CODER HOUSE

Data Science
Desafío Entregable
Alumna: Camila Varela

Modelo de predicción de incumplimento de préstamos

¿Cómo mitigar el riesgo de crédito mediante la manipulación de los datos?

AGENDA

1. Contexto, Objetivos y Audiencia
2. Hipótesis/Preguntas de Interés
3. Análisis Exploratorio
4. Data Wrangling
5. Conclusiones

CONTEXTO Y **AUDIENCIA**

Contexto

En el campo de la gestión del riesgo crediticio, los bancos utilizan técnicas de aprendizaje automático (ML) para crear varios modelos para predecir los incumplimientos de los préstamos. Los modelos de aprendizaje automático ayudan a detectar patrones en los datos, que luego se utilizan para categorizar nuevos registros. Si los préstamos no se reembolsan, los bancos sufren pérdidas financieras. Para minimizar este problema, el objetivo es mantener las tasas de impago lo más bajas posible.

Objetivos

El fin de este caso práctico es desarrollar una comprensión del análisis de riesgos en los servicios bancarios y financieros y entender cómo se utilizan los datos para mitigar o minimizar el riesgo de perder dinero al conceder préstamos a los clientes.

Este análisis pretende identificar patrones que indiquen si un cliente tiene dificultades para saldar su deuda, lo que es útil para tomar medidas cómo denegar el préstamo, reducir el importe del mismo, establecer un tipo de interés más alto a solicitantes de riesgo, entre otras medidas.

Para ello se extrajo una base de datos bancaria de Kaggle con el fin de poder analizar las cuestiones anteriormente mencionadas. Los datos se exploran y analizan inicialmente, seguido de un preprocesamiento para prepararlos para el modelado. Como última instancia, el objetivo de este caso práctico, es abordar finalmente su rendimiento.

Audiencia

Este proyecto está dirigido a profesionales del sector financiero, analistas de datos, científicos de datos y cualquier persona interesada en la aplicación de técnicas de análisis de datos y aprendizaje automático en el ámbito bancario. También puede ser relevante para gerentes y tomadores de decisiones en instituciones financieras que busquen optimizar sus procesos de evaluación de riesgo y mejorar su rentabilidad. Al proporcionar recomendaciones basadas en insights observados a partir del análisis de datos, este proyecto puede ayudar a los profesionales a tomar decisiones más informadas y mitigar el riesgo asociado al impago de tarjetas de crédito.

PREGUNTAS DE **INTERÉS**

Preguntas principales:
Las mismas que se podrán responder a partir del análisis del caso práctico

- ¿Por qué es importante la modelización del riesgo crediticio para las entidades financieras?
- ¿Con qué mètodos puede el modelado del riesgo crediticio ayudar a las entidades financieras a identificar riesgos potenciales y mitigarlos?
- ¿Cómo influye el modelado del riesgo crediticio en la determinación de la solvencia de una entidad financiera?
- ¿Cuáles son los pasos clave involucrados en el desarrollo de un modelo de riesgo crediticio?
- ¿Cómo pueden las entidades financieras evaluar la eficacia de sus modelos de riesgo crediticio y realizar mejoras si es necesario?

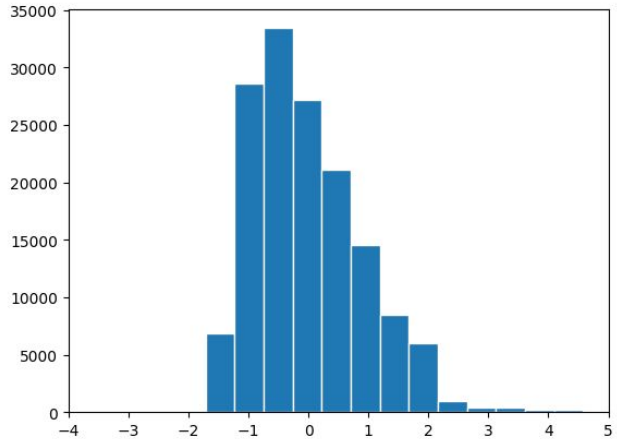
ANÁLISIS EXPLORATORIO DE DATOS

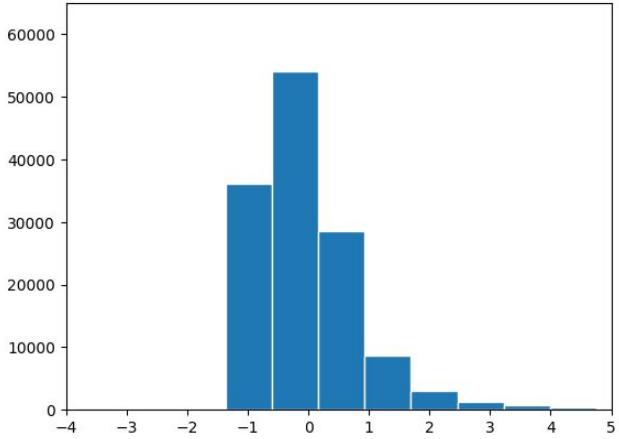
Aplicando MÉTODO Z SCORE

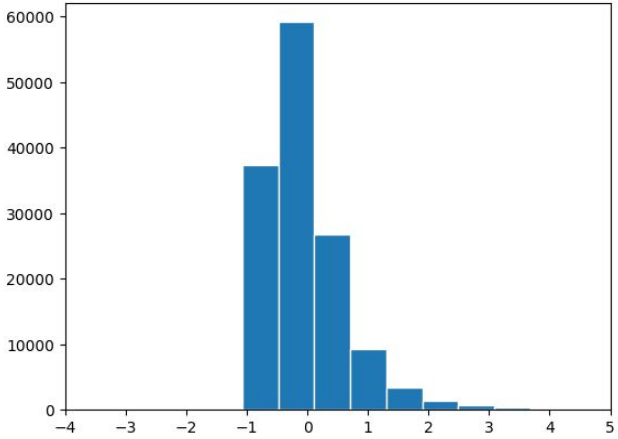
Representación del **monto de los préstamos** otorgados

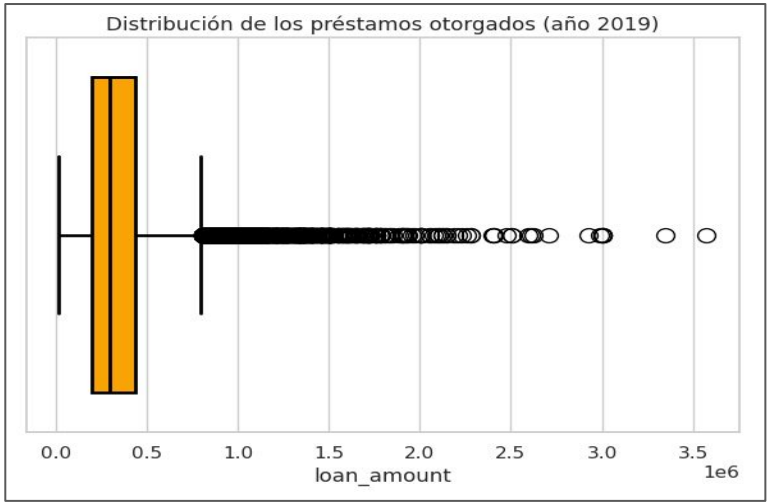
Representación del **valor de las propiedades** de los prestatarios

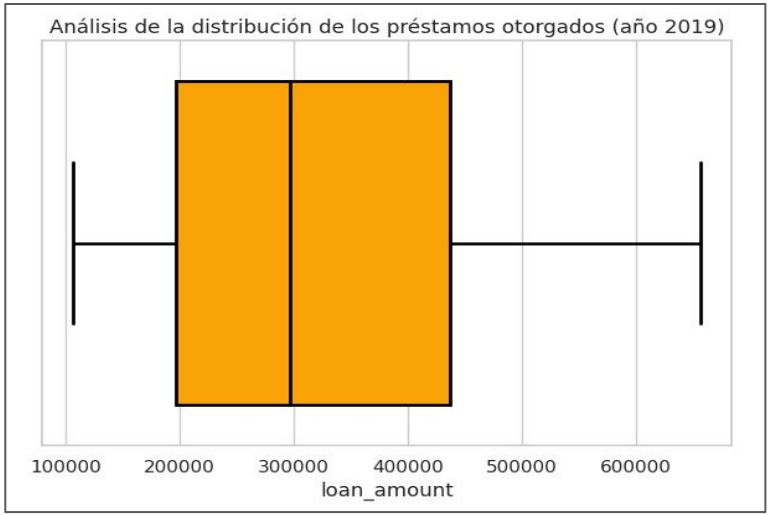
Representación de los **ingresos** de los prestatarios











ANÁLISIS EXPLORATORIO DE DATOS

En el gráfico podemos observar muchos outliers, estos son puntos de datos que se alejan significativamente del resto de observaciones en un conjunto de datos, los outliers pueden tener un gran impacto sobre los resultados de los análisis estadísticos al afectar medidas sensibles a los valores extremos, como la media y la desviación estándar.

Una vez identificados, debemos decidir qué hacer con ellos.

Existen diferentes métodos para gestionarlos, uno de ellos es la winsorización.

La winsorización es una técnica estadística que reemplaza los valores atípicos por el valor más cercano que no se considera un outlier según ciertos criterios, tiene como objetivo mitigar el impacto de los valores extremos en el análisis estadístico. Esta técnica proporciona robustez, es una alternativa sólida para manejar valores atípicos y datos sesgados. En lugar de descartar puntos de datos valiosos los preserva limitando sus valores.

Por ejemplo, podemos reemplazar los valores atípicos por el valor del percentil 95 o 5 más cercano, reduciendo el efecto de los outliers sin eliminarlos por completo.

En el caso de la primera imagen podemos ver como los datos de la variable **loan_aomunt** (monto del préstamo) se encuentran sesgados, El sesgo en los datos genera resultados que no son completamente representativos de la información que estás investigando.

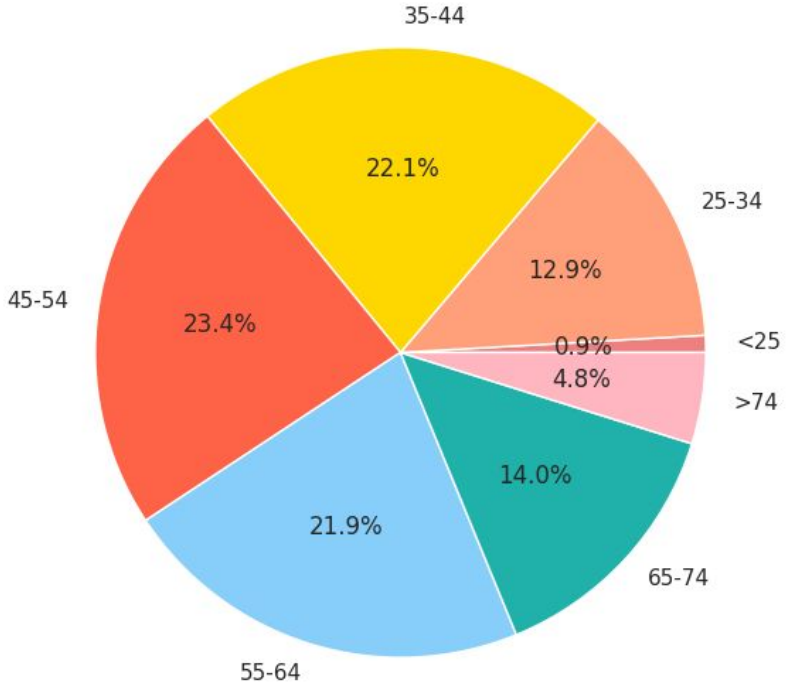
El hecho de haber aplicado esta técnica estadística nos ayudó a comprender la distribución de los datos, proporcionó información sobre la forma y la simetría del conjunto de datos, lo que nos permite descubrir patrones y anomalías que pueden no ser aparentes de inmediato.

ANÁLISIS EXPLORATORIO DE DAT

RANGO ETARIO

Los clientes de entre 35 y 65 años conforman aproximadamente un 68% de los préstamos otorgados.
Por otra parte los menores de 25 años conforman el grupo etario con menos préstamos otorgados.

La **edad, el score y antigüedad crediticia** es un factor CLAVE al momento de una análisis crediticio favorable.



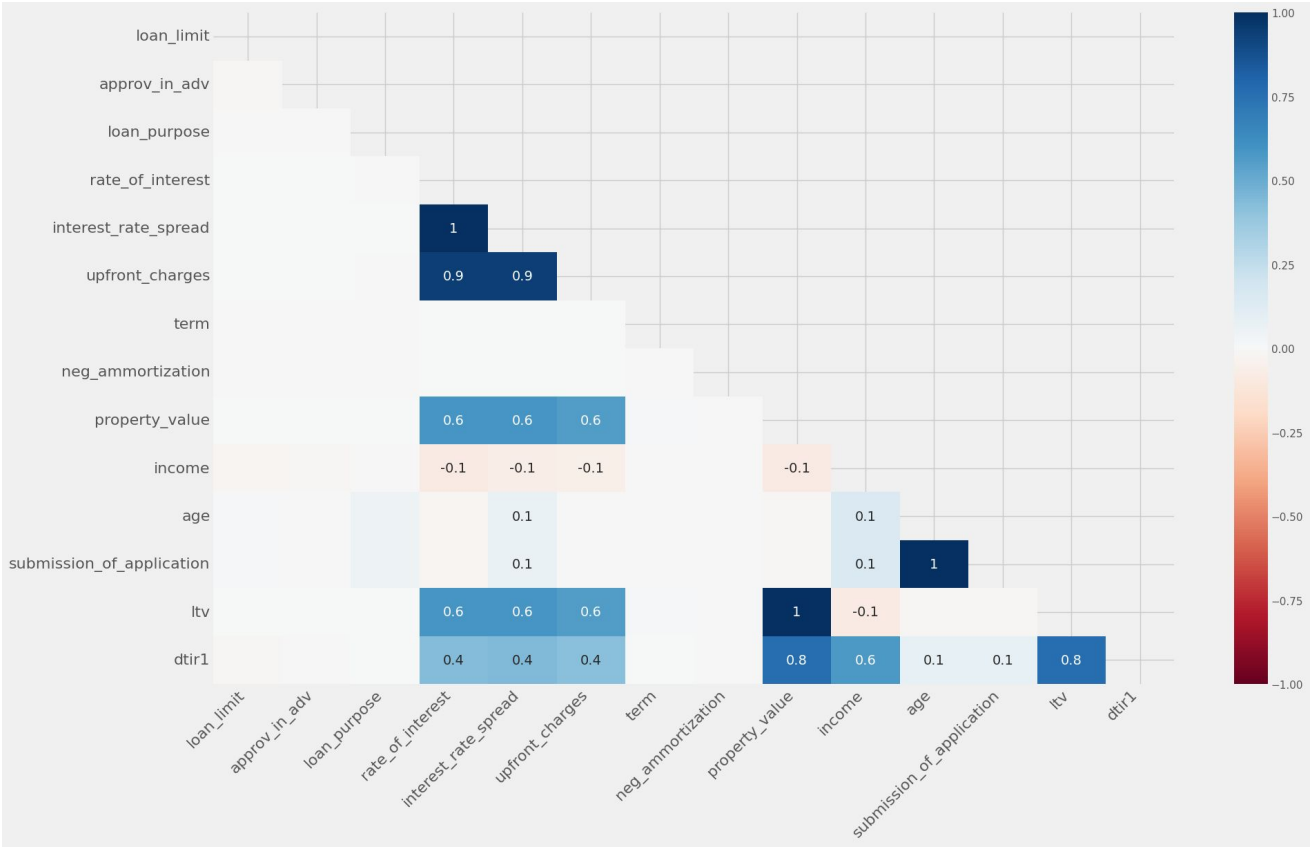
ANÁLISIS EXPLORATORIO DE DATOS

- Observamos los **valores nulos** y evalúo la correlación de los valores faltantes

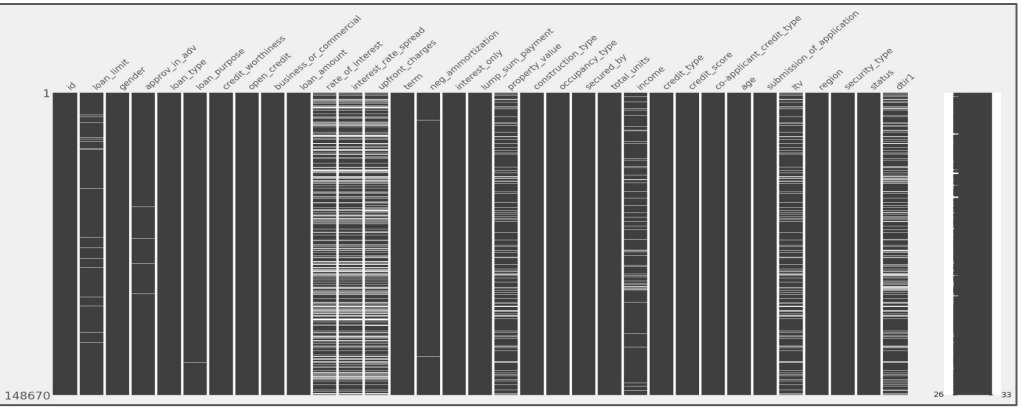
En esta gráfica se puede ver fácilmente el nivel de correlación entre los valores nulos de las diferentes features.

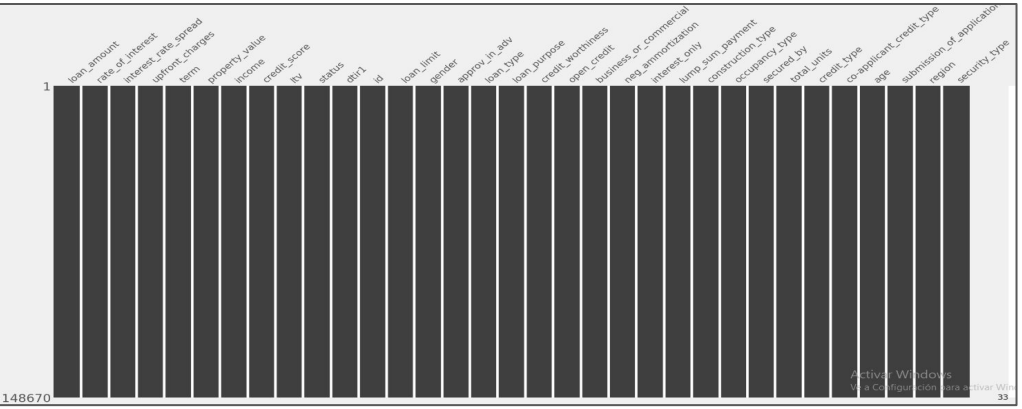
En la barra de la derecha se muestra el nivel de correlación positiva entre dos valores, el mismo se indica mediante la intensidad de azul.

A modo de ejemplo, se puede observar gran correlación entre la variable property_value' (valor de la propiedad) y 'LTV (1), lo cual no es casual dado que el ratio loan to value (LTV), o relación valor-préstamo se usa habitualmente en tramites hipotecarios para medir el porcentaje entre la cantidad del préstamo y el valor del inmueble que se utiliza como garantía.



Manipulación de valores nulos





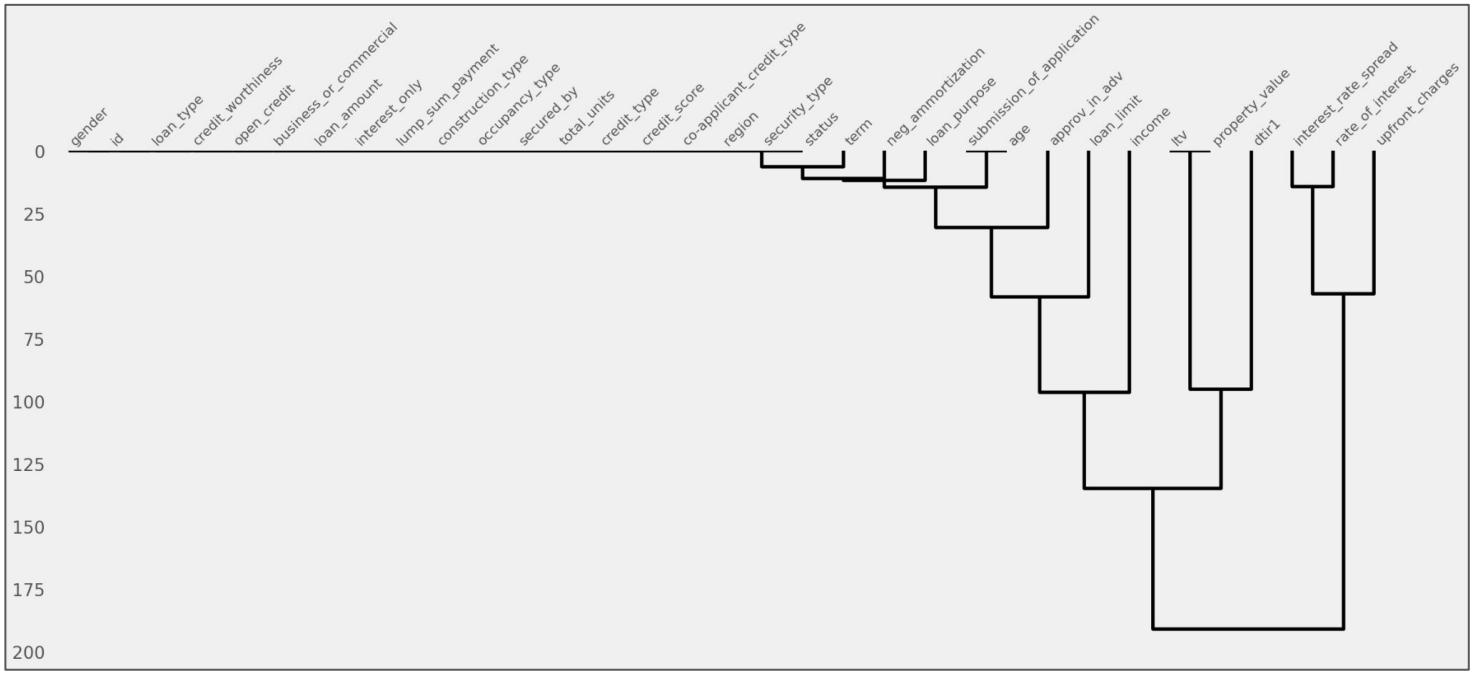
La función matrix de Missingno permite visualizar concretamente dónde se encuentran los valores nulos en un conjunto de datos.
Las marcas blancas del anterior gráfico representan los valores perdidos. Mediante este gráfico es más fácil encontrar patrones y vínculos existentes entre los missing values en las diferentes variables.

Datos numéricos. Podríamos haber completado los valores nulos utilizando la mediana de la columna, pero esto a menudo sesgará los datos, especialmente cuando faltan muchos valores. Una forma eficaz de completar valores NaN para datos numéricos es utilizar KNNImputer de sklearn, el mismo utiliza las columnas que tienen valores para estimar el valor nulo de una determinada celda.

Datos categóricos : Aquí tomamos el valor más probable (o el valor que aparece con más frecuencia) para esa columna. Una vez que se imputaron todos los valores nulos, por último se concatenó el dataframe numérico con el categórico

Clustering Jerárquico

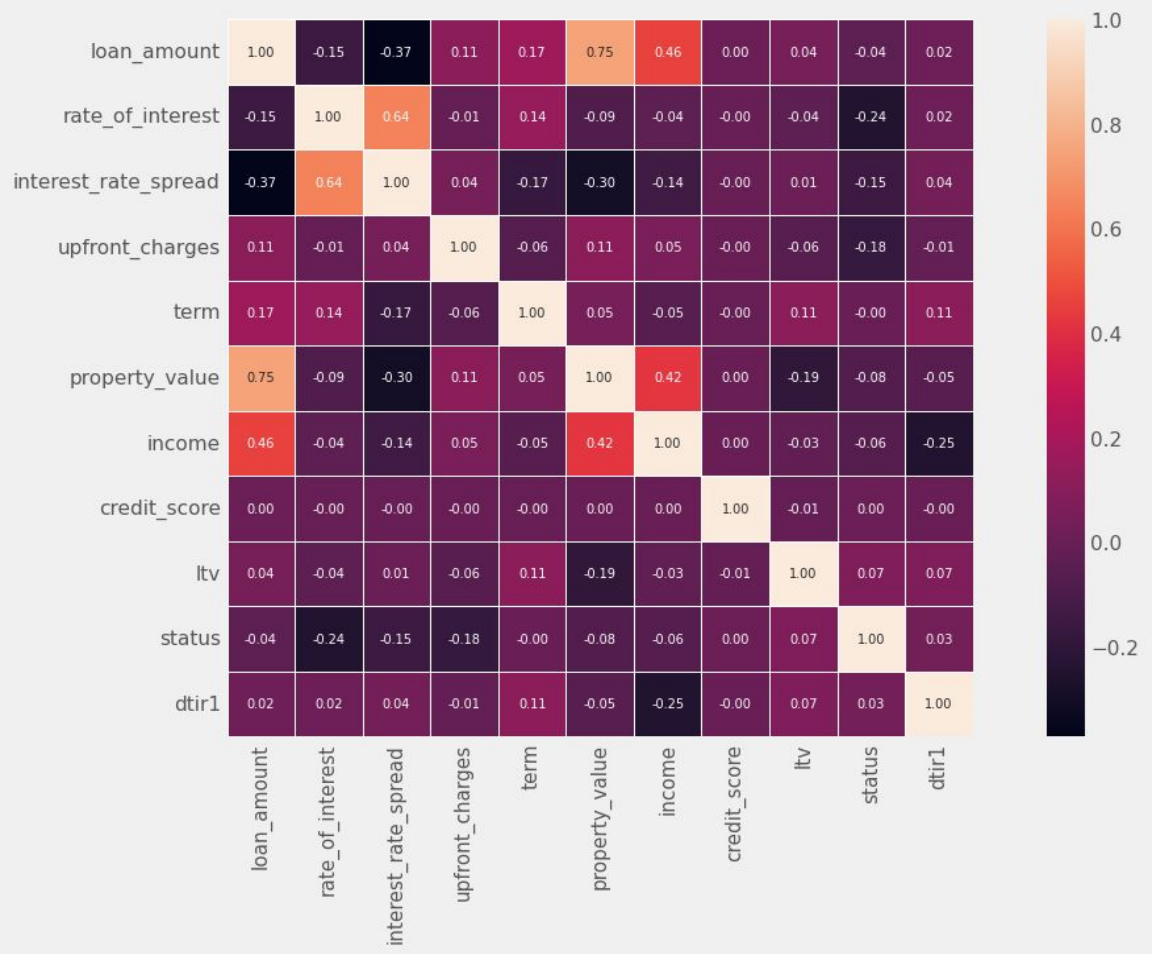
El **dendrograma** utiliza un algoritmo de agrupamiento jerárquico para agrupar variables entre sí según su correlación de nulidad (medida en términos de distancia binaria). En cada paso del árbol, las variables se dividen según qué combinación minimiza la distancia de los grupos restantes.



PCA - ANÁLISIS DE COMPONENTES PRINCIPALES

En primer lugar hay que tener en cuenta que NO se pòdra realizar un correcto análisis de componentes principales si no se eliminan los nulos y también es de gran importancia saber que esta técnica de reducción de dimensiones requiere de un dataframe conformado por variables cuantitativas, ya sea continuas o discretas.

En esta gráfica se puede ver fácilmente el nivel de correlación entre las variables cuantitativas.



APLICACIÓN DE ALGORITMO DE ML

XGBoostClassifier

De acuerdo a los resultados obtenidos en el gráfico:

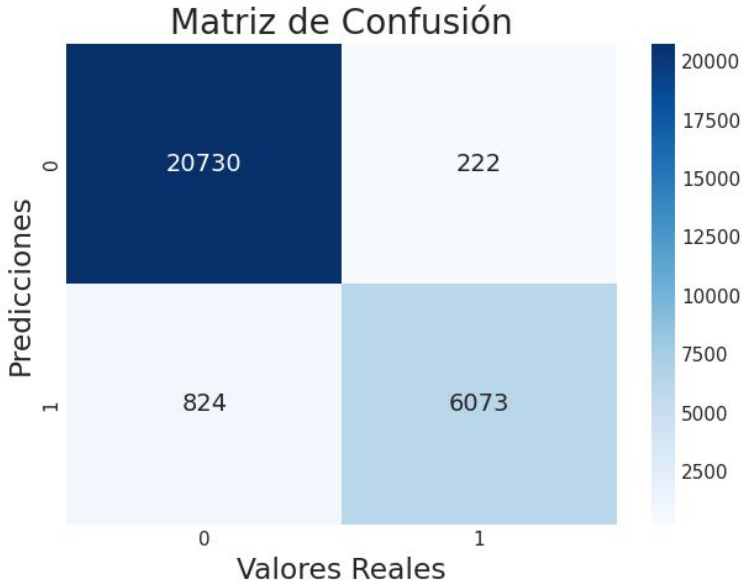
El valor 20730 representa la cantidad de casos negativos que fueron clasificados correctamente como negativos (verdaderos negativos, VN).

El valor 222 indica la cantidad de casos negativos que fueron clasificados incorrectamente como positivos (falsos positivos, FP).

El valor 824 indica la cantidad de casos positivos que fueron clasificados incorrectamente como negativos (falsos negativos, FN).

El valor 6073 representa la cantidad de casos positivos que fueron clasificados correctamente como positivos (verdaderos positivos, VP).

En resumen, el modelo clasificó correctamente 20730 casos negativos y 6073 casos positivos. Sin embargo, se cometieron 824 falsos negativos y 222 falsos positivos.



CONCLUSIONES

Se realizó el EDA sobre el dataset obteniendo una descripción general de las variables que la componen

Se logró obtener un dataset sólido, se limpieros valores nulos y se trabajó cautelosamente los outliers, dado que estos pueden indicar errores en los datos, influir en los análisis estadísticos e impactar en los procesos de toma de decisiones.

Hay seis columnas en las que un valor particular no se vio con frecuencia en el conjunto de datos y se consideraría un valor atípico. Sin embargo, para cada una de estas columnas, los casos en los que se encuentran valores poco frecuentes son casos en los que el prestatario incumplió. Por lo tanto, estos son datos importantes que debemos conservar.

Está claro que los altos valores de las propiedades NO resultan en más incumplimientos.

Podemos visualizar que los datos indican que la tasa de interés se incrementa porcentualmente al conceder préstamos a los jóvenes de menores de 25 años años, esto se debe a que no cuentan con historial crediticio consistente ni una antigüedad/score que le permita al banco asegurar el retorno de las obligaciones que hayan contraído, lo cual es indispensable para la entidad financiera para determinar cuán riesgoso es otorgarle a una persona un producto crediticio.
Por estas cuestiones la tasa de interés es ligeramente mayor.

Utilicé XGBoost ya que es un modelo de aprendizaje automático empleado para problemas de aprendizaje supervisado, en el que utilizamos los datos de entrenamiento para predecir una variable objetivo/respuesta. Tiene la capacidad de realizar relaciones complejas en datos, técnicas de regularización para evitar el sobreajuste y la incorporación de procesamiento paralelo para un cálculo eficiente.