UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

Regresión Logística y soluciones para
Multicolinealidad en Riesgo de Crédito.

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuaria

PRESENTA:

Delia Zaret Cárdenas Moreno

TUTORA

Dra. Lizbeth Naranjo Albarrán

Ciudad Universitaria, CD. MX, 2021

Dedicado a
mi familia

II

Índice general

Lista de figuras **V**

Lista de tablas **VII**

1. Bondad de Ajuste **3**
1.1. Devianza . . . 3
1.2. AIC . . . 4
1.3. Falsos Negativos y Falsos Positivos . . . 4
1.4. Sensibilidad y Especificidad . . . 4
1.5. Curva ROC y punto de corte . . . 5
1.6. Accuracy . . . 7

2. Modelos Lineales Generalizados **9**
2.1. Definición de Modelos Lineales Generalizados . . . 9
2.1.1. Componente Aleatoria . . . 9
2.1.2. Componente Sistemática . . . 10
2.1.3. Función de Enlace . . . 10
2.2. Modelos de Regresión para Datos Binarios . . . 10
2.2.1. Regresión Logística Binaria . . . 11

3. Reducción de Dimensión y Ponderación de Variables **17**
3.1. Multicolinealidad . . . 17
3.1.1. VIF y GVIF . . . 18
3.2. Componentes Principales . . . 19
3.2.1. Varianza explicada y selección del número de componentes . . . 24
3.2.2. Componentes principales para la matriz de correlación . . . 25
3.3. Métodos de regulación y ponderación de variables . . . 29
3.3.1. LASSO . . . 30
3.3.2. Group LASSO . . . 31
3.3.3. Ridge . . . 32

4. Aplicaciones en Riesgo de Crédito **33**
4.1. Análisis Descriptivo . . . 35
4.2. Ajuste de modelos . . . 48
4.2.1. Regresión Binaria . . . 48

4.2.2. Diagnóstico Multicolinealidad 51
4.2.3. Análisis de Componentes Principales 52
4.2.4. Regresión Binaria con componentes principales 54
4.2.5. Regresión LASSO 58
4.2.6. Regresión GROUP-LASSO 65
4.2.7. Regresión Ridge 68
4.3. Resultado 73
4.4. Conclusión 74

Apéndice A. Ajuste regresión LASSO y Ridge **77**

Índice de figuras

1.1. Ejemplo Curva ROC buen desempeño. 6
1.2. Ejemplo Curva ROC mal desempeño. 7

2.1. Cambio de probabilidad según el estimador en un modelo logístico. . 12

3.1. Representación geométrica componentes principales 20
3.2. Representación geométrica de Lasso. 30
3.3. Representación gráfica Group-Lasso. 31
3.4. Representación geométrica Ridge. 32

4.1. Distribución de clientes por edad dividido por la variable respuesta. . 36
4.2. Clientes por edad agrupada 36
4.3. Boxplot edad y límite de crédito. 37
4.4. Boxplot estado de cuenta por mes. 38
4.5. Histograma estado de cuenta por mes de abril a junio. 39
4.6. Histograma estado de cuenta por mes de julio a septiembre. 40
4.7. Histograma pagos por mes de abril a junio. 41
4.8. Histograma pagos por mes de julio a septiembre. 42
4.9. Gráfico de correlación variables continuas. 43
4.10. Diagrama dispersión estado de cuenta. 44
4.11. Matriz de confusión modelo 1, todas las variables. 50
4.12. Curva Roc modelo 1, todas las variables. 51
4.13. Componentes principales de la variable estado de cuenta. 53
4.14. Varianza acumulada variable estados de cuenta. 53
4.15. Matriz de confusión modelo 2, componentes principales. 56
4.16. Matriz de confusión modelo 3, componentes principales. 58
4.17. Curva Roc modelo 2 y 3, regresión logística con resultados de PCA. . 58
4.18. Logaritmo de λ respecto a devianza para regresión Lasso. 59
4.19. Coeficientes respecto $\log(\lambda)$ para regresión Lasso. 59
4.20. Coeficientes modelo 4 y 5, regresión Lasso λ min y λ se. 61
4.21. Matriz de confusión modelo 4, regresión lasso λ mín. 62
4.22. Matriz de confusión modelo 5, regresión Lasso λ se. 63
4.23. Curva Roc modelo 4 y 5, regresión Lasso λ mín y se. 63

4.24. Logaritmo de λ contra devianza para regresión Group Lasso y Logaritmo de λ contra coeficientes para regresión Group Lasso 65
4.25. Matriz de confusión modelo 6, Group Lasso. 66
4.26. Curva Roc modelo 6, Group Lasso. 67
4.27. Logaritmo de λ contra devianza para regresión Ridge. 68
4.28. Coeficientes contra devianza para regresión Ridge. 68
4.29. Coeficientes modelo 6 y 7, regresión Ridge λ mín. y λ se. 70
4.30. Matriz de confusión modelo 7, regresión Ridge λ mín. 71
4.31. Matriz de confusión modelo 8, regresión Ridge λ se. 71
4.32. Curva Roc modelo 7 y 8, regresión Ridge λ mín y se. 72

Índice de tablas

1.1. Matriz de confusión positivos y negativos. 4

4.1. Clasificación de variables. 35
4.2. Resumen variables de estado de cuenta. 38
4.3. Resumen variables pagos por mes. 41
4.4. Criterios prueba independencia. 45
4.5. Género *vs Y*. 45
4.6. Estado civil *vs*. *Y*. 46
4.7. Educación *vs*. *Y*. 46
4.8. Resultados prueba independencia Chi cuadrada. 47
4.9. Tabla comparativa de modelos por pruebas de bondad de ajuste. . . . 74

Introducción

En la vida cotidiana existen diversos problemas de clasificación en donde se tiene una variable respuesta que, por lo regular, nos expresa la ocurrencia de un evento o la no ocurrencia de dicho evento, esto es mejor conocido como un evento de respuesta binaria. Por ejemplo, imaginemos que tenemos una población de individuos que piden un préstamo y sólo parte de esta población paga su deuda y la otra no, o bien, existe una población de individuos que contrae una enfermedad y otros que no la contraen. Este problema se vuelve interesante en el momento en el que por medio de diversas técnicas estadísticas se puede predecir si el evento va a ocurrir o no basado en algunas características de los individuos o incluso características del evento, estas características se conocen como variables independientes y la variable respuesta binaria como variable dependiente. En este trabajo nos enfocaremos en describir y aplicar los fundamentos de algunas de las técnicas estadísticas que ayudan a hacer estas predicciones, como la regresión logística, y sus principales elementos conformados por la componente aleatoria y sistemática. Por otro lado hablaremos de un problema que se presenta al tener una alta correlación entre las variables independientes, esto se conoce como multicolinealidad y cuando esta característica está presente en nuestra base de estudio se necesita implementar otras alternativas ya que la multicolinealidad puede causar que la interpretación del modelo sea errónea, una de las técnicas que han mostrado buenos resultados para atenuar este problema son los modelos de penalización en particular la regresión ridge [10]. También es posible implementar modelos de reducción de dimensión como análisis de componentes principales y a partir de los resultados construir un nuevo modelo de regresión logística.

Estos modelos, a pesar de que tienen muchas otras aplicaciones, en este trabajo serán aplicados a una base de riesgo de crédito, donde el objetivo principal es clasificar de manera adecuada a los clientes que cumplen su deuda y a los que no lo hacen y poder aplicar el modelo a clientes futuros para poder identificar si ese cliente será un deudor o pagará su deuda. Antes de ajustar cualquier modelo es indispensable estudiar nuestra base de datos mediante un análisis exploratorio de los datos en donde nos familiarizaremos con las variables continuas y categóricas. Una vez que hayamos finalizado este proceso construiremos y evaluaremos mediante pruebas de bondad de ajuste diversos modelos para poder elegir el modelo que mejor se adecue a la base de datos, y finalmente podamos interpretar los resultados.

Capítulo 1

Bondad de Ajuste

En el momento que se requiere hacer un análisis estadístico, es sumamente importante elegir el modelo más apropiado, en este caso hablaremos de modelos de regresión. Es decir, una vez que tenemos la modelación de un conjunto de datos, nos interesa saber si el ajuste del modelo es adecuado o suficientemente bueno para el conjunto de datos que estamos estudiando.

Por lo regular durante el proceso de la modelación se obtienen varios modelos en donde se aplican distintos métodos de estimación. Para saber cuál es el modelo que tiene mejores resultados es necesario aplicar medidas de bondad de ajuste a todos los modelos y compararlos.

En otras palabras, lo que se busca es verificar si el ajuste de los modelos es bueno y si la propuesta que se tiene es consistente. Esta verificación se hace a través de ciertas medidas de bondad de ajuste, las más comunes son la devianza, el Criterio de información de Akaike (AIC) y la curva ROC, *receiver operating characteristic curve*, que engloba la especificidad y sensibilidad las cuales nos ayudan a identificar la proporción de falsos negativos y falsos positivos.

1.1. Devianza

La Devianza se utiliza como una medida de bondad de ajuste dada una secuencia de modelos anidados. En el caso de los modelos lineales generalizados de los cuales hablaremos más a detalle en el capítulo 2. Se puede representar de la siguiente forma:

$$D = \sum_{i=1}^{n} d_i,$$

donde d_i's se conocen como los componentes de la devianza.

Esta componente de la devianza se define como:

$$d_i = -2\log(f(x_i;\pi))$$

donde$f(x_i;\pi)$ es la función de verosimilitud del i-ésimo elemento, x_i.

1.2. AIC

El Criterio de Información de Akaike (AIC) fue propuesto en 1974 por el estadístico japonés Akaike en su libro *Information Theory an and Extension of the Maximum Likelihood Principle* [1] y se define como una medida de bondad de ajuste caracterizado por ser un estimador asintótico insesgado $E[\log(f(x;\pi))]$, dado por la esperanza de la función de log-verosimilitud, cuya expresión generalizada es la siguiente:

$$AIC = -2\ log(f(x;\pi)) + 2k,$$

donde k es el número de parámetros estimados bajo un modelo.

Para utilizar esta medida es necesario tener más de un modelo, dado que es una herramienta que ayuda a comparar los modelos entre sí evaluando el ajuste del modelo y su complejidad, entre más pequeño sea el AIC es mejor el ajuste, sin embargo se deben tomar en cuenta todos los resultados de las otras pruebas de bondad de ajuste para poder seleccionar un modelo óptimo.

1.3. Falsos Negativos y Falsos Positivos

Supongamos que existe un conjunto de individuos que se realizan una prueba para saber si están enfermos o no. En caso de que se presente un falso positivo (FP) quiere decir que a un individuo se le detectó la enfermedad pero en realidad no la tiene. Por otro lado un falso negativo (FN) quiere decir que un individuo fue catalogado como sano pero en realidad sí tiene la enfermedad. Por otro lado existen dos casos más verdaderos positivos (VP) en donde el individuo tiene la enfermedad y la prueba detecta la enfermedad, correspondiente a una clasificación correcta de los individuos. Por último tenemos los verdaderos negativos (VN) que son aquellos individuos que están sanos y la prueba los cataloga como sanos. Esto se puede apreciar en la tabla 4.4. Esta tabla también es conocida como matriz de confusión (*confusion matrix*).

Predicción/Prueba		
Observación	Negativo	Positivo
Positivo	FN	VP
Negativo	VN	FP

Tabla 1.1: Matriz de confusión positivos y negativos.

1.4. Sensibilidad y Especificidad

La sensibilidad y especificidad son medidas que nos ayudan a detectar la proporción de falsos positivos y falsos negativos que se obtienen bajo un modelo.

La sensibilidad (*sensitivity*) nos indica la capacidad de nuestra prueba para dar como casos positivos los casos en los que si se está presentando el suceso realmente. En otras palabras, mide la proporción de verdaderos positivos que son correctamente definidos como positivos. Se puede calcular de la siguiente manera:

$$Sensibilidad = \frac{VP}{VP+FN} = \frac{VP}{Positivos},$$

donde VP=Verdaderos positivos y FN=Falsos negativos.

La especificidad (*specificity*) nos indica la capacidad de nuestro estimador para dar como casos negativos los casos en los que no se está presentando el suceso. Es decir, mide la proporción de verdaderos negativos que son correctamente identificados como negativos. Se puede calcular de la siguiente manera:

$$Especificidad = \frac{VN}{VN+FP} = \frac{VN}{Negativos},$$

donde VN=Verdaderos negativos y FP=Falsos positivos.

En resumen la sensibilidad es la proporción de verdaderos positivos y la especificidad es la proporción de verdaderos negativos.

1.5. Curva ROC y punto de corte

La curva ROC compara para distintos puntos de corte la tasa de clasificación adecuada dado que es la representación gráfica de la sensibilidad contra la especificidad, esto es, si tenemos la proporción de verdaderos negativos cuál es la proporción de verdaderos positivos que podrían existir; es por ello que la gráfica de la curva ROC en el eje vertical tiene la sensibilidad (*sensitivity*) y en el eje horizontal 1-especificidad (*specificity*). Por ejemplo, si las proporciones son iguales entonces la curva no sería una curva sino una diagonal, lo cual indica que nuestro modelo no está diferenciando bien los casos ni de verdaderos positivos ni de verdaderos negativos. Lo que buscamos es que haya una gran cantidad de verdaderos positivos, así como una gran cantidad de verdaderos negativos. Esto querría decir que el modelo es eficiente pues hace una correcta identificación.

Cuando la curva ROC se forma surge, otro concepto importante llamado AUC (*Area Under the Cover*), área bajo la curva, que también ayuda a analizar el resultado de la prueba. Cuando el modelo tiene una gran cantidad de verdaderos positivos así como verdaderos negativos la curva estaría pegada a la esquina superior izquierda, similar al contorno de un triángulo y si calculáramos el área que se encuentra bajo ella obtendríamos $AUC = 1$, que es el resultado ideal que buscamos en una curva ROC sin embargo al ser una estimación los resultados serán cercanos a uno.

El análisis de la curva ROC ayuda a seleccionar modelos óptimos y a descartar modelos que no son tan adecuados, independientemente de la distribución de las clases en la población. Además ayuda a conocer cuál es la exactitud de la prueba.

El punto de corte separa una población de otra de acuerdo a sus características, en este caso, imaginemos que tenemos una población de enfermos y sanos. Supongamos que calculamos la probabilidad de que el individuo pertenezca a la población de enfermos y el resultado es 0.6. Sin embargo tenemos un punto de corte de 0.8 es decir la mínima probabilidad para considerarse parte de la población de enfermos, entonces el

individuo quedaría clasificado como sano ya que está debajo del punto de corte. Existen diversas técnicas para obtener el punto de corte óptimo sin embargo en este trabajo tomaremos el punto de corte de 0.5.

Veamos un ejemplo, vamos a simular el valor de glucosa en la sangre en una serie de individuos, los cuales se van a definir si son sanos o enfermos, después vamos a asignar aleatoriamente una etiqueta de sanos y enfermos. De esta forma vamos a calcular la curva ROC de los datos.

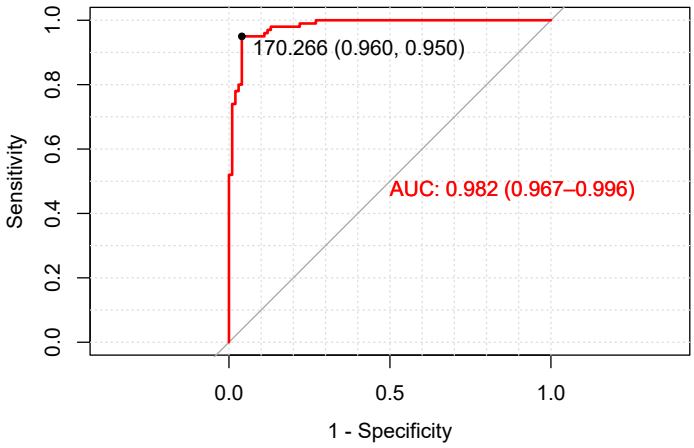

Figura 1.1: Ejemplo Curva ROC buen desempeño.

En la figura 1.1 podemos notar que tenemos un alto nivel de sensibilidad y de especificidad, por lo que se tiene una gran cantidad de verdaderos positivos y verdaderos negativos y en consecuencia una cantidad pequeña de falsos positivos y falsos negativos, lo que querría decir que los individuos están etiquetados correctamente y la clasificación tiene un buen desempeño. Por otro lado en la figura 1.2 podemos ver una curva que indica un mal desempeño.

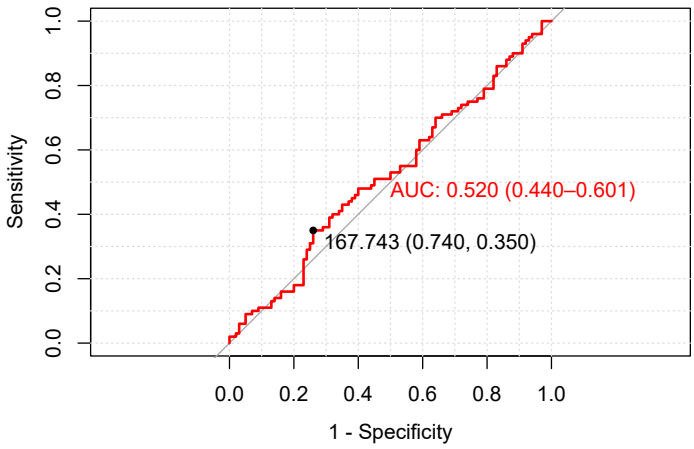

Figura 1.2: Ejemplo Curva ROC mal desempeño.

1.6. Accuracy

El accuracy mide la proporción de verdaderos positivos y verdaderos negativos sobre el total de observaciones, es decir, la proporción en el que el modelo clasifica de manera correcta. En términos de la matriz de confusión, está dado por la siguiente expresión:

$$Accuracy = \frac{VP + VN}{VN + FP + VP + FN}$$

Capítulo 2

Modelos Lineales Generalizados

2.1. Definición de Modelos Lineales Generalizados

Los modelos lineales generalizados (MLG) son una extensión de los modelos lineales en donde se permite que la variable de interés tenga una distribución distinta a la normal o Gaussiana.

Estos modelos se identifican por la relación que existe entre la variable respuesta o componente aleatoria y las variables explicativas por medio de un predictor lineal (combinación lineal de las variables). Esta relación está hecha a través de una función de enlace o función liga. Como ejempos, la variable respuesta puede ser binaria, es decir $Y \in \{0,1\}$, o puede ser un recuento es decir $Y \in \mathbb{N}$, entre otras. Esto establece la necesidad de una función de enlace para relacionar la variable respuesta Y con el predictor lineal β cuyo dominio son los números reales.

Los modelos lineales generalizados se caracterizan por tener tres componentes: componente aleatoria, componente sistemática y función de enlace, a continuación describiremos dichas componentes.

2.1.1. Componente Aleatoria

La componente aleatoria de los MLG, tiene que ver con la variable respuesta Y y su función de probabilidad. Y es una v.a. cuya función de densidad de probabilidad pertenece a la familia exponencial. La familia exponencial incluye diversas distribuciones como la Bernoulli, Binomial, Normal, Poisson, Gamma, Beta y Binomial Negativa, entre otras. La familia exponencial asociada a una v.a. X se caracteriza por tener la siguiente estructura:

$$f(x;\theta) = a(\theta)b(x)\exp\{c(\theta)d(x)\}$$

Donde θ es el parámetro de la distribución.

Respecto a las aplicaciones, si las observaciones de Y son binarias se identifican como 1 (éxito) y 0 (fracaso). También existen aplicaciones en las que la variable respuesta es un recuento y se asigna una distribución Poisson o una distribución binomial negativa.

Por último, si las observaciones de Y son continuas, Un ejemplo de distribución es una distribución normal, en cuyo caso se obtiene el caso particular de los Modelos Lineales.

2.1.2. Componente Sistemática

La componente sistemática de un MLG involucra a las variables explicativas $x_1, x_2, \ldots, x_k$ y en el caso de $k = 1$, se relacionan de la siguiente manera:

$$\eta = \alpha + \beta_1 x_1$$

A esta combinación lineal de las covariables, se le llama predictor lineal.

Cuando la varible respuesta Y es binaria η, debe estar en el intervalo [0,1] ya que estamos hablando de probabilidad, sin embargo la suma de $\alpha + \beta_1 x_1$ está en los números reales, por lo que es necesario una función de enlace que pueda mapear a los números reales en el intervalo [0,1].

La definición generalizada del predictor lineal en caso de tener k variables es la siguiente:

$$\eta = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

2.1.3. Función de Enlace

La función de enlace, relaciona el valor esperado de Y de la componente aleatoria con los predictores lineales incluidos en el modelo, es decir si $E(Y) = \mu$, existe $g(.)$ tal que:

$$g(\mu) = \alpha + \beta_1 x_1$$

De esta forma en caso de tener una regresión logística tendremos:

$$g(\cdot) : \mathbb{R} \longrightarrow [0, 1]$$

Esta función debe ser monótona y diferenciable.

2.2. Modelos de Regresión para Datos Binarios

Las variables respuesta binarias, cuentan con dos categorías: 1 éxito y 0 fracaso. La distribución de Y tiene probabilidades específicas:

$$\begin{aligned} P(Y = 1) &= \pi \\ P(Y = 0) &= 1 - \pi \end{aligned}$$

Para n observaciones independientes $Y_1, \ldots, Y_n$ el número de éxitos tiene una distribución Binomial de parámetros (n, π). Cada observación equivale a una variable binomial con $n = 1$. Esto quiere decir que tendría la siguiente distribución:

$$\begin{aligned} f(Y|\pi) &= \pi^y(1-\pi)^{1-y} \\ &= (1-\pi)\left(\frac{\pi}{1-\pi}\right)^y \\ &= (1-\pi)\exp\left[y\log\left(\frac{\pi}{1-\pi}\right)\right] \end{aligned}$$

Esta distribución, pertenece a la familia exponencial ya que se puede identificarr de la siguiente manera:

$$\begin{aligned} a(\pi) &= 1-\pi \\ b(y) &= \mathrm{I}^{(y)}_{\{0,1\}} \\ c(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) \\ d(y) &= y \end{aligned}$$

2.2.1. Regresión Logística Binaria

La regresión logística [2] cumple con las características del modelo para datos binarios. Nos ayuda a obtener la probabilidad de que cierto suceso ocurra, este suceso representa a la variable respuesta Y, la cual está en función de las variables independientes $X_1,\ldots,X_n$. Al ser un modelo cuya variable respuesta es binaria, de la sección 2.2, obtuvimos el siguiente resultado:

$$f(Y|\pi) = (1-\pi)\exp\left[y\log\left(\tfrac{\pi}{1-\pi}\right)\right]$$

Sin embargo $\pi(x) \in [0,1]$ y $\eta(x) \in \mathbb{R}$ por lo que necesitamos una función de enlace ver sección 2.1.3, que relacione las probabilidades con el coeficiente de regresión β. La función que tiene esta característica es cualquier función de probabilidad. En el caso de la regresión logística binaria, la función de enlace es la función probabilidad acumulada de una variable aleatoria logística.

Para una covariable x, para el caso de una variable explicativa:

$$\begin{aligned} \Rightarrow \pi(x) &= \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \\ \Rightarrow 1-\pi(x) &= 1-\left[\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right] \\ \Rightarrow \frac{\pi(x)}{1-\pi(x)} &= \exp(\alpha+\beta x) \\ \Rightarrow \log\left(\frac{\pi(x)}{1-\pi(x)}\right) &= \alpha+\beta x \end{aligned}$$

A esta función de enlace se le llama logit y es una función liga asociada a la regresión logística.

Si graficamos esta función podemos identificar fácilmente que si un estimador es positivo, entonces cuando x aumenta, la probabilidad aumenta. Por otro lado, si el estimador es negativo, si x aumenta entonces la probabilidad disminuye.

Es muy fácil ver esto mediante la figura 2.1. Del lado izquierdo de la figura en el caso $\beta > 0$ es claro que a medida que x aumenta la probabilidad aumenta y en el lado derecho que representa $\beta < 0$ sucede lo contrario, cuando x aumenta la probabilidad disminuye.

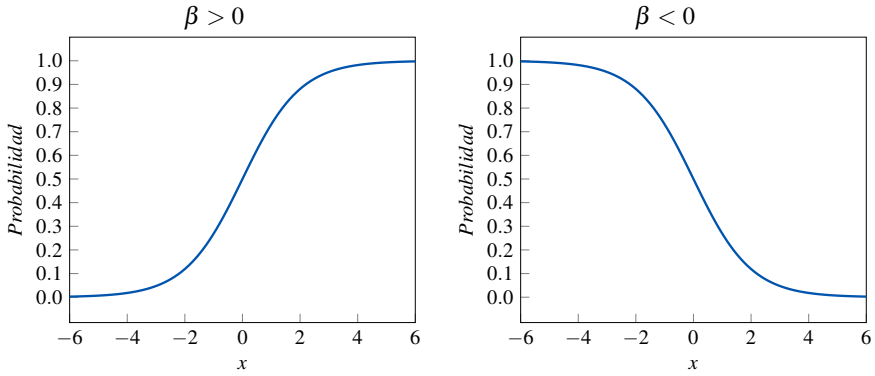

Figura 2.1: Cambio de probabilidad según el estimador en un modelo logístico.

Además esta función de enlace hace que surja un nuevo concepto definido como odd o momio:

$$momio = \frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x)$$

Este concepto es importante debido a que nos ayuda a interpretar los coeficientes de regresión.

Un momio o también llamado ratio se define como el cociente de la probabilidad de que el suceso ocurra entre la probabilidad de que el suceso no se presente.

Por ejemplo supongamos que existe el suceso de no pagar una deuda con probabilidad $\pi(x)$, donde: $\pi(x) = 0.7$.

$$\Rightarrow momio = \frac{0.7}{0.3} = 2.3$$

Lo que indica que es 2.3 veces más probable que este suceso se produzca a que no se produzca. Sin embargo, en este caso recordemos que nuestras probabilidades van a depender de nuestro estimador β y los valores que pueda tomar x. Por ejemplo, sea $\beta = 1.5$ y $\alpha = 0$.

$$\Rightarrow \frac{\pi(x)}{1-\pi(x)} = \exp(1.5x)$$

Si $x = 1$

$$\Rightarrow \frac{\pi(x)}{1-\pi(x)} = \exp(1.5) = 4.48,$$

esto quiere decir que por cada unidad que x aumente, entonces va a ser 4.48 veces más probable que la persona no pague.

Estos resultados se pueden generalizar para n variables explicativas:

$$\begin{aligned}
\Rightarrow \pi(x) &= \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)} \\
\Rightarrow 1-\pi(x) &= 1-\left[\frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}\right] \\
\Rightarrow \frac{\pi(x)}{1-\pi(x)} &= \exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k) \\
\Rightarrow \log\left(\frac{\pi(x)}{1-\pi(x)}\right) &= \alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k
\end{aligned}$$

Como ya mencionamos, para poder obtener la interpretación del modelo es importante primero encontrar los valores de los estimadores. Para estimar los parámetros $\beta_0, \beta_1, ...\beta_k$ podemos utilizar el método de estimación por máxima verosimilitud. Este método consiste en tomar como estimadores de los parámetros desconocidos aquellos que maximizan la verosimilitud de la muestra, es decir, la probabilidad de haber observado precisamente los datos estudiados, o bien los que tienen mayor probabilidad de ocurrir.

La función de verosimilitud $L(\beta)$ se puede calcular como el producto de las funciones de masa de probabilidad que intervienen en el modelo $L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$. La estimación de estos parámetros se puede simplificar aplicando logaritmo, ya que al ser una función creciente tiene el máximo en el mismo punto que la función de verosimilitud:

$$\begin{aligned}
\log(L(\beta)) &= \log\left(\prod_{i=1}^{n}(\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right) \\
&= \log\left(\prod_{i=1}^{n}(1-\pi_i)^1\right)\log\left(\prod \pi_i^{y_i}(1-\pi_i)^{-y_i}\right) \\
&= \log\left(\prod_{i=1}^{n}(1-\pi_i)^1\right)\log\left(\prod_{i=1}^{n}\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}\right) \\
&= \sum_{i=1}^{n}\log(1-\pi_i)+\sum_{i=1}^{n} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)
\end{aligned}$$

Sustituyendo $\pi(x)$ y $\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$:

$$\begin{aligned}
&= \sum_{i=1}^{n} \log\left(1 - \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}\right) + \sum_{i=1}^{n} y_i\left(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k\right) \\
&= \sum_{i=1}^{n} y_i \beta x_i + \log\left(\frac{1}{1+exp(\beta x_i)}\right) \\
&= \sum_{i=1}^{n} y_i \beta x_i - \log\left(1+exp(\beta x_i)\right)
\end{aligned}$$

Nos interesa obtener el valor de los parámetros $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ los cuales se estiman mediante métodos iteractivos ya que no existe una solución única. Existen diversos algoritmos, pero el método Newton-Raphson es muy común para estimar los parámetros en la regresión logística y es el que implementa *RStudio*. Este algoritmo consiste en iterar el valor de β hasta que la verosimilitud se maximice. Utilizamos la siguiente expresión:

$$\beta^{(t+1)} = \beta^{(t)} - \frac{\nabla_\beta l(\beta^t)}{\nabla_{\beta\beta} l(\beta^t)}$$

Calculemos $\nabla_\beta l(\beta^t)$

$$\begin{aligned}
\nabla_\beta l(\beta^t) &= \sum_{i=1}^{n} y_i \beta - \log\left(1+exp(\beta x_i)\right)) \\
&= \nabla_\beta \sum_{i=1}^{n} y_i \beta x_i - log(1+exp(\beta x_i)) \\
&= \nabla_\beta \sum_{i=1}^{n} y_i \beta x_i - log(1+exp(\beta x_i)) \\
&= \nabla_\beta \sum_{i=1}^{n} y_i \beta x_i - \nabla_\beta log(1+exp(\beta x_i)) \\
&= \sum_{i=1}^{n} y_i x_i - \left[\frac{1}{1+exp(-\beta x_i)} exp(\beta x_i) x_i\right] \\
&= \sum_{i=1}^{n} y_i x_i - \left[\frac{1}{1+exp(-\beta x_i)} x_i\right] \\
&= \sum_{i=1}^{n} y_i x_i - (\pi(x_i) x_i) \\
&= \sum_{i=1}^{n} (y_i - (\pi(x_i)) x_i
\end{aligned}$$

Con este resultado ahora podemos calcular $\nabla_{\beta\beta} l(\beta^t)$

$$\begin{aligned}
\nabla_{\beta\beta} l(\beta^t) &= \nabla_\beta \sum_{i=1}^{n} (y_i - (\pi(x_i))x_i \\
&= \sum_{i=1}^{n} \nabla_\beta (y_i - (\pi(x_i))x_i \\
&= \sum_{i=1}^{n} \nabla_\beta - (\pi(x_i))x_i \\
&= \sum_{i=1}^{n} \nabla_\beta - (\frac{1}{1+exp(-\beta x_i)})x_i \\
&= \sum_{i=1}^{n} \left[\frac{1}{1+exp(-\beta x_i)}\right]^2 exp(-\beta x_i)(-x_i)x_i \\
&= -\sum_{i=1}^{n} \left[\frac{exp(-\beta x_i}{1+exp(-\beta x_i)}\right] \left[\frac{1}{1+exp(-\beta x_i)}\right] X_i^T xi \\
&= -\sum_{i=1}^{n} \pi(x_i)(1-\pi(x_i))X_i^T xi
\end{aligned}$$

Transformamos estas expresiones a notación matricial de modo que

$$\begin{aligned}
\nabla_\beta l(\beta^t) &= \sum_{i=1}^{n} (y_i - (\pi(x_i))x_i = X^T(Y-Y) \\
\nabla_{\beta\beta} l(\beta^t) &= -\sum_{i=1}^{n} \pi(x_i)(1-\pi(x_i))X_i^T xiX^T P(1-P)X = X^T WX
\end{aligned}$$

Sustituyendo tenemos:

$$\begin{aligned}
\beta^{(t+1)} &= \beta^{(t)} - \frac{\nabla_\beta l(\beta^t)}{\nabla_{\beta\beta} l(\beta^t)} \\
&= \beta^{(t)} + (X^T WX)^{-1} X^T (Y-Y)
\end{aligned}$$

De esta manera podemos encontrar los valores de $\beta^{(t+1)}$.

Capítulo 3

Reducción de Dimensión y Ponderación de Variables

3.1. Multicolinealidad

Cuando se ajusta un modelo es imprescindible hacer uso e interpretación de los coeficientes individuales de regresión que nos ayuda a hscer inferencias como:

- Predicciones o estimaciones.
- Identificar aquellas variables significativas en el modelo.
- Explorar el efecto de las variables regresoras.

Estas inferencias serían relativamente fáciles si entre las variables regresora del modelo no existiera una dependencia lineal, es decir, fueran ortogonales. Pero frecuentemente por la naturaleza de los datos las variables tienden a no ser ortogonales lo cual no es un impedimento para hacer las inferencias antes mencionadas, el verdadero problema surge cuando la dependencia lineal que hay entre las variables regresoras es casi perfecta a esto se le conoce como multicolinealidad [4].

La multicolinealidad es un problema que se da cuando existen grandes correlaciones entre las variables de nuestras bases de datos, o bien existe cierta asociación entre dichas variables, esto hace que la información no sea confiable.

Las consecuencias de tener multicolinealidad en una base de datos y hacer un análisis de regresión surgen al momento de estimar los coeficientes del modelo (β_j), ya que pueden presentarse los siguientes casos:

- Los coeficientes de regresión pueden tener incongruencia en el signo, o bien que tengan poca precisión, es decir que exista un incremento de la varianza.
- Aceptar la hipótesis nula de que algún parámetro es cero, cuando en realidad la variable sí era relevante.
- Los intervalos de confianza para los estimadores tienden a ser muy grandes.

En este capítulo hablaremos de cómo diagnosticar la multicolinealidad y las técnicas que se utilizan frecuentemente para tratarla en caso de que exista en la base de datos que deseamos modelar. Existen varios libros y artículos [12] y [13] en los que se han propuesto y probado estas técnicas generando buenos resultados.

La multicolinealidad se puede detectar analizando los factores de inflación de la varianza VIF. Sin embargo, en caso de que la base de datos contenga variables categóricas, se deben analizar los factores de inflación de la varianza generalizada GVIF [3], esto se debe a que las correlaciones entre las variables son afectadas.

El GVIF es calculado a través de subconjuntos de las variables regresoras relacionados y de subconjuntos de variables regresoras mudas o dummy. Se plantea un escenario ideal en el que las variables regresoras no tienen correlación y existe una elipse basado en los intervalos de confianza de cada regresora, entre más grande sea el tamaño de dicha elipse existirá más variación en los datos [8].

También podemos fijarnos en la correlación que tienen nuestras variables continuas y en la tendencia de los datos en un gráfico de dispersión.

3.1.1. VIF y GVIF

La varianza estimada para un coeficiente (β_j) se calcula de la siguiente forma:

$$Var(\beta_j) = \frac{S^2}{(n-1)s_j^2}\frac{1}{1-R_j^2}$$

donde S^2 es el error estimado de la varianza, s_j^2 es la varianza muestral de x_j ,R_j^2 es el coeficiente de determinación obtenido al efectuar la regresión lineal múltiple de X_j sobre el resto del las regresoras.

Por lo que el factor de inflación de la varianza está representado por:

$$\frac{1}{1-R_j^2}$$

que es la razón entre la varianza observada y la varianza que se tendría en caso de que X_j estuviera no correlacionada con el resto de regresores del modelo.

Como ya habíamos mencionado el VIF no es aplicable a regresoras con múltiples grados de libertad (df) como variables definidas como factor. El GVIF se define de la siguiente manera, expresión dada por Fox and Monette (1992):

$$GVIF_1 \quad = \quad \frac{detR_{11}detR_{22}}{detR}$$

En donde R_{11} Representa la matriz de correlación del regresor en cuestión R_{22} Representa la matriz de correlación de los regresores restantes y R la matriz de correlación de todos los regresores del modelo. Cuando estudiamos una variable continua entonces $df = 1$ por lo que $GVIF^{\frac{1}{2df}}$ es equivalente a $\sqrt{VIF}$ pero también es equivalente aunque se tenga más de un grado de libertad [9].

En otras palabras, en caso de que las variables que se estén evaluando sean continuas entonces el VIF y GVIF serán iguales ya que sólo se está evaluando un regresor,

existe diferencia cuando se trata de variables categóricas puesto que se tiene regresores dummies correspondientes a cada valor que puede tomar la variable menos una categoría o factor y tienen distintas interpretaciones.

Cuando el valor de VIF es cercano a 1 indica que no existe multicolinealidad. Según algunos textos se tiene un problema de multicolinealidad cuando estos indicadores son mayores a 10 pero también existen artículos en donde el umbral es menor pues indica multicolinealidad cuando el VIF es mayor 5 en este caso particular si VIF es mayor a 10 diremos que existe multicolinealidad grave y si es mayor a 5 entonces la multicolinealidad es débil.

Para el GVIF la interpretación es distinta puesto que es la $GVIF^{\frac{1}{2df}}$ medida en la que se reduciría la precisión de los coeficientes debido a la multicolinealidad por lo que al ser equivalente a $\sqrt{VIF}$ buscaríamos $GVIF^{\frac{1}{2df}} < 2$ para descartar multicolinealidad débil y $GVIF^{\frac{1}{2df}} < 4$ para descartar multicolinealidad grave dependiendo cuál sea el criterio que deseamos utilizar.

Debido a estos problemas existen diversas técnicas para lidiar con la multicolinealidad y para conservar las variables más significativas de nuestro modelo como: análisis de componentes principales, regresión Ridge y Lasso, entre otras.

3.2. Componentes Principales

Esta técnica se utiliza para estudiar las relaciones que existen entre p variables correlacionadas, reducir la dimensión y así poder obtener un subconjunto de variables no correlacionadas. Este nuevo conjunto está hecho mediante combinaciones lineales del conjunto original y además contiene la mayor información posible, a este subconjunto se le llama componentes principales.

Al querer obtener un menor número de variables, entonces se tendría una reducción de dimensión en el conjunto de datos.

Existen diversos métodos para obtener las componentes principales. Pearson fue el primero en estudiar este análisis que consistía en buscar un subespacio que minimizara la suma de mínimos cuadrados, más tarde fue retomado por Hottelling quien propuso buscar una combinación lineal de aquellas variables que maximizaran la variabilidad utilizando multiplicadores de Lagrange [6]. Existen otros métodos como el de coordenadas principales y métodos biplot.

En esta sección nos enfocaremos en el método de Hottelling por lo que buscaremos una combinación lineal que contenga la mayor variabilidad posible.

Supongamos que tenemos dos variables, es decir, dos dimensiones y un conjunto de datos y nos interesa saber cuáles son sus componentes principales, al tener dos dimensiones entonces tendremos dos componentes principales, veamos la figura 3.1.

Para analizar los datos debemos buscar ciertas características que nos ayudarán a entender mejor el concepto de componentes principales, una de estas características es la covarianza muestral[1] ya que de esta forma sabremos cuál es la orientación de

[1]La covarianza nos permite identificar el comportamiento de dos variables aleatorias X,Y es decir la variación de manera conjunta respecto a sus medias, está definida por $cov(x,y) = \frac{\sum_{i=1}^{n}(x-x_i)(y-y_i)}{n}$ y tiene la siguiente interpretación:

los datos en el plano y cuáles son los valores propios que nos brinda la dirección de nuestros datos.

En este caso los datos tienen una covarianza positiva por la posición de los datos e indica que los valores propios también tienen una dirección positiva. Sin embargo. no solo nos interesa saber la dirección sino el valor propio asociado a ese vector ya que determinará la longitud del vector.

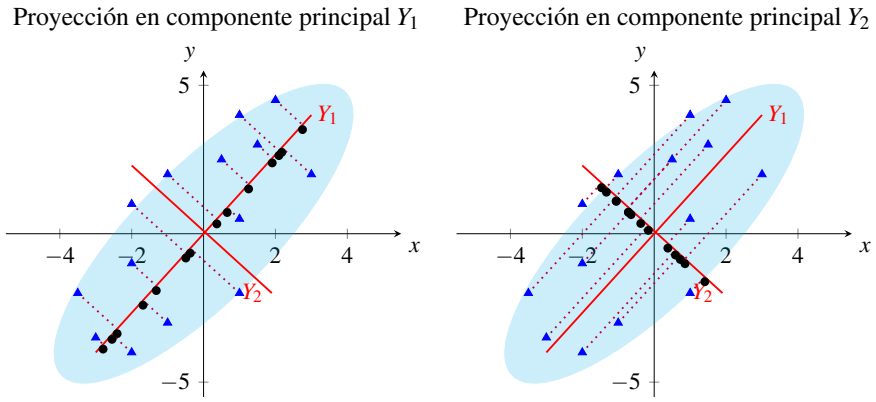

Figura 3.1: Representación geométrica componentes principales

.

En el lado izquierdo se encuentra una representación del conjunto de datos respecto a las variables originales X_1 y X_2, proyectadas sobre el vector propio o eigenvector de la componente principal Y_1. En el lado derecho tenemos la misma representación esta vez proyectando los datos sobre el vector propio de la componente Y_2. El objetivo es obtener la localización de los datos, tomando como ejes las componentes principales (Y_1, Y_2) con la condición de que Y_1 y Y_2 sean ortogonales esto se puede obtener mediante una combinación lineal de las variables originales, de esta forma podríamos proyectar el conjunto de datos en sus dos componentes principales y analizar en cuál componente existe mayor variabilidad. En este caso si analizamos las dos componentes podemos observar que existe mayor dispersión de los datos proyectados en la componente Y_1 que en la componente Y_2 por lo que tendríamos mayor varianza en la componente Y_1. Esta sería la componente principal que utilizaríamos si quisiéramos reducir la dimensión.

Esto se puede generalizar hasta n dimensiones dependiendo del número de variables con las que contemos. Consideremos un conjunto de datos de n observaciones con dimensión p de x_t variables $t = 1, 2, \ldots, n$, se desea obtener un nuevo conjunto de variables y_t no correlacionadas entre sí, que maximicen la varianza. Al hacer esto se garantiza que en el subconjunto de datos contenga la mayor variabilidad del conjunto de datos original y esa es la razón por la cual se elige el subconjunto con mayor varianza.

$cov(X,Y) < 0$ indica una relación negativa, cuando X aumenta Y disminuye.
$cov(X,Y) > 0$ indica una relación positiva, cuando X aumenta Y aumenta.
$cov(X,Y) = 0$ no existe relación entre las variables.

La variable y_t es una combinación lineal de x_t, es decir buscamos encontrar la coordenada del individuo i con $i = 1,\ldots,n$ en la componente j, con $j = 1,\ldots,n$:

$$\begin{aligned} Y_j &= a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p \\ y_{ij} &= a_{j1}x_{i1} + a_{j2}x_{i2} + \cdots + a_{jp}x_{ip} \\ &= a_j{}'X \end{aligned}$$

donde a_j=$(a_{j1}, a_{j2}, \ldots, a_{jp})'$ es un vector de constantes.

También podemos expresarlo de forma matricial, por medio de una multiplicación de matrices de modo que:

$$\begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} a_{j1} \\ \vdots \\ a_{jp} \end{pmatrix}$$

La forma más simple de reducir la dimensión es eligiendo sólo una observación del conjunto, pero esto generaría un sesgo en la información, o bien, aumentando el valor del vector de constantes ya que de esta forma aseguraríamos el aumento en la varianza pero no garantizaríamos ortogonalidad.

Es por ello que añadiremos la siguiente restricción para que los vectores propios tengan restricción de uno:

$$\|a'\| = a_j{}'a_j = \sum_{k=1}^{p} (a_{kj})^2 = 1$$

Se busca obtener a_j que se maximice la varianza de y_j respetando la ortogonalidad dada por la restricción anterior [7] que también significa que la norma del vector es igual a uno, es decir:

$$\max_{\{a:\|a'\|=1\}} Var(aX) = \max_{\{a:\|a\|=1\}} a'Var(X)a$$

De modo que la **primera componente principal** representa la mayor varianza y la segunda componente principal se obtiene calculando los valores a_2 de tal forma que y_1 y y_2 sean ortogonales.

Debido a que buscamos maximizar la varianza, utilizaremos los multiplicadores de Lagrange, que es el método habitual para maximizar una función sujeta a restricciones.

Para la primera primer componente principal deseamos encontrar el valor de a_1, entonces tenemos:

$$Var(y_1) = Var(a_1{}'x) = a_2{}'\Sigma a_1$$

donde Σ es la matriz de covarianzas [2]. Nuestra incógnita es a_1, construyendo la función $\Gamma(a_1)$:

[2]La matriz de covarianzas contiene la covarianza entre los elementos de un vector y de forma matricial se define como $\Sigma = E\left[(X - E[X])(X - E[X])^\top\right]$.

$$\Gamma(a_1) = a_1'\Sigma a_1 - \lambda(a_1'a_1 - 1)$$

derivamos e igualamos a cero para encontrar el máximo:

$$\begin{aligned} \frac{\partial\Gamma(a_1)}{\partial a_1} &= 2\Sigma a_1 - 2\lambda I a_1 \\ \Rightarrow (\Sigma - \lambda I)a_1 &= 0 \end{aligned}$$

Ahora, para que este sistema de ecuaciones tenga solución la matriz debe ser singular[3], que se reduce a que su determinante sea cero:

$$|(\Sigma - \lambda I)a_1| = 0$$

De esta forma λ es un valor propio de la matriz de covarianzas $\sum$ de orden p . En general esta ecuación $|(\sum - \lambda I)a_1| = 0$ tendrá p raices distintas $\lambda_1, \lambda_2, \ldots, \lambda_p$ si es definida positiva, ordenadas de mayor a menor $\lambda_1 > \lambda_2 > ... > \lambda_p$. Si desarrollamos la expresión anterior tenenemos:

$$\begin{aligned} (\Sigma - \lambda I)a_1 &= 0 \\ \Sigma a_1 - \lambda I a_1 &= 0 \\ \Sigma a_1 &= \lambda I a_1 \end{aligned}$$

Sustituyendo $\sum a_1$ en la ecuación:

$$\begin{aligned} Var(y_1) &= Var(a_1'x) = a_1'\Sigma a_1 \\ &= a_1'\lambda I a_1 \\ &= a'a_1\lambda \\ &= \sum_{k=1}^{p} a_{kj}^2\lambda \\ &= 1\lambda = \lambda \end{aligned}$$

Por lo que para maximizar la varianza tenemos que tomar λ_1. Debido a que están ordenados de mayor a menor, es el mayor valor propio que corresponde al vector a_1'.

Para obtener la segunda componente principal utilizaremos el mismo método, aunque agregaremos la restricción de que y_1 y y_2 estén no correlacionadas, es decir, que la covarianza entre estas variables sea cero:

$$\begin{aligned} Cov(y_1, y_2) &= Cov(a_1'x, a_2'x) \\ &= a_2'E\left[E[(x-\mu)(x-\mu)^\top]\right]a_1 \\ &= a_2'\Sigma a_1 \end{aligned}$$

en donde μ es la media, entonces podemos concluir que $Cov(y_1, y_2) = 0$ si y solo si $a_2'\Sigma a_1 = 0$ ahora recordemos que $\Sigma a_1 = \lambda a_1$, si sustituimos este resultado en la ecuación anterior tenemos:

$$a_2'\Sigma a_1 = a_2'\lambda a_1 = 0$$

[3]Una matriz es singular si y solo si su determinante es nulo.

Sin embargo, $\lambda > 0$ por lo que para cumplir $Cov(y_1, y_2) = 0$, basta con cumplir la siguiente expresión:

$$a_2{}' a_1 = 0$$

Por lo que para maximizar $var(y_2) = a_2{}' a_2$ tenemos las siguientes restricciones:

$$\begin{aligned} a_2{}' a_2 &= 1 \\ a_2{}' a_1 &= 0 \end{aligned}$$

Haremos el mismo proceso que realizamos para la primer componente principal. Utilizando multiplicadores de Lagrange. Este se compone de la función a maximizar que en este caso es la varianza de y_2 y las dos condiciones antes mencionadas:

$$\Gamma(a_2) = a_2{}' \Sigma a_2 - \lambda (a_2{}' a_2 - 1) - \gamma (a_1{}' a_2)$$

Donde Σ es la matriz de covarianzas. Derivamos esta función y obtenemos:

$$\frac{\partial \Gamma(a_2)}{\partial a_2} = 2\Sigma a_2 - 2\lambda a_2 - \gamma a_1 = 0$$

Ahora multipliquemos por $a_1{}'$ y recordemos que

$$\begin{aligned} a_1{}' a_2 = a_1{}' a_2 &= 0 \\ a_1{}' a_1 &= 1 \end{aligned}$$

De esta manera sustituimos en la ecuación

$$\begin{aligned} 2a_1{}' \Sigma a_2 - 2\lambda a_1{}' a_2 - \gamma a_1{}' a_1 &= 0 \\ 2\Sigma a_1{}' a_2 - 2\lambda (0) - \gamma (1) &= 0 \\ 2\Sigma a_1{}' a_2 - \gamma &= 0 \end{aligned}$$

Ahora sabemos que $cov(y_1, y_2) = 0$ por lo que

$$\begin{aligned} 2\Sigma a_1{}' a_2 &= \gamma \\ 2cov(y_1, y_2) &= \gamma \\ \Rightarrow \gamma &= 0 \end{aligned}$$

De esta manera si regresamos a la expresión $\frac{\partial \Gamma(a_2)}{\partial a_2}$ podemos concluir que

$$\begin{aligned} \frac{\partial \Gamma(a_2)}{\partial a_2} &= 2\Sigma a_2 - 2\lambda a_2 - \gamma a_1 \\ &= 2\Sigma a_2 - 2\lambda a_2 \\ \Rightarrow (\Sigma - \lambda I) a_2 &= 0 \end{aligned}$$

Siguiendo el mismo razonamiento que utilizamos para calcular la primera componente es fácil deducir que $var(y_2) = \lambda_2$, siendo el segundo valor propio más grande

de la matriz de covarianzas y a_2 su vector propio asociado. Con esta misma lógica podemos concluir que la varianza de cada componente principal corresponde al valor de cada valor propio de la matriz de covarianzas, es decir $var(y_1) = \lambda_1$, $var(y_2) = \lambda_2$, ..., $var(y_n) = \lambda_n$.

Como ya habíamos mencionado podemos ver el cálculo de componentes principales y de forma matricial multiplicando el vector de variables originales x por la matriz formada por los autovectores A que también se puede ver como una transformación lineal. De modo que:

$$\begin{aligned} y &= Ax \\ \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix} &= \begin{pmatrix} a_{11} & \dots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{np} \end{pmatrix} \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \end{aligned}$$

Y dado que $var(y_i) = \lambda_i$ y las variables son ortogonales es decir están no correlacionadas la matriz de covarianzas es:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_p \end{pmatrix}$$

Donde Λ representa a una matriz diagonal con los valores propios, definido esto existen una serie de propiedades [6]. Lo primero que se puede deducir es:

$$\Sigma A = \Lambda A$$

Puesto que

$$var(\Lambda) = var(Y) = Avar(X)A = Avar(X)A = A'\Sigma A$$

Donde $\Sigma = Cov(X,X)$ la matriz de covarianza de las variables originales, $\Lambda = Cov(Y,Y)$ la matriz de covarianza de los componentes principales también correspondiente a la diagonal con los autovalores, A la matriz ortogonal que contiene los autovectores. También puede ser representado de la siguiente manera

$$\Sigma = A\Lambda A'$$

expresión que utilizaremos para definir la varianza explicada.

3.2.1. Varianza explicada y selección del número de componentes

Una vez que tenemos el cálculo de la varianza nos interesa saber cuál es la varianza que aporta cada componente principal o bien analizar la varianza que se pierde al hacer esta reducción de dimensión. Sabemos que el total de la varianza es la suma de cada

autovalor de la matriz Λ que se puede expresar mediante el operador traza[4] de tal manera que:

$$\sum_{i=1}^{p} var(y_i) \quad = \quad tr(\Lambda)$$

Y dado que $\Lambda = A'\Sigma A$ y por propiedades del operador traza entonces

$$tr(\Lambda) = tr(\Sigma) = \sum_{i=1}^{p} var(y_i) = \sum_{i=1}^{p} \lambda_i$$

De esta manera podemos afirmar que el total de la varianza corresponde a la varianza de las variables originales por lo que podemos calcular el porcentaje de varianza que aporta cada componente principal

$$\frac{var(y_i)}{\sum_{i=1}^{p} var(x_i)} \quad = \quad \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

También se puede calcular el porcentaje acumulado de la varianza de m componentes principales con $m < p$

$$\frac{\sum_{i=1}^{m} var(y_i)}{\sum_{i=1}^{p} var(x_i)} \quad = \quad \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

La varianza explicada acumulada es de gran utilidad ya que nos ayuda a decidir cuál es el número mínimo de componentes que acumula la mayor variabilidad aunque exista una reducción de dimensión. Este número mínimo de componentes puede explicar los datos por lo que debe conservarse y por el contrario se deben descartar las componentes que no tengan una varianza relevante. El criterio del porcentaje de varianza acumulada puede variar pero si conservamos aquellas componentes que acumulan al menos el 90 % podemos asegurar una representatividad en los datos.

3.2.2. Componentes principales para la matriz de correlación

Otro método para calcular las componentes principales es utilizando la matriz de correlación R esto depende de que las variables estén estandarizadas. Esto es recomendable cuando las varianzas son muy grandes o si las unidades de las variables no son comparables, de esta manera podemos tener una representación absoluta. Por tanto, si hacemos la media cero y la varianza uno se da la misma importancia a las variables originales. Si este es el caso la matriz de covarianzas será igual a la matriz de correlaciones de tal modo que:

$$\Sigma = R$$

[4]El operador traza está definido como la suma de los elementos de la diagonal principal de una matriz cuadrada A, $tr(A) = a_{11} + a_{22} + \cdots + a_{nn}$.

donde

$$R = \begin{pmatrix} 1-\lambda & \rho & \dots & \rho \\ \rho & 1-\lambda & \dots & \rho \\ \vdots & \vdots & \dots & \vdots \\ \rho & \rho & \dots & 1-\lambda \end{pmatrix} = 0$$

Donde ρ es el coeficiente de correlación.

Veamos cuál es la relación de la matriz de los datos y la matriz de covarianza de los datos sin estandarizar.

Supongamos que queremos obtener las variables estandarizadas Z a partir de las variables originales X por lo que este vector se tendría que estandarizar restando su media y entre su desviación estándar esto se puede expresar matricialmente de la siguiente manera:

$$Z = D^{\frac{1}{2}}(X-\mu)$$

Donde D es una matriz diagonal con entradas las varianzas muestrales. Como ya habíamos visto si el vector está estandarizado la matriz de covarianzas es igual a la de correlaciones por lo que se cumple lo siguiente:

$$R = Corr(X) = Corr(Z) = Cov(Z) = Cov[D^{\frac{1}{2}}(X-\mu)(X-\mu)^{\top}D^{\frac{1}{2}}] = D^{\frac{1}{2}}\Sigma D^{\frac{1}{2}}$$

Por esta propiedad al momento de resolver el determinante para encontrar los valores propios y ρ es positiva:

$$|\Sigma - I\lambda| = |R - I\lambda| = 0$$

Desarrollando:

$$\begin{aligned} \Rightarrow \lambda_1 &= 1+(p-1)\rho \\ \lambda_2 &= ,\dots,= \lambda_p = 1-\rho \end{aligned}$$

Por otro lado los vectores propios tendrá una raíz máxima λ_1 y otra de rango $1-p$ por lo que los vectores propios se pueden expresar como:

$$\begin{aligned} a_1 &= (1,\dots,1)'\sqrt{p} \\ a_i &= (1,1,1,\dots,1-(i-1),0,\dots,0)'\sqrt{((i-1)*i)^{-1}} \\ a_p &= (1,1,1,\dots,1-(p-1))\prime\sqrt{((p-1)*p)^{-1}} \end{aligned}$$

Y dado que la diagonal de la matriz de correlaciones es uno entonces la proporción de varianza explicada se vería de la siguiente forma:

$$\frac{var(y_i)}{\sum_{i=1}^{p} var(x_i)} = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} = \frac{\lambda_i}{p}$$

Veamos un ejemplo en dos dimensiones usando ambos métodos la matriz de covarianzas y la matriz de correlaciones. Supongamos que tenemos un conjunto de clientes con dos variables la edad E, y el límite de crédito LC otorgado en un banco:

Datos originales										
E	27	55	40	38	42	30	25	50	45	33
LC	150	240	180	200	250	150	135	280	250	217
Datos estandarizados										
E	-1.16	1.67	0.15	-0.05	0.35	-0.86	-1.36	1.16	0.65	-0.55
LC	-1.10	0.69	-0.50	-0.10	0.89	-1.10	-1.39	1.49	0.89	0.23

Con este conjunto de datos podemos calcular tanto la matriz de covarianzas como la matriz de correlaciones:

$$\Sigma = \begin{pmatrix} 97.6111 & 426 \\ 426 & 2515.956 \end{pmatrix} \quad , \quad R = \begin{pmatrix} 1 & 0.8596 \\ 0.8596 & 1 \end{pmatrix}$$

En donde Σ es matriz de covarianza de los datos originales y R es la matriz de correlaciones de los datos estandarizados aunque también coincide con la matriz de correlaciones de los datos originales.

Ahora calculemos los valores propios para la matriz de covarianzas de los datos originales:

$$\Sigma - I\lambda = \begin{pmatrix} 97.6111 - \lambda & 426 \\ 426 & 2515.956 - \lambda \end{pmatrix}$$

Igualamos el determinante a cero

$$\begin{vmatrix} 97.6111 - \lambda & 426 \\ 426 & 2515.956 - \lambda \end{vmatrix} = 0$$

Desarrollamos y simplificamos:

$$\begin{aligned} (97.6111 - \lambda)(2515.956 - \lambda) - (426)(426) &= 0 \\ \lambda^2 - 2613.51\lambda + 64101 &= 0 \end{aligned}$$

Resolvemos la ecuación cuadrática:

$$\begin{aligned}\lambda_1 &= \frac{-(-2613.51)+\sqrt{(-2613.51)^2-4(64101)}}{2} = 2588.80\\ \lambda_2 &= \frac{-(-2613.51)-\sqrt{(-2613.51)^2-4(64101)}}{2} = 24.76\end{aligned}$$

Por lo que la solución estaría dada por:

$$\begin{aligned}\lambda_1 &= 2588.80\\ \lambda_2 &= 24.76\end{aligned}$$

Una vez que tenemos los valores propios podemos calcular los vectores propios restando λ_i a la diagonal de la matriz de covarianzas y resolviendo el sistema, para λ_1 el procedimiento sería el siguiente:

$$\Sigma - I\lambda_1 = \begin{pmatrix}97.6111-\lambda_1 & 426\\ 426 & 2515.956-\lambda_1\end{pmatrix} = \begin{pmatrix}-2491.19 & 426\\ 426 & -72.84\end{pmatrix}$$

Reducimos por eliminación gaussiana:

$$\begin{pmatrix}-2491.19 & 426\\ 426 & -72.84\end{pmatrix} \xrightarrow{\frac{f1}{-2491.1}} \begin{pmatrix}1 & -0.1710\\ 426 & -72.78\end{pmatrix} \xrightarrow{f2-f1\times 426} \begin{pmatrix}1 & -0.1710\\ 0 & 0\end{pmatrix}$$

Una vez que tenemos la mínima expresión podemos multiplicar por las incógnitas:

$$\begin{aligned}\begin{pmatrix}1 & -0.1710\\ 0 & 0\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix} &= a_1 - 0.1710a_2\\ &\Rightarrow\ a_1 = 0.1710a_2\\ &\Rightarrow\ \begin{pmatrix}a_1\\ a_2\end{pmatrix} = \begin{pmatrix}0.1710a_2\\ ya_2\end{pmatrix}\\ &\Rightarrow\ a_2\begin{pmatrix}0.1710\\ 1\end{pmatrix}\end{aligned}$$

El procedimiento para encontrar el valor de a_2 es análogo sólo que para reducir el sistema usaríamos λ_2. Los vectores propios estarían dados por

$$\begin{aligned}a_1 &= (0.1710, 1)\\ a_2 &= (-1, 0.1710)\end{aligned}$$

La simetría se debe a que a_1 y a_2 son ortogonales.

Para las variables estandarizadas el cálculo para los vectores propios y valores propios es similar con la diferencia que utilizaremos la matriz de correlación ya que es igual a la de covarianzas.

$$R - I\lambda = \begin{pmatrix}1-\lambda & 0.8596\\ 0.8596 & 1-\lambda\end{pmatrix}$$

Calculamos el determinante

$$\begin{aligned} |R - I\lambda| &= 0 \\ \lambda^2 - 2\lambda + 0.261088 &= 0 \end{aligned}$$

Y resolviendo la ecuación tenemos:

$$\begin{aligned} \lambda_1 &= 1.8596 \\ \lambda_2 &= 0.1404 \end{aligned}$$

Como esta R es positiva podemos usar las expresiones. Por lo que:

$\lambda_1 = 1 + (p-1)\rho = 1 + (2-1)0.8596 = 1.8596 \lambda_1 = 1 - \rho = 1 - 0.8596 = 0.1404$

Y siguiendo el mismo procedimiento que usamos para encontrar el vector propio a_1 en el ejemplo de las variables originales encontramos que:

$$\begin{aligned} a_1 &= (0.7071, 0.7071) \\ a_2 &= (-0.7071, 0.7071) \end{aligned}$$

También podemos usar las expresiones para vectores propios de modo que:

$$\begin{aligned} a_1 &= (1,1)\sqrt{2} = (2^{\frac{1}{2}}, 2^{\frac{1}{2}}) = (0.7071, 0.7071) \\ a_2 &= (1, -(2-1))\prime(\sqrt{((p-1)*p)^{-1}} \\ &= (1,-1)(\sqrt{(2-1)*2})^{-1} = (1,-1)\frac{\sqrt{2}}{2} = (0.7071, -0.7071) \end{aligned}$$

Podemos observar que ni los valores propios ni los vectores propios son iguales a pesar que son los mismos datos con la diferencia de estandarización pero tienen las mismas propiedades.

3.3. Métodos de regulación y ponderación de variables

El Ridge y LASSO son técnicas que se utilizan para reducir la complejidad del modelo por lo que son conocidas como *shrinkage' methods* y son especialmente útiles cuando tenemos problemas de multicolinealidad. Ambas técnicas son parecidas ya que ambas añaden una penalización a los estimadores del modelo, sin embargo la diferencia radica en la definición de la penalización. En la regresión LASSO *(least absolute shrinkage and selection operator)* los estimadores que aportan mayor información se reducen a cero mientras que en la regresión Ridge estos estimadores son muy cercanos a cero.

3.3.1. LASSO

Este método consiste en ajustar el modelo de regresión con p covariables aplicando restricciones o regularizaciones a los estimadores de modo que el valor de β dependerá de la forma de la región de restricción.

Cuando ajustamos un modelo de regresión logística requerimos estimar los cieficientes por el método de máxima verosimilitud pero al incluir una penalizaciñon lo que se desea es minimizar la siguiente expresión:

$$L(\beta) + \lambda \sum_{j=1}^{p} |\beta|$$

donde $L(\beta)$ es la función de verosimilitud, $\sum_{j=1}^{p} |\beta| = ||\beta||$ es la restricción, $||.||_1$ denota a la norma 1 o l_1

Para la regresión LASSO los coeficientes que aportan menos información son forzados a tener un estimador igual a cero. Por lo que sólo las variables más significativas serán conservadas en el modelo.

Geométricamente se podría expresar de la siguiente forma:

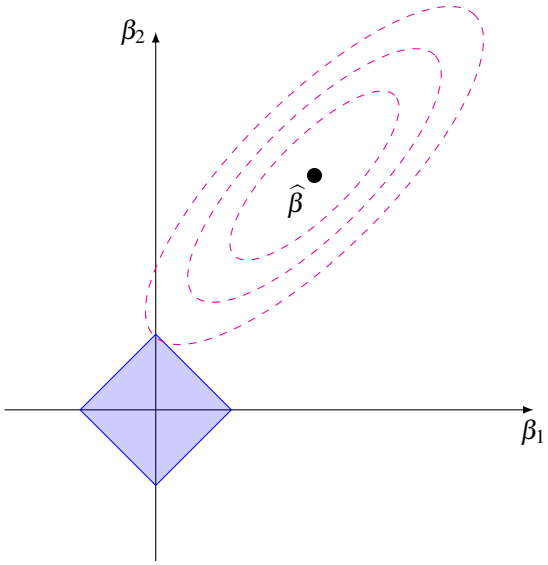

Figura 3.2: Representación geométrica de Lasso.

En la figura 3.2 podemos observar la idea intuitiva de la regresión Lasso, se observa cuál podría ser el valor de $\widehat{\beta}$ si se tuviera la restricción de la norma, los valores posibles de $\widehat{\beta}$ están dados por las líneas rojas punteadas que al intersecarse con la esquina del rombo en morado representando que la norma $L1$ resulta cero para β_1 dado que la solución ocurrió en una esquina, es por ello que en Lasso los coeficientes menos significativos se hacen cero.

Notemos que en estas penalizaciones está involucrado el parámetro libre λ el cual se selecciona de manera que el error cuadrático medio sea minimizado. Por lo regular se utiliza validación cruzada para encontrar dicho parámetro.

3.3.2. Group LASSO

Este método es similar a Lasso con la particularidad de que en lugar de evaluar una sola variable y determinar si el coeficiene es cero, se evalúa un grupo de variables. Esta técnica ayuda a tomar en cuenta las variables categóricas puesto que los grupos están preasignados a cada covariable, sin embargo en caso de que existan variables categóricas se hace un grupo formado por los valores de la variable categórica. Es decir nos ayuda a reducir a cero un grupo completo de variables.

La restricción se expresa de la siguiente forma:

$$L(\beta) + \lambda \sum_{k=1}^{k} ||\beta_{G_k}||_2$$

En donde $G = G_1, \ldots, G_K$, es una partición del conjunto original de datos.

Veamos un ejemplo geométrico en tres dimensiones [5] en la figura 3.3 en donde podemos observar en el primer cuadrante los vectores β_{11} y β_{12} y el escalar β_2 en este caso esta figura corresponde a a la región de restricción y si cortamos con planos podemos ver como se verían las regiones desde distintos ángulos. Cuando una variable perteneciente a algún grupo tiene su solución óptima en las esquinas de esta figura, entonces todas las variables que componen a ese grupo tendrían un coeficiente de cero en cambio si todas las variable de este grupo tienen soluciones óptimas fuera de las esquinas es decir en el plano cuyos ejes corresponden a (β_{11}, β_{12}) entonces se estimaría un coeficiente para cada una de estas variables respectivamente.

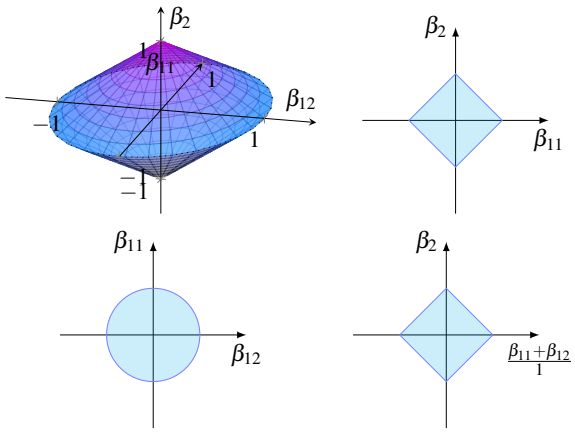

Figura 3.3: Representación gráfica Group-Lasso.

3.3.3. Ridge

En el caso de la regresión Ridge se tiene como objetivo penalizar los coeficientes que aportan menos información, son forzados a tener un coeficiente muy cercano a cero por lo que se estiman los coeficientes por máxima verosimilitud pero usamos una restricción diferente a la de LASSO. Se desea minimizar la siguiente expresión:

$$L(\beta) + \lambda \sum_{j=1}^{p} \beta^2$$

donde $L(\beta)$ es la función de verosimilitud, $\sum_{j=1}^{p} \beta^2 = ||\beta||_2$ es la restricción, $||.||_2$ denota a la norma 2 o l_2 Se puede observar su representación geométrica en la figura 3.4 en donde es claro que debido a que utilizamos una penalización l_2 entonces la región de restricción estará dada por un círculo.

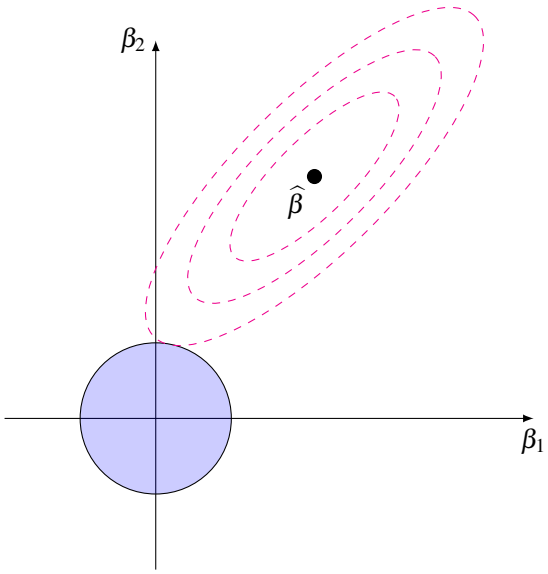

Figura 3.4: Representación geométrica Ridge.

Entonces la única diferencia entre los métodos Ridge y Lasso radica en la penalización si comparamos las figuras 3.4 y 3.2 podemos observar que en la figura 3.2 la restricción es representada por un rombo y en la figura 3.4 es representada por un círculo como consecuencia los valores de $\widehat{\beta}$ al intersecarse con el contorno del círculo tendrá infinitas soluciones por lo que el valor de $\widehat{\beta}$ nunca será cero sino un valor muy cercano a cero.

Capítulo 4

Aplicaciones en Riesgo de Crédito

Las aplicaciones de la regresión logística así como las otras técnicas que hemos descrito como regresión Lasso y Ridge, o análisis de componentes principales, son muy versátiles ya que se pueden utilizar en diversos ámbitos, por ejemplo en las ciencias de la salud o en análisis de riesgo de crédito. Debido a que sus resultados nos permiten hacer análisis explicativos y predictivos de las variables que son influyentes en la variable de estudio.

Para este capítulo utilizaremos una base de datos sobre los casos de clientes que no pagaron su deuda en un banco de Taiwan, es decir, nos enfocaremos en estudiar el riesgo de crédito, ya que ayuda a clasificar o predecir el riesgo de que un cliente no pague y de esta manera se pueden tomar medidas preventivas. En este caso específico es importante este tipo de modelos, puesto que Taiwan sufrió una crisis debido a la deuda de crédito.

A continuación vamos a describir las variables de nuestra base de datos:

- *Y default payment next month*: Pagó a tiempo (sí=0, no=1).
- *LIMIT_BALL*: monto del crédito.
- *SEX*: género (1=male, 2=female).
- *EDUCATION*: grado de escolaridad (escuela de posgrado=1, universidad=2, preparatoria=3, otros=4).
- *MARRIAGE*: estado civil (casado=1, soltero=2, otro=3).
- *AGE*: edad.
- *PAY_0*: historial de los pagos en septiembre.

  El historial tiene la siguiente escala:

  -1 = pagó a tiempo.

1 = retardo de pago por 1 mes.
2=retardo de pago por 2 meses.
3=retardo de pago por 3 meses.
4=retardo de pago por 4 meses.
5=retardo de pago por 5 meses.
6=retardo de pago por 6 meses.
7=retardo de pago por 7 meses.
8=retardo de pago por 8 meses.
9=retardo de pago por 9 meses o más.

- PAY_2: historial de los pagos en agosto. (escala igual a la del mes septiembre)
- PAY_3: historial de los pagos en julio. (escala igual a la del mes septiembre).
- PAY_4: historial de los pagos en junio. (escala igual a la del mes septiembre).
- PAY_5: historial de los pagos en mayo. (escala igual a la del mes septiembre).
- PAY_6: historial de los pagos en abril. (escala igual a la del mes septiembre).
- $BILL\_AMT1$: estado de cuenta en septiembre 2005.
- $BILL\_AMT2$: estado de cuenta en agosto 2005.
- $BILL\_AMT3$: estado de cuenta en julio 2005.
- $BILL\_AMT4$: estado de cuenta en junio 2005.
- $BILL\_AMT5$: estado de cuenta en mayo 2005.
- $BILL\_AMT6$: estado de cuenta en abril 2005.
- PAY_AMT1: monto del pago previo a septiembre 2005.
- PAY_AMT2: monto de pago previo a agosto 2005.
- PAY_AMT3: monto de pago previo a julio 2005.
- PAY_AMT4: monto de pago previo a junio 2005.
- PAY_AMT5: monto de pago previo a mayo 2005.
- PAY_AMT6: monto de pago previo a abril 2005.

A grandes rasgos nuestra base de datos contiene información del estado de cuenta de los clientes, así como el importe de pago mensual. Es decir, el monto que pagó el cliente de acuerdo a su deuda. Además incluye información acerca de la puntualidad en los pagos del cliente, contar con estas referencias nos ayuda a estudiar el comportamiento de los clientes y poder predecir el comportamiento, de futuros clientes potenciales. En este capitulo aplicaremos diversos modelos estadísticos con el fin de encontrar el que se ajusta mejor a nuestros datos.

4.1. Análisis Descriptivo

Iniciaremos el estudio de nuestra base de datos mediante un análisis exploratorio de los datos. Esto se realiza para identificar pistas clave que nos ayuden a manejar de forma correcta nuestros datos y también para familiarizarnos y conocer el comportamiento general de los tipos de variables con las que estamos tratando, es decir, si son categóricas o continuas. Además, es muy útil para identificar qué variables están correlacionadas o bien nos están brindando la misma información.

Observemos la tabla de clasificación 4.1 de las variables. En este caso en particular se dividieron las variables en categórica o continua según sea el caso y se hizo otra división para analizar las características del cliente que se relacionan con atributos personales y las características que se relacionan con el crédito, es decir aquellas que aportan información acerca del comportamiento del crédito como información de los pagos, estados de cuenta, etc.

Clasificación	Continua	Categórica
Característica cliente	▪ AGE	▪ EDUCATION ▪ SEX ▪ MARRIAGE
Característica crédito	▪ LIMITBALL ▪ BILLAMT1-BILLAMT6 ▪ PAYAMT1-PAYAMT6	▪ PAY0-PAY6

Tabla 4.1: Clasificación de variables.

Es importante hacer esta clasificación ya que la representación gráfica o tablas a usar depende del tipo de variable que estamos estudiando.

Ahora vamos a analizar algunas variables continuas pero a su vez utilizaremos la variable respuesta para poder estudiar su comportamiento en conjunto.

Comencemos con la edad, la edad es una variable cuyo mínimo es 21 años y su máximo es de 80 años. En la figura 4.1 en donde se muestra un histograma de los clientes distribuidos por edad y divididos en dos grupos aquellos que pagaron su deuda y los que no pagaron, es claro que existen más clientes que pagaron su deuda en contraste a los que no lo hicieron. También podemos observar que los clientes se encuentran acumulados en las edades de 25 a 35 años.

Aunque esta variable es continua y la representación gráfica ya nos aporta información importante podríamos acotar el problema de modo que decidimos agrupar de la edad por intervalos de 5 años para poder hacer más sencillo el análisis.

Enfoquémonos en la figura 4.2. Es evidente que la mayoría de la población se

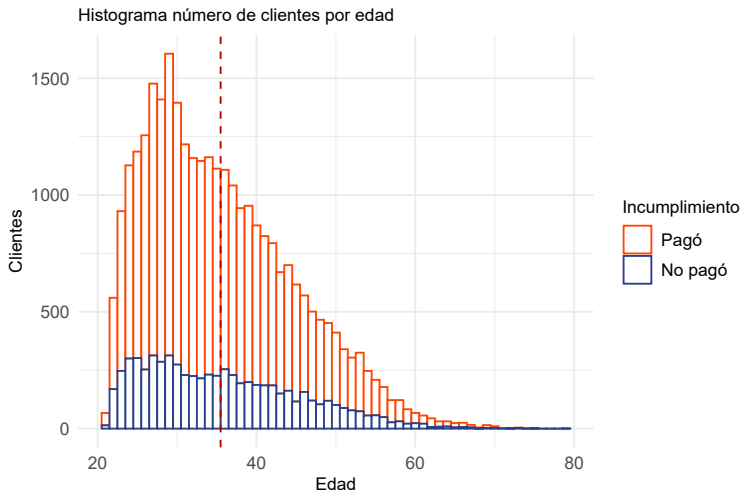

Figura 4.1: Distribución de clientes por edad dividido por la variable respuesta.

encuentra en la edad de 26 a 40 años, ya que si sumamos los porcentajes de estos grupos de edad tendríamos 60 % de los clientes totales, siendo el grupo de 26-30 años el que acumula más clientes con 23 % por el contrario el grupo de edad más joven de 21 a 25 años tan solo acumula el 9 % de los clientes totales.

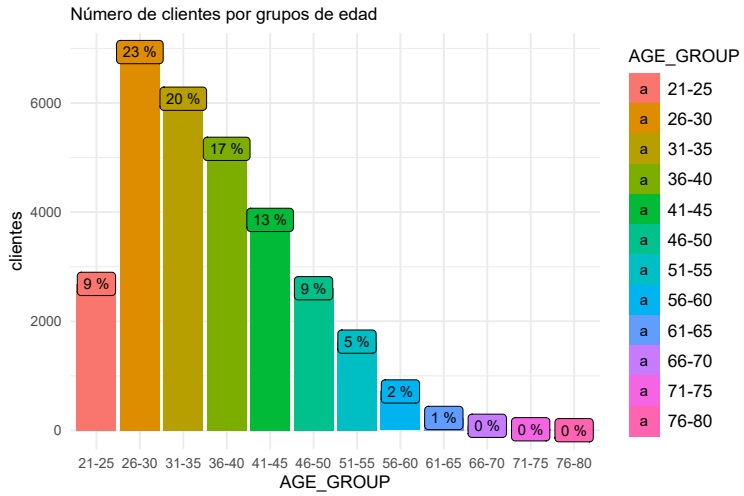

Figura 4.2: Clientes por edad agrupada

Veamos el comportamiento de las medias de ambas variables continuas divididas entre la variable respuesta en la figura 4.3, en la parte izquierda de la figura podemos

ver que no existe una diferencia entre las medias de las personas que pagaron su deuda y las que no pagaron su deuda respecto a la edad. Pero al hacer esta misma comparación con la variable de límite de crédito entonces notamos que existe una mayor cantidad de clientes que no pagaron su deuda con una media de límite de crédito menor respecto a la media de clientes que sí pagaron su deuda, es decir, entre menos elevado es el límite de crédito existen más clientes que no pagaron la deuda y viceversa.

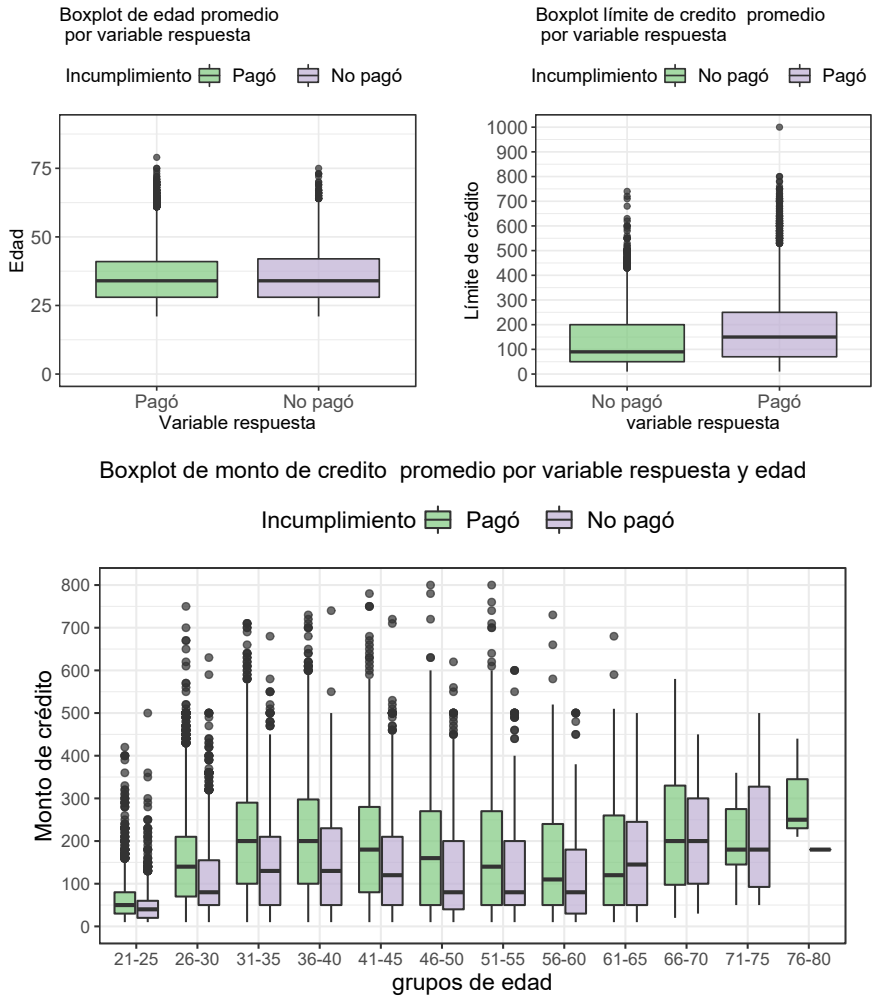

Figura 4.3: Boxplot edad y límite de crédito.

Sin embargo aún se puede hacer un análisis más detallado combinando estas variables continuas con la variable respuesta usando la división por grupos de edad que hicimos antes y estudiar su comportamiento en conjunto. Como se ve en la parte inferior de la figura 4.3 observamos que en los grupos de edades que se encuentran entre

los 25 y 65 años existe una diferencia de las medias de límite de crédito colocando la media de aquellos que no pagaron el crédito por encima de aquellos que si pagaron el crédito. Aunque existen ciertos grupos de edad en los que lo anterior no se cumple, de hecho sucede lo contrario, dado que en el grupo de 61 a 65 años la media del límite de crédito de las personas que pagaron la deuda es mayor a los que no pagaron, en el siguiente grupo de edad de 68 a 70 y de 71 a 75 no se ve una diferencia en las medias. Continuemos analizando las variables continuas que describen el comportamiento del crédito, comencemos con las variables de estado de cuenta que se componen por los estados de cuenta de cada mes de Abril a Septiembre del 2005. Primero veamos una tabla 4.3 con los montos promedios de estados de cuenta, máximos y mínimos según el mes y la representación gráfica 4.4.

Variable	Mes	Máximo	Mínimo	Media
BILLAMT1	Septiembre	964511	-165580	51223
BILLAMT2	Agosto	983931	-69777	49179
BILLAMT3	Julio	1664089	-157264	47013
BILLAMT4	Junio	891586	-170000	43262
BILLAMT5	Mayo	927171	-81334	891586
BILLAMT6	Abril	961664	-339603	38871

Tabla 4.2: Resumen variables de estado de cuenta.

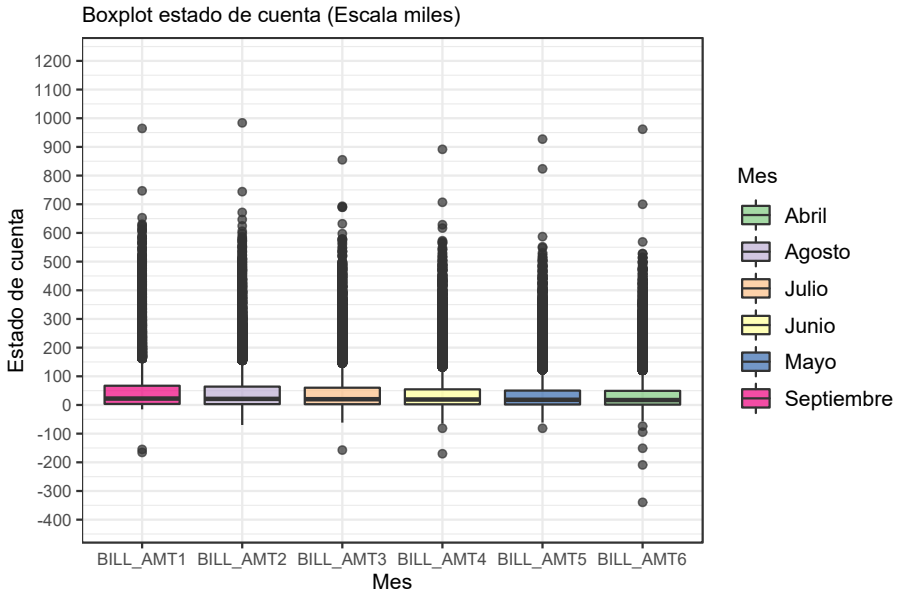

Figura 4.4: Boxplot estado de cuenta por mes.

En la tabla 4.3 podemos ver algunos datos relevantes por ejemplo en junio se en-

cuentra el cliente con el mayor estado de cuenta, pero si revisamos el promedio es en Mayo en donde tenemos el estado de cuenta promedio más elevado y Septiembre el que tiene el menor monto en promedio.

Enfocándonos en la gráfica 4.4 es notorio que existe gran cantidad de outliers para todos los meses por lo que para analizar su distribución vamos a descartar outliers, observemos los histogramas de cada mes en las figuras 4.5 y 4.6.

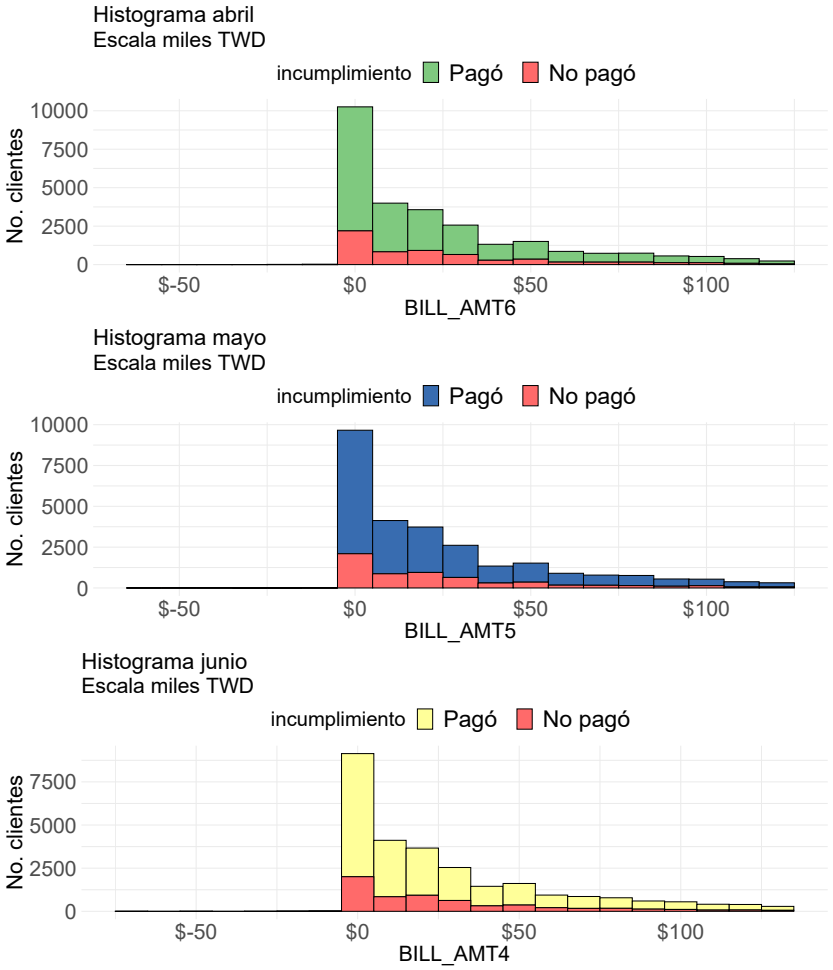

Figura 4.5: Histograma estado de cuenta por mes de abril a junio.

Podemos identificar que para todos los meses fijándonos en las figuras 4.5 y 4.6 una parte significativa de los clientes tienen un estado de cuenta en cero y la mayoría de los clientes se encuentra en el rango de cero a 50 mil NT dólares. Además para los 3 meses iniciales en la figura 4.5 la distribución de clientes que pagaron contra los que no pagaron es similar.

Continuemos analizando los siguientes meses figura 4.6. En julio el monto máximo

es 100 NT dólares mientras que en julio y agosto es de 150 mil NT, sin embargo, la distribución de clientes tanto los que pagaron y no pagaron es similar en los meses de abril a septiembre.

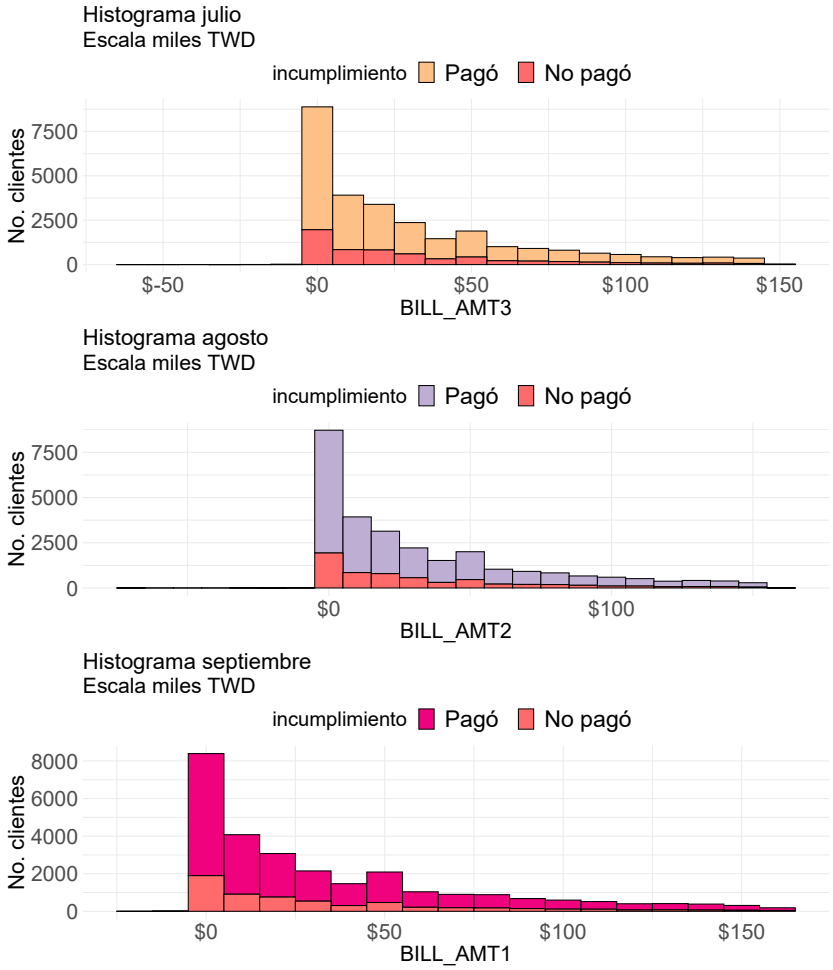

Figura 4.6: Histograma estado de cuenta por mes de julio a septiembre.

Ahora vamos a hacer un análisis análogo pero con los montos de pago del cliente del mes correspondiente, en este caso el monto mínimo para todos los meses es de cero por lo que tomaremos como mínimo aquel que sea mayor a cero.

Variable	Mes	Máximo	Mínimo	Media
PAYAMT1	Septiembre	873552	-165580	5663
PAYAMT2	Agosto	1684259	-69777	5921
PAYAMT3	Julio	896040	-157264	5225
PAYAMT4	Junio	621000	-170000	4826
PAYAMT5	Mayo	426529	-81334	4799
PAYAMT6	Abril	528666	-339603	5215

Tabla 4.3: Resumen variables pagos por mes.

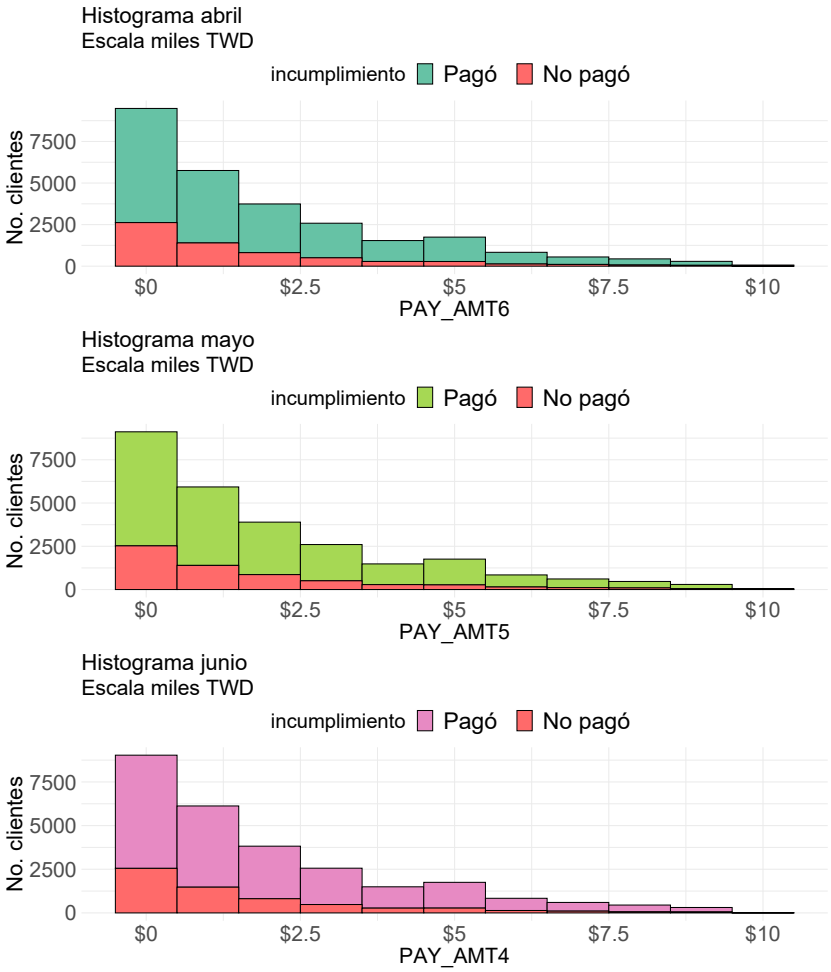

Figura 4.7: Histograma pagos por mes de abril a junio.

En el caso de los pagos de abril a junio figura 4.7 podemos observar que la mayoría

de los pagos están en cero. Sin embargo muestra una distribución similar los primeros tres meses en proporción de los clientes que pagaron y los que no pagaron.

Los siguientes meses figura 4.8 se observan cambios en la distribución de los pagos puesto que en julio existen más clientes que realizaron un pago de mil dólares comparando con agosto y septiembre, no obstante la proporción de clientes que no pagaron su deuda parece tener un comportamiento parecido en todos los meses además a diferencia de los pagos de abril a junio existe un pago de 12 mil dólares NT como máximo en este segundo trimestre.

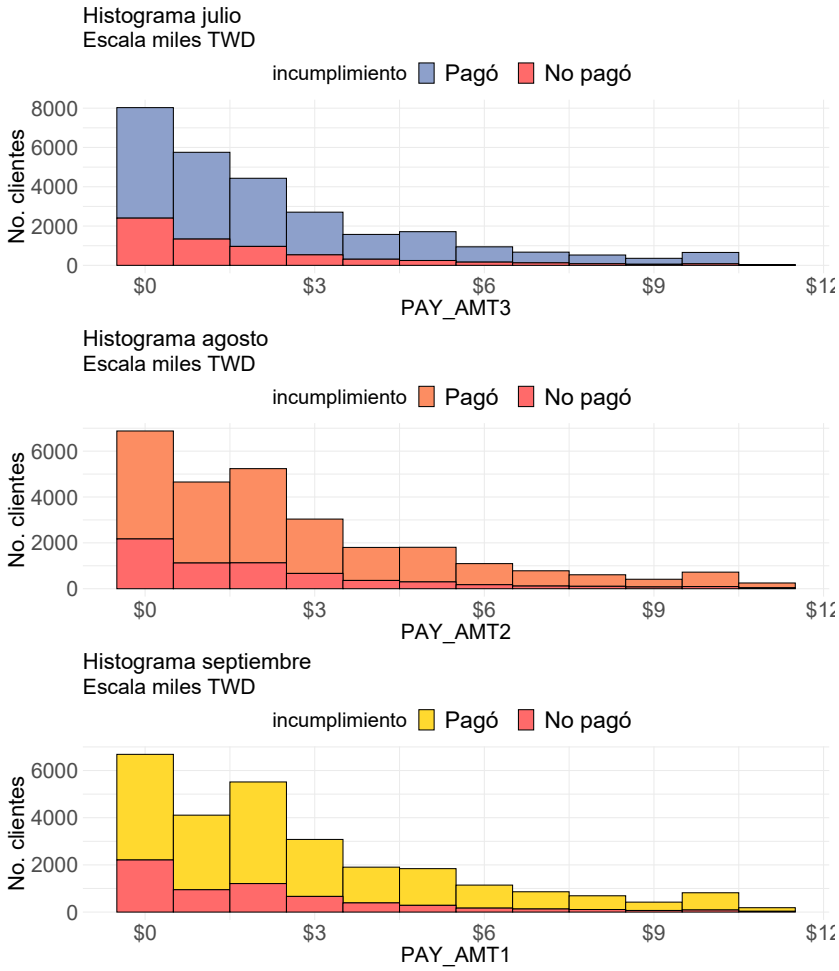

Figura 4.8: Histograma pagos por mes de julio a septiembre.

Como ya vimos las distribuciones de algunas variables parecen ser muy parecidas y debemos estudiar la correlación que existe entre las variables, al ser variables continuas podemos usar el gráfico de correlación veamos la figura 4.9 en este gráfico podemos apreciar la correlación de todas nuestras variables continuas, existe una correlación

cercana a uno en todas las variables de estado de cuenta puesto que podemos ver un conjunto de puntos de azul intenso acumulado en estas variables, esto podría ser un indicio de multicolinealidad, en las otras variables la correlación es muy baja, incluso en la variable edad se observa una correlación prácticamente de cero con las otras variables.

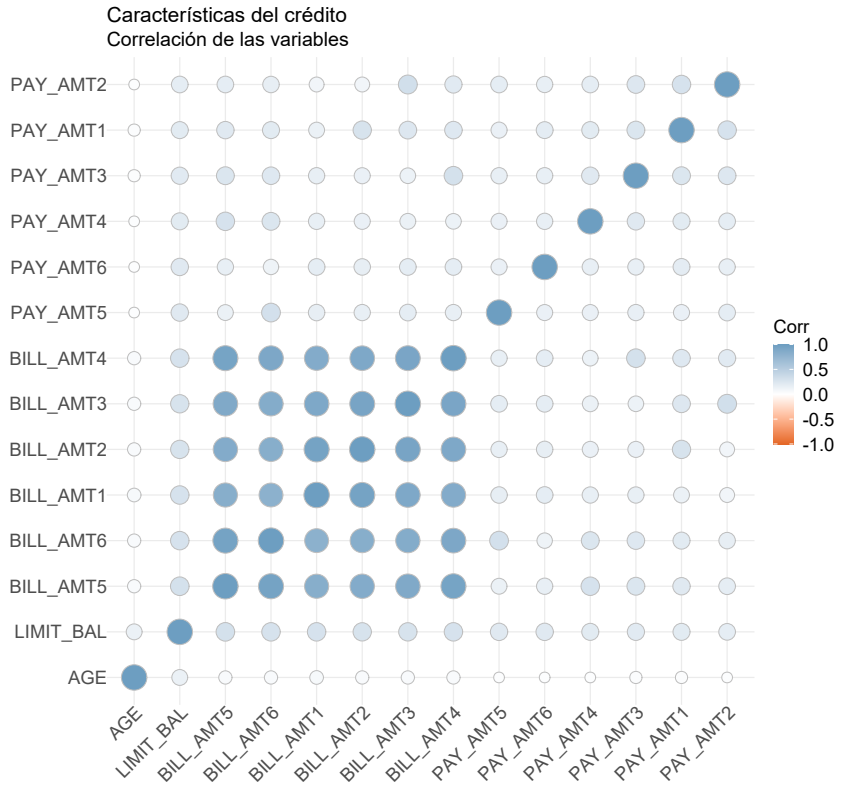

Figura 4.9: Gráfico de correlación variables continuas.

Debido a que los coeficientes de correlación para las variables de estado de cuenta son extremadamente altos haremos un gráfico de dispersión de estas variables para observar su comportamiento en la figura 4.10 en donde es claro que el comportamiento de las variables es similar y se puede ver una relación lineal entre todas las variables de estado de cuenta, en especial en la variable BILL_AMT5 y BILL_AMT6 en esta figura también podemos observar tanto a los clientes que no pagaron su deuda resaltados en rojo y aquellos que si lo hicieron resaltados en azul ambos grupos de clientes tienen una tendencia lineal. Este diagrama nos ayuda de confirmar que estas variables tienen una correlación extremadamente alta y probablemente exista multicolinealidad entre ellas.

Para finalizar analizaremos las variables categóricas: género, grado de escolaridad y estado civil, debido a que son variables categóricas, utilizaremos tablas de contingen-

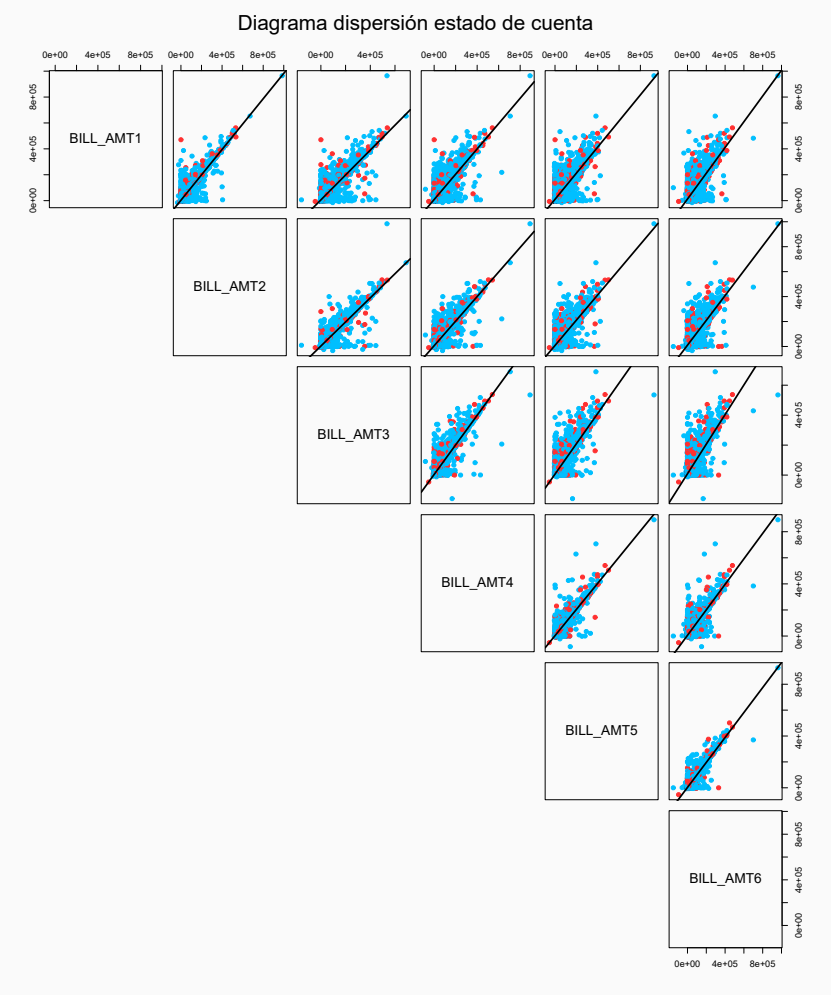

Figura 4.10: Diagrama dispersión estado de cuenta.

cia.

La tablas de contingencia pueden expresarse en términos diversos como frecuencias absolutas conjuntas condicionadas de una variable a valores de la otra, este caso la variable respuesta Y. También se pueden analizar las distribuciones marginales de las variables.

En este caso analizaremos la distribución conjunta de la variable respuesta y el resultado de la variable respuesta condicionado a la variable categórica que deseamos analizar, ya sea género, estado civil o educación.

Una vez que se tienen estas tablas se pueden analizar las proporciones y el comportamiento de las variables categóricas respecto a la variable respuesta o viceversa. Pero también nos interesa saber si la variable repuesta y la variable categórica son inde-

pendientes. Por lo que es necesario postular un modelo de independencia que se define como la tabla más probable pero construida al azar y nos muestra la forma que tendrían los conteos y las proporciones si las variables fueran independientes.

A través de una prueba de hipótesis con la estadística X^2 (ji-cuadrada) podemos comparar el modelo de independencia con el escenario real y saber si existe independencia entre las variables.

La prueba postula las siguientes hipótesis:

H_0 : Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.

H_a: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

De tal manera que dependiendo los resultados aceptaremos o rechazaremos la hipótesis nula ver tabla.

Prueba de independencia		
Resultado	Criterios	
Aceptar Ho,existe independencia	$X^2_{calculado} < X^2_{crtico}$	$pvalor > 0.05$
Rechazar Ho,existe dependencia	$X^2_{calculado} > X^2_{crtico}$	$pvalor < 0.05$

Tabla 4.4: Criterios prueba independencia.

Comencemos con la variable género y veamos las tablas de contingencia 4.5. Estas tablas pueden interpretarse de la siguiente forma. En general existe una mayor proporción de hombres no pagan conformando un 47 % del total de la población, cuando condicionamos la variable respuesta y dejamos fijo el valor del género es más claro que mientras que el 79 % del total de hombres no paga el 75 % del total de mujeres no paga, por lo que existe una mayor cantidad de hombres que no pagan por otro lado si esta vez condicionamos el valor del género y dejamos fijo el valores de los clientes que pagaron y los que no podemos observar que la proporción de hombres que pagaron es del 61 % mayor a la de las mujeres con 38 % pero también es mayor el porcentaje de los hombres que pagaron , esto se debe a como ya habíamos mencionado existe una mayor proporción de hombres que de mujeres en nuestra base de estudio.

Género	Género *vs*. Y		Condicionando Y		Condicionando Género	
	No pago	Pago	No pago	Pago	No pago	Pago
mujer	0.3005	0.0957	0.7583	0.2416	0.3858	0.4329
hombre	0.4783	0.1254	0.7922	0.2077	0.6141	0.5670

Tabla 4.5: Género *vs* Y.

Al llevar a cabo la prueba de hipótesis obtuvimos un p-value=4.945e-12 por lo que podemos afirmar que la variable respuesta Y está relacionada con el género.

Realizaremos el mismo análisis esta vez con la variable categórica estado civil, en la tabla 4.6, de forma general aquellas personas con estado civil soltero son la que cuentan con mayor proporción de incumplimiento de pago pero también son la mayoría de

los clientes son solteros en nuestra base de estudio. Cuando condicionamos la variable respuesta podemos ver que las proporciones son similares aunque existe una mayor proporción de solteros que no pagan con un 79 % seguido de el estado civil soltero con 76 % . Cuando condicionamos la variable estado civil, el valor soltero sigue dominando el incumplimiento de pago pero también cuanta con el mayor numero de clientes que pagaron su deuda con 50 % pero en este mismo apartado si nos enfocamos en los individuos casados podemos ver que el 48 % de los clientes casados pagaron contra el 44 % de clientes casados que no pagaron por lo que se puede afirmar condicionando al estado civil una mayor cantidad de clientes solteros pagaron respecto a los que no pagaron .

El resultado de la prueba de hipótesis $p - value = 7.791e - 07$ por lo que se puede concluir que la variable de estado civil y la variable respuesta están relacionadas.

Estado Civil	Estado civil *vs*. *Y*		Condicionando *Y*		Condicionando Estado Civil	
	No pago	Pago	No pago	Pago	No pago	Pago
Casado	0.3484	0.0001	0.7652	0.2347	0.4473	0.4831
Soltero	0.4207	0.1113	0.7907	0.2092	0.5402	0.5034
Otro	0.0096	0.0029	0.7639	0.2360	0.0123	0.0134

Tabla 4.6: Estado civil *vs*. *Y*.

La última variable categórica que analizaremos es grado de escolaridad en la tabla 4.7en donde aquellas personas que estudiaron hasta la universidad son aquellas que de forma general han incumplido más en el pago. Si condicionamos la variable educación entonces nuevamente observamos que aquellos que tenían un nivel universitario conformaban un 79 % de deudores que era el mayor porcentaje comparado con el nivel de preparatoria y posgrado, por último si obtenemos las proporciones respecto a la variable educación dejando fija la variable respuesta correspondiente a la columna marginal *y* podemos ver que en el grado de universidad existe 53 % de clientes que no pagaron pero también el 50 % si pagaron, esto es por que la mayoría de nuestra población tiene un grado universitario.

Educación	Educación *vs*. *Y*		Condicionando Y		Condicionando Educación	
	No pago	Pago	No pago	Pago	No pago	Pago
Escuela posgrado	0.3483	0.1068	0.7652	0.2347	0.4472	0.4831
Universidad	0.4204	0.1113	0.7905	0.2094	0.5398	0.5034
Preparatoria	0.0096	0.0029	0.7639	0.2360	0.0123	0.0134
Otro	0.0004	0	1	0	0	0

Tabla 4.7: Educación *vs*. *Y*.

Una vez que estudiamos las variables categóricas y su relación con la variable respuesta. Es necesario hacer este mismo análisis entre todas las variables categóricas. Esto es para analizar si existe relación entre ellas, debido a que tenemos un número significativo de variables categóricas resumimos los resultados de la prueba ji-cuadrada en la tabla 4.8 en este caso los p values son menores a 0.05 por lo que rechazaríamos

la hipótesis nula en cada caso lo cual implica que cada variable tiene una dependencia significativa.

Tabla 4.8: Resultados prueba independencia Chi cuadrada.

Variable 1	**Variable 2**	**P-value**	**Chi square**	**df**
*PAY*_0	*PAY*_2	$p < 0.0001$	131600	100
*PAY*_0	*PAY*_3	$p < 0.0001$	81016	100
*PAY*_0	*PAY*_4	$p < 0.0001$	52182	100
*PAY*_0	*PAY*_5	$p < 0.0001$	34553	90
*PAY*_0	*PAY*_6	$p < 0.0001$	NaN	100
*PAY*_0	SEX	$5.113157e-25$	140	10
*PAY*_0	MARRIAGE	$1.155239e-14$	111	20
*PAY*_0	EDUCATION	$8.453647e-235$	1207	30
*PAY*_0	default.payment.next.month	$p < 0.0001$	5366	10
*PAY*_2	*PAY*_3	$p < 0.0001$	117809	100
*PAY*_2	*PAY*_4	$p < 0.0001$	69740	100
*PAY*_2	*PAY*_5	$p < 0.0001$	40259	90
*PAY*_2	*PAY*_6	$p < 0.0001$	NaN	100
*PAY*_2	SEX	$6.959317e-33$	178	10
*PAY*_2	MARRIAGE	$8.380233e-16$	117	20
*PAY*_2	EDUCATION	$3.209062e-276$	1402	30
*PAY*_2	default.payment.next.month	$p < 0.0001$	3474	10
*PAY*_3	*PAY*_4	$p < 0.0001$	115139	100
*PAY*_3	*PAY*_5	$p < 0.0001$	70796	90
*PAY*_3	*PAY*_6	$p < 0.0001$	NaN	100
*PAY*_3	SEX	$2.157327e-29$	161	10
*PAY*_3	MARRIAGE	$4.933582e-13$	"102	20
*PAY*_3	EDUCATION	$1.917657e-257$	1314	30
*PAY*_3	default.payment.next.month	$p < 0.0001$	2622	10
*PAY*_4	*PAY*_5	$p < 0.0001$	116027	90
*PAY*_4	*PAY*_6	$p < 0.0001$	NaN	100
*PAY*_4	SEX	$1.204529e-23$	133	10
*PAY*_4	MARRIAGE	$1.813151e-16$	121	20
*PAY*_4	EDUCATION	$4.111076e-214$	1109	30
*PAY*_4	default.payment.next.month	$p < 0.0001$	2341	10
*PAY*_5	*PAY*_6	$p < 0.0001$	NaN	90
*PAY*_5	SEX	$1.805984e-18$	105	9
*PAY*_5	MARRIAGE	$1.335414e-15$	112	18
*PAY*_5	EDUCATION	$6.623549e-17$	927	27
*PAY*_5	default.payment.next.month	$0.00e+$	2198	9
*PAY*_6	SEX	$3.009732e-12$	NaN	10
*PAY*_6	MARRIAGE	$7.092356e-12$	NaN	20

Continúa en la página siguiente

Tabla 4.8 – Continúa en la página anterior

Variable 1	Variable 2	P-value	Chi square	df
PAY_6	EDUCATION	$1.213365e-167$	NaN	30
PAY_6	default.payment.next.month	$p < 0.0001$	NaN	10
SEX	MARRIAGE	$5.381729e-07$	29	2
SEX	EDUCATION	$2.603103e-05$	24	3
SEX	default.payment.next.month	$4.944679e-12$	48	1
EDUCATION	*MARRIAGE*	$6.334696e-232$	1089	6
EDUCATION	default.payment.next.month	$1.495065e-34$	160	3
MARRIAGE	default.payment.next.month	$7.79072e-07$	28	2

4.2. Ajuste de modelos

Vamos a ajustar diversos modelos contemplando las técnicas que estudiamos: regresión logística, análisis de componentes principales, regresión Lasso y Ridge. El primer modelo constará de todas las variables de la base de datos, primero dividiremos la base de datos en dos, trainning set (conjunto de entrenamiento) que consta de 75 % del total de los datos y test set (conjunto de prueba) 25 %, con la primera partición ajustaremos el modelo y con la segunda haremos las pruebas de bondad de ajuste para evaluar qué tan eficiente es el modelo.

4.2.1. Regresión Binaria

Es importante ajustar el modelo de regresión logística con todas las variables en un inicio, esto es para tener un punto de comparación de la devianza y los estimadores. Incluso aunque sabemos que la base tiene multicolinealidad iniciaremos el ajuste de la base de datos desde cero y vamos a monitorear cómo va evolucionando el resultado. Revisemos el resumen de este primer modelo:.

```
Call:
glm(formula = ‘default payment next month‘ ~ SEX + EDUCATION +
    MARRIAGE + LIMIT_BAL + AGE + PAY_0 + PAY_2 + PAY_3 + PAY_4 +
    PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 +
    BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
    PAY_AMT4 + PAY_AMT5 + PAY_AMT6, family = binomial("logit"),
    data = training_set_m1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3022  -0.5999  -0.5054  -0.3013   3.4033

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.130e+00  1.290e-01  -8.759  < 2e-16 ***
SEX2        -1.350e-01  3.736e-02  -3.614 0.000302 ***
EDUCATION2   1.195e-02  4.321e-02   0.277 0.782052
EDUCATION3  -8.616e-03  5.766e-02  -0.149 0.881213
EDUCATION4  -1.245e+00  2.347e-01  -5.305 1.13e-07 ***
MARRIAGE2   -1.752e-01  4.204e-02  -4.168 3.07e-05 ***
MARRIAGE3   -1.025e-01  1.604e-01  -0.639 0.522676
```

```
LIMIT_BAL   -2.100e-06  2.031e-07 -10.340  < 2e-16 ***
AGE          3.457e-03  2.267e-03   1.525 0.127302
PAY_0-1      4.521e-01  1.249e-01   3.620 0.000295 ***
PAY_00      -2.888e-01  1.349e-01  -2.140 0.032318 *
PAY_01       7.747e-01  9.771e-02   7.929 2.21e-15 ***
PAY_02       1.983e+00  1.226e-01  16.176  < 2e-16 ***
PAY_03       2.013e+00  1.923e-01  10.465  < 2e-16 ***
PAY_04       1.628e+00  3.481e-01   4.677 2.92e-06 ***
PAY_05       2.035e+00  6.102e-01   3.335 0.000852 ***
PAY_06       8.488e-01  9.467e-01   0.897 0.369923
PAY_07       2.566e+00  1.717e+00   1.495 0.134990
PAY_08      -1.187e+01  3.247e+02  -0.037 0.970832
PAY_2-1     -2.361e-01  1.312e-01  -1.800 0.071888 .
PAY_20      -3.796e-02  1.607e-01  -0.236 0.813226
PAY_21      -3.409e-01  6.033e-01  -0.565 0.572084
PAY_22      -1.345e-02  1.360e-01  -0.099 0.921233
PAY_23       8.600e-02  2.073e-01   0.415 0.678224
PAY_24      -7.954e-01  3.823e-01  -2.080 0.037492 *
PAY_25       5.030e-01  8.041e-01   0.626 0.531627
PAY_26       7.199e-02  1.756e+00   0.041 0.967289
PAY_27       1.665e+01  5.022e+02   0.033 0.973556
PAY_28      -2.570e+00  5.418e+02  -0.005 0.996216
PAY_3-1      6.344e-02  1.244e-01   0.510 0.610184
PAY_30       1.642e-01  1.451e-01   1.131 0.257939
PAY_31      -1.082e+01  3.247e+02  -0.033 0.973413
PAY_32       4.582e-01  1.468e-01   3.122 0.001798 **
PAY_33       5.460e-01  2.548e-01   2.143 0.032106 *
PAY_34       3.027e-01  4.821e-01   0.628 0.530019
PAY_35      -9.031e-01  9.746e-01  -0.927 0.354101
PAY_36      -2.490e+00  3.831e+02  -0.006 0.994815
PAY_37       1.357e-01  8.847e-01   0.153 0.878084
PAY_38      -1.497e+00  1.899e+00  -0.788 0.430718
PAY_4-1     -8.656e-02  1.256e-01  -0.689 0.490642
PAY_40      -6.992e-02  1.400e-01  -0.499 0.617561
PAY_41       7.045e-01  4.593e+02   0.002 0.998776
PAY_42       2.486e-01  1.501e-01   1.656 0.097678 .
PAY_43      -4.622e-02  2.810e-01  -0.165 0.869332
PAY_44       4.098e-01  4.985e-01   0.822 0.411002
PAY_45      -1.811e+00  9.331e-01  -1.941 0.052310 .
PAY_46      -1.070e+01  4.337e+02  -0.025 0.980310
PAY_47      -2.239e+01  2.875e+02  -0.078 0.937927
PAY_48      -3.699e+01  4.337e+02  -0.085 0.932028
PAY_5-1     -7.568e-02  1.228e-01  -0.616 0.537847
PAY_50       6.111e-02  1.361e-01   0.449 0.653491
PAY_52       3.560e-01  1.530e-01   2.327 0.019984 *
PAY_53       1.774e-01  2.812e-01   0.631 0.528251
PAY_54      -2.230e-01  5.159e-01  -0.432 0.665544
PAY_55       1.435e+00  1.017e+00   1.411 0.158188
PAY_56       2.362e+01  2.875e+02   0.082 0.934525
PAY_57       2.367e+01  2.875e+02   0.082 0.934385
PAY_6-1     -1.026e-01  9.411e-02  -1.091 0.275427
PAY_60      -4.129e-01  1.021e-01  -4.045 5.22e-05 ***
PAY_62      -1.244e-01  1.196e-01  -1.041 0.298077
PAY_63       6.013e-01  2.806e-01   2.142 0.032158 *
PAY_64      -4.249e-01  5.288e-01  -0.804 0.421668
PAY_65      -1.610e-01  9.053e-01  -0.178 0.858875
PAY_66       4.692e-01  9.682e-01   0.485 0.627960
PAY_67      -8.180e-01  1.803e+00  -0.454 0.650024
PAY_68              NA         NA      NA       NA
BILL_AMT1   -1.922e-06  1.284e-06  -1.497 0.134470
BILL_AMT2    3.595e-06  1.633e-06   2.202 0.027671 *
BILL_AMT3    1.482e-06  1.445e-06   1.026 0.304898
BILL_AMT4   -6.246e-07  1.529e-06  -0.408 0.682973
BILL_AMT5    1.703e-07  1.716e-06   0.099 0.920945
BILL_AMT6   -2.888e-07  1.366e-06  -0.211 0.832563
PAY_AMT1    -1.054e-05  2.576e-06  -4.092 4.28e-05 ***
PAY_AMT2    -1.037e-05  2.592e-06  -3.999 6.35e-05 ***
PAY_AMT3    -6.674e-07  1.888e-06  -0.353 0.723724
```

```
PAY_AMT4    -2.426e-06  2.097e-06  -1.157 0.247313
PAY_AMT5    -3.109e-06  2.024e-06  -1.536 0.124517
PAY_AMT6    -2.326e-06  1.534e-06  -1.516 0.129509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23779  on 22499  degrees of freedom
Residual deviance: 19487  on 22423  degrees of freedom
AIC: 19641
Number of Fisher Scoring iterations: 11
```

En este primer resumen del modelo 1 podemos observar que ciertas variables parecen ser no significativas ya que cuentan con *p-values* muy grandes, éstas son las variables correspondientes a los estados de cuenta de cada mes, es decir, las variables de estado de cuenta. Además, si nos fijamos en el resultado de los factores de la varianza generalizados en la sección confirmamos que nuestra base tiene multicolinealidad por lo que es necesario hacer nuevos modelos basados en distintos métodos de selección de variables. También es importante revisar el valor del AIC en este caso como se ve en el resumen es de 19641 y el valor de la devianza es 19487 los cuales son valores elevados y lo que se busca es que el valor de la devianza sea "pequeñorespecto a los grados de libertad, sin embargo estos valores sólo nos ayudan a hacer comparaciones con otros modelos. Por otro lado revisaremos la cantidad de verdaderos positivos y verdaderos negativos en el lado derecho de la figura 4.11 así como la matriz de confusión representada en una gráfica de mosaico que se encuentra en el lado izquierdo de la figura 4.11 en donde es claro que este modelo cuenta con una gran cantidad de verdaderos negativos que representa a aquellas personas que no incumplieron en el pago, es decir que pagaron su deuda puesto que el área del cuadrado que representa estos casos es la más grande de la gráfica de mosaico, pero la cantidad de verdaderos positivos que representa a aquellas personas que incumplieron en el pago no es tan relevante.

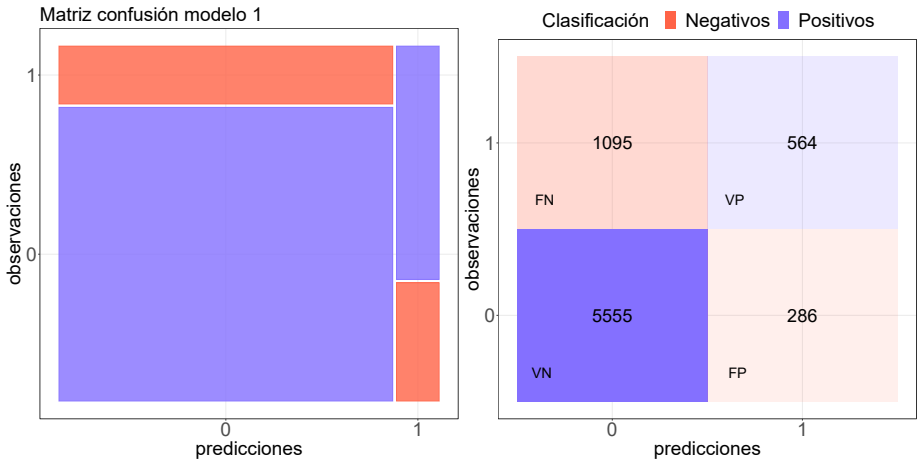

Figura 4.11: Matriz de confusión modelo 1, todas las variables.

Para revisar más a detalle enfoquémonos en el lado izquierdo de la figura en donde

podemos ver los casos totales, existen 5555 clientes que pagaron su deuda formando un 74 % del set de prueba y un porcentaje 7 % de verdaderos positivos, lo que indicaría que la suma de falsos positivos y falsos negativos es de 9 % en total. Sin embargo aún tenemos que tomar en cuenta la sensibilidad y especificidad.

Por último podemos analizar la curva ROC ver figura 4.12 aquí podemos ver la línea roja pegada a la esquina superior izquierda lo cual indica que el modelo tiene un buen nivel de especificidad y sensibilidad, esto se refuerza con el valor del área bajo la curva ya que es de 0.7524.

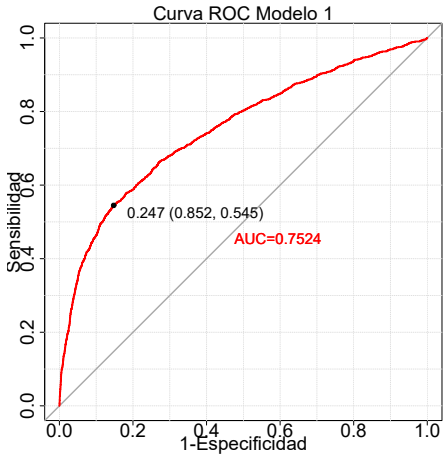

Figura 4.12: Curva Roc modelo 1, todas las variables.

Estos datos en conjunto describen que en general el ajuste del modelo es bueno respecto a los niveles de verdaderos positivos y verdaderos negativos, a pesar de ello existe aún el problema de multicolinealidad que podría afectar en los estimadores es por ello que continuaremos analizando diversos modelos que ayudan a mitigar la multicolinealidad de acuerdo a la literatura [11] y de los cuales se han tenido buenos resultados de forma empírica.

4.2.2. Diagnóstico Multicolinealidad

En el análisis exploratorio de los datos identificamos que existe una alta correlación entre las variables de estado de cuenta, esto indica que probablemente haya multicolinealidad, además si analizamos los factores de inflación de la varianza generalizada que se obtiene a partir del modelo modelo 1 obtenemos los siguientes resultados:

```
                GVIF         Df         GVIF^(1/(2*Df))
SEX             1.030018      1         1.014898
EDUCATION       1.247831      3         1.037590
MARRIAGE        1.347745      2         1.077462
LIMIT_BAL       1.704269      1         1.305476
AGE             1.397080      1         1.181981
PAY_0           1.357088e+08  10        2.550530
```

PAY_2	2.232899e+09	10	2.933887
PAY_3	4.223928e+12	10	2.278444
PAY_4	4.801059e+13	10	2.231332
PAY_5	9.669708e+08	9	2.156383
PAY_6	4.325180e+03	9	1.592210
BILL_AMT1	23.02733	1	4.798680
BILL_AMT2	38.38520	1	6.195579
BILL_AMT3	28.20098	1	5.310460
BILL_AMT4	25.49297	1	5.049057
BILL_AMT5	29.88799	1	5.466991
BILL_AMT6	17.67697	1	4.204399
PAY_AMT1	1.478114	1	1.215777
PAY_AMT2	1.490935	1	1.221039
PAY_AMT3	1.532763	1	1.238048
PAY_AMT4	1.477487	1	1.215519
PAY_AMT5	1.535143	1	1.239009
PAY_AMT6	1.155647	1	1.075010

Recordemos que existían distintos criterios para determinar si existía multicolinealidad ver sección 3.1.1, en este caso hay variables que tiene el valor de $GVIF^{\frac{1}{2df}}$ mayor a 2 lo que es equivalente a un VIF mayor a 4, incluso existen valores $GVIF^{\frac{1}{2df}}$ mayores a 5 por lo que notamos que existe multicolinealidad grave entre las variables que corresponden a las variables de estado de cuenta, estas variables son continuas por lo que podemos tomar directamente como el valor del VIF, siendo las variables de BILL_AMT2 y BILL_AMT3 las variables con un vif más elevado con 38 y 28, respectivamente . Por otro lado, en el diagrama de dispersión de estas variables en la figura 4.10 confirmamos que existe cierta tendencia y en conclusión la base de datos tiene multicolinealidad.

Notemos que las variables categóricas *PAY*_3 y *PAY*_4 son ligeramente mayores a 4 cuando elevamos al cuadrado $GVIF^{\frac{1}{2df}}$ para obtener el VIF por lo que en este caso no las tomaremos cómo presencia de multicolinealidad.

4.2.3. Análisis de Componentes Principales

Una vez que identificamos que nuestra base de datos tiene multicolinealidad, es importante manejarla, en este caso utilizaremos componentes principales para obtener las variables que aportan mayor información al modelo y las componentes principales. Veamos cuales son los vectores propios de cada componente principal los cuales nos darían la combinación lineal de cada componente si se multiplica por las variables originales.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
BILL_AMT1	0.4475841	0.55364282	-0.44992027	0.1928128	-0.50352325	-0.002719602
BILL_AMT2	0.4417654	0.39161138	-0.03350715	-0.1688348	0.78851233	0.010209092
BILL_AMT3	0.4296866	0.07512402	0.71786681	-0.4137713	-0.33476845	-0.105434387
BILL_AMT4	0.3979791	-0.26727471	0.29121344	0.7064547	0.06789333	0.426243795
BILL_AMT5	0.3699188	-0.45057905	-0.17374957	0.1561488	0.05263381	-0.776399988
BILL_AMT6	0.3530636	-0.50992086	-0.40758078	-0.4895412	-0.07254586	0.451985899

Para elegir cual es la variable que representa mejor los datos analizaremos la varianza de las componentes principales de las variables que tienen multicolinealidad, para este caso la varianza es mayor en el primer componente que corresponde a la variable $\text{BILL}_A MT1 que es el estado de cuenta en el mes de agosto$:

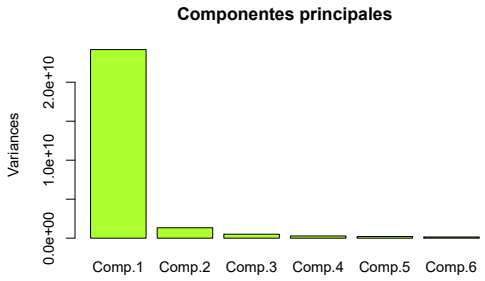

Figura 4.13: Componentes principales de la variable estado de cuenta.

Enfocándonos en la varianza acumulada podemos ver que si nos quedamos con la primera componente principal podemos obtener mayor parte de la varianza. En la figura 4.14 es claro que el primer componente acumulan 91 %. de la varianza.

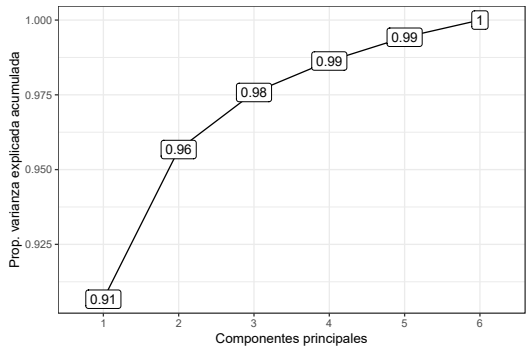

Figura 4.14: Varianza acumulada variable estados de cuenta.

Por lo que la primera componente principal representa mejor a nuestros datos y estaría dada por

$$PC1 = 0.4475BILL\_AMT1 + 0.4417BILL\_AMT2 + 0.4296BILL\_AMT3 + 0.3979BILL\_AMT4 + 0.3699BILL\_AMT5 + 0.3530BILL\_AMT6$$

Este resultado se puede implementar para construir un nuevo modelo de regresión logística sin el problema de multicolinealidad.

4.2.4. Regresión Binaria con componentes principales

Como vimos en la secciones anteriores la base fue diagnosticada con multicolinealidad por lo que ajustaremos dos modelos basándonos en el análisis de componentes principales, existen dos métodos. El primero, al que llamaremos modelo 2, consiste en una selección de variables, es decir, tomamos la variable que representa más peso en la componentes principal que acumula mayor varianza en este caso sería la variable *BILL_AMT*1 y otro ,al que llamaremos modelo 3, construido con la primer componente principal es decir la combinación lineal *PC*1 dado que aporta 91 % de la varianza, ambos métodos resuelven el problema de multicolinealidad.

Aplicaremos las mismas pruebas del modelo 1 a los modelos basados en los resultados de componentes principales.

El resumen de nuestro de modelo 2, es el siguiente:

```
Call:
glm(formula = `default payment next month` ~ SEX + EDUCATION +
    MARRIAGE + LIMIT_BAL + AGE + PAY_0 + PAY_2 + PAY_3 + PAY_4 +
    PAY_5 + PAY_6 + BILL_AMT1 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
    PAY_AMT4 + PAY_AMT5 + PAY_AMT6, family = binomial("logit"),
    data = training_set_m1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2977  -0.5991  -0.5053  -0.3061   3.3936

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.150e+00  1.286e-01  -8.944  < 2e-16 ***
SEX2        -1.332e-01  3.733e-02  -3.568 0.000360 ***
EDUCATION2   9.639e-03  4.319e-02   0.223 0.823375
EDUCATION3  -1.267e-02  5.762e-02  -0.220 0.826015
EDUCATION4  -1.244e+00  2.344e-01  -5.308 1.11e-07 ***
MARRIAGE2   -1.750e-01  4.203e-02  -4.164 3.12e-05 ***
MARRIAGE3   -1.039e-01  1.604e-01  -0.648 0.517219
LIMIT_BAL   -2.092e-06  2.016e-07 -10.374  < 2e-16 ***
AGE          3.556e-03  2.267e-03   1.568 0.116826
PAY_0-1      4.589e-01  1.249e-01   3.675 0.000238 ***
PAY_00      -2.939e-01  1.347e-01  -2.182 0.029085 *
PAY_01       7.800e-01  9.765e-02   7.987 1.38e-15 ***
PAY_02       1.990e+00  1.225e-01  16.249  < 2e-16 ***
PAY_03       2.017e+00  1.923e-01  10.490  < 2e-16 ***
PAY_04       1.647e+00  3.478e-01   4.736 2.18e-06 ***
PAY_05       2.063e+00  6.113e-01   3.375 0.000737 ***
PAY_06       8.530e-01  9.446e-01   0.903 0.366463
PAY_07       2.569e+00  1.714e+00   1.498 0.134037
PAY_08      -1.184e+01  3.247e+02  -0.036 0.970920
PAY_2-1     -2.728e-01  1.304e-01  -2.093 0.036365 *
PAY_20      -8.206e-02  1.591e-01  -0.516 0.605944
PAY_21      -4.945e-01  5.968e-01  -0.828 0.407390
PAY_22      -5.629e-02  1.346e-01  -0.418 0.675769
PAY_23       4.039e-02  2.062e-01   0.196 0.844727
PAY_24      -8.423e-01  3.818e-01  -2.206 0.027380 *
PAY_25       4.533e-01  8.023e-01   0.565 0.572083
PAY_26       3.965e-02  1.753e+00   0.023 0.981958
PAY_27       1.678e+01  5.024e+02   0.033 0.973364
PAY_28      -2.746e+00  5.422e+02  -0.005 0.995959
PAY_3-1      7.479e-02  1.234e-01   0.606 0.544558
PAY_30       2.010e-01  1.434e-01   1.402 0.160959
PAY_31      -1.069e+01  3.247e+02  -0.033 0.973746
PAY_32       4.984e-01  1.453e-01   3.431 0.000602 ***
PAY_33       5.842e-01  2.540e-01   2.300 0.021462 *
PAY_34       3.553e-01  4.820e-01   0.737 0.461017
PAY_35      -8.544e-01  9.752e-01  -0.876 0.380967
PAY_36      -2.609e+00  3.834e+02  -0.007 0.994569
```

```
PAY_37       1.679e-01  8.849e-01   0.190 0.849516
PAY_38      -1.475e+00  1.903e+00  -0.775 0.438475
PAY_4-1     -7.243e-02  1.249e-01  -0.580 0.561858
PAY_40      -4.824e-02  1.391e-01  -0.347 0.728664
PAY_41       7.026e-01  4.593e+02   0.002 0.998779
PAY_42       2.690e-01  1.493e-01   1.802 0.071595 .
PAY_43      -2.179e-02  2.806e-01  -0.078 0.938104
PAY_44       4.217e-01  4.986e-01   0.846 0.397719
PAY_45      -1.806e+00  9.339e-01  -1.934 0.053146 .
PAY_46      -1.055e+01  4.342e+02  -0.024 0.980612
PAY_47      -2.231e+01  2.882e+02  -0.077 0.938288
PAY_48      -3.692e+01  4.342e+02  -0.085 0.932226
PAY_5-1     -7.709e-02  1.223e-01  -0.630 0.528470
PAY_50       6.096e-02  1.352e-01   0.451 0.652036
PAY_52       3.561e-01  1.521e-01   2.341 0.019250 *
PAY_53       1.760e-01  2.810e-01   0.626 0.531235
PAY_54      -2.179e-01  5.155e-01  -0.423 0.672575
PAY_55       1.444e+00  1.018e+00   1.418 0.156131
PAY_56       2.361e+01  2.882e+02   0.082 0.934710
PAY_57       2.358e+01  2.882e+02   0.082 0.934772
PAY_6-1     -1.000e-01  9.383e-02  -1.066 0.286347
PAY_60      -4.009e-01  1.006e-01  -3.984 6.79e-05 ***
PAY_62      -1.140e-01  1.177e-01  -0.969 0.332685
PAY_63       6.071e-01  2.801e-01   2.167 0.030213 *
PAY_64      -4.194e-01  5.290e-01  -0.793 0.427857
PAY_65      -1.594e-01  9.060e-01  -0.176 0.860339
PAY_66       5.097e-01  9.617e-01   0.530 0.596092
PAY_67      -7.930e-01  1.805e+00  -0.439 0.660472
PAY_68              NA         NA      NA       NA
BILL_AMT1    2.041e-06  3.540e-07   5.765 8.16e-09 ***
PAY_AMT1    -7.178e-06  2.279e-06  -3.149 0.001638 **
PAY_AMT2    -9.672e-06  2.357e-06  -4.104 4.06e-05 ***
PAY_AMT3    -1.541e-06  1.637e-06  -0.942 0.346268
PAY_AMT4    -2.690e-06  1.858e-06  -1.447 0.147772
PAY_AMT5    -3.281e-06  1.732e-06  -1.894 0.058223 .
PAY_AMT6    -2.344e-06  1.500e-06  -1.563 0.118119
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23779  on 22499  degrees of freedom
Residual deviance: 19499  on 22428  degrees of freedom
AIC: 19643

Number of Fisher Scoring iterations: 11
```

En el resumen de este segundo modelo podemos identificar el valor del AIC que es de 19643 y el valor de la devianza que es de 19499. Ahora vamos a evaluar cuál es la capacidad del modelo conformado con las componentes principales del estado de cuenta para clasificar las observaciones. Es decir, cuál es la cantidad de verdaderos positivos y verdaderos negativos. Para esto utilizaremos la matriz de confusión, figura 4.15, en donde podemos notar que tenemos una cantidad significativa de verdaderos negativos al igual que en el primer modelo. Notemos que la mayor cantidad de observaciones se encuentra en el área del mosaico formado por $x = 0$ y $y = 0$ que representa a aquellas personas que pagaron su deuda, seguido del cuadrado $x = 0$ y $y = 1$ que representa a aquellas personas que incumplieron en el pago sin embargo el modelo las clasificó como personas que pagarán a tiempo su deuda si observamos la tabla de la izquierda podemos observar a que estos mosaicos corresponden a 5555 y 1095 clientes respectivamente.

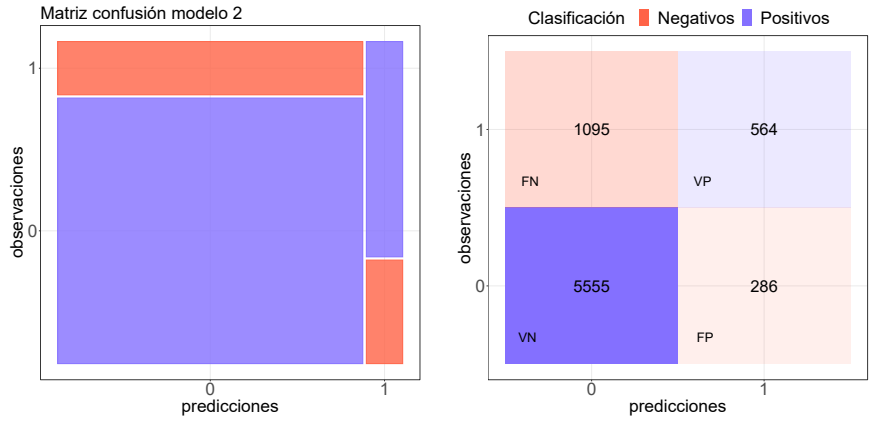

Figura 4.15: Matriz de confusión modelo 2, componentes principales.

Los siguientes dos casos representan un porcentaje mínimo del subconjunto de datos en el que se aplicó la prueba. El 7 % conformado por aquellos casos observados como incumplimientos clasificados correctamente por el modelo y el 3 % de los casos que representa a clientes que pagaron su deuda pero el modelo los clasificó como incumplimiento de pago.

Ahora revisemos el modelo 3 que incluye la primer componente principal, $PC1$.

```
Call:
glm(formula = `default payment next month` ~ SEX + EDUCATION +
    MARRIAGE + LIMIT_BAL + AGE + PAY_0 + PAY_2 + PAY_3 + PAY_4 +
    PAY_5 + PAY_6 + PC1 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 +
    PAY_AMT5 + PAY_AMT6, family = binomial("logit"), data = training_set_m7)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3017  -0.5985  -0.5059  -0.3029   3.4184

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.128e+00  1.290e-01  -8.748  < 2e-16 ***
SEX2        -1.361e-01  3.733e-02  -3.647 0.000265 ***
EDUCATION2   1.124e-02  4.319e-02   0.260 0.794598
EDUCATION3  -9.085e-03  5.760e-02  -0.158 0.874687
EDUCATION4  -1.241e+00  2.347e-01  -5.289 1.23e-07 ***
MARRIAGE2   -1.747e-01  4.204e-02  -4.155 3.25e-05 ***
MARRIAGE3   -9.902e-02  1.605e-01  -0.617 0.537300
LIMIT_BAL   -2.131e-06  2.023e-07 -10.530  < 2e-16 ***
AGE          3.503e-03  2.267e-03   1.545 0.122317
PAY_0-1      4.495e-01  1.248e-01   3.600 0.000318 ***
PAY_00      -2.995e-01  1.348e-01  -2.223 0.026243 *
PAY_01       7.734e-01  9.770e-02   7.916 2.45e-15 ***
PAY_02       1.979e+00  1.226e-01  16.145  < 2e-16 ***
PAY_03       2.015e+00  1.923e-01  10.480  < 2e-16 ***
PAY_04       1.637e+00  3.479e-01   4.706 2.53e-06 ***
PAY_05       2.056e+00  6.111e-01   3.364 0.000768 ***
PAY_06       8.510e-01  9.451e-01   0.900 0.367917
PAY_07       2.563e+00  1.721e+00   1.489 0.136466
PAY_08      -1.188e+01  3.247e+02  -0.037 0.970808
PAY_2-1     -2.570e-01  1.305e-01  -1.970 0.048864 *
PAY_20      -5.203e-02  1.591e-01  -0.327 0.743612
PAY_21      -4.430e-01  5.982e-01  -0.740 0.458997
PAY_22      -3.081e-02  1.344e-01  -0.229 0.818673
```

```
PAY_23       6.958e-02  2.061e-01   0.338 0.735618
PAY_24      -8.141e-01  3.818e-01  -2.132 0.032989 *
PAY_25       4.808e-01  8.025e-01   0.599 0.549094
PAY_26       3.417e-02  1.759e+00   0.019 0.984504
PAY_27       1.649e+01  5.021e+02   0.033 0.973802
PAY_28      -2.472e+00  5.416e+02  -0.005 0.996358
PAY_3-1      7.587e-02  1.237e-01   0.613 0.539611
PAY_30       1.977e-01  1.438e-01   1.374 0.169291
PAY_31      -1.070e+01  3.247e+02  -0.033 0.973726
PAY_32       4.894e-01  1.457e-01   3.360 0.000779 ***
PAY_33       5.767e-01  2.542e-01   2.269 0.023249 *
PAY_34       3.480e-01  4.816e-01   0.723 0.469943
PAY_35      -8.673e-01  9.743e-01  -0.890 0.373338
PAY_36      -2.318e+00  3.830e+02  -0.006 0.995170
PAY_37       1.832e-01  8.838e-01   0.207 0.835805
PAY_38      -1.424e+00  1.893e+00  -0.752 0.451960
PAY_4-1     -6.980e-02  1.251e-01  -0.558 0.576791
PAY_40      -5.567e-02  1.394e-01  -0.399 0.689681
PAY_41       6.823e-01  4.593e+02   0.001 0.998815
PAY_42       2.596e-01  1.496e-01   1.736 0.082640 .
PAY_43      -3.519e-02  2.807e-01  -0.125 0.900253
PAY_44       4.164e-01  4.984e-01   0.835 0.403527
PAY_45      -1.792e+00  9.325e-01  -1.922 0.054646 .
PAY_46      -1.081e+01  4.334e+02  -0.025 0.980094
PAY_47      -2.245e+01  2.870e+02  -0.078 0.937666
PAY_48      -3.705e+01  4.334e+02  -0.085 0.931871
PAY_5-1     -7.942e-02  1.224e-01  -0.649 0.516457
PAY_50       4.958e-02  1.355e-01   0.366 0.714371
PAY_52       3.390e-01  1.524e-01   2.225 0.026105 *
PAY_53       1.624e-01  2.810e-01   0.578 0.563220
PAY_54      -2.439e-01  5.161e-01  -0.473 0.636500
PAY_55       1.417e+00  1.017e+00   1.393 0.163666
PAY_56       2.361e+01  2.870e+02   0.082 0.934448
PAY_57       2.372e+01  2.870e+02   0.083 0.934133
PAY_6-1     -1.045e-01  9.387e-02  -1.113 0.265591
PAY_60      -4.260e-01  1.011e-01  -4.215 2.50e-05 ***
PAY_62      -1.444e-01  1.181e-01  -1.223 0.221378
PAY_63       5.851e-01  2.803e-01   2.087 0.036864 *
PAY_64      -4.440e-01  5.286e-01  -0.840 0.400914
PAY_65      -1.816e-01  9.064e-01  -0.200 0.841183
PAY_66       4.346e-01  9.734e-01   0.447 0.655205
PAY_67      -8.343e-01  1.799e+00  -0.464 0.642841
PAY_68              NA         NA      NA       NA
PC1          1.080e-06  1.737e-07   6.220 4.96e-10 ***
PAY_AMT1    -8.041e-06  2.292e-06  -3.508 0.000452 ***
PAY_AMT2    -1.036e-05  2.377e-06  -4.360 1.30e-05 ***
PAY_AMT3    -1.956e-06  1.653e-06  -1.183 0.236938
PAY_AMT4    -2.980e-06  1.881e-06  -1.584 0.113170
PAY_AMT5    -3.529e-06  1.751e-06  -2.016 0.043840 *
PAY_AMT6    -2.169e-06  1.507e-06  -1.440 0.149972
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23779  on 22499  degrees of freedom
Residual deviance: 19494  on 22428  degrees of freedom
AIC: 19638

Number of Fisher Scoring iterations: 11
```

En este modelo podemos visualizar el AIC con un valor de 19638 y una devinza de 19494, el AIC es menor al modelo 2 por una cantidad mínima al igual que la devianza, por lo que hasta el momento el modelo construido con $PC1$, en lugar de la selección de variables es mejor. Para estar seguros realizemos la matriz de confusión, figura 4.16, en este modelo podemos ver el mismo patrón la mayor cantidad de clientes se encuentra

clasificados en verdaderos negativos conformado por los clientes que pagaron su deuda clasificados correctamente.

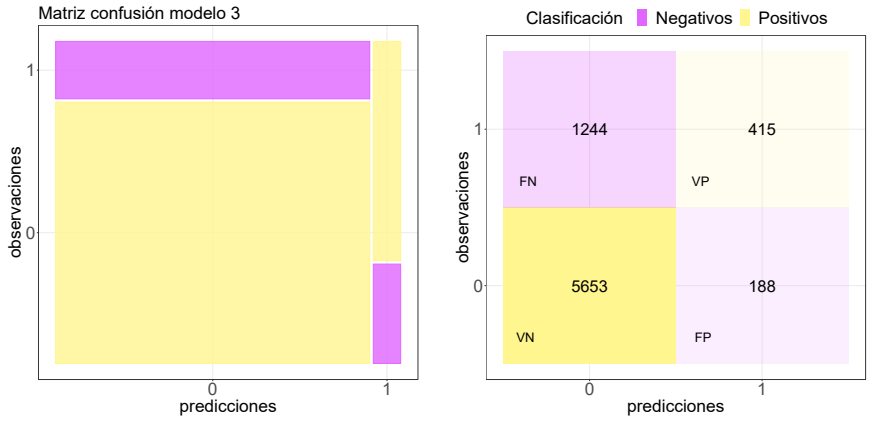

Figura 4.16: Matriz de confusión modelo 3, componentes principales.

Veamos la curva ROC de ambos modelos. En el caso del modelo 2 podemos visualizar del lado izquierdo de la figura 4.17 una curva que no está pegada a la esquina superior izquierda lo que significa que el ajuste no es tan eficiente, si revisamos el área bajo la curva *AUC* tiene un valor de 0.602 con lo que corroboramos que el desempeño del modelo no es el deseado, al otro lado se encuentra la curva ROC del modelo 3 en donde claramente existe un mejor desempeño ya que tiene un *AUC* de 0.71.

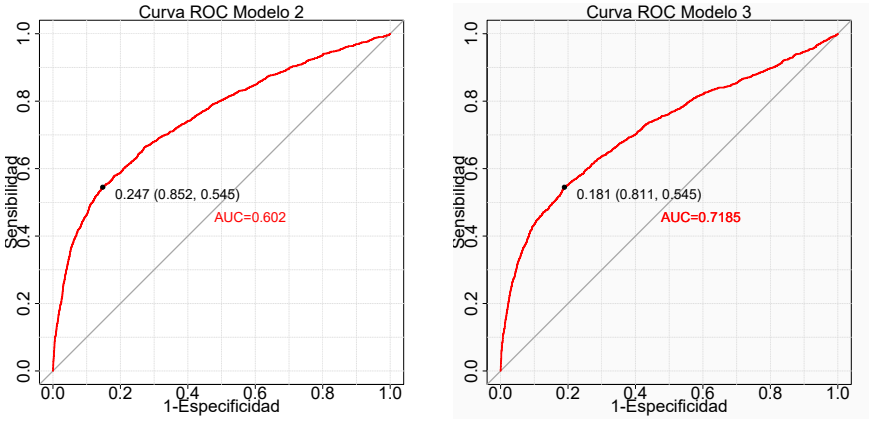

Figura 4.17: Curva Roc modelo 2 y 3, regresión logística con resultados de PCA.

4.2.5. Regresión LASSO

Otra opción para lidiar con la multicolinealidad es la regresión LASSO. Como revisamos en la parte teórica 3.3.1, consiste en reducir ciertos coeficientes a cero.

Recordemos que para este método es necesario encontrar el parámetro λ, por lo que haremos un análisis de este parámetro. Primero veamos qué valores podría tomar λ en la figura 4.18 observamos que entre más grande es el logaritmo de λ, más grande es la devianza y nos interesa tener una devianza relativamente pequeña por lo que el valor óptimo de lambda podría ser cercano $log(\lambda) = -8$.

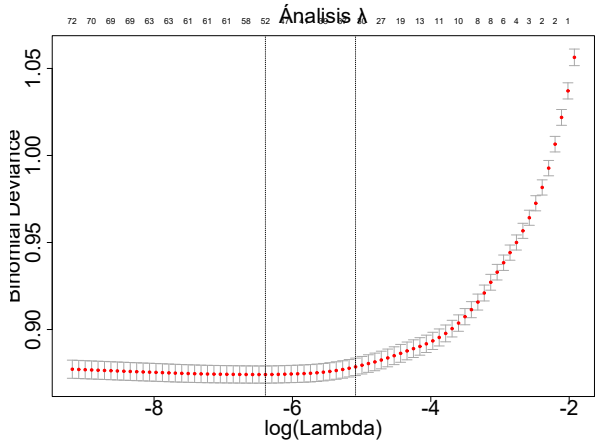

Figura 4.18: Logaritmo de λ respecto a devianza para regresión Lasso.

Por otro lado podemos ver el comportamiento de los coeficientes de regresión según los valores de λ, en la figura 4.19 se puede apreciar cómo los coeficientes toman el valor de cero a medida que el valor de λ aumenta.

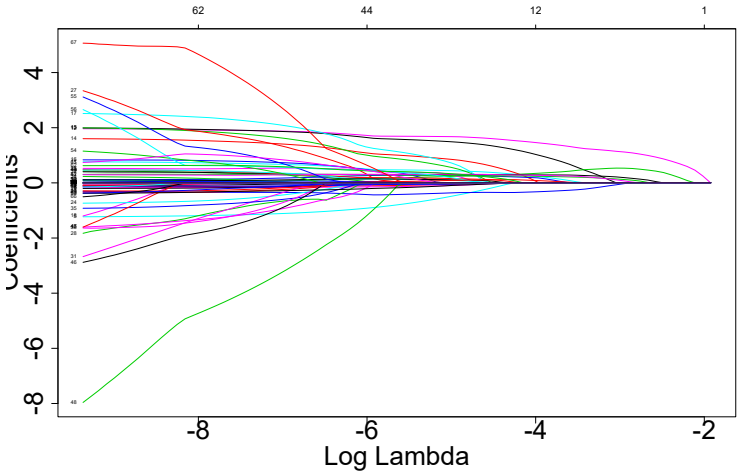

Figura 4.19: Coeficientes respecto log(λ) para regresión Lasso.

Una vez que hicimos este análisis podemos ajustar el modelo tomando el λ mínimo que es aquel que minimiza la devianza o bien tomando λ con el menor error estándar asociado de los coeficientes estimados, ajustaremos un modelo para ambos casos.

Veamos los coeficientes que obtenemos al ajustar del modelo con λ mínimo y λ con el menor error estándar podemos ver que ambos modelos redujeron varios coeficientes a cero y a diferencia del modelo de componentes principales ambos modelos conservaron la variable de BILL_AMT2 en lugar de BILL_AMT1 ya que esta variable fue propuesta por tener mayor peso en el la primera componente principal, PC1.

```
80 x 2 sparse Matrix of class "dgCMatrix"
                 MIN        SE                    MIN        SE
(Intercept) -1.061133e+00 -1.124995e+00    PAY_67       .            .
LIMIT_BAL   -2.061176e-06 -1.724158e-06    PAY_68       1.244541e+00 .
SEX2        -1.183374e-01 -6.128150e-02    BILL_AMT1    .            .
EDUCATION2   .             .               BILL_AMT2    1.927760e-06 3.335560e-07
EDUCATION3   .             .               BILL_AMT3    .            .
EDUCATION4  -1.032349e+00 -5.369640e-01    BILL_AMT4    .            .
MARRIAGE2   -1.567080e-01 -1.050380e-01    BILL_AMT5    .            .
MARRIAGE3    .             .               BILL_AMT6    .            .
AGE          2.309727e-03  5.613281e-05    PAY_AMT2    -6.832709e-06 -2.216235e-06
PAY_0-1      2.041335e-01  .               PAY_AMT3    -1.065086e-06  .
PAY_00      -3.627554e-01 -3.716825e-01    PAY_AMT4    -2.186196e-06 -6.500090e-07
PAY_01       6.122623e-01  4.472973e-01    PAY_AMT5    -2.577339e-06 -7.081755e-08
PAY_02       1.825321e+00  1.679598e+00    PAY_AMT6    -1.559015e-06  .
PAY_03       1.788276e+00  1.506816e+00    PAY_55       .             .
PAY_04       1.284089e+00  8.772182e-01    PAY_56       1.328854e-01  .
PAY_05       1.365055e+00  6.163101e-01.   PAY_57       4.978171e-01  2.305426e-01
PAY_06       6.499276e-01  .               PAY_58       .             .
PAY_07       1.787474e+00  7.993827e-01.   PAY_6-1       .             .
PAY_08       .             .               PAY_62       3.552315e-03  8.210150e-02
PAY_2-1     -9.553264e-02  .               PAY_63       5.094018e-01  3.481635e-01
PAY_20       .             .               PAY_64      -4.180517e-02  .
PAY_21      -1.233513e-01  .               PAY_65       .             .
PAY_22       1.083779e-01  2.275177e-01.   PAY_66       6.430696e-01  .
PAY_23       2.155109e-01  2.781347e-01.   AIC 19632.38     19760.83
PAY_24      -2.282324e-01  .               Deviance 19528.38 19696.83
PAY_25       5.087558e-01  3.414433e-01
PAY_26       .             .
PAY_27       8.800842e-01  .
PAY_28      -6.241722e-01  .
PAY_3-1     -5.261741e-02 -5.058803e-02
PAY_30       .             .
PAY_31       .             .
PAY_32       3.100338e-01  3.002982e-01
PAY_33       2.558001e-01  1.617990e-01
PAY_34       .             .
PAY_35      -3.585505e-01  .
PAY_36       .             .
PAY_37       .             .
PAY_38      -1.451798e-02  .
PAY_4-1     -1.694867e-02  .
PAY_40       .             .
PAY_41       .             .
PAY_42       2.976705e-01  2.665264e-01
PAY_43       6.156765e-02  .
PAY_44       2.533087e-01  9.995716e-02
PAY_45      -6.013352e-01  .
PAY_46       .             .
PAY_47       .             .
PAY_48      -2.235265e+00  .
PAY_5-1     -9.823254e-02 -1.135521e-02
PAY_50       .             .
PAY_52       3.015053e-01  3.129330e-01
PAY_53       1.324859e-01  1.155206e-01
PAY_54      -7.107195e-02  .
```

Revisemos más a detalle el comportamiento de los coeficientes de ambos modelos con una representación gráfica en la figura 4.20

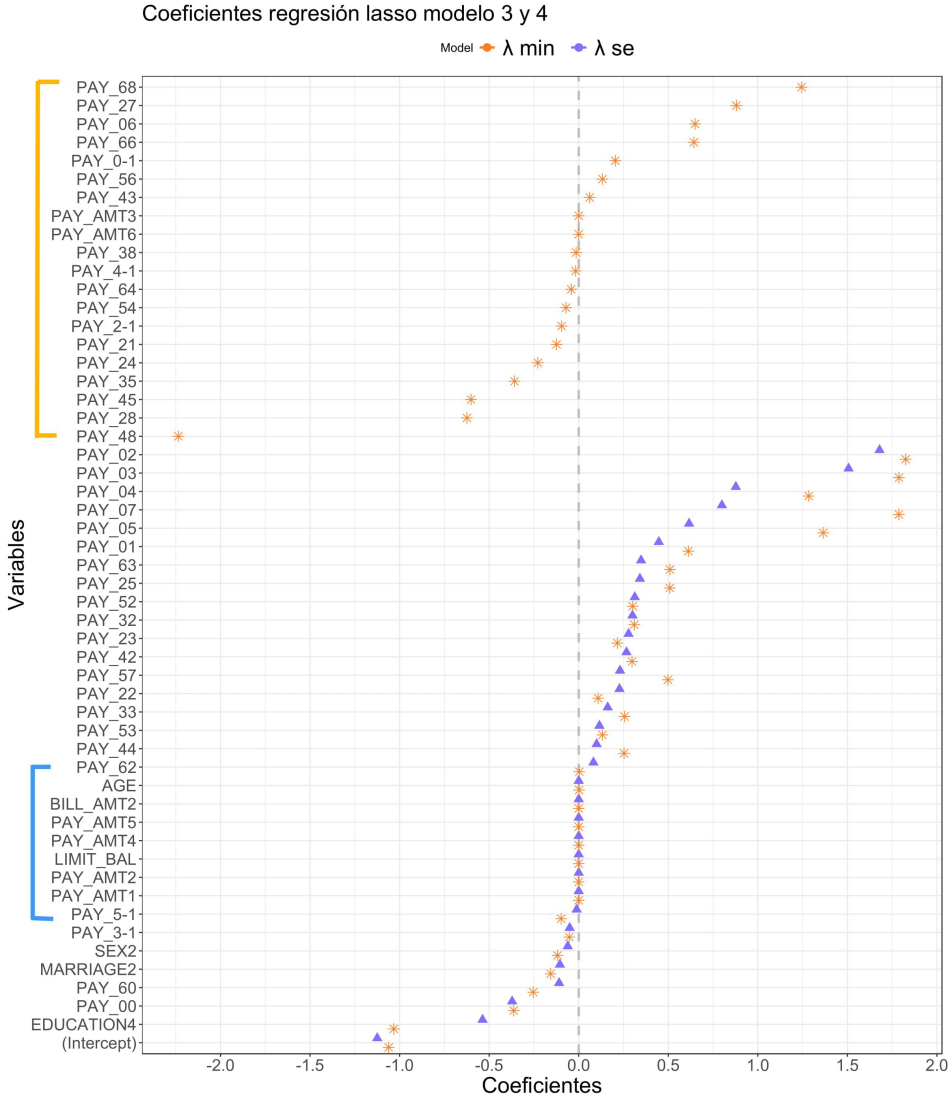

Figura 4.20: Coeficientes modelo 4 y 5, regresión Lasso λ min y λ se.

Es evidente que en las variables marcadas con una línea amarilla en el modelo de λ mínimo tenemos coeficientes estimados algunos con valores muy significativos como *PAY*4_8 o *PAY*6_8 pero para el modelo λ de estas variables tienen un coeficiente

de cero por lo que no son relevantes en la estimación y no aparecen en la gráfica. Revisemos los siguientes casos en donde ambos modelos estimaron un coeficiente, existen variables en los que los coeficientes fueron casi idénticos que son las variables marcadas en azul en donde en ambos modelos los coeficientes son muy cercanos a cero, por otro lado tenemos las variables sin marcar en donde *lambda* mínimo son mayores que en el modelo de λ se. Aunque ya analizamos las diferencias que existen entre los coeficientes de cada modelo, ahora tenemos que calcular las medidas de ajuste para ver que tan eficientes son dichos modelos. Iniciemos con el modelo de λ mínimo. En la figura 4.16 podemos visualizar la matriz de confusión del lado derecho veremos que tiene una gran cantidad de verdaderos negativos 5556 representando a aquellos clientes que pagaron su deuda y 578 de verdaderos positivos que representa a los clientes que incumplieron su deuda y el modelo los clasificó de forma correcta.

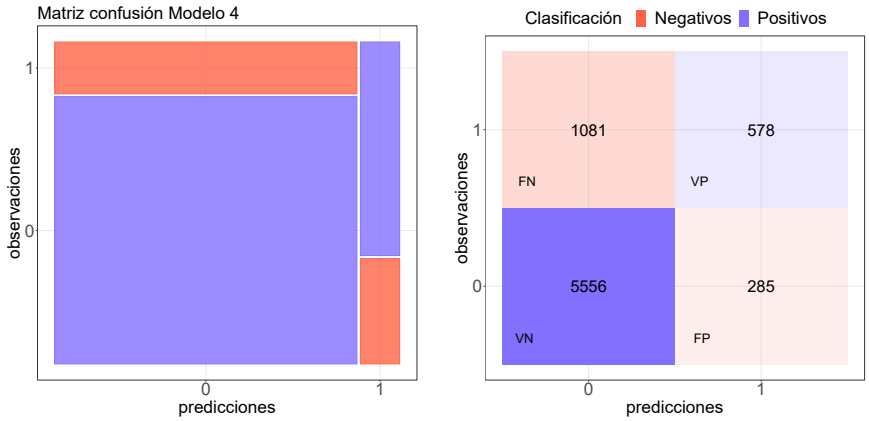

Figura 4.21: Matriz de confusión modelo 4, regresión lasso λ mín.

Ahora enfoquémonos en los casos en el que el modelo no acertó en la clasificación que es la cantidad de falsos negativos, que en este caso son los clientes que incumplieron su deuda pero se clasificaron como clientes que cumplieron su deuda, podemos ver que es el segundo mosaico más grande del lado izquierdo con un valor de 1081 clientes, y por otro lado el mosaico más pequeño representa a los clientes que pagaron su deuda pero fueron clasificados como deudores con 285 clientes.

Veamos el mismo análisis, esta vez para el modelo creado con que representa el menor error estándar asociado, en la figura 4.22 podemos observar que la mayor cantidad de observaciones se encuentra en el área correspondiente a verdaderos negativos con 5571 de individuos correctamente clasificados como clientes que pagaron su deuda, seguido de aquellos clientes que no pagaron su deuda. Sin embargo la predicción dicta que si pagaron su deuda, los casos restantes representan una mínima cantidad dada por verdaderos positivos que son 557 clientes que no pagaron su deuda y los modelo los clasificó correctamente en deudores, y por último los falsos positivos que son aquellos clientes que pagaron su deuda, no obstante el modelo los clasificó como deudores.

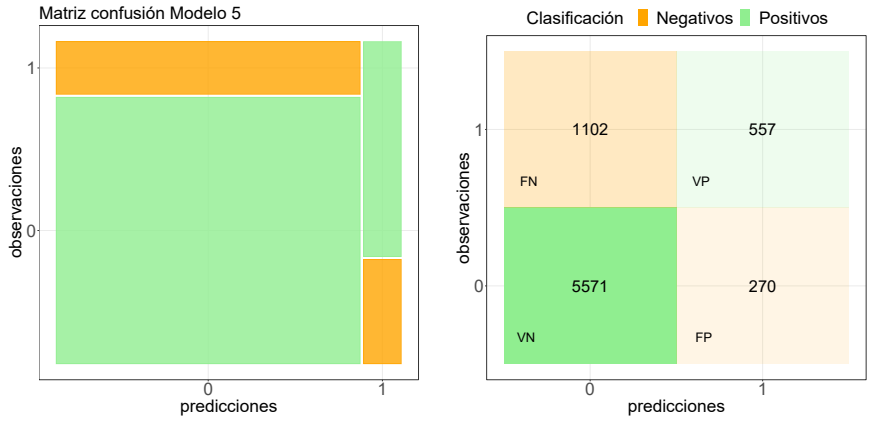

Figura 4.22: Matriz de confusión modelo 5, regresión Lasso λ se.

Finalmente revisamos cuál es el ajuste de la curva ROC para ambos modelos, el valor del área bajo la curva en la figura 4.23 en el modelo 4 correspondiente a λ min se aprecia una curva significativamente pegada a la esquina superior izquierda y tiene un AUC de 0.7627 lo cual indica un ajuste relativamente bueno.

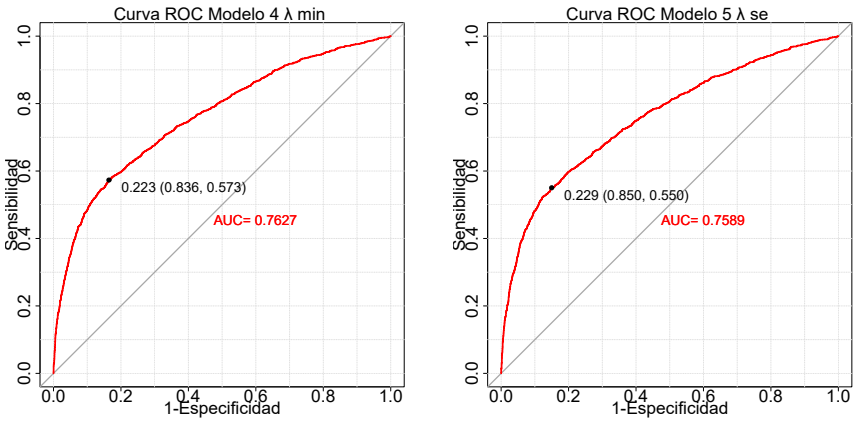

Figura 4.23: Curva Roc modelo 4 y 5, regresión Lasso λ mín y se.

Del lado derecho de la figura tenemos la curva ROC del modelo 5 construido con λ se en donde la curva se observa relativamente menos pegada a la esquIna superior derecha comparada con el modelo 3 y cuenta con un AUC de 0.7589, por lo que en este caso el modelo 3 tiene un mejor desempeño.

Por último analicemos el resultado del VIF para descartar la ausencia de multicolinealidad. Como podemos ver todos los valores son menores a 5 por lo que no se tiene multicolinealidad grave ni multicolinealidad débil para este modelo λ se.

```
Variable        MIN      SE
```

```
1   LIMIT_BAL 1.526377 1.478993
2        SEX2 1.024022 1.022474
3  EDUCATION4 1.005905 1.004399
4   MARRIAGE2 1.291765 1.290920
5         AGE 1.321384 1.320171
6     PAY_0.1 4.491926       NA
7      PAY_00 4.972666 2.817313
8      PAY_01 2.823238 1.666319
9      PAY_02 2.784974 1.900474
10     PAY_03 1.402256 1.265953
11     PAY_04 1.294696 1.251225
12     PAY_05 1.373788 1.358535
13     PAY_06 1.852858       NA
14     PAY_07 1.876112 1.228259
15    PAY_2.1 3.610174       NA
16     PAY_21 1.015617 1.012359
17     PAY_22 2.783496 2.473502
18     PAY_23 1.451977 1.404870
19     PAY_24 1.973914 1.949618
20     PAY_25 2.584352 1.657826
21     PAY_27 2.675332       NA
22     PAY_28 1.029838       NA
23    PAY_3.1 3.068781 1.454162
24     PAY_32 2.181334 2.041398
25     PAY_33 1.666329 1.598416
26     PAY_34 2.470092 2.053811
27     PAY_35 2.263431       NA
28     PAY_38 1.455366 1.106237
29    PAY_4.1 2.737360       NA
30     PAY_42 2.219484 2.007296
31     PAY_43 1.573410       NA
32     PAY_44 2.340894 1.732775
33     PAY_45 3.067331 1.203544
34     PAY_48 1.565474       NA
35    PAY_5.1 2.249636       NA
36     PAY_52 2.275752 2.127225
37     PAY_53 1.507880 1.321545
38     PAY_54 2.467506       NA
39     PAY_56 1.134493 1.104036
40     PAY_57 1.218679 1.205093
41     PAY_60 2.297086 2.024788
42     PAY_62 2.231431 2.181589
43     PAY_63 1.601601 1.315508
44     PAY_64 1.433938 1.238959
45     PAY_66 1.200583 1.162496
46     PAY_68 1.487149       NA
47  BILL_AMT2 1.759585 1.721014
48   PAY_AMT1 1.296250 1.230157
49   PAY_AMT2 1.209992 1.161295
50   PAY_AMT3 1.229090       NA
```

```
51    PAY_AMT4 1.191051 1.106172
52    PAY_AMT5 1.118328 1.107233
53    PAY_AMT6 1.115299       NA
```

4.2.6. Regresión GROUP-LASSO

En la sección anterior definimos las variables categóricas como *dummies* por lo que al momento de realizar las penalizaciones convierte en cero los coeficientes de algunas variables *dummy* como es el caso de las variable $EDUCATION2$ y $EDUCATION3$ que se encuentran en el *summary* del modelo de regresión Lasso.

El método de Group Lasso consiste en reducir a cero las variables menos significativas pero en caso de ser categórica reduce a cero todos los factores que pertenecen a dicha variable a diferencia de la regresión Lasso.

Al igual que en la regresión Lasso, esta regresión depende del parámetro , a medida que el λ aumenta el número de regresores estimados disminuye hasta que todos los coeficientes son penalizados y se reducen cero, esto sucede con el λ máximo que en este caso es de 1462.99. Por otro lado se puede observar el comportamiento de la λ en función de la devianza en la figura 4.24 además podemos ver que la gráfica es muy similar a la de la regresión Lasso que se encuentra en la figura 4.19.

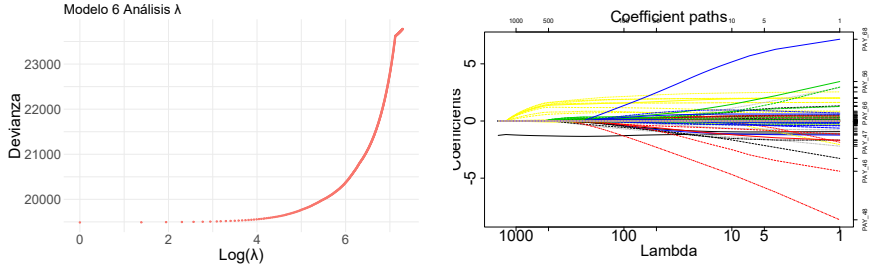

Figura 4.24: Logaritmo de λ contra devianza para regresión Group Lasso y Logaritmo de λ contra coeficientes para regresión Group Lasso

También se puede evaluar el comportamiento del parámetro λ contra los coeficientes dando como resultado la figura 4.24 en donde podemos observar que a medida que λ aumenta entonces existen más coeficientes que se reducen a cero.

Ajustemos el modelo para analizar el comportamiento de los regresores. En este caso tomaremos el λ mínimo con el que tenemos los siguientes resultados:

```
fit.splice$coefficients
                      100.
(Intercept) -1.278244e+00.
LIMIT_BAL   -1.792541e-06.      PAY_41      -2.973113e-01
SEX2        -7.423403e-02.      PAY_42       1.711263e-01
EDUCATION2   2.183644e-02.      PAY_43       1.001119e-01
EDUCATION3   2.665763e-02.      PAY_44       1.704600e-01
EDUCATION4  -4.134875e-01.      PAY_45      -2.974080e-01
MARRIAGE2   -9.263010e-02.      PAY_46      -3.501896e-01
MARRIAGE3   -5.831749e-02.      PAY_47       1.495687e-01
AGE          1.652187e-03.      PAY_48      -1.316148e+00
PAY_0-1      2.812699e-01.      PAY_5-1     -8.130970e-02
PAY_00      -1.535737e-01.      PAY_50      -6.017733e-02
```

PAY_01	7.431875e-01.	PAY_52	2.137543e-01
PAY_02	1.880107e+00.	PAY_53	1.776311e-01
PAY_03	1.904465e+00.	PAY_54	-3.516271e-02
PAY_04	1.571576e+00.	PAY_55	1.999779e-01
PAY_05	1.579816e+00.	PAY_56	3.976373e-01
PAY_06	1.159449e+00.	PAY_57	2.885898e-01
PAY_07	2.146542e+00.	PAY_6-1	-4.045538e-02
PAY_08	8.099949e-01.	PAY_60	-7.847065e-02
PAY_2-1	-3.484501e-02.	PAY_62	1.017157e-01
PAY_20	-1.507765e-02.	PAY_63	3.051650e-01
PAY_21	-1.573454e-01.	PAY_64	-1.094051e-01
PAY_22	1.068361e-01.	PAY_65	1.437953e-01
PAY_23	1.741953e-01.	PAY_66	4.396232e-01
PAY_24	-1.064678e-01.	PAY_67	7.988206e-02
PAY_25	1.740752e-01.	PAY_68	1.355135e+00
PAY_26	1.361742e-01.	BILL_AMT1	0.000000e+00
PAY_27	1.642371e-01.	BILL_AMT2	6.806647e-07
PAY_28	-6.939290e-01.	BILL_AMT3	0.000000e+00
PAY_3-1	-5.024822e-02.	BILL_AMT4	0.000000e+00
PAY_30	-1.100854e-02.	BILL_AMT5	0.000000e+00
PAY_31	-5.709212e-01.	BILL_AMT6	0.000000e+00
PAY_32	2.369227e-01.	PAY_AMT1	-4.452381e-06
PAY_33	2.431449e-01.	PAY_AMT2	-3.583336e-06
PAY_34	1.280573e-01.	PAY_AMT3	-2.968225e-07
PAY_35	-4.053958e-01.	PAY_AMT4	-1.300147e-06
PAY_36	1.408378e-01.	PAY_AMT5	-5.903695e-07
PAY_37	1.814275e-01.	PAY_AMT6	-3.151476e-07
PAY_38	-4.804351e-01		
PAY_4-1	-4.561914e-02		
PAY_40	-3.020860e-02		

Podemos ver que sólo penalizó a las variables de estado de cuenta, conservando sólo la que corresponde a *BILL_AMT*2 a diferencia de la regresión Lasso que penalizó a las categorías de la variable *EDUCACION*, *PAY*_0, entre otras, la regresión Group Lasso si calculó un coeficiente para cada una de estas variables.

Ahora analicemos el desempeño de este modelo, comenzando con la matriz de confusión en la figura 4.25 podemos ver que al igual que en los modelos anteriores la mayor parte de clientes se encuentra concentrada en la sección de verdaderos negativos que son los clientes que pagaron su deuda incluso este número es mayor que el resultado que obtuvimos en la regresión lasso, tiene una cantidad de verdaderos positivos también similar aunque menor a los modelos antes vistos.

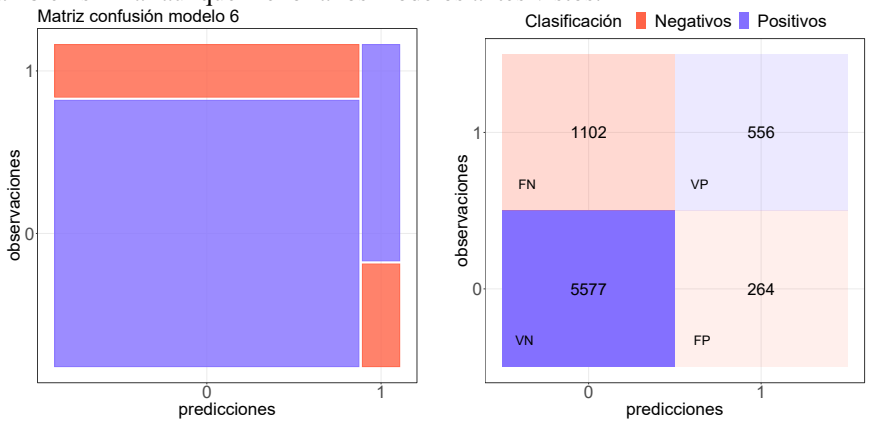

Figura 4.25: Matriz de confusión modelo 6, Group Lasso.

Por último analicemos la curva ROC, en la figura 4.26 podemos ver que es una curva bastante buena con un AUC de 0.7616 muy parecido al de la regresión lasso por lo que el ajuste del modelo es potencialmente bueno además con esta técnica lidiamos con las consecuencias de la multicolinealidad dado que al acotar los coeficientes reducimos los casos en los que los que los coeficientes resultan grandes o con signo contrario ver sección 3.1. Además tiene un mejor manejo de las variables categóricas.

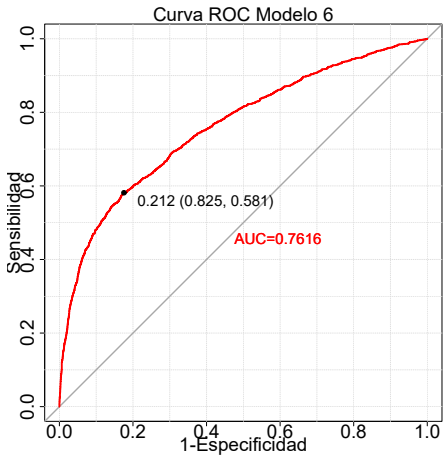

Figura 4.26: Curva Roc modelo 6, Group Lasso.

Por último analicemos el resultado del VIF y GVIF para descartar la ausencia de multicolinealidad. Como podemos ver ya no existe un GVIF con valor mayor a 4, por lo que ya no existiría presencia de multicolinealidad grave pero si una multicolinealidad débil. en las variables PAY_3,PAY_4 y PAY_6.

```
                GVIF Df GVIF^(1/(2*Df))
SEX        1.028777e+00  1        1.014286
EDUCATION  1.243547e+00  3        1.036996
MARRIAGE   1.347211e+00  2        1.077355
LIMIT_BAL  1.674148e+00  1        1.293889
AGE        1.396666e+00  1        1.181806
PAY_0      1.337009e+08 10        2.548630
PAY_2      2.157806e+09 10        2.928874
PAY_3      4.083677e+12 10        3.971226
PAY_4      4.670127e+13 10        3.824657
PAY_5      9.507421e+08  9        3.153416
PAY_6      4.150339e+03  9        1.588564
BILL_AMT2  1.951165e+00  1        1.396841
PAY_AMT1   1.266888e+00  1        1.125561
PAY_AMT2   1.203219e+00  1        1.096914
PAY_AMT3   1.170623e+00  1        1.081953
PAY_AMT4   1.154212e+00  1        1.074342
PAY_AMT5   1.137554e+00  1        1.066562
PAY_AMT6   1.112420e+00  1        1.054713
```

4.2.7. Regresión Ridge

La regresión Ridge es el siguiente método que aplicaremos para tratar la multicolinealidad, en este caso al igual que en la regresión Lasso al ser un método que aplica penalizaciones también debemos analizar los parámetros de λ, sin embargo la diferencia será que las penalizaciones van a reducir los coeficientes menos significativos a una cantidad muy cercana a cero pero nunca será cero como en la regresión Lasso.

Comencemos analizando los parámetros, en la figura 4.28 podemos observar que el logaritmo de λ con la menor devianza es -4 y a medida que el valor de logaritmo avanza el valor de la devianza aumenta.

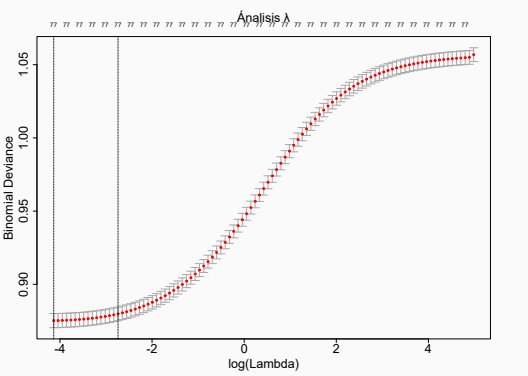

Figura 4.27: Logaritmo de λ contra devianza para regresión Ridge.

Por otro lado podemos observar el comportamiento de los coeficientes dependiendo del parámetro λ en la figura 4.28 en donde podemos apreciar que a medida que el logaritmo de λ aumentan los coeficientes tienden a cero, es decir que entre más grande sea el valor de λ que seleccionemos para construir el modelo van a existir más variables con coeficientes relativamente pequeños dependiendo de la importancia que tenga la variable.

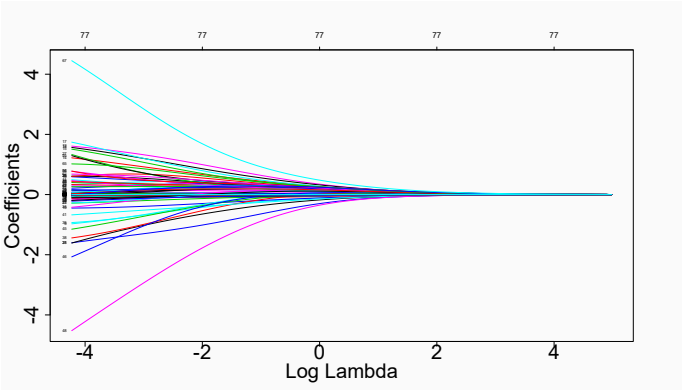

Figura 4.28: Coeficientes contra devianza para regresión Ridge.

Ahora veamos los coeficientes que obtuvimos con λ mínimo y con λ con el menor error estándar asociado

```
80 x 2 sparse Matrix of class "dgCMatrix"
                      s0            s0
(Intercept) -1.090487e+00 -1.123174e+00.    PAY_66       1.012596e+00  7.969635e-01
LIMIT_BAL   -1.799157e-06 -1.299074e-06.    PAY_67      -3.895620e-03  1.947983e-01
SEX2        -1.246073e-01 -9.769964e-02.    PAY_68       4.351781e+00  2.511890e+00
EDUCATION2   2.958570e-02  4.281792e-02.    BILL_AMT2    1.028753e-06  4.791636e-07
EDUCATION3   2.234912e-02  4.962927e-02.    BILL_AMT3    4.590901e-07  2.081726e-07
EDUCATION4  -9.535733e-01 -6.117559e-01.    BILL_AMT4    1.365072e-07  1.164983e-07
MARRIAGE2   -1.503758e-01 -1.092006e-01.    BILL_AMT5    6.332771e-08  6.155058e-08
MARRIAGE3   -7.762681e-02 -3.809323e-02.    BILL_AMT6   -5.363457e-08  1.723464e-08
AGE          2.920884e-03  2.256645e-03.    PAY_AMT1    -6.596627e-06 -4.305987e-06
PAY_0-1      1.352565e-01  3.815850e-03.    PAY_AMT2    -5.892024e-06 -3.244815e-06
PAY_00      -4.451946e-01 -3.689410e-01.    PAY_AMT3    -2.093085e-06 -2.141527e-06
PAY_01       4.800908e-01  2.802867e-01.    PAY_AMT4    -3.351215e-06 -3.028232e-06
PAY_02       1.594052e+00  1.245057e+00.    PAY_AMT5    -3.285711e-06 -2.678786e-06
PAY_03       1.543445e+00  1.111281e+00.    PAY_AMT6    -2.461409e-06 -2.160150e-06
PAY_04       1.201771e+00  8.532887e-01.    AIC        19688.81        19849.59
PAY_05       1.480801e+00  9.692492e-01.    deviance     19534.81       19695.59
PAY_06       5.840975e-01  4.443792e-01
PAY_07       1.705848e+00  1.090014e+00
PAY_08      -1.899838e-01  1.175339e-01
PAY_2-1     -7.568128e-02 -5.847661e-02
PAY_20       3.518074e-02 -5.498009e-02
PAY_21      -2.795011e-01 -1.820656e-01
PAY_22       2.193571e-01  3.375294e-01
PAY_23       3.352730e-01  4.323369e-01
PAY_24      -4.048177e-01 -1.030930e-01
PAY_25       6.103182e-01  5.088723e-01
PAY_26       6.160231e-01  6.581733e-01
PAY_27       1.259867e+00  6.120337e-01
PAY_28      -1.557905e+00 -1.228351e+00
PAY_3-1     -4.994207e-02 -7.117488e-02
PAY_30       4.645410e-02  2.865264e-03
PAY_31      -1.562192e+00 -9.165586e-01
PAY_32       3.305012e-01  2.902829e-01
PAY_33       3.962774e-01  3.378580e-01
PAY_34       1.605403e-01  1.567483e-01
PAY_35      -9.202487e-01 -5.852360e-01
PAY_36       4.828599e-01  2.645447e-01
PAY_37       1.691613e-02  1.083330e-01
PAY_38      -1.412342e+00 -8.710652e-01
PAY_4-1     -3.318375e-02 -4.284285e-02
PAY_40      -2.163502e-02 -2.576392e-02
PAY_41      -6.629310e-01 -4.841476e-01
PAY_42       2.870937e-01  2.771175e-01
PAY_43       6.289934e-02  1.299395e-01
PAY_44       4.242867e-01  3.428461e-01
PAY_45      -1.117773e+00 -6.227022e-01
PAY_46      -2.021536e+00 -8.508009e-01
PAY_47      -3.377870e-02  2.374522e-01
PAY_48      -4.405600e+00 -2.613936e+00
PAY_5-1     -1.048832e-01 -7.093988e-02
PAY_50      -2.971048e-02 -2.910470e-02
PAY_52       2.625586e-01  2.679226e-01
PAY_53       1.252971e-01  1.993653e-01
PAY_54      -2.714364e-01 -1.125272e-01
PAY_55       7.494951e-01  4.499122e-01
PAY_56       1.243547e+00  6.305796e-01
PAY_57       7.539054e-01  4.493745e-01
PAY_58       .             .
PAY_6-1      1.200806e-02  1.820818e-02
PAY_60      -2.119626e-01 -1.102753e-01
PAY_61       .             .
PAY_62       7.829486e-02  1.722727e-01
PAY_63       6.759356e-01  5.709165e-01
```

```
PAY_64      -1.863530e-01 -6.424111e-02
PAY_65       3.305181e-02  1.652403e-01
```

En el caso de la regresión Ridge podemos observar que tanto las variables de estado de cuenta de abril a septiembre y las variables de pago por mes tienen coeficientes muy cercanos a cero en ambos modelos, en donde existen diferencias más notorias es en las variables categóricas puesto que para la variable de educación, $EDUCATION4$, los coeficientes del modelo con el λmn y con el λse muestran una diferencia significativa.

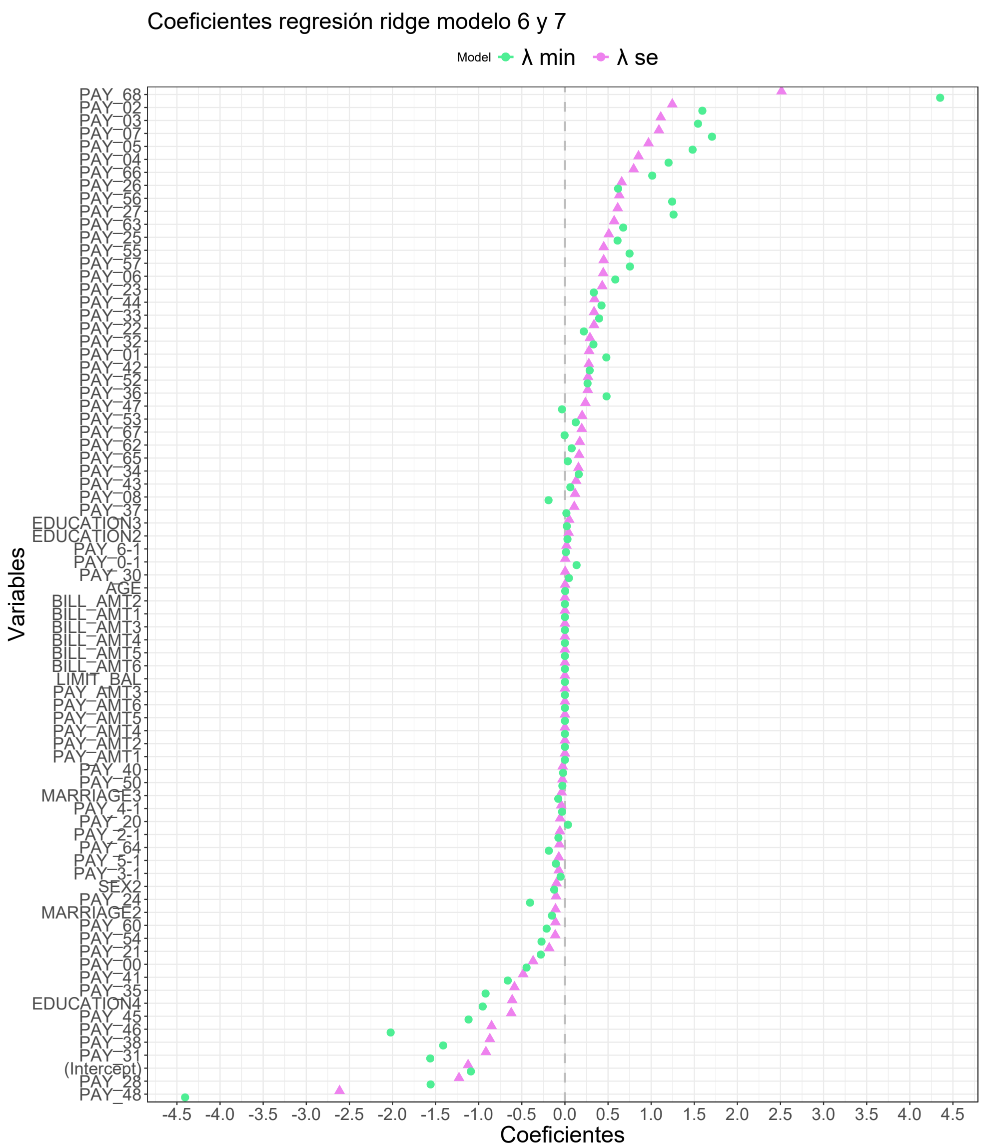

Figura 4.29: Coeficientes modelo 6 y 7, regresión Ridge λ mín. y λ se.

Revisemos cuáles son las medidas de ajuste para estos modelos, iniciemos con el modelo construido a partir de λ mínimo, en la figura 4.30, en este modelo el desempeño es similar a los modelos de Lasso y Group Lasso ya que la mayoría de las observaciones se encuentran clasificados como verdaderos negativos, es decir, individuos que pagaron su deuda y el modelo los clasificó correctamente con un total de 5562 observaciones, el siguiente caso en el que el modelo hizo una clasificación correcta corresponde a 572 clientes que representa a los clientes que incumplieron su deuda y el modelo los clasificó como deudores, por otro lado revisemos los casos en el que el modelo no acertó, dicho de otra forma los falsos positivos con 279 casos y falsos negativos con 1087 casos.

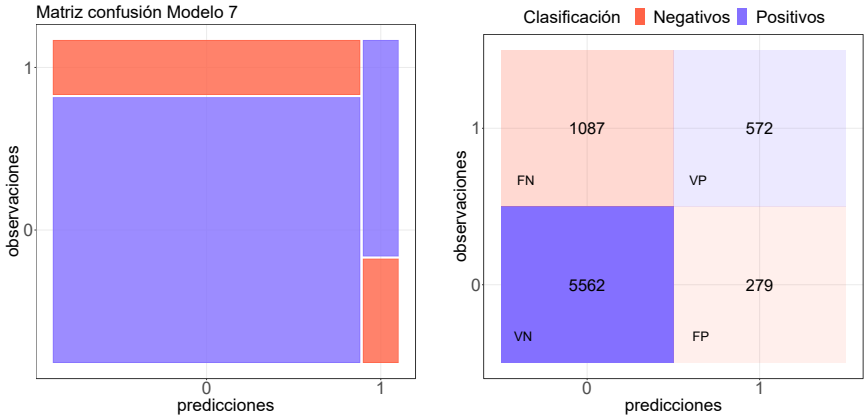

Figura 4.30: Matriz de confusión modelo 7, regresión Ridge λ mín.

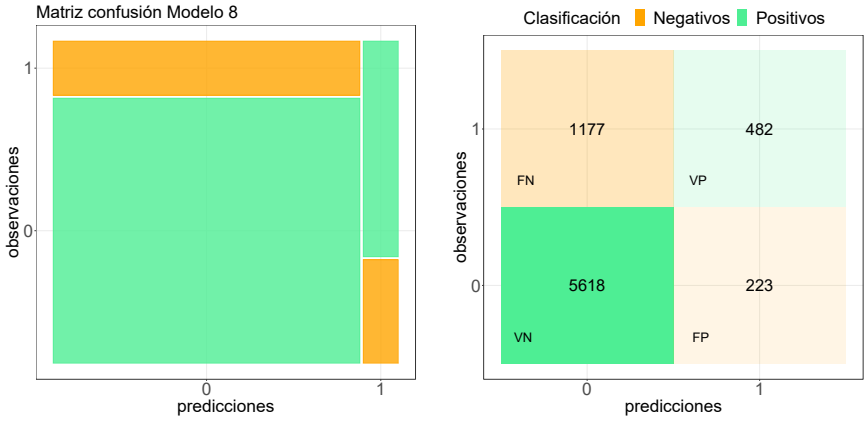

Figura 4.31: Matriz de confusión modelo 8, regresión Ridge λ se.

Por otro lado, evaluando el modelo de regresión Ridge con λ con el menor error estándar asociado podemos ver su matriz de confusión en la figura 4.31 vemos el mismo

comportamiento puesto que agrupa la mayor cantidad de clientes en verdaderos negativos, pero en este caso la cantidad es mayor comparado con el modelo 7, sin embargo en la clasificación de verdaderos positivos el modelo 7 es mejor ahora revisemos los casos en el que el modelo no acertó en las predicciones formado por los falsos negativos con 1177 casos, es decir clientes deudores clasificados como que pagarían su deuda y falsos positivos con 223 observaciones representando a los clientes que pagaron su deuda pero fueron clasificados como deudores por el modelo.

Las cantidades de las matrices de confusión varían por lo que es importante comparar las curvas ROC ver figura 4.32 de ambos modelos para decidir cuál tiene mejor desempeño. Notemos que gráficamente no se observa diferencia a simple vista en la curva ROC, pero si revisamos el valor del AUC notamos que el modelo 8 construido con λse tiene una área bajo la curva de 0.7638 que es mayor al área bajo la curva del modelo 7 con un valor de 0.7634 por lo que dicho modelo tiene un mejor desempeño.

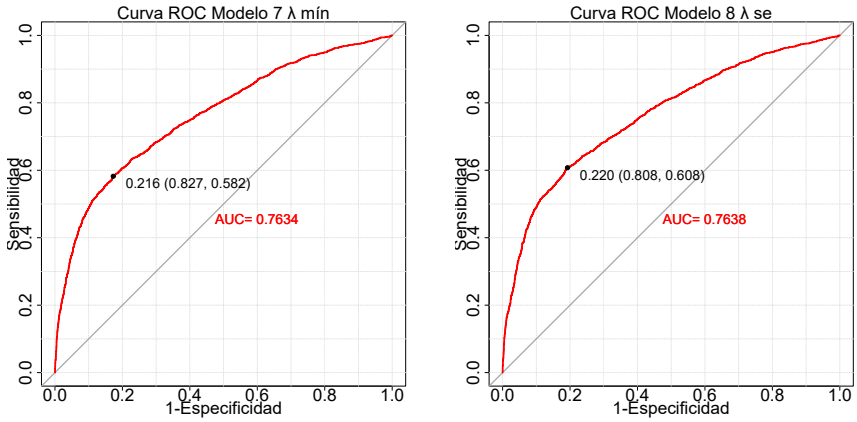

Figura 4.32: Curva Roc modelo 7 y 8, regresión Ridge λ mín y se.

Por último analicemos el valor del VIF, en este caso observamos valores menores a 5 por lo descartamos la presencia de multicolinealidad.

```
LIMIT_BAL     0.70109 PAY_40       0.52208
SEX2          0.61118 PAY_41       0.66223
EDUCATION2    0.64512 PAY_42       0.61215
EDUCATION3    0.64726 PAY_43       0.64717
EDUCATION4    0.61165 PAY_44       0.67861
MARRIAGE2     0.64572 PAY_45       0.70751
MARRIAGE3     0.61297 PAY_46       0.53350
AGE           0.65364 PAY_47       0.29601
PAY_0.1       0.63049 PAY_48       0.61570
PAY_00        0.51356 PAY_5.1      0.64570
PAY_01        0.60744 PAY_50       0.52019
PAY_02        0.60436 PAY_52       0.62552
PAY_03        0.63158 PAY_53       0.64258
PAY_04        0.63280 PAY_54       0.68459
```

```
PAY_05      0.64712 PAY_55      0.64006
PAY_06      0.66967 PAY_56      0.62950
PAY_07      0.60364 PAY_57      0.26574
PAY_08      0.42776 PAY_6.1     0.66976
PAY_2.1     0.65187 PAY_60      0.58045
PAY_20      0.48198 PAY_62      0.62977
PAY_21      0.62805 PAY_63      0.66466
PAY_22      0.62244 PAY_64      0.65826
PAY_23      0.63058 PAY_65      0.64321
PAY_24      0.67081 PAY_66      0.60413
PAY_25      0.67275 PAY_67      0.48060
PAY_26      0.55560 PAY_68      0.59055
PAY_27      0.24304 BILL_AMT1   0.53293
PAY_28      0.61236 BILL_AMT2   0.37714
PAY_3.1     0.66680 BILL_AMT3   0.39581
PAY_30      0.53412 BILL_AMT4   0.39570
PAY_31      0.66660 BILL_AMT5   0.38404
PAY_32      0.59613 BILL_AMT6   0.50742
PAY_33      0.65401 PAY_AMT1    0.64477
PAY_34      0.67773 PAY_AMT2    0.62235
PAY_35      0.69911 PAY_AMT3    0.63989
PAY_36      0.26645 PAY_AMT4    0.63177
PAY_37      0.70368 PAY_AMT5    0.62282
PAY_38      0.63431 PAY_AMT6    0.63457
PAY_4.1     0.65465
```

4.3. Resultado

Una vez que tenemos la evaluación de todos los modelos haremos un comparativo para elegir cuál es el modelo que mejor se ajusta a los datos en la tabla 4.9 podemos observar los valores de las pruebas de bondad de ajuste dependiendo del modelo.

Iniciemos identificando el modelo con el menor AIC, éste corresponde al modelo 4 construido mediante regresión Lasso con λ mínimo, este modelo a su vez también tiene la menor devianza y la mayor cantidad de verdaderos negativos y verdaderos positivos asociados, además si observamos la cantidad de falsos negativos es la más baja de todos los modelos, sin embargo al fijarnos en sus falsos positivos no cuenta con el menor número comparado con otros modelos en especial si lo comparamos con el modelo 8, Ridge λ con el menor error estándar asociado, ya que la diferencia es de 61 observaciones mal clasificadas, si adicional a esto observamos el valor del AUC en el modelo 8 es mayor que en el modelo 4 por décimas, pero si ponderamos el resultado que obtuvimos del AIC y la devianza en el modelo 8 es más adecuado el modelo 4 correspondiente a la regresión Lasso evaluado con el λ mínimo. Es importante mencionar que aunque el modelo 6 construido mediante regresión Group Lasso a pesar de que tiene un AUC significativo y tiene un mejor manejo de las variables categóricas, es uno de los modelos con mayor devianza y AIC por lo que por esta razón lo descartaríamos.

Comparación de modelos								
	Regresión Logística			Regresión Lasso		Group-Lasso	Regresión Ridge	
	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7	Modelo 8
	variables	Sel ACP	ACP	λ min	λ se	λ	λ min	λ se
AIC	19641	19643	19638	19632.388	19760.83	19793.32	19688.81	19849.59
Devianza	19487	19499	19494	19528.38	19696.83	19647.32	9534.81	19695.59
VP	564	540	415	578	557	556	572	482
VN	5555	5581	5653	5557	5571	5577	5562	5518
FP	286	280	188	284	270	264	279	223
FN	1095	1119	1244	1081	1102	1102	1087	1177
Accuracy	0.8325	0.8139	0.8090	0.818	0.8170	0.8178	0.8178	0.8108
AUC	0.7524	0.602	0.7185	0.7621	0.7589	0.7616	0.7634	0.7638
P.M.	si	no	no	no	no	no	no	no

Tabla 4.9: Tabla comparativa de modelos por pruebas de bondad de ajuste.

Donde PM es prescencia de multicolinealidad.

También podemos observar que la mayoría de los modelos tiene un ajuste relativamente bueno, pero en el caso del modelo 2 que es aquel que se construyó con regresión logística y la selección de variables mediante el análisis de componentes principales tiene un AUC muy bajo en comparación a los demás, el AIC es el mayor de todos los modelos, además, es uno de los que cuentan con la mayor devianza por lo que este sería el modelo menos adecuado para nuestra base de datos seguido del modelo 2 construido incluyendo la variable $PC1$ resultado del análisis de componentes principales. Por otro lado el modelo que mejor se ajusta a nuestras base es el modelo 4 que está construido con regresión lasso utilizando el λ mínimo.

Para la función de enlace logit, el logaritmo natural de las probabilidades depende de los coeficientes estimados.

A medida que aumentan las probabilidades logarítmicas aumentan las probabilidades de que el evento en cuestión suceda, en es que caso estudiemos los coeficientes de regresión Lasso para la variable de límite de crédito vemos un coeficiente negativo lo cual indica que por cada unidad que aumente el límite de crédito disminuye las probabilidades logarítmicas de que el individuo no pague su deuda en $-2.061176e-06$ por lo que a mayor límite de crédito menor probabilidad de incumplir, veamos la interpretación de algunas variables categóricas como el género en este caso sabemos que esta variable tiene los valores 1=hombre, 2=mujer y el coeficiente SEX2 es negativo esto indica que el cambio de hombre a mujer disminuye en $-1.183374e-01$ la probabilidad logarítmica de que el individuo no pague su deuda de modo que es más probable que una mujer pague su deuda a un hombre, por otro lado tenemos las variables que tiene más de dos categorías como PAY_0 aquí podemos visualizar los coeficientes en donde el cambio de pagar a tiempo a tener un mes de retraso aumenta la probabilidad logarítmica de incumplimiento de pago en $2.041335e-01$.

4.4. Conclusión

Aunque existe gran variedad de modelos que se pueden ajustar a una base de datos con variable respuesta Y binaria, no todas las técnicas serán las adecuadas u óptimas para la base de estudio. Es sumamente necesario familiarizarse con los datos a través

del análisis exploratorio de los datos y hacer las pruebas pertinentes para encontrar el modelo adecuado.

En lo particular, la base de datos que estudiamos mostró mejores resultados aplicando regresión Lasso, lo que indica que aunque el modelo de regresión logística es un excelente modelo de clasificación existen ocasiones en el que es necesario reforzarlo, en este caso, se reforzó con la penalización de los coeficientes como describimos en el capítulo 3 de esta manera pudimos observar que las pruebas de bondad de ajuste mostraron mejores resultados. Además notamos que algunas de las variables cuyos coeficientes convergieron a cero fueron las variables que tenían mayor VIF por lo que para este caso en particular, este método, ayudó a mitigar la multicolinealidad. Si lo viéramos desde el punto de vista de un banco utilizar el modelo adecuado significa una ganancia y disminuir significativamente el riesgo de que un cliente se convierta en deudor, de tal manera que al momento de aplicar el modelo a un conjunto de clientes potenciales se va a otorgar el crédito a los clientes con mayor probabilidad de pagar, o bien accionar planes para aquellos clientes que probablemente sean deudores con el objetivo de que dichos planes ayuden al cliente a pagar su deuda, estas decisiones depende de la entidad que esté utilizando el modelo estadístico. Es por ello que este tipo de modelos es de gran utilidad para tomar las mejores decisiones si se tiene un problema de clasificación con el problema de multicolinealidad.

Apéndice A

Ajuste regresión LASSO y Ridge

```
library(glmnet)
base <- fread("ruta/default of credit card clients (1).csv", header = TRUE)

base$MARRIAGE<-ifelse(base$MARRIAGE==0,3,base$MARRIAGE) #corregimos el valor de los cero
base$EDUCATION<-ifelse(base$EDUCATION==0,4,ifelse(base$EDUCATION==6,4,ifelse(base$EDUCATION==5,4,
base$SEX <- factor(base$SEX, levels = c("1","2"))
base$EDUCATION <- factor(base$EDUCATION, levels = c("1","2","3","4"))
base$MARRIAGE <- factor(base$MARRIAGE, levels = c("1","2","3"))
base$PAY_0 <- factor(base$PAY_0, levels = c("-2","-1","0","1","2","3","4","5","6","7","8"))
base$PAY_2 <- factor(base$PAY_2, levels = c("-2","-1","0","1","2","3","4","5","6","7","8"))
base$PAY_3 <- factor(base$PAY_3, levels = c("-2","-1","0","1","2","3","4","5","6","7","8"))
base$PAY_4 <- factor(base$PAY_4, levels = c("-2","-1","0","1","2","3","4","5","6","7","8"))
base$PAY_5 <- factor(base$PAY_5, levels = c("-2","-1","0","2","3","4","5","6","7","8"))
base$PAY_6 <- factor(base$PAY_6, levels = c("-2","-1","0","1","2","3","4","5","6","7","8"))
base$`default payment next month`<- factor(base$`default payment next month`, levels = c("0","1")

#Dividimos en training y test set

set.seed(123)
split = sample.split(base$`default payment next month`, SplitRatio = 0.75)
training_set_m1 = subset(base, split == TRUE)
test_set_m1 = subset(base, split == FALSE)

training_set_m1<-training_set_m1%>%select(-ID)
test_set_m1<-test_set_m1%>%select(-ID)
y<- training_set_m1$`default payment next month`
x <- model.matrix(`default payment next month`~.,training_set_m1)[,-1]
```

```
#valores de lambda contra la devianza
set.seed(123)
cv.lasso <- cv.glmnet(x,as.factor(y), alpha = 1, family = "binomial")
plot(cv.lasso,main="?nalisis ??",cex.axis=2.5,cex.lab=2.5,cex.main=2.5,font.main=1,adj

#gráficas lambda contra coeficientes
#vamos a revisar los coeficioentes de lamda sin  restricción de lamda
model<- glmnet(x,as.factor(y), alpha = 1, family = "binomial",
               standardize=TRUE)
plot(model,xvar=c("lambda"), label = TRUE,cex.axis=2.5,cex.lab=2.5,cex.main=2.5)+axis(s

# coeficioentes de lamda restringido
modelres<- glmnet(x,y, alpha = 1, family = "binomial",
                  standardize=TRUE,lower=-0.8,upper=0.8)
plot(modelres,xvar=c("lambda"), label = TRUE)
plot(modelres,xvar = "dev", label = TRUE)

#ajustamos el modelo LASSO con el lambda minimo y lambda se

modelmin <- glmnet(x,as.factor(y), alpha = 1, family = "binomial",
                lambda = cv.lasso$lambda.min)
modelse <- glmnet(x,as.factor(y), alpha = 1, family = "binomial",
                lambda = cv.lasso$lambda.1se)

#ajustamos el modelo Ridge con el labda minimo y se,
#cambia el valor de alpha

modelmin <- glmnet(x,as.factor(y), alpha = 0, family = "binomial",
                   lambda = cv.ridge$lambda.min)
modelse <- glmnet(x,as.factor(y), alpha = 0, family = "binomial",
                  lambda = cv.ridge$lambda.1se)
```

Bibliografía

[1] *AKAIKE, H. (1973)*. [*Information Theory an Extension of the Maximum Likelihood Principle*]. in Petrov, B. N.; Csáki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971,Budapest: Akadémiai Kiadó, pp. 267–281; republished in Kotz, S.; Johnson, N. L., eds. (1992), Breakthroughs in Statistics, I, Springer-Verlag, pp. 610–624.

[2] *Hastie, T., Tibshirani, R.,, Friedman, J. (2001).* [*The Elements of Statistical Learning.*] New York, NY, USA: Springer New York Inc pp 119-120.

[3] *Fox, J. and Monette, G. (1992)* [*Generalized collinearity diagnostics. JASA, 87, pp 178–183.*]

[4] *Douglas C. Montgomery/ Elizabeth A. Peck/ G. Geoffrey Vining. Tercera reimpresión (2006)*. [*Introducción al análisis de regresión lineal*]. Editorial Continental,Mexico DF. Título original Introduction to linear regressions analysis / Douglas C. Mongoery , John Wiley Sons 2002.

[5] *Kim, S. and Xing, E. Sparsity*. [*Tree-Guided Group Lasso for Multi-Task Regression with Structured*].Proceedings of the 27th International Conference on Machine Learning, 2010.

[6] *Jolliffe, I. (1986)*. [*Principal Component Analysis.*].Springer Verlag.

[7] *George H Dunteman. (1989)*. [*Principal components analysis. Number 69.*].Sage Publications, Inc.

[8] *Fox, J., Weisberg, S. (2011).* [*An R companion to applied regression.*]. Sage Publications, Inc.

[9] *Fox, J. (2008).* [*Applied regression analysis and generalized linear models (2nd ed.)*]. Sage Publications, Inc.

[10] *N. A. M. R. Senaviratna and T. M. J. A. Cooray. (2019)*. [*Diagnosing Multicollinearity of Logistic Regression Model*].Asian Journal of Probability and Statistics.

[11] *Habshah Midi , S.K. Sarkar Sohel Rana. (2010)*. [*Collinearity diagnostics of binary logistic regression model*].Journal of Interdisciplinary Mathematics.

[12] *Society of Photo-optical Instrumentation Engineers, Yidong Chen, Andreas N. Dorsel, Edward R. Dougherty(2001)*. [*Optical Technologies and Informatics*].SPIE.

[13] *Jianqing Fan, Runze Li, Cun-Hui Zhang, Hui Zou (2020)*. [*Statistical Foundations of Data Science*].Chapman and Hall/CRC.