PREDICCIÓN

PRÁCTICA 2 MEJORADA

JORGE CASAÑ VÁZQUEZ

Predicción

Profesor: Ricardo Queralt

EXECUTIVE SUMARY

El objetivo de este informe es desarrollar un modelo estadístico que ayude a los inversores a tomar la mejor decisión en base a análisis previos y calcular los tipos de interés asociados a cada tipología de préstamos. Para realizar nuestro estudio resulta imprescindible realizar una depuración en los datos, de tal forma que nos quedemos con las variables y observaciones que mejor expliquen el modelo de regresión. Hemos elegido como variable endógena loan_status, la cual deberemos convertirla en booleana.

En primer lugar, realizaremos la selección de modelos Best Subset y selección Stepwise, dentro de este último está el Forward Stepwise, Backward Stepwise y Mixto. En segundo lugar realizaremos una regresión GLM, en donde realizaremos una cross-validation del modelo, seguidamente de un análisis de la Curva ROC buscando la óptima probabilidad cut-off.. Posteriormente realizaremos regresiones regularizadas, explicando las regresiones Ridge, Lasso y Elastic Net, en donde se constatará que las variables estadísticamente más significativas son “revol_util” e “int_rate”. Finalmente, realizaremos un análisis de los modelos no lineales a través de regresiones polinomiales. En base a diferentes modelos de regresión, tomaremos decisiones estratégicas para la cartera del inversor.

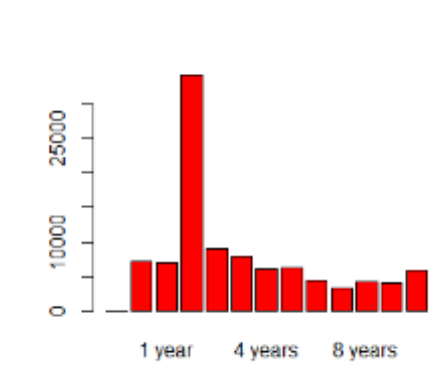
INTRODUCCIÓN

Vamos a trabajar con los datos recogidos de préstamos de la tabla LoanStats_2016Q3, del cual tenemos 99.122 observaciones y 111 variables. Lo primero que haremos será la depuración en los datos, y para ello descartaremos las variables de las cuales la mayoría son NA´s y las variables cualitativas que no aportan información relevante a nuestro estudio estadístico.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

GLM. MODELOS DE REGRESIÓN

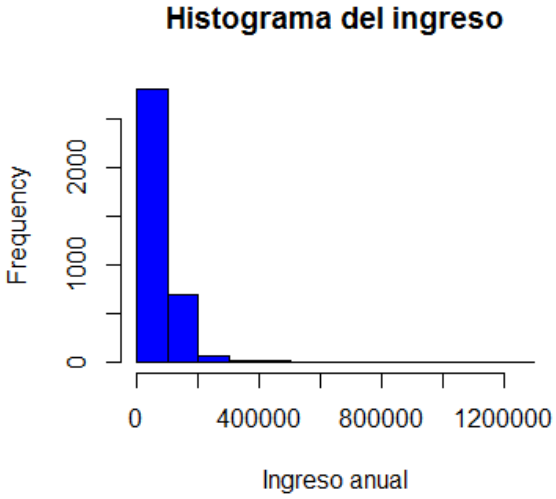
Nuestras variables elegidas son las siguientes: “loan_status", "grade", "sub_grade", "open_acc","pub_rec", "dti", "delinq_2yrs","inq_last_6mths", "emp_length", "annual_inc", "home_ownership", "purpose", "addr_state", "loan_amnt","int_rate", "installment", "issue_d", "revol_bal", "revol_util”.



Gráficamente a través de la columna “emp_lenght” observamos que gran parte de los préstamos concedidos son a corto plazo, de hasta 3 años fundamentalmente.

Nos interesa convertir los missing values de las variables “int_rate” y “revol_util” y “revol_bal” por la media, de los cuales, los dos primeros tendremos que convertirlos en formato porcentual, para poder hacer nuestros cálculos estadísticos.

De la variable “loan_status” seleccionaremos solamente las filas que sean fully Paid o Cargado. Haciendo booleana la variable, hacemos que nos devuelva el valor 1 si el préstamo está totalmente pagado y 0 en caso contrario.

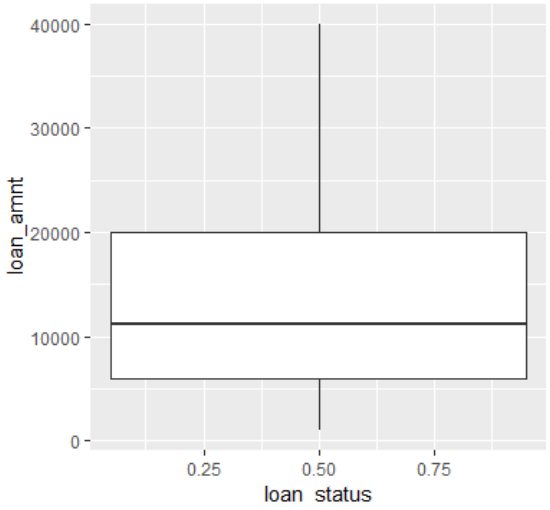


El tipo de interés representado a través de un histograma nos muestra que los préstamos de menor “anual income” son los que tienen una mayor frecuencia.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

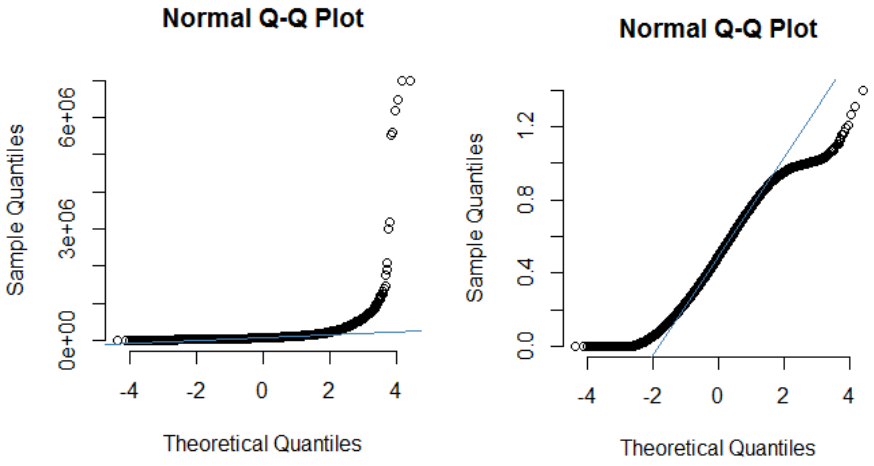
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0532  0.1099  0.1349  0.1453  0.1799  0.3099
```

A través de un boxplot podemos relacionar el importe del préstamo y su estado.

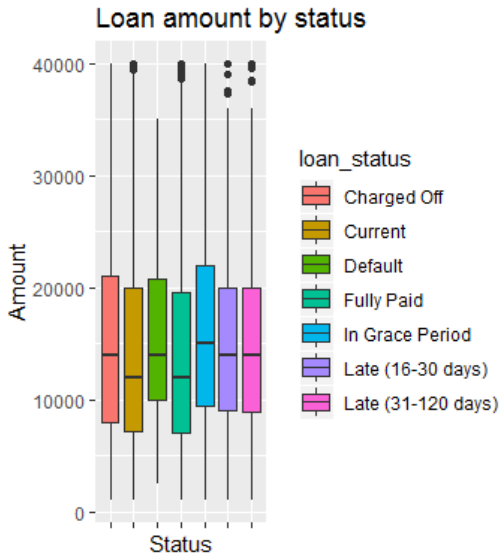


La mediana representa un valor ligeramente superior a los 10000, el importe de los préstamos, el valor queda centrado en 0,5, no habiendo outliers en los extremos

Los residuos se distribuyen de la siguiente manera:



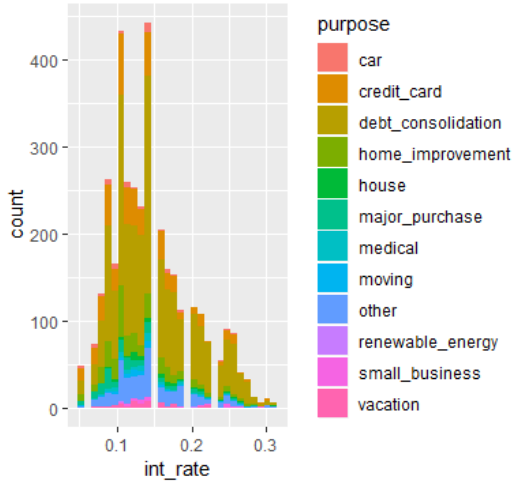
CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu



A través de un boxplot podemos ver la comparación entre el status con el amount con el objetivo de ver los outliers. La variable "Charge-off" no tiene outliers, en cambio, "Current", "Fully Paid" son los que más outliers presentan.

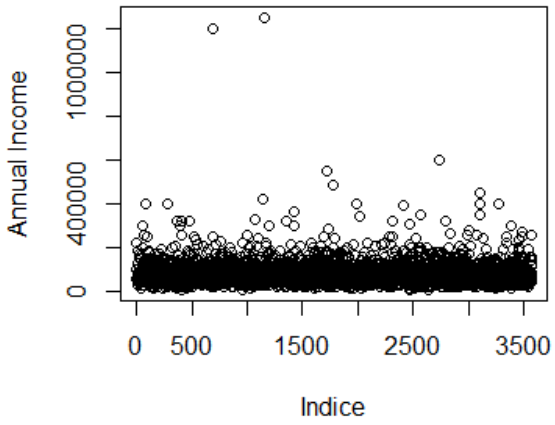
Sería interesante ver la relación entre el tipo de interés y el propósito por el cual se solicita el préstamo, por lo que seleccionaremos la variable "int_rate" y "purpose".

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu



Observamos que tienen un mayor tipo de interés las financiaciones de la tarjeta de crédito, la consolidación de la deuda y las mejoras del hogar. El resto de propósitos como el médico, para la compra de vivienda y de movilidad tienen asociados tipos de interés más bajos que los primeros.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10000   50000   70000   80717   97000 1250000
```



Seleccionando nuestra variable endógena, convertida en valor booleano, el estado del préstamo "loan_status", observamos que las variables "revol_bal", "pub_rec" y "dti" son las estadísticamente más significativas arrojando un p-valor próximo a 0, con un AIC de 198,63 y un BIC de 664,59.

```
revol_bal           -7.581e-05  4.580e-05  -1.655   0.0979 .
```

```
pub_rec             -1.913e+00  7.964e-01  -2.401   0.0163 *
dti                 -1.835e-01  1.104e-01  -1.661   0.0966 .
```

A través de la selección de modelos, en el modelo "Forward" las variables que son las más significativas estadísticamente son:

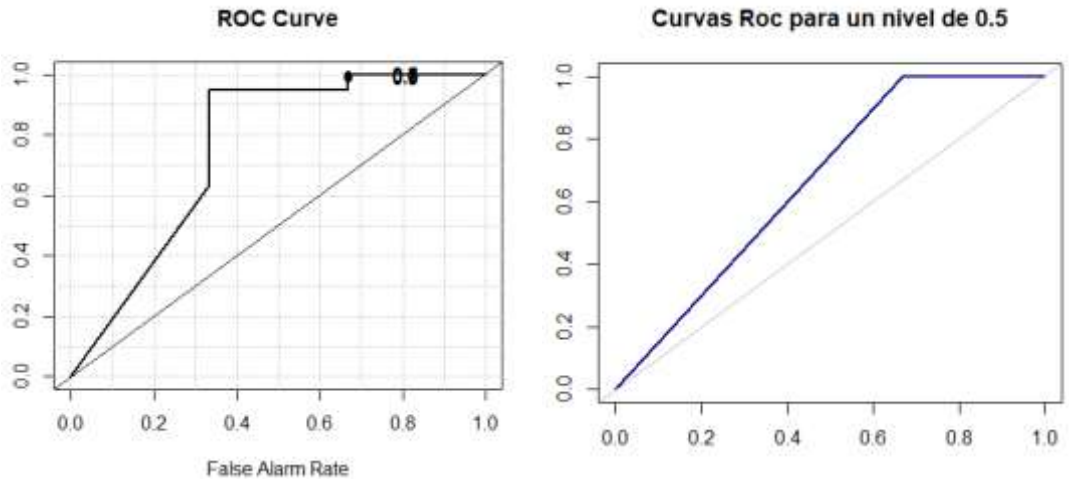
```
emp_length8 years   -1.917e-01  7.941e-02  -2.414 0.015759 *
emp_length9 years   -1.501e-01  7.914e-02  -1.896 0.057933 .
annual_inc          -1.779e-02  6.302e-03  -2.823 0.004759 **
```

```
purposesmall_business  5.096e-01  1.768e-01   2.882 0.003948 **
purposevacation        4.696e-01  2.083e-01   2.254 0.024196 *
```

A través del modelo "Both" obtenemos un AIC de 40758,44, el cual arroja el mismo resultado que con el análisis hecho a través del modelo "Backward".

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

El análisis Cutt-off es el valor a partir del cual se acepta el modelo. Si aumentamos el valor de corte, disminuirá el número de falsos positivos y por otra parte, el número de falsos negativos aumentará. Como se puede observar en el gráfico, utilizando el 0,52 el error es el más bajo de todos. Además con el análisis de la matriz de confusión nos indica que no tenemos un elevado número de falsos positivos ni negativos.

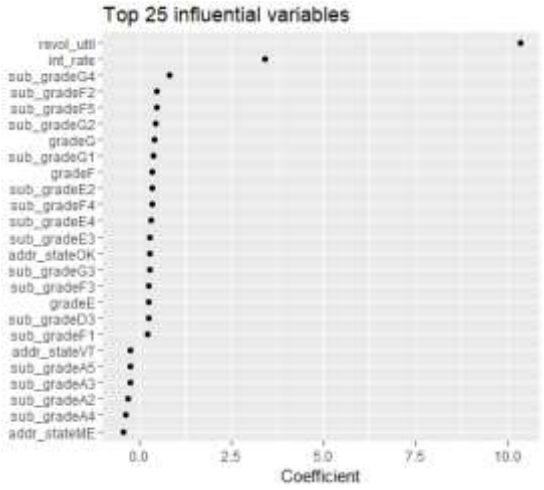


MODELOS DE REGULARIZACIÓN

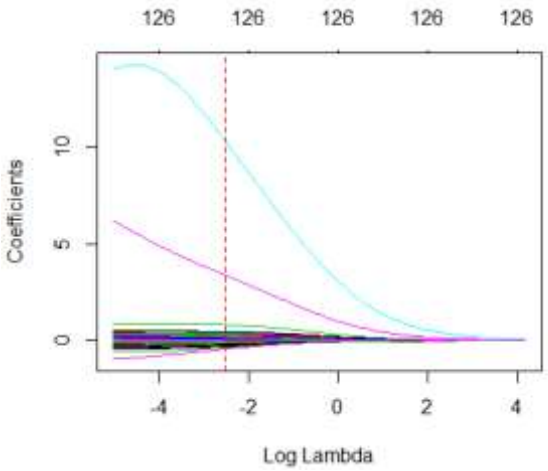
1. REGRESIÓN RIDGE

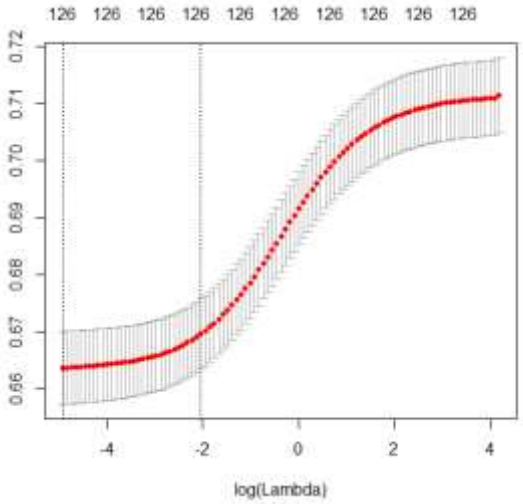
El modelo de regresión Ridge es similar al ajuste por mínimos cuadrados en cuanto a que ambos tratan de minimizar el RSS. La principal diferencia reside en que el Ridge incorpora un término llamado "shrinkage penalthy" que fuerza a que los coeficientes de los predictores tiendan a 0, si bien nunca llegarán a ser iguales a 0. Independiente de la significatividad de los predictores, ninguno se elimina. Este método consigue minimizar la influencia sobre el modelo de los predictores menos relacionados con la variable respuesta, pero en el modelo final va a seguir apareciendo. Aunque esto no supone un problema para la precisión del modelo, si para su interpretación.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu



De entre las 25 variables estadísticamente más significativas observamos que revol_util e int_rate son las que mejor explican el modelo de regularización Ridge





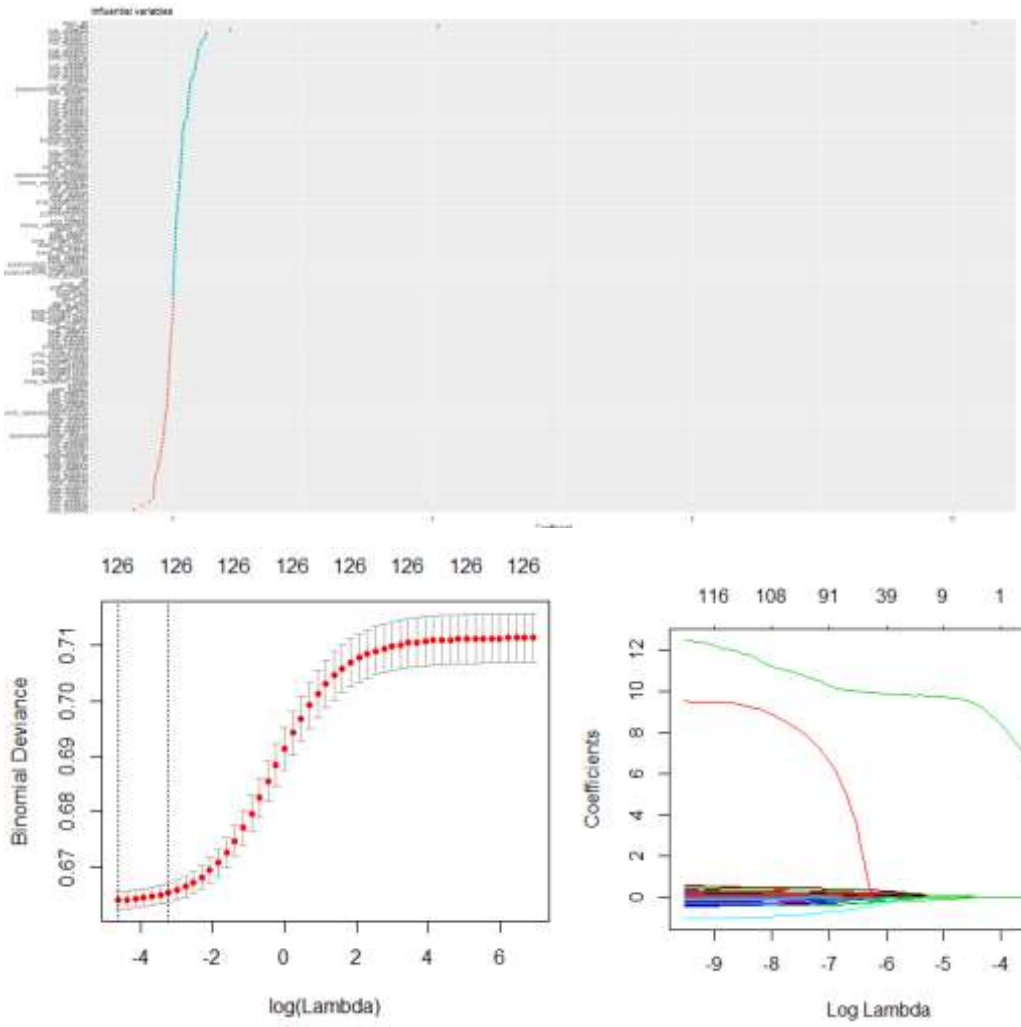
Podemos observar a través del último gráfico que a medida que aumenta el lambda, el error cuadrático medio aumenta. El punto de corte estará en -2 aproximadamente, sabiendo que cuando se cambia de pendiente es cuando tenemos que elegir el valor, por ser el estadísticamente más significativo.

2. REGRESIÓN LASSO

El modelo Lasso, al igual que en el Ridge, fuerza a que los coeficientes de los predictores tiendan a 0. La diferencia es que en Lasso sí que es posible fijar algunos de ellos iguales a 0, lo que permite además de reducir la varianza, realizar la selección de los predictores. Como resultado, el modelo Lasso tiende a generar modelos más fáciles de interpretar.

En el gráfico de abajo observamos que las variables estadísticamente más significativas coinciden con las variables que mejor explican el modelo Ridge, esto es, revol_util e int_rate.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu



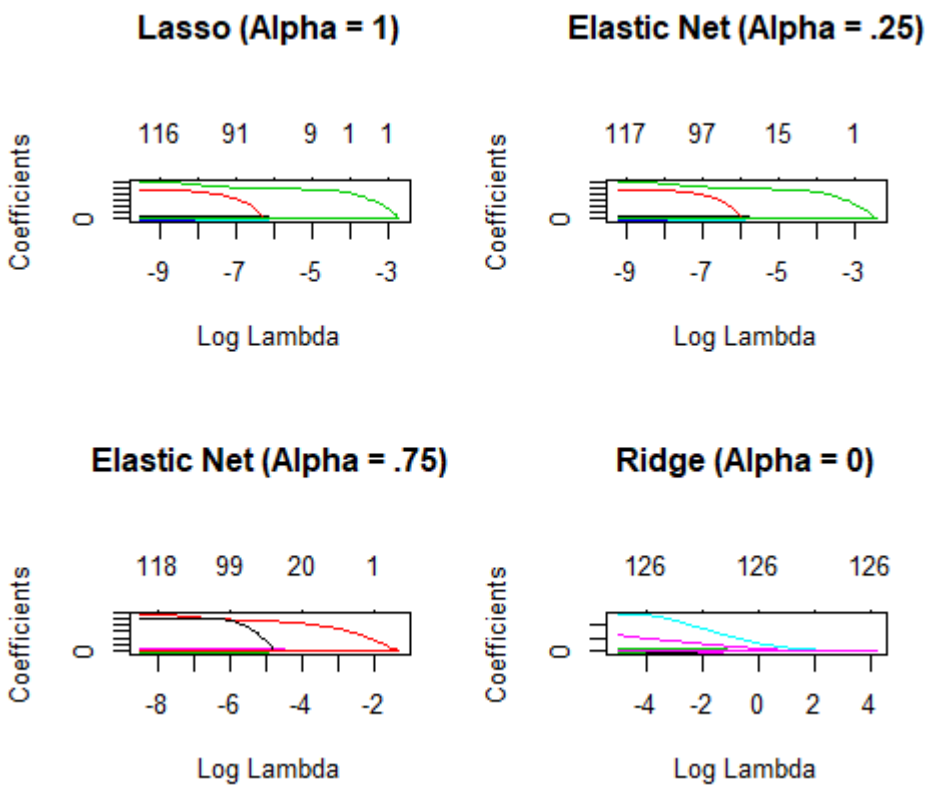
Al igual que en el modelo anterior se establece un punto mínimo de manera gráfica fuera de la vista por lo que para hallarlo volveremos a usar fórmulas. En este caso vuelve a ser 0.01 para un error medio del modelo de 0.664, el cual lleva un lambda de 0.039

3. REGRESIÓN ELASTIC NET

Este modelo es una combinación de los dos modelos anteriores, en donde incorpora la selección de variables del Lasso y la contracción de los predictores correlacionados como la regresión del Ridge.

Representamos 4 gráficos para tener una visión global de los modelos calculados anteriormente pero incorporando un alfa de 0.25 y 0.75.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu



MODELOS NO LINEALES

El objetivo de los modelos no lineales es obtener los valores de los parámetros asociados con la mejor curva de ajuste, calculada a través de los mínimos cuadrados. Comparando con los modelos lineales, los cuales tienen la ventaja de ser fácilmente interpretables tienen limitaciones importantes en su capacidad predictiva, debida fundamentalmente a la asunción de linealidad. Los modelos no lineales son los siguientes:

1. REGRESIÓN POLINOMIAL: la cual consiste en añadir curvatura al modelo introduciendo nuevos predictores que se obtienen al elevar todos o alguno de los predictores originales a distintas potencias.
2. LAS FUNCIONES DE PASO: Se divide el rango del predictor en K subintervalos de forma que, en cada uno, se emplean únicamente las observaciones que pertenecen a la región para ajustar al modelo.
3. LOS SPLINES DE REGRESIÓN: Se trata de una extensión a la regresión polinómica y de las step functions que consiguen una mayor flexibilidad. Consiste en dividir el rango del predictor X en K subintervalos. Para cada una de las nuevas regiones se ajusta una función polinómica, introduciendo una serie de

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

restricciones que hacen que los extremos de cada función se aproximen a los de las funciones de las regiones colindantes.

4. LOS SPLINES LOCALES: También se realizan ajustes por regiones, pero en este método las regiones se solapan las unas con las otras.
5. LOS SPLINES SUAVIZADOS: El concepto es similar a los splines locales pero consigue la aproximación de los extremos de las funciones colindantes de forma distinta.
6. LOS MODELOS DE ADITIVOS GENERALZIADOS (GAM): es el resultado de extender los métodos anteriores para emplear múltiples predictores.

CONCLUSIONES

A través de este informe hemos depurado los datos que eran los más relevantes y seleccionado las variables independientes que mejor explican el modelo de regresión, en donde la variable dependiente es loan_status, aspecto crucial para un análisis estadístico en un científico de datos. Por medio análisis como el cut-off, la matriz de confusión, la curva ROC, los modelos de regularización Ridge, Lasso y Elastic net hemos podido realizar pruebas en R-studio para que los inversores puedan obtener decisiones estratégicas a la hora de solicitar un préstamo.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto, 2
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu