Credit One EDA Analysis

El siguiente reporte presenta la información rescatada del EDA realizado sobre los datos reportados por Credit One. El siguiente será en forma de lista de manera que se vayan detallando todos los puntos analizados durante el estudio.

Primer análisis de los datos

- Observando las primeras filas de los datos destaca que las variables dependientes del análisis son: balance límite, sexo, nivel educativo, estado civil, edad, el tiempo en que se hicieron los últimos pagos (al día, atrasado por X meses, etc) y, los cobros y pagos de los últimos meses. Estas son las variables que esperamos que nos puedan ayudar a predecir si un cliente recurrirá a una falta de pago o no del crédito asignado. Por otro lado, la variable independiente es si el cliente incurre en una falta de pago o no.
- Al observar la descripción de los datos destaca que se tienen 30000 muestras para el análisis. Al tener 23 variables independientes se busca mínimo tener 10 veces la cantidad de muestras para poder tener un análisis de valor. Por lo tanto, se confirma que los datos entregados por Credit One se pueden utilizar para el estudio.
- Este mismo método para describir los datos nos permite ver que el balance límite del crédito ronda desde los 10 mil al millón de dólares. También cabe destacar que los datos reportan más casos en que los clientes cumplieron con sus pagos que de clientes que faltaron con los mismos. Esto último no quiere decir que sea algo positivo porque cualquier falta de pago de un solo cliente es importante para Credit One.
- Al observar los datos destacan varias variables categóricas como lo son: sexo, nivel educativo, estado civil, edad y el tiempo en que se hicieron los últimos pagos (al día, atrasado por X meses, etc). Esto debe ser tomado en consideración en las siguientes etapas del estudio.

Limpieza de datos

- Fue primordial primero eliminar cualquier fila de datos que tuviera variables NA. Esto es posible dado que la cantidad de observaciones es suficiente para el estudio, si se contara con una cantidad más limitada de datos entonces se debería utilizar un método que haga un cálculo de los datos NA y los reemplace.
- Los datos presentan una columna ID que no es relevante para el estudio y es eliminada.
- Al observar los datos de la columna de tiempo en que se realizaron los últimos pagos se observa que hay variables negativas (de -2 a 9). Al ser una columna categórica y al preferir trabajar solamente con valores positivos o cero, se toma la decisión de cambiar los valores -1 a 10 y -2 a 11 y recordar esta convención para el resto del estudio.
- En el atributo de nivel educativo es posible observar que hay varias variables (0, 4, 5 y 6) que significan “otro tipo de nivel educativo”. Al no tener información detallada sobre que es cada categoría se decide crear una sola categoría bajo el número 0 y reemplazar los otros 3 valores por este número. De esta forma se elimina complejidad del estudio.

Matriz de correlación

- Se plantea una matriz de correlación con los datos como se encuentran de forma que esta se salta las variables categóricas. A partir de esta se implementa un código que permite observar los 10 pares de variables con la correlación más alta. De aquí se puede observar que todas las variables de BILL_AMT* tienen alta correlación entre sí. Normalmente se decide eliminar las variables con correlación más alta pero primero debemos analizar los pares. En este caso es esperado que haya alta correlación entre las columnas dato que es un valor acumulativo en el tiempo y los meses. Por eso, se decide no eliminar ninguna variable y esperar a próximos métodos de feature engineering.
- Desea observarse también la matriz de correlación con todas las variables incluidas. Para realizar esto se deben crear dummy variables para las variables categóricas de forma que se crea una columna por cada categoría de cada variable y se elimina la columna original. A partir de esta matriz de correlación se implementa el mismo código utilizado anteriormente y se ven los 20 pares de correlación más altos. Nuevamente los pares indicados son categorías de los mismos atributos y por lo tanto se decide no eliminarlos.

Preparación de datos para el modelado

- Al no eliminar ningún atributo hasta el momento aparte de ID se decide dejar los datos preparados para el modelado. Para ello se deben convertir las variables categóricas utilizando la función de astype('category'). En este paso se convierten tanto las variables independientes como la variable dependiente.

Visualización de datos

A partir de los gráficos generados y los análisis estadísticos destacan las siguientes observaciones:

- De histogramas
    - Hay más mujeres en la muestra que hombres.
    - La mayoría de los clientes de Credit One están solteros, seguido por los casados.
    - La mayoría de los clientes se han graduado de un nivel académico. Los más comunes son aquellos con un bachillerato seguidos por aquellos que alcanzaron estudios avanzados.
    - Al analizar los pagos de los últimos meses lo más común es que los clientes paguen el crédito revolvente o que hayan pagado el costo total.
    - El rango de edades de los clientes va de 21 a 79 con una media de ~35. Es posible observar que la mayoría de los clientes tiene una edad entre los 25 y 40 años. Esto se puede tomar como un mercado meta y en proyectos futuros considerar otros elementos específicos para este rango de edades.
- Box plots
    - Se utiliza un box plot de las edades que refleja el mismo comportamiento que el observado en el histograma.
    - El box plot del Limit balance muestra que la mayoría de valores se encuentra bajo los 250 mil dólares.