Formulación de un modelo de scoring para solicitantes de créditos, no clientes de una entidad financiera

David Barrientos Guillén

Septiembre 2018

Este documento se trata de un resumen del trabajo de fin de master "Formulación de un modelo de scoring para solicitantes de créditos, no clientes de una entidad financiera". Por otro lado la realización de este documento se debe que no se autoriza la publicación del TFM debido a motivos de confidencialidad de la empresa en la que se realizó este trabajo.

Este trabajo aborda el problema de resolución de un caso concreto de un modelo de scoring. En el ámbito bancario, un modelo de scoring es aquel puntúa o estima la probabilidad de riesgo de mora. El problema a resolver es estimar la probabilidad de mora en el primer año de vida de los préstamos de tipo consumo para personas físicas no vinculadas a la entidad. Por otro lado se define un préstamo moroso como aquel que tiene un pago pendiente superior a tres meses. Además el segmento de clientes al que se aplica son aquellos no vinculados a la entidad, esto quiere decir que tienen una pequeña relación con el banco o directamente no la tienen.

La muestra que se utiliza para resolver el problema son aproximadamente 5000 datos. Además se dispone de datos sociodemográficos del mejor titular y datos referentes al préstamo. Estas variables son tanto continuas como categóricas.

El primer paso es el preprocesamiento de la muestra. Este consta de dos partes, la primera es identificar y eliminar de la muestra los outliers. La segunda es realizar un clustering para aquellas variables categóricas con un alto número de variables.

Después se pasa a determinar si existe una relación entre el indicador de mora en el primer año del préstamo y las variables existentes en la muestras. Por lo que se realizan contrastes estadísticos para saber si la distribución de los casos morosos y no morosos son diferentes para cada variable. Se realiza el test de Kolmogorov-Smirnov para las variables continuas y el contraste de χ^2 para las variables categóricas. De esta forma se selección aquellas variables que serán utilizadas en los modelos.

Tras escoger las variables que finalmente se utilizan en el modelo, se proceda a separar la muestra de datos en entrenamiento y test, repartiendo los datos en

80 % y 20 %, respectivamente.

El problema planteado trata de determinar la probabilidad de que ocurra un suceso, esto es lo que en estadística se conoce como clasificación supervisada. Por lo que el siguiente paso es aplicar diversas técnicas de clasificación supervisada sobre la muestra de entrenamieto. Estas técnicas o modelos estadísticos serían:

- Regresión logística.
- GAM.
- KNN.
- Support vector machine.
- Redes neuronales.
- Árboles de decisión.
- Random forest.
- Gradient boosting machine.

Por otro lado, la mayoría de estas técnicas tienen varios parámetros ajustables. Los parámetros óptimos serán obtenidos aplicando validación cruzada y obteniendo el valor del criterio AUC.

El siguiente paso es obtener el punto de corte para cada modelo de cara a realizar un dictamen de si se acepta o se rechaza el scoring. Por lo que se realizan las estimaciones de probabilidad de mora para la muestra de entrenamiento, sobre estas se obtienen las funciones de distribución de los casos morosos y no morosos, después se la distancia de Kolmogorov-Smirnov y el valor de la estimación de probabilidad de mora a la que se encuentra, siendo este el punto de corte.

Para finalizar el trabajo se realiza un test para ver que comportamiento tendrían los modelos con una muestra no utilizada en el entrenamiento. A partir de las estimaciones de probabilidad obtenidas se pueden calcular los estadísticos AUC, el error cuadrático medio y la distancia de Kolmogorov-Smirnov. Después se utilizan los puntos de corte para realizar dictámenes sobre la muestra del test, estos se pueden comparar con los resultados en una matriz de confusión de la que se van a sacar los resultados de imprecisión, sensibilidad, especificidad y la mora de los dictámenes aprobados.

También cabe indicar que los datos utilizados en este trabajo fueron extraídos mediante lenguaje SQL y después fueron analizados con el software estadístico R.