DESARROLLO DE UN MODELO DE RIESGO DE CRÉDITO

Helga Zambrana

CODER HOUSE

Agenda

CODER HOUSE

1. Presentación del problema
2. Objetivo de la investigación
3. Descripción del dataset elegido
4. Hallazgos del análisis exploratorio de datos (EDA)
5. Visualizaciones

1 Presentación del problema

El riesgo de crédito analiza la probabilidad de que un prestatario no reembolse un préstamo solicitado, generando pérdidas en el prestamista.

Los bancos comerciales y de inversión, los fondos de capital riesgo, las empresas de gestión de activos y las compañías de seguros, entre otras entidades financieras, tienen la necesidad de gestionar el riesgo de crédito para mitigar las pérdidas.

A través del desarrollo de un modelo de aprendizaje automático se busca facilitar las aprobaciones de créditos, ayudando a predecir qué clientes son más propensos a dejar de pagar sus deudas.

2 Objetivo de la investigación

CODER HOUSE

Analizar la información disponible de los aplicantes a un préstamo e identificar las posibles variables que determinen que el solicitante no pagará el préstamo

Analizar diferentes opciones de modelos de clasificación y seleccionar el que tenga la mejor performance en predecir si el solicitante pagará o no el préstamo

Desarrollar un modelo de riesgo de crédito en Python para predecir las probabilidades de impago (PD) y asignar puntuaciones de crédito a los solicitantes existentes o potenciales

3 Descripción del dataset elegido

El dataset elegido fue Home Equity (HMEQ) de Kaggle.

HMEQ presenta información sobre las características y la morosidad de 5960 préstamos con garantía hipotecaria. Un préstamo con garantía hipotecaria es un préstamo en el que el deudor utiliza el capital de su vivienda como garantía subyacente.

Los criterios de selección del dataset fueron la claridad de los datos recolectados y su robustez, lo que facilita el análisis, el procesamiento y la generación de un modelo.

3 Descripción del dataset elegido

CODER HOUSE

Nombre de la variable	Tipo de variable	Descripción
BAD	Binaria	1, corresponde a un prestamista con deuda impaga o con mora; 0, corresponde al candidato con los pagos al día. Se considera morosidad cuando han pasado 90 días desde que el prestamista no ha pagado
LOAN	Numérica continua	Monto en USD del préstamo solicitado
MORTDUE	Numérica continua	Monto en USD adeudado de la hipoteca existente
VALUE	Numérica continua	Monto en USD de la propiedad hipotecada
REASON	Categórica	Motivo para solicitar el préstamo DebtCon, corresponde a consolidación de deudas; HomeImp, corresponde a mejoras para el hogar
JOB	Categórica	Profesión o categoría profesional
YOJ	Numérica discreta	Cantidad de años en el trabajo actual

3 Descripción del dataset elegido

Nombre de la variable	Tipo de variable	Descripción
DEROG	Numérica discreta	Número de informes derogatorios importantes. Es información proporcionada por una institución financiera a las agencias de crédito y se relaciona con la morosidad o la cancelación de una cuenta de una línea de crédito
DELINQ	Numérica discreta	Número de líneas de crédito morosas
CLAGE	Numérica discreta	Edad expresada en meses de la línea de crédito de mayor antigüedad, los modelos de puntuación consideran un mínimo de 6 meses y recién a partir de 2 años es fiable
NINQ	Numérica discreta	Número de veces que ha solicitado un nuevo crédito en los últimos 2 años
CLNO	Numérica discreta	Número de líneas de crédito abiertas
DEBTINC	Numérica continua	Ratio deuda-ingreso (DTI), es la cantidad de ingresos brutos mensuales que una persona genera frente a la deuda que debe pagar por mes. Los prestamistas generalmente buscan ratios no mayores al 36%, aunque un DTI del 43% puede calificar para una hipoteca

4 Hallazgos del análisis exploratorio de datos (EDA)

CODER HOUSE

Durante el análisis y pre-procesamiento de los datos se eliminaron y transformaron las variables que no aportarán al modelo y se encontraron los siguientes hallazgos

Variable Target	El dataset cuenta con una variable target definida 'BAD'. El valor 1 corresponde al candidato con préstamo incumplido o con mora y el valor 0 corresponde al candidato que paga su deuda.
Valores nulos y duplicados	No existen registros duplicados. 11 de 13 columnas tienen datos faltantes. 'DEBTINC' es la variable con mayor cantidad de valores nulos con un 21.3% de su total. Se reemplazan los valores nulos declarándolos como valores desconocidos.
Desbalanceo de datos	Se observa que de 5960 aplicantes, solo el 19.9% representa a personas con deuda morosa.

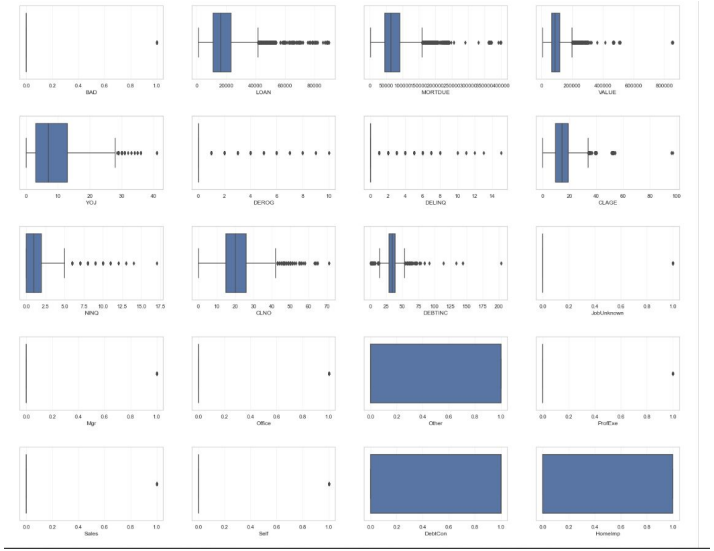
4 Hallazgos del análisis exploratorio de datos (EDA)

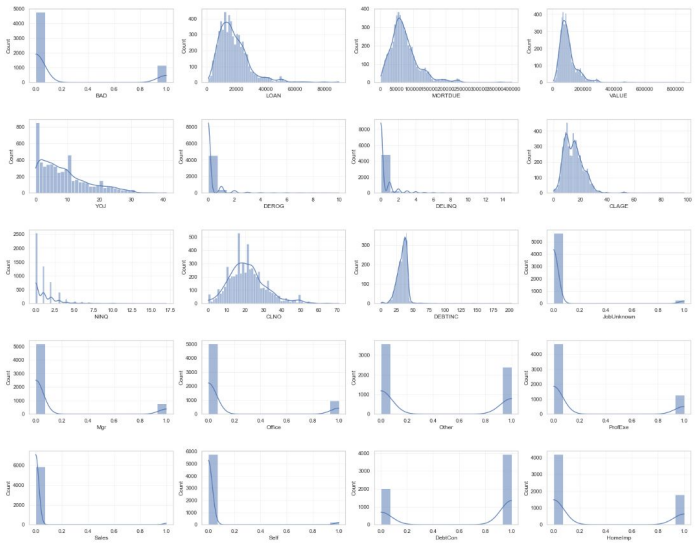
CODER HOUSE

Variables Categóricas vs Variables Numéricas

Variables Categóricas	Según la variable 'REASON', 'DebtCon' (consolidación de deudas) es el motivo principal para solicitar un préstamo, entre quienes pagan o no sus deudas. A partir de la muestra, inferimos que si la razón es 'HomeImp' (mejoras del hogar), la diferencia entre quienes pagan y no pagan disminuye. Según la variable 'JOB', y teniendo en cuenta la diferencia dentro de cada categoría, inferimos que la diferencia es poca entre quienes tienen su hipoteca al día y quienes no entre los profesionales de 'Sales' (Ventas) y 'Self' (Autónomos). Transformar 'JOB' y 'REASON' de variables categóricas a variables numéricas con GETDUMMIES
Variables Numéricas	Pasar la variable 'CLAGE' de meses a años para igualar con la unidad de medida de 'YOJ' La media de las variables en USD 'LOAN', 'MORTDUE' y 'VALUE' es mayor en las muestras de préstamos PAGADOS. Como es esperable, la media de la variable 'DEBTINC' (ratio de deuda-ingresos) es mayor en el caso de préstamos MOROSOS, con un posible Outlier de 203.3121. La media de las variables relacionadas con el historial crediticio de la persona que solicita el préstamo, que incluye 'DEROG', 'DELINQ' y 'NINQ', es mayor en el caso de los préstamos MOROSOS. La media de la variable 'CLNO' (líneas de crédito abiertas) es similar en ambos casos.

5 Visualizaciones

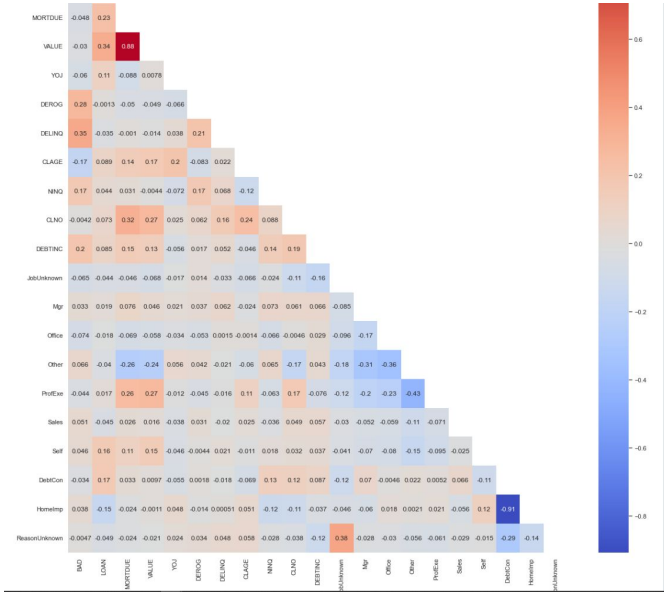




Conclusiones

Se necesita normalizar todas las características porque la escala de cada atributo es diferente.
Varios atributos tienen una distribución sesgada.
'LOAN', 'MORTDUE', 'VALUE', 'DEBTINC' cuentan con gran cantidad de outliers.

5 Visualizaciones



Conclusiones

El mapa de correlaciones puede indicar una relación predictiva a ser explotada en el modelo de clasificación. Los colores más fríos corresponden a una baja correlación, los más cálidos corresponden a una alta correlación.

Las variables relacionadas con el historial crediticio ('DELINQ', 'DEROG', 'NINQ') son las más correlacionadas con la variable target ('BAD'). Es un indicio de que pueden ser utilizadas como variables de clasificación, aunque como están ligeramente correlacionadas entre sí sugiriendo que la información podría ser redundante.

El monto adeudado de la hipoteca existente ('MORTDUE') y el valor de la garantía subyacente ('VALUE') están correlacionadas entre si, podría ser datos redundantes

GRACIAS

Helga Zambrana | Data Science
LinkedIn