Universidad Nacional Autónoma de México

Facultad de Ciencias

Algunos modelos de clasificación estadística usados en "Credit Scoring"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuaria

PRESENTA:

Ana Laura Medina Pérez

DIRECTOR DE TESIS:

Dra. Guillermina Eslava Gómez

2013

1. Datos del alumno:
Medina
Pérez
Ana Laura
53 35 08 02
Universidad Nacional Autónoma de México
Facultad de Ciencias Actuaría
303033211

2. Datos del tutor
Dra
Guillermina
Eslava
Gómez

3. Datos del sinodal 1
Dra en E
Ruth Selene
Fuentes
García

4. Datos del sinodal 2
Dra en Adm
María del Pilar
Alonso
Reyes

5. Datos del sinodal 3
Act
Jaime
Vázquez
Alamilla

6. Datos del sinodal 4
Mat
Margarita Elvira
Chávez
Cano

7. Datos del trabajo escrito
Algunos modelos de clasificación estadística usados en *Credit Scoring*
Aplicación en una base de clientes con créditos de consumo
124 p
2013

Contenido

Introducción

Para una entidad financiera la necesidad de **calificar** a un individuo u otra entidad con el fin de otorgarle un crédito, ampliar su límite crediticio o simplemente conocer su capacidad de cubrir deudas con base en su historia crediticia, le es primordial para conocer los posibles escenarios económicos a los que está sujeta. El modelar esta toma de decisiones de la manera más rápida y precisa genera un menor impacto en las posibles pérdidas esperadas.

Modelar estos escenarios es un factor importante para cualquier entidad financiera pues le permitirán aplicar las estrategias operacionales que aumentarán la rentabilidad del préstamo o incluso del cliente. Cualquier entidad que confía en pagos regulares de clientes sabe que existe un riesgo de incumplimiento del solicitante debido a una gran diversidad de circunstancias.

A lo largo de la historia han existido diversos métodos para evaluar todas estas opciones, por ejemplo, desde tiempos muy antiguos los prestamistas buscaban la manera de reducir el riesgo valiéndose de diversas fuentes de información, que iban desde libros que contenían el historial de sus clientes hasta el preguntar a otros prestamistas si tenían referencias sobre el nuevo solicitante.

Conforme fue creciendo la población resultó más complicado obtener información a través de estas fuentes, sin embargo, gracias a los avances tecnológicos se ha logrado analizar la información de aspirantes más rápido. Hoy en día, las entidades financieras analizan grandes bases de datos con el objetivo, en este caso, de discriminar a un buen solicitante de crédito de uno malo. Pero el problema esencial es definir lo que es un buen solicitante.

En el contexto del negocio de concesión de crédito, un buen solicitante es aquel que puede cumplir la obligación de pago del mismo, hecho que se relaciona a diversos atributos que favorecen o ponen en ventaja a dicho cliente entre muchos otros solicitantes, un ejemplo de esto sería el ingreso bruto con el que cuenta para solventar sus gastos y deudas, la edad y estado civil, su grado de estudios y el tipo de trabajo que tiene, el número de personas que dependen económicamente de él, etc. Por otro lado, un mal solicitante es aquel que no logra dicho cometido y tiene en su contra diversos factores que no lo hacen merecedor del crédito ya que indican que los medios

constantes o seguros que tiene para cubrir la deuda no son suficientes, de acuerdo a las políticas de cada entidad financiera. Pero todos estos datos no responden al cien por ciento si es un buen solicitante o uno malo, pues siempre existirán factores "de la vida diaria" que ayudarán a estimar qué tipo de solicitante es.

La metodología empleada para alcanzar este fin es la de los modelos estadísticos que ayudan a clasificar a solicitantes de crédito de acuerdo a la definición de bueno y malo para cada entidad financiera.

Suponga que ya se cuenta con toda la información necesaria, o bien, que está disponible para cada uno de los candidatos a crédito (que pueden ser desde un par hasta millones de ellos), entonces algunos ejemplos de esto son el modelo de *Credit Scoring*, así como los siguientes modelos:

- Projection pursuit
- Análisis de conglomerados
- Discriminante lineal
- Discriminante cuadrático
- Regresión logística
- Redes neuronales
- Árboles de decisión

Cada uno de los modelos anteriores necesita de diversas hipótesis para desarrollarse y dependen totalmente del tipo de información con la que cuenten, del tipo de variables explicativas con los que se trate o simplemente del objetivo a analizar en la base de datos. Los primeros dos modelos mencionados sirven principalmente para el reconocimiento de patrones en la muestra, así como para encontrar datos atípicos o correlaciones inesperadas. Los siguientes buscan de alguna manera obtener una predicción y porcentaje de mala clasificación de datos, siendo así que se buscan aquellos modelos que arrojen mejores predicciones y menores tasas de clasificación errónea. Con el análsis de los resultados de estos modelos se pueden dar argumentos para una buena toma de decisiones, mismas que dependen de otros factores que se ilustrarán adelante.

Por otro lado, para dimensionar el comportamiento de la población objetivo[1] se necesita comparar el modelo que se desempeñe mejor en las tomas de decisiones, es por ello que se deben tener en cuenta métodos como las *curvas ROC* (*Receiver*

[1]Se entiende por población objetivo a aquellos solicitantes a discriminar.

Operator Characteristic) para discriminar modelos.

Ahora bien, para ejemplificar el desarrollo de estos procesos que ayudarán a clasificar una base de solicitantes de crédito se considerará el método de *Credit Scoring* que es muy utilizado actualmente en bancos e instituciones financieras, el cual consiste principalmente en conocer, mediante el comportamiento de solicitantes previos, la propensión o probabilidad de que un nuevo solicitante sea un buen cliente. Es de esta forma que mediante el modelo de *Credit Scoring* se liga el conocimiento previo del negocio y la teoría estadística.

Para ejemplificar algunos de estos modelos se considera una muestra de 1,000 clientes que solicitaron créditos de consumo a un banco alemán, de esta manera en el primer capítulo se plantean las diferencias entre discriminar y clasificar; los modelos a utilizar para clasificar a los clientes entre clientes con capacidad crediticia y clientes sin capacidad crediticia y, finalmente, se define el modelo de *Credit Scorig*, así como un poco de historia del mismo.

El segundo capítulo se iniciará con un análisis descriptivo de la base (que en adelante se llamará *base DM*) para resumir las características de la población con el objetivo de dar a conocer sus atributos; se aplicarán modelos estadísticos para el reconocimiento de patrones como son *projection pursuit* y, mediante el método de componentes principales, encontrar las principales características que discriminan a los solicitantes.

En el tercer capítulo se desarrollarán dos modelos: regresión logística y redes neuronales, con el objetivo de obtener las mejores predicciones y tasas de clasificación de los datos observados para finalmente probar si existe la posibilidad de obtener un mejor resultado al mezclar ambos modelos.

Finalmente, se planteará una breve discusión de los principales problemas que se presentaron en la implementación de los modelos mencionados y, para terminar con las conclusiones generales de la presente tesis, del cómo se solucionaron dichos problemas. En la ultima sección también se podrán encontrar distintos anexos que dan soporte y complemento a los resultados aquí presentes.

Cabe mencionar que a lo largo de cada capítulo se hará mención de terminología en el idioma inglés por la familiaridad con que se conoce en la literatura, sin embargo, se procurará dar a conocer su traducción al español más usada.

Debido a que resulta más sencillo entender y tomar decisiones sobre los análisis basándose en una buena visualización de los datos, a lo largo de esta tesis se hará uso de paquetes estadísticos como son R y sus librerías *rggobi, gplots, MASS, lmtest, nnet* y *ROCR* así como del *software* SAS a fin de facilitar la compresión de los aná-

lisis efectuados. Además, para tener un pleno conocimiento de cómo se obtuvieron los datos, en los anexos se encontrarán los códigos empleados.

Capítulo 1

Clasificación *vs* Discriminación

Una de las principales aplicaciones de la estadística en cualquier ámbito consiste en lograr **clasificar** adecuadamente las observaciones de una población, para ello se utiliza una función que **discrimine** lo mejor posible las categorías de los elementos esperando siempre no ajustar modelos que sobre ajusten a las observaciones.

Un sobre ajuste ofrece estimaciones que responden a la muestra observada pero que generalmente no ajustarán a la muestra de una población. De esta necesidad surge de igual manera el lograr discriminar dichos elementos en grupos bien definidos. Dado lo anterior, haciendo referencia a Venables and Ripley (2002, p. 331), se define como *clasificación* al hecho de asignar nuevos elementos en grupos previamente bien definidos, así como *discriminar* al hecho de definir las variables o causas de separación de los grupos para que se logre una clasificación adecuada.

Ahora bien, no se debe de perder la idea que el clasificar un nuevo elemento también da cabida a dos situaciones: la primera es que se espera obtener predicciones de futuros eventos y la segunda que dichas clasificaciones y predicciones tengan la menor tasa de error esperada.

Por ejemplo, suponga una población que puede conformarse de diversas poblaciones, y que en cada elemento de la población de interés se ha observado una variable aleatoria $p-$dimensional denotada por x, cuya distribución se conoce mediante la población procedente, entonces un nuevo elemento se presenta para ser clasificado en alguna de las poblaciones consideradas anteriormente; sin embargo, puede que se tenga toda la información de este nuevo elemento para su discriminación o quizá no se cuente con ella. Este último evento se presenta con gran frecuencia en un banco al momento de intentar evaluar a un nuevo solicitante al cual le falta información, o bien, simplemente por un error humano se capturen mal los datos.

En ingeniería, en su rama de *Machine Learning*, a este problema de clasificación, a pesar de faltantes de información, se le conoce con el nombre de reconocimiento de

patrones, del inglés *Pattern Recognition*, en donde el uso principal de los modelos es la construcción de máquinas que clasifiquen de manera automática la llegada de nueva información. La aplicación de estos modelos en el medio financiero es conocido como *Credit Scoring*.

Ahora bien, existen dos vertientes de clasificación de observaciones:

- **Clasificación supervisada.** En esta técnica se cuenta previamente con una muestra de elementos bien clasificados de tal forma que sirven como pauta o modelo para la clasificación de las siguientes observaciones. El objetivo en sí consiste en "aprender" de esta muestra de datos para que el modelo a elaborar sea capaz de etiquetar una futura observación o elemento con base en la construcción de reglas establecidas de la muestra previamente clasificada para así predecir a la variable de respuesta.

- **Clasificación no supervisada.** Por el contrario, en esta técnica no se tiene una muestra *a priori* definida, por lo que es necesario determinar el número de clases o poblaciones a las cuales puede pertenecer una observación para posteriormente aplicar métodos estadísticos que permitan la clasificación de los datos. Esta técnica resulta más compleja que la anterior debido a que en la clasificación supervisada se pueden extraer los distintos perfiles de cada población ya definida, mientras que en la clasificación no supervisada es necesario definir cuantas categorías se presentan en la población con base en los diversos perfiles encontrados, además de tener que encontrar la regla de clasificación adecuada para discriminar las observaciones de la base.

Se conocen varios modelos de clasificación supervisada y predicción de datos tales como: discriminante lineal, discriminante cuadrático, regresión logística, redes neuronales, árboles de decisión, entre otros. Mientras que por modelos de clasificación supervisada se puede puntualizar principalmente análisis de conglomerados, componentes principales y *projection pursuit*. Saber qué modelo aplicar a determinada base de datos a menudo tiene que ver con el tipo de variables con las que se trabaja: cualitativas, cuantitativas o de ordenación y es importante tener en cuenta que la potencia de predicción y clasificación de un modelo se basa, en parte, en el tipo de variables con las que se desarrolle.

1.1. Algunos modelos de clasificación supervisada

Para el desarrollo de la presente tesis se han considerado, basándose en la estructura de la base con la que se trabajó, los siguientes modelos.

1.1.1. Regresión Logística

Una de las primeras formas de manejar la clasificación de datos es el modelo de discriminante lineal planteado por Ronald Aylmer Fisher $(1890-1962)$, sin embargo, cuando se trata de datos categóricos una opción es aplicar el modelo de regresión logística y, dado que en la base se cuenta con una variable de respuesta de dos categorías, se decidió desarrollar este modelo.

El modelo de regresión logística pertenece a la familia de modelos lineales generalizados y es de gran utilidad cuando las observaciones a clasificar no se distribuyen como una normal, aunque tiene la habilidad de hacerlo también cuando este supuesto no sucede. El modelo se basa en variables dicotómicas para describir una función de la media como una función de las variables explicativas.

En general, los modelos lineales generalizados son una gran herramienta para la discriminación de datos. De hecho, es posible extenderlos a variables de respuesta con distribuciones no normales y modelos en funciones de medias, claro ejemplo de esto son el modelo de regresión logística y el modelo loglineal.

Según expresa Agresti (2002, p.116), los modelos lineales generalizados tienen tres atributos principales:

1. Un componente aleatorio (*random component*), que se identifica con la variable de respuesta Y y su distribución de probabilidad; la variable Y presenta p observaciones indepedientes $(y_1, \ldots, y_p)$ con una "función natural" de la familia exponencial de la siguiente forma:

$$f(y_i; \theta_i) = a(\theta_i) b(y_i) e^{[y_i Q(\theta_i)]}, \tag{1.1}$$

   donde el parámetro de θ_i puede variar de $i = 1, \ldots, N$ dependiendo de los valores de las variables explicativas y el término $Q(\theta_i)$ es llamado "parámetro natural". Sin embargo, las distribuciones *Poisson* y Binomial que son la base de los Modelos Loglineal y de Regresión Logística, respectivamente, se consideran como casos especiales.

2. Un componente sistemático (*systematic component*), que define a las variables explicativas $(x_1, \ldots, x_p)$ usando una función de predicción lineal tal que si se denota a η_{ij} como el valor predictivo de j en la observación i tomando j valores en $(1, \ldots, n)$, entonces:

$$x_i = \textstyle\sum_j \beta_j \eta_{ij} \quad i = 1, \ldots, p. \tag{1.2}$$

A la combinación lineal de las variables explicativas se le conoce como "predictor lineal", donde usualmente $\eta_{ij} = 1$ para toda i en el coeficiente de una intersección llamada β_0 en el modelo. Y finalmente,

3. Una función liga (del inglés *link function*), que conecta al componente aleatorio y al componente sistemático. Sea $\mu_i = \mathbb{E}(Y_i)$, $i = 1, \ldots, p$ la función liga entre μ_i y x_i se denota por $x_i = g(\mu_i)$, donde la función g es monótona y diferenciable. Por lo que g es la liga entre cada x_i y $\mathbb{E}(Y_i)$ a través de la fórmula:

$$g(\mu_i) = \sum_j \beta_j x_{ij} \quad i = 1, \ldots, p. \tag{1.3}$$

La función $g(\mu) = \mu$ es llamada liga identidad (del inglés *identity link*) cuando $\eta_i = \mu_i$ la cual sería la función liga para un modelo de regresión lineal ordinario con distribución normal del vector de variables explicativas $\mathbf{Y}$, y para el caso donde $g(\mu_i) = Q(\theta_i)$ y $Q(\theta_i) = \sum_j \beta_j x_{ij}$ a esta función liga se le conoce como liga canónica (del inglés *canonical link.*)

Ahora bien, una gran ventaja entre el modelo de regresión logística y el modelo de regresión lineal consiste en que en el primero las variables dependientes son variables categóricas o binarias, mientras que en el segundo se suponen continuas. De acuerdo con Hosmer and Lemeshow (2000, p.1), las diferencias entre la regresión logística y la regresión lineal se basan principalmente en dos puntos: los parámetros del modelo y las hipótesis iniciales. Una vez dadas estas diferencias el modelo se desarrolla bajo los mismos principios generales que usa la regresión lineal.

En la tabla 1.1.1 se aprecian como resumen los distintos modelos lineales generalizados indicando el componente aleatorio, la función liga, el componente sistemático y el modelo al cual se aplica; esta tabla permite tener un panorama general de este tipo de modelos.

Tabla 1.1.1. Tipos de modelos lineales generalizados para análisis estadístico

Componente aleatorio	**Liga**	**Componente sistemático**	**Modelo**
Normal	Identidad	Continua	Regesión
Normal	Identidad	Categórica	Análisis de Varianza
Normal	Identidad	Mixta	Análisis de Covarianza
Binomial	Logit	Mixta	Regresión Logística
Poisson	Log	Mixta	Loglineal
Multinomial	Logit Generalizada	Mixta	Respuesta Multinomial

Fuente: Agresti (2002, p.118)

1.1.2. Redes Neuronales

Las redes neuronales o NN (por sus siglas en inglés *Neural Networks*) se desarrollan bajo el concepto de *métodos de aprendizaje*, es decir, tanto bajo el concepto de clasificación supervisada como de la no supervisada, y a su vez en el campo estadístico como en el de inteligencia artificial, basándose en la idea de extraer combinaciones lineales de las variables independientes o características para después modelar la variable dependiente como una función no lineal de las características. Cook and Swayne (2007, p.88) lo describen como un modelo aditivo que transforma las variables independientes, usualmente a través de una función logística, agrega otras variables independientes, realiza la transformación de nuevo y agrega una vez más para producir predicciones de clase, o bien, categoría.

Según menciona Rojas (1996, p.476), este modelo fue propuesto inicialmente en los años 60s por Widrow and Hoff (1960) y Rosenblatt (1962), y para los años 80s comenzó a tener mayor atención, sin embargo, su implementación como hoy se conoce va de la mano con los avances computacionales, ya que que se basa en cálculos recursivos de funciones.

El modelo de redes neuronales para la clasificación de elementos se basa en ponderadores que están asociados a un peso, en inglés *weight*, y cada capa de nodos resulta ser una nueva función. La interconexión entre neuronas o funciones con las diversas cargas o ponderadores da como resultado final una clasificación de elementos (Ripley, 1996, p.146)

Este método está muy relacionado con la rama de ingeniería *machine learning* donde se maneja otro tipo de terminología que también será necesario definir. Por ejemplo, a los parámetros se les conoce como ponderadores, o bien del inglés *weights*; a las variables independientes, del inglés *predictors*), se les conoce como *inputs* y a la variable dependiente, del inglés *response*, se le conoce como variable objetivo, o bien, del inglés *target*. No debe confundirse el significado de *output* con el de *target*, pues el primer concepto se refiere al actual valor observado en la muestra, mientras que el segundo, al resultado de aplicar cierto modelo y obtener una predicción del mismo.

Otra diferencia singular entre la estadística y la rama *machine learning* es que la primera supone que se conoce o se quiere conocer la distribución de los elementos a clasificar, mientras que realmente en cuanto a clasificación se trata las redes neuronales se basan en la experiencia y la "prueba y error" para desarrollar el modelo.

Hastie et al. (2009) compara el método de redes neuronales con un "modelo de aproximación libre" similar o apegado a la estadística clásica, especialmente cuando el tamaño del *conjunto de entrenamiento* es pequeño comparado con el problema real a resolver, esto debido a que mientras más sencillo sea el modelo NN menos recursivi-

dad de funciones presentará. Pero, ¿a qué se referiere con *conjunto de entrenamiento*?

Como se ha remarcado en el párrafo anterior, existen diversos tipos de conjuntos o bien subconjuntos para desarrollar este método. Hastie et al. (2009) menciona lo siguiente: "Estos subconjuntos son extraídos de la población inicial y pueden ser tantos como se desee o necesite de hecho, el cómo tomar cada subconjunto equivale a un método diferente de trabajo, el más básico radica en dividir la base en dos, un conjunto de prueba o entrenamiento (*training set*) y un conjunto de validación (*validation set*) y se definen de la siguiente forma: el subconjunto de entrenamiento es empleado para estimar los parámetros del modelo y el subconjunto de validación es usado para diversos propósitos, si es lo suficientemente grande y representativo de la población se puede estar casi seguro de que el error obtenido en este subconjunto es un estimador razonable del error que arrojaría el modelo de ser aplicado a toda la población. Bajo estos mismos supuestos de representatividad y tamaño se puede utilizar el conjunto de validación para seleccionar el modelo adecuado de clasificación y evaluación."

En el caso en el que el subconjunto de validación sea pequeño, el autor aconseja evaluar diversos modelos con diversos niveles de complejidad sobre el subconjunto de entrenamiento para seleccionar aquel modelo con menor desviación en el subconjunto de validación, cuidando siempre que el modelo elegido no esté sobre estimando al subconjunto de validación.

Existe otro tipo de partición muy implementado en la práctica y consiste en segmentar en tres subconjuntos a la población, los dos primeros ya mencionados y un tercero llamado el conjunto de prueba (*test set*), de tal forma que el conjunto de entrenamiento es usado para estimar los parámetros, el conjunto de validación es utilizado para seleccionar el mejor modelo y el subconjunto de prueba o *test set* es para obtener una desviación de la estimación de la predicción del error.

De igual forma se pueden clasificar las diversas técnicas de partición de la población de acuerdo al número de subconjuntos con los que se quiera trabajar, o bien si se quieren tomar estos subconjuntos con o sin reemplazo. Sin embargo, en general estas técnicas se pueden aplicar a cualquier método de clasificaión siempre y cuando no se tienda a sobre estimar la clasificación.

1.2. *Credit Scoring*

Para responder a la pregunta ¿qué es *Credit Scoring*? se debe entender primero el significado de cada una de las palabras que lo conforman: *Credit* y *Scoring*, o por su traducción al español "crédito" y "puntaje".

Crédito

Por crédito se entiende el **acuerdo** en el cual un comprador **recibe algo de valor a cambio** de la promesa de **pagar** al prestamista en una **fecha posterior**, en otras palabras "comprar hoy, pagar después". (Anderson, 2007, p. 3)

Por lo anterior se puede deducir que *comprador* o *solicitante* es aquella persona que solicita el crédito y puede ser de carácter física o moral, y *prestamista* u *otorgante*, el banco o la institución financiera que lo otorga, misma que debe cumplir con los estatutos que establece la Ley General de Títulos y Operaciones de Créditos.

Sin embargo, esta operación como ya se vió se desarrolla en un **tiempo establecido** por ambas partes involucra también cierta **recompensa, o tasa de interés** (al tratarse de un monto), que el prestamista ganará por dicho servicio.

Por otro lado, a pesar de que los créditos resultan deudas a saldar a distintos plazos (dependiendo del tipo de crédito y objetivo del mismo) e inclusive el acordar la tasa de interés conveniente puede resultar incómodo para alguna de las dos partes (más frecuentemente para el solicitante), se observa que gran parte de la economía se mueve mediante la concesión de créditos y, por supuesto, del interés que se pueda generar de ellos, debido a que facilita la compra de bienes y/o servicios a un consumidor que a pesar de que, en el momento de la adquisición no cuenta con el dinero en efectivo puede disfrutar de la compra, lograr beneficios de ella y saldar su deuda con "cómodos pagos". Son estas algunas de las razones por las que se ha vuelto extensa la rama de los créditos, abarcando desde créditos personales hasta créditos hipotecarios, pues se busca su mejor comercialización y acoplamiento a las tendencias y necesidades de la sociedad, sin olvidar las utilidades para la entidad emisora del crédito.

Ahora bien, es un hecho que a pesar de que en la antigüedad contraer una deuda monetaria implicaba también contraer una deuda de honor, hoy en día el crédito ha tomado como primera instancia el derecho a poseerlo y en segunda la obligación a pagarlo. Es por ello que actualmente en la práctica las instituciones financieras han tenido que catalogar a sus clientes, y a los que podrían llegar a serlo, de acuerdo a su comportamiento de pago con otras instituciones y con la propia institución que lo evalúa, es decir, basándose en su morosidad; de tal forma que un cliente es considerado como moroso si ha incumplido en alguno de los pagos de sus créditos (vigentes o cerrados, dependiendo de qué tan estrictos sean los criterios utilizados) a partir del primer día después de su fecha de pago. Esto se logra mediante una Sociedad de Información Creditica (SIC) o Burós de Crédito, los cuales administran las bases de datos con los historiales de crédito de las personas, incluyendo los créditos que han obtenido y si han cubierto su deuda a tiempo o no.[1]

[1]El objeto de las SIC es vender información del historial crediticio de las personas a estas

Scoring

Scoring se refiere al uso de herramientas numéricas para alinear casos de orden (personas, compañías, países) de acuerdo a una cualidad verdadera o percibida (funcionamiento, conveniencia, posibilidad de venta, riesgo) para discriminar entre ellos y asegurar decisiones objetivas y constantes (seleccionar, descartar, exportar, vender). Los datos disponibles son integrados en un solo valor que implica alguna cualidad, usualmente relacionada convenientemente (Anderson, 2007, p.7). En otras palabras, *Scoring*, o bien "Puntaje", es una manera de ordenar por grado de importancia (ya antes definida) una lista de objetos o personas a clasificar a través de una única cifra llamada puntaje. La unión de todas las posibles cualidades o ventajas de cada observación aporta cierto peso a su puntaje y, por consiguiente, mejora su lugar en la escala definida dando como resultado un solo valor numérico que lo identifica ante los demás miembros a clasificar.

Si se asigna un puntaje a datos similares de periodos pasados, se logra generar un modelo predictivo basado en esos puntajes históricos para evaluar la probabilidad relativa de un futuro acontecimiento. Dado lo anterior, un prestamista debe tomar dos tipos de decisiones, la primera es si hay que concederle el crédito a un nuevo aspirante, para lo cual se utiliza como referencia el comportamiento de su puntaje, y la segunda es cómo tratar con aspirantes existentes (lo cual puede implicar considerar incrementar el límite de los créditos), y esto se logra al evaluar su puntaje a través del método de *Credit Scoring*. Sin importar la decisión a considerar, el objetivo es que exista una larga muestra de consumidores previos con especificaciones y subsecuentemente exista una historia crediticia. Todas las técnicas usan esta muestra para identificar las conexiones entre las características de los consumidores y por consiguiente qué tan buena o mala es su historia crediticia. (Anderson, 2007, p.9).

Credit Scoring

Esencialmente este método se basa en el comportamiento de consumidores similares que fueron evaluados bajo las mismas restricciones, para lo cual se toma una muestra de datos recientes. La información con la que trabaja el modelo dispone de varias covariables o características sociodemográficas y de comportamiento crediticio, como son: edad, sexo, dirección, ocupación, saldos en cuentas corrientes, morosidad, etc., las cuales se consideran como atributos de cada *i-ésimo* consumidor.

Con estos datos acumulados se puede crear una *Scorecard*, donde se van asignando los puntos que equivalen a cada característica según la política de cada entidad fi-

entidades y empresas siempre y cuando los usuarios así lo autoricen. Las SIC presentan dicha información en forma estandarizada en un "Reporte de Crédito", el cual es utilizado comúnmente, junto con otra información de la persona, para determinar si ésta es o no sujeto de crédito. El reporte de crédito puede ser un factor relevante para la aprobación o rechazo de las solicitudes de crédito de las personas. Fuente: http://www.banxico.org.mx/

nanciera o prestamista; o bien, usar técnicas estadísticas que se ayudan de técnicas de minería de datos para poder encontrar la mejor manera de discriminar los datos. De esta forma se define *Credit Scoring* como el conjunto de modelos de decisión que ayuda a prestamistas o entidades financieras a maximizar un beneficio derivado de la evaluación de un riesgo, asociado a la concesión de un crédito, es decir, ayuda a decidir si se concede o no el crédito a una persona en función de modelos matemáticos y algunos algoritmos relacionados a minería de datos. Este método también puede ayudar a estimar el monto del crédito que se debe otorgar y, de acuerdo al análisis de los datos, qué estrategia operacional incrementará la rentabilidad de los prestatarios ante los prestamistas. Cabe señalar que no evalúa la capacidad acreedora de un consumidor ni intenta responder el porqué de la situación financiera de alguna persona o entidad financiera, punto que será abordado más adelante.

1.2.1. Breve historia de método de *Credit Scoring*

El método de *Credit Scoring* no se remonta más que a 60 años atrás. Según menciona Thomas et al. (2002, p.3), los primeros planteamientos a este tipo de problemas de clasificación en estadística los realizó Fisher (1936), sin embargo en 1941 Durand fue el primero en reconocer que se pueden usar los mismos modelos para discriminar entre un buen o mal préstamo, siendo éste un proyecto de investigación que realizó para *National Bureau of Economic Research* de los Estados Unidos, aunque realmente no fue utilizado para algún propósito predictivo.

En la década de los 30s algunas compañías introdujeron sistemas de puntajes para tratar de encontrar inconsistencias en las decisiones de crédito a través del análisis de los mismos y al iniciar la Segunda Guerra Mundial se comenzaron a utilizar dentro del servicio militar, sin embargo existían muy pocas personas especializadas en el tema. No fue sino hasta finales de la guerra que se empezaron a automatizar las decisiones de crédito y las técnicas de clasificación comenzaron a desarrollarse en la estadística de tal forma que, al notar el gran aporte y beneficios obtenidos en los modelos de decisiones de crédito, se empezó a creer que los modelos estadísticos podían hacer un mejor trabajo que aquellas personas que se formaron con bases empíricas por falta de personal experimentado. Por lo que en los 50s se forma la primer consultoría en San Francisco por Bill Fair y Earl Isaac.

Más adelante, en los 60s, con la aparición de las tarjetas de crédito y gracias al avance computacional de la época, los bancos y otras instituciones de créditos comenzaron a usar con mayor frecuencia el método de *Credit Scoring*, ya que existía conocimiento previo y además les generaba dos grandes beneficios: reducir las pérdidas y hacer más rápida la toma de decisiones; lo cual hizo crecer al mercado de manera acelerada. De esta manera, el método de *Credit Scoring* comienza a tomar presencia conforme evoluciona la sociedad.

1.2.2. Implementaciones actuales del método de *Credit Scoring*

El método de *Credit Scoring* generalmente se utiliza en cualquier tipo de entidad financiera que posea una base con suficientes datos y variables para la clasificación en el otorgamiento de créditos. Sin embargo, en general es usado como una guía de decisión que afecta directamente al consumidor al marcar vías o parámetros para aceptar o rechazar las solicitudes de un cliente, para aumentar el valor de un préstamo o mensualidad, ajustar una tasa de interés, el periodo del mismo e inclusive obtener una predicción del riesgo que se corre al aplicarle este método a una cartera morosa.

La presente tesis implementa el método de *Credit Scoring* utilizando los modelos de regresión logística y redes neuronales para predecir el adecuado otorgamiento de créditos a través de la clasificación de observaciones anteriores. Sin embargo, este método se puede usar en cualquier base de datos con otro tipo de objetivo debido a que una de las herramientas con las que trabaja es el minado de datos que se define como el proceso de seleccionar, explorar y modelar grandes volúmenes de datos para descubrir información previamente desconocida e inclusive encontrar patrones para la toma de decisiones.

Se puede decir que la minería de datos trabaja mediante un análisis de segmentación que indica cuál de los segmentos es más propenso a exhibir un tipo particular de comportamiento, fungiendo de esta forma como fuente de conocimiento acerca de la base de datos a tratar para así facilitar el entendimiento del comportamiento de las observaciones, el resumen de las estadísticas más representativas de éstas (media, varianza, frecuencia y correlación con otras variables), la redución de variables a sólo aquellas que aportan más información mediante el modelo de componentes principales, la observación de conglomerados que pueden acumular observaciones con características similares utilizando la predicción y considerando condiciones similares del pasado e inclusive una explicación basada en el comportamiento de los datos.

Estas mismas técnicas son utilizadas para construir modelos donde se considera información de otras áreas de estudio por lo que se requieren aplicar distintas decisiones, algunos ejemplos de esto serían lo siguientes:

- Predecir el riesgo que corre una empresa de ir a la quiebra. Sin embargo, sólo se pueden tomar pequeñas muestras para calcular los históricos y la acumulación de información puede ser manipulada por sus directores.
- Desarrollar el cálculo de un puntaje de morosidad asignado a toda persona que actualmente posea o haya adquiridó en algún momento un crédito. Este tipo de datos actualmente se pueden obtener al solicitar un Reporte Especial del Buró de Crédito mexicano.

- En la detección de fraude dentro de cualquier tipo de institución financiera. Por ejemplo, en un banco y la posible duplicidad de tarjetas si en estos momentos se registra una compra en México y en dos minutos otra en Nueva York con el mismo número de plástico.

Fuera del ámbito financiero se puede pensar en los siguientes problemas:

- El conocer los posibles donadores a cierto evento de beneficencia de este año dadas las aportaciones del año anterior en el mismo evento o eventos similares.
- Intentar encontrar las causas o principales factores que influyen en la propagación de una enfermedad.
- Conocer la campaña política que favorece en el momento a algún candidato debido a las peticiones de la población.
- Estimar el *stock* que debe tener una empresa en determinadas épocas del año de acuerdo a los comportamientos de consumo de años pasados.
- Conocer las futuras necesidades de un cliente al cual se le ofrece algún bien o servicio de acuerdo al comportamiento de otros clientes con perfiles similares.
- Encontrar comportamientos genéticos que ayuden a la investigación científica;

Otros ejemplos se pueden consultar en Crook and Swayne (2002).

1.2.3. Alcances y limitaciones del método de *Credit Scoring*

A continuacón se enumeran algunos alcances y limitaciones que se observan en el método de *Credit Scoring* al momento de llevarlo a la práctica. Para mayor detalle consultar Crook and Swayne (2002).

- Se ha señalado ya que el método de *Credit Scoring* no evalúa la capacidad acreedora de un consumidor ni intenta responder el porqué de la situación financiera de alguna persona o entidad financiera, debido a que sólo maneja datos para encontrar patrones o tendencias y a partir de ellos no es posible conocer el pasado o porvenir de los solicitantes a crédito y los posibles sucesos a los que estarán sujetos.
- El método de *Credit Scoring* permite evaluar de manera rápida una base de datos de gran volumen de solicitantes de créditos, sin embargo no es fácil de adaptar a las reglas de decisión de cada entidad financiera.
- Por otro lado, desde el punto de vista estadístico, al desarrollar el método de *Credit Scoring* el analista asume diversas características sobre la estructura de los datos y después usa modelos y pruebas estadísticas para probar o desaprobar lo que asumió inicialmente, no obstante la calidad del modelo depende de la habilidad del analista.

- La labor del analista de encontrar las relaciones entre los datos que son valuados con la ayuda de los procesos del minado de datos permite que éste no sea un experto en la construcción de modelos y eso reduce el costo del análisis.
- Debido a que los mecanismos de evaluación dentro de una compañía se diversifican y obtienen información de diversas fuentes, a veces no del todo fiables, pueden presentarse casos de inestabilidad en la organización y manejo de información.
- Muchas veces al tener una base de datos de gran volumen se vuelve más complejo el analizarla, inclusive este hecho puede facilitar el tener errores al momento de aplicar algún método estadístico que a nivel operativo puede generar pérdidas de gran tamaño.
- Para el desarrollo del método de *Credit Scoring* se debe contar con un panorama general de la situación financiera en la que vive el país y la institución financiera sobre la cual se realiza el análisis, pues el no estar bien informados sobre el clima económico durante y antes del análisis puede implicar una mala toma de decisiones, generando malas predicciones y clasificaciones de futuros acreedores de crédito, lo cual creará futuras pérdidas económicas.

Capítulo 2

Análisis exploratorio de los datos

El objetivo de este capítulo es el de implementar los modelos y métodos antes mencionados con la ayuda de una base de datos de clientes con crédito de consumo tomada de Fahrmeir et al. (1996)[1], quien la utilizó para ejemplificar modelos de regresión logística. La base se puede encontrar con el nombre *Determining the solidness of borrowers via credit scoring.*

Con el fin de ilustrar el análisis exploratorio de datos en la sección 2.1 se usará un análisis descriptivo de la base para conocer detalles de las variables, para mayor información se debe consultar el Anexo D. En la sección 2.2 se implementará el método de *projection pursuit* con el objetivo de encontrar proyecciones interesantes. Por otro lado, en la sección 2.3, con la ayuda de los análisis de las primeras dos secciones, se realiza una recodificación de variables para reducir el número de variables que participen en los modelos que se desarrollarán en el siguiente capítulo. Finalmente, en la sección 2.4 se desarrollará el modelo de componentes principales para saber si es posible reducir la dimensionalidad de la base de datos, así como para identificar aquellas variables de mayor peso entre los clientes.

2.1. Descripción de la base de trabajo, la *base DM*

La base de datos a utilizar, que en adelante será llamada *base DM*, consta de 1,000 registros de clientes con créditos de consumo. Para cada cliente se tiene una covariable binaria de respuesta tal que:

$$kredit = \begin{cases} \text{si } x_i = 1 \Rightarrow \text{Cliente con capacidad crediticia} \\ \text{si } x_i = 0 \Rightarrow \text{Cliente sin capacidad crediticia} \end{cases}$$

[1]La base de datos se encuentra en el sitio:
http : //www.stat.uni − muenchen.de/service/datenarchiv/kredit/kredit_e.html.
La descripción de sus variables se encuentra en Fahrmeir, Hamerle and Tutz. (1996). *Multivariate statistische Verfahren.* Gruyter. Berlín. Second Edition. p. 390.

Variable	**Nombre en la base**	**Nombre en el texto**	**Descripción de la variable**
1	kredit	Crédito	Indica la capacidad crediticia del cliente
2	laufkont	Saldos	Saldos de cuentas
3	laufzeit	Duración	Duración en meses del crédito
4	moral	Morosidad	Pagos de créditos previos
5	verw	Propósito	Propósito del crédito
6	hoehe	Monto	Monto del crédito en DM
7	sparkont	Ahorros	Valor en cuentas de ahorro
8	beszeit	Antiguedad_E	Antigüedad en el empleo
9	rate	Capacidad	Porcentaje de ingresos disponible para el pago
10	famges	Género	Estado Marital / Sexo
11	buerge	Aval	Aval / Garante
12	wohnzeit	Antiguedad_V	Antigüedad en la vivienda
13	verm	Garantías	Posibles bienes en garantía
14	alter	Edad	Edad del cliente en años
15	weitkred	Créditos_O	Otros créditos activos
16	wohn	Vivienda	Tipo de vivienda
17	bishkred	Créditos_DM	Número de créditos previos en el banco DM (incluyendo el que corre)
18	beruf	Ocupación	Ocupación del cliente
19	pers	Dependientes	Número de dependientes
20	telef	Teléfono	Indica si el cliente cuenta con teléfono
21	gastarb	Trabajo	Indica si el cliente tiene trabajo foráneo

Tabla 2.1: Tabla que describe las 21 variables de la *base DM*.

En general, se etiqueta a los créditos de clientes con capacidad crediticia como "créditos buenos" y a los créditos de clientes sin capacidad crediticia como "créditos malos" con el fin de explicar mejor los resultados obtenidos en el desarrollo de la presente tesis.

Adicional a esto, se define que la *base DM* consiste en una muestra estratificada donde, de los $1,000$ registros que se tienen, 700 corresponden a clientes con capacidad crediticia (*i.e.* la variable *kredit* $= 1$) y 300 a clientes sin capacidad crediticia; además cabe mencionar que la base no presenta valores faltantes o *missings*.

Por otro lado, las 20 variables de la base en su mayoría son de carácter cualitativo; tres de ellas se presentan tanto continuas como categóricas y el resto únicamente categóricas. En el Anexo C se encuentra una tabla que desglosa dichas categorías y los posibles valores que pueden tomar.

En la tabla 2.1 se describe el nombre de cada variable y se define el nombre de trabajo (nombre en la base) que se utilizará para el desarrollo de los modelos y métodos.

2.1.1. Estadística descriptiva de la *base DM*

Con base en la tabla del Anexo C se obtiene una idea de los parámetros de la institución financiera para otorgar un crédito y utilizando las últimas columnas referentes al procentaje de frecuencia relativa de los clientes con capacidad crediticia y clientes sin capacidad crediticia se concluye lo siguiente:

- **Información demográfica**: se observa que aproximadamente el 80 % de los créditos son otorgados a hombres, de los cuales el 53 % son casados o viudos, con edades que oscilan entre los 26 y 39 años, que pueden llegar a tener a lo más 2 dependientes económicos y que puede considerarse como trabajador o empleado calificado o con algún puesto menor como funcionario ejerciendo su empleo en su lugar de origen. Por otro lado, se tiene preferencia en clientes con antigüedad en el empleo de entre 1 y 4 años y antigüedad en la vivienda actual de más de 7 años, o bien, entre 1 y 4 años con vivienda rentada; además de no ser aval.

- **Información interna**: el 75 % de los clientes con capacidad crediticia no tiene algún otro crédito corriendo o saldo pendiente en el banco, son solicitantes sin créditos previos en otros bancos o en el mismo banco, o bien, con créditos previos saldados (73 %). Además, no muestran ahorros y el porcentaje de pago que implicaría de sus ingresos es menor al 20 %; cuentan con algún seguro de vida como garantía.

- **Información de los créditos**: en su mayoría son créditos con duración de entre 6 y 24 meses (67 % en promedio entre créditos buenos y créditos malos), aproximadamente el 40 % de los créditos solicitados son para comprar autos usados o muebles con monto solicitado de entre 1,500 y 5,000 marcos alemanes[2].

Sin embargo, este tipo de comparaciones y observaciones es sólo lo que a simple vista los datos muestran, la información será más clara y de mayor veracidad en las siguientes secciones al entrar en detalle mediante estadística descriptiva y exploración de datos.

[2]Se considera esa moneda por la antigüedad de la muestra de datos, nótese que el 60 % de los clientes no tiene teléfono en casa. Para tener un parámetro del monto mencionado, 1.95583 marcos equivalen a un euro, i.e. que aproximadamente dos marcos alemanes son equivalentes a un euro y al proponer el tipo de cambio de un euro en $16.00 M.N. se concluye que el monto promedio a prestar por el banco oscila entre 24,000 y 80,000 pesos; de todas formas se debe considerar el año y situación financiera correspondiente en la cual se otorgaron estos créditos.

Variables continuas

A continuación se muestra la información de las tres variables continuas presentes en la *base DM*:

- Duración en meses del crédito (*laufzeit*)
- Monto del crédito (*hoehe*)
- Edad en años del cliente (*alter*)

Se desarrolló estadística descriptiva para conocer más sobre su comportamiento, los resultados se muestran en la tabla 2.2, basándose en esta información se concluye lo siguiente:

- El valor mínimo en la variable que representa la duración en meses del crédito (*laufzeit*) es de 4 meses y el valor máximo de 72 meses, se analizó junto con la variable de propósito del crédito (*verw*) y se observó que 4 meses no es de gran frecuencia, además de que cae en categorías de préstamos para carros usados, muebles y otros créditos, siendo todos clientes con capacidad crediticia. Por otro lado, la máxima duración se presenta en un crédito para muebles y resulta de un cliente sin capacidad crediticia, éste puede tratarse como un dato atípico.
- La suma total de la variable de monto del crédito (*hoehe*) acumula 3,271,248 marcos alemanes, siendo el monto mínimo de 250 marcos, el máximo de 18,424 y la media de 3,271 marcos.
- Finalmente, en cuanto a la edad de los clientes representada por la variable *alter*, los créditos se han solicitado por clientes con edad mínima de 19 años y máxima de 75, siendo la media de 36 años.

Adicional a esta información, utilizando diagramas de caja, histogramas, tablas de frecuencia y diversos gráficos presentes en el Anexo D se obtiene lo siguiente:

- La variable de duración en meses del crédito solicitado (*laufzeit*) presenta gran dispersión en los datos, además de que la población se encuentra sesgada hacia intervalos de pocos meses (la media es 21 meses), inclusive el histograma presenta dos modas: en los 14 y 24 meses; no se puede decir que sea una variable con distribución semejante a una distribución normal.
- La variable relacionada al monto del crédito (*hoehe*) presenta una sola moda en su histograma, además de estar sesgada totalmente a la izquierda (créditos de montos bajos); realmente sólo el 8.6 % de los créditos presentan montos mayores a 10,000 marcos alemanes.

Estadísticos básicos de variables continuas

Estadístico	Duración en meses del crédito ($laufzeit$)	Monto del crédito ($hoehe$)	Edad en años del cliente ($alter$)
N	1,000	1,000	1,000
Rango	68	18,174	56
Mínimo	4	250	19
Máximo	72	18,424	75
Suma	20,903	3,271,248	35,542
Media	21	3,271	36
Err. Estándar	0	89	0
Des. Estándar	12	2,823	11
Varianza	145	7,968,000	129

Tabla 2.2: Tabla que describe las 3 variables continuas de la *base DM*.

- La variable asignada a la edad en años del cliente (*alter*) presenta una media de 28 años y una concentración en un rango de edad más amplio de entre 24 y 38 años. Es la variable con distribución más semejante a una normal.

Finalmente, se analiza si las tres variables continuas presentan alguna correlación, para ellos se utiliza la gráfica *Pairs* (figura 2.1) y la variable *Crédito* (*kredit*) a la cual se define nuevamente bajo el siguiente esquema:

$$kredit = \begin{cases} B & \text{si } x_i = 1, \forall i \\ \\ M & \text{si } x_i = 0, \forall i \end{cases}$$

Donde B significa buena calidad crediticia y M mala calidad crediticia. Sin embargo, para una mejor visialización de los datos también se asignan colores:

$$kredit = \begin{cases} rojo & \text{si } x_i = M, \forall i \\ \\ azul & \text{si } x_i = B, \forall i \end{cases}$$

De esta forma, la figura 2.1 muestra que la duración en meses del crédito graficado contra la edad del cliente (*alter*, *laufzeit*) es mayor mientras tenga mayor edad. Sin embargo, disminuye el número de clientes con créditos buenos (puntos color azul) conforme avanza la edad y el número de meses; y para los clientes de créditos malos (puntos color rojo) se nota que a edades cortas hay un gran número de clientes sin

capacidad crediticia sin importar el monto del mismo. Conforme avanza la edad se ven menos clientes sin capacidad crediticia (puntos color rojo), aunque realmente el monto del crédito no parece relacionarse tanto con la edad. Por último, se observan las coordenadas de las variables de duración en meses del crédito y monto del crédito otorgado (*laufzeit*, *hoehe*), donde se aprecia que no hay clientes con créditos malos con montos de entre 400 y 600 marcos alemanes y pocos meses de duración del crédito, inclusive en montos un poco menores a 600 marcos alemanes son casi todos créditos buenos con duración menor o igual a 25 meses.

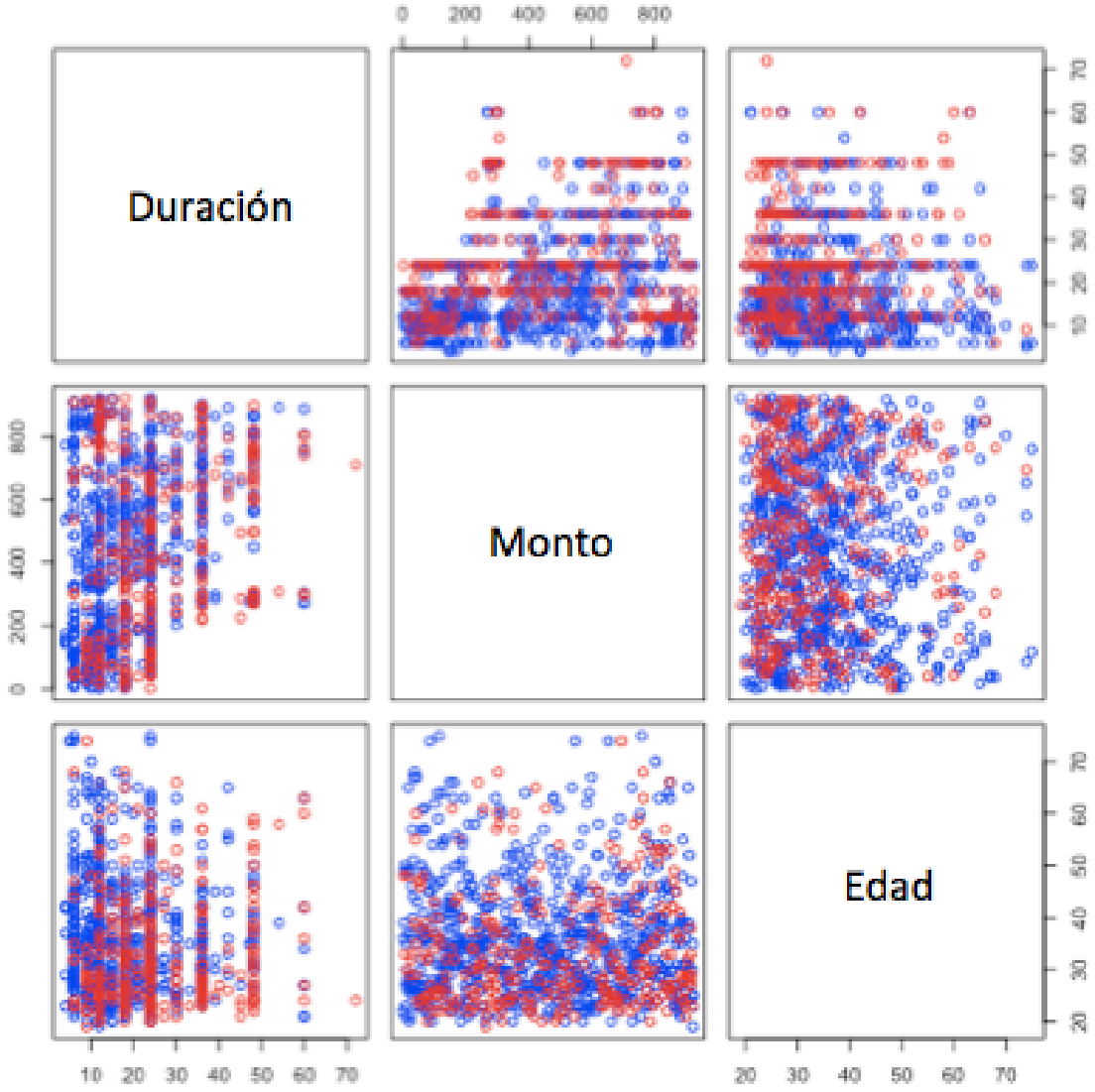

Figura 2.1: Gráfica *pairs* de las 3 variables continuas contenidas en la *base DM*: duración en meses del crédito (*laufzeit*), monto del crédito (*hoehe*) y edad en años del cliente (*alter*), representados en color azul los clientes con capacidad crediticia y en color rojo los clientes sin capacidad crediticia.

Variables discretas

Para el caso de las variables discretas se ha obtenido nuevamente una tabla que resume los estadísticos descriptivos, tabla 2.3, sin embargo esta información actúa principalmente como auditor de datos que indicará en su momento si la base cuenta con inconsistencias.

Estadísticos básicos de variables discretas

Nombre en el texto	Nombre en la base	Rango	Mínimo	Máximo	Moda
Crédito	kredit	1	0	1	1
Saldos	laufkont	3	1	4	4
Morosidad	moral	4	0	4	2
Propósito	verw	10	0	10	3
Ahorros	sparkont	4	1	5	1
Antiguedad_E	beszeit	4	1	5	3
Capacidad	rate	3	1	4	4
Género	famges	3	1	4	3
Aval	buerge	2	1	3	1
Antiguedad_V	wohnzeit	3	1	4	4
Garantías	verm	3	1	4	3
Créditos_O	weitkred	2	1	3	3
Vivienda	wohn	2	1	3	2
Créditos_DM	bishkred	3	1	4	1
Ocupación	beruf	3	1	4	3
Dependientes	pers	1	1	2	1
Teléfono	telef	1	1	2	1
Trabajo	gastarb	1	1	2	1

Tabla 2.3: Tabla que describe las variables discretas de la *base DM*.

Intentar obtener gráficas de dispersión de variables categóricas resulta poco ilustrativo, sin embargo, se pueden analizar las relaciones entre las variables discretas y las continuas revisando las mayores frecuencias de cada categoría en el caso de las variables discretas, así como las modas, varianzas y demás estadísticas de las variables continuas. Como ejemplo obsérvense en la figura 2.2 los histogramas de seis de las variables de la *base DM*: saldo de cuentas en el banco (*laufkont*), duración en meses del crédito (*laufzeit*), pago de créditos previos (*moral*), monto del crédito en marcos alemanes (*hoehe*), estado marital y género del cliente (*famges*) y edad en años del cliente (*alter*) donde se tiene que:

- En el histograma de la variable de duración en meses del crédito (*laufzeit*), la categoría 4 se refiere a créditos para al menos un año de duración o que éste

sea de un monto mayor a 200 marcos, mientras que en menor proporción la categoría 3 que es equivalente a créditos de entre 0 y 200 DM.

- En el histograma de la variable de saldo de cuentas en el banco (*laufkont*) se observa que la mayoría de los créditos están dentro de una periodicidad de 5 a 25 meses.
- En el histograma de la variable de pago de créditos previos (*moral*) se tiene que la mayoría de los créditos son por los clientes que pertenecen a la categoría 2, es decir aquellos que no presentan solicitudes de crédito previas o que tienen todos sus créditos anteriores saldados.
- El histograma de la variable del monto del crédito en marcos alemanes (*hoehe*) indica que la mayoría de los créditos son de montos menores a 50,000 marcos alemanes.
- El histograma de la variable del estado marital y género del cliente (*famges*) menciona que la mayoría de créditos son por personas de la categoría 3, hombres casados o viudos.
- Y finalmente en el histograma de la variable de edad en años del cliente (*alter*) se observa que la mayoría de los clientes están entre las edades de 20 y 40 años, siendo los de 25 a 30 años los de mayor frecuencia y de 40 en adelante la frecuencia disminuye considerablemente.

2.2. Exploración de la *base DM* mediante *projection pursuit*

Después de realizar el análisis de las variables con estadística descriptiva se procederá a la exploración mediante la técnica de *projection pursuit*.

Vos and Evers (2004, p.28) refieren a Friedman y Tukey (1974) para exponer la exploración de *projection pursuit* como una técnica para la visualización de datos de alta dimensión. A su vez, Vos and Evers (2004) comentan: "Con esto se refieren a tomar una base de $k - dimensiones$ y transformarla a 1 o 2 dimensiones a razón de facilitar su visualización, de tal manera que definen esta técnica como *una técnica para encontrar proyecciones lineales* interesantes."

Por otro lado, citando de nuevo a Vos and Evers (2004, p.28): "Diaconis y Freedman (1984) muestran que una proyección seleccionada aleatoriamente de un conjunto de datos de alta dimensión tendrá un aspecto similar a una muestra de una distribución normal multivariante. Lo anterior implica que la mayoría de las proyecciones de grandes dimensiones a una o dos dimensiones de datos parecerá similar a una

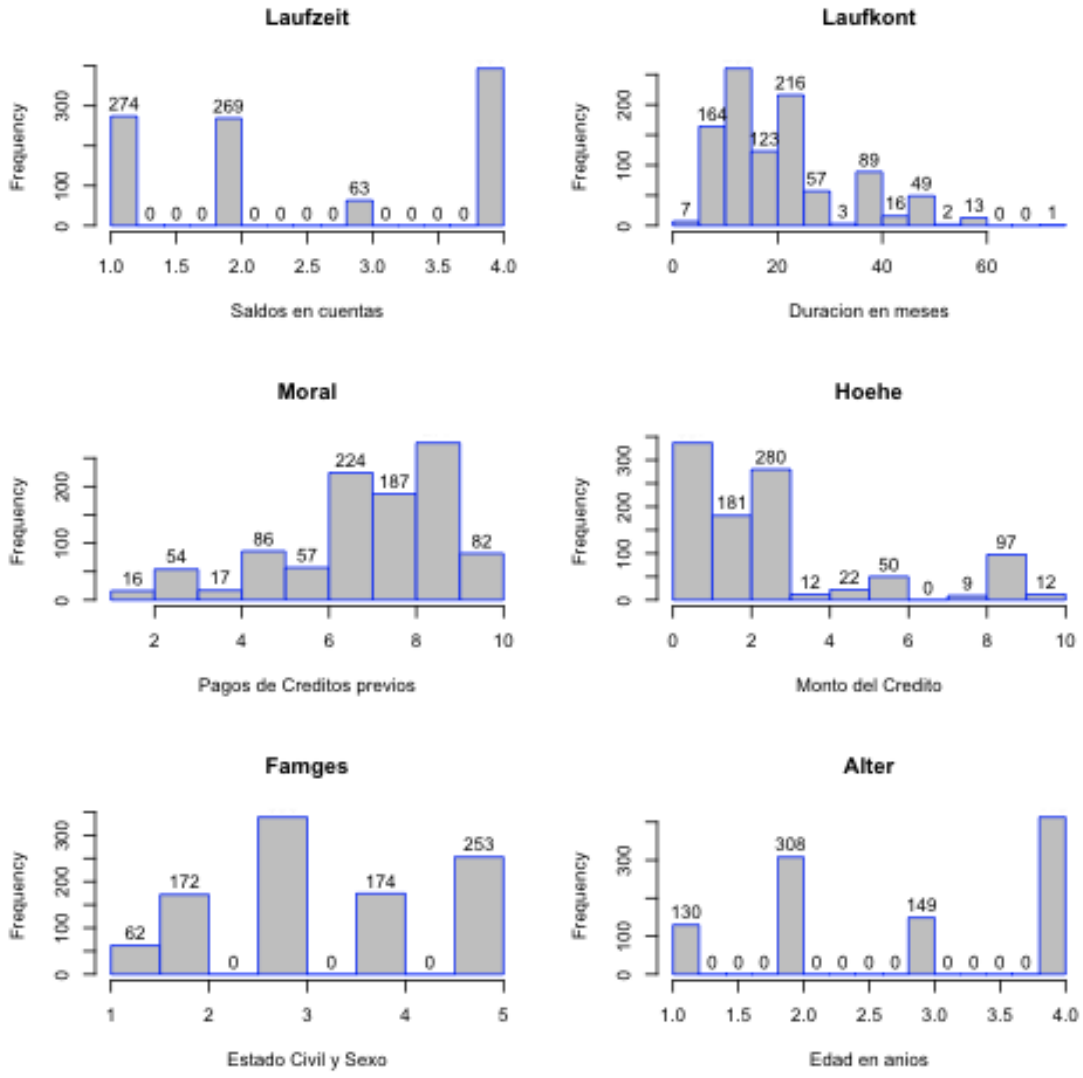

Figura 2.2: Histograma de seis variables presentes en la *base DM*: Saldos de cuentas (*laufkont*), duración del crédito en meses (*laufzeit*), pagos de créditos previos (*moral*), monto del crédito (*hoehe*), estado civil y sexo del cliente (*famges*) y edad en años del cliente (*alter*), donde las variables de la columna derecha son de tipo continuas.

muestra proveniente de una distribución gaussiana."

De lo anterior se puede definir *projection pursuit* como una proyección interesante si ésta es muy diferente o lo más diferente a una distribución gaussiana, por lo que esta técnica se basa en generar indicadores que midan la no-normalidad de la distribución. Para lo ya mencionado deben de tenerse en cuenta las siguientes consideraciones:

1. El tiempo computacional que será requerido para ejecutarlo, así como las herramientas adecuadas y,

2. La sensibilidad a *outlies*[3].

Ripley (1996) comenta que una vez encontrada una proyección interesante es importante quitar la estructura que revela para permitir otras visualizaciones interesantes, es por ello que este método es particularmente importante para el mejor conocimiento de la base, pues permite una mejor toma de decisiones del cómo reestructurar variables, agrupar observaciones, o bien, darle algún seguimiento especial a determinados datos o variables.

Existen diversos criterios para obtener una *proyección interesante*, uno de ellos es el de componentes principales que es una proyección de p dimensiones en 2, ésta es una de varias aplicaciones de este método que se abordará más adelante.

Ahora bien, no se debe confundir la técnica de *projection pursuit* con el de *minado de datos*, ya se mencionó que en el minado se buscan *patrones o características interesantes* y en *projection pursuit* se buscan *proyecciones lineales interesantes*, los cuales son dos conceptos que van de la mano para la mejor comprensión de los datos.

Por otro lado, no se debe de perder como objetivo en la exploración de datos que a su vez se busca generalizar y no particularizar el resultado. Es decir, que una explotación adecuada con hallazgo de patrones útiles es aquella en que estos patrones serán similares si se toma otra muestra de los datos y se explora, por lo que no se trata de patrones particulares de una muestra, sino de la población en sí.

Para ejemplificar la técnica de *projection pursuit* se desarrollaron dos visualizaciones utilizando la librería *rggobi* del *software R* con la vista *2D Tour*, el índice *Holes* para optimizar y con la opción de una manipulación *Oblique*. Estas proyecciones se aprecian en las figuras 2.3 y 2.4 de las páginas 23 y 24 respectivamente. Estos gráficos tienen como objetivo proyectar la relación entre las variables de propósito del crédito (*verw*), monto del crédito (*dhoehe*), duración en meses del crédito (*dlauzeit*), estado

[3]Se define un *outlier* como un valor inusual que normalmente corresponde a una observación errónea dentro de la base de datos. Dichos datos suelen extraerse de la muestra para que su valor no afecte los resultados de los análisis.

marital y género del cliente (*famges*) y la capacidad crediticia del cliente (*kredit*)

En la figura 2.3 se aprecia la dispersión de la variable de duración en meses del crédito (*dlauzeit*) y su tendencia a la poca duración (se debe recordar que la categoría 1 pertenece a créditos mayores a 54 meses), mientras que en la figura 2.4 se aprecia que la categoría 3 (hombre casado o viudo) es la de mayor presencia en la base, que presenta créditos de mayor monto, así como mayor variabilidad en el propósito del crédito.

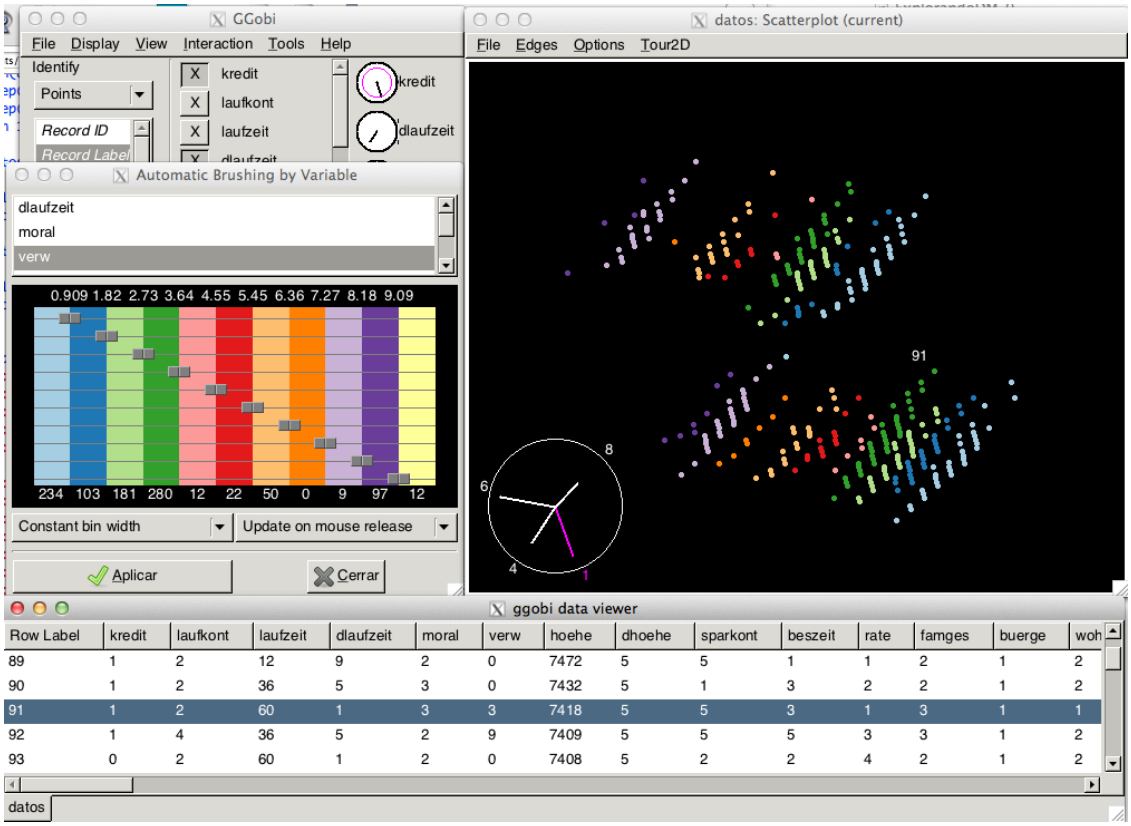

Figura 2.3: Exploración de la base con ayuda del *software R* y su librería *rggobi*. Observando la gráfica con fondo negro: en la parte inferior izquierda se observan los ejes de la gráfica donde se denota con 1 a la variable *kredit*, con 4 la variable de duración en meses del crédito (*dlauzeit*), con 6 la variable del propósito del crédito (*verw*) y con 8 la variable relacionada al monto del crédito (*dhoehe*). La gráfica está coloreada con base en la variable que indica el propósito del crédito: de azul claro (primera capa de color de derecha a izquierda) se muestra la categoría cero: otros, de color azul la categoría 1: carro nuevo; de verde claro la categoría 2: carro usado; de verde la 3: muebles; de rosa la 4: radio/televisión; de rojo la 5: electrodomésticos; de naranja claro la 6: reparaciones; de naranja la 8: vacaciones; de lila la 9: capacitación y de morado la 10: negocios. Con este gráfico se aprecia la dispersión de la variable de duración en meses del crédito (*dlauzeit*) y su tendencia a la poca duración (se debe recordar que la categoría 1 pertenece a créditos mayores a 54 meses). Se ha puntualizado el registro 91 para ejemplificar la lectura de la gráfica.

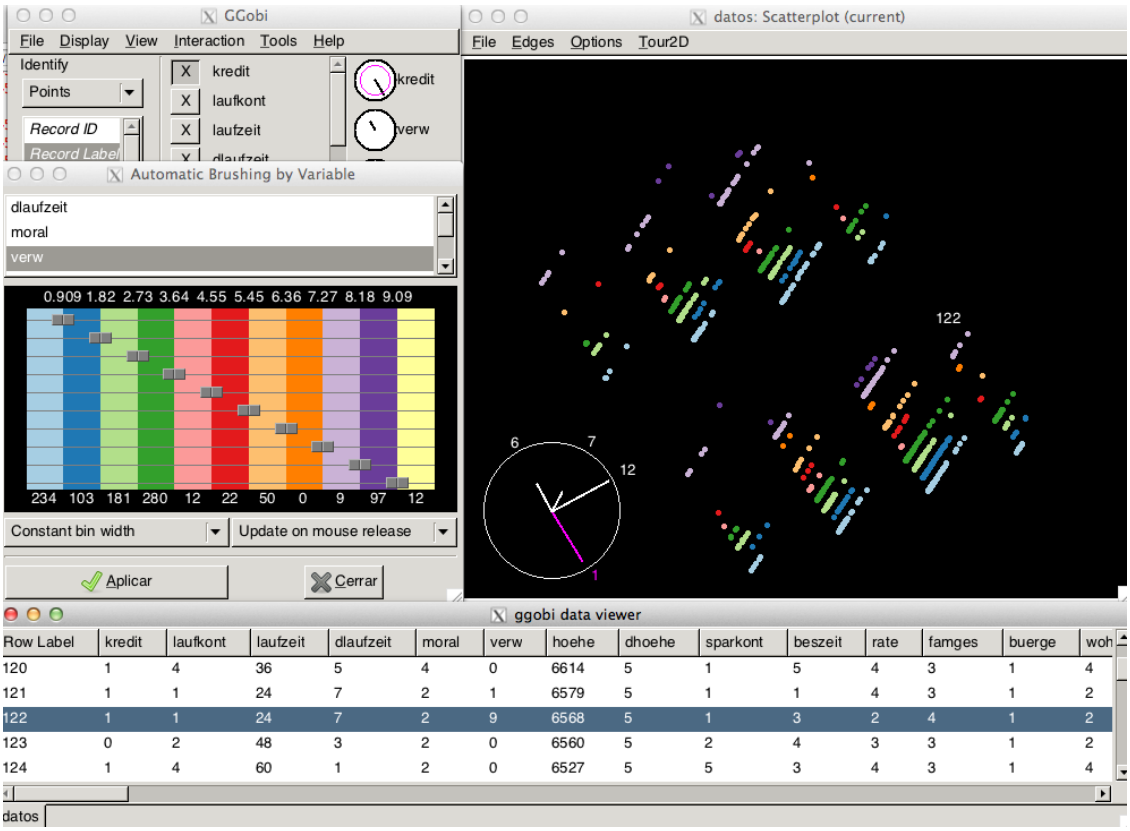

Figura 2.4: Exploración de la base con ayuda del *software R* y su librería *rggobi*. Observando la gráfica con fondo negro: en la parte inferior izquierda se observan los ejes de la gráfica donde se denota con 1 a la variable *kredit*, con 6 la variable del propósito del crédito (*verw*), con 7 la variable de tipo continua de monto del crédito (*hoehe*) y con 12 la variable relacionada al estado marital y sexo del cliente (*famges*). La gráfica está coloreada con base en la variable que indica el propósito del crédito: de azul claro (primera capa de color de derecha a izquierda) se muestra la categoría cero: otros, de color azul la categoría 1: carro nuevo; de verde claro la categoría 2: carro usado; de verde la 3: muebles; de rosa la 4: radio/televisión; de rojo la 5: electrodomésticos; de naranja claro la 6: reparaciones; de naranja la 8: vacaciones; de lila la 9: capacitación y de morado la 10: negocios. Con este gráfico se aprecia que la categoría 3 (hombre casado o viudo) es la de mayor presencia en la base, que presenta créditos de mayor monto, así como mayor variabilidad en el propósito del crédito. Se ha puntualizado el registro 122 para ejemplificar la lectura de la gráfica.

2.3. Recodificación de variables

Con ayuda de las exploraciones estadísticas anteriores, los histogramas, tablas de frecuencia del Anexo D y un poco de exploración de las variables, se buscó disminuir el número de categorías por variable al colapsar categorías con frecuencias muy bajas o que se refieran a características similares. De igual forma se descartaron aquellas variables que eran muy generales y no marcaban alguna diferencia en la base. Por ejemplo, el 96 % de los clientes tienen trabajo foráneo (variable *gastarb*), esto indica una característica general de la población pero no caracteriza al segmento de clientes que tienen capacidad crediticia de los clientes que no la tienen, sucede de igual manera con la variable que indica si el cliente es aval o garante (*buerge*) o la variable que indica el saldo en cuenta de ahorro en el banco DM (*sparkont*), por otro lado se descartan también variables que proporcionalmente son iguales en ambos tipos de cliente (con capacidad crediticia y sin capacidad crediticia), por ejemplo las variables que indican si el cliente tiene teléfono (*telef*), el tipo de vivienda (*wohn*), antigüedad en la vivienda y empleo (*wonhzeit* y *beszit*). Dado lo anterior se llegó a las siguientes conclusiones:

Selección de variables de interés y reducción de categorías de las mismas:

1. Variable de clasificación de capacidad crediticia del cliente (*kredit*), sin modificación.

2. Saldos de cuentas en el *banco DM* (*laufkont*), se colapsan las categorías 3 y 4 pues la categoría 3 es muy pequeña y el 56.7 % de los clientes con capacidad crediticia cae en estas categorías: 1= Sin cuenta vigente, 2 = Sin saldo, 3 = Saldo $0 \geq$ o cuentas de cheques con al menos un año.

3. Duración en meses del crédito (*laufzeit* y *dlaufzeit*), se colapsó en una sola las categorías 1, 2, 3, 4 y 5, pues como se ha visto más del 80 % de las observaciones se encuentran en periodos cortos menores a los 30 meses: $1 = < 6$, $2 = 6 < \ldots \leq 12$, $3 = 12 < \ldots \leq 18$, $4 = 18 < \ldots \leq 24$, $5 = 24 < \ldots \leq 30$ y $6 = >30$.

4. Pago de créditos previos (*moral*), se colapsan las categorías 0 y 1, pues ambas reflejan problemas en créditos previos e incumplimiento con otras instituciones, además de que ambas categorías son pequeñas; también se colapsan las categorías 3 y 4, pues como contraparte de las categorías 0 y 1, ambas categorías califican al cliente como capaz de adquirir otro crédito: 1 = Atraso en pagos de créditos previos / Problemas en créditos vigentes / Con créditos vigentes en otros bancos, 2 = Sin créditos previos / Créditos previos saldados y 3 = Sin problemas con créditos vigentes en banco DM / Saldados todos los créditos previos en el banco DM.

5. Propósito del crédito (*verw*), se colapsan las categorías 4 y 5 por tratarse de electrodomésticos y aparatos de línea blanca para el hogar; las categorías 6 y 8

por presentar similitudes en las características que definen al crédito como son monto y duración en meses del crédito (se puede usar como apoyo las figuras 2.3 y 2.4); finalmente también se colapsan las categorías 9 y 10 tomando como base las razones anteriores y que ambos créditos están enfocados al mismo giro o ramo. Cabe mencionar que la categoría 7 desaparece por tener frecuencia igual a cero: 1 = Otros créditos, 2 = Carro nuevo, 3 = Carro usado, 4 = muebles, 5 = Radio / Televisión / Electrodomésticos, 6 = Reparaciones / Vacaciones y 7 = Capacitación / Negocios.

6. Monto del crédito (*hoehe* y *dhoehe*), para la variable categórica se colapsan en una sola las categorías 1, 2, 3 y 4, estas 4 categorías acumulan sólo el 8.6 % de las observaciones, además de que va de la mano de la política de estos créditos donde se muestran cortos periodos de duración, así como montos menores a los $7,500$ marcos alemanes: $1 = \leq 500$, $2 = 500 < \ldots \leq 1,000$, $3 = 1,000 < \ldots \leq 1,500$, $4 = 1,500 < \ldots \leq 2,500$, $5 = 2,500 < \ldots \leq 5,000$, $6 = 5,000 < \ldots \leq 7,500$ y $7 = > 7,500$.

7. Porcentaje de ingresos disponible para el pago (*rate*), se invirtió el orden de las categorías para generar una variable categórica ordenada: $1 = < 20$, $2 = 20 \leq \ldots < 25$, $3 = 25 \leq \ldots < 35$ y $4 = \geq 35$.

8. Estado marital y género (*famges*), sin modificar.

9. Edad en años del cliente (*alter* y *dalter*), para la variable categórica se agrupan las categorías 4 y 5, pues sólo el 5 % del total de la muestra caen en estas dos categorías: $1 = 0 \leq \ldots \geq 25$, $2 = 26 \leq \ldots \geq 39$, $3 = 40 \leq \ldots \geq 59$ y $4 = \leq 60$.

10. Otros créditos corriendo (*weitkred*), sin modificar.

11. Ocupación (*beruf*), se agrupan las categorías 1 y 2 pues la primer categoría es muy pequeña: 1 = Desempleado / No residente / Sin Preparación / Residencia Permanente, 2 = Trabajador calificado / Funcionario menor y 3 = Ejecutivo / Autoempleo / Alto funcionario.

Adicional a esto se ordenaron todos los números de las categorías de menor a mayor iniciando con el número uno. A continuación se muestra la tabla que describe la recodificación de las 10 variables de la *base DM* previamente seleccionadas para desarrollar los modelos en las siguientes secciones. También se presentan las variables binarias que se obtiene al dicotomizar las categorías de cada covariable, para este desarrollo se deja una categoría de referencia de tal manera que de n categorías en una variable se obtienen $n - 1$ covariables binarias.

Tabla de transformación de variables a utilizar de la *base DM*, parte 1

Variable	Nombre en base	Nombre en texto	Descripción	Valor	Categorías recodificadas	Dicotomización de variables Dc1	Dc2	Dc3	Dc4	Dc5	Dc6
1	kredit	Crédito	Capacidad crediticia del cliente	1	Cliente con capacidad crediticia	1	0				
				0	Cliente sin capacidad crediticia	0	0				
2	laufkont	Saldos	Saldo de cuentas	1	Sin cuenta vigente	1	0				
				2	Sin saldo	0	1				
				3	Saldos $0 \geq$ o cuenta de cheques con al menos un año	0	0				
3	laufzeit	Duración	Duración en meses del crédito (continua)								
3b	dlaufzeit	Saldos	Duración en meses del crédito	1	< 6	0	0	0	0	0	
				2	$6 < \ldots \leq 12$	1	0	0	0	0	
				3	$12 < \ldots \leq 18$	0	1	0	0	0	
				4	$18 < \ldots \leq 24$	0	0	1	0	0	
				5	$24 < \ldots \leq 30$	0	0	0	1	0	
				6	> 30	0	0	0	0	1	
4	moral	Morosiodad	Pago de créditos créditos previos	1	Atraso en pagos de créditos previos/ Problemas en créditos vigentes / Con créditos vigentes en otros bancos	0	0				
				2	Sin créditos previos / Créditos previos saldados	1	0				
				3	Sin problemas con créditos vigentes en banco DM / Saldados todos los créditos previos en el banco DM	0	1				
5	verw	Propósito	Propósito del crédito	1	Otros Créditos	0	0	0	0	0	0
				2	Carro nuevo	1	0	0	0	0	0
				3	Carro nuevo	0	1	0	0	0	0
				4	Muebles	0	0	1	0	0	0
				5	Radio / Televisión / Electrodomésticos	0	0	0	1	0	0
				6	Reparaciones / Vacaciones	0	0	0	0	1	0
				7	Capacitación / Negocios	0	0	0	0	0	1

Para cada variable, la categoría de referencia que aparece con el valor de cero en todas sus variables dicotómicas derivadas corresponde a la categoría de referencia.

Tabla de transformación de variables a utilizar de la *base DM*, parte 2

Variable	Nombre en base	Nombre de en texto	Descripción	Valor	Categorías recodificadas	Dicotomización de variables Dc1	Dc2	Dc3	Dc4	Dc5	Dc6
6	hoehe	Monto	Monto del crédito otorgado por el banco DM (continua)								
6b	dhoehe	Monto	Monto del Crédito otorgado por el banco DM	1	≤ 500	1	0	0	0	0	
				2	$500 < \ldots \leq 1,000$	0	1	0	0	0	
				3	$1,000 < \ldots \leq 1,500$	0	0	1	0	0	
				4	$1,500 < \ldots \leq 2,500$	0	0	0	1	0	
				5	$2,500 < \ldots \leq 5,000$	0	0	0	0	0	
				6	$5,000 < \ldots \leq 7,500$	0	0	0	0	1	
				7	$> 7,500$	0	0	0	0	0	
7	rate	Capacidad	Porcentaje de ingresos disponibles para el pago	1	< 20	1	0	0			
				2	$20 \leq \ldots < 25$	0	1	0			
				3	$25 \leq \ldots < 35$	0	0	1			
				4	≥ 35	0	0	0			
8	famges	Género	Estado Marital/ Género	1	Hombre divorciado / vive aparte	1	0	0			
				2	Mujer divorciada / vive aparte / Hombre soltero	0	0	0			
				3	Hombre casado / viudo	0	1	0			
				4	Mujer soltera	0	0	1			
9	alter	Edad	Edad en años del cliente (continua)								
9b	dalter	Edad	Edad del cliente en años	1	$0 \leq \ldots \geq 25$ años	0	0	0			
				2	$26 \leq \ldots \geq 39$ años	1	0	0			
				3	$40 \leq \ldots \geq 59$ años	0	1	0			
				4	≤ 60 años	0	0	1			
10	weitkred	Créditos_O	Otros Crédito	1	En otros bancos	0	0				
				2	Tiendas departamentales	1	0				
				3	Sin otros créditos vigentes	0	1				
11	beruf	Ocupación	Ocupación del cliente	1	Desempleado / No residente / Sin Preparación/ Residencia Permanente	0	0				
				2	Trabajador calificado / Funcionario menor	1	0				
				3	Ejecutivo / Autoempleo / Alto funcionario	0	1				

Para cada variable, la categoría de referencia que aparece con el valor de cero en todas sus variables dicotómicas derivadas corresponde a la categoría de referencia.

2.4. Componentes principales

A continuación se menciona el análisis de componentes principales y se implementa sobre la *base DM* a la cual previamente se le han recodificado las variables según la tabla "Transformación de variables a utilizar de la *base DM*" de la sección anterior, para este desarrollo se utilizan las variables dicotomizadas previamente definidas.

Para definir el análisis de componentes principales se hace referencia a Jolliffe (2002, p.7), quien comenta que el análisis de componente principales es quizá la más vieja técnica del análisis multivariado. Según menciona Jolliffe, esta técnica fue introducida por Pearson al desarrollar ajustes ortogonales por mínimos cuadrados y después fue desarrollada y llevada a la técnica por Hotelling.

Como es el caso de muchos métodos multivariados, esta técnica no fue cotidianamente usada hasta contar con la presencia de avances tecnológicos en equipos de cómputo. Sin embargo, actualmente se puede encontrar en cualquier *software* estadístico, ya sea como "análisis de componentes principales", o bien, por sus siglas en inglés *PCA* o *PC* (*principal components analysis*).

2.4.1. Definición del modelo de componentes principales

El desarrollo del análisis de componentes principales se basa en reducir la solución a *eigenvalores* y *eigenvectores* de la matriz de varianzas y covarianzas o de la matriz de correlación, de tal forma que permite representar de manera óptima, en un espacio de dimensión pequeña, a observaciones de un espacio general $p-$dimensional con base en la varianza de las variables. En este sentido, el análisis de componentes principales es el primer paso para identificar las posibles variables *latentes*, o no observadas, que generan los datos.

Lo anterior es posible plantearlo desde tres enfoques diferentes:

- **Enfoque geométrico.** Se desea encontrar un subespacio de menor dimensión p tal que al proyectar sobre él los puntos éste conserve su estructura con la menor distorsión posible. Esto se logra trasponiendo una línea recta sobre la nube de puntos tal que se exija que las distancias entre los posibles puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles, *i.e.* se busca minimizar las distancias ortogonales (figura 2.5).

- **Enfoque estadístico.** Se basa en representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión uno que es equivalente a sustituir las p variables originales por una nueva variable, Y_1, que resuma óptimamente la información y que para minimizar la pérdida de

información de los datos observados se utilizó la variable de máxima varianza. Este enfoque se desarrollará a continuación.

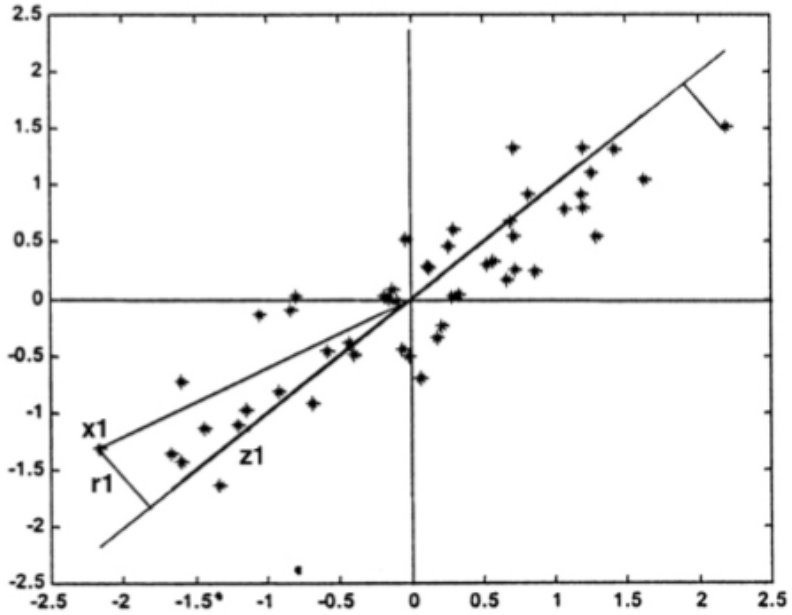

Figura 2.5: Ejemplo de la recta que minimiza las distancias ortogonales de los puntos a ella denotada por z_1 y siendo x_1 el punto a proyectar se define a r_1 como la distancia entre el punto y su proyección. Este método se ocupa para encontrar proyecciones lineales interesantes que permitan representar de manera óptima un espacio de dimensión pequeña, dicho método es conocido como componentes principales. Fuente: Peña (2002, p. 135).

Tomando como base el enfoque estadístico, considere que por el momento se conoce la matriz de covarianzas del vector $\mathbf{X}$ de p variables aleatorias[4] *i.e.* $\mathbf{V}$ y que se ha restado a cada variable su media, de manera que las variables de la matriz $\mathbf{X}$ tienen media cero y su matriz de covarianzas viene dada por $\mathbf{V} = (\mathbf{1/n} * \mathbf{X'X})$ [5] por lo que ahora el problema es encontrar el vector $\mathbf{a}$ que maximiza la forma cuadrática $\mathbf{a}'\mathbf{Va}$ sujeta a la restricción $\mathbf{a}'\mathbf{Ia} = \mathbf{1}$, *i.e.* Max $\mathbf{a}'\mathbf{Va}$. Una aproximación estándar a este problema de maximización consiste en aplicar la técnica de Multiplicadores de Lagrange, lo que lleva a la siguiente expresión:

$$Max[\mathbf{a_1}'\mathbf{Va_1} - \lambda(\mathbf{a_1}'\mathbf{a_1} - \mathbf{1})]. \tag{2.1}$$

[4]En un caso más realista donde la matriz de covarianzas es desconocida se reemplaza por una matriz de covarianzas de una muestra, es decir, una matriz de dispersión muestral; para conocer más al respecto ver [8], capítulo 3.

[5]En este desarrollo no necesariamente se hacen supuestos distribucionales sobre las $x_1, \ldots, x_p$ variables y no existe algún modelo subyacente que implique hacer inferencia alguna, sino que es una forma diferente de expresar los datos.

Se desarrolla la expresión y al considerar que λ es el multiplicador de Lagrange se obtiene:

$$\begin{aligned} Max[\mathbf{a_1'Va_1} - \lambda(\mathbf{a_1'a_1} - \mathbf{1})] &= Max[\mathbf{a_1'Va_1} - \lambda\mathbf{a_1'Ia_1} - \lambda] \\ &= Max[\mathbf{a_1'}(\mathbf{V} - \lambda\mathbf{I})\mathbf{a_1} - \lambda] \\ &= Max[\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{a_{1j}V_{ij}a_{1j}}], \end{aligned} \tag{2.2}$$

de tal forma que la ecuación 2.1 queda como sigue:

$$\begin{aligned} &Max_{a_{1j}}[\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{a_{ij}V_{ij}a_{1j}} - \lambda\sum_{\mathbf{i=1}}^{\mathbf{p}}\mathbf{a_{1i}^2} - \lambda] \\ &= f(a_{11}, a_{12}, \ldots, a_{1p}), \end{aligned} \tag{2.3}$$

donde además $\mathbf{V_{ij}} = \mathbf{V_{ji}}\ \forall\ i, j$ por simetría de la matriz $\mathbf{V}$. Derivando con respecto a $\mathbf{a_1}$ se tiene:

$$\mathbf{Va_1} - \lambda\mathbf{a_1} = \mathbf{0} \tag{2.4}$$

o bien,

$$(\mathbf{V} - \lambda\mathbf{I_p})\mathbf{a_1} = \mathbf{0} \tag{2.5}$$

$$\Longleftrightarrow |\mathbf{V} - \lambda\mathbf{I}| = \mathbf{0}. \tag{2.6}$$

De esta forma, λ es el *eigenvalor* de $\mathbf{V}$ y $\mathbf{a_1}$ es el *eigenvector* correspondiente tal que a cada λ le corresponde un *eigenvector* de $\mathbf{V}$. En términos generales la ecuación tiene asociadas p raíces características o valores propios o bien *eigenvalores*, $\lambda_1, \lambda_2, \ldots, \lambda_p$ y gracias a la propiedad de la matriz $\mathbf{V}$ de simetría y al ser positiva definida se logra que sus raíces sean reales y positivas pero, en el caso de existir dos raíces iguales se dice que la elipsoide presenta un corte transversal o *cross-section*.[6]

Ahora, se tiene $\mathbf{X}$= $(x_1, x_2 \ldots x_p)$ y al retomar que la estructura de las covarianzas o correlaciones entre las p variables resulta interesante[7], entonces se comienza a partir de una función lineal $a_1'\mathbf{X}$ tal que los elementos de $\mathbf{X}$ tengan varianza máxima, a_1 es el vector de p constantes $a_{11}, a_{12} \ldots a_{1p}$ y$'$ denota una matriz transpuesta tal que:

[6]Para mayor desarrollo de esta sección consultar Jolliffe (2002, Capítulo 2).

[7]Cabe mencionar que este método se concentra en las varianzas de esta población.

$$\mathbf{Y_1} = \mathbf{a_1^{'}X} = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p = \sum_{j=1}^{p} \mathbf{a_{ij}x_j}, \tag{2.7}$$

visto como vector:

$$\mathbf{Y_1} = (a_{11}, \ldots, a_{1p}) \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = \mathbf{a_1^{'}X}.$$

De tal manera que la siguiente función lineal $\mathbf{a_2^{'}X}$ donde de igual forma los elementos de $\mathbf{X}$ presentan varianza máxima y, sucesivamente la función lineal $bfa_k^{'}X$ tal que ninguna de las funciones definidas con anterioridad presenta correlación y además, la k-ésima función, $\mathbf{a_k^{'}X}$, representa al k-ésimo componente principal:

$$\begin{aligned} \mathbf{Y_1} = \mathbf{a_1^{'}X} &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \\ \mathbf{Y_2} = \mathbf{a_2^{'}X} &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \\ &\vdots \\ \mathbf{Y_k} = \mathbf{a_k^{'}X} &= a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kp}x_p, \end{aligned} \tag{2.8}$$

donde:

$$Var(\mathbf{Y_1}) > Var(\mathbf{Y_2}) > \ldots > Var(\mathbf{Y_k}) \tag{2.9}$$

se define

$$Var(\mathbf{Y_1}) = Var(\mathbf{a_1^{'}X}) = \mathbf{a_1^{'}}Var(\mathbf{X})\mathbf{a_1} = \mathbf{a_1^{'}Va_1}, \tag{2.10}$$

tal que se expresa la norma del vector como:

$$\sum_{j=1}^{p} \mathbf{a_{ij}^2} = 1 = \mathbf{a^{'}I_p a}. \tag{2.11}$$

Donde $\mathbf{I_p}$ es la matriz identidad $(p \times p)$, $\mathbf{V}$ es una matriz tal que $(i,j)-$ésimo elemento representa la covarianza (la cual se supone conocida), con el $i-$ésimo y $j-$ésimo elementos de $\mathbf{X}$ donde $i \neq j$ y la varianza del $j-$ésimo elemento de $\mathbf{X}$ cuando $i = j$.

Dado el conjunto de ecuaciones 2.8 y suponiendo $k = 1, 2, \ldots, p$ el $k-$ésimo componente principal está dado por $\mathbf{Y_k} = \mathbf{a_k'X}$ donde $\mathbf{a_k'}$ es un *eigenvector* de $\mathbf{V}$ que corresponde al $k-$ésimo *eigenvalor* de mayor tamaño, λ_k. Además, si $\mathbf{a_k}$ está representado por una unidad, i.e. $\mathbf{a_k'a_k} = \mathbf{1}$, entonces $Var(\mathbf{Y_k}) = \lambda_\mathbf{k}$ donde $Var(\mathbf{Y_k})$ representa la varianza de la función lineal $\mathbf{Y_k}$.

2.4.2. Interpretación del análisis de componentes principales

Al interpretar los resultado del modelo usualmente se usa la frase: "Las primeras k componentes explican el x % de la varianza de la muestra", esta expresión nace como una de las propiedades de los componentes principales y se debe relacionar a la traza de $\mathbf{V}$ donde:

$$traza(\mathbf{V}) = \sum_{i=1}^{p} Var(\mathbf{x_i}) = \sum_{i=1}^{p} \lambda_\mathbf{i} = \sum_{i=1}^{p} Var(\mathbf{Y_i}), \tag{2.12}$$

donde se dice que el componente inicial $i-$ésimo explica el:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} 100\,\% \tag{2.13}$$

del total de la variabilidad contenida en los datos.

Por otro lado, si $\lambda_i = 0$ para alguna i, entonces el conjunto de variables originales es linealmente independiente, *i.e.* el conjunto de observaciones están contenidas en un subespacio de $\mathbb{R}^p$.

En la práctica se calculan $\lambda_\mathbf{1}, \lambda_\mathbf{2}, \ldots, \lambda_\mathbf{p}$ si $\exists\ \lambda_i \approx 0$ pues ocurre que si

$$\begin{gathered} \frac{\sum_{i=1}^{k} \lambda_\mathbf{i}}{\sum_{i=1}^{p} \lambda_\mathbf{i}} 100\,\% \approx 100 \\ k < p \end{gathered} \tag{2.14}$$

se pueden descartar $p-k$ dimensiones *i.e.* trabajar con $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_k}$ componentes principales.

2.4.3. Comentarios sobre el modelo de componentes principales

1. **Estandarización de variables originales.** Respecto a este procedimiento pueden presentarse dos situaciones: la primera sucede cuando alguna de las

variables tiene una varianza mucho mayor a las demás, entonces el primer componente principal coincidirá con mucha proximidad a esta variable. Por otro lado, aquellas variables que presentan valores que varían mucho en magnitud, por ejemplo cuando se tratan variables de tiempo, edad, entre otros, donde las escalas de medida de las variables son muy distintas la maximización de la ecuación 2.3 dependerá decisivamente de las escalas de medida y las variables con valores más grandes tendrán más peso en el análisis. Si se quiere evitar este problema se deben estandarizar las variables antes de calcular los componentes principales, de tal manera que se logre que los valores numéricos de las variables originales sean similares, sus varianzas iguales a uno y sus covarianzas serán los coeficientes de correlación con lo que ahora la solución depende de las correlaciones y no de las varianzas.

$$\frac{(x_i - \hat{x})}{\sqrt{\sum_{i=1}^{p} \frac{(x_i - \hat{x})}{\mu}}} \tag{2.15}$$

2. **Esferado de componentes principales.** De manera informal, con esferado se refiere a "estirar" o "reducir" la variabilidad de las covariables de tal manera que todas tengan la misma varianza y gráficamente la varianza de cada variable forme una esfera, tal que:

$$Y_{i-esferada} = \frac{Y_i}{\sqrt{\lambda_i}} \tag{2.16}$$

donde:

$$V(Y_{i-esferada}) = 1. \tag{2.17}$$

3. ***"Rectification"*** se refiere a la interpretación de los coeficientes asociados a la combinación lineal. Ocasionalmente los primeros k componentes, con $k < p$ pueden ser identificados con las características contenidas en las observaciones originales, esto sirve sólo de guía para entender el conjunto de datos. Jolliffe (2002) observó lo siguiente: "Cuando todas o casi todas las correlaciones entre variables originales son positivas entonces el primer componente principal tiene coeficientes de igual signo y casi de igual magnitud y puede interpretarse como un promedio ponderado de todas las variables mientras que, el segundo componente principal tiene coeficientes de signo + y − de tal suerte que el primer componente refleja el tamaño de las observaciones mientras que el segundo refleja la *forma* de las observaciones".

4. **Reducción de dimensionalidad.** Suponga que se tiene $\mathbf{X_1}, \ldots, \mathbf{X_p}$ variables originales con las cuales se calcula $\mathbf{Y_1}, \ldots, \mathbf{Y_p}$ componentes principales y sólo se trabaja con $\mathbf{Y_1}, \ldots, \mathbf{Y_k}$ tal que, $k < p$ acumulan la mayor variabilidad de la muestra para realizar análisis estadísticos subsecuentes, por ejemplo, alguna

regresión, regresión logística, discriminante lineal, discriminante cuadrático, redes neuronales, etc.

5. Es importante tener en cuenta que las covarianzas (o correlaciones) miden únicamente las relaciones lineales entre las variables. Cuando entre ellas existan relaciones fuertes no lineales el análisis de componentes principales puede dar una información muy parcial de las variables.

2.4.4. Sección práctica

A continuación, se comienza esta sección con la matriz de correlaciones de las variables continuas recodificadas ya que de existir poca correlación entre los datos no sería viable este método. Al final de la tabla se muestra una fila de datos que representa la desviación estándar de cada variable (Std Dev), es decir, la distancia media que tienen los datos con respecto a su media aritmética:

```
> round(cor(datos_recofid),digits=3)

          laufzeit hoehe  alter
laufzeit    1.000 0.625 -0.038
hoehe       0.625 1.000  0.032
alter      -0.038 0.032  1.000

> round(sd(datos_recofid),digits=3)

laufzeit    hoehe    alter
  12.059 2822.752   11.353
```

En general, el factor que presentan la mayor correlación tienen sentido (*hoehe, laufzeit*, 0.625), monto del crédito *vs* su duración en meses.

Por otro lado, al observar la desviación estándar de las tres variables continuas se observa que la variable que describe el monto del crédito (*hoehe*) presenta una desviación mucho más elevada que las otras dos variables continuas y, considerando que como solución a este tipo de problemas, Jolliffe (2002, p.42) hace referencia a Naik y Khattree (1996) en su estudio para transformar los tiempos récord de las Olimpiadas en velocidades, ya que ocuparon la matriz de covarianzas; en la sección práctica de este tema se siguió el mismo camino.

Número de variables usadas en el análisis de componentes principales

De acuerdo a las características de la base de datos de trabajo, la *base DM*, sólo se cuenta con 10 variables categóricas, donde sólo 3 se puede visualizar en su extracción original como variables continuas y de las cuales se ha analizado su correlación

en la sección anterior. Dado lo anterior, esta sección práctica se enfoca a la idea de utilizar el análisis de componentes principales para describir la base de datos, tomando de referencia los comentarios de Jollifer (2002, p.339), quien menciona que cuando el análisis de componentes principales es usado como técnica descriptiva, no hay razón por la que las variables en el análisis deban de ser de algún tipo en especial, siendo un caso extremo o particular cuando se presentan una mezcla de variables continuas, categóricas o incluso binarias (0/1).

Jollifer (2002, p.339) también menciona que es cierto que las varianzas, covarianzas y correlaciones tienen especial relevancia para una variable $p-$dimensional $\mathbf{x}$, y que las funciones lineales de variables binarias son menos fáciles de interpretar que las funciones de variables continuas, sin embargo, tomando como objetivo principal del análisis de componentes principales el sumarizar la mayoría de la "variación" que se presenta en el conjunto original de p variables, usando un menor número de variables derivadas, se puede lograr el objetivo antes descrito independientemente de la naturaleza de las variables originales.

Por otro lado, Jollifer (2002, p.339) cita a Gower (1966), quien señala que el uso de análisis de componentes principales en datos binarios, no proporciona una representación de pocas dimensiones debido a que en estas condiciones, este análisis es equivalente a un análisis de coordenadas principales (del inglés, *principal coordinate analysis*), que se basa en la similitud entre dos individuos. De lo anterior, Jollifer también menciona a Cox (1972), quien sugiere su idea a la cual llama *permutational principal componentes*, que se basa en el hecho de que un conjunto de datos, elaborado por p variables binarias, puede ser expresado en un número de diferentes pero equivalentes caminos, de tal forma que sugiere que una alternativa para el análisis de componentes principales es el transformar a variables binarias independientes usando tales permutaciones.

Dicho lo anterior, a continuación se toman las 10 variables categóricas recategorizadas previamente en la sección anterior; el total de categorías de estas 10 covariables son 44 categorías y, al dicotomizarlas tomando una categoría de referencia por variable, se obtienen 34 variables binarias. Estas 34 variables son las que se utilizarán para aplicar y desarrollar el análisis de componentes principales.

Resultados de generar el análisis de componentes principales a la *base DM*

Habiendo definido ya el número de variables utilizadas para aplicar el modelo de componentes principales, a continuación se utiliza el *software R* con el comando *princomp* para obtener la siguiente información sumarizada:

```
> PCA3=princomp(datos.pca,cor=F)
> summary(PCA3)
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     0.6930861 0.6244289 0.5695144 0.55852452 0.52865536 0.51446370
Proportion of Variance 0.1017639 0.0826010 0.0687114 0.06608515 0.05920585 0.05606978
Cumulative Proportion  0.1017639 0.1843649 0.2530763 0.31916145 0.37836730 0.43443708

                           Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12
Standard deviation     0.47802474 0.47362796 0.46579701 0.44826316 0.43565096 0.42710804
Proportion of Variance 0.04840833 0.04752193 0.04596346 0.04256822 0.04020654 0.03864514
Cumulative Proportion  0.48284541 0.53036734 0.57633080 0.61889902 0.65910556 0.69775070

                          Comp.13    Comp.14    Comp.15    Comp.16    Comp.17    Comp.18
Standard deviation     0.40326372 0.38065868 0.34824637 0.34269518 0.32870251 0.31304945
Proportion of Variance 0.03445067 0.03069664 0.02569168 0.02487914 0.02288893 0.02076086
Cumulative Proportion  0.73220137 0.76289801 0.78858969 0.81346883 0.83635775 0.85711861

                          Comp.19    Comp.20    Comp.21    Comp.22    Comp.23    Comp.24
Standard deviation     0.3020761 0.26549527 0.25768820 0.24447093 0.23401910 0.22721066
Proportion of Variance 0.0193309 0.01493251 0.01406723 0.01266117 0.01160171 0.01093646
Cumulative Proportion  0.8764495 0.89138203 0.90544925 0.91811042 0.92971213 0.94064859

                           Comp.25     Comp.26     Comp.27     Comp.28     Comp.29
Standard deviation     0.204423057 0.193907188 0.188604988 0.183763294 0.175555265
Proportion of Variance 0.008852769 0.007965392 0.007535737 0.007153802 0.006529006
Cumulative Proportion  0.949501357 0.957466750 0.965002486 0.972156288 0.978685294

                           Comp.30     Comp.31     Comp.32     Comp.33     Comp.34
Standard deviation     0.170832605 0.155201738 0.148240482 0.141375687 0.073353128
Proportion of Variance 0.006182454 0.005102847 0.004655357 0.004234175 0.001139873
Cumulative Proportion  0.984867748 0.989970594 0.994625951 0.998860127 1.000000000
```

Se observa que con los primeros 12 componentes se obtiene un 69.8 % de información del conjunto de datos y con 16 componentes se obtiene el 81.3 % de la variación total de la base, es decir, de 34 variables resultantes de dicotimizar las originales se puede trabajar sólo con 16, o inclusive 12. Otra forma de verificar esto es revisar los *eigenvalores* o valores propios resultantes, es decir, la desviación estándar de cada componente.

Al observar los *eigenvectores*[8] en forma matricial con respecto a las características de la población (matriz factorial), donde cada entrada es llamada *factor*, se obtiene la correlación actual entre cada variable y los componentes resultantes. Cabe mencionar que para que cada factor sea fácilmente interpretable se espera que cumplan las siguientes características que no son fáciles de conseguir:

- Cada factor sea próximo a uno

[8]Vectores que no se vieron afectados por la transformación lineal aplicada.

- Una variable debe de tener coeficientes elevados sólo con un factor.
- No deben de existir factores con coeficientes similares.

Se puede apreciar en la matriz de factores presente en el Anexo E que la mayoría de las características están bien representada por un factor mayor a 0.5. En la componente número 20 se encuentra el factor de mayor tamaño (0.833) correspondiente a la categoría 4 de la variable *Género* que indica el género femenino y estado civil soltera (*famges*).

2.4.5. Conclusiones del análisis de componentes principales

Dado lo anterior se describen las primeras dos componentes para conocer que características son de mayor peso al clasificar un elemento:

- La primer componente está determinada principalmente por la variable *Morosidad* (*moral*) que describe el comportamiento del cliente al saldar créditos previos, lo cual hace sentido pues la institución requiere conocer su capacidad de pago en otras ocasiones. De no conocer esta información es muy probable que el banco sea conservador al calificar a un nuevo solicitante. Esta componente también tiene peso sobre la variable *Género* relacionada al estado civil y género del cliente (*famges*), donde presenta mayor significancia la categoría de hombre casado o viudo y completando el perfil demográfico del cliente se observa significancia en la variable *Edad* que expresa la edad en años del cliente (*dalter*), dando mayor importancia a aplicantes con edades en el rango de 26 a 39 años; finalmente este componente presenta significancia en la variable *Duración* que representa la duración en meses del crédito (*dlafzeit*), en particular en la categoría 2: de 6 a 12 meses, lo que refuerza el análisis exploratorio de datos donde se nota que en su mayoría se otorgaron créditos de plazos cortos.
- La segunda componente está determinada principalmente por la variable *Capacidad*, que expresa el porcentaje de ingreso que el cliente tiene disponible para el pago del crédito (*rate*). Es importante notar que todas las categorías de esta variable son significativas, pero en especial la categoría 1 equivalente a "$< 20\,\%$", lo cual puede hacer pensar que se buscan clientes propensos a fallar con sus pagos para generar utilidad mediante los intereses. Sin embargo, hay que tener en cuenta que en la base el $47.6\,\%$ de los clientes caen en esta categoría. El resto de factores de mayor tamaño en esta componente están relacionados al crédito: duración en meses (*dlaufzeit*) predominando de nuevo periodos cortos, propósito del crédito (*verw*) con mayo peso en carro nuevo, carro usado, muebles y, finalmente, monto del crédito (*dhoehe*) con montos entre 500 y 5,000 marcos alemanes.

Las últimas dos componentes sirven para conocer la dirección o tendencia de las decisiones y se observa que se concentran en las variables de duración y monto del crédito solicitado (*dlaufzeit* y *dhoehe*).

Finalmente, en la figura 2.6 se pueden apreciar las gráficas de las primeras cuatro componentes, así como de las últimas cuatro, todas elaboradas con ayuda del *software R*, donde el color rojo equivale a clientes sin capacidad crediticia y el color azul a clientes con capacidad crediticia. En las gráficas superiores, del lado izquierdo se observa la gráfica de las primeras dos componentes, donde se presentan muy marcados cuatro grupos en las componentes pero, complicado para discriminar a simple vista. Sin embargo, la concentración de valores se encuentra entre 1 y -1 de ambos ejes, dejando en 0 una franja vertical casi vacía y otra horizontal no tanto, aunque en estas franjas predominan observaciones sin capacidad crediticia. Por otro lado, de lado derecho se presenta la gráfica de las componentes 3 y 4, donde se observa mayor dispersión en los datos. No obstante se presenta mayor concetración de datos en la esquina superior izquierda. En las gráficas inferiores se observa del lado izquierdo la gráfica con las últimas dos componentes donde se presentan dos conglomeraciones muy marcadas, teniendo mayor peso las observaciones sobre la coordenada $(0, 0)$ y en una de las esquinas de este gráfico un conjunto aislado de observaciones que posiblemente tengan que ver con el peso con signo negativo de la covariable *Monto*, que expresa el monto del crédito (*dhoehe*). Ahora, se observa del lado derecho la gráfica de las componentes 31 y 32 donde se ve la tendencia de las observaciones a estar alrededor de la misma coordenada.

A su vez, se ha obtenido mediante la librería *rggobi* del *software R* la exploración en tres dimensiones de las tres primeras y últimas componentes, figura 2.7. En el conjunto de gráficas se aprecia de color verde a clientes sin capacidad crediticia y el color naranja a clientes con capacidad crediticia. De lado izquierdo, se encuentra la gráfica de las primeras tres componentes donde con ayuda de otra dimensión se aprecia que las primeras componentes se conglomeran en cuatro conjuntos marcados principalmente por las primeras dos componentes (con ayuda del círculo de la esquina inferior izquierda se puede ubicar la posición de cada eje, *i.e.* de cada componente). Del lado derecho, se presenta la gráfica de las componentes 32, 33 y 34, donde se aprecia que las observaciones no se condensan solamente alrededor del origen, sino que la componente 32 marca tres subconjuntos y la componente 33 marcó un poco más cierto grupo de datos, en su mayoría clientes con capacidad crediticia.

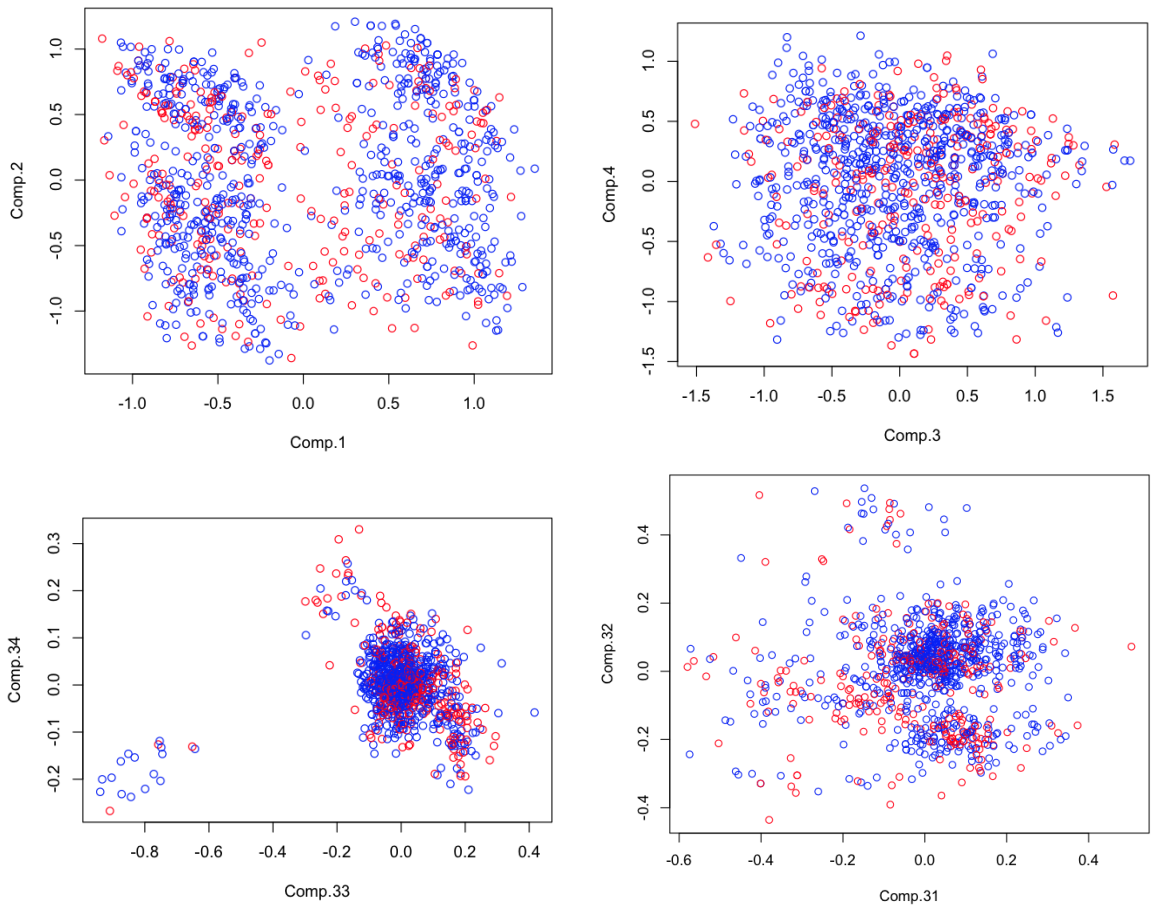

Figura 2.6: *Gráficas superiores:* del lado izquierdo, gráfica de las primeras dos componentes y de lado derecho, gráfica de las componentes 3 y 4.*Gráficas inferiores:* del lado izquierdo, gráfica con las últimas dos componentes. Del lado derecho, gráfica de las componentes 31 y 32. Se aprecia que el color rojo equivale a clientes sin capacidad crediticia y el color azul a clientes con capacidad crediticia.

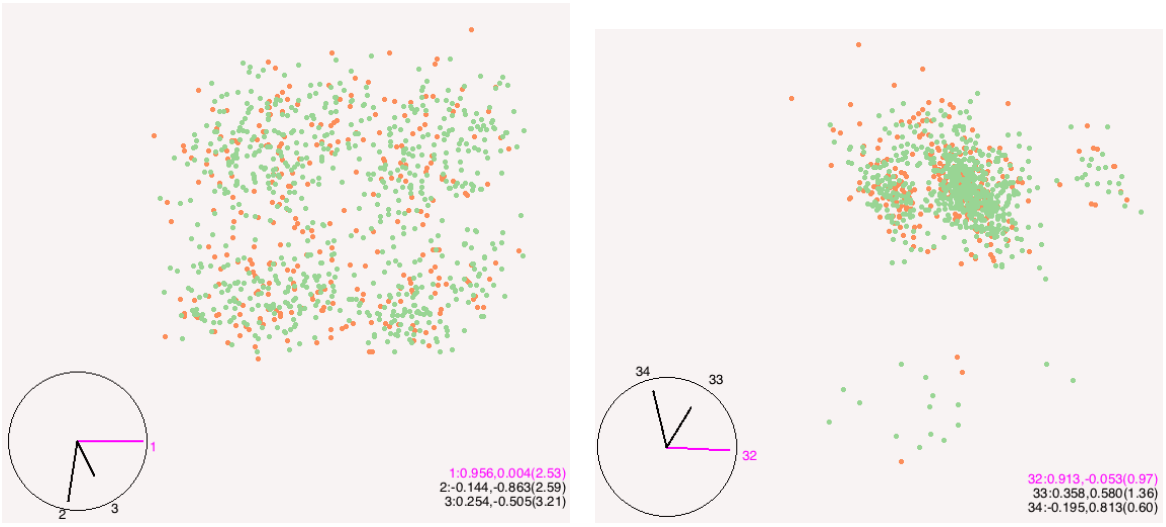

Figura 2.7: Del lado izquierdo: gráfica de las primeras tres componentes. Del lado derecho: gráfica de las componentes 32, 33 y 34, donde se aprecía de color verde el equivale a clientes sin capacidad crediticia y de color naranja a clientes con capacidad crediticia

2.5. Conclusiones del capítulo

En este capítulo se determinó el perfil de un cliente con capacidad crediticia, según las políticas del la institución financiera, de tal manera que mediante la obtención de estadísticas descriptivas, la exploración de *projection pursuit* (para detalle consultar el Anexo D) y el análisis de los componentes principales de esta sección se definió como clientes de preferencia de género masculino, edades referentes a un adulto joven, con estado civil soltero o viudo, estable en su trabajo y hogar, por tener permanencia en estos mayor a dos años, sin deudas actualmente, con una capacidad de pago ajustada y preferentemente con referencias crediticias dentro de la misma institución. En cuanto a los créditos otorgados en su mayoría son de montos pequeños, por periodos de corto a mediano plazo y en su mayoría con fines para compra de muebles.

Capítulo 3

Modelos de clasificación

3.1. Introducción

A continuación se plantean los modelos de regresión logística y redes neuronales, así como algunas pruebas para valuación de los modelos obtenidos con el objetivo de aplicar dichos modelos para la clasificación de clientes con capacidad crediticia de aquellos que no tienen capacidad crediticia y provenien de la *base DM*.

Una vez desarrollados estos modelos con las observaciones de la *base DM*, se seleccionarán aquellos que presenten las mejores tasas de clasificación de clientes para orientar la toma de decisiones al estimar nuevas clasificaciones.

Las tasas de clasificación se definen como sigue: dada una población con una variable de respuesta binaria se busca aprender del comportamiento y perfil de cada categoría[1] para que cuando ingrese una nueva observación se pueda clasificar al grupo al cual las características que lo definen son más afines. De esta manera, al aplicar un modelo ajustado a las observaciones se podrá medir qué porcentaje de clientes fue clasificado correctamente, es decir, clientes con capacidad crediticia clasificados como tal, o bien clientes sin capacidad crediticia clasificados de esta manera, llamadas *observaciones bien clasificadas* y, por otro lado, clientes con o sin capacidad crediticia clasificados en su inversa, también llamadas *observaciones mal clasificadas*. De esta manera en la siguiente tabla se aprecia la lectura de estas definiciones:

$clasificación$	$bien$	mal
$bien$	a	b
mal	c	d

(3.1)

[1] La definición de perfiles de cada grupo de la población resulta más sencilla al identificar aquellas variables que definen a cada grupo de la población mediante los modelos y métodos mencionados en el capítulo anterior.

En estricta teoría se busca que las tasas de observaciones bien clasificadas sean lo más próximas a uno. Por otro lado, otra razón para revisar la potencia de clasificación del modelo es a través de *tasa global de correcta clasificación* tal que $tgdc = a/(a+b+c+d)$.

3.2. Regresión Logística

Retomando el problema inicial de clasificar bien o mal a solicitantes de crédito discriminando entre dos posibles poblaciones previamente definidas como "con capacidad crediticia" y "sin capacidad crediticia", que no necesariamente tiene un orden establecido[2], se plantea el uso de un modelo de regresión logística con base en una *variable de clasificación*, Y, en este caso la variable *kredit*, tal que $\mathbf{y_i} \in \{1,0\}$, donde $Y = 1$ indica que es un cliente con capacidad crediticia y $Y = 0$ indica cuando el cliente no tiene capacidad crediticia.

La muestra consiste de n elementos del tipo (y_i,x_i) donde x_i es el valor de la variable independiente para la $i-$ésima observación, y y_i denota el valor de la variable dicotómica dependiente, en un caso general se tiene lo siguiente:

$$\{(y_1, x_{11}, x_{12}, \ldots, x_{1p}), (y_2, x_{21}, x_{22}, \ldots, x_{2p}), \ldots, (y_n, x_{n1}, x_{n2}, \ldots, x_{np})\}. \tag{3.2}$$

Se define a $\pi(x_i)$ como la probabilidad de éxito de la variable $\mathbf{x}$ tal que:

$$\begin{aligned} P(Y = 1|\mathbf{x}) &= \pi(\mathbf{x}) \\ &= \frac{exp(\beta \mathbf{x}')}{1+\beta \mathbf{x}'} \\ &= \frac{exp(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}{1+exp(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}, \end{aligned} \tag{3.3}$$

donde $\beta = (\beta_0, \ldots, \beta_p)$ es el vector de parámetros desconocidos del modelo. Sin embargo, se observa que la ecuación de los parámetros no es lineal, por lo que se utiliza una transformación logito, o por sus siglas en inglés *logit*, para obtener la función lineal de un caso generalizado:

$$logit[\pi(\mathbf{x})] = log\Big(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\Big) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{3.4}$$

[2]Aunque sí puede suceder que para el estudio una población sea de mayor interés que la otra.

Donde *logit*, representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a una población. Al ser una función lineal de las variables explicativas facilita la estimación e interpretación del modelo.

Por otro lado, la expresión:

$$\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})} = exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p), \tag{3.5}$$

proveniente de la ecuación 3.4, es conocido como momio, o bien, por sus siglas en inglés *odd*, y expresa una proporción de riesgo con base en la probabilidad de que un solicitante sea clasificado como un cliente con capacidad crediticia.

Sea $\Omega = \frac{\pi(x)}{1-\pi(x)}$ tal que $\Omega \in \mathbb{R}^+$, y $\exists\ \Omega_1$ y $\Omega_2 \in \mathbb{R}^+$ se define:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{(\pi_1)/(1-\pi_1)}{(\pi_2)/(1-\pi_2)} \tag{3.6}$$

como el cociente llamado razón de momios, o bien *odds ratio*, que cae en los $\mathbb{R}^+$ y donde si $\Omega = 1$ expresa independencia entre la variable explicativa y la dependiente, y si $\Omega > 1$ expresa que es más probable el éxito que el fracaso, o bien el ser un cliente con capacidad crediticia que no serlo. Para mayor desarrollo consultar Hosmer and Lemeshow (2000, p.47).

3.2.1. Estimación del modelo *logit*

La estimación del modelo nos lleva a estimar los parámetros, desarrollo que se lleva a cabo por el método de máxima verosimilitud, la cual expresa la probabilidad del dato observado como función de los parámetros desconocidos tal que maximice los valores de ésta.

Ahora bien, asuma que $\mathbf{Y} \sim$ *Bernoulli* tal que y_i representa la respuesta de la $i-$ésima observación, a continuación considere una muestra aleatoria de datos $(\mathbf{x_i}, \mathbf{y_i})$ con $i = (1, \ldots, p)$ y denote a $\beta = (\beta_\mathbf{0}, \beta_\mathbf{1})$ tal que:

$$P[\mathbb{Y} = y_i] = \pi_i^{y_i}(1-\pi_i)^{(1-y_i)}, \tag{3.7}$$

donde $P[\mathbb{Y} = 0] = (1-\pi_i)$ y $P[\mathbb{Y} = 1] = \pi_i$. Al asumir independencia entre las observaciones se tiene:

$$\ell(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} [1 - \pi_i]^{1-y_i}, \tag{3.8}$$

y mediante la estimación por máxima verosimilitud se obtiene:

$$L(\beta) = ln(\ell(\beta)) = \sum_{i=1}^{n} y_i ln(\pi) + \sum_{i=1}^{n} (1 - y_i) ln(1 - \pi)) \tag{3.9}$$

$$\begin{aligned} &= \sum_{i=1}^{n} y_i ln(\pi) + \sum_{i=1}^{n} ln(1 - \pi) + \sum_{i=1}^{n} -y_i ln(1 - \pi) \\ &= \sum_{i=1}^{n} y_i ln\Big(\tfrac{\pi}{(1-\pi)}\Big) + \sum_{i=1}^{n} ln(1 - \pi). \end{aligned} \tag{3.10}$$

Por ser β un vector se calculan las derivadas parciales con respecto a β_0, β_1, ..., β_p entonces:

$$\hat{\beta} = \begin{cases} \sum_{i=1}^{n} \Big[y_i - \pi_i\Big] = 0 \\ \sum_{i=1}^{n} x_1 \Big[y_i - \pi_i\Big] = 0 \\ \vdots \end{cases}$$

donde a $\hat{\beta}$ se le conoce como la estimación del parámetro β y de la primer expresión se obtiene que:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}_i. \tag{3.11}$$

Sin embargo, las ecuaciones de los valores de $\hat{\beta}$ son no lineales por lo que β_0, β_1, ..., β_p requiere de *software* especializado para ser calculados.

3.2.2. Evaluación de la eficacia del modelo

En esta sección se busca contestar las siguientes preguntas:

1. *¿El modelo que incluye la variable en cuestión dice más acerca de la variable de respuesta que un modelo que no incluya dicha variable?*
2. *¿Las variables en dicho modelo son significativas para la clasificación de los datos?*

Deviance

Al desarrollo de la primer pregunta se le conoce como *goodness-of-fit*. Para el caso de los modelos de regresión lineal es equivalente a calcular la suma de los cuadrados de los residuos del modelo, de tal forma que para el caso de regresión logística la comparación entre los valores observados y los predichos se obtiene mediante la devianza o *deviance* que se basa en el logaritmo de la función de verosimilitud de logaritmos que procede de la expresión 3.9 y se define por:

$$Deviance = D = -2ln\ell(\beta), \tag{3.12}$$

es decir,

$$D = -2\sum_{i=1}^{p}(y_i log\pi_i + (1-y_i)log(i-\pi_i)), \tag{3.13}$$

donde se considera que $\beta^{`} = (\beta_0, \beta_1, \ldots, \beta_p)$ es un vector de $p+1$ variables. Dado lo anterior, se define la desviación para cada observación como:

$$d_i = -2(y_i log\pi_i + (1-y_i)log(1-\pi_i)). \tag{3.14}$$

Prueba del Cociente de Verosimilitudes

Esta prueba también busca responder si el modelo que incluye la variable en cuestión dice más acerca de la variable de respuesta que un modelo que no incluya dicha variable, y se basa en probar las siguientes hipótesis:

$$H_0 : \beta_1 = \ldots = \beta_m = 0 \quad vs \quad H_a : \beta_i \neq 0$$

para algún $i = 1, \ldots, m$ donde m representa el número de variables incluidas en el modelo ajustado. Esta prueba se basa en el concepto de *deviance* antes mencionado tal que se compara D con y sin la variable independiente a probar:

G = D (Modelo sin la variable) - D (Modelo con la variable).

Donde $G \sim \chi^2_{p-k}$, tal que p es el número de variables explicativas del modelo inicial y k el número de variables del submodelo ajustado (Hosmer and Lemeshow, 2000,

p.14).

De esta forma, el $p-valor$ asociado a esta prueba indica que cuando $P[\chi^2(p-k) > G] < \alpha$ se rechaza la hipótesis nula, es decir, no hay evidencia para suponer que $\beta_1 = \ldots = \beta_m = 0$, por lo que la variable en cuestión es significativa en el modelo.

Prueba de χ^2 *de Pearson*

Otra prueba importante para evaluar la eficacia de un modelo es la **Prueba de χ^2 de Pearson** donde nuevamente se siguen los caminos posibles para obtener la diferencia entre los valores observados y los valores ajustados $(y - \hat{y})$, de tal forma que para el caso de la regresión logística se busca enfatizar la importancia de la probabilidad estimada de cada variable, tal que suponiendo que el modelo ajustado tiene p variables explicativas y denotando por J al número de categorías que puede contener cada covariable se nombra como *configuraciones* a cada pareja de categorías de éstas, de tal forma que se denota por m_j al número de observaciones que toma cada configuración específica con $j = 1, \ldots, J$ donde $\sum_{j=1}^{J} m_j = n$. Nótese que algunas observaciones tendrán los mismos valores, entonces $J < n$ y además se pueden identificar entre los m_i observados con respecto al valor de y_i señalando sólo éxitos, n_1, o fracasos n_2 para hacer más fácil la medida de ajuste del modelo.

Dicho lo anterior, se denota para cada $j-ésima$ configuración su valor ajustado, $\hat{y}_j$ como:

$$\hat{y}_j = m_j\hat{\pi}_j = \frac{\exp^{\hat{\pi}(x_j)}}{1 + \exp^{\hat{\pi}(x_j)}}, \tag{3.15}$$

donde $\hat{\pi}(x_i)$ es el valor estimado de la función *logit*. Y se define al *Residuo de Pearson* de la $j-ésima$ configuración como:

$$r(y_j, \hat{\pi_j}) = \frac{(y_j - m_j\hat{\pi_j})}{\sqrt{m_j\hat{\pi_j}(1 - \hat{\pi_j})}}, \tag{3.16}$$

con lo anterior, la estadística de Prueba de χ^2 de Pearson basada en este residual resulta

$$\chi^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi_j})^2 \tag{3.17}$$

donde la distribución de la estadística χ^2 bajo el supuesto de que ajustó bien a los datos supone ser χ^2 con $J - (p + 1)$ grados de libertad donde J es el número de

categorías que puede contener cada covariable.

Finalmente, si el $p-valor$ asociado a esta prueba es menor a α se rechaza la prueba de hipótesis $H_0 : \beta_1 = \ldots = \beta_m = 0$ vs $H_a : \beta_i \neq 0$, es decir, no hay evidencia para suponer que $\beta_1 = \ldots = \beta_m = 0$.

Prueba de Hosmer - Lemeshow

En esta prueba, Hosmer and Lemeshow (2000, p. 147) proponen agrupar en percentiles, usualmente deciles, a las observaciones con base en las probabilidades estimadas de dos maneras distintas, suponga $g = 10$, donde g expresa el número de grupos con aproximadamente el mismo número de observaciones en cada uno, las observaciones deben tener la misma configuración específica aproximadamente igual al número total del observaciones:

- En la primera opción se ordenan las observaciones de menor a mayor con base en sus probabilidades estimadas, de tal forma que el primer grupo contiene las menores estimaciones y el último las mayores.
- En el segundo camino se consideran puntos de corte definidos, $k/10$ donde $k = 1, 2, \ldots, 9$, de tal forma que el primer grupo contiene todas las estimaciones que poseen una probabilidad asignada ≤ 0.1.

Así, para cada grupo se obtiene el estadístico de *Hosmer - Lemeshow* definido como sigue:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k^{'}\bar{\pi}_k)^2}{(n_k^{'}\bar{\pi}_k)(1 - \bar{\pi}_k)}, \tag{3.18}$$

donde $n_k^{'}$ representa el número total de observaciones en el grupo k y c_k el número de configuraciones específicas en el k-*ésimo* decil.

$$o_k = \sum_{j=1}^{c_k} y_i \tag{3.19}$$

representa el número de respuestas entre c_k y

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k^{'}} \tag{3.20}$$

es la probabilidad promedio estimada en el *k-ésimo* grupo.

Con base en extensas simulaciones, Hosmer and Lemeshow (2000, p. 148) demostraron que cuando $J = n$ y el modelo de regresión ajustado es el correcto, entonces $\hat{C} \sim \chi^2_{g-2}$. Concluyendo, se dice que a valores grandes de la estadística $\hat{C}$ (valores pequeños del *p-valor*) indica una falta de ajuste en el modelo.

De esta manera, la prueba de hipótesis $H_0 : \beta_1 = \ldots = \beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ se rechaza cuando el $p - valor$ asociado es menor a α, es decir, no hay evidencia para suponer que $\beta_1 = \ldots = \beta_m = 0$.

Prueba de Wald

Existen otro tipo de pruebas estadísticas como son el estadístico de Wald, o bien *Wald test*, donde se busca conocer si un parámetro es significativo o no, para lo cual se contrasta la hipótesis:

$$H_0 : \beta_j = 0 \quad vs \quad H_a : \beta_j \neq 0.$$

Siendo la estadística de prueba:

$$\omega_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}. \tag{3.21}$$

Este estadístico se basa en obtener la comparación de la probabilidad estimada del parámetro de la pendiente, β_j, con respecto a la estimación de su error estándar $\widehat{SE}(\hat{\beta}_j)$ tal que bajo la hipótesis de que $\beta_j = 0$ sea cierta se distribuye normal. De tal forma que se rechaza H_0 cuando $\omega_j < \alpha$, es decir, no hay evidencia para suponer que $\beta_1 = \ldots = \beta_j = 0$, lo que indica que la variable es significativa para el modelo.

Otros criterios

Además de las pruebas de significancia, un par de modelos y tal vez de los más conocido son el Criterio de Información de Akaike, o por sus siglas en inglés *Akaike Information Criterion* (AIC), así como el Criterio de Schwarz (SC). Estos dos criterios sirven también para comparar modelos, mientras menor sea el valor nos indica un modelo más deseable. Según Agresti (2002, p.141), un modelo óptimo es aquel que contiene valores más cercanos a los verdaderos valores de las probabilidades. A su vez, define al criterio de información de Akaike como:

$$AIC = -2(ln\ell(\beta) - n), \tag{3.22}$$

donde n es el número de parámetros del modelo, de tal manera que este criterio se ve afectado por el número de parámetros del modelo, mientras que el criterio de Schwarz ajusta, además de para el número de parámetros, para el número de observaciones.

3.2.3. Procedimientos para la selección de variables

El proceso de selección del modelo varía considerablemente conforme el número de variables explicativas aumenta debido a que esto se refleja en el incremento de las iteraciones del modelo, sin embargo siempre se buscan dos principales objetivos:

- Que el modelo sea lo suficientemente complejo como para ajustar adecuadamente al conjunto de datos.
- Y que a su vez sea sencillo al momento de interpretarlo.

Con base en esto, comúnmente se ocupan procedimientos automatizados fáciles de encontrar en diversos *softwares* estadísticos: *Forward*, *Backward* o bien *Stepwise*.

La selección mediante el procedimiento *Forward* consiste en agregar términos secuencialmente hasta que una nueva adición no mejore el ajuste. La eliminación mediante el procedimiento *Backward* inicia con un modelo complejo y elimina secuencialmente términos de tal forma que afecte lo menos posible el modelo ajustado (*i.e.* se remueve el mayor *p-valor*), este método se detiene cuando una eliminación conduce a un peor ajuste. Finalmente, y tal vez el más utilizado, se tiene el procedimiento *Stepwise*, que es una mezcla de los dos anteriores, sin embargo toma como base el procedimiento *Forward* añadiendo elementos y, en su caso, eliminándolos cuando el ajuste empeora para dar oportunidad de agregar otra variable.

Varios estadísticos prefieren el procedimiento *Backward* sobre el *Forward* pues sienten que es mejor eliminar términos de un modelo complejo que comenzar añadiendo términos con base en un modelo muy simple, sin embargo los tres procedimientos se deben de manejar con cautela.

3.2.4. Estrategias para la selección del modelo

Lograr seleccionar el modelo que ajuste mejor a los datos, clasifique mejor a las estimaciones de los mismo y además sea fácil de interpretar, depende de los criterios,

pruebas y procedimientos utilizados.

Agresti (2002, p. 216) cita lo siguiente: "Cualquier modelo es una simplificación de la realidad. Sin embargo, un modelo simple que ajusta adecuadamente es ventajoso sobre un modelo complejo, si un modelo tiene relativamente pocos puntos en contra, describiendo bien la realidad, éste proveerá de buenas estimaciones de las probabilidades esperadas y de los *odds ratios* que describen los efectos de los predictores".

Dado lo anterior, se debe tomar en cuenta al ajustar un modelo que se tiene que comparar con el modelo saturado para poder definir si el modelo ajustado resulta mejor. Por otro lado, al ajustar un modelo, o bien, simplificar un modelo, es recomendable realizar la toma de decisiones sobre la significancia de las variables al comparar los *p-valores*, basarse en la prueba del cociente de verosimilitudes (sobre todo al comparar modelos anidados, así como el criterio de AIC), y también la Prueba de *Wald*.

Además de esto, se debe comparar la bondad de ajuste del modelo mediante la Prueba de la χ^2 de *Pearson* y por medio de la Prueba de *Hosmer - Lemeshow* conocer si las variables ajustan bien al modelo.

En la selección de modelos es recomendable utilizar los distintos procedimientos de selección de variables *Forward*, *Backward* o bien *Stepwise*, sin olvidar que procedimientos como *Stepwise* resultan complementarios ya que pueden existir otros submodelos que estos procedimientos dejen a un lado. Sin embargo, la disponibilidad de éstos en los *softwares* estadísticos nos dan oportunidad de obtener más opciones y ayudan a la exploración de los datos.

Finalmente, no se debe perder de vista que otro factor para la toma de decisiones en la selección de modelos se basa en obtener bajas tasas de clasificación errónea en la estimación de las observaciones buenas clasificadas como malas y las malas clasificadas como buenas. Aquellos modelos que sean de fácil interpretación, buen ajuste y además buenas tasas serán los que llamen más la atención del investigador.

Debe de tenerse en cuenta que a pesar de todas las pruebas y procedimientos no se debe de dejar de lado que el análisis profundo de la base de datos ayudará a la mejor toda de decisiones, ya que algunas de ellas se pueden basar en el conocimiento estadístico, o bien del negocio.

3.2.5. Sección práctica

A continuación se ilustra el modelo de regresión logística con ayuda del *software R*. Se comenzó por definir el modelo saturado que se tomará para comparativos

futuros, para esto se establecieron cuatro modelos iniciales con las 10 variables en cuestión, pero combinando el tipo de las variables con las que se cuenta, dicotómica, categórica y continua, es decir:

- Modelo 1: En el primer modelo se introdujeron 33 variables binarias.
- Modelo 2: En el segundo modelo se conservan las tres variables continuas (*alter:* edad en años del cliente, *laufzeit:* duración en meses del crédito, *hoehe:* monto del crédito) y el resto sigue considerándose binario (3 continuas y 20 binarias, en total 23 variables).
- Modelo 3: En el tercer modelo se consideran las variables tal y como son: tres variables continuas y siete variables categóricas, se debe tener en cuenta que las variables categóricas son ordinales (3 continuas y 7 categóricas, en total 10 variables).
- Modelo 4: En el cuarto modelo, con base en lo observado en los tres modelos anteriores, se mantuvo una combinación de variables binarias (*verw:* propósito del crédito, *rate:* porcentaje de ingreso disponible para el pago, *famges:* estado marital y género, *alter:* edad en años del cliente), variables categóricas (*laufkont:* saldo de cuentas en el banco DM, *moral:* pago de créditos previos, *weitkred:* otros créditos vigentes, *beruf:* ocupación) y variables continuas (*alter:* edad en años del cliente, *laufzeit:* duración en meses del crédito, *hoehe:* monto del crédito) (11 binarias, 4 categóricas y 3 continuas, en total 21 variables).

Para tomar la decisión de selección se toma en cuenta la siguiente información:

Cuadro C.1. Selección del modelo saturado

Modelo	**Num. de variables**	**Cociente de Verosimilitudes**			**Hosmer-Lemeshow**			**AIC**
		DF	G	$Pr(\chi^2 > G)$	**DF**	χ^2	$Pr > \chi^2$	
Modelo 1	33	33	265.84	<.0001	8	7.7789	0.4554	1025.9
Modelo 2	23	23	249.89	<.0001	8	10.6196	0.2242	1019.8
Modelo 3	10	10	208.33	<.0001	8	5.5799	0.6942	1035.4
Modelo 4	21	21	246.94	<.0001	8	10.6783	0.2206	1018.8

En el Cuadro C.1 se observa el cociente de verosimilitud de los cuatro modelos saturados propuestos previamente. Para el Modelo 1 se esta comparando la prueba de hipótesis $H_0 : \beta_1 = \ldots = \beta_{34} = 0$ vs $H_a : \beta_i \neq 0$ para algún $i = 1, \ldots, 34$ donde el estadístico $G = 265.84$ y $Pr[\chi^2 > 265.84] < \alpha = 0.05$ por lo que las variables del modelo 1 saturado seleccionado son significativas y se rechaza H_0. De igual forma, se presenta para cada modelo saturado, para el modelo 2 saturado se tiene la prueba de hipótesis $H_0 : \beta_1 = \ldots = \beta_{23} = 0$ vs $H_a : \beta_i \neq 0$ para algún $i = 1, \ldots, 23$ donde el estadístico $G = 249.89$ y $Pr[\chi^2 > 249.89] < \alpha = 0.05$ por lo que se rechaza H_0. De igual forma para los otros dos modelos saturados.

Adicional a esto se observa que para los cuatro modelos saturados la hipótesis nula del cociente de verosimilitudes sobre la varianza nula y la varianza residual se rechaza, $H_0 : \beta_1 = \ldots = \beta_p = 0$ *vs* $H_a : \beta_i \neq 0$ para $i = 1, \ldots, m$ donde m representa el número de variables en el modelo, lo que implica que no todos los coeficientes son cero. Por otro lado, en la prueba de Hosmer-Lemeshow se observa que ninguno de los modelos presenta falta de ajuste. En cuanto al valor de la prueba AIC se observa que varía considerablemente de modelo a modelo siendo el modelo 4 el de menor valor, sin embargo no el más complejo.

Cuadro $C.2$. Matriz de confusión de modelos saturados

Modelo	**mal→ mal**	**mal → bien**	**bien → mal**	**bien → bien**	**tgdc**
Modelo 1	46.00 %	54.00 %	9.57 %	90.43 %	90.86 %
Modelo 2	**49.00 %**	**51.00 %**	**9.29 %**	**90.71 %**	**90.29 %**
Modelo 3	44.33 %	55.67 %	10 %	90.00 %	90.43 %
Modelo 4	45.33 %	54.67 %	9.57 %	90.43 %	89.71 %

Se marca con negritas el modelo saturado con mejores tasas de clasificación.

Un factor que se debe de tener en cuenta para la toma de decisiones es la matriz de confusión que indica las tasas de clasificación errónea y clasificación correcta de los modelos, estas matrices se aprecian en el Cuadro $C.2$ donde se observa una tasa alta de clasificación correcta (bien → bien), clientes buenos clasificados como buenos, del 90 % y una tasa de error de clientes buenos clasificados mal (bien → mal) máximo del 10 % que es buena, el problema radica en las malas observaciones clasificadas bien (mal → bien) donde las tasas están por encima del 50 %. El modelo 3, que no presentó falta de ajuste, tiene la mayor tasa de clientes mal clasificados como bien (mal → bien) con un 56 %. Este factor se puede deber a la composición de la base de datos donde sólo el 30 % de los clientes son clientes sin capacidad crediticia. Una acción que puede equilibrar esta situación, si la base contuviera mayor número de observaciones, es el remover aquellos clientes que se sabe son buenos y que clasifican bien para introducir un poco de variabilidad al modelo.

Finalmente, se decidió trabajar con el modelo 2 que, a pesar de que las tasas de clasificación errónea son elevadas (realmente todas lo son), es el de menor error al clasificar clientes sin capacidad crediticia, o bien clientes malos, es decir, su tasa global de clasificación correcta es buena (*tgdc*), ya que a pesar de que el modelo 1 presente la mejor *tgdc*, sus tasas de clasificación errónea son muy altas. En el Cuadro $C.3$ se aprecia el modelo seleccionado y se resaltan con negritas las variables representativas en el modelo con base en el *p-valor* de cada una.

A partir del modelo 2 se comenzarán a desglosar diversos submodelos con el objetivo de encontrar aquel que ajuste mejor y a su vez clasifique mejor. Para esto se ocuparán diversas técnicas de selección de modelos antes mencionadas (*Backward, Fordward* y *Stepwise*), así como el conocimiento aprendido en la sección exploratoria para mantener o excluir variables de acuerdo a las necesidades del negocio.

Cuadro $C.3$. Modelo 2: modelo saturado

Nombre en el texto	Estimación	Error Estándar	Wald	$Pr(>\|z\|)$	Límites de Confianza al 95 % Inferior	Superior
Intercept	0.8863	0.5125	2.9912	0.0837	-0.1181	1.8907
SaldosDc1	**-1.6981**	**0.1995**	**72.424**	**<.0001**	**-2.0892**	**-1.307**
SaldosDc2	**-41.2208**	**0.2025**	**36.3366**	**<.0001**	**-1.6177**	**-0.8238**
Duración	**-0.0292**	**0.00853**	**11.7238**	**0.0006**	**-0.046**	**-0.0125**
MorosidadDc2	**0.7588**	**0.2813**	**7.2744**	**0.007**	**0.2074**	**1.3103**
MorosidadDc3	**1.236**	**0.2923**	**17.8763**	**<.0001**	**0.663**	**1.809**
PropósitoDc2	**1.5531**	**0.3449**	**20.2807**	**<.0001**	**0.8771**	**2.229**
PropósitoDc3	**0.6071**	**0.2422**	**6.2803**	**0.0122**	**0.1323**	**1.0818**
PropósitoDc4	**0.8277**	**0.2294**	**13.0162**	**0.0003**	**0.378**	**1.2773**
PropósitoDc5	0.2273	0.4343	0.274	0.6007	-0.6239	1.0785
PropósitoDc6	-0.00069	0.3496	0.000	0.9984	-0.6858	0.6844
PropósitoDc7	**0.7031**	**0.2983**	**5.5549**	**0.0184**	**0.1184**	**1.2879**
CapacidadDc1	**-0.8682**	**0.2833**	**9.3932**	**0.0022**	**-1.4234**	**-0.313**
CapacidadDc2	**-0.6903**	**0.3212**	**4.6194**	**0.0316**	**-1.3197**	**-0.0608**
CapacidadDc3	-0.3432	0.2926	1.3754	0.2409	-0.9167	0.2303
Edad	**0.0152**	**0.00772**	**3.8731**	**0.0491**	**0.000062**	**0.0303**
Monto	**-0.0001**	**0.000041**	**5.9439**	**0.0148**	**-0.00018**	**-0.00002**
GéneroDc1	-0.2841	0.3608	0.6204	0.4309	-0.9912	0.4229
GéneroDc3	**0.5494**	**0.1859**	**8.7359**	**0.0031**	**0.1851**	**0.9137**
GéneroDc4	0.3069	0.2973	1.0657	0.3019	-0.2758	0.8896
Créditos_ODc2	-0.1173	0.3966	0.0874	0.7675	-0.8946	0.66
Créditos_ODc3	0.3629	0.2296	2.4993	0.1139	-0.087	0.8129
OcupaciónDc2	-0.0391	0.2018	0.0376	0.8463	-0.4347	0.3565
OcupaciónDc3	-0.1093	0.2854	0.1467	0.7017	-0.6687	0.4501

Con base en lo anterior, del modelo 2 se obtuvo el submodelo 2.1 aplicando la selección de variables mediante la eliminación hacia atrás, de tal manera que en el Cuadro $C.4$ se realiza el monitoreo de las pruebas de bondad de ajuste aplicadas para cada paso bajo la prueba de Hosmer-Lemeshow, y se observa que los submodelos no presentan falta de ajuste. Por otro lado, se logró obtener un modelo más pequeño con 11 variables, esto se ve expresado en los valores de AIC que disminuyeron en cada paso. Al momento de extraer el cociente de verosimilitud para el Submodelo final 2.1 se obtiene que $P(\chi^2_{12} > G) = 7.22e-44$ por lo que se rechaza la hipótesis nula $H_0 : \beta_1 = \ldots = \beta_p = 0$ *vs* $\beta_a : \beta_i \neq 0$ para alguna $i = 1, \ldots, m$ donde m representa el número de variables en el modelo, es decir, no se excluyen variables significativas del modelo.

Con el Cuadro $C.5$ se nota que las variables que quedaron en el modelo van acorde con las características que describen al conjunto de buenos solicitantes de crédito definida en la sección exploratoria sigue constante la idea de que el género y estado civil (hombre soltero o viudo) es significativa en el modelo, así como la capacidad de pago expresada por la variable de porcentaje de ingreso disponible para el pago del crédito (*CapacidadDc1*), aunque cabe mencionar que falta la variable de edad en años del cliente (*Edad*).

Cuadro $C.4$. Submodelo 2.1, pasos del procedimiento de eliminación hacia atrás

Paso	Variable removida	AIC	df	Devianza residual	G	$P(\chi^2 > G)$	Hosmer-Lemeshow HL	Hosmer-Lemeshow p-valor	Number de Vars.
0	Saturado	1019.8	976	971.84	249.89	4.754428e-40	10.6196	0.2241977	22
1	PropósitoDc6	1017.8	977	971.84	239.89	1.375963e-38	10.6197	0.224189	21
2	OcupaciónDc2	1015.9	978	971.87	249.86	4.051238e-41	6.8862	0.5489625	20
3	Créditos_ODc2	1014.0	979	971.96	249.77	1.18022e-41	10.1096	0.2574194	19
4	OcupaciónDc3	1012.1	980	972.08	249.65	3.397361e-42	10.6280	0.223678	18
5	PropósitoDc5	1010.4	981	972.39	249.34	1.040117e-42	9.1868	0.3267808	17
6	GéneroDc1	1009.0	982	973.03	248.70	3.615185e-43	9.7782	0.2809372	16
7	GéneroDc4	1008.4	983	974.45	247.28	1.764057e-43	9.5092	0.3011677	15
8	CapacidadDc3	1007.8	984	975.82	245.91	8.175973e-44	10.2923	0.2451076	14
9	CapacidadDc2	1008.9	985	978.86	242.87	8.131681e-44	10.2380	0.2487206	13
10	Edad	1009.9	986	981.87	239.86	7.761513e-44	6.5934	0.5810702	12
11	Créditos_ODc3	1010.9	987	984.89	236.84	7.224229e-44	8.6609	0.3717013	11

Null deviance: 1221.73 on 999 degrees of freedom
Hosmer-Lemeshow df: 8

En el Cuadro $C.5$ se presenta el submodelo 2.1 final ajustado, que comparado con el modelo 2 saturado se observa que se respetan casi todas las variables que en un inicio se marcaron con negritas y parecían significativas en el modelo.

Cuadro $C.5$. Submodelo 2.1, eliminación hacia atrás

Nombre en el texto	Estimación	Error Estándar	Wald	$Pr(>\|z\|)$	Límites de Confianza al 95 % Inferior	Superior
Intercept	1.2799	0.3458	13.703	0.0002	0.6051	1.9627
SaldosDc1	-1.6864	0.1968	73.4599	<.0001	-2.0773	-1.3051
SaldosDc2	-1.2116	0.2	36.685	<.0001	-1.6074	-0.8222
MorosidadDc2	0.8889	0.2685	10.9617	0.0009	0.3664	1.4212
MorosidadDc3	1.3788	0.2808	24.1164	<.0001	0.8328	1.9355
PropósitoDc2	1.4934	0.3311	20.3394	<.0001	0.8649	2.1677
Duración	-0.0324	0.0083	15.2044	<.0001	-0.0488	-0.0162
PropósitoDc3	0.4818	0.2223	4.6982	0.0302	0.0495	0.9218
PropósitoDc4	0.7555	0.2084	13.1458	0.0003	0.3505	1.1683
PropósitoDc7	0.5601	0.2822	3.9394	0.0472	0.0147	1.123
CapacidadDc1	-0.4682	0.1707	7.5233	0.0061	-0.8045	-0.1347
Monto	-0.00009	0.000038	5.4582	0.0195	-0.00016	-0.00001
GéneroDc3	0.5162	0.1635	9.9653	0.0016	0.197	0.8385

Al observar en el cuadro $C.6$ las tasas de clasificación errónea y clasificación correcta en cada paso de la selección se observa una transición de una tasa de 51 % de clientes mal clasificados como con capacidad crediticia, a una tasa de 55 %. Cabe mencionar que entra en cuestión el desarrollo a partir del paso 8, pues se incrementa nuevamente el valor de AIC y G del Cuadro $C.4$, y en el Cuadro $C.6$ se observa que las tasas de clasificación errónea no mejoran, lo que permite pensar que puede existir un mejor modelo; sin embargo, la tasa global de clasificación correcta mejora 0.57 %.

Cuadro C.6. Matriz de Confusión en % de cada paso del Submodelo 2.1

Step	mal→ mal	mal → bien	bien → mal	bien → bien	tgdc
0	49.00	51.00	9.29	90.71	90.85714
1	49.00	51.00	9.29	90.71	90.71429
2	48.67	51.33	9.57	90.43	90.42857
3	48.33	51.67	9.29	90.71	90.71429
4	48.33	51.67	9.43	90.57	90.57143
5	48.00	52.00	9.57	90.43	90.42857
6	48.00	52.00	9.57	90.43	90.42857
7	46.33	53.67	9.57	90.43	90.42857
8	47.67	52.33	8.57	91.14	91.14286
9	47.67	52.33	9.43	90.57	90.57143
10	45.33	54.67	9.29	90.71	90.71429
11	44.67	55.33	9.71	90.29	90.28571

Dado que en el Cuadro C.4 y Cuadro C.6 nuevamente se aprecia que en el paso 8 las tasas se elevan se generó el Submodelo 2.1.1, en el cual en vez de remover la variable *CapacidadDc3* se removió la variable *Edad* por fines de negocio. Sin embargo, aunque la prueba de Hosmer-Lemeshow muestra que este submodelo no presenta falta de ajuste (HL = 8.6609, $Pr(> |z|) = 0.3717$, $DF = 8$) se obtiene un valor de AIC igual (AIC = 1010.9) y las tasas de clasificación errónea no varían.

Por otro lado, aplicando el método de selección de variables *Stepwise* a partir del modelo 2 saturado se obtuvo el Submodelo 2.2. En el Cuadro C.7 se presenta el submodelo final ajustado.

Cuadro C.7. Submodelo 2.2, eliminación mediante Stepwise Selection

Nombre en el texto	Estimación	Error Estándar	Wald	$Pr(> \|z\|)$	Límites de Confianza al 95 % Inferior	Superior
Intercept	-1.5134	0.3344	20.4843	<.0001	-2.1749	-0.8619
SaldosDc1	1.676	0.1948	74.0356	<.0001	1.2985	2.0629
SaldosDc2	1.1949	0.1985	36.2543	<.0001	0.8085	1.5875
MorosidadDc2	-0.8436	0.2659	10.0683	0.0015	-1.3705	-0.326
MorosidadDc3	-1.3312	0.2797	22.6547	<.0001	-1.8856	-0.7872
PropósitoDc2	-1.2338	0.315	15.3453	<.0001	-1.878	-0.6382
Duración	0.0299	0.00815	13.4428	0.0002	0.014	0.046
PropósitoDc4	-0.5239	0.1889	7.6918	0.0055	-0.899	-0.1575
Monto	0.000087	0.000038	5.3662	0.0205	0.000014	0.000162
CapacidadDc1	0.4957	0.1695	8.5478	0.0035	0.1647	0.8299
GéneroDc3	-0.4987	0.1625	9.4192	0.0021	-0.8189	-0.1814

Al momento de aplicar la Prueba de Hosmer-Lemeshow se obtiene que el submodelo 2.2 no presenta falta de ajuste (HL = 7.7821, p-valor=0.4550, Número de Vars.= 8) y que el valor de AIC = 1017.3 es mayor al valor del submodelo 2.1 y 2.1.1.
A su vez se desarrolló otro submodelo, el cual también parte del modelo 2, sin embargo ahora se introducen interacciones entre las variables con mayores correlaciones tales como:

- *Saldos vs morosidad:* Saldo de cuentas en el *banco DM vs* pago de créditos previos.
- *Duración vs monto:* Duración en meses del crédito *vs* monto del crédito.
- *Morosidad vs edad:* Pago de créditos previos *vs* edad en años del cliente.
- *Morosidad vs créditos_ O:* Pago de créditos previos *vs* otros créditos vigentes.
- *Monto vs capacidad:* Monto del crédito *vs* porcentaje de ingresos disponibles para el pago.
- *Capacidad vs género:* Porcentaje de ingresos disponibles para el pago *vs* estado marital / género.

De tal manera que se genera el submodelo 2.3 que se presenta en el Cuadro $C.8$, el cual se obtuvo mediante el proceso de selección *Stepwise* y donde se observa que únicamente agrega la interacción *Monto*Duración* (*Duración:* duración en meses del crédito, *Monto:* monto del crédito); ninguna variable fue removida en el proceso.

Cuadro $C.8$. Submodelo 2.3, eliminación mediante Stepwise Selection con interacciones

Nombre en el texto	**Estimación**	**Error Estándar**	**Wald**	$Pr(>\|z\|)$	**Límites de Confianza al 95 %** **Inferior**	**Superior**
Intercept	-1.923	0.3799	25.6235	<.0001	-2.6752	-1.184
Monto	0.000219	0.000069	10.0636	0.0015	0.000086	0.000358
Duración	0.0497	0.0118	17.864	<.0001	0.0267	0.0729
SaldosDc1	1.6598	0.1957	71.9607	<.0001	1.2805	2.0483
SaldosDc2	1.1883	0.1991	35.6384	<.0001	0.8007	1.5821
MorosidadDc2	-0.8772	0.2667	10.8205	0.001	-1.4061	-0.3584
MorosidadDc3	-1.3667	0.2806	23.719	<.0001	-1.9233	-0.8213
PropósitoDc2	-1.31	0.317	17.079	<.0001	-1.958	-0.7103
PropósitoDc4	-0.5065	0.1902	7.0911	0.0077	-0.8842	-0.1375
CapacidadDc1	0.5063	0.1704	8.8341	0.003	0.1738	0.8422
GéneroDc3	-0.5213	0.1633	10.1943	0.0014	-0.8432	-0.2026
Monto*Duración	-0.00000448	0.000001921	5.4306	0.0198	-0.00000828	-0.000000737

El valor de AIC de este modelo es 1010.3, el cual es mejor al AIC de los submodelos anteriores, sin embargo considerando que el método de Selección *Stepwise* es complementario también se trabajó el submodelo 2.4 que introduce interacciones con base en las mismas correlaciones pero mediante el método de eliminación hacia atrás, de tal forma que en el Cuadro $C.9$ se aprecia el submodelo final.

Cuadro C.9. Submodelo 2.4, eliminación hacia atrás con interacciones

Nombre en el texto	Estimación	Error Estándar	Wald	$Pr(>\|z\|)$	Límites de Confianza al 95 % Inferior	Superior
Intercept	2.261	0.3948	5.725	1.03e-08	1.486678	3.034413
SaldosDc1	-1.815	0.1944	-9.336	< 2e-16	-2.196401	-1.434197
SaldosDc2	-1.251	0.1971	-6.346	2.21e-10	-1.636975	-8.644338e-01
Duración	-0.04768	0.01145	-4.164	3.13e-05	-7.011614e-02	-2.523396e-02
PrósitoDc2	1.411	0.3201	4.406	1.05e-05	7.830526e-01	2.037964
PrósitoDc3	0.4470	0.2132	2.096	0.036050	2.908638e-02	8.650075e-01
PrósitoDc4	0.6736	0.1974	3.413	0.000642	2.867902e-01	1.060441
Monto	-0.00211	0.0006754	-3.121	0.001803	-3.431749e-04	-7.841113e-05
CapacidadDc1	-0.4940	0.1686	-2.93	0.003384	-8.244396e-01	-1.636216e-01
GéneroDc3	0.5183	0.1635	3.170	0.001523	1.978687e-01	8.387096e-01
Edad	0.01489	0.007352	2.025	0.042860	4.787001e-04	2.929855e-02
Duración*Monto	0.00004	0.00001837	2.207	0.027297	4.542189e-07	7.654378e-06

Al aplicar la Prueba de Hosmer-Lemeshow se observa que el modelo no presenta falta de ajuste (HL = 2.4292, p-valor=0.9650, Número de Vars.= 8) y al verificar la Prueba de Ji-cuadrada con respecto de un paso anterior, donde se elimina la variable *CapacidadDc2*, se rechaza la prueba de hipótesis $H_0 : \beta_1 = \ldots = \beta_m = 0$ vs $H_a : \beta_i \neq 0$ para alguna $i = 1, \ldots, m$ donde m representa el número de variables en el modelo, ya que $\chi^2_{(12)} = 2.9974$ y $[Pr > \chi^2_{(12)}] = 0.0834$ entonces no todas las variables son cero, mientras que para el modelo final se tiene $\chi^2_{(11)} = 2.8095$ y $[Pr > \chi^2_{(11)}] = 0.0937$. Las variables significativas entre el submodelo 2.3 y el submodelo 2.4 varían principalmente en que el segundo mantiene las variables *MorosidadDc2* y *MorosidadDc3*, *PropósitoDc3* y *Edad*.

Cuadro C.10. Submodelo 2.4.1, eliminación hacia atrás con iteraciones eliminando la variable *alter*

Nombre en el texto	Estimación	Error Estándar	Wald	$Pr(>\|z\|)$	Límites de Confianza al 95 % Inferior	Superior
Intercept	1.707	0.3880	4.398	1.09e-05	0.9461	2.4670
SaldosDc1	-1.671	0.1977	-8.452	< 2e-16	-2.0588	-1.2837
SaldosDc2	-1.203	0.2007	-5.996	2.02e-09	-1.5966	-0.8099
Duración	-0.05359	0.01185	-4.521	6.16e-06	-0.07682	-0.0304
MorosidadDc2	0.9255	0.2695	3.434	0.000595	0.3973	1.4538
MorosidadDc3	1.416	0.2817	5.028	4.96e-07	0.8643	1.9686
PropósitoDc2	1.597	0.3351	4.767	1.87e-06	0.9405	0.0225
PropósitoDc3	0.5256	0.2247	2.339	0.019335	0.08517	0.9660
PropósitoDc4	0.7526	0.2097	3.589	0.000332	0.3416	1.1635
PropósitoDc7	0.5771	0.2814	2.051	0.040257	0.0256	1.1286
Monto	-0.004313	7.008e-05	-3.300	0.000967	-3.686e-04	-9.3906e-05
CapacidadDc1	-0.4771	0.1716	-2.781	0.005425	-0.8134	-0.1408187
GéneroDc3	0.5420	0.1644	3.296	0.000980	0.2197	0.8642
Duración*Monto	4.804e-06	1.924e-06	2.497	0.012531	1.033e-06	8.5753e-06

Se recurrirá nuevamente a la modificación de las variables de salida donde se observa en qué paso se presenta el menor ajuste y el incremento del valor AIC, y en vez de la variable en cuestión se retira la variable *alter*, esto debido a que es una variable que se puede filtrar desde un inicio en la base de datos, ya que existen políticas de la institución en la que solamente proceden solicitantes de cierto rango de edad. Dado lo anterior, se obtiene el submodelo 2.4.1, del cual el modelo final se encuentra en la tabla $C.10$.

En el submodelo 2.4.1 se observa un mejor AIC = 1006.5 *vs* el valor de AIC = 1028.1 del submodelo 2.4, lo cual indica que se tiene un modelo más complejo. Al obtener la prueba de Hosmer-Lemeshow se obtiene que el modelo ajusta bien a los datos (HL = 3.3944, p-valor=0.9072, Número de Vars.= 8) por lo que hay evidencia para rechaza la hipótesis nula $H_0 : \beta_1 = \ldots = \beta_m = 0$ vs $H_a : \beta_i \neq 0$ para alguna $i = 1, \ldots, m$ donde m representa el número de variables no incluidas en el modelo. Adicional se obtiene $P(\chi^2_{13} > G) = 1.5765e - 44$. Como conclusión, resulta un modelo factible excluir la variable de edad del cliente ya que se puede acotar la base de solicitantes desde un inicio, sin embargo se tiene que validar que la significancia de la variable no afecte en futuras implementaciones.

Tratando de encontrar un mejor modelo se sometió el modelo 2 saturado al método de selección por los valores de AIC mediante la función *step()* del *software R* y se obtuvo el submodelo 2.5. En el Cuadro $C.11$ se observa el modelo final junto con sus estadísticos.

Cuadro $C.11$. Submodelo 2.5, eliminación mediante AIC

Nombre en el texto	**Estimación**	**Error Estándar**	**Wald**	$Pr(>\|z\|)$	**Límites de Confianza al 95 % Inferior**	**Superior**
Intercept	0.7558	0.4520	1.672	0.094504	-0.1301	1.6418
SaldosDc1	-1.682	0.1980	-8.492	< 2e-16	-2.0697	-1.2935
SaldosDc2	-1.191	0.2008	-5.933	2.98e-09	-1.5849	-0.7978
Duración	-0.03031	0.08346	-3.631	0.000282	-0.0467	-0.0139
MorosidadDc2	0.7634	0.2799	2.728	0.006380	0.2148	1.3119
MorosidadDc3	1.226	0.2907	4.219	2.45e-05	0.6566	1.7959
PropósitoDc2	1.524	0.3345	4.556	5.22e-06	0.8683	2.1794
PropósitoDc3	0.5245	0.2252	2.329	0.019871	0.0831	0.9659
PropósitoDc4	0.8064	0.2113	3.816	0.000136	0.3922	1.2206
PropósitoDc7	0.6299	0.2825	2.230	0.025775	0.0762	1.1836
Monto	-0.0010	3.836e-05	-2.668	0.007635	-0.0002	-2.71e-05
CapacidadDc1	-0.6339	0.1939	-3.270	0.001076	-1.01387	-0.2539
CapacidadDc2	-0.43961	0.2507	-1.753	0.079519	-0.9309	0.0518
GéneroDc3	0.50791	0.1675	3.033	0.002425	0.1796	0.8361
Edad	0.01336	0.077537	1.773	0.076175	-0.0014	0.0281
Créditos_ODc3	0.3758	0.2041	1.842	0.065543	-0.0242	0.7758

Para el submodelo 2.5 se obtiene un $AIC = 1007.8$, la prueba de Hosmer-Lemeshow indica que el modelo ajusta bien a los datos (HL = 10.2922, p-valor=0.2451, Nú-

mero de Vars.= 8), además la prueba de cociente de verosimilitudes con respecto al modelo 2 saturado muestra una $G = 245.91$ con $P[\chi^2_{15} > G] = 8.175973e - 44$ lo cual implica que la prueba de hipótesis nula se rechaza, $H_0 : \beta_1 = \ldots = \beta_m = 0$ vs $H_a : \beta_i \neq 0$ para alguna $i = 1, \ldots, m$ donde m representa el número de variables incluidas en el modelo.

Finalmente, se trabajaron dos submodelos adicionales: submodelo 2.6 y submodelos 2.7, en el primero se introdujeron únicamente las variables continuas duración en meses del crédito (*Duración*), monto del crédito (*Monto*) y edad en años del cliente (*Edad*), mientras que en el segundo submodelo se consideraron únicamente las variables binarias en base al modelo 2 saturado esto con el objetivo de conocer si al mezclar variables binarias con continuas, las segundas, no afectaban en mayor proporción el comportamiento de las binarias o viceversa. En otras palabras, se busca conocer si el grupo de variables continuas o el grupo de variables binarias son suficientes para discriminar la base y clasificar adecuadamente para así disminuir el número de variables no incluidas en el modelo.

Con base en lo anterior, en el submodelo 2.6 se obtiene un $AIC = 1176.3$, mientras que para el submodelo 2.7 el valor es AIC =1063.3, según la prueba de Hosmer-Lemeshow ambos submodelos ajustan bien a los datos (HL = 9.0642, p-valor=0.3369, Número de Vars.= 8 para el submodelo 2.6 y HL = 6.5440, p-valor=0.5865, Número de Vars.= 8 para el submodelo 2.7); por otro lado, el submodelo 2.6 muestra una $G = 53.43$ con $P[\chi^2_3 > G] = 1.4576e - 11$ mientras que el submodelo 2.7 presenta un valor de $G = 200.13$ con $P[\chi^2_{20} > G] = 9.662721e - 32$, lo cual implica que la prueba de hipótesis nula se rechaza para ambos casos, $H_0 : \beta_1 = \ldots = \beta_m = 0$ vs $H_a : \beta_i \neq 0$ para alguna $i = 1, \ldots, m$ donde m representa el número de variables incluidas en el modelo. Sin embargo, al analizar las tasas de mal clasificación, mal $\rightarrow$ bien y bien $\rightarrow$ mal, el submodelo 2.6 presenta tasas de 88 % y 3.9 % respectivamente, mientras que el submodelo 2.7 presenta tasas de 62 % y 8.9 %, las cuales están por encima de cualquier otro submodelo e inclusive del modelo saturado, con lo que se concluye que el manejo del conjunto de variables binarias así como el conjunto de variables continuas, generan una mejor clasificación al trabajar en conjunto.

3.2.6. Conclusiones del modelo

Se han resumido los valores de las pruebas de cada submodelo presente anteriormente en la Cuadro $C.12$, así como los valores del modelo saturado seleccionado en un inicio con el fin de detectar el mejor modelo. Por otro lado, también son de gran ayuda las matrices de confusión finales de cada modelo y submodelo que se muestran en el cuadro $C.13$.

Dado los datos de los Cuadros $C.12$ y $C.13$, se tiene que el modelo saturado 2 con 23 variables presenta las mejores tasas de mal clasificación, seguidas del submodelo 2.5 que además es el segundo modelo con el mejor valor AIC, sin embargo se prefiere el submodelo 2.5 debido a que su tasa global de clasificación correcta es la mejor de todos los modelos presentados. Además, las variables consideradas en el modelo se sustentan en su totalidad con la conclusión de la sección exploratoria. Finalmente se marcan con negritas estos dos modelos (modelo 2 saturado y submodelo 2.5) además del submodelo 2.1 que sirve de contraste.

Cuadro $C.12$. Comparación de modelos y submodelos

				Pruebas estadísticas			
Modelo/	**Num. de**	**Método de**	**Valor**	**Hosmer-Lemeshow**		**Razón de Verosimilitudes**	
Submodelo	**variables**	**selección**	**AIC**	HL	p-valor	χ^2_x	$[Pr > \chi^2_x]$
Modelo 2	**23**	**Saturado**	**1019.8**	**10.6196**	**0.2242**		
Submodelo 2.1	**12**	**Backward**	**1010.9**	**8.6609**	**0.3717**	**3.0141**	**0.0825**
Submodelo 2.1.1	10	Backward	1010.9	8.6609	0.3717	3.0141	0.0825
Submodelo 2.2	10	Stepwise	1013.8	7.7821	0.4550	5.4151	0.0199
Submodelo 2.3	11	Stepwise	1010.3	7.1137	0.5244	10.9280	0.0042
Submodelo 2.4	11	Backward	1028.1	2.4292	0.9650	2,8095	0.0937
Submodelo 2.4.1	13	Backward	1006.5	3.3944	0.9072	2.8804	0.0897
Submodelo 2.5	**15**	**AIC**	**1007.8**	**10.2923**	**0.2451**	**2.8803**	**0.0896**

Cuadro $C.13$. Matriz de Confusión en % de modelos y submodelos ajustados

Modelo	**mal→ mal**	**mal → bien**	**bien → mal**	**bien → bien**	**tgdc**
Modelo 2	**49.00**	**51.00**	**9.29**	**90.71**	**90.71429**
Submodelo 2.1	**44.67**	**55.33**	**9.71**	**90.29**	**90.28571**
Submodelo 2.1.1	44.67	55.33	9.71	90.29	90.28571
Submodelo 2.2	45.00	55.00	9.86	90.14	90.14286
Submodelo 2.3	46.33	53.67	10.00	90.00	90.00000
Submodelo 2.4	43.00	57.00	11.29	88.71	88.71429
Submodelo 2.4.1	46.33	53.67	10.57	89.43	89.42857
Submodelo 2.5	**47.67**	**52.33**	**8.86**	**91.14**	**91.14286**

Por otro lado, mediante el gráfico 3.1 y las curvas ROC se puede apreciar que a pesar de que la curva del modelo 2 saturado (de color rojo) sobrepasa a las otras dos curvas casi en cada valor del las tasa falsas positivas, existe un pequeño rango de

puntos cercanos a la coordenada (0.1, 0.38) donde el submodelo 2.5 presenta mejores tasas.

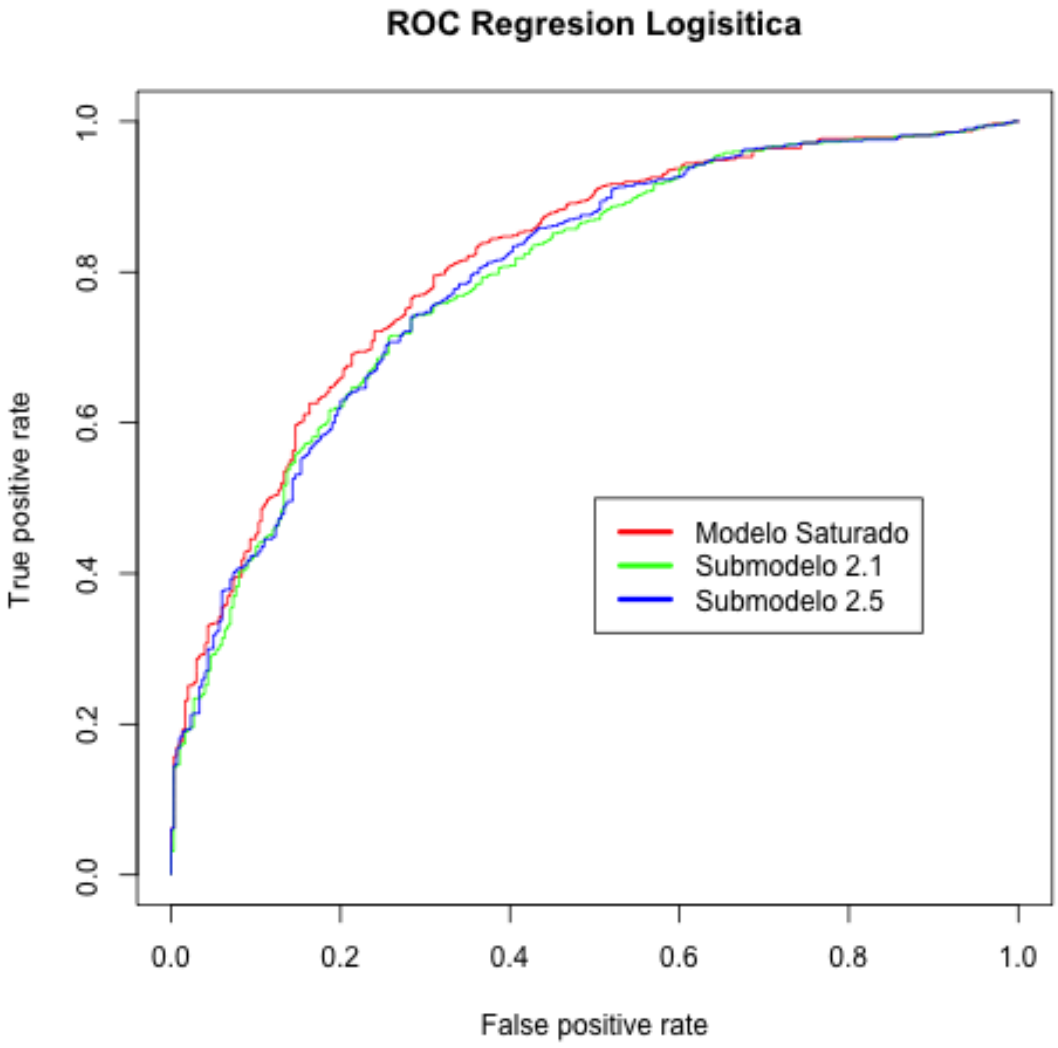

Figura 3.1: Curvas ROC para comparar los modelos de regresión logística: modelo 2 saturado (colo rojo), submodelo 2.1 (color verde) y submodelo 2.5 (color azul).

Finalmente, es posible enlazar estas conclusiones mediante la interpretación de los *momios*. Se toma como ejemplo el submodelo 2.5 que fue el de mayor atención en el análisis, se aplican las exponenciales de los coeficientes así como de los valores de los intervalos de confianza obtenidos, de tal manera que en el Cuadro $C.14$ se aprecian los valores.

Cuadro $C.14$. Submodelo 2.5 con exponenciales en coeficientes e intervalos de confianza

Variable	exp(coef)	exp(lim inf)	exp(lim sup)
SaldosDc1	5.374	3.645	7.922
SaldosDc2	3.291	2.221	4.879
Duración	1.031	1.014	1.048
MorosidadDc2	0.466	0.269	0.807
MorosidadDc3	0.293	0.166	0.519
PropósitoDc2	0.218	0.113	0.42
PropósitoDc3	0.592	0.381	0.92
PropósitoDc4	0.446	0.295	0.676
PropósitoDc7	0.533	0.306	0.927
Monto	1	1	1
CapacidadDc1	1.885	1.289	2.756
CapacidadDc2	1.552	0.95	2.537
GéneroDc3	0.602	0.433	0.836
Edad	0.987	0.972	1.001
Créditos_ODc3	0.687	0.46	1.024

En el cuadro $C.14$ se aprecia que para el caso de variables dicotómicas los intervalos de confianza ayudan a identificar si la variable puede llegar a tomar el valor de 1, mismo valor en el cual el incremento en la razón de riesgo es indefinido, en el caso de que el intervalo de confianza toque este valor es recomendable excluir la variable.

Por ejemplo, en la variable de pago de créditos previos en su categoría 2 equivalente a sin créditos previos o con créditos previos saldados (*MorosidadDc2*), el incremento en el riesgo al ocurrir el evento de que el cliente sea mal acreedor presenta una tasa del 46.6 % mientras que para la variable de porcentaje de ingresos disponible para el pago en su categoría uno equivalente a $< 20\,\%$ (*CapacidadDc2*), que resulta mayor a uno, indica que en la presencia de la variable la ocurrencia de que se trate de un cliente sin capacidad crediticia sucede menos que en la categoría de referencia, ya que el incremento en una unidad se obtiene la siguiente categoría. Es interesante notar que para la variable que representa el monto del crédito solicitado al *banco DM* (*Monto*) el valor del *momio* $= 1$ tanto para el valor del coeficiente como para los valores del intervalo de confianza, lo que expresa independencia entre los eventos, es decir, que no es significativo el monto del crédito solicitado para deducir si un solicitante a crédito será un buen acreedor o uno malo.

En el gráfico 3.2 se pueden apreciar los valores de los *momios* para cada variable del submodelo 2.5.

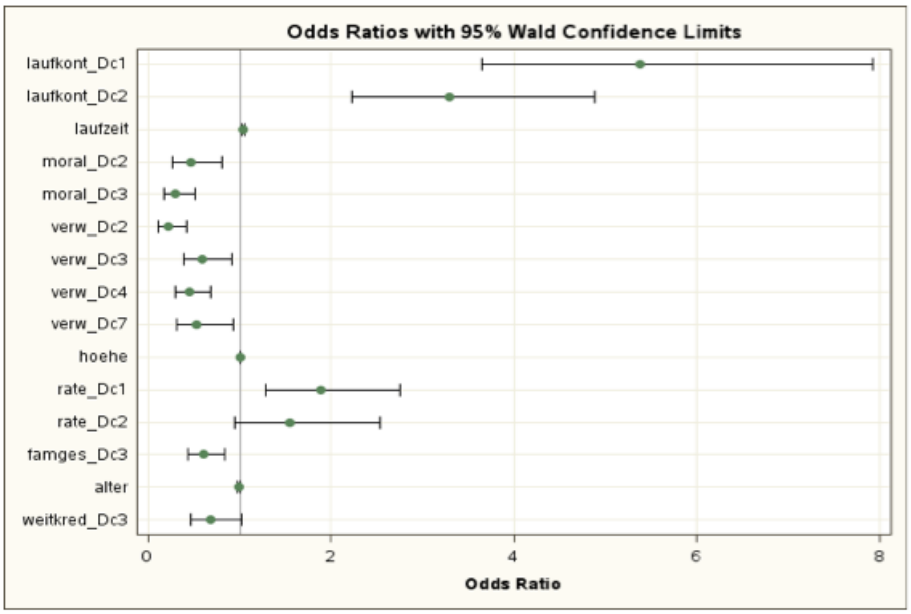

Figura 3.2: Gráfica de *momios* para las variables del submodelo 2.5.

Finalmente, si se observa la *razón de momios* de las variables *GéneroDc3* con respecto a la variable *CapacidadDc1* se tiene que $\theta = 1.885 \ / \ 0.602 = 3.1312$, lo cual implica que aumenta un 313 % el hecho de ser un buen acreedor el ser un hombre soltero o viudo con una capacidad de pago menor al 20 %.

Este tipo de análisis es muy útil a nivel negocio, pues se pueden encontrar razones de peso para clasificar a los solicitantes.

3.3. Redes Neuronales

La estructura del modelo de redes neuronales se construye a partir de elementos llamados nodos o neuronas, los cuales reciben un conjunto de entradas, o *inputs*, a través de una variable vectorial $\mathbf{x} = (x_0, x_1, \ldots, x_p)$, posteriormente a estas unidades se les aplica una ponderación para calcular variables escalares de salida, o *outputs*, de tal forma que se le añade una constante de sesgo w_0 para así transformar el resultado a una forma no lineal:

$$z = f(\mathbf{w}'\mathbf{x}), \tag{3.23}$$

donde $\mathbf{w}$ es un vector de ponderaciones y f una función que generalmente es no lineal, además al vector $\mathbf{x}$ se le incorpora la variable $x_0 = 1$, de tal manera que acompaña al término de sesgo w_0, incluido en el vector de pesos $\mathbf{w}' = (w_0, \ldots, w_p)$

$\Longrightarrow$

$$z = f\left(\sum_{i=0}^{p} \mathbf{w_i x_i}\right) = f\left[(w_0, \ldots, w_p)\begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_p \end{pmatrix}\right].$$

En la figura 3.3 se muestra un modelo abstracto de una neurona con n *inputs* tal que el $i-$ésimo canal transmite el valor del $i-$ésimo *input* x_i además, a cada $i-$ésimo canal se le asocia un ponderador, o *weight* w_i, de tal forma que cada valor de x_i es multiplicado por este peso y toda esta información es "transmitida" por los canales hasta llegar al centro de la neurona o *nodo*, donde la información será evaluada bajo la función f, a la cual se le conoce como *función primitiva*. En este modelo abstracto, al cual se le llama *simple perceptron*, sólo se menciona un nodo; sin embargo, en una red neuronal de mayor tamaño conocida como *multilayer perceptron* este modelo actuaría como una neurona de una capa intermedia.

Por otro lado, en las variables iniciales la función de respuesta es la unidad, es decir $\mathbf{x} = f(x)$, pues es el inicio de la intercomunicación entre los nodos. Sin embargo, existe otra función $\mathbf{g}$ que se implementa al incrementar el número de capas de nodos ocultas y a la cual se le conoce por *función de activación* que actúa como una función de liga de una regresión logística de hecho, usualmente se puede designar esta función como la función logística tal que $\pi(x) = (1 + \exp^{-x})^{-1}$. (Figura 3.4).

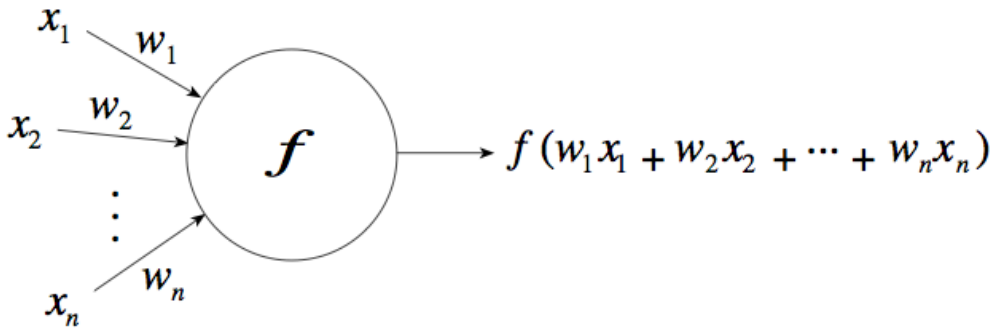

Figura 3.3: Esquema de una neurona abstracta donde cada flecha denota un canal por donde viajan los valores de cada x_i bajo una ponderación específica w_i de cada canal, de tal forma que toda esta información se une en el centro de la neurona donde se la aplica un función primitiva f antes definida. Fuente: Rojas, R. (1996, p.26)

Ahora bien, retomando el modelo de regresión lineal simple,

$$\begin{aligned} Z(\mathbf{x_j}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \\ &= (\beta_0 + \beta_{\mathbf{i}}' \mathbf{x_i}) = \textstyle\sum_{j=1}^{p} \beta_{\mathbf{j}} \mathbf{x_j}, \end{aligned} \tag{3.24}$$

tal que cada característica, x_j, está representada como una combinación lineal, $Z(x_j)$, de tal forma que la variable *target* está modelada como función de combinaciones lineales de Z_j, es decir:

$$\begin{aligned} Z_j &= s(\omega_{0j} + \omega_{\mathbf{j}}' \mathbf{x}) \quad j = 1, \ldots, N, \\ f_i &= \omega_{0i} + \omega_{\mathbf{i}}' \mathbf{Z} \qquad i = 1, \ldots, H, \\ \Rightarrow & \qquad\qquad y_j(x) = g_j(f) \end{aligned} \tag{3.25}$$

se denota la estructura del modelo de NN's como:

$$y(\mathbf{x}, \omega, H) = g\left(\omega_0 + \sum_{j=1}^{H} \omega_{0j} f\left(\omega_{j0} + \sum_{k=1}^{N_{in}} \omega_{jk} x_k\right)\right), \tag{3.26}$$

donde H representa el número de capas ocultas y N_{in} el número de nodos ocultos dado lo anterior esta ecuación es asociada a un modelo MLP (*Multilayered Perceptron*).

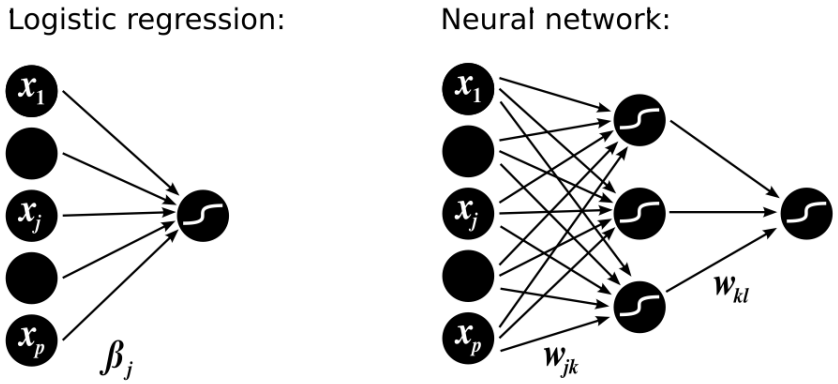

Figura 3.4: La función logística que fue desarrollada en la sección anterior puede ejemplificarse por la figura de la izquierda, mientras que aplicando esta misma función como función de respuesta del modelo de redes neuronales con una capa oculta se obtiene la imagen de la derecha. Fuente: Vos and Evers (2004, p. 89).

3.3.1. Estimación de parámetros

Ahora bien, note que en la ecuación 3.26 existen dos tipos de parámetros, los cuales requieren tratos distintos: los ponderadores, que sus estimaciones son llamados *training* o prueba y H, que es el número de nodos ocultos presentes en el modelo.

La estimación de los parámetros puede ser enfocada a cualquier modelo de selección o a un modelo que involucre medias. Balakrishnan (2010, p.347) cita a Titterington, quien dice que otras tendencias con asociaciones estadísticas han incluido el uso o alternativas a los mínimos cuadrados que pueden ser interpretadas como reglas de máxima verosimilitud, y el uso de diversos métodos de regularización para hacer frente a la multiplicidad frecuente de óptimos locales así como a la parametrización típicamente alta de los modelos; estos incluyen el uso de términos destinados a eliminar pesos no influyentes (llamada *weight decay* en la literatura de redes neuronales).

Dicho lo anterior, surge la duda del por qué H es considerado como parámetro en el modelo y la respuesta va de la mano con conocer la importancia del número de capas ocultas que contenga el modelo de redes neuronales con estructura de *perceptron*.

Como indica Titterington (Balakrishnan, 2010, p.347), en primer lugar se analizan los caminos para estimar los ponderadores para lo cual, como en cualquier modelo de regresión (retomando la relación entre NN y regresión), el conjunto de prueba intenta minimizar el error de la función de tal forma que se busca la manera de maximizar la probabilidad de obtener los datos, es decir, calcular el error cuadrático medio:

$$\mathbb{E}(\omega) = \frac{1}{n}\sum_{i=1}^{N_{in}}\Big[y(x_i,\omega) - t\Big]^2 = \frac{1}{n}\sum_{i=1}^{N_{in}}\Big[y - t\Big]^2, \tag{3.27}$$

mientras que para un problema de clasificación la ecuación anterior es renombrada como *cross-entropy* o bien *deviance*,

$$\begin{aligned}\mathbb{E}(\omega) &= -\tfrac{1}{n}\textstyle\sum_{i=1}^{p}\Big[t\log y(x_i,\omega) + (1-t)\log(1-y(x_i,\omega))\Big] \\ &= -\tfrac{1}{n}\textstyle\sum_{i=1}^{p}\Big[t\log(y) + (1-t)\log(1-y)\Big],\end{aligned} \tag{3.28}$$

donde t reemplaza a la probabilidad p en la ecuación 3.13 y hace referencia al valor de las variables *target*, $\frac{1}{n}$ al tamaño de la muestra de prueba con n igual al número de elementos considerados y $y(x_i,\omega)$ está dado por la ecuación 3.26, de tal forma que el primer modelo se basa en la diferencia entre el valor verdadero y la predicción, mientras que el segundo busca (el logaritmo) de su razón o *ratio*.

Por otro lado ¿por qué es que se considera al número de nodos ocultos, H, como otro parámetro? Según expresa Hastie et al. (2009, p. 400): "es mejor tener un gran número de nodos ocultos, ya que las ponderaciones extras pueden ser reducidas a cero con un método apropiado. De lo contrario, al usar pocos nodos ocultos el modelo puede no tener la suficiente flexibilidad para acotar la no-linealidad de los datos. De hecho, este número de unidades ocultas oscila entre 5 y 100 y algunos investigadores aplican el modelo de *cross-validation* para dicho cometido. Sin embargo, el número de capas ocultas va de la mano del conocimiento de los datos por parte del analista, ya que cada capa oculta es un extracto de las características construidas a partir de los *inputs* del modelo, lo cual implica que a su vez cada capa se traduce como una construcción jerárquica de las características a diversos niveles de análisis."

3.3.2. Sección práctica

Para el desarrollo práctico del modelo de redes neuronales se utilizó el *software* *R* y su librería *nnet*, el cual permite variar los parámetros de decaimiento (*decay*), rango (*rang*) y número de capas (*size*). Se programó un ciclo anidado mediante 3 iteraciones anidadas (*fors*), con el objetivo de correr una red neuronal por cada posible combinación de los valores que toman los 3 parámetros antes mencionados, tal que:

- *decay* = { 0.1, 0.2, ..., 1},

- *rang* = { 0.1, 0.2, ..., 1 },
- *size* = { 1, 2, ..., 20 }.

En el mismo ciclo anidado se pide que los valores de cada parámetro, así como las tasas de clasificación errónea, se guarden en una matriz para poder realizar el análisis de las salidas de cada red, dicha matriz tuvo una salida de 2,000 registros. Cabe mencionar que la función *nnet* también permite decidir un número máximo de iteraciones permitidas en cada Red, sin embargo para fines prácticos se decidió fijar el parámetro a máximo 500 iteraciones.

Posteriormente se definió el conjunto de variables del modelo saturado seleccionado en la regresión logística (modelo 2 saturado, que contiene 3 variables continuas y 20 dicotómicas) para correr el ciclo anidado antes mencionado, y se seleccionó por cada capa la combinación de parámetros que generaron las tasas de error más bajas. En el Cuadro $C.15$ se muestran las tasas de clasificación errónea donde m⟶b indica la tasa de clientes sin capacidad crediticia clasificados como clientes con capacidad crediticia y b⟶m indica la tasas de clientes con capacidad crediticia clasificados como clientes sin capacidad crediticia, a su vez, la fila marcada con negritas indica las menores tasas de clasificación errónea, la cual se obtiene en la última capa inclusive, a su vez se aprecia que conforme se agregan capas al modelo éste ajusta mejor ya que la primer capa se puede relacionar a la regresión logística del modelo 2 saturado donde las tasa son muy altas. En la figura 3.5 se observan las curvas ROC de las capas 1, 5, 10, 15 y 20 con base en los parámetros del Cuadro $C.15$.

Cuadro $C.15$. Matriz de salidas de la red neuronal 1 con variables de entrada del modelo 2 saturado

Size	**Fila**	***Rang***	***Decay***	**m⟶b** (%)	**b⟶m** (%)
1	[13,]	0.2	0.3	52.67	10.57
2	[101,]	0.1	0.1	47.33	9.43
3	[231,]	0.4	0.1	39.33	10.43
4	[321,]	0.3	0.1	37.33	9.71
5	[481,]	0.9	0.1	37.00	8.00
6	[561,]	0.7	0.1	32.33	9.29
7	[641,]	0.5	0.1	27.33	11.29
8	[711,]	0.2	0.1	24.67	5.43
9	[881,]	0.9	0.1	25.67	6.00
10	[951,]	0.6	0.1	21.33	4.00
11	[1031,]	0.4	0.1	17.33	3.71
12	[1171,]	0.8	0.1	14.00	2.86
13	[1201,]	0.1	0.1	13.33	2.43
14	[1381,]	0.9	0.1	11.00	3.29
15	[1401,]	0.1	0.1	10.00	1.86
16	[1571,]	0.8	0.1	7.33	1.57
17	[1671,]	0.8	0.1	7.00	1.29
18	[1791,]	1	0.1	6.33	1.00
19	[1811,]	0.2	0.1	6.00	1.57
20	**[1901,]**	**0.1**	**0.1**	**5.33**	**1.29**

Cuadro C.16. Matriz de salidas de la red neuronal 2 con variables de entrada del submodelo 2.1

Size	**Fila**	***Rang***	***Decay***	**m⟶b** (%)	**b⟶m** (%)
1	[1,]	0.1	0.1	49.33	0.1129
2	[111,]	0.2	0.1	41.33	10.57
3	[241,]	0.5	0.1	40.00	10.57
4	[361,]	0.7	0.1	33.67	9.14
5	[451,]	0.6	0.1	29.33	7.29
6	[571,]	0.8	0.1	22.00	5.43
7	[611,]	0.2	0.1	18.67	4.00
8	[751,]	0.6	0.1	15.00	4.43
9	[881,]	0.9	0.1	10.67	2.00
10	[921,]	0.3	0.1	8.67	1.86
11	[1011,]	0.2	0.1	8.00	1.86
12	[1141,]	0.5	0.1	4.33	1.00
13	[1141,]	0.5	0.1	4.33	1.00
14	[1341,]	0.5	0.1	2.00	0.86
15	[1411,]	0.2	0.1	2.33	0.14
16	[1501,]	0.1	0.1	2.67	0.14
17	[1671,]	0.8	0.1	1.67	0.14
18	[1761,]	0.7	0.1	1.33	0.14
19	[1881,]	0.9	0.1	1.67	0.00
20	**[1961,]**	**0.7**	**0.1**	**0.33**	**0.29**

Cuadro C.17. Matriz de salidas de la red neuronal 3 con variables de entrada del submodelo 2.5

Size	**Fila**	***Rang***	***Decay***	**m⟶b** (%)	**b⟶m** (%)
1	[1,]	0.1	0.1	50.33	0.12
2	[101,]	0.1	0.1	43.00	12.71
3	[201,]	0.1	0.1	38.00	8.57
4	[311,]	0.2	0.1	36.00	9.14
5	[441,]	0.5	0.1	31.67	8.29
6	[551,]	0.6	0.1	26.33	6.43
7	[661,]	0.7	0.1	22.00	5.71
8	[741,]	0.5	0.1	16.67	3.43
9	[861,]	0.7	0.1	15.00	3.71
10	[911,]	0.2	0.1	12.00	3.57
11	[1031,]	0.4	0.1	10.00	2.00
12	[1161,]	0.7	0.1	8.67	1.71
13	[1221,]	0.3	0.1	5.67	1.43
14	[1371,]	0.8	0.1	4.33	1.71
15	[1461,]	0.7	0.1	4.33	1.00
16	[1521,]	0.3	0.1	4.00	0.86
17	[1671,]	0.8	0.1	3.67	0.71
18	[1721,]	0.3	0.1	4.00	0.57
19	[1841,]	0.5	0.1	2.33	0.57
20	**[1981,]**	**0.9**	**0.1**	**3.33**	**0.43**

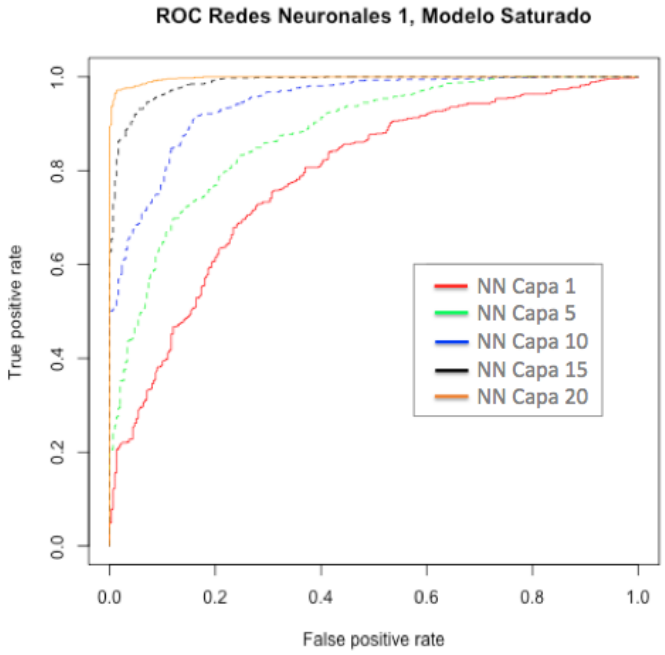

Figura 3.5: Curvas ROC de redes neuronales con base en las variables del modelo 2 saturado. Se grafican las menores tasas de error presentes en las capas 1, 5, 10, 15 y 20 con base en el Cuadro $C.15$.

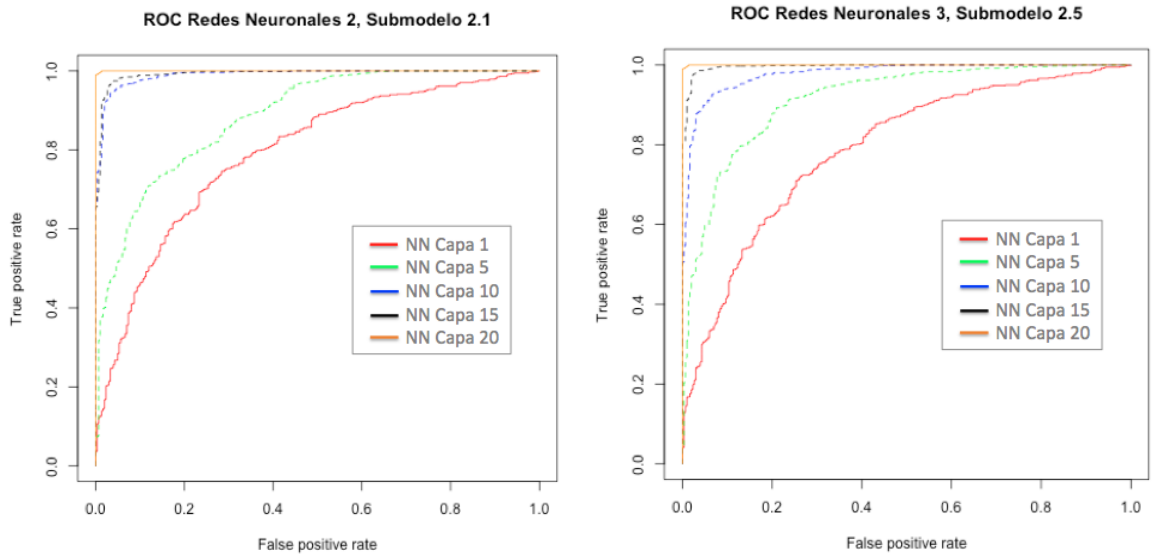

Figura 3.6: Del lado izquierdo: gráfica ROC de NN 2, conjunto de variables del submodelo 2.1, se grafican las menores tasas de error presentes en las capas 1, 5, 10, 15 y 20 con base en el Cuadro $C.16$. Del lado derecho: gráfica ROC de NN 3, conjunto de variables del submodelo 2.5, se grafican las menores tasas de error presentes en las capas 1, 5, 10, 15 y 20 con base en el Cuadro $C.17$

Finalmente, como sugiere Hastie et al. (2009), se tomó aleatoriamente una muestra de la base del 30 % (30 % con *kredit* = 1 y 30 % con *kredit* = 0) con el objetivo de estimar sobre ella los parámetros y probarlos sobre el 70 % del resto de la base. Los resultados obtenidos al ejecutar el programa con 2,000 iteraciones resultó con las mejores tasas de clasificación errónea (las más bajas) en la capa 8 con una tasa = 0 de clientes sin capacidad crediticia clasificados como clientes con capacidad crediticia y una tasa 0.48 % de clientes con capacidad crediticia clasificados como clientes sin capacidad crediticia, en el Cuadro $C.18$ se aprecian dichas tasas.

Cuadro $C.18$. Matriz de salidas de la red neuronal *test* con variables de entrada del modelo 2 saturado y el 30 % de la base tomada aleatoriamente

Size	**Fila**	***Rang***	***Decay***	**m⟶b** (%)	**b⟶m** (%)
8	**[761,]**	**0.7**	**0.1**	**0.00**	**0.47**
9	[861,]	0.7	0.1	0.00	0.47
10	[901,]	0.1	0.1	0.00	0.47
11	[1001,]	0.1	0.1	0.00	0.47
12	[1101,]	0.1	0.1	0.00	0.47
13	[1201,]	0.1	0.1	0.00	0.47
14	[1301,]	0.1	0.1	0.00	0.47
15	[1401,]	0.1	0.1	0.00	0.47
16	[1501,]	0.1	0.1	0.00	0.47
17	[1601,]	0.1	0.1	0.00	0.47
18	[1701,]	0.1	0.1	0.00	0.47
19	[1801,]	0.1	0.1	0.00	0.47
20	[1901,]	0.1	0.1	0.00	0.47
6	[561,]	0.7	0.1	2.22	1.42
7	[661,]	0.7	0.1	2.22	1.42
5	[401,]	0.1	0.1	10.00	2.38
4	[361,]	0.7	0.1	15.55	3.33
3	[241,]	0.5	0.1	23.33	5.23
2	[101,]	0.1	0.1	34.44	11.42
1	[41,]	0.5	0.1	34.44	15.23

Estas son muy buenas tasas, al aplicarlas sobre el resto de la base (700 registros), se obtienen las siguientes tasas:

$clasificación$	$bien$	mal
$bien$	88,5 %	11,4 %
mal	50 %	50 %.

La gráfica ROC de la red neuronal test se aprecia a la izquierda en la figura 3.7, mientras que la comparación de esta estimación con la estimación obtenida del modelo de redes en el total de la población y la estimación obtenida mediante la regresión logística en el submodelo 2.5 se encuentra a la derecha de la misma figura, de esta manera se aprecia al parecer la mejor estimación de los dos modelos junto con el modelo saturado de regresión logística.

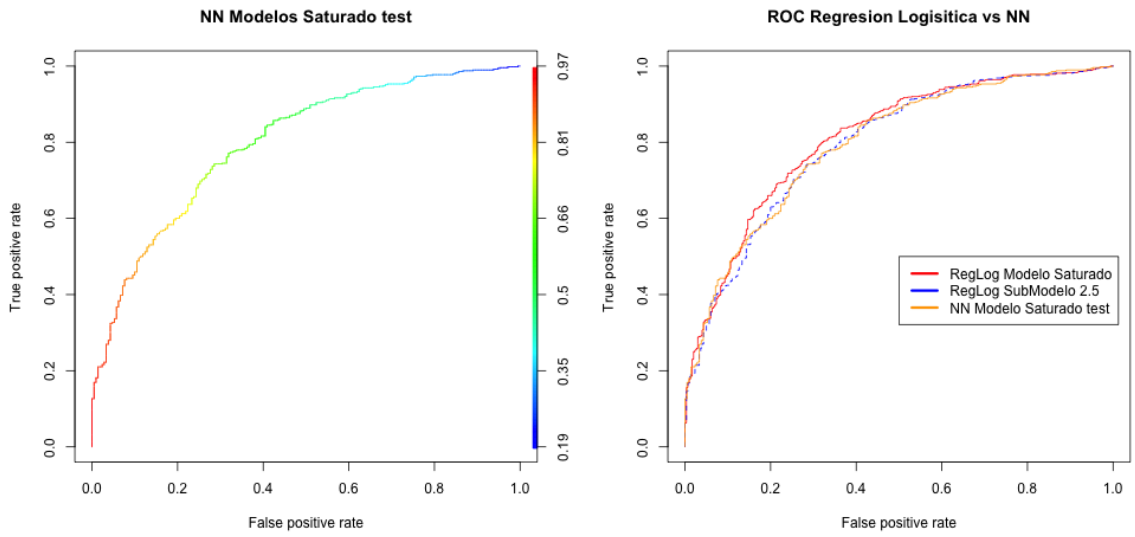

Figura 3.7: Del lado izquierdo: gráfica ROC de una muestra aleatoria de un 30 % de la base que se utiliza para estimar los parámetros. Del lado derecho: gráfica ROC del modelo saturado y el submodelo 2.5 de regresión logística así como NN del modelo saturado con el 70 % de los datos, se grafica la primera capa para comparar con las curvas de la regresión logística.

3.3.3. Conclusiones del modelo

La información de los primeros tres modelos de redes se puede contrastar nuevamente mediante las curvas ROC, de tal manera que en la figura 3.8 se observan las capas 1, 5, 10, 15 y 20 con sus menores tasas de error para los modelos NN 1 (rojo), NN 2 (azul) y NN 3 (naranja). En este gráfico se aprecia que las curvas se aproximan con mayor velocidad hacia la esquina superior izquierda con el modelo NN 1, notándose un salto grande de la capa 5 a la 10, le sigue el modelo NN 2 y finalmente el modelo NN 3 es más moderado.

Es interesante observar que al contrastar la curva del modelo generado con base en las variables del modelo 2 saturado con la curva del conjunto de variables provenientes del submodelo 2.1 y 2.5 del modelo de regresión logística, el modelo saturado presenta tasas de clasificación errónea competitivas con respecto a un conjunto de variables previamente seleccionado mediante otro modelo.

Por otro lado, se presenta el cuadro $C.19$ que contiene, por modelo, la capa con menores tasas de clasificación errónea, donde se considera permisible que esta tasa no exceda del 15 % (este porcentaje variará en cada institución financiera de acuerdo a sus políticas de riesgos); lo anterior debido a que las mayores tasas de clasificación errónea se encuentran en la tasa de clientes con capacidad crediticia clasificados como clientes sin capacidad crediticia. De esta manera es posible seleccionar un

submodelo por cada modelo de NN aplicado, *i.e.*, para el Modelo NN 1 procedente del modelo 2 saturado de regresión logística, la capa 12 presenta una tasa de clientes con capacidad crediticia clasificados como clientes sin capacidad crediticia de 14 % mientras que la tasa de clientes sin capacidad crediticia clasificados como clientes con capacidad crediticia de 2.86 %, para el Modelo NN 2 procedente del submodelo 2.1 de regresión logística las tasa son 15 % y 4.43 %, respectivamente, se sigue la misma idea para los modelos NN 3 y NN 1 *test*.

Cuadro *C*.19. Matriz de salidas de los cuatro modelos de red neuronales

Modelo de Red Neuronal	***Modelo de regresión log. procedente***	***Capa***	**Fila**	***Rang***	***Decay***	**m⟶b** (%)	**b⟶m** (%)
NN 1	Modelo 2	12	[1171,]	0.8	0.1	14.00	2.86
NN 2	Submodelo 2.1	8	[751,]	0.6	0.1	15.00	4.43
NN 3	Submodelo 2.5	9	[861,]	0.7	0.1	15.00	3.71
NN 1 *test*	**Modelo 2**	**8**	**[761,]**	**0.7**	**0.1**	**0.00**	**0.47**

Finalmente, se busca que el modelo seleccionado tenga el menor número de capas y las menores tasas de clasificación errónea, por lo que el modelo NN 1 *test* resulta el mejor modelo con las menores tasas de clasificación errónea (marcado en el cuadro con negritas), seguido del modelo NN 2.

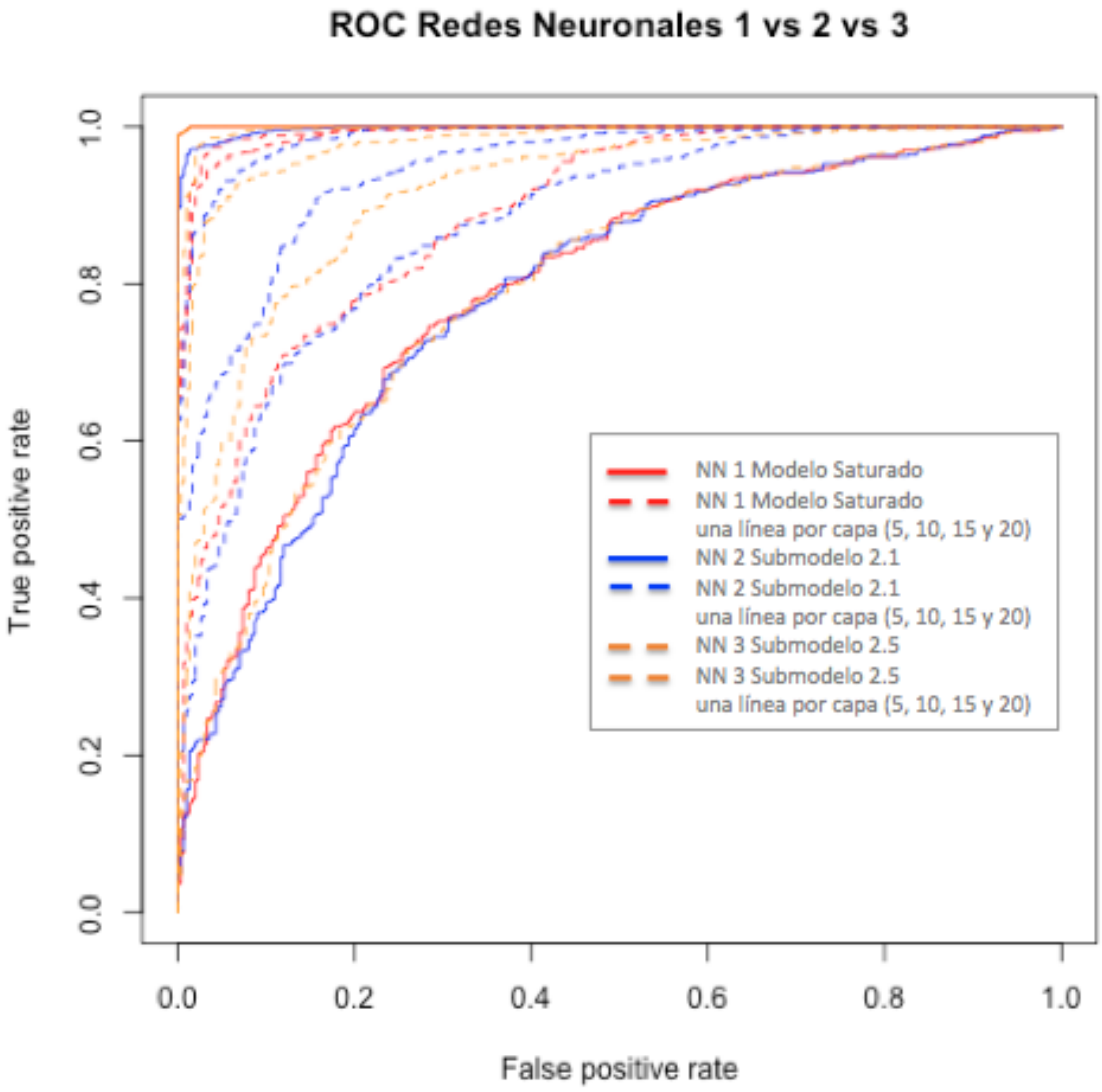

Figura 3.8: Curvas ROC de los modelos de NN 1, 2 y 3 con base en las variables del modelo 2 saturado (23 variables, color rojo), Submodelo 2.1 (12 variables, color azul) y submodelo 2.5 (15 variables, color naranja), respectivamente. Cada línea expresa las menores tasas de error presentes en las capas 1, 5, 10, 15 y 20 de cada uno de los modelos mencionados, con base en los Cuadros $C15$, $C.16$ y $C.17$.

3.4. Conclusiones del capítulo

En la figura 3.9 se aprecian las curvas ROC de redes neuronales 1, 2 y 3 con base en las variables del modelo 2 saturado (que contiene 3 variables continuas y 20 binarias), submodelo 2.1 (que contiene 2 variables continuas y 10 binarias) y submodelo 2.5 (que contiene 3 variables continuas y 12 binarias), respectivamente, y se grafican las menores tasas de error presentes en las capas 1, 5 y 10 así como las curvas ROC de regresión logística de los modelos: modelo 2 saturado, submodelo 2.1 y submodelo 2.5. Con base en esta información y la obtenida en los desarrollos se aprecia que aunque las tasas de clasificación de los modelos de redes neuronales son muy buenas al compararlas con las del modelos de regresión logística se pierde interpretabilidad y se tiene mayor propensión a sobreajustar el modelo que en un modelo de regresión logística.

Sin embargo, al introducir conjuntos de variables previamente seleccionados mediante el modelo de regresión logística se presenta un mejor ajuste y éste se obtiene al considerar una red con al menos cinco capas en su desarrollo. Esta idea va de la mano con los comentarios de Hastie et al. (2009), por lo cual es un buen argumento para considerar como finales los resultados de estos modelos presentes en el cuadro $C.19$.

Cabe mencionar que a pesar de la poca interpretabilidad del modelo de redes neuronales es una herramienta útil para aproximar una función que discrimine a los datos debido a la recursividad de las funciones aplicadas, sin embargo, se debe tener en cuenta que las redes neuronales no tienen como característica ser un modelo que permite evaluar la contribución de las variables en la clasificación de las observaciones, es por ello que se utilizó el conjunto de variables obtenidas previamente de los modelos de regresión logística con mejores valores de AIC y menores tasas de clasificación errónea.

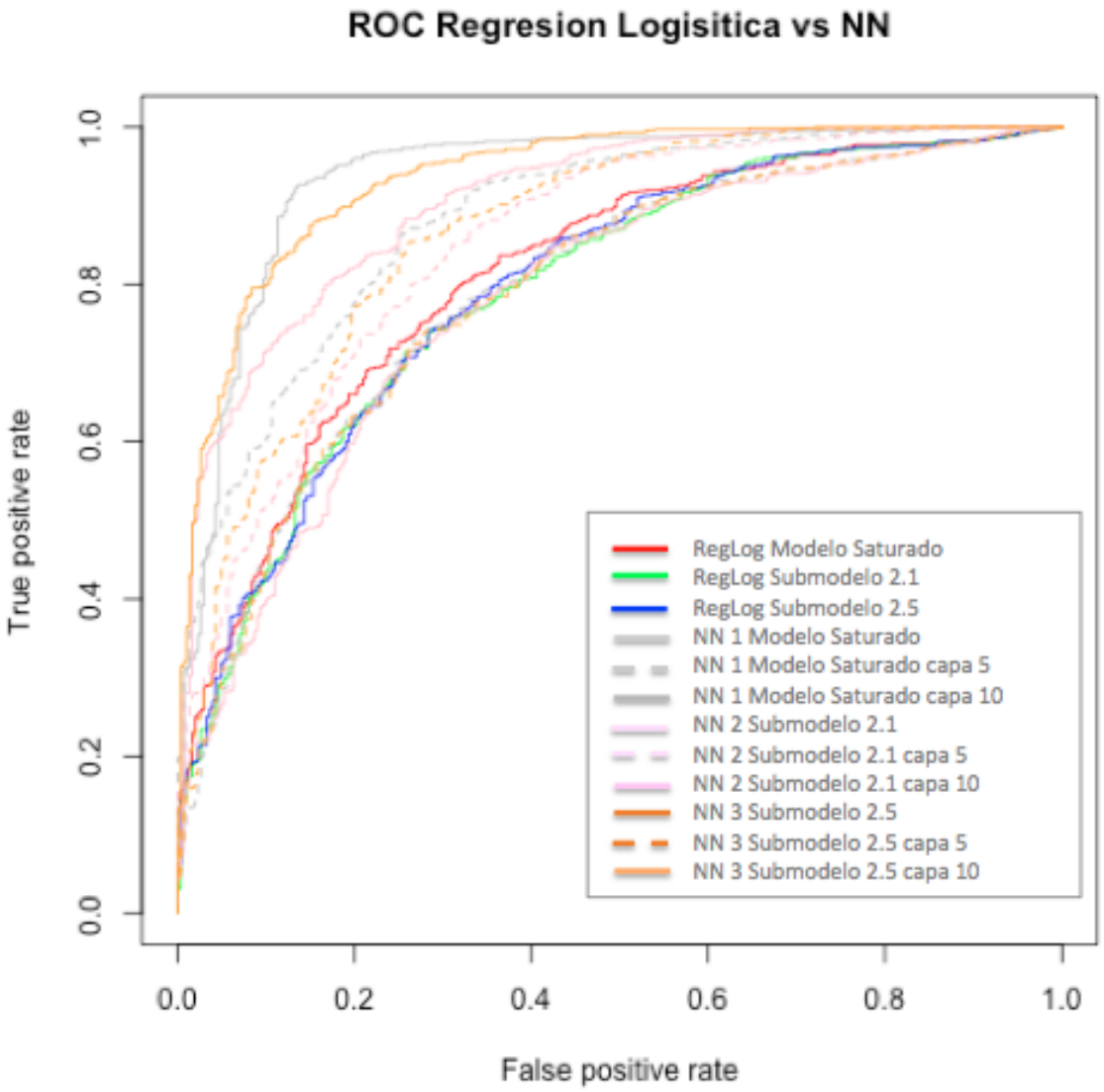

Figura 3.9: Curvas ROC de los modelos de NN 1, 2 y 3 con base en las variables del modelo 2 saturado (23 variables, color rojo), Submodelo 2.1 (12 variables, color azul) y submodelo 2.5 (15 variables, color naranja), respectivamente, así como las curvas de los modelos de regresión logística antes mencionados. Cada línea expresa las menores tasas de error presentes en las capas 1, 5, 10, 15 y 20 de cada uno de los modelos de NN, con base en los Cuadros $C15$, $C.16$ y $C.17$.

Conclusión y discusión

Conclusión

En la presente tesis se han implementado dos modelos: el modelo de regresión logística y el modelos de redes neuronales. Estos modelos parten de la variabilidad de modelos estadísticos existentes para la mejor compresión y clasificación de un conjunto de observaciones, los cuales permite dejar de lado posibles suposiciones que se puedan generar de los atributos de una base al observarla rápidamente. Lo anterior mediante el análisis de las observaciones para la toma de decisiones orientadas.

En el caso aquí presente, sobre la muestra de datos previamente clasificada (buen acreedor o mal acreedor) se realizó un análisis exploratorio de datos y reconocimiento de proyecciones interesantes (mediante la técnica de *projection pursuit*), para definir los puntos de corte en la población. Adicionalmente se implementó el análisis de componentes principales para averiguar si era posible disminuir el número de variables que describieran mejor a la población e introducirlas en los próximos modelos, o bien, reemplazar las variables por las componentes que acumularan al menos el 70 % de la variabilidad de la base, esto debido a que mientras mayor sea el conocimiento de una población mejor será la selección del modelo de clasificación a aplicar. A su vez, este tipo de análisis permite reducir el conjunto de variables y colapsar algunas categorías para trabajar, abriendo paso a un conjunto más reducido y preciso.

Con ayuda del análisis exploratorio previo se decidió utilizar 10 variables recodificadas en el modelo de regresión logística para encontrar un modelo ajustado que permitiera clasificar futuras observaciones de solicitantes a crédito (que tuvieran un perfil similar a clientes con capacidad crediticia). Para ello se partió de un modelo saturado con 23 variables (3 continuas y 7 categóricas, en total 10 variables, que al generar variables categóricas {0,1} se obtiene 20 covariables binarias) y mediante tres procedimientos distintos de selección de variables: *backward, stepwise* y *AIC* se seleccionó aquel modelo con menores tasas de clasificación errónea.

Como siguiente paso, y usando nuevamente las 10 variables previamente seleccionadas en la recodificación, se ejemplificó el modelo de redes neuronales que ayudó a afinar el modelo ajustado, sin llegar a sobreajustar, para lo cual se tomaron los tres modelos de regresión logística con menores tasas de clasificación errónea, ya

que el conjunto de variables de cada uno de ellos se utilizaron para generar modelos de redes neuronales. Es de esta manera que es posible integrar y retroalimentar un modelo de los resultados de otro modelo para obtener mayor interpretabilidad, en especial al tratarse con una red neuronal.

Finalmente, para poder analizar los distintos resultados que arrojó cada modelo se utilizaron curvas ROC para la discriminación de éstos, así como la comparación de tasas de clasificación errónea.

Realizados los análisis anteriores se puede concluir que el mayor tiempo empleado al análisis se utilizó para la limpieza y reconocimiento de la base de datos, al analizar variable por variable y variable contra variable para la disminución de las mismas, e incluso de categorías para su entrada en los modelos, así como el poder colapsar categorías similares o complementarias. En este caso, la ejecución del modelo de componentes principales generó una reducción a 12 componentes, sin embargo no provocó un impacto al introducir estas 12 componentes en los modelos de regresión logística y redes neuronales, por lo que se decidió seguir utilizando las variables originales reestructuradas para mantener la interpretabilidad de los resultados obtenidos.

Por otro lado, las tasas de clasificación errónea obtenidas en los distintos submodelos de regresión logística resultaron mayores a las tasas del modelo 2 saturado. Sin embargo, al aplicar a este conjunto de variables el modelo de redes neuronales se obtiene una mejor clasificación de las observaciones, sin esperar sobreajuste, de la capa 5 a la 10, siendo éstas las menores tasas de clasificación errónea obtenidas.

Es importante aterrizar los resultados de todos los modelos aplicados, así como el aprendizaje obtenido de estos desarrollos mediante procesos y acciones reales y prácticas para que retroalimenten al negocio. Para lograr lo anterior se hace referencia al modelo de *Credit Scoring*, el cual se basa en la retroalimentación del esfuerzo realizado para resumirlo en ponderadores de diversas situaciones o características del cliente, resultando de esta manera un procedimiento recursivo que va desde el análisis hasta la ejecución de las decisiones.

Hoy en día el mercado de créditos se ha vuelto variado ya que se cuenta con diversos productos de crédito específicos (hipotecario, automotriz, de nómina, personal, etc.), por lo que resulta relativamente más sencillo adquirir algún bien o servicio. Por ello se puede hablar entonces de la influencia del modelo de *Credit Scoring* dentro del ciclo económico, pues la automatización de las decisiones y velocidad con que se lleva a cabo dicho mecanismo ha acelerado la economía del país, aunque se debe aclarar que no en todos los sentidos ha sido bueno. Claro ejemplo es la más reciente crisis económica internacional generada por la desmesurada concesión de créditos hipotecarios para aprender de ella.

Discusión

En esta tesis se presentaron principalmente dos problemas, dada una base de 1,000 clientes con crédito, donde previamente se tenían identificados aquellos clientes con capacidad crediticia de los clientes sin capacidad crediticia. Primero se debía identificar el perfil de ambos grupos de clientes y como segundo paso, y siendo el objetivo de esta tesis, se buscó obtener modelos de predicción, clasificación y discriminación que se ajusten adecuadamente a la base de estudio con el objetivo de que al tomarse una muestra futura dichos modelos predigan a qué solicitantes se les debe otorgar o no el crédito, con base a diversas características que los describen.

Dicho lo anterior, primero se identificó el perfil de clientes con capacidad crediticia definido por las principales características de acuerdo a diversos factores como son el objetivo del crédito solicitado, el tiempo y monto del mismo, características sociodemográficas, capacidad de pago, entre otras. De tal forma que, según las políticas del la institución financiera, el perfil de un cliente con capacidad crediticia se definió como clientes preferentemente de género masculino, edades referentes a un adulto joven, con estado civil soltero o viudo, estable en su trabajo y hogar, por tener permanencia en estos mayor a dos años, sin deudas actualmente, con una capacidad de pago ajustada y preferentemente con referencias crediticias dentro de la misma institución.

El principal problema presente en esta sección fue el manejo de un gran conjunto de variables que a su vez contenían bastantes categorías, de tal forma que se llegó a tener una base de más de 90 variables generando como objetivo a corto plazo el reducir el número de las mismas y a su vez, de las variables que participarían en los modelos subsecuentes. De tal forma que mediante estadísticos descriptivos, análisis exploratorio de datos, *projection pursuit* y el análisis de componentes principales se obtuvieron proyecciones interesantes que permitieron seleccionar las variables de mayor peso para definir a cada grupo, así como colapsar categorías similares o de igual comportamiento al observarlas con el resto de variables. En resumen, de 23 variables (de las cuales 3 eran continuas y de esas mismas 3 se presentaba las covariables categóricas) que sumaban 94 categorías en la base, se decidió trabajar con 10 variables (10 categóricas, 3 de las cuales también se presentan continuas, en total 13), donde de las 10 variables categóricas se generan 60 categorías y al colapsar categorías, con bases en los procedimientos ya mencionados, se obtuvieron 44 categorías; a su vez al transformar todas a variables binarias y dejando una categoría de referencia por variable se obtuvieron 34 variables binarias que explican la base de datos. El resumen de este desarrollo se encuentra en la "Tabla de transformación de variables" a utilizar de la *base DM*, ubicada en las páginas 27 y 28 del Capítulo 2.

Una de las razones por las cuales se decidió emplear el modelo de regresión logística para la predicción y clasificación de datos fue que la variable de respuesta, *kredit* o

Crédito, es binaria. De esta forma el siguiente problema consistió en seleccionar el mejor modelo saturado con base en las variables continuas o categóricas, es decir, se buscaron las mejores tasas de clasificación errónea y correcta de cuatro modelos saturados en los cuales se mezclaban las mismas 10 variables preseleccionadas, pero considerando sólo variables categóricas, únicamente variables continuas o la mezcla de ambas, obteniendo de esta forma que el modelo saturado con variables continuas y binarias presentaba, de entre los cuatro modelos, las mejores tasas. Sin embargo, las tasas de clasificación aún eran muy grandes, ya que las tasas de clasificación errónea eran mayores o iguales al 50 %, lo que indica que aplicar el modelo de clasificación y lanzar una moneda resultaba igual de probable.

Posteriormente, se buscó un modelo más complejo, o bien, uno con mejores tasas de clasificación mediante tres métodos de selección de variables: *backward, stepwise* y *AIC* y se obtuvo, mediante el método AIC, el submodelo 2.5 con 15 variables (3 continuas y 12 binarias). Este modelo arrojó una tasa global de clasificación correcta del 91 %, que resulta la mejor de todos los modelos desarrollados, sin embargo con una tasa de clientes sin capacidad crediticia clasificados como clientes con capacidad crediticia del 52 %, lo que hace notar que la tasa global de clasificación correcta no es un indicador absoluto por sí sólo. Dicho lo anterior se retoma la idea de que el mejor modelo presente en esta sección resultó del modelo 2 saturado que presenta una *tgdc* del 90.7 % y tasas de clasificación errónea de clientes sin capacidad crediticia clasificados como clientes con capacidad crediticia del 51 % y de clientes con capacidad crediticia clasificados como clientes sin capacidad crediticia del 49 %.

Con el objetivo de ejemplificar el modelo de redes neuronales, obtener un mejor ajuste con la búsqueda de menores tasas de clasificación errónea, y ya que este modelo también va de la mano con el modelo de regresión logística empleado anteriormente, se introdujeron al modelo de NN las variables del modelo 2 saturado (que contiene 3 variables continuas y 20 binarias), las variables del modelo obtenido mediante el método de selección AIC y un tercer modelo que tuvo buenas tasas de clasificación (90 % como tasa global de clasificación correcta y 55 % de clientes sin capacidad crediticia clasificados como clientes con capacidad crediticia), el cual se obtuvo mediante la selección de variables por eliminación sobre el *p-valor* (*backward*). El problema enfrentado con el modelo de redes neuronales, es que al ser como una "caja negra", sólo arrojó (como se esperaba) las tasas de clasificación errónea de una red con cierto número de capas ocultas, cierto valor del parámetro *decay*, cierto valor del parámetro *rango* y bajo cierto número de interaciones preestablecidas, y para seleccionar la red con mejor predicción se debía de correr manualmente cada posible red. De tal manera que se programaron mediante el *software R* varios ciclos anidados que movían estos parámetros en un rango de cero a uno incrementando cada parámetro de 0.1 en 0.1, el único parámetro fijo fue el número de iteraciones permitidas que se estableció en máximo 500, de esta manera se obtuvieron 2,000 iteraciones en un lapso de 40 minutos. Posteriormente, de cada capa se extrajo la

combinación de parámetros con mejores tasas de mal clasificación, las más bajas, y se graficaron con ayuda de las Curvas ROC.

En este desarrollo se comprobó la idea de Hastie et al. (2009) que dice que empleando este modelo a partir de la quinta capa oculta se obtienen las mejores estimaciones, ya que para los tres modelos desarrollados a partir de esta capa se obtienen tasas de clasificación errónea de 37 %, 32 % y 29 %, respectivamente, obtenidas con el modelo de regresión logística saturado, con el modelo de selección de variables mediante el método *backward* y el método AIC, las cuales son mucho mejores que las tasas del 51 %, 52 % y 55 %, respectivamente, para cada modelo mencionado. Inclusive, es posible marcar una cota superior en las tasas de clasificación errónea de acuerdo a las políticas de riesgos de cada entidad financiera para poder seleccionar una capa más arriba, de tal forma que colocando esta cota en un 15 % para tasas de clientes con capacidad crediticia clasificados como clientes sin capacidad crediticia se seleccionan las capas 12, 9 y 8, respectivamente.

Dado que se introdujeron al modelo de redes neuronales conjuntos de variables previamente seleccionadas y analizadas, no resulta tanto una "caja negra". Sin embargo, este modelo, a diferencia del modelo de regresión logística, no permite conocer la ponderación que se le dio a cada característica como para poder aplicar un *score* final a cada variable y así tener un modelo más puntual y poder generalizarlo, esto principalmente al momento de ejecutarlo dentro de los sistemas de una institución financiera, ya que para otras áreas sería poco clara dicha aplicación.

Finalmente, y con ayuda de las Curvas ROC, así como las tablas y cuadros que indican las mejores tasas de clasificación errónea, se puede concluir que efectivamente el mayor tiempo absorbido en este tipo de análisis se emplea en la primera etapa de preselección de variables, colapso de categorías y selección de modelos, además de que el modelar no es sencillo y conlleva a su vez un buen conocimiento de las necesidades de cada modelo, sus limitantes y puntos a favor, así como un buen conocimiento de la base a tratar y de un claro objetivo de análisis.

Es también relevante mencionar el aprendizaje obtenido de ejecutar estas técnicas y modelos en distintos *softwares* como son SAS y principalmente R, así como de conocer también sus limitantes y ventajas. A lo largo de este trabajo se implementó en ambos *softwares* los distintos modelos y por la familiaridad que se tiene con el *software R*, que maneja un lenguaje en parte orientado a objetos, se desarrollaron algoritmos con iteraciones anidadas para un aprendizaje más completo en la sección práctica de redes neuronales y poder generar 2,000 salidas de predicción en un corto periodo.

Anexo A: Códigos implementados en el trabajo

.1. *Projection pursuit*

```
library('rggobi')
library('MASS')
library('klaR')

>gb<-ggobi(datos_recofid)[1]
```

.2. Componentes principales

```
> round(cor(datos_recofid),digits=3)
> round(sd(datos_recofid),digits=3)

> PCA3=princomp(datos.pca,cor=F)
> summary(PCA3)

> round(PCA3$sdev,digits=3)
```

Generación de gráficas de componentes 1, 2, 3, 32, 33 y 34:

```
plot(PCA3\$scores[,1:2],col=color)
plot(PCA3\$scores[,3:4],col=color)
plot(PCA3\$scores[,33:34],col=color)
plot(PCA3\$scores[,31:32],col=color)
```

.3. Regresión logística

```
 hosmerlem <- function (y,yhat,g=10)
```

```
{
cutyhat <- cut(yhat,breaks = quantile(yhat,probs = seq(0,1,1/g)), include.lowest = T)
obs <- xtabs(cbind(1-y,y)~cutyhat)
expect <- xtabs(cbind(1-yhat,yhat)~cutyhat)
chisq <- sum((obs-expect)^2/expect)
P <- 1 - pchisq(chisq, g-2)
c("X^2" = chisq, Df = g-2, "P(>Chi)" = P)
}

 step(x1)
 c1 <- glm(kredit ~ laufkont_Dc1+laufkont_Dc2+laufzeit +
    moral_Dc2+moral_Dc3+verw_Dc2+verw_Dc3 + verw_Dc4 +
    verw_Dc7+hoehe+rate_Dc1 + rate_Dc2 + famges_Dc3 + alter +
    weitkred_Dc3, data = datos, family = binomial(link = "logit"))
summary(c1)

prediccion.scores.c1 <- predict(c1,type="response")
prediccion.c1 <- round(prediccion.scores.c1,0)

tab.c1 <- table(datos$kredit,prediccion.c1)
tab.c1/rowSums(tab.c1)
tab.c1

hosmerlem(y=datos$kredit,yhat=fitted(h1))
lrtest(h1)
dchisq(262.75,32)*2

pred.c1 <- prediction(prediccion.scores.c1, datos$kredit)
perf.c1 <- performance(pred.c1, "tpr", "fpr")
plot(perf.c1, colorize=T)
plot(perf.c1,print.cutoffs.at=seq(0,1,by=0.1), colorize=T)
```

.4. Redes neuronales

```
 set.seed(1)

datos.clasifica=datos[,1]
datos.clasifica.grupos=factor(datos.clasifica,labels=c('m','b'))
datosSaturados3=datos[,-c(1,15,16,3,22,23,24,25,26,29,7,36,37,40,11,46)]
datosNett3 <- data.frame(datosSaturados3,datos.clasifica.grupos)
```

```
 set.seed(1)
Red2.1 <- nnet(datos.clasifica.grupos ~ ., data = datosNett3,
size = 1, rang = 0.2,decay = 0.1, maxit = 500)
Red2.5 <- nnet(datos.clasifica.grupos ~ ., data = datosNett3,
size = 5, rang = 0.5,decay = 0.1, maxit = 500)
Red2.10 <- nnet(datos.clasifica.grupos ~ ., data = datosNett3,
size = 10, rang = 0.2,decay = 0.1, maxit = 500)
Red2.15 <- nnet(datos.clasifica.grupos ~ ., data = datosNett3,
size = 15, rang = 0.7,decay = 0.1, maxit = 500)
Red2.20 <- nnet(datos.clasifica.grupos ~ ., data = datosNett3,
size = 20, rang = 0.9,decay = 0.1, maxit = 500)

prediccion.scores2.1 <- predict(Red2.1,type="response")
f2.1<-Red2.1$fitted.values
pred2.1 <- prediction(f2.1, datos$kredit)
perf2.1 <- performance(pred2.1, "tpr", "fpr")
plot(perf2.1, colorize=T)
```

Anexo B: Nociones básicas

En esta sección se busca dar un panorama básico de los conceptos que se deben tener claros para abordar y comprender la teoría de cada capítulo, para lo cual se desarrollarán un par de ejemplos mientras se plantean conceptos de relevancia.

Para este análisis se cuenta con una población P donde cada observación o solicitante tiene asignado un vector $p-$dimensional de variables:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

de tal forma que el $j-$ésimo elemento del vector representa una característica de la observación y a su vez se denota al vector transpuesto de $\mathbf{x}$ como $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$, dicho lo anterior se puede plantear la siguiente matriz denotando cada elemento por $x_{i,j}$ tal que el $i-$ésimo solicitante presenta la $j-$ésima característica es decir:

$$\mathbf{P} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{10001} & x_{10002} & \ldots & x_{1000p} \end{pmatrix}$$

Se ha acotado a 1,000 observaciones pues es el número de observaciones presentes en la base de datos con la que se trabajará, tal que se define N =1,000 ahora, se definen los estadísticos básicos.

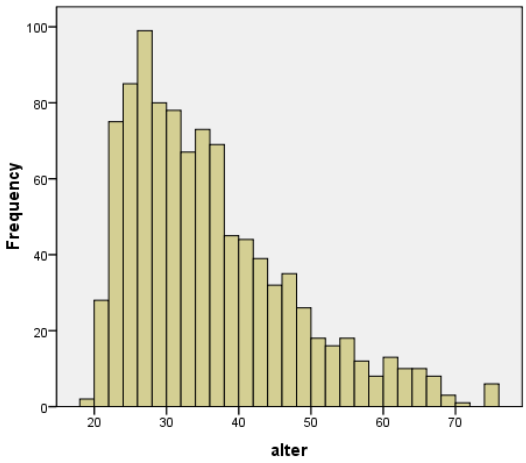

Figura 10: Histograma de la variable *alter* que denota la edad en nuestra base de datos donde la media se encuentra en los 28 años.

Suponga que se busca conocer la edad promedio de los $1,000$ aspirantes a crédito tal que se define la **media muestral** como:

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \tag{29}$$

$$\Rightarrow \bar{x}_{edad} = \frac{1}{1,000}\sum_{i=1}^{n} x_{i,edad} = 28$$

Nótese en la figura 10 que un histograma es una excelente herramienta para mostrar las frecuencias de las clases de una variable, en este caso las mayores frecuencias se encuentras entre los 26 y 28 años. De hecho, se nota que la mayor parte de la población se encuentra entre los 22 y 38 años.

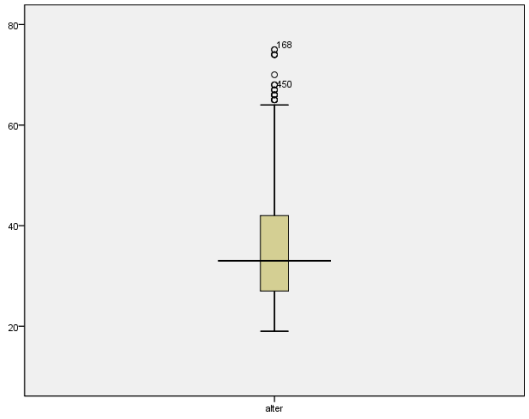

Figura 11: Diagrama de caja de la variable *alter* que denota la edad en nuestra base de datos donde se nota gran dispersión en los datos, pero cierta concentración entre los 25 y 42 años.

Por otro lado, una medida de dispersión básica, la **varianza muestral**, da a conocer qué tan dispersos o distantes se encuentran las observaciones de la media muestral tal que:

$$s_j^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2 \tag{30}$$

por lo que para conocer qué tan dispersas son las edades en la muestra se desarrolla:

$$\Rightarrow s_{edad}^2 = \frac{1}{1,000}\sum_{i=1}^{n}(x_{i,edad} - \bar{x}_{edad})^2$$

$$\Rightarrow s_{edad}^2 = 128{,}88$$

En la figura 11 se observa un diagrama de caja donde muestra en sus "bigotes" los valores mínimos y máximos; en la caja, la concentración de la mayoría de la dispersión; al inicio y fin de la caja, el primer y tercer cuartil mientras que la línea que intersecta la caja representa la media de los datos, o bien, el segundo cuartil.

Ahora se define la **covarianza muestral** considerada una medida de asociación entre la $i-$ésima y la $k-$ésima variables como:

$$\begin{gathered} s_{jk}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k) \\ i,k = 1,2,\ldots,p \end{gathered} \tag{31}$$

de tal forma que cuando $j = k$ la expresión anterior se reduce a la covarianza muestral. Finalmente, la última estadística descriptiva que se considerará será el **coeficiente de correlación muestral** también conocido como **coeficiente de correlación de Pearson**, el cual es una medida de asociación lineal entre dos variables que no dependen de la unidad de medición. El coeficiente de correlación, para la $i-$ésima y la $k-$ésima variables, es definido como sigue:

$$\begin{aligned} r_{ik} = & \quad \frac{s_{ik}^2}{\sqrt{s_i^2}\sqrt{s_k^2}} \\ r_{ik} = & \quad \frac{\sum_{j=1} N(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)}{\sqrt{\sum_{j=1} N(x_{ji}-\bar{x}_i)^2}\sqrt{\sum_{j=1} N(x_{jk}-\bar{x}_k)^2}} \\ para & \quad i = 1,2,\ldots,p; \\ & \quad k = 1,2,\ldots,p \end{aligned} \tag{32}$$

La covarianza y correlación proporcionan medidas de asociación lineal, sin embargo, ambas estadísticas son sensibles a valores inusuales, o bien, valores atípicos que en su mayoría corresponden a errores al tomar la muestra o población y que usualmente generan "ruido" en los resultados, a estos valores normalmente se les conoce como ***outliers*** y en general lo más recomendable es extraerlos del resto de las observaciones para no sesgar los resultados.

La notación asignada con anterioridad responde a 1 variable con p observaciones, si ahora se piensa en manipular n medidas con p variables, que es como se trabajará en adelante, resulta mejor organizar la información en arreglos o matrices de tal forma que se definen la siguiente notación:

- Dependencia linear, un conjunto de vectores $\mathbf{x_1}, \ldots, \mathbf{x_p}$ es linealmente dependiente si existen escalares $c_1, \ldots, c_p$, no todos nulos, tales que: $c_1\mathbf{x_1} + \ldots + \mathbf{c_p x_p} = \vec{\mathbf{0}}$, donde $\vec{\mathbf{0}}$ representa un vector nulo que tiene todas sus componentes iguales a cero. Si un conjunto de vectores no es linealmente dependiente se dirá que los vectores son linealmente independientes.

- Se llama matriz $\mathbf{A}$ de dimensiones $(n \times p)$ a un conjunto de nxp números reales, ordenados en n filas y p columnas y se llama matriz traspuesta $\mathbf{A'}$ a la matriz obtenida a partir de $\mathbf{A}$ intercambiando filas por columnas.

- Matriz identidad, $\mathbf{I_n}$, de dimensión n, como la matriz de dimensiones $n\ x\ n$ que tiene unos en la diagonal y cero fuera de ella.

- Matriz cuadrada, se dice que una matriz es cuadrada si $n = p$, y este número se denomina orden de la matriz.

- Matriz simétrica, dentro de las matrices cuadradas se llaman simétricas a las que tienen cada fila igual a la correspondiente columna, es decir $a_{ij} = a_{ji}$. Una matriz simétrica es, por lo tanto, idéntica a su transpuesta, de esta forma se dirá que $\mathbf{A}$ es simétrica si $\mathbf{A'} = \mathbf{A}$.

- Matriz diagonal, es una matriz cuadrada y simétrica que tiene únicamente términos no nulos en la diagonal principal. Un caso importante de una matriz diagonal es la matriz identidad o unidad $\mathbf{I}$. En particular, los productos $\mathbf{AA'}$ y $\mathbf{A'A}$ conducen a matrices simétricas.

Ahora bien, se definen las estadísticas básicas con base en los conceptos antes mencionados como:

El vector de media muestral:

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}.$$

La matriz de covarianza muestral:

$$\mathbf{S} = \begin{bmatrix} s_{11}^2 & s_{12}^2 & \ldots & s_{1p}^2 \\ s_{21}^2 & s_{22}^2 & \ldots & s_{2p}^2 \\ \vdots & \vdots & & \vdots \\ s_{p1}^2 & s_{p2}^2 & \ldots & s_{pp}^2 \end{bmatrix}.$$

La matriz de correlación muestral:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}.$$

Finalmente, se pueden simplificar las expresiones anteriores como sigue:

$$\bar{X} = \frac{1}{N}\mathbf{X'}\vec{\mathbf{1}} \tag{33}$$

$$\mathbf{S} = \frac{1}{N}\Big(\mathbf{X'}\ \mathbf{X} - \frac{1}{N}\mathbf{X'}\vec{1}\vec{1}'\ \mathbf{X}\Big) = \frac{1}{N}\mathbf{X'X}\ \text{-}\ \bar{X}\bar{X}' \tag{34}$$

donde $\vec{\mathbf{1}}$ representa un vector columna de N unos.

Anexo C: Variables de la *base DM*

A continuación se presenta la descripción de las 21 variables presentes en la *base DM* donde sólo 3 de ellas son de tipo continuo. La base cuenta con categorías preestablecidas, siendo así que en la columna *Nombre* se encontrará el nombre de la variable en su idioma origen (alemán), en la columna *Nombre de trabajo* se encuentra el nombre que se ha asignado a cada variable para identificarla con mayor facilidad en el momento de análisis, en la columna *Descripción* el significado de dicha variable, en la columna *Valor* las diversas categorías que posee y en la columna *Categorías* se da la descripción de cada categoría; las últimas dos columnas se refieren a la Frecuencia Relativa calculada tanto para los buenos créditos, o bien clientes con capacidad crediticia, como para los malos créditos, o bien clientes sin capacidad crediticia. Para el caso las variables continuas se ha agregado una variable categórica por cada una de ellas. Finalmente, con el símbolo $\star$ se señalan las variables que se utilizaron en el desarrollo de los distintos modelos. Se debe de tener en cuenta que las variables aquí mostradas se encuentran en su forma original y que en las páginas 27 y 28 del Capítulo 2 se puede encontrar una tabla resumen con la recodificación de las variables a utilizar en los modelos.

No.	**Nombre en base**	**Nombre en texto**	**Descripción**	**Valor**	**Categorías**	**Frec. Relativa en % de: Créditos Buenos**	**Créditos Malos**
1	kredit$\star$	Crédito	Capacidad crediticia del cliente	0	Con capacidad	0.00	30.00
				1	Sin capacidad	70.00	0.00
2	laufkont$\star$	Saldos	Saldos de cuentas en el banco DM	1	Sin cuenta Vigente	45.00	19.86
				2	Sin saldo	35.00	23.43
				3	$0 \geq \ldots \leq 200$ DM o cuenta de cheques con al menos un año	4.67	7.00
				4	≥ 200 DM o más	15.33	49.71
3	laufzeit$\star$	Duración	Duración en meses del crédito solicitado (continua)				

No.	Nombre en base	Nombre en texto	Descripción	Valor	Categorías	Frec. Relativa en % de: Créditos Buenos	Créditos Malos
				1	> 54	2,33	1,00
3b	dlaufzeit⋆	Duración	Duración en	2	$48 < \ldots \leq 54$	0.33	0.14
			meses	3	$42 < \ldots \leq 48$	10.67	3.14
			del crédito	4	$36 < \ldots \leq 42$	1.67	1.71
			solicitado	5	$30 < \ldots \leq 36$	12.67	6.86
			(categórica)	6	$24 < \ldots \leq 30$	6.33	5.43
				7	$18 < \ldots \leq 24$	22.00	22.57
				8	$12 < \ldots \leq 18$	18.67	18.71
				9	$6 < \ldots \leq 12$	22.33	30.00
				10	≤ 6	3.00	10.43
				0	Atraso en pagos de créditos previos	8.33	2.14
4	moral⋆	Morosidad	Pago de créditos previos	1	Problemas en créditos vigentes/Con créditos vigentes en otros Bancos	9.33	3.00
				2	Sin créditos previos/ Créditos previos saldados	56.33	51.57
				3	Sin problemas con créditos vigentes en Banco DM	9.33	8.57
				4	Saldados todos los créditos previos en el Banco DM	16.67	34.71
				0	Otros Créditos	29.67	20.71
5	verw⋆	Propósito	Propósito	1	Carro nuevo	5.67	12.29
			del	2	Carro usado	19.33	17.57
			crédito	3	Muebles	20.67	31.14
				4	Radio/Televisión	1.33	1.14
				5	Electrodomésticos	2.67	2.00
				6	Reparaciones	7.33	4.00
				7	Educación	0.00	0.00
				8	Vacaciones	0.33	1.14
				9	Capacitación	11.33	9.00
				10	Negocios	1.67	1.00
6	hoehe⋆	Monto	Monto del Crédito solicitado al Banco DM (continua)				
				1	$> 20,000$	0,00	0,00
6b	dhoehe⋆	Monto	Monto del	2	$15,000 < \ldots \leq 20,000$	1.00	0.29
			crédito	3	$10,000 < \ldots \leq 15,000$	7.00	2.00
			solicitado	4	$7,500 < \ldots \leq 10,000$	6.67	3.71
			al banco	5	$5,000 < \ldots \leq 7,500$	11.33	9.71
			DM	6	$2,500 < \ldots \leq 5,000$	25.00	28.57
				7	$1,500 < \ldots \leq 2,500$	19.67	24.57
				8	$1,000 < \ldots \leq 1,500$	17.00	19.86
				9	$500 < \ldots \leq 1,000$	11.33	9.14
				10	≤ 500	1.00	2.14

No.	Nombre en base	Nombre en texto	Descripción	Valor	Categorías	Frec. Relativa en % de: Créditos Buenos	Créditos Malos
7	sparkont	Ahorros	Monto de los Ahorros	1	No disponible/ Sin ahorros	72.33	55.14
				2	< 100	9.33	3.00
				3	$100 \leq \ldots < 500$	3.67	7.43
				4	$500 \leq \ldots < 1,000$	2.00	6.00
				5	$\geq 1,000$	10.67	21.57
8	beszeit	Antiguedad_E	Antigüedad en el empleo	1	Sin empleo	7.67	5.57
				2	≤ 1 año	23.33	14.57
				3	$1 \leq \ldots < 4$ años	34.67	33.57
				4	$4 \leq \ldots < 7$ años	13.00	19.29
				5	≥ 7 años	21.33	27.00
9	rate⋆	Capacidad	Porcentaje de ingresos disponibles para el pago	1	≥ 35	11.33	14.57
				2	$25 \leq \ldots < 35$	20.67	24.14
				3	$20 \leq \ldots < 25$	15.00	16.00
				4	< 20	53.00	45.29
10	famges⋆	Género	Estado Marital/ Sexo	1	Hombre divorciado/ vive aparte	6.67	4.29
				2	Mujer divorciada/ vive aparte	11.33	10.29
				2	Hombre soltero	25.00	18.43
				3	Hombre casado/ viudo	48.67	57.43
				4	Mujer soltera	8.33	9.57
11	buerge	Aval	Aval o Garante	1	Ninguno	90.67	90.71
				2	Co-solicitante	6.00	3.29
				3	Garante	3.33	6.00
12	wohnzeit	Antiguedad_V	Antigüedad en la vivienda	1	≤ 1 año	12.00	13.43
				2	$1 \leq \ldots < 4$ años	32.33	30.14
				3	$4 \leq \ldots < 7$ años	14.33	15.14
				4	≥ 7 años	41.33	41.29
13	verm	Garantías	Posibles bienes en garantía	1	No disponbles/ Sin activos	20.00	31.71
				2	Carro / Otro	23.67	23.00
				3	Cuenta de ahorro /Seguros de vida	34.00	32.86
				4	Casa o Terreno	22.33	12.43
14	alter⋆	Edad	Edad en años del solicitante (continua)				
14b	dalter⋆	Edad	Edad en años del solicitante	1	$0 \leq \ldots \geq 25$ años	26.67	15.71
				2	$26 \leq \ldots \geq 39$ años	47.33	52.72
				3	$40 \leq \ldots \geq 59$ años	21.67	26.14
				4	$60 \leq \ldots \geq 64$ años	2.33	3.00
				5	≥ 65 años	2.00	2.43
15	weitkred⋆	Créditos_O	Otros créditos vigentes	1	En otros bancos	19.00	11.71
				2	Tiendas Departamentales	6.33	4.00
				3	Sin otros créditos vigentes	74.67	84.29

No.	Nombre en base	Nombre en texto	Descripción	Valor	Categorías	Frec. Relativa en % de: Créditos Buenos	Créditos Malos
16	wohn	Vivienda	Tipo de vivienda	1	Departamento libre	62.00	75.43
				2	Piso rentado	14.67	9.14
				3	Propietario	23.33	15.57
17	bishkred	Créditos_DM	No. de créditos previos en el Banco DM incluyendo los que corren	1	Uno	66.67	61.86
				2	Dos o tres	30.67	34.43
				3	Cuatro o cinco	2.00	3.14
				4	Seis o más	0.67	0.57
18	beruf⋆	Ocupación	Ocupación	1	Desempleado/ No residente	2.33	2.14
				2	Sin preparación/ Residencia permanente	18.67	20.57
				3	Trabajador calificado /Funcionario menor	62.00	63.43
				4	Ejecutivo/Autoempleo /Alto funcionario	17.00	13.86
19	pers	Dependientes	Número de Dependientes	1	3 o más	15.33	15.57
				2	0 o 2	84.67	84.43
20	telef	Teléfono	Teléfono	1	No	62.33	58.43
				2	Sí	37.67	41.57
21	gastarb	Trabajo	Trabajo Foráneo	1	Sí	1.33	4.71
				2	No	98.67	95.29

Anexo D: Frecuencias de las variables de la *base DM*

A continuación se presentan las tablas de frecuencia de cada una de las variables de la *base DM*, así como los histogramas de las 20 variables de la *base DM* donde en las gráficas de lado izquierdo se encuentran las cifras globales y del lado derecho los histogramas con base en la respuesta de la variable *kredit* (Crédito), que designa 1 a los clientes considerados como con capacidad crediticia (barras de color rojo) y cero a los clientes sin capacidad crediticia (barras de color azul). Estos histogramas sirven para conocer las categorías de las variables de mayor frecuencia y de mayor relevancia para otorgar o no un crédito a un futuro solicitante. Las gráficas se encuentran en el orden en que se presentan las variables en la tabla de descripción del Anexo C.

Saldo en cuentas (*laufkont*)

Variable *laufkont*

Categoría	**Frecuencia**	**%**	**% Acum**
1: Sin cuenta corriendo	214	27.4	27.4
2: Sin saldo o deuda	269	26.9	54.3
3: $0 \leq \ldots < 200$	63	6.3	60.6
4: ≥ 200 o verificado hasta por lo menos 1 año	394	39.4	100
Total	1,000	100	

La categoría de mayor frecuencia es la 4 que pertenece a solicitantes con 200 marcos alemanes como saldo en sus cuentas, o bien que el banco tiene certeza de ellos por una vigencia de un año, a su vez en el histograma de la izquierda de la figura 12 se aprecia que la mayoría de los clientes con capacidad crediticia caen en esta categoría, de hecho el no tener cuentas en el banco DM parece factor relevante para otorgar el crédito. Debido a la baja frecuencia de la categoría 3, y de acuerdo a las exploraciones realizadas sobre esta variable y el resto, se puede agrupar la categoría 2 y 3.

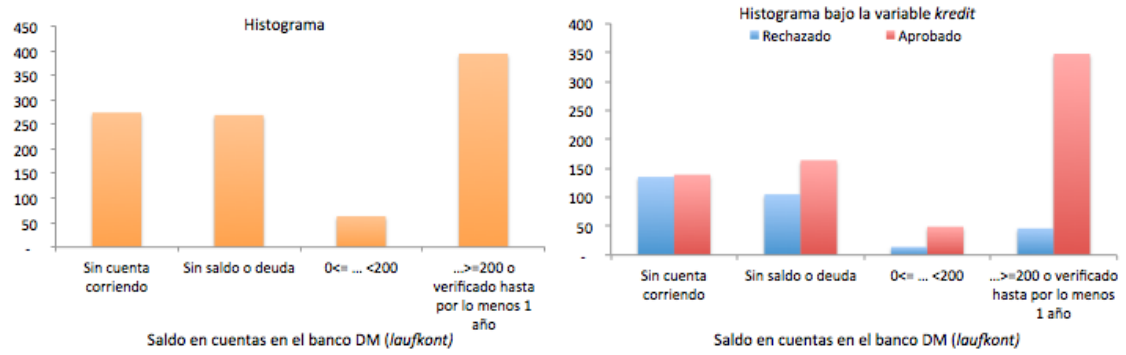

Figura 12: A la izquierda el histograma de la variable Saldos, *laufkont*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Duración en meses del crédito (*laufzeit*, continua)

Esta es la primera de las tres variables continuas que presenta la base y por la naturaleza de la misma se pueden plantear otro tipo de observaciones:

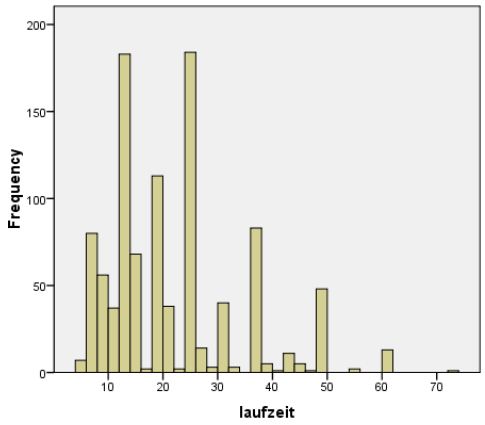

Figura 13: En el histograma las barras muestran que la base es bimodal, una moda la tiene en los 14 meses y otra en los 24 meses. Sin embargo, se observa que existe gran dispersión en los datos. Si se hace un recorrido por el eje de las ordenadas se observa que las mayores frecuencias se muestran en préstamos anuales.

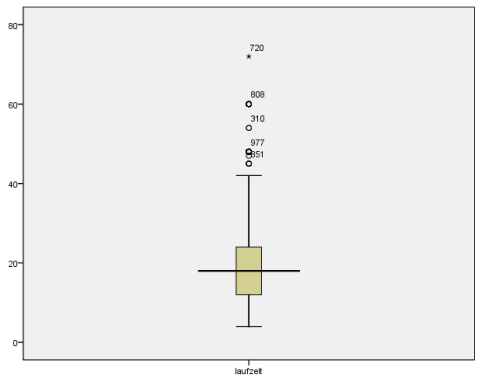

Figura 14: Existen varios *outliers* en la muestra, y la caja se encuentra sesgada hacia "pocos meses", esto es más claro si se observan los cuartiles del gráfico: Q_1 : 12 meses, Q_2 : 18 meses (mediana) y Q_3 : 24 meses, cuanto el valor máximo está en los 72 meses.

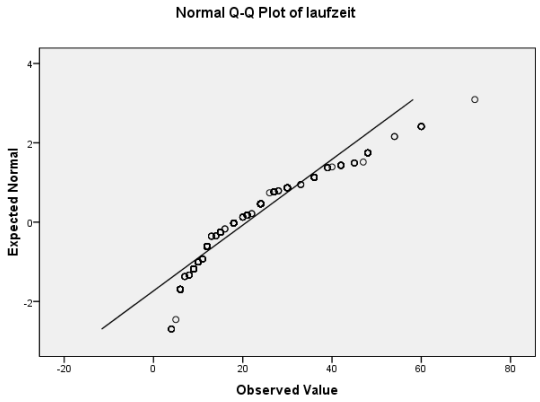

Figura 15: Q-Q plot de la variable *laufzeit*, duración en meses del crédito, que muestra una media de 21 meses

El Q-Q plot de la variable *laufzeit* con respecto de una distribución normal resulta poco apegado a ésta, sin embargo se observará que tiene mayor similitud que otras variables. En la realidad, no se espera que se otorguen créditos de ciertos plazos en particular, sino que se busca una cartera de clientes saludables que puedan pagar sus deudas en un plazo no tan mayor al pactado.

Duración en meses del crédito (*dlaufzeit*, categórica)

Variable *dlaufzeit*

Categoría	**Frecuencia**	**%**	**% Acum**
1: > 54	14	1.4	1.4
2: $48 < \ldots \leq 54$	2	0.2	1.6
3: $42 < \ldots \leq 48$	54	5.4	7
4: $36 < \ldots \leq 42$	17	1.7	8.7
5: $30 < \ldots \leq 36$	86	8.6	17.3
6: $24 < \ldots \leq 30$	57	5.7	23
7: $18 < \ldots \leq 24$	224	22.4	45.4
8: $12 < \ldots \leq 18$	187	18.7	64.1
9: $6 < \ldots \leq 12$	277	27.7	91.8
10: ≤ 6	82	8.2	100
Total	1,000	100	

Con ayuda de la variable categórica Duración, *dlaufzeit*, se aprecia en la tabla de frecuencias que la mayor frecuencia es en la categoría 9, de 6 a 12 meses, y en segundo lugar la categoría 7 que corresponde a un periodo de entre 18 a 24 meses. Las categorías 1 y 2 muestran la menor temporalidad solicitada y no se aprecian frecuencias muy altas. Por otro lado, no se debe olvidar que la media en esta variable es de 21 meses y su varianza de 145 meses, lo cual indica dispersión en los datos.

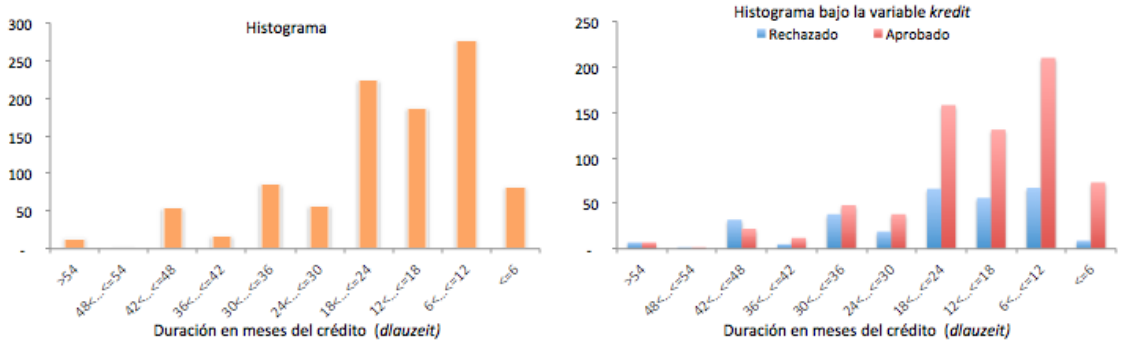

Figura 16: A la izquierda el histograma de la variable Duración, *dlaufzeit*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Pagos de crédito previos(*moral*)

Variable *moral*

Categoría	**Frecuencia**	**%**	**% Acum**
0: Atraso en pago de créditos previos	40	4.0	4.0
1: Problemas con créditos vigentes/ Con créditos vigentes en otros bancos	49	4.9	8.9
2: Sin créditos previos/ Créditos previos saldados	530	53.0	61.9
3: Sin problemas con créditos vigentes en el banco DM	88	8.8	70.7
4: Saldados todos los créditos previos en el banco DM	293	29.3	100
Total	1,000	100	

La mayor frecuencia se encuentra en la categoría 2 y junto con la categoría 4 son dos factores que permiten se otorgue el crédito. Las categorías 0 y 1 que equivalen a problemas con créditos previos no prevalece en general en la muestra, se pueden agrupar estas categorías.

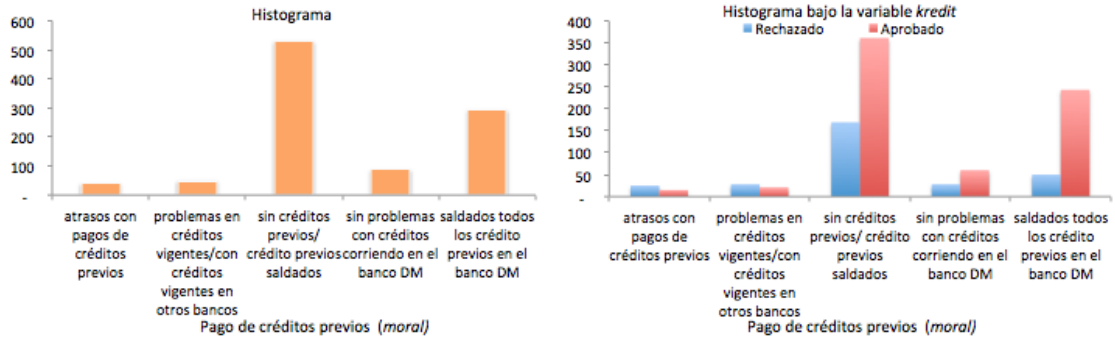

Figura 17: A la izquierda el histograma de la variable Morosidad, *moral*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Propósito del crédito (*verw*)

Variable *verw*

Categoría	**Frecuencia**	**%**	**% Acum**
0: otros	234	23.4	23.4
1: carro nuevo	103	10.3	33.7
2: carro usado	181	18.1	51.8
3: muebles	218	28.0	79.8
4: radio / televisión	12	1.2	81.0
5: electrodomésticos	22	2.2	83.2
6: reparaciones	50	5.0	88.2
7: educación	0	0.0	88.2
8: vacaciones	9	0.9	89.1
9: capacitación	97	9.7	98.8
10: negocios	12	1.2	100
Total	1,000	100	

Es posible unir categorías sobre esta variable al hacer un análisis más profundo en relación con otras variables como son: monto del crédito otorgado y duración del mismo, así como la decisión de si es un cliente con o sin capacidad crediticia. Todo esto es posible al aplicar técnicas como *projection pursuit* la cual se analiza en el Capítulo 2.

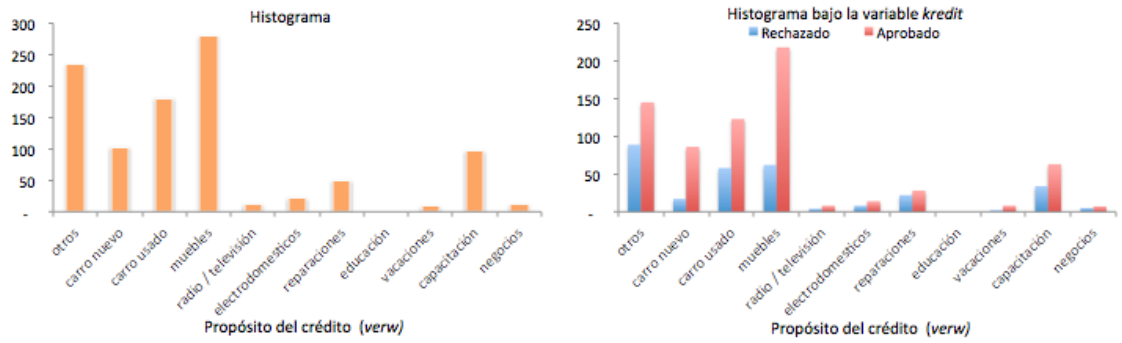

Figura 18: A la izquierda el histograma de la variable Propósito, *verw*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Monto del crédito (*hoehe*, continua)

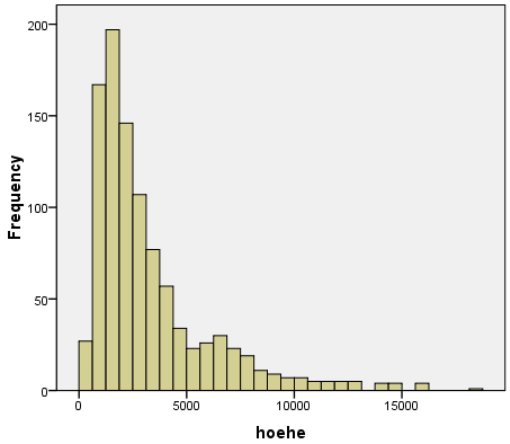

Figura 19: En el histograma de la variable, las barras muestran curtosis positiva sesgada totalmente a la izquierda, además de una sola moda muy bien marcada, pues se tiene que la mayoría de los préstamos solicitados están por debajo de los 5,000 marcos, con lo cual se confirma lo visto en el diagrama de caja anterior.

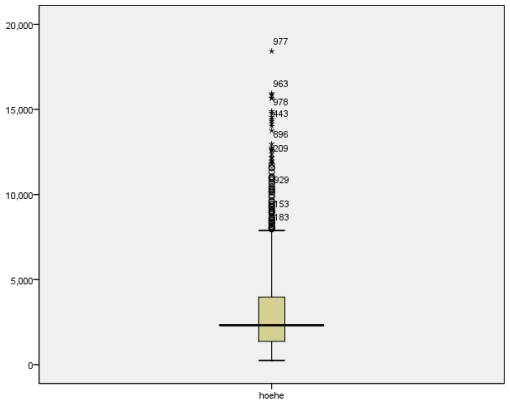

Figura 20: En el diagrama de caja de la variable se aprecia que existen muchos *outlayers* en la muestra y además un gran sesgo a cantidades menores a los 5,000 DM; se confirma esto echando un vistazo a los cuartiles de la caja: Q_1 : 1,364,5 marcos, Q_2 : 2,319,5 marcos (mediana) y Q_3 : 3,972,75 marcos

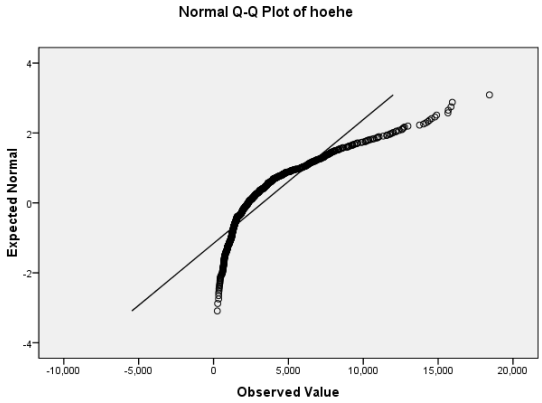

Figura 21: Q-Q plot de la variable *hoehe*, monto del crédito, que muestra una media de 3,271 marcos alemanes

El Q-Q plot de la variable *hoehe* con respecto de una distribución normal resulta poco agradable pues no existe presencia de similitud con esta distribución. De igual manera que con la variable Duración, no se espera, o bien, no es necesario que esta variable se asemeje a una distribución normal.

Monto del crédito (*dhoehe*, categórica)

Variable *dhoehe*

Categoría	**Frecuencia**	**%**	**% Acum**
1: $> 20,000$	0	0.0	0.0
2: $15,000 < \ldots \leq 20,000$	5	0.5	0.5
3: $10,000 < \ldots \leq 15,000$	35	3.5	4
4: $7,500 < \ldots \leq 10,000$	46	4.6	8.6
5: $5,000 < \ldots \leq 7,500$	102	10.2	18.8
6: $2,500 < \ldots \leq 5,000$	275	27.5	46.3
7: $1,500 < \ldots \leq 2,500$	231	23.1	69.4
8: $1,000 < \ldots \leq 1,500$	190	19	88.4
9: $500 < \ldots \leq 1,000$	98	9.8	98.2
10: ≤ 500	18	1.8	100
Total	1,000	100	

En la tabla de frecuencias de la variable se tiene que la mayor frecuencia se encuentra en la categoría 6 ($2,500 \leq \ldots \leq 5,000$ DM), seguida de la 7 ($5,000 \leq \ldots \leq 7,500$). Cabe mencionar que la media de esta variable es de 3,271 DM y que son muy pocos los préstamos ≤ 500 DM y $> 20,000$, de hecho se puede ver que las solicitudes tienden a ser de montos mayores o iguales a 500 DM, pero no alcanzan valores tan grandes con gran frecuencia.

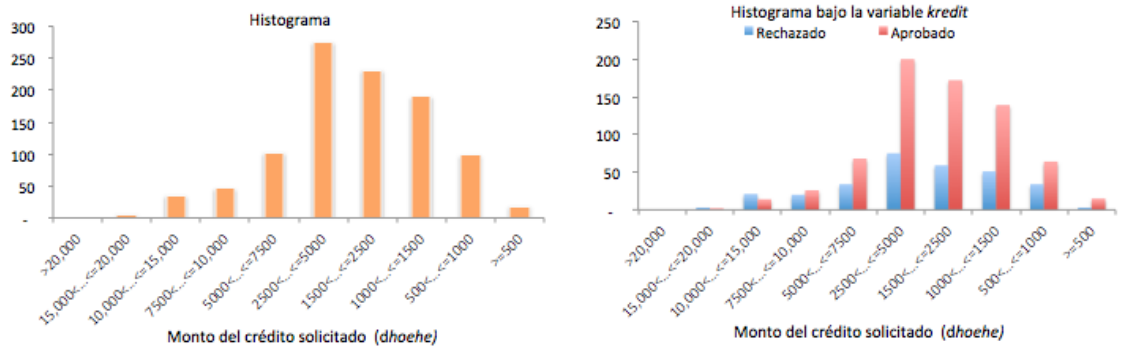

Figura 22: A la izquierda el histograma de la variable Monto, *dhoehe*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Valor de las cuentas de ahorro (*sparkont*)

Variable *sparkont*

Categoría	**Frecuencia**	**%**	**% Acum**
1: No disponible/Sin ahorros	603	60.3	60.3
2: < 100	103	10.3	70.6
3: $100 \leq \ldots < 500$	63	6.3	76.9
4: $500 \leq \ldots < 1,000$	48	4.8	81.7
5: $\geq 1,000$	183	18.3	100
Total	1,000	100	

La categoría con mayor frecuencia es la primera que nos indica que el cliente no cuenta con Ahorros en el *banco DM*, o bien no es un dato disponible lo cual implica que son clientes con otro tipo de relación en el banco, o bien son clientes nuevos. Posiblemente las tres categorías intermedias podrían agruparse.

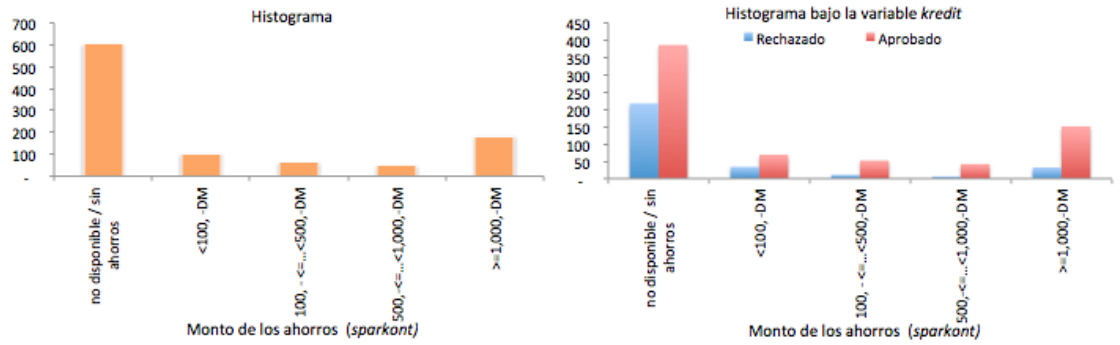

Figura 23: A la izquierda el histograma de la variable Ahorros, *sparkont*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Antigüedad en el empleo (*beszeit*)

Variable *beszeit*

Categoría	**Frecuencia**	**%**	**% Acum**
1: Sin empleo	62	6.2	6.2
2: ≤ 1 año	172	17.2	23.4
3: $1 \leq \ldots < 4$ años	339	33.9	57.3
4: $4 \leq \ldots < 7$ años	174	17.4	74.7
5: ≥ 7 años	253	25.3	100
Total	1,000	100	

La categoría de mayor frecuencia es aquella con antigüedad entre 1 y 4 años, debido a que esta variable es indispensable para conocer si el cliente tiene los recursos para saldar la deuda.

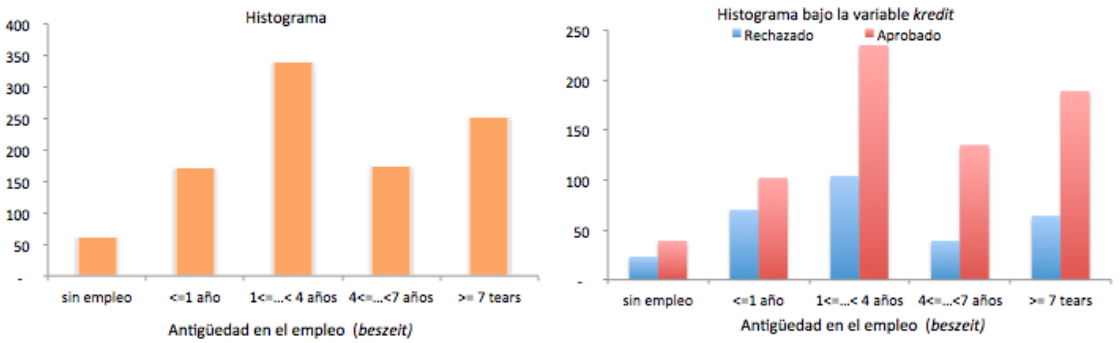

Figura 24: A la izquierda el histograma de la variable Antiguedad_E, *beszeit*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Porcentaje de los ingresos disponibles para el pago (*rate*)

Variable *rate*

Categoría	Frecuencia	%	% Acum
1: ≥ 35	136	13.6	13.6
2: $25 \leq \ldots < 35$	231	23.1	36.7
3: $20 \leq \ldots < 25$	157	15.7	52.4
4: < 20	476	47.6	100
Total	1,000	100	

Nótese que la mayoría de los clientes (47. %) cuenta con menos del 20 % de sus ingresos disponibles para pagar la posible deuda a contraer, sería lógico el por qué están buscando un crédito para financiarse, en estos casos son clientes potenciales a generar revolvencia y a su vez ganancias a la institución.

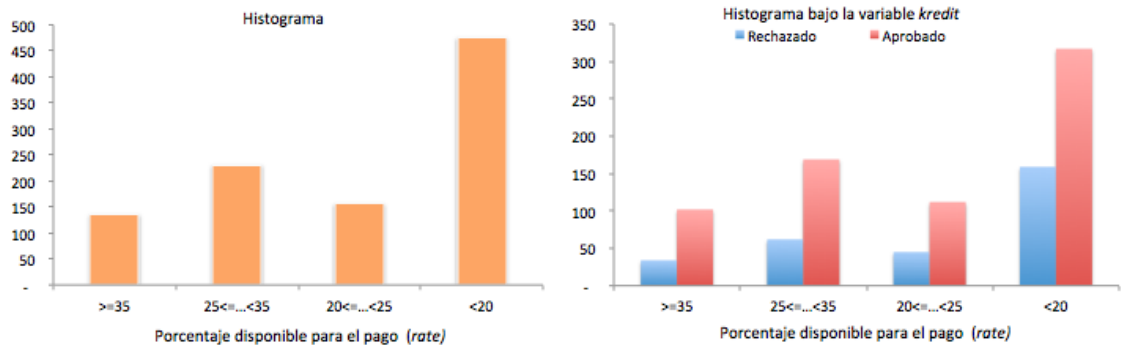

Figura 25: A la izquierda el histograma de la variable Capacidad, *rate*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Estado marital y sexo (*famges*)

Variable *famges*

Categoría	Frecuencia	%	% Acum
1: Hombre divorciado/vive aparte	50	5.0	5.0
2: Mujer divorciada/vive aparte			
Hombre soltero	310	31.0	36.0
3: Hombre casado/viudo	548	54.8	90.8
4: Mujer soltera	92	9.2	100
Total	1,000	100	

Esta covariable ya se encuentra categorizada desde el origen de la base y se observa que se han mezclado los géneros con el estado civil de los clientes, por lo que es preferible dejar la variable sin agrupar categorías, pues seguramente por análisis previos se ha categorizado de esta manera. Se aprecia que casi el 55 % de los clientes

son hombres casados o viudos; en segundo, lugar con un 31 %, hombres solteros o mujeres divorciadas o solteras.

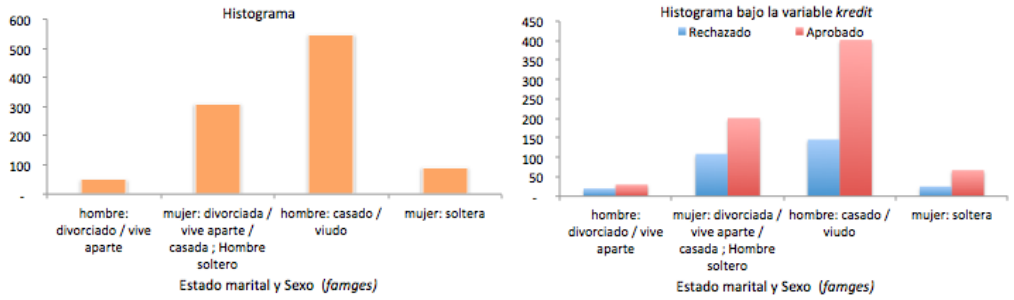

Figura 26: A la izquierda el histograma de la variable Género, *famges*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Aval o garante (*buerge*)

Variable *buerge*

Categoría	**Frecuencia**	**%**	**% Acum**
1: Ninguno	907	90.7	90.7
2: Co-Solicitante	41	4.1	94.8
3: Garante	52	5.2	100
Total	1,000	100	

Nótese que el 90 % de los clientes no actúan como aval o garante, el otro 10 % que sí lo hace puede representar ante la institución mayor riesgo.

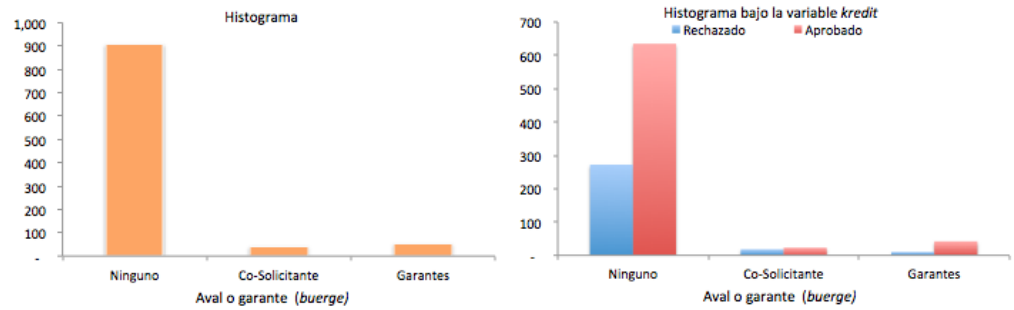

Figura 27: A la izquierda el histograma de la variable Aval, *buerge*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Antigüedad en la vivienda (*wonhzeit*)

Variable *wonhzeit*

Categoría	Frecuencia	%	% Acum
1: <1 año	130	13.0	13.0
2: $1 \leq \ldots < 4$ años	308	30.8	43.8
3: $4 \leq \ldots < 7$ años	149	14.9	58.7
4: ≥ 7 años	413	41.3	100
Total	1,000	100	

La gran mayoría de los clientes tiene más de siete años viviendo en su domicilio actual (41.3 %), lo cual puede representar puntos a su favor pero realmente esta variable dice mucho más al momento de cruzarla con el resto.

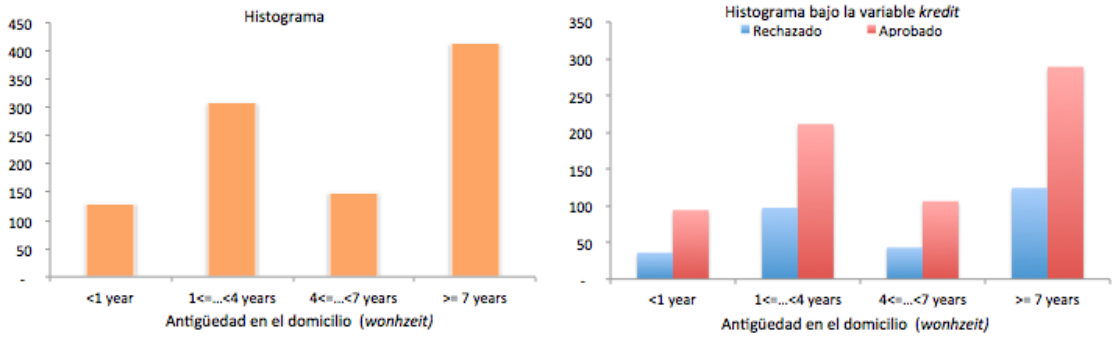

Figura 28: A la izquierda el histograma de la variable Antiguedad_V, *wonhzeit*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Posibles bienes en garantía (*verm*)

Variable *verm*

Categoría	Frecuencia	%	% Acum
1: No disponible/Sin activos	282	28.2	28.2
2: carro / otro	232	23.2	51.4
3: contrato de ahorro/seguro de vida	332	33.2	84.6
4: propietario de casa o terreno	154	15.4	100
Total	1,000	100	

Nótese que el 30 % de los clientes no cuenta con algún bien para dejar en garantía; el origen de esta variable puede referirse a cuando se solicitan créditos para autos o algún bien raíz, ya que los mayores porcentajes de créditos autorizados son para clientes sin algún bien de respaldo o para aquellos que tienen saldos de ahorro en el banco.

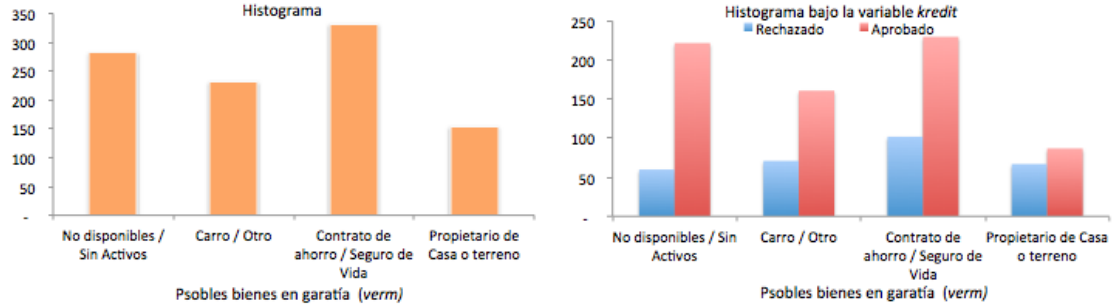

Figura 29: A la izquierda el histograma de la variable Garantías, *verm*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Edad en años del cliente (*alter*, continua)

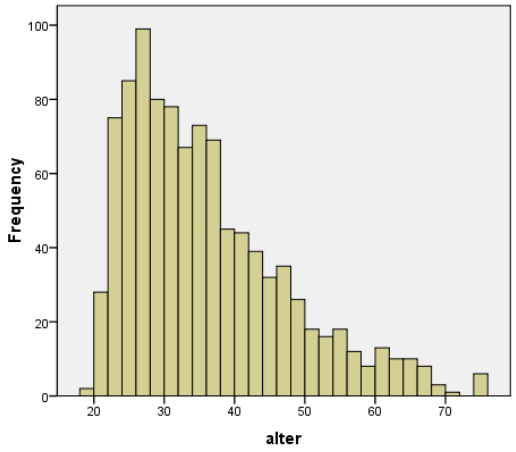

Figura 30: En el histograma de la variable Edad, *alter*, (edad en años del cliente), las barras muestran curtosis positiva sesgada totalmente a la izquierda, además de una sola media muy bien marcada en los 28 años, con gran presencia en edades de entre 28 y 38 años.

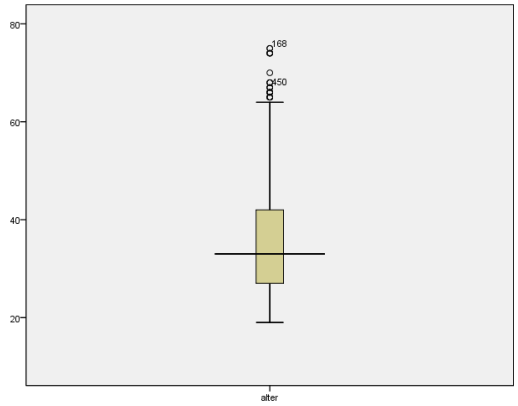

Figura 31: Diagrama de Caja de la variable Edad, se presenta más amplia con edades en su mayoría por debajo de los 40 años aproximadamente y con presencia de pocos *outlayers* en la muestra. Se confirma esto con los cuartiles de la caja: Q_1 : 27 años, Q_2 : 33 años (mediana) y Q_3 : 42 años.

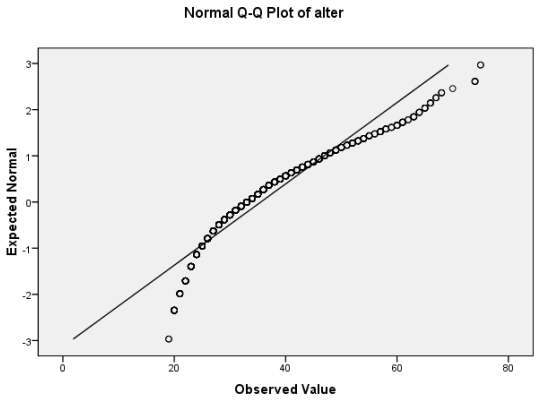

Figura 32: Q-Q plot de la variable *alter*, edad en años del cliente, que muestra una media de 3,271 marcos alemanes

El Q-Q plot de la variable Edad (*alter*) con respecto de una distribución normal resulta no tan apegado como se quisiera a una distribución normal, sin embargo sí en un tramo de la curva donde se encuentra la mayor población de esta muestra. Realmente se espera que las mayores frecuencias de edad de solicitantes a crédito se encuentren en edad productivas y de estabilidad, de esta forma es más probable que los solicitantes paguen sus deudas.

Edad en años del cliente (*dalter*, categórica)

Siguiendo con la última variable continua: *alter*, edad en años del cliente:

Variable *dalter*

Categoría	**Frecuencia**	**%**	**% Acum**
1: $0 \leq \ldots \geq 25$ años	190	19.0	19.0
2: $26 \leq \ldots \geq 39$ años	511	51.1	70.1
3: $40 \leq \ldots \geq 59$ años	248	24.8	94.9
4: $60 \leq \ldots \geq 64$ años	28	2.8	97.7
5: ≥ 65 años	23	2.3	100
Total	1,000	100	

Como bien se puede apreciar en la distribución de los valores de la variable continua, en su mayoría los clientes se encuentran en las categorías 2 y 3, sin embargo aquí se aprecia que la institución prefiere otorgar créditos a clientes de entre 40 y 59 años que a clientes con edades menores a 25 años.

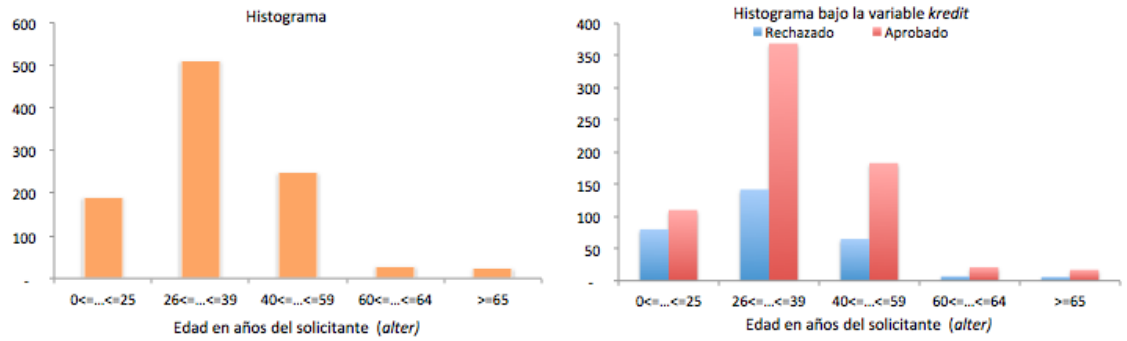

Figura 33: A la izquierda el histograma de la variable Edad, *dalter*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Otros créditos vigentes (*weitkred*)

Variable *weitkred*

Categoría	**Frecuencia**	**%**	**% Acum**
1: en otros bancos	139	13.9	13.9
2: tiendas departamentales	47	4.7	18.6
3: sin otros créditos vigentes	814	81.4	100
Total	1,000	100	

Nótese que la mayoría de los "buenos créditos" se otorgan a clientes sin otros créditos vigentes y esto es fácil de entender considerando que en su mayoría son clientes con menos del 20 % de sus ingresos para cubrir deudas.

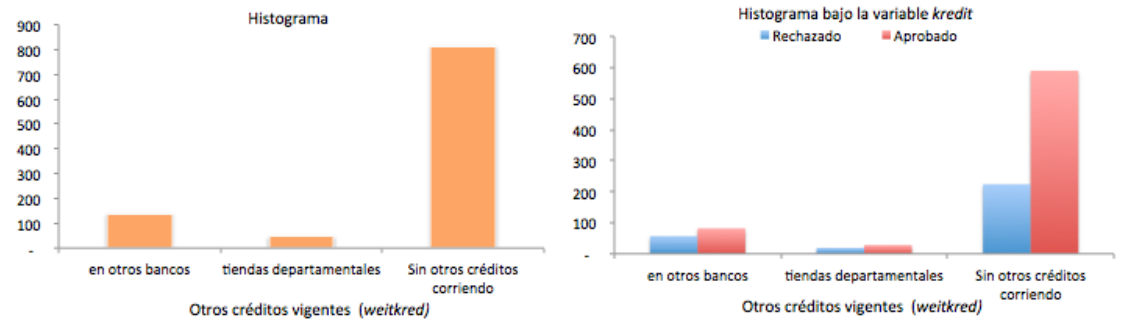

Figura 34: A la izquierda el histograma de la variable Créditos_O, *weitkred*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Tipo de vivienda (*wohn*)

Variable *wohn*

Categoría	Frecuencia	%	% Acum
1: departamento libre	179	17.9	17.9
2: piso rentado	714	71.4	89.3
3: propietario	107	10.7	100
Total	1,000	100	

A grandes rasgos, esta variable indica que el 89.3 % de los clientes vive en pisos rentados.

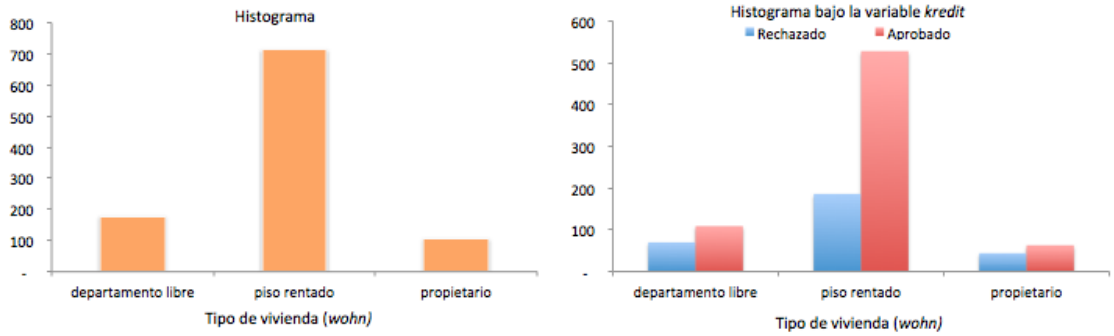

Figura 35: A la izquierda el histograma de la variable Vivienda, *wohn*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Número de créditos previos en el banco DM incluyendo los que actualmente se tienen (*brishkred*)

Variable *brishkred*

Categoría	Frecuencia	%	% Acum
1: uno	633	63.3	63.3
2: dos o tres	333	33.3	96.6
3: cuatro o cinco	28	2.8	99.4
4: seis o más	6	0.6	100
Total	1,000	100	

Se observa que el 63.3 % de los clientes ya cuenta con un crédito vigente en la institución, lo cual hace pensar en algún proceso de fidelización o retención de clientes, pues un gran porcentaje de clientes tienen capacidad crediticia en esta muestra.

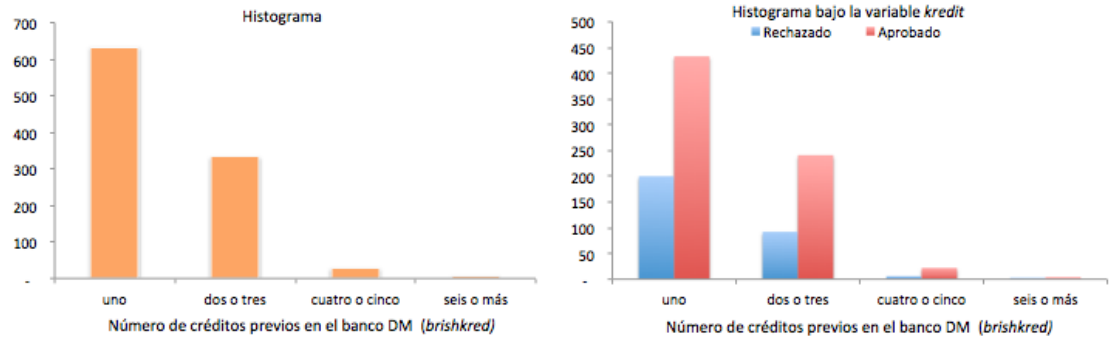

Figura 36: A la izquierda el histograma de la variable Créditos_DM, *brishkred*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Ocupación (*beruf*)

Variable *beruf*

Categoría	**Frecuencia**	**%**	**% Acum**
1: desempleado/no residente	22	2.2	2.2
2: sin preparación/residencia permanente	200	20.0	22.2
3: trabajador calificado/funcionario menos	630	63.0	85.2
4: ejecutivo/autoempleo/alto funcionario	148	14.8	100
Total	1,000	100	

Se aprecia que el 63 % de los clientes cae en la categoría 3. No se tiene algún factor que haga más propenso el otorgamiento de créditos con base en esta variable para esta base de datos en particular. En un plano más actual esta variable puede ser de gran ayuda si la información es correcta y actualizada.

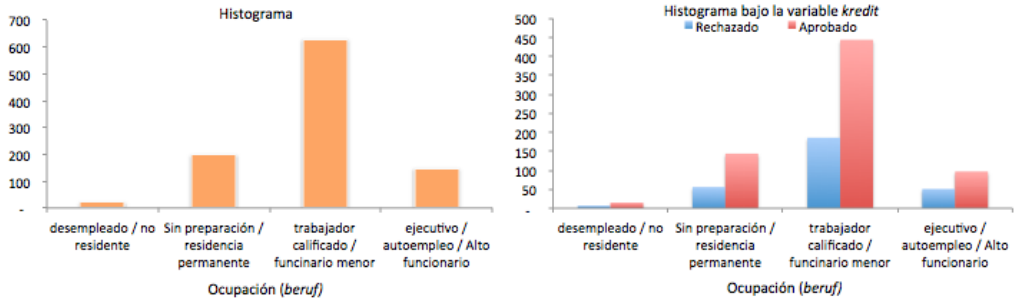

Figura 37: A la izquierda el histograma de la variable Ocupación, *beruf*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Número de dependientes (*pers*)

Variable *pers*

Categoría	Frecuencia	%	% Acum
1: 3 y más	845	84.5	84.5
2: 0 a 2	155	15.5	100

Se observa que el 85 % de la base cuenta con 3 o más dependientes económicos, esto se puede atribuir al hecho de la época social en la cual fue generada la base de datos ya que en su mayoría estos créditos fueron solicitados por padres de familia, o bien, la cabeza de la familia.

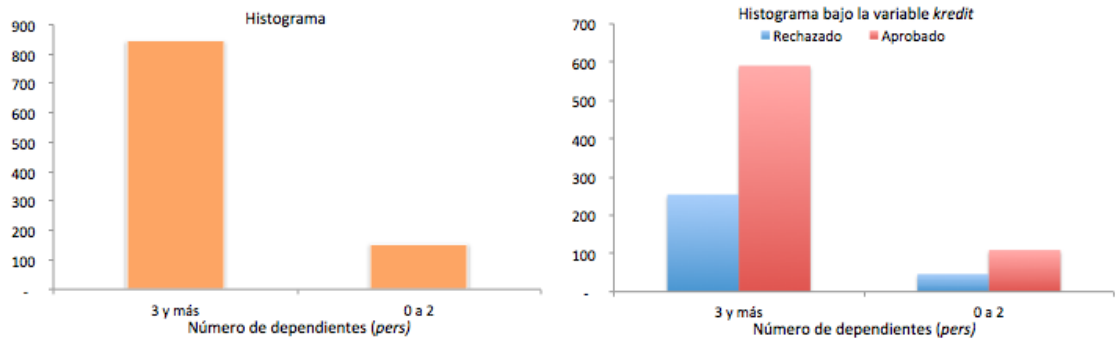

Figura 38: A la izquierda el histograma de la variable Dependientes, *pers*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Teléfono (*telef*)

Variable *telef*

Categoría	Frecuencia	%	% Acum
1: no	596	59.6	59.6
2: si	404	40.4	100

Para describir esta variable se puede tomar el mismo argumento de la variable anterior, la época social a la cual pertenece la información, pues en la actualidad en su mayoría los solicitantes a crédito cuentan ya con algún teléfono. Algo que sí es muy frecuente de encontrar es que debido a la mala calidad de datos no se tenga esta información clara, la cual es realmente importante para realizar campañas de colocación y contacto al cliente.

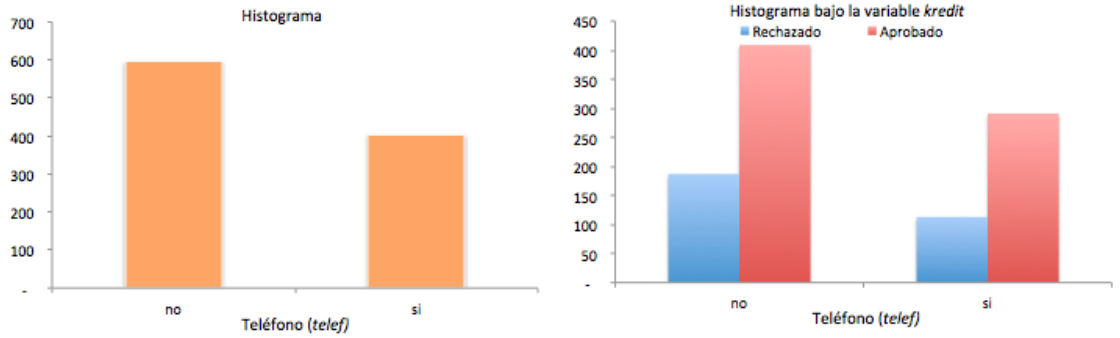

Figura 39: A la izquierda el histograma de la variable Teléfono, *telef*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Trabajo foráneo (*gastarb*)

Variable *gastarb*

Categoría	Frecuencia	%	% Acum
1: sí	963	96.3	96.3
2: no	37	3.7	100

Es relevante notar que el 96 % de los clientes tiene trabajo foráneo. Con un mayor contexto se puede explicar esta covariable, dado la falta del mismo se asumirá que esta característica es particular del total de la población y se puede evitar introducir la variable a los modelos planteados.

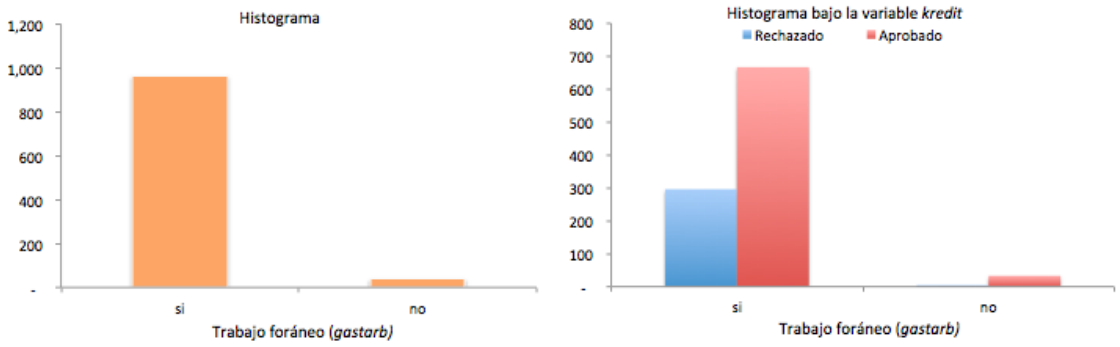

Figura 40: A la izquierda el histograma de la variable Trabajo, *gastarb*, con los 1,000 casos. A la derecha el histograma de la misma variable marcando en azul los 300 clientes sin capacidad crediticia y en rojo los 700 clientes con capacidad crediticia (clasificación con base en la variable Crédito, *kredit*).

Anexo E: Matriz de vectores propios al implementar el análisis de componentes principales en la *base DM*

A continuación se presenta la matriz de vectores propios o *eigenvectores* (Matriz Factorial) resultante de aplicar a las 10 variables seleccionadas y recodificadas de la *base DM* el método de componentes principales utilizando la matriz de Covarianzas; cabe mencionar que la base de trabajo constó de 34 variables pues las 10 variables originales categóricas se transformaron en *n-1* variables dicotómicas siendo n el número de categorías de cada variable.

Nota: aquellos valores que no aparecen son valores menores a cero.

Parte 1:

```
Loadings:
              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
laufkont_Dc1                 0.357         0.507 -0.107         0.127  0.167  0.110        
laufkont_Dc2                -0.322 -0.188 -0.402  0.232        -0.171 -0.382               
dlaufzeit_Dc1                                                          0.107               
dlaufzeit_Dc2 -0.140  0.193 -0.353         0.179 -0.505                -0.208 -0.283   0.153
dlaufzeit_Dc3                       0.139         0.194  0.101  0.145         0.664   0.210
dlaufzeit_Dc4                0.336                0.291        -0.110  0.130 -0.535  -0.264
dlaufzeit_Dc5                                                                            
moral_Dc2     -0.660         0.115 -0.135 -0.116         0.105                          
moral_Dc3      0.622  0.107 -0.142  0.178  0.108                                        0.117
verw_Dc2             -0.119                                                              
verw_Dc3             -0.102  0.125  0.118  0.220                0.147 -0.414          0.198
verw_Dc4              0.245 -0.123  0.112 -0.355               -0.101  0.613          0.138
verw_Dc5                                                                                 
verw_Dc6                                                                                 
verw_Dc7                                                                -0.103         -0.139
dhoehe_Dc1                                                                               
dhoehe_Dc2            0.122                                                              
dhoehe_Dc3            0.192 -0.167                      -0.346                0.197  -0.360
dhoehe_Dc4            0.193                      0.129  0.686  0.271        -0.206   0.183
dhoehe_Dc5           -0.365  0.327        -0.185        -0.302 -0.211                0.436
```

```
dhoehe_Dc6                                                                  -0.199
rate_Dc1             0.644  0.346 -0.175 -0.107  0.122 -0.117        -0.220
rate_Dc2            -0.156                                           0.220  0.318
rate_Dc3            -0.383 -0.223  0.112                0.211               -0.381
famges_Dc1
famges_Dc3     0.260         0.294 -0.246 -0.174 -0.557  0.376 -0.314         0.201  -0.108
famges_Dc4                  -0.121                 0.147         0.109
dalter_Dc2     0.155         0.118 -0.156 -0.375 -0.275 -0.188  0.761
dalter_Dc3                                                     -0.153
dalter_Dc4                                                     -0.186
weitkred_Dc2
weitkred_Dc3                       0.155        -0.151                  -0.126   0.216
beruf_Dc2                   0.138  0.694 -0.281 -0.162               -0.213         -0.212
beruf_Dc3                        -0.428         0.130                0.119
```

Parte 2:

```
             Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17 Comp.18 Comp.19 Comp.20
laufkont_Dc1           0.218   0.155  -0.263   0.577                  0.175
laufkont_Dc2                           -0.370   0.455           0.105   0.287
dlaufzeit_Dc1                          -0.125  -0.195   0.582  -0.244   0.264
dlaufzeit_Dc2           0.156           0.161          -0.291
dlaufzeit_Dc3          -0.125  -0.124   0.155          -0.202   0.379
dlaufzeit_Dc4 -0.106           -0.267                          0.417
dlaufzeit_Dc5                   0.106                          -0.319           0.134
moral_Dc2              -0.144
moral_Dc3              -0.100
verw_Dc2       -0.143  -0.186   0.244   0.172  -0.131  -0.383  -0.103   0.465   0.109
verw_Dc3                       -0.332  -0.516  -0.278                  -0.166   0.156
verw_Dc4        0.316   0.171  -0.117  -0.278          -0.107          -0.104
verw_Dc5
verw_Dc6                        0.110                                   0.109  -0.234
verw_Dc7                0.142           0.329   0.345   0.224          -0.566   0.157
dhoehe_Dc1
dhoehe_Dc2              0.167   0.334          -0.239   0.228   0.509
dhoehe_Dc3    -0.192    -0.162  -0.555                          -0.100
dhoehe_Dc4                     -0.141   0.192   0.164          -0.120   0.141
dhoehe_Dc5     0.300           -0.146   0.203
dhoehe_Dc6    -0.122            0.345  -0.336          -0.322          -0.356
rate_Dc1       0.178
rate_Dc2      -0.607   0.143
rate_Dc3       0.485                                           0.122
famges_Dc1                                                    -0.106          -0.290
famges_Dc3                     -0.100  -0.108           0.130   0.121           0.179
famges_Dc4                                                     0.124           0.833
dalter_Dc2
dalter_Dc3
dalter_Dc4                                                                     -0.117
weitkred_Dc2            0.231                  -0.103
weitkred_Dc3           -0.770   0.177           0.245   0.176          -0.142
beruf_Dc2     -0.153            0.144                  -0.144  -0.186
```

```
beruf_Dc3             -0.158                          -0.244  -0.244
```

Parte 3:

```
             Comp.21 Comp.22 Comp.23 Comp.24 Comp.25 Comp.26 Comp.27 Comp.28 Comp.29
laufkont_Dc1
laufkont_Dc2
dlaufzeit_Dc1                 0.148  -0.277
dlaufzeit_Dc2                 0.135  -0.132
dlaufzeit_Dc3
dlaufzeit_Dc4                                                                -0.100
dlaufzeit_Dc5          0.270  -0.250   0.634                          0.302
moral_Dc2                                      0.106           0.517          0.412
moral_Dc3                                      0.101           0.524          0.419
verw_Dc2       0.315   0.282          -0.131  -0.235  -0.236          -0.134
verw_Dc3       0.140                         -0.143  -0.186
verw_Dc4                                     -0.119  -0.141
verw_Dc5                              0.105   0.483   0.399  -0.201  -0.175
verw_Dc6      -0.487  -0.433  -0.115   0.264  -0.335  -0.351
verw_Dc7       0.227                 -0.187  -0.252  -0.125
dhoehe_Dc1                                                   -0.154  -0.200
dhoehe_Dc2     0.379          -0.118   0.146
dhoehe_Dc3
dhoehe_Dc4
dhoehe_Dc5    -0.148           0.119
dhoehe_Dc6    -0.369   0.115   0.281  -0.155
rate_Dc1                                                     -0.180  -0.339   0.245
rate_Dc2              -0.101           0.256                 -0.188  -0.379   0.239
rate_Dc3                      -0.104   0.188                 -0.191  -0.308   0.245
famges_Dc1     0.144                                0.555  -0.608          -0.278  -0.199
famges_Dc3                                          0.104  -0.122                  -0.102
famges_Dc4    -0.125  -0.231   0.156           0.182  -0.208                  -0.114
dalter_Dc2                                                                   0.127
dalter_Dc3    -0.192           -0.513  -0.287   0.308  -0.220  -0.100   0.352   0.248
dalter_Dc4     0.163           0.656   0.286           -0.227  -0.170   0.319   0.136
weitkred_Dc2                          0.104           0.126   0.458  -0.289  -0.458
weitkred_Dc3                                                                 -0.252
beruf_Dc2      0.211  -0.350
beruf_Dc3      0.326  -0.644  -0.100                   0.165           0.118
```

Parte 4:

```
             Comp.30 Comp.31 Comp.32 Comp.33 Comp.34
laufkont_Dc1
laufkont_Dc2
dlaufzeit_Dc1          0.485   0.133           0.252
dlaufzeit_Dc2          0.262   0.134  -0.204   0.181
dlaufzeit_Dc3          0.265   0.111  -0.215   0.150
dlaufzeit_Dc4          0.216   0.132  -0.198   0.126
```

dlaufzeit_Dc5	0.136	0.335	0.134	-0.229	
moral_Dc2					
moral_Dc3					
verw_Dc2			0.281		
verw_Dc3			0.243		
verw_Dc4			0.242		
verw_Dc5		-0.201	0.662	0.137	
verw_Dc6			0.354		
verw_Dc7			0.305		
dhoehe_Dc1				-0.730	-0.599
dhoehe_Dc2		0.199		0.234	-0.384
dhoehe_Dc3		0.170		0.225	-0.349
dhoehe_Dc4		0.201		0.214	-0.312
dhoehe_Dc5		0.244		0.192	-0.251
dhoehe_Dc6		0.352		0.157	-0.205
rate_Dc1	-0.130	0.167	-0.101		0.111
rate_Dc2	-0.140	0.156	-0.126		
rate_Dc3	-0.140	0.165			
famges_Dc1	0.211				
famges_Dc3	0.104				
famges_Dc4					
dalter_Dc2	-0.188				
dalter_Dc3	-0.478				
dalter_Dc4	-0.409	-0.108			
weitkred_Dc2	-0.593			-0.105	
weitkred_Dc3	-0.199				
beruf_Dc2					
beruf_Dc3					

Bibliografía

[1] Anderson R. (2007). *The Credit Socring Toolkit, Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford. First published.

[2] Agresti A. (2002). *Categorical Data Analysis*. John Wiley & Sons. Second Edition.

[3] Balakrishnan N. (2010). Methods and Applications of Statistics in Business, Finance, and Management Science. Wiley. Article: *Neural Networks for Michael Titterington*. pp. 347-356.

[4] Cook D. and Swayne D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis with R and GGobi*. Springer.

[5] Hand D. J. (2005). Good Practice in retail credit scorecard assessment. *Journal of the Operation Research Society*, **56**(9): 1109 - 1117.

[6] Hastie T. , Tibshirani R. and Friedman J. (2009). *Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer. Second Edition.

[7] Hosmer D. W. and Lemeshow S. (2000). *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Second Edition.

[8] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Second Edition.

[9] Peña D. (2002). *Análisis de Datos Multivariantes*. Mc Graw Hill, Interamericana de España. Primera Edición.

[10] Ripley B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press. First Edition.

[11] Rojas R. (1996). *Neural Networks A Systematic Introduction*. Springer.

[12] Thomas L. C., Edelman D. B. and Crook J. N. (2002). *Credit Scoring and its applications*, Monographs on mathematical modelind and computation, SIAM.

[13] Venables W. N. and Ripley B. D. (2002). *Modern applied statistcs with S*. Springer.

[14] Vos W. and Evers L. (2004). *MSc in Bioinformatics: Statistical Data Mining*. Oxford University.