ENTREGA FINAL IA

Presentado por:

Sebastian Castro Bolaños

Jorge Andres Cardeño Devia

Profesor

Raúl Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín 2023-2

RIESGO DE INCUMPLIMIENTO DE CRÉDITO HIPOTECARIO

Nuestro propósito en este proyecto es predecir la capacidad de pago de una población no bancarizada mediante el uso de machine learning. Utilizaremos datos históricos de solicitudes de préstamos para anticipar si un solicitante podrá cumplir con un préstamo o no. Home Credit busca mejorar su capacidad para evaluar el riesgo de impago y tomar decisiones informadas sobre la aprobación o denegación de solicitudes de préstamo.

Este proyecto se alinea con la competición "Home Credit Default Risk" de Kaggle, la cual desafía a los participantes a desarrollar modelos de aprendizaje automático para prever la probabilidad de incumplimiento de pago por parte de los clientes de Home Credit, una institución financiera no bancaria. La evaluación del riesgo crediticio en este grupo se complica debido a la falta de datos crediticios tradicionales.

Este informe se sumerge en el riesgo de incumplimiento de crédito hipotecario, centrándose en el análisis de datos proporcionados para la competición. El conjunto de datos incluye información diversa sobre los solicitantes de crédito, abarcando aspectos demográficos, historiales de pago y más. Los participantes deben desarrollar modelos predictivos precisos que estimen la probabilidad de incumplimiento de pagos de crédito.

Durante el análisis exploratorio de datos, evaluaremos la distribución de la variable objetivo, identificaremos desequilibrios de clases y analizaremos la calidad del conjunto de datos. Además, exploraremos relaciones y patrones dentro de los datos para extraer conocimientos relevantes. La competición "Home Credit Default Risk" en Kaggle presenta un desafío crucial en el desarrollo de modelos efectivos de aprendizaje automático para prever el riesgo de incumplimiento de crédito hipotecario, y este informe ofrece un enfoque analítico detallado para abordar este desafío.

Importancia de los archivos:

application_train.csv: Este archivo contiene los datos de entrenamiento principales, que incluyen información sobre los solicitantes de crédito.

application_test.csv: Este archivo contiene los datos de prueba, y se utiliza para evaluar el rendimiento de los modelos predictivos.

POS_CASH_balance.csv: Contiene información mensual sobre los saldos de los préstamos anteriores relacionados con créditos en tiendas.

bureau.csv: Proporciona datos adicionales sobre los préstamos anteriores de los solicitantes de crédito, si están disponibles.

previous_application.csv: Proporciona datos sobre las aplicaciones previas de los solicitantes de crédito en Home Credit.

installments_payments.csv: Contiene información sobre los pagos mensuales de los préstamos anteriores.

credit_card_balance.csv: Proporciona datos mensuales sobre los saldos de las tarjetas de crédito de los solicitantes de crédito.

bureau_balance.csv: Este archivo contiene información mensual sobre el saldo de deuda de los préstamos anteriores del archivo "bureau.csv".

Exploración de datos

En la fase de exploración de datos, nos encontramos con un conjunto amplio y diverso de información, compuesto por un total de 122 variables. La variable que se busca predecir y que representa un componente crítico en nuestra tarea se denomina "TARGET".

Desglosando la composición de estas variables, observamos:

65 Variables Continuas: Estas son aquellas que pueden tomar un rango infinito de valores dentro de un intervalo determinado, proporcionando una gama diversa de información numérica.

41 Variables Discretas: A diferencia de las continuas, las variables discretas toman valores específicos y finitos.

16 Variables Categóricas: Estas variables representan categorías o grupos distintos y no tienen un orden inherente.

La diversidad de estas variables proporciona una riqueza de información que será exhaustivamente explorada durante el análisis de datos. Este proceso nos permitirá identificar patrones, relaciones y posibles correlaciones que serán fundamentales para el desarrollo de modelos predictivos robustos en nuestro objetivo de evaluar el riesgo de incumplimiento de crédito hipotecario.

Variables de interés

En el análisis de datos, se identifican varias variables clave que arrojarán luz sobre la capacidad de pago de los clientes y contribuirán significativamente al desarrollo de modelos predictivos. A continuación, se presentan algunas de las variables de interés destacadas:

NAME_INCOME_TYPE: Representa el tipo de ingreso del cliente, proporcionando información crucial sobre la fuente de ingresos y su estabilidad.

NAME_EDUCATION_TYPE: Indica el nivel educativo alcanzado por el cliente, lo cual puede tener implicaciones directas en la capacidad financiera y la estabilidad laboral.

NAME_FAMILY_STATUS: Describe el estado familiar del cliente, proporcionando insights sobre el entorno familiar y sus posibles efectos en la capacidad de pago.

NAME_HOUSING_TYPE: Refleja la situación de vivienda del cliente, indicando si vive en una propiedad arrendada, con sus padres, entre otras opciones. Esto puede influir en la estabilidad financiera.

DAYS_BIRTH: Representa la edad del cliente en días desde que aplicó. Este dato es esencial para evaluar la relación entre la edad y la capacidad de pago.

DAYS_EMPLOYED: Indica cuántos días antes de solicitar el préstamo el cliente comenzó su empleo actual. Este dato proporciona información sobre la estabilidad laboral.

OCCUPATION_TYPE: Describe el tipo de ocupación del cliente, lo cual es crucial para entender la naturaleza de su empleo y sus ingresos asociados.

ORGANIZATION_TYPE: Indica el tipo de organización en la que trabaja el cliente, ofreciendo perspectivas adicionales sobre la estabilidad laboral y el entorno de trabajo.

CODE_GENDER: Refleja el género del cliente, un factor importante que puede influir en las características financieras y en la capacidad de pago.

Valores nulos

La gestión de valores nulos en nuestro conjunto de datos es fundamental para el análisis, y hemos establecido un umbral del 25% como criterio para determinar la relevancia de las variables. Aquellas con más del 25% de valores nulos plantean desafíos para el análisis, y nuestro enfoque se centra en variables que consideramos importantes.

Es crucial analizar por qué faltan valores en estas variables específicas. Por ejemplo, la variable "OWN_CAR_AGE" puede ser intrigante de analizar, a pesar de tener un 65% de datos faltantes. La ausencia de estos datos podría indicar que la falta de información está directamente relacionada con la inexistencia de un vehículo propio. Este enfoque selectivo nos permite gestionar eficientemente los valores nulos mientras extraemos información valiosa y contextualizada de variables clave para nuestro objetivo de evaluación del riesgo crediticio.

Preprocesamiento de datos:

Iniciamos el proceso cargando los datos en un entorno de programación mediante el uso de bibliotecas como Pandas. Posteriormente, exploramos los datos cargados para comprender las columnas y tipos de datos presentes. La fase de limpieza de datos incluyó acciones como el tratamiento de valores faltantes y la corrección de errores.

Para facilitar el modelado, llevamos a cabo la codificación de variables categóricas, transformándolas en representaciones numéricas adecuadas. Además, normalizamos o escalamos las variables para garantizar que tengan una escala comparable, contribuyendo a un análisis más consistente.

Como último paso en este proceso, nos enfocamos en la selección o ingeniería de características, identificando las más relevantes o creando nuevas características según sea necesario.

Modelo supervisado

En el proceso de modelado supervisado, comenzamos por separar las características (variables independientes) de la variable objetivo (variable dependiente). Luego, dividimos los datos en conjuntos de entrenamiento y prueba. Seleccionamos un modelo supervisado apropiado para el problema, que puede incluir opciones como regresión logística, árboles de decisión o redes neuronales. Posteriormente, entrenamos el modelo utilizando los datos de entrenamiento y realizamos predicciones con el modelo entrenado utilizando los datos de prueba. Finalmente, evaluamos el rendimiento del modelo empleando métricas pertinentes, tales como precisión, recall o error cuadrático medio. Este enfoque sistemático garantiza una implementación efectiva de modelos supervisados en la resolución de problemas específicos.

Los modelos que utilizamos son:

- Random Forest Classifier
- SGD Classifier

Las metricas utilizadas:

Númericas:

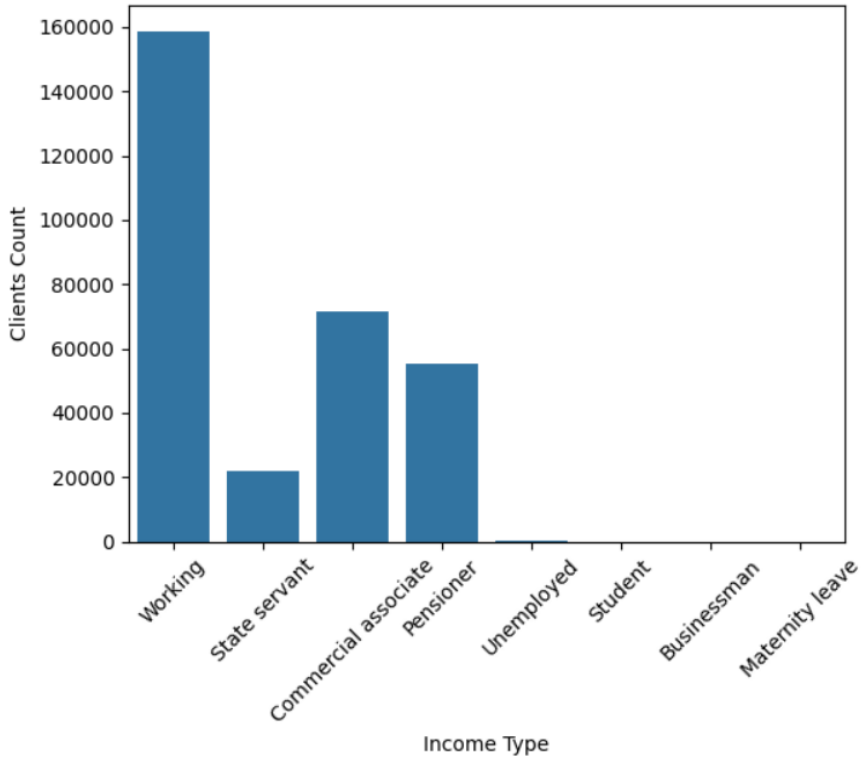
- F1 Score
- Recall
- Precisión
- Matriz de confusión

Gráficas:

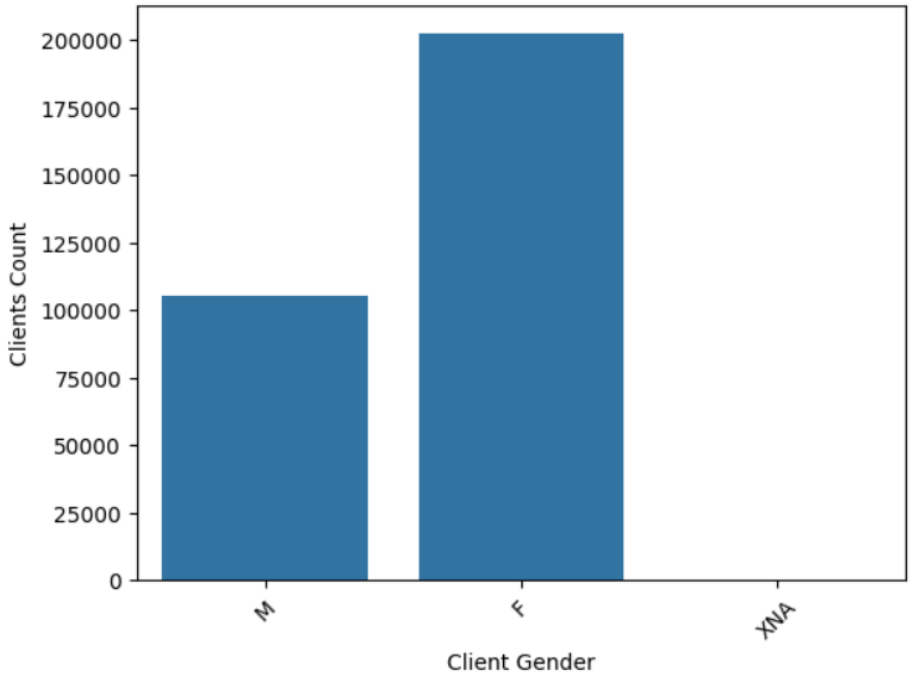
- Precision - Recal
- Curva ROC.l

Analisis de las gráficas

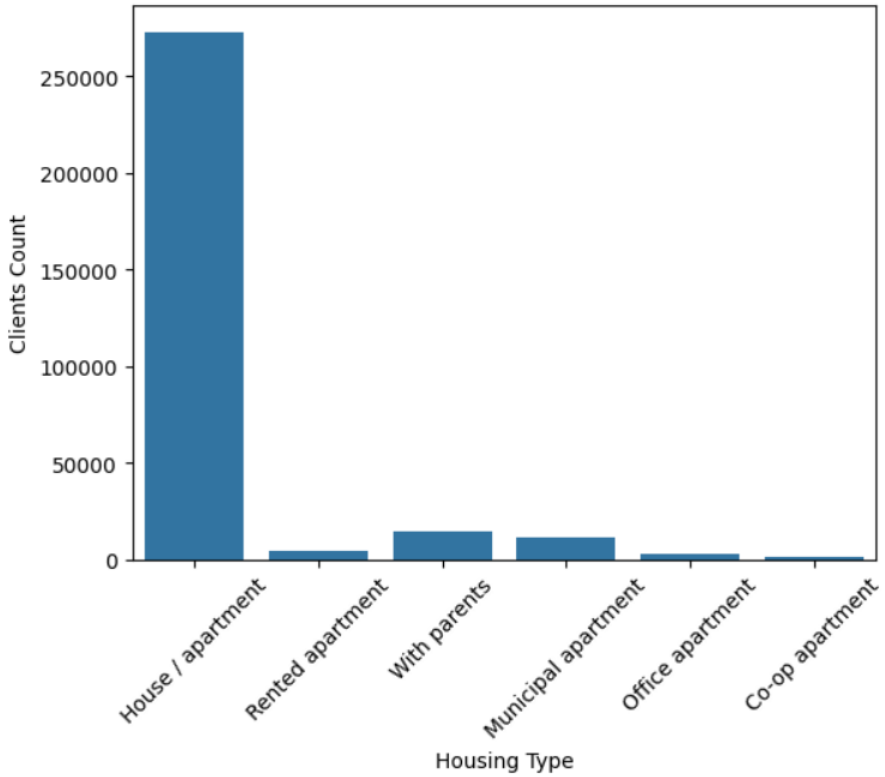
Tipo de ingreso: De esta gráfica podemos ver que la mayoria de los clientes que están solicitando los prestamos son trabajadores del común.



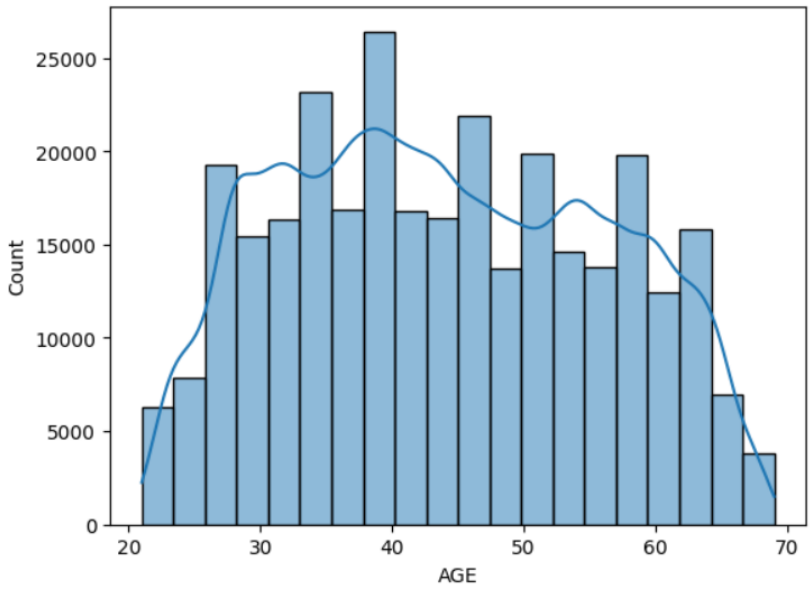
Genero de los clientes : De esta gráfica podemos ver que la mayoria de los clientes que solicitan prestamos son mujeres.



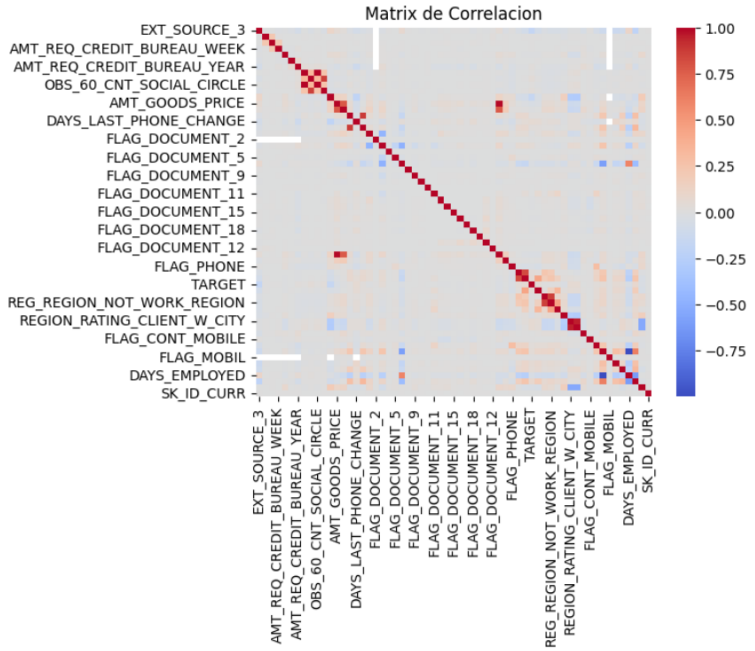
Tipo de vivienda: De esta gráfica podemos ver que la mayoria de los clientes que están solicitando los prestamos ya tiene una vivienda tipo casa o apartamento propia.



Variabilidad de las edades: De esta gráfica podemos ver la variabilidad de las edades de los clientes que están solicitando los prestamos.



Matriz de correlación:
La matriz de correlación revela que varias de las variables numéricas seleccionadas no presentan relaciones lineales entre sí. Este hallazgo destaca la diversidad y complejidad de estas variables, indicando la necesidad de enfoques más avanzados y no lineales en nuestro análisis.



Retos y consideraciones:

Pre procesamiento complejo: El conjunto de datos puede contener una amplia variedad de características, algunas de las cuales pueden estar incompletas o requerir una limpieza y transformación exhaustiva. Fue necesario realizar un pre procesamiento adecuado, como tratar los valores faltantes, manejar características categóricas y normalizar las variables numéricas.

La efectividad de nuestro modelo se mide por su precisión, la cual debería superar el 80%. Este umbral es crucial para asegurar que la empresa no otorgue préstamos a personas con dificultades, protegiendo tanto los intereses individuales como las ganancias de la compañía.

Conclusiones

La gestión de valores nulos es esencial para nuestro análisis, y hemos establecido un umbral del 25% como criterio para determinar la relevancia de las variables. Aunque las variables con más del 25% de valores nulos plantean desafíos para el análisis, nuestro enfoque se centra en aquellas que consideramos importantes.

La exploración de datos revela un conjunto diverso y extenso de información, compuesto por un total de 122 variables. La variable objetivo, denominada "TARGET", representa el foco central de nuestra tarea de predicción. Dentro de este conjunto, observamos una variedad de 65 variables continuas, 41 variables discretas y 16 variables categóricas. La diversidad en la naturaleza de estas variables proporciona una oportunidad valiosa para analizar en profundidad el conjunto de datos, identificar patrones significativos y desarrollar modelos predictivos efectivos.

La relevancia del preprocesamiento de datos se destaca en la gestión de datos desafiantes y complejos en nuestra tarea de evaluar el riesgo de incumplimiento de crédito hipotecario. Dedicamos una considerable cantidad de tiempo a la limpieza y transformación de datos, una etapa crucial que influyó directamente en la calidad de los modelos finales desarrollados. Este proceso ha demostrado ser fundamental para obtener resultados precisos y confiables.

Elegimos detalladamente las características más cruciales para anticipar el riesgo de incumplimiento crediticio. A través de un proceso de selección de características, logramos potenciar la precisión de los modelos y, al mismo tiempo, disminuir su complejidad.

El análisis de la matriz de correlación revela que muchas de las variables numéricas seleccionadas carecen de relaciones lineales significativas entre sí. Este hallazgo es crucial, ya que indica la diversidad y complejidad del conjunto de datos. La falta de correlaciones lineales sugiere que estas variables pueden contener información única y complementaria, lo que destaca la importancia de un enfoque más holístico y no lineal en nuestro análisis.

Durante el desarrollo del proyecto, hemos reconocido la vitalidad del aprendizaje constante. Exploramos innovadoras técnicas, probamos diversos enfoques y nos mantenemos al día con los avances en el ámbito del aprendizaje automático. Esta vivencia ha contribuido a nuestro crecimiento profesional y ha generado la motivación necesaria para continuar perfeccionándonos en futuros proyectos.

Elección del modelo adecuado: Evaluamos y probamos varios modelos supervisados para determinar cuál era el más adecuado para nuestro problema. Encontramos que algunos modelos tenían un mejor desempeño que otros en términos de precisión y capacidad para manejar los desafíos específicos de los datos.

Selección de características significativas: Identificamos las características más relevantes para predecir el incumplimiento crediticio. Al realizar una cuidadosa selección de características, pudimos mejorar la precisión de los modelos y reducir la complejidad.

Interpretación de los resultados: Además de obtener predicciones precisas, fue importante comprender y explicar los factores clave que influyen en el incumplimiento crediticio. Esto nos permitió brindar información valiosa a los interesados y tomar decisiones más informadas en futuras estrategias de gestión de riesgos crediticios.

En resumen, este proyecto nos ha brindado la oportunidad de desarrollar competencias en el manejo de datos complejos, la selección de características, la elección de modelos y la

interpretación de resultados. También hemos reconocido la importancia de abordar desafíos particulares y hemos experimentado el proceso iterativo y colaborativo inherente al trabajo en un proyecto de aprendizaje automático.

Enlace de la competición en kaggle:
https://www.kaggle.com/c/home-credit-default-risk/overview

Enlace del video entrega final:
https://youtu.be/d3g4VstmpBc