PREDICCIÓN

Práctica 2: Préstamos

V. Miguel Sempere Navarro

En la presente práctica se plantea analizar un dataset que almacena los datos de una empresa de préstamos online.

Lending Club es una Fintech con origen en San Francisco 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.

En primer lugar, se ha descargado el dataset a estudiar desde el siguiente link: https://www.lendingclub.com/info/download-data.action. Este archivo contiene datos desde 2007 a 2011, con más de 99.000 observaciones y un total de 11 variables. Por ello, el primer paso a realizar es una limpieza de variables y una ordenación del dataset. El objetivo es explicar la variable "LOAN STATUS", la cual se escoge como variable dependiente para tratar de predecir si se van a devolver o no los préstamos concedidos.

Todas las variables que, a priori, no aportan valor a la investigación, han sido suprimidas; escogiéndose como las óptimas para trabajar las 18 siguientes:

"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".

Sin embargo, para la construcción del modelo solo se tomarán algunas de ellas, tal y como se puede observar en el código de GitHub.

Será objetivo de la práctica averiguar si las variables escogidas para el modelo a realizar son válidas o no.

Para poder empezar con el análisis, el primer paso es dividir las observaciones en dos grupos, uno con el 70% de las observaciones como modelo de entrenamiento y otro con el 30% como modelo de test. Esto tiene como fin crear dos modelos que permitan predecir la variable dependiente. Seguidamente, y con las variables escogidas para la

construcción del modelo, se plantea una regresión del modelo de entrenamiento para estudiar la relación entre las variables y así poder realizar el modelo en cuestión.

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.

Tal y como se observa en el gráfico a continuación, el resultado de la curva es positivo y por tanto **se da por válido el modelo**.