Detección de comportamientos financieros a partir de Machine Learning

Trabajo práctico Final – Cluster AI – Ciencia de Datos

Martín Moro, Pablo Wahren e Ivan Weigandi
Universidad Tecnológica Nacional, Buenos Aires, Argentina

Resumen
El objetivo del presente trabajo es aplicar modelos de machine learning a fines de predecir el comportamiento financiero de las personas. En particular, a partir de la aplicación de los modelos de aprendizaje supervisado Support Vector Machine (SVM), Logistic Regression, y KNN classification se busca predecir si una persona tomó crédito en un banco u otra institución financiera a partir de diversas características de las mismas.

Palabras claves
Inclusión Financiera, Crédito, Machine learning, Clasificación, Regularización.

1 INTRODUCCÍON

La inclusión financiera es clave para reducir la pobreza. Sin embargo, alrededor de 2.500 millones de personas en el mundo no utilizan servicios financieros formales y el 75% de los pobres no cuentan con cuenta bancaria. Uno de los principales aspectos que permiten determinar si una persona está incluida financieramente es si accede a crédito bancario o de otra institución financiera (Banco Mundial, 2019).

La base "Global Findex" del Banco Mundial abarca datos provenientes de encuestas de más de 140 países que permiten observar diversas características de las personas y sus comportamientos financieros (Demirgut et al., 2018).

El objetivo del presente trabajo es desarrollar un modelo predictivo que permita reconocer las características de las personas que tomaron un crédito en una institución financiera en los últimos 12 meses. Por un lado, esto contribuirá a conocer mejor los determinantes de la toma de préstamos en entidades financieras por parte de los individuos. Por el otro, le permite a las instituciones financieras conocer mejor el comportamiento de sus clientes o potenciales clientes.

2 ANÁLISIS EXPLORATORIO DE DATOS

2.1 Procesamiento de la base de datos

La base "Global Findex" cuenta con 154.923 observaciones (individuos) y 105 características. Entre estas últimas se encuentran características generales (género, edad, nacionalidad, entre otras) y de tipo financiero (ahorra, pidió prestado, entre otras).

A fines de depurar la base para poder realizar en análisis estadístico y posteriormente los modelos de clasificación se procedió a: quitar duplicados, eliminar aquellas características cuyas respuestas nulas superaban el 50%. Para los nulos restantes en las características que se mantuvieron se le imputan los valores con un método multivariado, utilizando la estrategia de métodos más frecuentes. Para las observaciones donde esto no fue posible se procedió a eliminarlas. Finalmente, resultó una base con 152.466 observaciones y 55 características.

Para transformar las respuestas en categorías binarias, se conservaron como "sí" aquellas respuestas afirmativas y se consideraron como "no" las respuestas negativas y, los "no sé" y los "rechazo".

Asimismo, se realizaron agrupamientos de los datos estadísticos por país y región a fines de poder realizar un análisis agregado. Se corroboró que cada país cuenta con al menos 1.000 observaciones, a excepción de Haití y Túnez.

2.2 Estadísticas descriptivas

Se obtuvieron diversas estadísticas descriptivas como la media, desvío standard, cuantiles, mínimos y máximos de cada variable.

Se realizó un análisis por país donde se vio la distribución de las respuestas positivas en preguntas claves como tiene cuenta, ahorra, pidió prestado y pidió prestado en una institución financiera. Visualizándose los siguientes resultados:

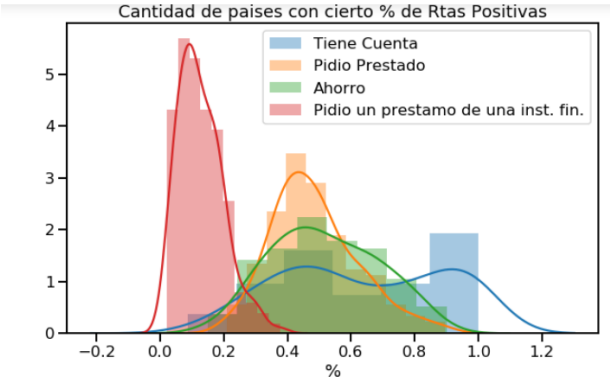

Figura 1. Primera visualización

Se puede apreciar como pedir prestado en una institución financiera tiene un grado de respuestas afirmativas inferior al resto de las variables, ya que da cuenta de una inclusión financiera más profunda.

El análisis por región nos permitió corroborar que la profundidad del sistema financiero es superior en los países de mayores ingresos. Por último, realizamos un mapa de calor para visualizar que porcentaje de la población pidió prestado en una institución financiera según país.

Figura 2. Visualización de préstamos a inst financieras

3 MATERIALES Y METODOS

3.1 Jupyter

Para cumplir nuestro objetivo, utilizamos diversas herramientas de entorno de programación, el principal es Jupyter, donde desarrollamos íntegramente el código para el análisis de la información.

Contamos dentro del mismo con las librerías Numpy (para el cálculo con matrices), Matplotlib (para la visualización), Scikit-Learn (para los algoritmos de regresión), Pandas (para la gestión de los datasets en dataframes) y GeoPandas (para georeferenciar los datos).

3.2 Clasificación

Para poder predecir la posibilidad de tomar créditos a partir de conocer el índice de deuda privada del país y al conocer los labels del data set, elegimos un aprendizaje supervisado a partir de la clasificación de nuestras variables con funciones de decisión.

Por medio de estas, las cuales toman un vector input X con "n" features y le asigna una de las K clases.

Antes de poder utilizar un modelo de clasificación, se debe de dividir el data set en train y test, para que el modelo pueda aprender la regla de decisión, clasificando las muestras de test luego de haber aprendido un modelo en train y comparándose con el valor real, obteniéndose una exactitud.

A fin de lograr esto, debimos de escalar nuestras variables mediante el autoScaler (herramienta de Scikit Learn), con una media de 0 y un desvío estándar de 1

Para obtener el mejor modelo existente y maximizar la precisión, utilizamos un método de Cross-Validation en el set de training, el cual consiste en dividir nuestro training set en K porciones e iterar sobre el mismo K veces, para luego obtener un promedio de precisión y evitar el Overfitting, que haría que nuestro modelo clasificador se encuentre entrenado perfectamente sólo para nuestros datos y no para nuevos valores.

Empleamos 3 modelos, los cuales consisten en hiper-parámetros que hemos seleccionado, para encontrar los mejores para nuestro modelo se ha generado una lista de estos y probamos las mejores combinaciones posibles de estos (Grid-Search).

Los modelos de clasificación utilizados son los siguientes:

3.2.1 Regresión Logística

Es una regresión lineal precedida de una función de activación sigmoide, lo que genera un output binario. A cada muestra clasificada le asigna una probabilidad de pertenecer a cada clase existente en el problema, si esta probabilidad es mayor a un cierto threshold, entonces pertenece a una clase y viceversa.

$$p(y_i|X) = \sigma(w^T, X) \quad (1)$$

$$\sigma(x) = \frac{1}{1+\exp(-x)} \quad (2)$$

3.2.2 Support Vector Machine

Este modelo consta de un clasificador lineal, en el cual se busca el hiperplano separador que maximiza el margen entre clases, siendo cada muestra mal clasificada penalizada por una función de costo C (seleccionada como hiper-parámetro). El margen separador queda definido por "s" muestras, llamadas support vectors.

Utilizamos un Kernel (función de similaridad entre muestras) gaussiano, para determinar una frontera no líneal de clasificación.

$$K_{gaussiano}(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \quad (3)$$

3.2.3 KNN Classifier

Clasifica cada nuevo dato en el grupo que corresponda, según tenga K vecinos mas cerca de un grupo u otro. Calcula la distancia del elemento nuevo a cada uno de los existentes y ordena esas distancias para seleccionar a que grupo pertenece. Se selecciona como hiper-parámetro los K-vecinos

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2} \quad (4)$$

3.2.4 Error

Para determinar cuan bien nuestros modelos clasifican muestras, hemos utilizado funciones de error definidas como:

Error cuadrático medio (MSE)

$$MSE = \frac{|\sum(\widehat{y_t} - y_t|}{n} \quad (5)$$

Raíz cuadrada del error cuadrático medio (RMSE)

$$MSE = \frac{\sum(\widehat{y_t} - y_t)^2}{n} \quad (6)$$

Media del error (MAE)

$$MSE = \sqrt{\frac{\sum(\widehat{y_t} - y_t)^2}{n}} \quad (7)$$

3.3 Regularización

Debido a que nuestro data set se compone de 105 columnas, hemos decidido eliminar los duplicados y las columnas con un % mayor de nulos del 15%.

No obstante, nuestro data set aún poseía 63 columnas, para lo cual, decidimos quedarnos con las variables de mayor poder explicativo para nuestro fin, para lograr esto regularizamos el problema mediante LASSO, eliminando de gran manera el ruido.

Este modelo asigna un parámetro beta a cada feature. El parámetro lambda de regularización penaliza al modelo y obliga a llevar a cero los betas (o pesos) de las features menos importantes utilizando la norma L1 sobre los pesos de las features. A medida que variamos el parámetro penalizador, algunos pesos se anulan y quedan en 0, quedando con un valor mayor a cero las variables más importantes.

$$min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta^2\right\} sujeto\ a\ \sum_{j=1}^{p}|\beta_j| \le t \quad (8)$$

3.4 Curva AUC-ROC

El área bajo la curva ROC(AUC) da una idea de cuan bueno es mi clasificador, independientemente de la precisión. Contemplándose la relación entre verdaderos positivos y falsos positivos.

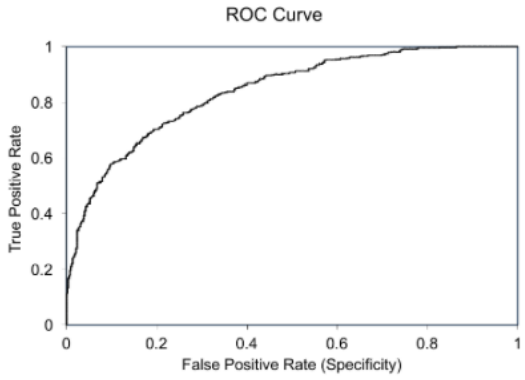

Figura 3. Ejemplo de curva ROC

3.5 Matriz de Confusión

Es un elemento para evaluar los resultados de clasificación. En cada posición se encuentran los verdaderos positivos (TP), los verdaderos negativos (TN), los falsos positivos (FP) y los falsos negativos (FN). Luego se obtienen los valores de precisión, sensitividad y especificidad.

$$Precisión = \frac{(TN+TP)}{Total} \quad (9)$$

$$Sensitividad = \frac{TP}{(TP+FN)} \quad (10)$$

$$Especificidad = \frac{TN}{TN+FP} \quad (11)$$

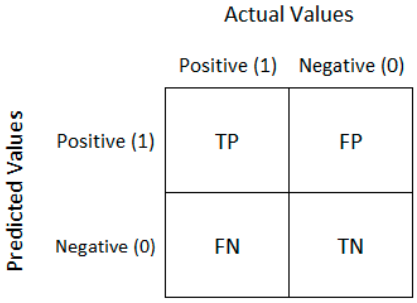

Figura 2. Ejemplo de Matriz de Confusión

4 RESULTADOS

Empleando las herramientas descritas, obtuvimos una clasificación de variables con una precisión promedio de 88% arrojada equitativamente en nuestros 3 modelos de clasificación.

Los errores en clasificación también fueron consistentes entre los tres modelos, observándose un RSME de 0.33, un MSE de 0.11 y un MAE de 0.11.

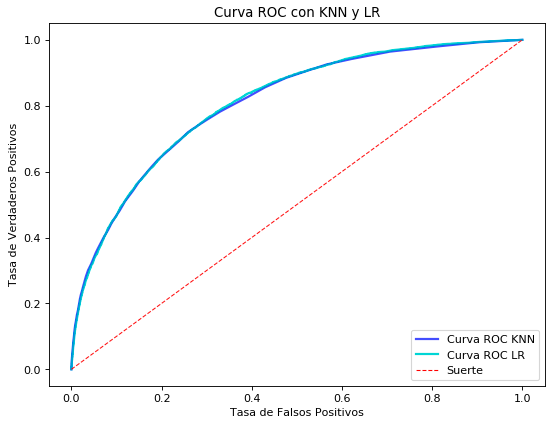

Figura 4. Curva ROC observada

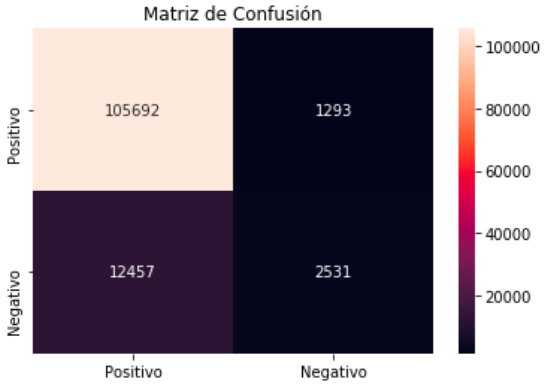

Figura 5. Matriz de confusión observada

5 CONCLUSIONES

A partir de los datos provistos por el Banco Mundial en la base “Global Findex” hemos logrado construir un modelo de clasificación robusto, que permite predecir con un cierto margen de error la probabilidad de que una persona haya pedido un préstamo en una institución financiera en los últimos doce meses.

6 REFERENCIAS

[1] Banco Mundial (2019). . Consultado el 16/11/2019. Disponible en https://www.bancomundial.org/es/topic/financialinclusion/overview

[2] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Science+ Business Media.

[3] Demirguc-Kunt, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2018). Global Findex Database 2017 [La base de datos Global Findex 2017]. *World Bank Publications*.

[4] Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85.