



|  |
| --- |
| **Universidad Cenfotec**  **Certificación XTOL Data Analytics & Big Data** |
| Yurguen Peñaranda Thomas  2 de Marzo del 2020 |
| Course 5: Data Science with Python  Task 2: Prepare and Explore the Data |
|  |





Contenido

[Introducción 2](#_Toc34070495)

[1. Resultados del Análisis Exploratorio de Datos (EDA) 2](#_Toc34070496)

[2. Lecciones aprendidas del EDA 3](#_Toc34070497)

# Introducción

Credit One es una autoridad de calificación crediticia que brinda servicios tercerizados a empresas para la aprobación de crédito de sus clientes minoristas. En el último año, se ha visto un incremento en los clientes minoristas que están incumpliendo en el pago de sus préstamos, lo cual está afectando a varios de las empresas a las cuales Credit One les brinda el servicio de evaluación de crédito. Esta situación debe de ser atacada de manera urgente, ya que pone en riesgo la reputación y por ende los negocios de Credit One con sus clientes.

Este documento es uno de los entregables del proyecto que está desarrollando el área de Data Science para generar un modelo que con base en ciertas características demográficas de un cliente, permita predecir si el próximo mes el cliente incumplirá o no en el pago de su préstamo. Lo anterior con el fin de que las empresas clientes de Credit One, puedan tomar acciones proactivas ante los posibles incumplimientos de sus clientes. Dichas estrategias pueden ser como por ejemplo, disminuir el monto máximo de crédito o incluso implementar estrategias más agresivas de cobro, entre otras.

Este entregable consiste en realizar un análisis exploratorio de los datos, con el fin de poder realizar labores de preprocesamiento y poder depurar el set de datos, y así reducirlos tiempos de procesamiento requeridos para el entrenamiento de los modelos y optimizar los niveles de calidad de predicción de los modelos.

# Resultados del Análisis Exploratorio de Datos (EDA)

Previo a la realización de los modelos de predicción de datos se procedió a realizar un Análisis Exploratorio de Datos (EDA), esto para depurar el set de datos a entrenar buscando así reducir los tiempos de procesamiento requeridos para el entrenamiento de los modelos y tratar de optimizar el nivel de predicción de los modelos.

Como se detalló en el entregable anterior, el set de datos incluye 30000 registros y 24 columnas, donde 23 columnas correspondes a los predictores y la variable a predecir seria el comportamiento del cliente, que se clasificaría como "Incumple" o "No Incumple“.

Primeramente se realizó un análisis del tipo de dato de cada variable y se identificó que todas las variables eran de tipo “int64” que es un tipo de variable para números enteros. Sin embargo, las variables SEX, EDUCATION, MARRIAGE, PAY\_0, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6 y default payment next month realmente eran categorías, por lo que se procedió a cambiar el tipo de estas variables a “category”.

Luego se removió la variable ID, ya que se considera que esta variable no agrega valor para el análisis ni el entrenamiento de los datos, sino que solo es el identificador único de cada registro. También, se cambió el nombre de la columna “default payment next month” por “DefaultPaymentNextMonth” pues se considera más sencillo que las variables tengan nombres sin espacios.

A las variables SEX, EDUCATION, MARRIAGE y DefaultPaymentNextMonth se les sustituyó los valores numéricos por la etiqueta que correspondía (por ejemplo a la variable SEX se le sustituyó el 1 por “Masculino” y el 2 por “Femenino”), esto con el fin de facilitar la visualización de los gráficos.

Al hacer un gráfico que resuma de manera general la cantidad de clientes que incumplieron vs los que no incumplieron se puede ver que solo 6636 incumplieron contra 23364 que no incumplieron en el pago.

Al hacer un análisis con gráficos boxplot se puede ver que las variables de edad, límite de crédito, estado de cuenta mensual y pago mensual presentan un comportamiento asimétrico en sus datos, pues los gráficos no están centrados en el eje vertical si no que están más cerca del límite inferior, y en el límite superior se muestran datos atípicos. Esto quiere decir que para estas variables, la mayoría concentración de los datos es en valores bajos, sin embargo hay algunos valores altos. Se graficó por aparte los datos que Incumplieron y los que No Incumplieron, y se pudo ver que para las variables analizadas, ambos grupos de datos tienen un comportamiento similar en estos gráficos. Los datos atípicos no se pueden eliminar, pues son parte del comportamiento real del set de datos. Al tener un comportamiento asimétrico, se debe analizar la mediana de los datos en vez de la media.

A nivel de estado civil, el volumen de incumplimiento entre casados y solteros es bastante similar, mientras que el volumen de incumplimiento de divorciados es muy bajo, pero se debe a que la proporción general de divorciados es muy bajo y la proporción de clientes en otro estado civil es prácticamente despreciable. También se puede ver que a nivel general hay más clientes solteros que casados.

A nivel de género, el volumen de incumplimiento entre mujeres es ligeramente mayor al de los hombres. También se puede ver que a nivel general hay más clientes femeninos que masculinos.

A nivel de grado académico, el volumen de incumplimiento es mayor en los de Universidad, disminuye para los de Postgrado y más bajo para los de Secundaria. También se puede ver que a nivel general hay una mayor proporción de clientes de Universidad y que la proporción de clientes con otro grado académico es prácticamente despreciable.

A nivel de límite de crédito otorgado, se puede ver que está entre 10 mil y 1 millón de NTD (nuevo dólar taiwanés). Al hacer una agrupación en 10 grupos por esta variable, se puede ver un comportamiento descendente en la frecuencia de incumplimiento de cada clase, ósea entre mayor es el límite de crédito, hay menor nivel de incumplimiento. La mediana de los datos es de 140 mil NTD.

A nivel de edad, se puede ver que está entre 21 y 79 años. Al hacer una agrupación en 6 grupos por esta variable, se puede ver un comportamiento constante en la frecuencia de incumplimiento cada clase. La mediana de los datos es de 34 años.

Al hacer un análisis de correlación, se identificó una correlación casi perfecta entre el seguimiento conjunto de variables: BILL\_AMT1 y BILL\_AMT2; BILL\_AMT2 y BILL\_AMT3; BILL\_AMT3 y BILL\_AMT4; BILL\_AMT4; y BILL\_AMT5; BILL\_AMT5 y BILL\_AMT6. Estas variables representan el estado de cuenta mensual (cuanto debe de pagar el cliente cada mes). Por lo tanto, se puede inferir que hay una correlacion entre el estado de cuenta de un mes y el del mes siguiente. Debido a que tener dos variables que estén fuertemente correlacionadas no agrega valor para el proceso de entrenamiento del set de datos, se procederá a eliminar las variables BILL\_AMT2, BILL\_AMT4 y BILL\_AMT6, de esta manera se eliminará la correlación.

# Lecciones aprendidas del EDA

De acuerdo a esta experiencia, se puede concluir que no existe una "receta de cocina" para la aplicación de algoritmos de Machine Learning para la predicción de datos. Ya que aun cuando se tengan 2 set de datos con las mismas variables y enfocados a estudiar el mismo problema, cada set de datos tiene un comportamiento diferente en cuanto a la cantidad y distribución de datos.

Por lo tanto, es importante invertir el tiempo que sea necesario en "poder entender los datos", estos ayudará a poder contextualizar mejor el problema y poder dirigir el abordaje adecuado para el mismo, al mismo tiempo que permite seleccionar cuales son las variables que realmente agregan valor al problema en específico que se esté analizando, disminuyendo así los tiempos de procesamiento requeridos (los cuales dependiendo del tamaño del set de datos, pueden llegar a ser de varias horas o incluso días) y buscando optimizar los resultados de predicción de los modelos.

El Análisis Exploratorio de Datos (EDA) no es algo que se pueda realizar de manera mecánica, ya que existen gran cantidad de herramientas estadísticas que se puedan utilizar en este tipo de análisis. Sin embargo, se puede empezar por hacer un análisis de las estadísticas básicas del set de datos, ver la media, desviación estándar, mediana, máximo y mínimo. Esto permitirá tener una idea general del nivel de dispersión de los datos. Para complementar este tipo de análisis se pueden realizar gráficos boxplot, que permite analizar de una manera visual si los datos están centrados o no y si tienen datos atípicos o no (los cuales dependiendo del caso en análisis, se podría considerar en eliminarlos o no).

También se recomienda analizar cuál es el tipo de variable que tiene cada atributo (numérico o categoría), ya que puede que el ambiente de programación esté interpretando un atributo de tipo numérico cuando realmente sea una categoría y si no se corrige, estará induciendo cierto sesgo al proceso de entrenamiento de los set de datos.

Para variables numéricas, se recomienda clasificarlas en ciertos grupos, lo cual facilita el análisis de la distribución de esta variable, en vez de tener que analizar la distribución de cada uno de los posibles valores por separado.

Para variables categóricas, es importante tener claro que significa cada uno de los posibles valores (por ejemplo; que 1 signifique masculino y 2 femenino) y si los valores actuales no son claros, se puede considerar sustituir por etiquetas de datos que sean más claros para el analista.

Otro buena técnica, es el análisis de correlación que ayuda en la disminución del número de variables, ya que cuando 2 variables tienen una fuerte correlación (muy cercana a 1) quiere decir que ambas dan el mismo aporte al proceso de entrenamiento, por lo que si se elimina una de las 2, no se verá afectada la calidad de predicción pero si se optimizará el tiempo de procesamiento requerido.

Existen también, otras técnicas para la visualización de datos, como: gráficos de pastel, histogramas, gráficos de serie de tiempo, etc. Y otras técnicas de selección de variables, como PCA (Análisis de componentes principales) o RFE (Recursive Feature Elimination) que son útiles cuando se tiene un set de datos con muchas variables y se quiere utilizar una herramienta más estructurada para la selección de las variables importantes.

Sin embargo, aun cuando todas estas herramientas (y muchas más) están disponibles, lo más importante es que el analista las utilice de manera lógica y con un objetivo, utilizando su criterio de experto para ir generando conclusiones con base en las herramientas utilizadas. Dicho criterio de experto se va afinando conforme el analista va adquiriendo más y más experiencia en el análisis de datos.