Predicción de que alguien presente dificultades financieras usando métodos de aprendizaje supervisado

Federico Ocampo Ortiz, Andres Felipe Orrego Quintero

Especialización Analítica y Ciencia de Datos.

Universidad de Antioquia.

ABSTRACT Los avances tecnológicos han hecho que muchas labores que anteriormente eran arduas y tediosas, hoy sean más eficientes, uno de esos casos lo encontramos en la evaluación de perfiles para otorgar créditos, gracias a la ciencia de datos hoy se tienen modelos que pueden predecir si una persona va a entrar en mora o no, en este trabajo se crea un modelo que predice si una persona va a entrar en mora por más de 90 días, para esto se hace una preparación del dataset y posteriormente se emplean varias técnicas de modelado y evaluación

INDEX TERMS aprendizaje automático, competición, dataset, evaluación modelo, give me some credit, kaggle, mora, puntaje crediticio,

I. INTRODUCTION

En la actualidad, el otorgamiento de créditos se ha convertido en una parte fundamental del sistema financiero, impulsando el crecimiento económico y facilitando el acceso a bienes y servicios para individuos y empresas. Sin embargo, uno de los principales desafíos que enfrentan las instituciones crediticias es determinar la probabilidad de que un prestatario cumpla con sus obligaciones de pago. La capacidad de predecir de manera precisa si una persona va a pagar o no un crédito resulta vital para minimizar los riesgos y mantener la estabilidad del sistema financiero.

En este contexto, el uso de técnicas de aprendizaje automático y modelos predictivos ha surgido como una herramienta prometedora para evaluar la solvencia crediticia de los solicitantes. Estos modelos aprovechan la disponibilidad de grandes cantidades de datos históricos, que incluyen información sobre el perfil financiero de los prestatarios, antecedentes crediticios, características demográficas y otros factores relevantes. Al utilizar algoritmos sofisticados, estos modelos pueden identificar patrones ocultos y generar predicciones precisas sobre la capacidad de pago de un individuo.

El objetivo de este estudio es desarrollar un modelo predictivo que permita determinar si una persona va a pagar o no un crédito, utilizando un enfoque basado en el aprendizaje automático. Para lograrlo, se utilizará un conjunto de datos históricos de préstamos y se aplicarán diferentes técnicas de preprocesamiento y selección de características. A continuación, se entrenarán varios algoritmos de aprendizaje automático, como árboles de decisión, regresión logística y redes neuronales, con el fin de evaluar su rendimiento en la predicción de la capacidad de pago de los prestatarios.

El resultado de este estudio tiene el potencial de proporcionar a las instituciones financieras una herramienta efectiva para evaluar el riesgo crediticio de manera más precisa, lo que puede tener un impacto significativo en la toma de decisiones relacionadas con la aprobación o rechazo de solicitudes de crédito. Además, el modelo propuesto puede contribuir al desarrollo de estrategias de gestión de riesgos más efectivas, ayudando a mitigar las pérdidas asociadas con préstamos incobrables y mejorar la rentabilidad de las instituciones crediticias.

II. ESTADO DEL ARTE

En el campo de la predicción de la capacidad de pago de créditos, se ha observado un progreso significativo en los últimos años gracias al uso de técnicas avanzadas de aprendizaje automático. Se han destacado diversos avances relevantes en esta área.

En cuanto a las técnicas de aprendizaje automático, los algoritmos supervisados como los árboles de decisión, las redes neuronales, la regresión logística y los bosques aleatorios han ganado popularidad como opciones para

modelar la solvencia crediticia. Estos algoritmos se caracterizan por su capacidad para abordar conjuntos de datos complejos y capturar relaciones no lineales, lo que los hace adecuados para abordar el problema de predicción de pago de créditos.

En un estudio realizado por el banco central de Grecia, demostró que el uso de Gradient boosting generaba mejores resultados a la hora de encontrar personas que pueden tener problemas financieros que los de regresión lineal y regresión logística [1].

En términos de evaluación de modelos, se ha utilizado la validación cruzada y métricas como la precisión, el área bajo la curva ROC (AUC-ROC) y la ganancia acumulativa de precisión (CAP) para evaluar el rendimiento y comparar diferentes enfoques.

Por último, se ha incrementado el interés en la interpretabilidad y aplicabilidad de los modelos. Comprender los factores clave que influyen en las predicciones resulta crucial para tomar decisiones transparentes y éticas en el ámbito crediticio.

La predicción de la capacidad de pago de créditos ha experimentado avances notables gracias al uso de algoritmos supervisados, técnicas de selección de características, métodos de ensamble y procesamiento del lenguaje natural. Aunque se han logrado mejoras en la precisión y eficacia de los modelos, todavía existen desafíos por enfrentar, como el manejo de desequilibrios en los datos y la interpretación de modelos complejos. Estos temas continúan siendo áreas de interés para futuras investigaciones en este campo.

Por otro lado, el algoritmo de referencia para esta competición consistía en un random forest ensamblado con muy buen rendimiento y una área bajo la curva de 0.864. En comparación con uno de los mejores modelos comerciales para el puntaje de crédito que cuenta con un área bajo la curva de 0.870, se puede evidenciar que la diferencia está por debajo de 0.006 [2].

El top 3 de ganadores de la competición de kaggle fueron el equipo australiano Perfect Storm, un equipo de 1 persona en singapur llamado GXAV y un equipo de 1 persona de Boston, MA llamado Occupy. El equipo Perfect Storm, obtuvo una medida de desempeño AUC de 0.86955, el segundo equipo fue GXAV con 0.86929 y por último el equipo Occupy obtuvo 0.86928.

TABLA I
RESULTADOS DE LOS GANADORES DE LA COMPETICIÓN

Ranking	Equipo	Puntaje
1	Perfect Storm	0.86955
2	GXAV	0.86929
3	Occupy	0.86928

Estos equipos obtuvieron una mejor área bajo la curva de 0.004 en comparación con el algoritmo de referencia. Hicieron uso de un enfoque mixto con Random Forest, Maquinas de soporte vectorial y GradientBoosting.

III. ENTRENAMIENTO Y EVALUACIÓN DEL MODELO

La base de datos utilizada se denomina Give Me Some Credit y fue obtenida del reto de Kaggle propuesto en el año 2011. Cuenta con un total de 150.000 muestras, 11 variables de entrada y una variable de salida SeriousDlqin2yrs. El área sobre la curva (AUC) fue utilizada como la métrica del rendimiento del modelo para la competición. Esta fue seleccionada ya que se ha encontrado que la dominancia de la tasa de verdaderos positivos y falsos positivos en la curva ROC es equivalente a la máxima utilidad y la dominancia de la frontera eficiente y óptima para los objetivos del negocio. [3].

La distribución de muestras por clase tiene una proporción de 14:1, contando con 139.974 muestras para la clase negativa y 10.026 muestras para la clase positiva. En (1) se puede observar la distribución de muestras entre clases.

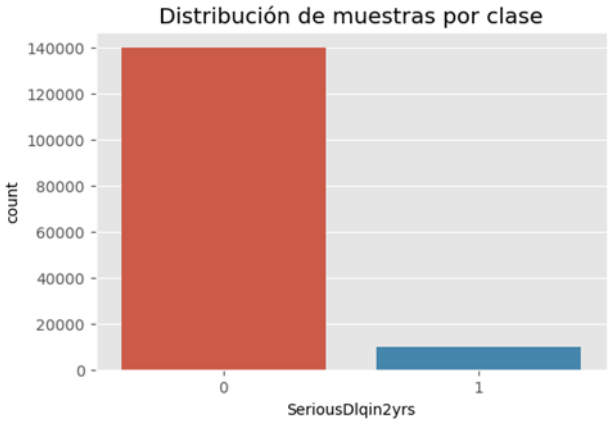

FIGURE 1. **Distribución de muestras por clase en la base de datos original.**

A causa de este desbalanceo en las clases, se realizó un submuestreo en la clase negativa para poder tener una distribución homogénea de las clases. Después de aplicado el submuestreo se redujo la cantidad de muestras de la base de datos a 20.052. En la figura 2 se muestra la nueva distribución de muestras después de la aplicación del método de balanceo

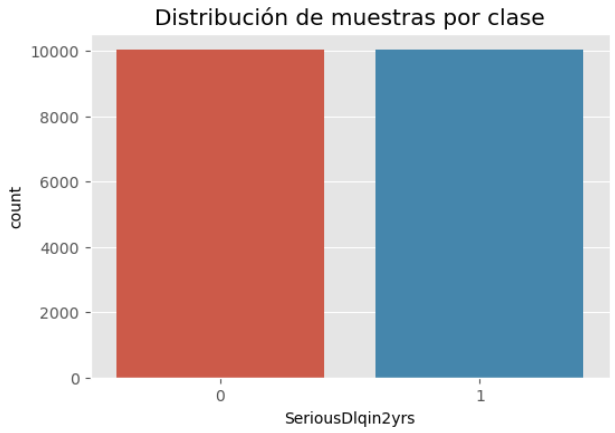

FIGURE 2. **Distribución de muestras por clase en la base de datos después del undersampling.**

La medida de desempeño a utilizar será la curva ROC-AUC, aprovechando que esta métrica no es sensible al desequilibrio entre las clases. Se espera una AUC encima del 80% para considerar el modelo como aceptable. Con este valor se busca aumentar la sensibilidad, para disminuir los falsos negativos, clientes que van a tener estrés financiero, pero son clasificados sin estrés financiero en dos años.

De manera alternativa también se hace uso de la métrica de exactitud (Accuracy) para medir la capacidad del modelo de hacer las predicciones correctas.

Con el fin de garantizar la independencia del modelo con la partición de los datos de prueba y evaluación, se aplicó una validación cruzada en el modelo usando 6 combinaciones.

IV. RESULTADOS

A continuación, se presentan los resultados obtenidos con diferentes técnicas y su correspondiente métrica de exactitud calculada. Se puede observar que las técnicas con mejores resultados fueron Gradient Boosting, XGBoost y Random Forest.

TABLA II
PRECISIÓN PARA LOS MODELOS IMPLEMENTADOS

Equipo	Puntaje
Gradient Boosting	0.76856298
XGBoost	0.76837828
Random Forest	0.76800887
Neural Network	0.74898412
SVC	0.71499815
Regresión lineal	0.70650166
Naive Bayes Complemento	0.70594754
KNN	0.68027337
Arboles de decisión	0.6784263

En comparación con el top 3 de los ganadores de la competición en Kaggle, se ve una diferencia de 0.02 puntos con el mejor modelo evaluado en este trabajo, que fue Random Forest.

TABLA III
COMPARACIÓN CON MODELOS GANADORES DE LA COMPETICIÓN

Equipo	Puntaje
Este trabajo	**0.84626**
#1 Perfect Storm	0.86955
#2 GXAV	0.86929
#3 Occupy	0.86928

Aunque las técnicas de optimización, limpieza y transformación de los datos en este trabajo distan mucho de lo realizado por los equipos expertos y ganadores de la competición de kaggle, Se encuentra una pequeña diferencia de 0.02 puntos, lo que indica que este trabajo puede seguir mejorando y evolucionar con el fin de obtener mejores resultados. Por ejemplo, la aplicación de métodos para la optimización de hiperparametros como GridSearchCV puede ser utilizada para definir la mejor configuración que pueda tener el modelo.

Otros resultados a destacar en la implementación y evaluación de los modelos son los siguientes:

TABLA IV
COMPARACIÓN CON MODELOS GANADORES DE LA COMPETICIÓN

Técnica	AUC / ROC
RandomForest	0.846257618963919
XGBoost	0.8216096634063742
GradientBoost Classifier	0.7874088672021275

V. REFERENCIAS

[1] Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2018). A robust machine learning approach for credit risk analysis of large loan-level datasets using deep learning and 40 extreme gradient boosting. BIS International Workshop on Big Data for Central Bank Policies (pp. 1-20). Bali: Bank of Greece.

[2] Sharma, Dhruv, Elements of Optimal Predictive Modeling Success in Data Science: An Analysis of Survey Data for the 'Give Me Some Credit' Competition Hosted on Kaggle (March 2, 2013). Available at SSRN: https://ssrn.com/abstract=2227333 or http://dx.doi.org/10.2139/ssrn.2227333

[3] P Beling, Z Covaliu & R M Oliver (2005) Optimal scoring cutoff policies and efficient frontiers, Journal of the Operational Research Society, 56:9, 1016-1029, DOI: 10.1057/palgrave.jors.2602021