José Calderón T.

Abstracto

Análisis de clasificación de los clientes que incumplen con los pagos de la tarjeta de crédito

Credit one

Clasificación de clientes

## Limpieza y depuración de datos

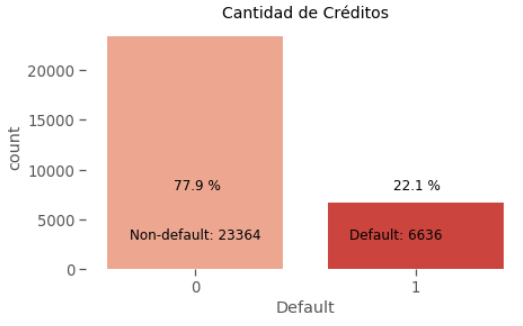
Se elimina la columna ID, ya que no es un atributo que pueda brindar información relevante en el análisis.

En el atributo EDUCATION se reemplazaron los valores 0, 5 y 6 con el valor 4, esto con el fin de dejar un único valor a la clasificación “Others”.

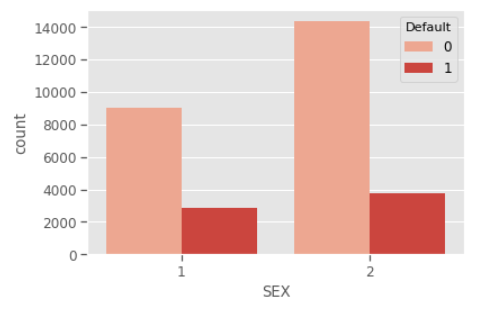
Se ajusta el nombre “default payment next month” a “Default” para que sea más sencillo de manipular en el proceso.

## EDA

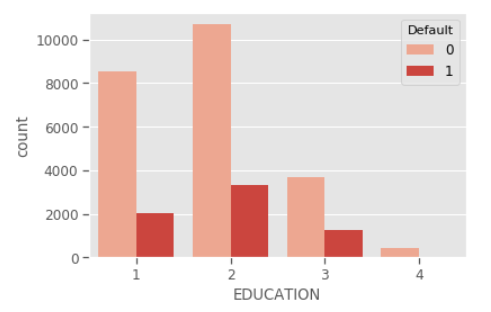
El dataset está conformado por 30 mil registros, de los cuáles un 22.1% (6636) representan a los clientes que han incumplido con el pago de la tarjeta de crédito.



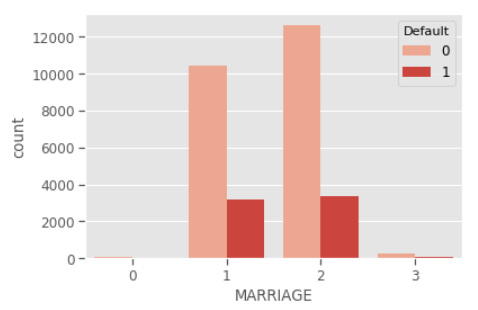
Al validar si el incumplimiento de pago está relacionado con el sexo del deudor, nos damos cuenta de que la variación es muy breve entre ambos sexos.



En otra validación de datos, se tomó como base el grado de educación del deudor. En este caso el gráfico nos muestra que los deudores que incumplen con su pago son mayoritariamente los que cuentan con un grado de escolaridad universitario (2), seguidos por los que finalizaron la escuela (1) y por último los que finalizaron el colegio (3).



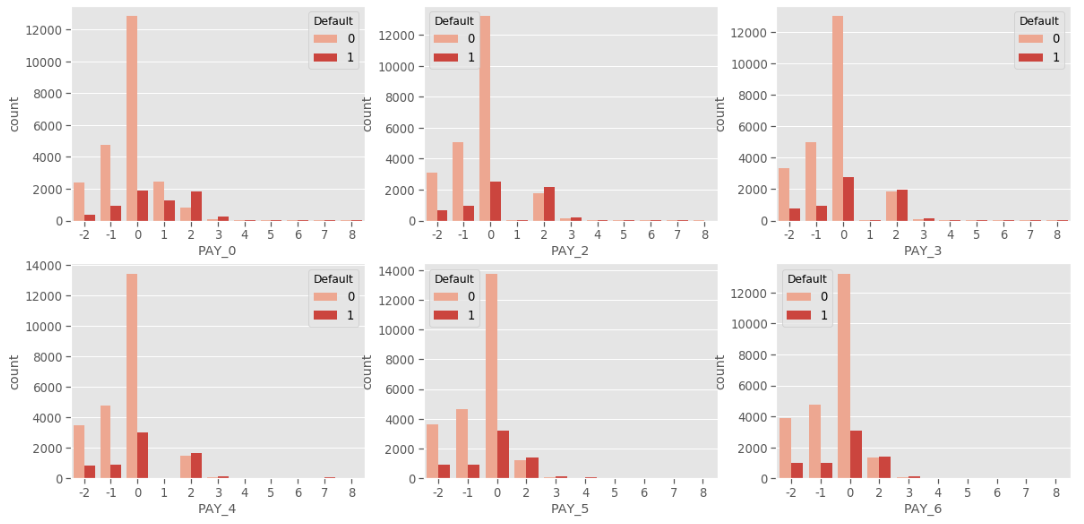
En el siguiente gráfico se muestra la comparación por estado civil del deudor, la cual no detalla mayor diferencia entre los casados (1) y los solteros (2)



Por último, se realizó una comparación entre el comportamiento de pago de los deudores durante 6 meses consecutivos.

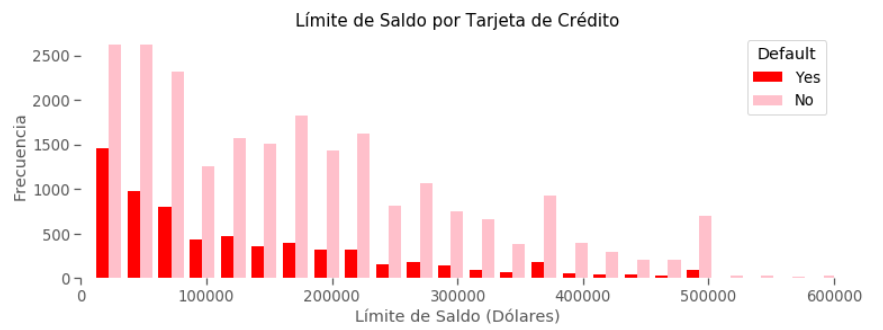
Al graficar los atrasos mensuales en los pagos, se pudo observar que había gráficos que se repetían, como se puede apreciar en los gráficos de los atributos PAY\_2, PAY\_3 y PAY\_4, así como los gráficos del PAY\_5 y PAY\_6.

El resultado fue el siguiente:



Mi teoría al revisar esa información es que, muy probablemente, los datos que corresponden a los meses de abril a agosto fueron obtenidos de manera errónea, por lo que en este caso sería bueno consultar con el cliente o el dueño del dataset, si la información suministrada es correcta o no.

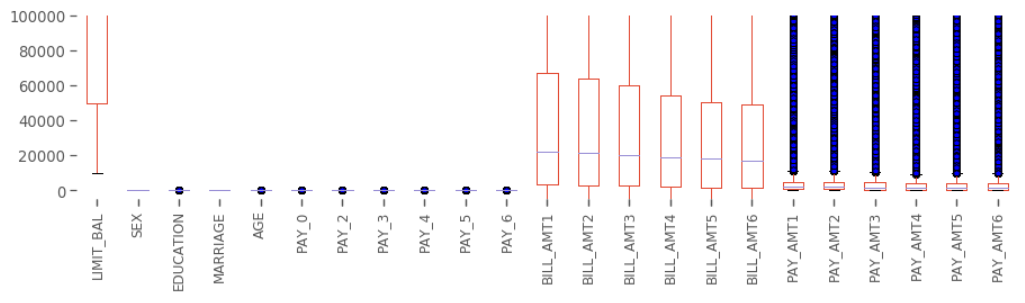
En el siguiente gráfico se puede observar el comportamiento de incumplimiento de los clientes según su límite de crédito aprobado.



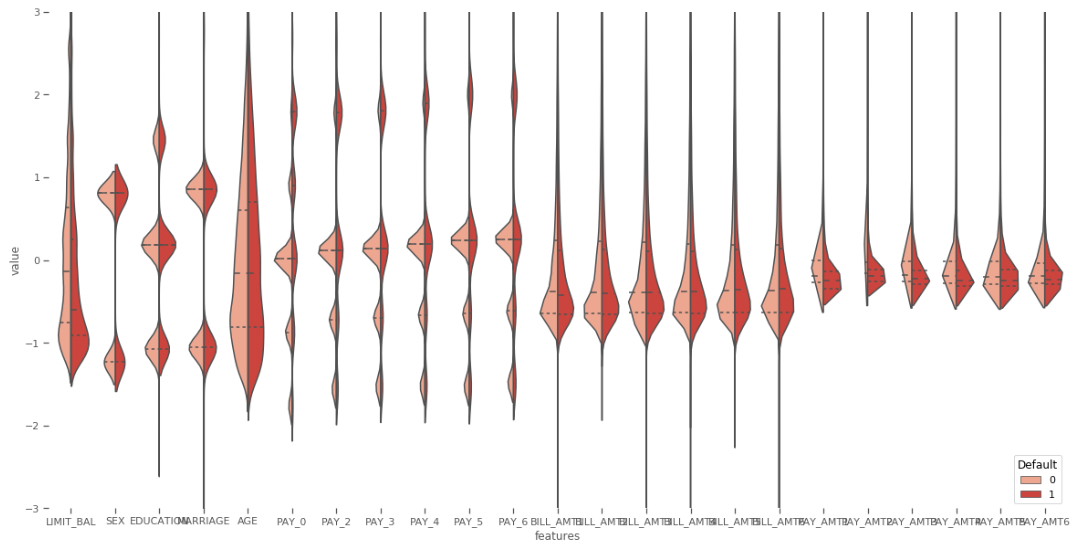
Se puede apreciar que las tarjetas de crédito cuyo límite es más bajo, son los que tienen una mayor incidencia con el incumplimiento de los pagos.

## RFE

Al revisar los datos como tal, se puede observar que a pesar de que los atributos son tipos de datos numéricos, los valores que llevan van desde -3 hasta 100.000, por lo que se debe de normalizar la data.

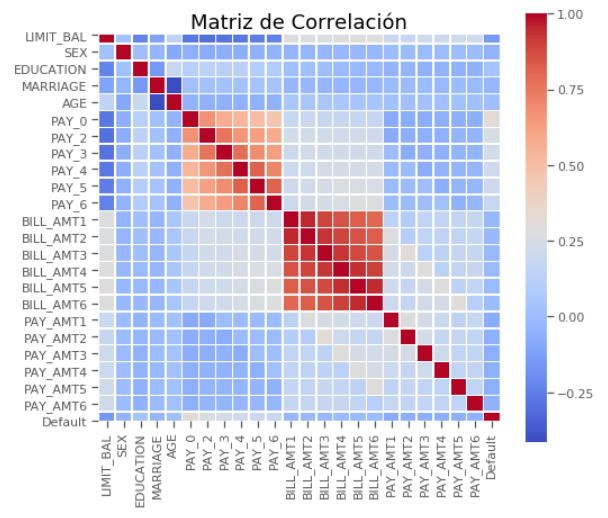


Posterior al proceso de normalización, los valores quedaron comprendidos entre un rango de -3 y 3, mostrándose la data de la siguiente manera:



## Matriz de Correlación

La matriz de correlación se muestra de la siguiente manera:



Se puede apreciar que la mayor correlación se encuentra en los atributos de pagos PAY\_0… PAY\_6 y los BILL\_AMT1… BILL\_AMT6.

## Modelos de Clasificación

Los modelos elegidos para realizar el proceso de clasificación son:

* Logistic Regression
* Decision Tree
* Random Forest

Al ejecutar los procesos los resultados fueron los siguientes:

|  |  |
| --- | --- |
| **Modelo** | **Accuracy** |
| Logistic Regression | 0.8078 |
| Decision Tree | 0.808 |
| Random Forest | 0.817 |

Entre los ajustes realizados para probar los modelos estuvieron:

* Modificar el Cross-Validation, pasándolo de 5 a 10, sin embargo, el resultado fue el mismo.
* Se utilizó la función RandomizedSearchCV para buscar los parámetros más óptimos para ejecutar los modelos

Por lo tanto, el modelo elegido fue el Random Forest, ya que ofreció mejores valores de precisión.

Preguntas para investigar:

1. ¿Cómo se asegura de que los clientes puedan pagar sus préstamos?

Existen muchas variables externas que pueden ocasionar que los deudores no puedan cumplir con sus obligaciones mensuales y pagar a tiempo. Sin embargo, con base en los datos obtenidos y en los análisis realizados, se puede utilizar el modelo entrenado para validar los clientes que pueden incurrir en un incumplimiento del pago o no.

1. ¿Podemos aprobar clientes con alta certeza?

Utilizando el modelo obtenido, se puede dar una certeza de un 81.7% para conocer si un cliente puede llegar a incumplir con el pago o no.