base_BM), n_BM)

dim(base_BM[train_BM,])
table(base_BM[train_BM,11])


########

lossmatrix=matrix(c(0,1,1,1,0,1,2,1,0), ncol=9, nrow=9)
cart_BM= rpart( BYM~ Clearing2 + Cant_veces_opero2 + Edad +
                    Sexo + Antiguedad  +
                    Ocu5 + Cuotas.totales +
                    Valor.cuota.TotIng , base_BM ,
                  subset =train_BM,  na.action=na.rpart,
                  method="class",
                  control=rpart.control(cp=0.0001))


plotcp(cart_BM, main="")#para ver donde podar
printcp(cart_BM)
```

```
pruned_BM=prune(cart_BM, cp=0.00501720)
plot(pruned_BM, uniform=TRUE, margin=0.1,
     main='Arbol de Clasificacion')
text(pruned_BM, use.n=TRUE, cex=0.85,
     splits=TRUE, pretty=0,col=c("red","green","blue"))

## Prediccion BM

base.test=N_base[train ,]
BYM.test=N_base$BYM[train ]


tree.pred_BM=predict(cart_50, base.test ,type ="class")

table(tree.pred_BM ,BYM.test)

(prop=(prop.table(table(tree.pred_BM ,BYM.test), m=1))*100)
```