PREDICCIÓN – MDS

HM2 – LENDING CLUB STATISTICS

1. INTRODUCCIÓN:

"Lending Club" es una compañía FinTech con origen y sede en San Francisco, CA, que trabaja para facilitar préstamos online a través de su plataforma peer-to-peer. Su web permite a los usuarios publicar referencias del proceso de préstamo online, lo cual puede ser visto por otros individuos y así elegir el producto que más encaje con su perfil y condiciones.

Como ejemplo llamativo, esta compañía vendió en 2015 préstamos online por más de $15.98 billones, lo que la situó como la plataforma de préstamos online más grande del mundo.

2. PREPARACIÓN DATASET:

He descargado el dataset a estudiar desde el link:
https://www.lendingclub.com/info/download-data.action (Enlaces a un sitio externo.)Enlaces a un sitio externo.

Dicho archivo contiene datos desde 2007 a 2011, con más de 99.000 observaciones y un total de 111 variables por lo que tendremos que comenzar limpiando y ordenando dicho dataset para realizar este estudio de predicción.

Las características u observaciones están claramente divididas en dos grupos:

a) Préstamo (Cantidad, tipo de interés, plazos)

b) Cliente o usuario del préstamo, la cual ocupa el mayor volumen de observaciones en el dataset y son las más importantes por la Fintech a la hora de aceptar o no la concesión de sus préstamos. Dichas características incluyen el empleo y su duración, el historial crediticio, patrimonio, etc.

Para empezar a preparar los datos, nos hemos quedado con las variables que consideramos más significativas a la hora de explicar la que vamos a tratar como variable dependiente "LOAN STATUS", ya que el objetivo es predecir si se van a pagar o no dichos préstamos en función de ello.

Hemos suprimido muchas variables por lo tanto, tales como "url", "loan id" o "average current balance" entre otras, ya que es imposible trabajar con 111 y estas son claramente menos significativas a la hora de explicar LOAN STATUS.

Por lo tanto, nos quedamos con las siguientes 18 variables independientes:

"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").

- Representacion del “interest rate”:

Para comenzar examinando los datos, empiezo por investigar la distribución de cada observación numérica mediante histogramas, segmentada y representada según el Loan Outcome (Resultado del Préstamo) y esto se puede observar en el código R adjunto a este informe.

Aquí podemos comprobar que el tipo de interés fue particular y claramente significativo a la hora de explicar la variable independiente. Examinando el gráfico, “fully paid loans” están claramente asociados con bajos tipos de interés, mientras que “charged off” tienen una distribución mucho más estable, tendiendo a altos tipos de interés. Esto tiene mucho sentido, ya que elevados tipos de interés son asignados a inversiones con mayor riesgo.

- Construcción del modelo de regresión seleccionando las columnas necesarias:

Observando el análisis categórico y tras nuestra regresión lineal, comenzamos a mirar la distribución de las variables que afectan a nuestra variable dependiente “Loan Status”, empezando por los grados más bajos. Como cabría esperar, un alto número de préstamos con grados bajos, fallaron a ser repagados lo cual indica con un 0.99 de Pr. que la variable Grade sea muy significativa a la hora de interpretar el “Loan Outcome”.

En todo caso, existen muchas observaciones con valores *NA* y por lo tanto, dividimos el modelo anterior en 2 submodelos para el estudio, uno de entrenamiento, que llevará el 70% (train.data) de los datos y otro de test que llevará solo el 30% de los datos para luego practicar la regresión con su correspondiente estudio de predicción.

Calculamos el BIC y el AIC del modelo y luego buscamos la óptima cut-off probability tanto dentro como fuera de la muestra de entrenamiento.

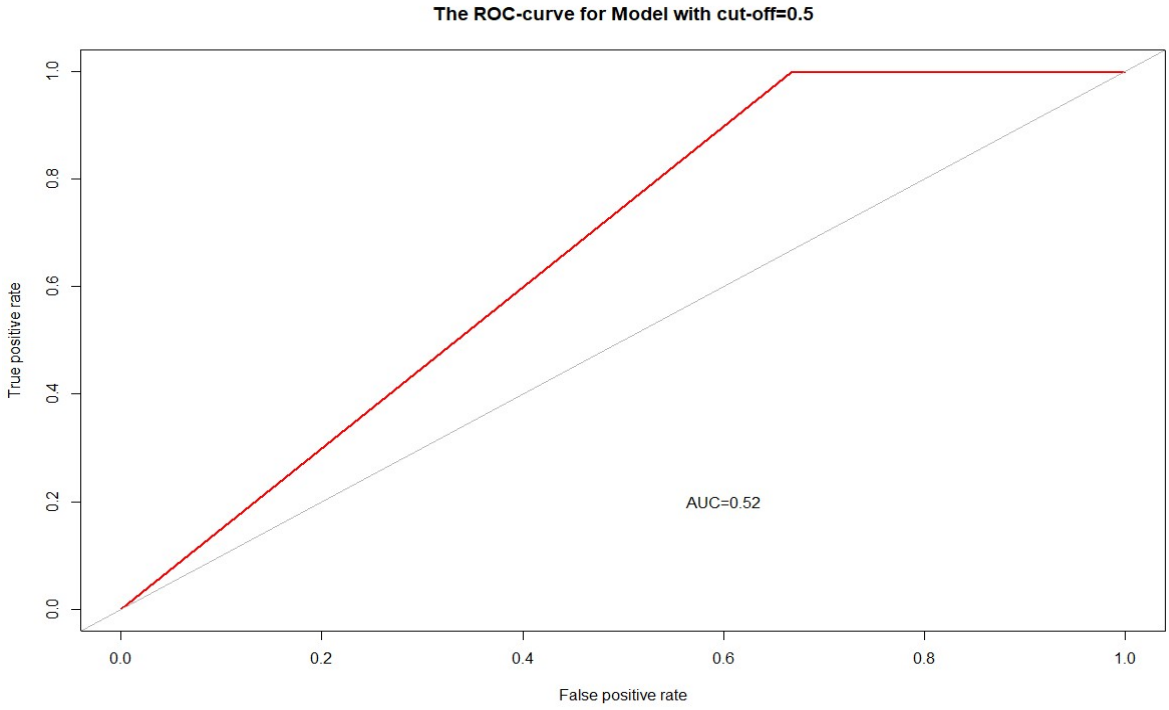
- Cut – Off = 0.02

En la mayoría de ocasiones se escoge un Cut Off de 0.5 pero para este estudio hemos elegido un Cut Off de 0,02, lo que significa que el p_cut es igual a 0,98.

Esto se hace con el objetivo de minimizar el coste total en el conjunto de entrenamiento. Estoy utilizando una función de coste asimétrica suponiendo que dar un mal préstamo cuesta 10 veces más que el coste de rechazar la solicitud de alguien que puede pagar.

Una vez hecha la matriz de regresión y estudiadas cuales son las variables más significativas, se plantean dos matrices de confesión para ver qué grupo da menos error. Tal y como se esperaba, el modelo de entrenamiento es el que menos probabilidad de error tiene.

El siguiente y último paso es plantear la curva ROC para validar que el modelo de test creado es válido y, por tanto, las variables escogidas para el estudio son las adecuadas.



3.CONCLUSIÓN:

Haciendo una valoración general del código, parece que los factores más importantes en la predicción de resultados en préstamos son las características fundamentales del mismo:

Grado, sub-grado, tipo de interés y plazos de repago.

Entre las características que definen mejor el perfil de cliente pagador, el ingreso anual parece ser la más destacada, mientras que otra que parecía relevante tal y como es el “dti” (Debt to income ratio), sorprendentemente tuvo un impacto muy bajo en términos relativos.

Este tipo de modelo de entrenamiento es muy importante. Mejorar la capacidad de ganar predicción en como varían los estados de un préstamo hace que la plataforma Fintech sea inteligente y puede complementar mejor el tradicional “credit scoring” que muchos prestamistas llevan a cabo hoy en día.