ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

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

PROYECTO INTEGRADOR

Previo a la obtención del Título de:

Matemático

Presentado por:

Gorki Salomón Freire Zambrano

GUAYAQUIL - ECUADOR

Año: 2024

DEDICATORIA

A Mabe y a mi familia por su apoyo incondicional, y por supuesto, a Pan.

Gorki Freire Z.

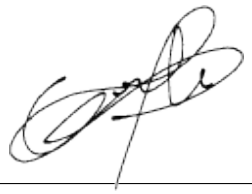
AGRADECIMIENTOS

Quisiera expresar mi más profundo agradecimiento a mis padres, quienes han sido la base sobre la cual he construido todo lo que soy hoy. A María Belén, por su infinita paciencia y constante apoyo a lo largo de la carrera. A mis tutores, cuyo acompañamiento y guía han sido esenciales para culminar con éxito este proyecto. A mis amigos de la carrera, quienes con su apoyo incondicional y colaboración en grupos de estudio hicieron que el proceso de aprendizaje fuera más llevadero. Gracias totales.

Gorki Freire Z.

DECLARACIÓN EXPRESA

"Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; *Gorki Salomón Freire Zambrano,* doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"



Gorki Freire Z.

EVALUADORES

Firmado electrónicamente por:
LUZ ELIMAR MARCHAN MENDOZA

Luz Elimar Marchán Mendoza

PROFESOR DE LA MATERIA

Firmado electrónicamente por:
CARLOS ANIBAL SUAREZ HERNANDEZ

Carlos Aníbal Suaréz Hernández

TUTOR DEL PROYECTO

RESUMEN

Este proyecto aborda la necesidad de mejorar la clasificación del riesgo crediticio en una institución financiera a través de la implementación de algoritmos de Machine Learning. Actualmente, la evaluación del riesgo crediticio en Ecuador se realiza de manera operativa, lo que puede conducir a clasificaciones incorrectas, incrementando la morosidad o limitando el acceso a créditos. El objetivo de este trabajo es desarrollar modelos de clasificación del riesgo crediticio utilizando algoritmos de Machine Learning para evaluar su eficacia en la predicción de incumplimientos crediticios en una institución financiera. Se emplearon los modelos de regresión logística, tanto lineal como no lineal, utilizando el método de Newton para optimizar los parámetros de los modelos. Se compararon los resultados obtenidos mediante la librería scikit-learn utilizando métricas como Precisión, Sensibilidad, Especificidad y la Curva ROC-AUC. Los resultados mostraron que el modelo No Lineal proporciona mayor precisión, dada su capacidad para ajustarse a los datos. Finalmente, se concluye que el uso de estos modelos puede mejorar la rentabilidad de la institución financiera al agilizar la toma de decisiones con respecto a créditos y reducir la probabilidad de incumplimiento en los préstamos.

Palabras Clave: Machine Learning, Riesgo Crediticio, Regresión Logística No Lineal, Curva ROC-AUC.

ABSTRACT

This project addresses the need to improve credit risk classification in a financial institution through the implementation of Machine Learning algorithms. Currently, credit risk assessment in Ecuador is performed in an operational manner, which can lead to incorrect classifications, increasing probability of default or limiting access to credit. The aim of this project is to develop credit risk classification models using Machine Learning algorithms to evaluate their effectiveness in predicting credit defaults in a financial institution. Both linear and nonlinear logistic regression models were employed, using Newton's method to optimize the model parameters. The results obtained and those generated using the scikit-learn library were compared using metrics such as Accuracy, Sensitivity, Specificity and the ROC-AUC Curve. The results showed that the nonlinear model provides higher accuracy, given its capacity to adapt to complex data. Finally, it is concluded that the use of these models can improve the profitability of the financial institution by streamlining credit decision making and reducing the probability of loan default.

Keywords: *Machine Learning, Credit Risk, Nonlinear Logistic Regression, ROC-AUC Curve.*

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

El propósito del presente proyecto es implementar modelos de clasificación de riesgo crediticio utilizando algoritmos de Machine Learning (ML) y comparar su desempeño entre ellos. La clasificación precisa de este riesgo es fundamental para la estabilidad y rentabilidad de las instituciones financieras. Uno de los objetivos más importantes de estas entidades es brindar crédito y que el prestatario no deje de pagar su deuda, es decir, que no caiga en *default*. Por tanto, el riesgo crediticio se mide a través de la probabilidad de *default*, ya sea de una empresa o individuo. La institución financiera se encarga de decidir, con base en el riesgo y otros condicionamientos, si brinda o no el crédito y las condiciones que exige para darlo.

La hipótesis que este trabajo propone, es que la implementación de modelos de ML mejora significativamente la precisión y eficiencia al medir este riesgo, garantizando la rentabilidad y estabilidad de la institución crediticia. La metodología incluye varias etapas como la selección de la base de datos, el preprocesamiento de los mismos, y su división en conjuntos de entrenamiento y prueba, la implementación de los algoritmos para la construcción de los modelos, y el análisis de los resultados.

1.1 Descripción del problema

El desafío que afronta el presente proyecto es la incorrecta clasificación de riesgo crediticio, la cual impacta a individuos e instituciones financieras en varios aspectos. Puede llevar a 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. Se requiere implementar varios modelos predictivos utilizando algoritmos de ML para comparar su precisión y efectividad. Los modelos cuentan con algunas restricciones como: proteger los datos personales de los clientes cumpliendo con regulaciones legales, asegurar que los modelos no discriminen afectando a grupos vulnerables y que sean adaptables a fluctuantes condiciones económicas. Los indicadores de interés son las variables históricas que impactan el riesgo crediticio, la precisión del modelo, su sensibilidad y especificidad. El proyecto busca superar las limitaciones de los métodos manuales de concesión de crédito mejorando la precisión y eficiencia en la evaluación del riesgo, aprovechando los avances tecnológicos y de cómputo utilizando algoritmos de ML.

1.2 Justificación del problema

La determinación del riesgo crediticio solamente de forma manual o utilizando empresas calificadoras de riesgo a través de puntaje o *score* crediticio, puede no ser tan efectivo para determinar nuevas concesiones de crédito dentro de la institución financiera. A nivel internacional, cada vez es más común que las instituciones financieras realicen el análisis de sus clientes *in-house*, es decir, dentro de la misma institución (Siddiqi, 2017). Debido a esto, se

propone el uso de algoritmos de ML como Regresión Logística (RL) Lineal y No Lineal para mejorar la capacidad de las instituciones en la determinación del riesgo de crédito del individuo. Esto expande los métodos mediante los cuales la institución financiera mide el riesgo, por tanto, su aplicación es crucial para reducir tasas de morosidad e incrementar la rentabilidad, contribuyendo a la sostenibilidad económica de las instituciones y promoviendo un acceso más justo al crédito. El estudio tiene impactos sociales, aumentando el acceso al crédito y estimulando el crecimiento económico, pudiendo reducir desigualdades sociales y contribuyendo a la estabilidad del sistema financiero. En resumen, esta investigación se realiza para mejorar la precisión y eficiencia en la evaluación del riesgo crediticio, aprovechar los avances tecnológicos y contribuir positivamente al desarrollo económico y social.

1.3 Objetivos

1.3.1 Objetivo General

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

1.3.2 Objetivos Específicos

- Analizar datos históricos de riesgo crediticio en una institución financiera, identificando las variables más relevantes para la construcción de los modelos predictivos, utilizando técnicas de análisis de datos y estadística.
- Implementar modelos de clasificación del riesgo crediticio, a través de algoritmos como Regresión Logística Lineal y No Lineal.

- Comparar la precisión y efectividad de los modelos desarrollados, mediante la aplicación de métricas de rendimiento.

1.4 Marco teórico

1.4.1 Riesgo crediticio

El riesgo crediticio o de crédito, se define como el riesgo de impago causado por cambios en la calidad crediticia de los emisores o contrapartes (Duffie and Singleton, 2012). En este contexto, el riesgo crediticio corresponde al riesgo de que el prestatario incumpla o tenga dificultades con su obligación hacia la institución financiera. Generalmente, se lo mide como la probabilidad de *default*, es decir, la probabilidad de que el prestatario no pague o tenga dificultades con su deuda.

1.4.2 Métodos para la evaluación del riego crediticio

Métodos tradicionales

Antiguamente, para evaluar si un individuo era apto para un crédito se tomaba en cuenta su carácter, habilidad de repago y el colateral, es decir, un activo en garantía (Thomas et al., 2017). Tradicionalmente, otro criterio que utilizan las instituciones financieras para decidir si conceder o no un crédito, es el puntaje o *score* crediticio. Este se define como el conjunto de modelos de decisión y sus técnicas subyacentes que ayudan a los prestamistas en la concesión de crédito al consumo (Thomas et al., 2017). Este puntaje es calculado generalmente por empresas calificadoras de crédito y abarcan variables como el tiempo con crédito, si paga sus deudas a tiempo, si tiene tarjeta de crédito, su cupo, etc.

Adicionalmente, las instituciones financieras suelen utilizar un puntaje de corte, es decir,

si el puntaje es menor al corte, no se concede crédito y si es mayor o igual al corte, es apto para crédito y se evalúa el monto, que también depende del tipo de producto financiero (crédito de consumo, hipoteca, carro, etc.) que se esté solicitando. A esto se suma una aplicación manual de forma escrita o digital donde se preguntan características de índole personal como edad, ingresos, empleo, estabilidad laboral, etc.

De acuerdo con Makhado (2023), las principales métricas para evaluar crédito son:

1. **Historial crediticio**: Se refiere al compartamiento financiero pasado. Este historial refleja disciplina y confianza financiera. Sus componentes son:
    - Historial de pagos: Contiene registros del individuo acerca de sus pagos a tiempo en préstamos pasados, tarjetas de crédito o cualquier otro tipo de crédito comercial. Si falló pagos, los hizo tarde, o no pagó; son condiciones que afectan negativamente este rubro.
    - Utilización de crédito: Se refiere al porcentaje de crédito utilizado sobre el cupo disponible. Si es elevado (más del 50%), esto sugiere que se tiene mucha confianza sobre el crédito y puede ser que no se lo pueda pagar, por tanto, afecta el puntaje de crédito de forma negativa.
    - Tiempo de historial crediticio: Se refiere al tiempo que un individuo ha tenido cuentas de crédito abiertas. Mientras más antiguas las cuentas se da mayor confianza, ya que indican experiencia manejando crédito.
    - Tipos de crédito: Existen diferentes tipos ya sean tarjetas de crédito, casas comerciales, hipotecas, autos, esto indica que el individuo puede manejar distintas formas de crédito.
    - Revisiones recientes: Cada vez que un individuo aplica para una nueva línea de

crédito se consulta a su reporte de crédito. Muchas consultas pueden afectar negativamente este rubro ya que pueden sugerir desesperación o estrés financiero por conseguir crédito.

2. **Ratio de Deuda-Ingresos**: Corresponde a una métrica de las obligaciones financieras del individuo con respecto a sus ingresos. Un valor alto del ratio puede indicar que el individuo tiene problemas de cumplir con sus obligaciones en el futuro, ya sea por gastos inesperados o reducción de ingresos. Los individuos con bajo nivel del ratio Deuda-Ingresos pueden tener mejores tasas de interés y condiciones de deuda, ya que se perciben como individuos de menor riesgo para el prestamista.

Algoritmos de Machine Learning (ML)

A continuación, se exponen los diferentes algoritmos de ML a implementar para la evaluación de riesgo crediticio.

Regresión Logística Lineal

El modelo de regresión logística es utilizado para la predicción de resultados binarios (0 o 1) y se estructura con la siguiente función sigmoide (Hosmer Jr et al., 2013):

$$\pi(X) = \frac{1}{1 + e^{-g(X)}} \tag{1.1}$$

Esta función toma valores entre 0 y 1. Esto sirve para aproximar el modelo al problema de ML supervisado, ya que generalmente, los datos de la variable objetivo de comportamiento de crédito vienen dados de forma binaria. Por otro lado, se define la transformación logit como:

$$g(X) = log\left(\frac{\pi(X)}{1 - \pi(X)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \tag{1.2}$$

donde $X = (x_1, \ldots, x_p)$ es el vector de atributos o características de los individuos, en este caso, son los atributos del aplicante a crédito o diferentes condicionamientos que tenga el cliente; y $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ son los $p + 1$ parámetros a estimar, empleando un método de optimización. Una forma de estimar la regresión logística es utilizando el método de máxima verosimilitud, que se realiza con la ayuda de la función de verosimilitud (Hosmer Jr et al., 2013):

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i}, \tag{1.3}$$

donde $i = 1, \ldots, n$ es el número de registros o individuos. Lo que se está estimando es la probabilidad de que el individuo i pueda pagar el préstamo, de acuerdo con las características y sus parámetros, por tanto, se estima usando la función sigmoide. Dada la forma de la función de verosimilitud se facilita la estimación de los parámetros al aplicarle logaritmo:

$$l(\beta) = log\left(L(\beta)\right) = \sum_{i=0}^{n} y_i log(\pi(x_i)) + (1 - y_i) log(1 - \pi(x_i)). \tag{1.4}$$

Para obtener los β_i que maximizan a (1.4) se deriva e iguala a 0 con cada uno de los parámetros y se obtienen las $p + 1$ ecuaciones de verosimilitud.

$$\sum_{i=0}^{n} (y_i - \pi(X_i)) = 0 \tag{1.5}$$

$$\sum_{i=0}^{n} X_{ij}(y_i - \pi(X_i)) = 0 \quad j = 1, \ldots, p \tag{1.6}$$

Por tanto, el problema de optimización radica en encontrar los valores de los parámetros β_i que cumplan con las condiciones anteriores. Una de las formas para optimizar estos parámetros de la regresión logística es utilizando el método de Newton, que es un algoritmo que viene dado de la siguiente forma (Luenberger and Ye, 2021):

$$\beta_{k+1} = \beta_k - H^{-1} \boldsymbol{\nabla} f(\beta_k). \tag{1.7}$$

Esto nos permite encontrar β^* que minimiza (maximiza al colocar un menos por delante) la función (1.4), donde H es la matriz Hessiana, cuya inversa multiplica al gradiente de la función. Por tanto, las variables en forma matricial son de la siguiente forma:

$$\nabla f(\beta_k) = X^T (Y - \pi_\beta(X)) \tag{1.8}$$

$$H = X^T \pi_\beta(X)(\pi_\beta(X) - 1)X \tag{1.9}$$

$$\pi_\beta(X) = \frac{1}{1 + e^{-X\beta}}. \tag{1.10}$$

En este caso, X es una matriz nxp, donde n corresponde al número de filas o registros, y p son las columnas o atributos. El problema a minimizar empleando el método de Newton es de la siguiente manera:

$$\min_{\beta} \quad -Y^T log(\pi(X) - (1 - Y)^T log(1 - \pi(X)). \tag{1.11}$$

Este modelo se adapta en gran forma para medir el riesgo crediticio y forma parte del estado del arte para clasificarlo (Lessmann et al., 2015). Ya que la variable dependiente y_i es una variable binaria, que en este problema representa si un individuo tuvo problemas con su deuda, o si pagó su deuda oportunamente, por ejemplo.

Entre las ventajas de utilizar este modelo están:

- Facilidad de implementación, interpretación y eficiencia para entrenar.
- No realiza ninguna suposición acerca de la distribución de clases.
- Es muy fácil que se extienda a múltiples clases (regresión multinomial).
- Se obtienen buenos resultados cuando la base de datos es linealmente separable.

Algunas de las desventajas de utilizar este algoritmo son:

- Limitante por el supuesto de linealidad de la variable dependiente y las independientes.
- Solo puede ser utilizada para predecir valores discretos.
- El cálculo de la matriz Hessiana puede acarrear un alto costo computacional, si se emplea el método de Newton.

Regresión Logística No Lineal

La RL No Lineal se estima de la misma forma que la RL Lineal, sin embargo, se utiliza una función no lineal en la transformación logit, por ejemplo utilizando la función $-log(x)$.

$$g(X) = log\left(\frac{\pi(X)}{1-\pi(X)}\right) = -log(\beta_0) - log(\beta_1)x_1 - \ldots - log(\beta_p)x_p. \tag{1.12}$$

La ventaja principal de utilizar este método en comparación con la RL Lineal, es la posibilidad de utilizar e interpretar datos no lineales. También, se la puede considerar empleando sus atributos o características de forma no lineal, por ejemplo, generando variables polinómicas.

1.4.3 Comparación de los modelos

Para poder realizar comparación entre la RL lineal y no lineal, se describen las siguientes métricas.

Curva ROC-AUC (acrónimo de Receiver Operating Characteristic-Area Under the Curve)

Es un gráfico que ilustra el rendimiento de un modelo de clasificador binario (también se puede utilizar para clasificación de clases múltiples) en valores con respecto a distintos umbrales de corte para las predicciones.

La curva ROC es la gráfica de la tasa de verdaderos positivos (TPR, por sus siglas en

inglés) frente a la tasa de falsos positivos (FPR, por sus siglas en inglés) en cada configuración de umbral.

Matriz de confusión

Es una matriz de números que indica dónde se confunde un modelo, corresponde a una distribución por clase del rendimiento predictivo de un modelo de clasificación, es decir, la matriz de confusión es una forma organizada de mapear las predicciones a las clases originales a las que pertenecen los datos. A partir de esta matriz, se pueden obtener los indicadores de Sensibilidad y Especificidad. Sensibilidad corresponde a la tasa de los verdaderos positivos, mientras que Especificidad es la tasa de verdaderos negativos.

CAPÍTULO 2

2. METODOLOGÍA

2.1 Preprocesamiento de datos

Para este proyecto se utilizó la base de datos *Home Credit Risk* del sitio web de *Kaggle*, que posee 307511 filas y 122 columnas. Esta base de datos se obtuvo de una institución financiera que concede créditos hipotecarios. Se utilizó código de Python para preprocesar los datos, utilizando principalmente las librerías de numpy y pandas. Se empezó por revisar el dataset de entrenamiento que contiene a la variable objetivo cuyo valor es 1, si el cliente tuvo dificultades con el pago de su deuda, y 0 caso contrario. Se emplearon datos históricos de aplicación a préstamos para predecir si nuevos individuos pueden o no pagar el crédito. Esto convirtió al proyecto en un problema de clasificación supervisado.

A continuación, se describen las variables más relevantes de la base de datos, basandóse en el trabajo de Vizhñay y Samaniego (2019) , quienes calcularon variables determinantes para obtener crédito en Ecuador.

- **ID**: ID único del préstamo.
- **CREDIT**: Variable binaria dependiente del modelo, es 1 si tuvo dificultades para pagar su préstamo (se atrasó en sus pagos o no pagó), caso contrario es 0.

- **GENDER**: Género del cliente.
- **CASA**: Si el cliente es dueño de un bien inmueble.
- **CARRO**: Si el cliente es dueño de un carro.
- **INGRESO**: Ingreso del cliente.
- **VALOR**: Valor nominal del préstamo.
- **EDUCACION**: Máximo nivel de educación del cliente.
- **DAYSAGE** : Edad del cliente en días desde la aplicación al crédito.
- **DAYSWORK** : Días que el cliente ha estado empleado desde la aplicación al crédito.
- **PROFESION**: Clase de trabajo del cliente.
- **SCORE**: Variable normalizada de puntaje crediticio de una fuente externa.

La variable de interés es la variable de **CREDIT**. Esta es la variable dependiente del modelo.

2.2 Análisis Exploratorio de Datos

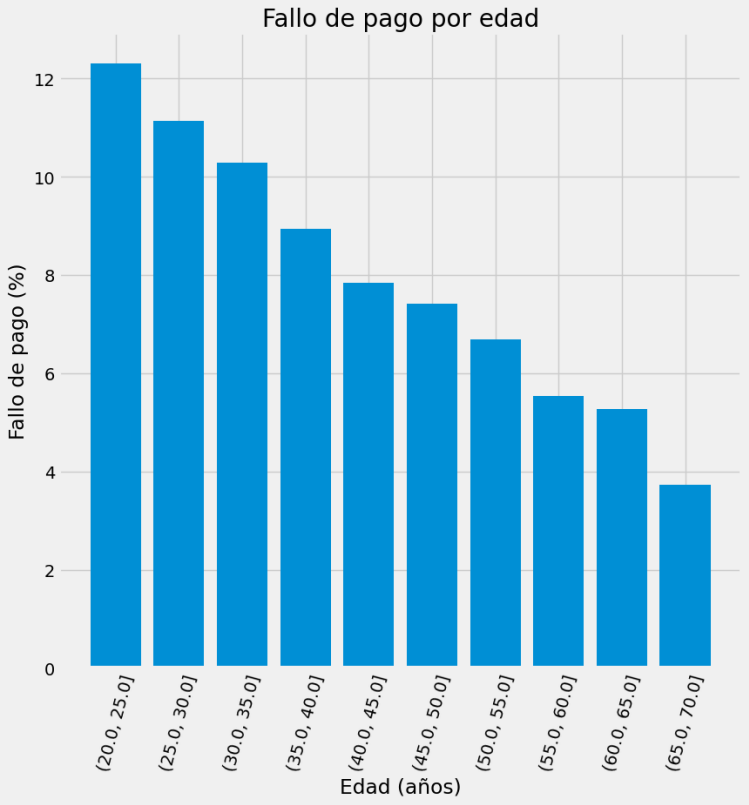
Luego de la carga de datos, se realizó un Análisis Exploratorio de Datos, mediante el cual se utilizaron gráficos para encontrar tendencias, anomalías, patrones o relaciones dentro de los datos. El objetivo de este análisis fue encontrar información en los datos. Lo que se encuentra ayuda a construir el modelo, por ejemplo, decidir sobre las variables a utilizar. Al examinar la variable dependiente u objetivo, se encontró que 24825 registros contienen (CREDIT=1), lo que quiere decir que hay muchos más préstamos que fueron pagados a tiempo, en contraste con los

que no. Al examinar por valores nulos se encontró que la base de datos posee 122 columnas, de las cuales 67 tienen valores nulos. Una opción es eliminar las columnas o variables con un porcentaje alto de valores nulos, sin embargo, estos valores pueden ser útiles para el modelo, así que se conservaron. Luego, dentro de las variables categóricas se realizó el proceso de etiquetado y para las variables que tenían más de dos valores se utilizó la librería de pandas con la función *getdummies*, la cual permite crear varias variables dummies con respecto a las frecuencias que posea la variable categórica. La base quedó con 307511 filas y 240 columnas. Al revisar por anomalías, se creó una nueva columna binaria que indica si el valor es anómalo y sus valores pasaron a ser nulos (np.nan en Python).

Luego de lidiar con las anomalías y las variables categóricas, se realizó un análisis de correlación entre la variable de interés, y las demás. Se encontró que la variable **DAYSAGE** es la que tiene la mayor correlación con la variable **CREDIT**. En la base de datos esta variable está con valor negativo porque resta días desde el día de la aplicación, lo que significa que a mayor edad tenga el cliente, es menos la posibilidad de que no pague su deuda. Se muestra en la Figura 2.1 la relación que posee con la edad del cliente.

Las variables que más se correlacionaron con la variable dependendiente fueron las de **SCORE** con correlaciones negativas, lo que indicó que si el puntaje crediticio se incrementa, es más probable que se pague la deuda. También se observó que la variable de edad **DAYSAGE** se encuentra de forma positiva correlacionada con **SCORE1**, indicando que uno de los factores en este puntaje es la edad. En ML es común emplear una técnica llamada *Feature Engineering* que consiste en crear nuevas variables a partir de la base de datos o a partir de conocimiento técnico del tema a modelar. En este caso, se utilizaron los datos para crear nuevas variables basados en

Figura 2.1.
Comparación de falla de pagar el préstamo y la edad del cliente

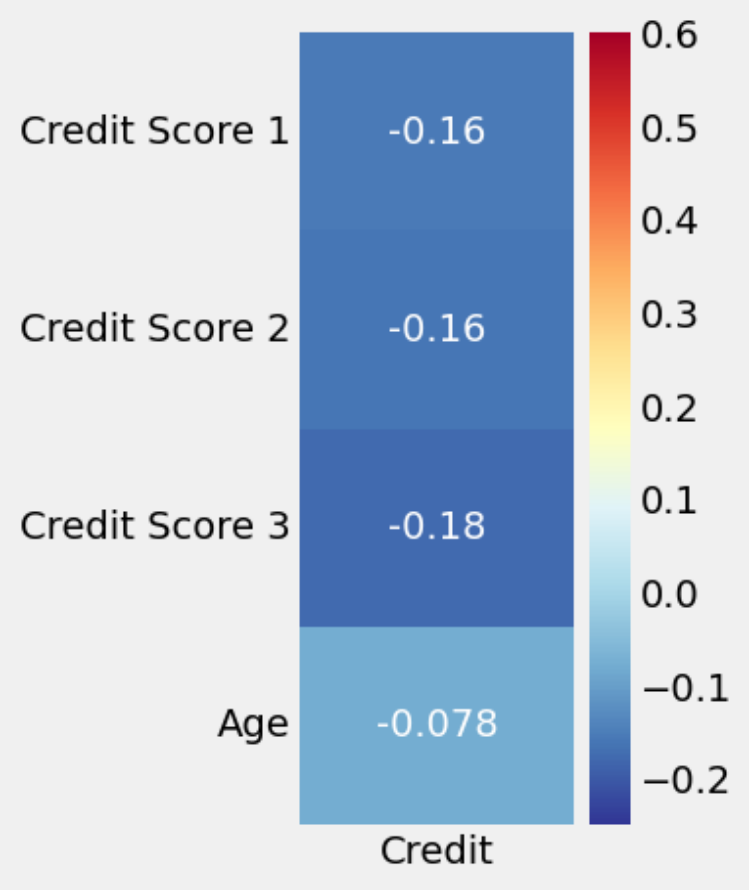


características relevantes para determinar el riesgo de crédito, estos son:

- **Porcentaje Ingresos**: Es el porcentaje de crédito relativo al ingreso del individuo.
- **Tiempo Crédito**: Es la cantidad de tiempo que le toma al cliente cancelar el crédito.
- **Porcentaje de Empleo Edad**: Es el porcentaje de días empleado con respecto a la edad del cliente.

Adicionalmente, en base a las variables de la Figura 2.2, se crearon variables polinomiales de hasta tercer grado, para estimar de forma más precisa a la variable objetivo.

Figura 2.2.
Correlaciones con la variable objetivo



2.3 Implementación de los modelos

Empleando Python como lenguaje de programación para implementar los modelos, se desarrollaron los algoritmos *from scratch*, es decir, sin utilizar ninguna librería de ML para emplearlos. Para la visualización de los resultados se utilizaron las librerías **seaborn** y **matplotlib**. Se implementaron tres modelos. El primero, fue estimado de forma Lineal. Luego, se utilizó la librería de Scikit-Learn para compararlo. El tercero, se lo estimó de forma No Lineal en los atributos o características, empleando polinomios de las variables que tienen más

correlación con la variable objetivo (Figura 2.2). Todos los modelos se estimaron añadiendo pesos sobre la variable objetivo, debido a que es un problema de tipo desbalanceado. Esto ocurre cuando hay más registros de unas clases que de otras, por tanto, los clasificadores tienden a verse abrumados por las clases grandes e ignoran las pequeñas (Zhao and Cen, 2013). En efecto, la distribución de la variable objetivo es más de un 90% de casos negativos con respecto a los positivos. Por tanto, se añadieron pesos a la parte positiva para lograr un mejor balance dentro de la estimación de los modelos. Adicionalmente, se empleó el criterio de *trust region* (Moré and Sorensen, 1983), el cual consiste en añadir una matriz diagonal muy pequeña a la matriz Hessiana. Esto es necesario para que el algoritmo pueda calcular la inversa de la matriz Hessiana sin ningún problema.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

3.1 Resultados

Para el entrenamiento de los modelos se dividió la base de datos de manera aleatoria. El 90% de los datos se empleó para entrenamiento y el restante para el testeo. Una vez entrenados los modelos, se realizaron las predicciones mediante testeo haciendo uso de la función sigmoide junto con los parámetros estimados.

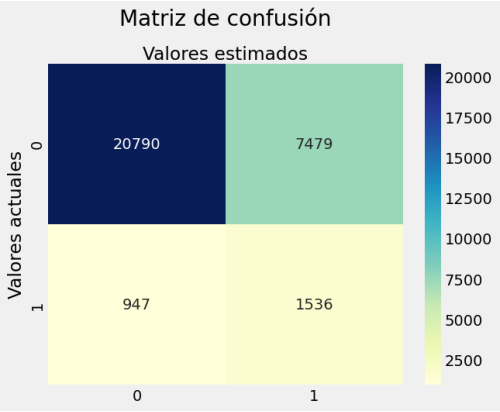
Para evaluar los modelos, se utilizó el indicador de Precisión, que mide las predicciones correctas sobre el total de predicciones. Adicionalmente, se emplearon las métricas de la Matriz de Confusión y la Curva ROC-AUC. De la Matriz de Confusión se obtuvieron los indicadores de Sensibilidad, que corresponden 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 cómo cambia la tasa de verdaderos positivos en función de la tasa de falsos positivos cuando cambia el umbral de decisión del modelo. Mientras más cercano a 1 sea el valor de AUC, el test es mejor. Por defecto, en los modelos de regresión logística la línea de base es de 0.5.

Tabla 3.1.
Resultados de los modelos con sus respectivos indicadores.

Modelos	Precisión	Sensibilidad	Especificidad	AUC
Lineal	0.7260	0.6186	0.7354	0.6770
Scikit-Learn	0.7263	0.6210	0.7356	0.6783
No Lineal	0.7272	0.6251	0.7361	0.6806

Figura 3.1.
Resultados para el modelo Lineal.

(a)
Matriz de confusión



(b)
Curva ROC-AUC

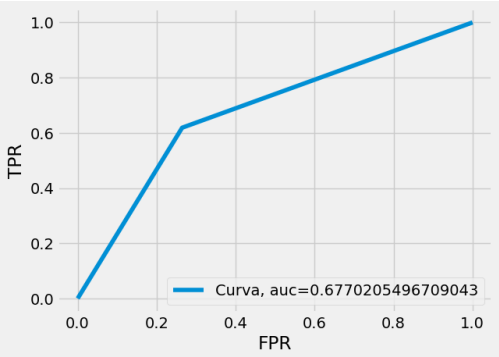
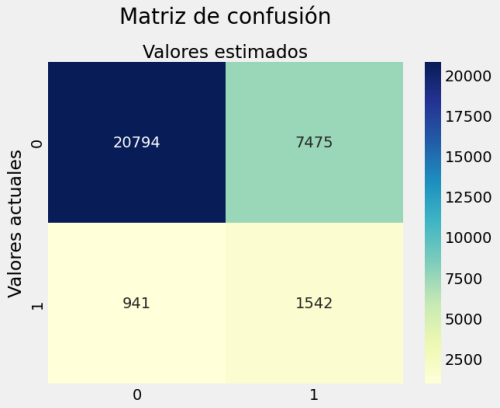


Figura 3.2.
Resultados para el modelo con Scikit-Learn

(a)
Matriz de confusión



(b)
Curva ROC-AUC

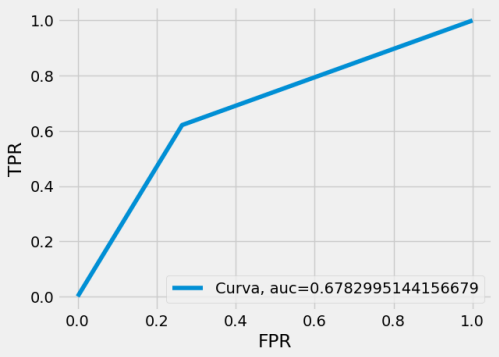
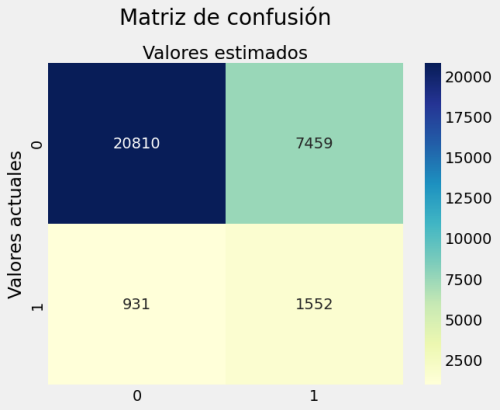
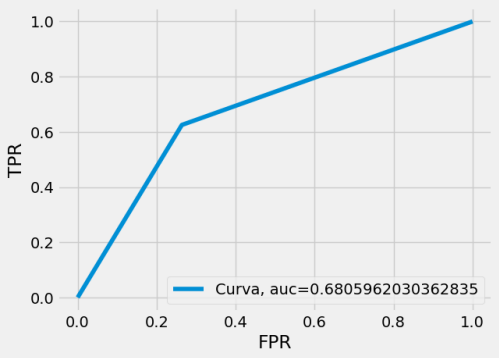


Figura 3.3.
Resultados para el modelo No Lineal

(a)
Matriz de confusión



(b)
Curva ROC-AUC



3.2 Análisis de Resultados

Es importante destacar que se realizó una estimación utilizando pesos relativos para la variable **CREDIT**, debido a que sus datos están desbalanceados, es decir, más del 90% son préstamos sin problemas, lo que hace que al entrenar el modelo no se considere la otra clase. Se recuerda que lo que busca la institución financiera es reducir la clasificación de falsos negativos para evitar la repercusión de esto en su rentabilidad y tasa de morosidad.

A partir de la comparación de los indicadores de los modelos, en la Tabla 3.1 se visualiza que todos se comportan de forma similar, no obstante, el modelo No Lineal alcanza un puntaje ligeramente más alto en todas las métricas en comparación con los otros modelos. Esto se debe a que las variables polinómicas seleccionadas capturan mejor la complejidad de los datos.

La precisión se mantiene casi constante en tres modelos, con valores muy cercanos (0.7264 en Scikit-learn, 0.7260 en el modelo Lineal y 0.7272 en el No Lineal). Esto sugiere que todos los modelos tienen un rendimiento similar. Sin embargo, el modelo No Lineal tiene una ligera ventaja, lo que indica que tiene una mejor capacidad para identificar correctamente.

La sensibilidad tiene resultados similares. Esto indica que estos modelos tienen un mejor desempeño al identificar las instancias positivas reales, lo cual es crucial en el contexto de los mercados financieros donde la detección de verdaderos positivos es prioritaria.

En cuanto a la especificidad, los tres modelos presentan valores muy cercanos. El modelo Scikit-learn tiene una especificidad de 0.7356, el Lineal de 0.7354 y el No Lineal de 0.7361. Aunque las diferencias son mínimas, el modelo polinómico se destaca, sugiriendo una mayor habilidad para evitar falsos positivos.

El AUC es un indicador general de la capacidad del modelo para distinguir entre clases.

En este caso, el modelo No Lineal presenta el valor más alto 0.6806, seguido por Scikit-learn 0.6783 y el modelo Lineal 0.6770. Aunque las diferencias son pequeñas, un mayor valor de AUC en el modelo sugiere un mejor rendimiento general en la discriminación entre clases positivas y negativas.

En general, el modelo No Lineal muestra un desempeño superior o comparable en todos los indicadores clave, especialmente en términos de sensibilidad y AUC. Esto sugiere que el modelo polinómico es más efectivo para capturar la complejidad de los datos y distinguir entre clases, lo que lo convierte en la mejor opción entre los enfoques evaluados.

Aunque las diferencias en algunos indicadores, como la precisión y la especificidad, son mínimas, la mejora en la sensibilidad y el AUC del modelo No Lineal tiene un impacto significativo en aplicaciones donde es crucial maximizar la detección de verdaderos positivos sin sacrificar demasiado en términos de falsos positivos. Por tanto, para el contexto del proyecto donde se requiere una mayor sensibilidad y una mejor discriminación general entre clases, el modelo No Lineal es el de mejor desempeño.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- Se ha logrado identificar que las variables de calificación externa y la edad son relevantes para la construcción de los modelos predictivos de riesgo crediticio. Además, el rendimiento superior del modelo de regresión logística No Lineal sugiere que las relaciones polinomiales entre las variables pueden ser significativas en la predicción del riesgo crediticio.
- Se implementaron los modelos de regresión logística tanto Lineal como No Lineal. Los resultados mostraron que el modelo No Lineal tiene un rendimiento ligeramente superior en términos de sensibilidad y AUC. Esto indica que, en el contexto de la predicción de riesgo crediticio, las relaciones capturadas por el modelo polinómico pueden proporcionar una mejor comprensión y predicción del comportamiento de los clientes.
- Al comparar las métricas de rendimiento, se observó que los modelos lineales son competitivos en cuanto a precisión y especificidad. Sin embargo, la regresión No Lineal se destaca en términos de sensibilidad y AUC, lo que sugiere que es más efectiva para identificar casos de alto riesgo (verdaderos positivos), lo cual es relevante en el contexto financiero.

- El modelo No Lineal fue el más efectivo, sin embargo, al hacer estimaciones polinomiales puede incurrir en *overfitting*. Lo cual hace que el modelo no siempre sea replicable.
- Los modelos contribuyen a que la toma de decisiones de crédito sea mucho más ágil y rápida, y se realice con base en el análisis de cada institución financiera y sus clientes.

Recomendaciones

- Dado que el modelo de regresión logística No Lineal mostró un rendimiento superior en términos de sensibilidad y AUC, se recomienda priorizar este modelo en la implementación práctica para la predicción del riesgo crediticio. Este enfoque permitirá una mejor identificación de clientes de alto riesgo, lo cual es esencial para mitigar pérdidas financieras.
- Se sugiere investigar otras posibles relaciones entre variables que no fueron capturadas por el modelo actual. Asimismo, abordar el problema con otros modelos de ML como Random Forest, Redes Neuronales, Support Vector Machine (SVM),XGBoost, etc., que permitan obtener resultados adicionales u otros criterios al momento de predecir riesgo crediticio.
- Se recomienda realizar un ajuste al criterio de paro para reducir tiempos y emplear técnicas adicionales de regularización para optimizar más los resultados.
- Es recomendable para futuras investigaciones, implementar técnicas para verificar si los algoritmos al clasificar el riesgo o la probabilidad de default están incurriendo en algún tipo de discriminación injusta, ya sea étnica, de género, etc.

- Se puede emplear una heurística que facilite la obtención de una mejor solución inicial, lo que permite reducir el número de iteraciones en la convergencia del algoritmo. Asimismo, utilizar métodos de reducción de dimensión para reducir el tamaño de la base de datos, y de modo que se pueda enfocar mejor el problema.

BIBLIOGRAFÍA

Duffie, D. and Singleton, K. J. (2012). Credit risk: pricing, measurement, and management. In *Credit Risk*. Princeton university press.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Luenberger, D. G. and Ye, Y. (2021). *Linear and nonlinear programming*. Springer.

Makhado, P. (2023). The limitations of traditional credit scoring systems. https://medium.com/@phindulo60/the-limitations-of-traditional-credit-scoring-systems-e92833fdfa8a [Accessed: 04/08/2024].

Moré, J. J. and Sorensen, D. C. (1983). Computing a trust region step. *SIAM Journal on scientific and statistical computing*, 4(3):553–572.

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons.

Thomas, L., Crook, J., and Edelman, D. (2017). *Credit scoring and its applications*. SIAM.

Vizhñay, A y Samaniego, A (2019). Determinantes del acceso al crédito en el ecuador. *Revista Espacios*, 40(13):25–36.

Zhao, Y. and Cen, Y. (2013). *Data mining applications with R*. Academic Press.