

Clasificación de riesgo crediticio mediante técnicas de Machine Learning para una institución financiera

PROBLEMA

En Ecuador, usualmente, se evalúa el riesgo crediticio de forma operativa, es decir, mediante aplicaciones manuales y con base en puntajes crediticios. Esto puede llevar a una incorrecta clasificación de riesgo con concesiones de crédito a individuos con alta probabilidad de incumplimiento, incrementando las tasas de morosidad y pérdidas financieras, o puede negar crédito a individuos con buen potencial para cumplir sus obligaciones, limitando su acceso a oportunidades económicas o de negocio. El uso algoritmos de Machine Learning (ML) como alternativa para las instituciones al momento de evaluar riesgo de crédito, permite superar las limitaciones de los métodos manuales de concesión de crédito, mejora la precisión y eficiencia en la evaluación del riesgo y aprovecha los avances tecnológicos y de cómputo.

OBJETIVO GENERAL

Desarrollar modelos de clasificación del riesgo crediticio utilizando algoritmos de ML, para evaluar su eficacia en la predicción de incumplimientos crediticios en una institución financiera.

PROPUESTA

Se propone modelar el riesgo crediticio de una institución financiera utilizando los algoritmos de Regresión Logística Lineal y No Lineal, empleando el método de Newton como herramienta de optimización del modelo. Estos algoritmos son sencillos de implementar, interpretar y eficientes para entrenar. A partir de un análisis de correlación **Figura 1**, se determinan las variables más relevantes para implementar en el modelo No Lineal. Esto ayuda a determinar si las variables de balance de la cuenta y duración del crédito contribuyen a mejorar la precisión del modelo.

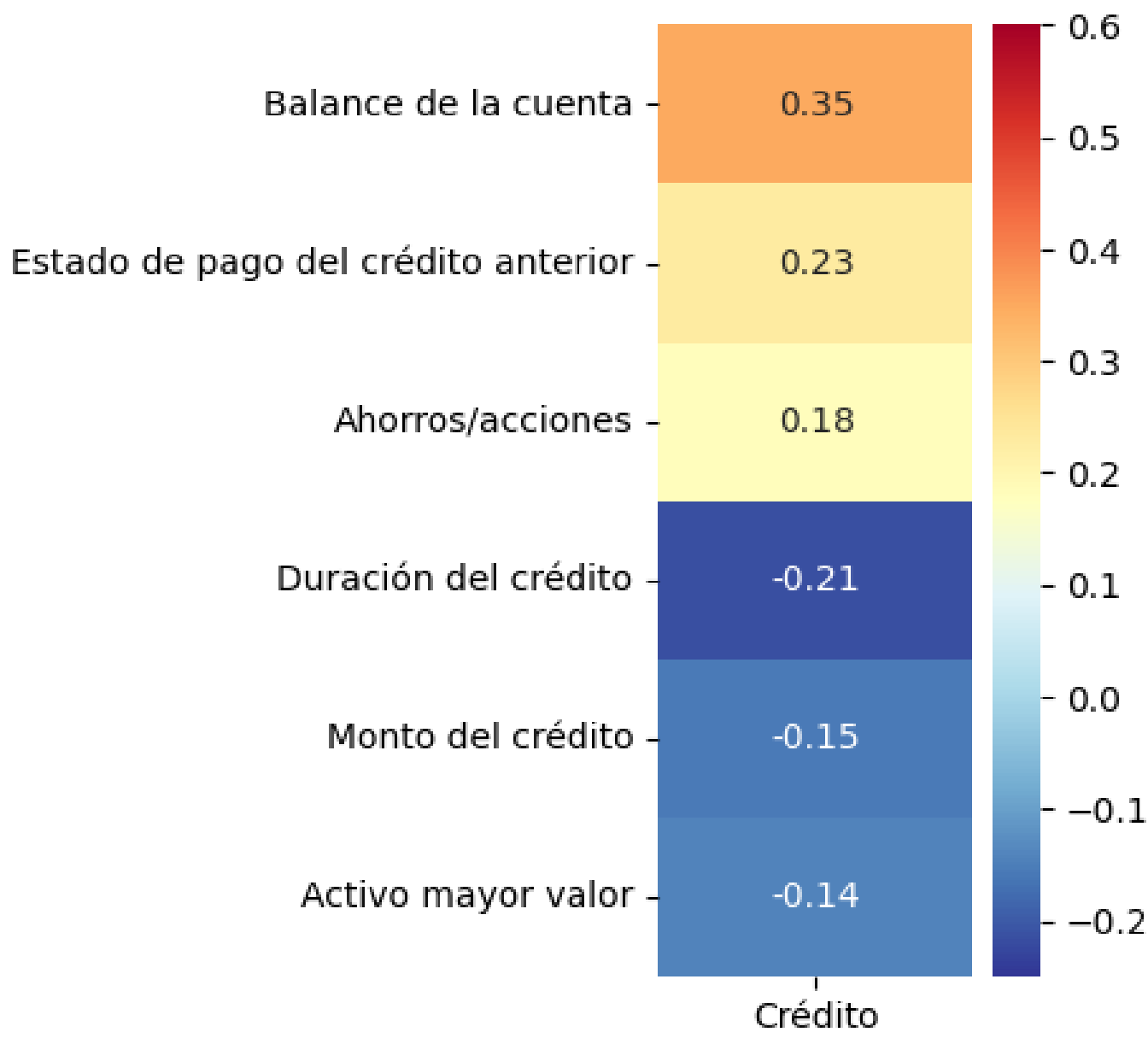


Figura 1. Correlación de Pearson entre la variable objetivo y las variables independientes más relevantes.

RESULTADOS

Los algoritmos se implementaron desde cero, es decir from scratch, y sus resultados se compararon con los obtenidos mediante scikit-learn. Para evaluar los modelos se utilizó el indicador de Precisión, que mide las predicciones correctas sobre el total de predicciones. Adicional, se emplearon las métricas de la Matriz de Confusión y la Curva ROC-AUC (acrónimo de Receiver Operating Characteristic-Area Under the Curve). De la Matriz de Confusión se obtuvieron los indicadores de Sensibilidad, que corresponden a la capacidad para dar con casos positivos empleando la tasa de verdaderos positivos y de Especificidad, que emplea la tasa de verdaderos negativos.

La Curva ROC-AUC indica como cambian estos dos indicadores cuando cambia el umbral de decisión del modelo. Mientras más cercano a 1 esté el valor de AUC, el test es mejor. La diferencia en la precisión entre el modelo de scikit-learn y el modelo Lineal es menor a 2%, esto se debe a que scikit-learn utiliza regularización y mejores técnicas de optimización. Mientras que el modelo No Lineal superó a los demás modelos en todas las métricas.

RESULTADOS	Precisión	Sensibilidad	Especificidad	AUC
Scikit-Learn	0.772	0.8579	0.5373	0.8024
Lineal	0.756	0.8688	0.4478	0.7815
No Lineal	0.991	0.9515	0.8942	0.9346

Tabla 1. Resultados de los modelos con sus respectivos indicadores.

CONCLUSIONES

- Los modelos permiten estimar de forma precisa el riesgo crediticio, de tal manera que la institución financiera fortalece su rentabilidad, reduciendo la probabilidad de *default* en sus préstamos.
- Los modelos contribuyen a que la toma de decisiones de crédito sea mucho más ágil y rápida.
- En general, la incorporación de no linealidades en el modelo de regresión logística mejora significativamente su capacidad para clasificar los casos de manera precisa.
- El modelo No Lineal es el más preciso, sin embargo, puede incurrir en *overfitting*.