MEMORIA “EXPLORATORY DATA ANALYSIS”

INFORMACIÓN DEL PROYECTO.

En nuestro proyecto partimos de un “dataset” con una serie de columnas que aportan información sobre créditos concedidos y sobre el prestatario que ha recibido estos créditos.

Tratamos de conocer qué variables inciden sobre el impago de un préstamo, si la incidencia es significativa y si estas variables pueden servir de base para la creación de un modelo de Machine Learning que permita anticipar la probabilidad de impago de un préstamo.

Los datos de los que partimos son proporcionados por el reconocido “Lending Club”. Se trata de una compañía estadounidense, con sede en San Francisco, de préstamos entre particulares, que opera por internet y que provee de financiación al margen de la banca tradicional. Esta compañía no opera en España, por lo que su comunicación con ellos solo puede hacerse vía telefónica.

En la siguiente página de Kaggle pueden obtenerse datos de los préstamos que otorgaron el “Lending Club” de los ejercicios 2007-2018 con más de 3 Gigas de datos https://www.kaggle.com/datasets/imsparsh/lending-club-loan-dataset-2007-2011.

No obstante, para nuestro análisis, y a efectos de simplificar la carga de datos, hemos optado por utilizar una versión reducida de estos datos, con unos 40.000 registros de 2011 que se usa a modo de demostración del servicio Amazon Sagemaker Canvas perteneciente al entorno AWS de Amazon.

https://catalog.us-east-1.prod.workshops.aws/workshops/80ba0ea5-7cf9-4b8c-9d3f-1cd988b6c071/en-US/1-use-cases/4-finserv

Los datos originales están repartidos en 2 archivos de tipo “csv” llamados “loans-part-1” y “loans-part-2”.

El archivo “loans-part-1” se compone de las siguientes columnas e información:

- Id: número de identificación del préstamo.
- loan_status: indica si el préstamo está vigente, pagado o impagado.
- loan_amount: importe del préstamo
- funded_amount_by_investors: cantidad recibida por el prestatario descontadas comisiones.
- loan_term: número de pagos para amortizar el préstamo (36 o 60)
- interest_rate: tipo de interés del préstamo
- installment: cuota mensual
- grade: clasificación del préstamo según criterio del LC
- subgrade: subclasificación del crédito según criterio del LC
- verification_status: indica si los ingresos del prestatario han sido verificados
- issued_on: fecha de concesión del préstamo.
- purpose: propósito para el que se pide el préstamo

- dti: ratio del total de cuotas exceptuando la hipoteca del prestatario dividido por sus ingresos anuales.
- Inquiries_last_six_months: número de consultas del préstamo en los últimos 6 meses.
- open_credit_lines: número de préstamos vigentes del prestatario .
- derogatory_public_records: registros públicos negativos del prestatario
- revolving_line_utilization_rate: se trata de una tasa de uso de crédito mediante tarjetas “revolving”
- total_credit_lines: total de créditos del prestatario en la base de datos del LC

El archivo “loans-part-2” se compone de las siguientes columnas e información:

- id: número de identificación del préstamo
- employment_length: años trabajados por el prestatario (la cifra 10 indica 10 o más años)
- employer_title: sector del empleador del prestatario
- home_ownership: régimen de la vivienda del pretatario.
- annual_income: ingresos anuales del prestatario

HIPÓTESIS DEL PROYECTO

Tratamos de contrastar si algunas de las variables de ambos dataset, que consideramos relevantes, tienen una incidencia significativa en el impago de un préstamo.

En concreto, analizaremos si las siguientes variables tienen incidencia significativa o no en el impago de un préstamo:

1. El importe del préstamo
2. El tipo de interés
3. El importe de la cuota
4. El ratio “dti”
5. El acceso a créditos “revolving”.
6. El tiempo que lleva trabajando el prestatario
7. El régimen de su vivienda
8. El propósito para el que fue concedido el préstamo
9. Los ingresos anuales del prestatario

ANÁLISIS INICIAL DE DATOS

En el análisis inicial de datos hacemos las siguientes operaciones y comprobaciones:

1. Comprobamos que los índices (columna “id”) de ambos archivos coinciden a fin de comprobar si podemos aplicarles el método “merge”. Efectivamente ambos dataset tienen los mismo índices y le aplicamos el “merge” para unirlos en uno solo.
2. Comprobamos características del dataset obtenido (“shape”, “info()”, etc). Comprobamos que las columnas “revolving_line_utilization_rate” y “employment_length” tienen valores nulo.
3. Comprobamos la cardinalidad de las variables categóricas. La columna “employer_title” tiene una alta cardinalidad.
4. Comprobamos que no hay filas duplicadas.

5. Consideramos que las siguientes columnas no son relevantes, o no las vamos a necesitar, para nuestro análisis, por lo que procedemos a eliminarlas:
    - funded_amount_by_investors: no es más que el importe prestado menos comisiones.
    - grade: es una clasificación dada por el LC
    - subgrade: idem al anterior
    - verification_status: si los ingresos del prestatario han sido comprobados.
    - issued_on: fecha de concesión del préstamo. Es irrelevante.
    - inquiries_last_month: veces que se ha consultado el crédito en la BD.
    - open_credit_lines: créditos abiertos por el prestario
    - derogatory_public_records: registros públicos negativos.
    - total_credit_lines: número de créditos del prestatario en la BD
    - employer_title: alta cardinalidad. No es práctico su análisis.
6. Renombramos la columna “revolving_line_utilization_rate” por “revolving_rate”
7. Hacemos un análisis estadístico descriptivo con el método “describe” tanto de las variables numéricas como paraméstricas.
8. Comprobamos valores únicos de cada columna del dataset restante (df_final). El más importante de todos, ya que va a constituir nuestra variable a estudiar es “loan_status”. Se compone de 3 valores:
    - “fully_paid”: 32.920 filas (préstamos pagados)
    - “charged off”. 5.627 filas (préstamos impagados)
    - “current”: 1.140 filas (préstamos vigentes)

LIMPIEZA DE DATOS

➢ LIMPIEZA DE VALORES NULOS

Tenemos 2 columnas que contienen valores nulos:

- “revolving_rate”: 50 filas Nan
- “employment_length”: 1075 filas Nan

Empleamos 2 estrategias para asignar los valores Nan en cada una de ellas.

a) Columna “revolving_rate”. Hacemos un estudio de la distribución de esta columna y vemos que los valores 0 son significativamente superiores al resto. En concreto, los valores 0 son el 2,46% del total, el siguiente valor tiene el 0,16% del total. En este caso suponemos que los valores Nan son simplemente valores 0 que no se han completado en la BD, por lo que procedemos a asignar el valor 0 a los Nan.
b) Columna “employment_lenth”. Tiene 1075 columnas Nan. De un análisis gráfico y numérico no sacamos conclusiones claras sobre si los Nan pueden ser una valor en concreto. Tampoco de un análisis de 30 primeras filas con valor Nan sacamos un patrón concreto. Con esta columna, procedemos a estudiar el peso relativo de cada uno de los 10 valores que la componen, y procedemos a repartir los valores Nan en las filas correspondientes de manera aleatoria en función de la proporción inicial de cada valor.

➢ LIMPIEZA DE OUTLIERS.

De un análisis inicial de gráficos del tipo “boxplot” a las variables numéricas contínuas, vemos que las columnas “loan_amount”, “interest_rate” y “annual_income” tienen numerosos “outliers”.

Para reducir su número, procedemos a considerar “outlier” los valores que sean superiores a 6 veces el rango intercuartílico de cada una de estas variables. Con esta metodología, solamente la columna “annual_income” pasa a tener “outliers”.

Hacemos un estudio específico de esta variables, y vemos que de 39.717 filas, solamente 144 valores son “outliers”, lo que representa el 0,36% del total, siendo el valor mínimo de los “outliers” de 334.000 y el máximo de 6.000.000. Con esta información, pasamos a otorgar el importe calculado de 6 veces el rango intercuartílico a cada unos de estos 144 valores.

De esta manera conseguimos eliminar todos los “outliers” de nuestro dataset.

ANÁLISIS UNIVARIANTE Y BIVARIANTE

Una vez limpiado nuestro dataset, hacemos una división del mismo en los siguientes dataframes:

1) df_final_current: dataframe con los préstamos vigentes
2) df_final: dataframe con los créditos pagados e impagados
3) df_final_pais: dataframe con los créditos pagados
4) df_final_charged_off: dataframe con los créditos impagados

A continuación, mostramos un gráfico de tipo “pairplot” de las columnas que tenemos en el dataframe final. De este gráfico no podemos extraer visualmente conclusiones claras.

A continuación seguimos con un análisis de las distintas variables, tanto de su distribución como de su relación con nuestra variable de estudio que es el “loan_status”

Distinguimos si la variable es numérica continua o categórica (o numérica discreta con pocos valores)

a) **Variables numéricas continuas.** Hacemos 3 tipos de gráficos en cada una de ellas
   - → Histograma de 20 bins de la variable continua para ver qué forma tiene su distribución.
   - → Histograma de 20 bins de la variable continua comparando los valores de préstamos pagados e impagados para cada tramo.
   - → Mapa de calor con los porcentajes relativos de préstamos impagados por tramos de la variable continua (20 tramos)
b) **Variables categóricas y numéricas discretas con pocos valores.** En este caso hacemos 2 tipos de gráficos de cada una de ellas:
   - → Valores totales de cada categoría de la variable diferenciando si el préstamo ha sido pagado o no.
   - → Mapa de calor de los valores relativos de impagados por cada categoría de la variable

ANÁLISIS MULTIVARIANTE

En este apartado hacemos un estudio y mapa de calor de las posibles correlaciones entre las variables numéricas continuas calculadas según el método de Pearson.

Seguimos viendo un gráfico de las correlaciones entre estas variables con su distribución por puntos y la línea de tendencia que siguen.

CONTRASTE DE HIPÓTESIS

En nuestro análisis queremos analizar la posible incidencia que cada una de las variables tiene sobre la variable que queremos estudiar, que es si el préstamo deviene en impagado o no.

De acuerdo con esta premisa, en todas las variables usamos la siguiente lógica para hacer el contraste de hipótesis:

- Hipótesis Nula: No hay evidencia significativa de que la variable en estudio tenga incidencia en el impago o no del préstamo
- Hipótesis Alternativa: Hay una incidencia significativa de la variable en estudio en el impago o no del crédito

De acuerdo con el estudio gráfico, no podemos concluir que ninguna de las variables siga una Distribución Normal. Por otro lado, la variable que queremos contrastar es la columna "loan_status" que tiene 2 posibles valores: "paid" o "charged off"

De acuerdo con esta información, las alternativas que tenemos para hacer nuestro contraste son dos:

a) Test de Mann-Whitney: Para contrastar la incidencia de las variables numéricas en la variable objetico ("loan_status"), ya que de ninguna de ellas podemos asegurar que es una distribución normal.
b) Test de independencia Chi-Cuadrado: para contrastar si las variables categóricas o numéricas discretas con pocos valores.

Aplicados los test de hipótesis, en todas las variables rechazamos la hipótesis nula, y por tanto todas ellas tienen incidencia en el impago o no del préstamo

CONCLUSIONES

Las principales conclusiones que podemos tomar de nuestro análisis serían las siguientes:

1. Todas las variables analizadas tienen una en nuestra variable de estudio (préstamo impagado o no).
2. Dado que tenemos una serie de variables independientes, unas de tipo numérico continuas, otra numérica discreta con pocos valores, y otras categóricas, y puesto que todas ellas inciden en la variable dependiente ("loan_status"), podríamos hacer una regresión de tipo logístico que nos sirviera de modelo para predecir la probabilidad de pago de un préstamo.

3. Una vez que tengamos desarrollado el modelo, podemos a plicarlo a los créditos aún vigentes, a fin de tomar las medidas preventivas necesarias sobre aquellos que presentan mayor probabilidad de impago.
4. El modelo debe difundirse entre los empleados de LG que tengan responsabilidades en la aprobación de créditos y en aquellos que velen por el cumplimiento de los contratos vigentes, a fin de que enfoquen sus esfuerzos en los clientes de más riesgo.