



|  |
| --- |
| **Universidad Cenfotec**  **Certificación XTOL Data Analytics & Big Data** |
| Yurguen Peñaranda Thomas  16 de marzo del 2020 |
| Course 5: Data Science with Python  Task 3: Build and Evaluate Models |
|  |





Contenido

[Introducción 2](#_Toc34938123)

[1. Análisis Inicial de los datos: 2](#_Toc34938124)

[2. Preprocesamiento de los datos: 2](#_Toc34938125)

[2.1. Eliminación de columna de “ID” 2](#_Toc34938126)

[2.2. Cambio de tipo de dato de las variables 2](#_Toc34938127)

[2.3. Análisis de correlación 2](#_Toc34938128)

[2.4. Set de datos final a entrenar: 3](#_Toc34938129)

[3. Labores generales para el Entrenamiento de los Modelos: 4](#_Toc34938130)

[3.1. Valor de aleatorización: 4](#_Toc34938131)

[3.2. Algoritmos a utilizar: 5](#_Toc34938132)

[3.3. Métricas de evaluación de la calidad de predicción de los modelos: 5](#_Toc34938133)

[3.4. Partición de los datos en el Set de Train y Test: 5](#_Toc34938134)

[3.5. Parámetros utilizados en los modelos: 5](#_Toc34938135)

[4. Entrenamiento y optimización de los Modelo de Predicción 6](#_Toc34938136)

[5. Selección del modelo de predicción a utilizar: 6](#_Toc34938137)

[5.1. Afinamiento del modelo 6](#_Toc34938138)

[6. Resultados y análisis de las predicciones 6](#_Toc34938139)

# 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.

En el documento se detallan todos los pasos realizados para generar el modelo de predicción: análisis inicial y preprocesamiento de los datos, definición de parámetros y selección de los algoritmos a utilizar, entrenamiento y validación de los resultados de predicción en el set de test, selección del mejor algoritmo y análisis de resultados.

# Análisis Inicial de los datos:

Se tiene un set de datos cuyo objetivo es investigar si los clientes financieros de Taiwán llegan a incumplir en las obligaciones de pago de sus préstamos bancarios o no (default payment).

El set de datos incluye 30000 registros y 25 columnas, donde 24 columnas correspondes a los predictores y la variable a predecir seria el comportamiento del cliente, que se clasificaría como "Incumple" o "No Incumple“.

# Preprocesamiento de los datos:

Con el fin disminuir los tiempos de procesamiento de entrenamiento de los modelos y buscar optimizar los resultados de predicción se realizaron diversas tareas de preprocesamiento. A continuación se detallan estas tareas realizadas.

## Eliminación de columna de “ID”

El set de datos incluía una columna del **ID** del registro. Esta columna no agrega valor para el análisis en cuestión, por lo tanto se procede a eliminarla.

## Cambio de tipo de dato de las variables

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**”.

## Análisis de correlación

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 correlación 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.

## Set de datos final a entrenar:

Con las labores de preprocesamiento realizadas anteriormente, el dataset a entrenar de los clientes está conformado por 30000 registros y 21 variables, que se detallan a continuación.

Figura 1 Detalle del set de datos final a entrenar

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Variable** | **Descripción** | **Posible Valor** | **Tipo de dato final** | **Uso** |
| 1 | LIMIT\_BAL | Monto de crédito otorgado | Monto en NTD (NT dollar - nuevo dólar taiwanés) | int64 (Número entero) | Variable de entrada (Predictor) |
| 2 | SEX | Género | 1 – Masculino  2 - Femenino | category (Categoría) | Variable de entrada (Predictor) |
| 3 | EDUCATION | Nivel educativo | 1 – Postgrado  2 – Universidad  3 – Secundaria  0, 4, 5 6 - Otros | category (Categoría) | Variable de entrada (Predictor) |
| 4 | MARRIAGE | Estado civil | 1 – Casado  2 – Soltero  3 – Divorciado  0 - Otros | category (Categoría) | Variable de entrada (Predictor) |
| 5 | AGE | Edad | Edad (Años) | int64 (Número entero) | Variable de entrada (Predictor) |
| 6-11 | PAY\_0  PAY\_2  PAY\_3  PAY\_4  PAY\_5  PAY\_6 | Historial del status de pagos anteriores:   * PAY\_0: Status de pago realizado en septiembre 2005 * PAY\_2: Status de pago realizado en agosto 2005 * PAY\_3: Status de pago realizado en julio 2005 * PAY\_4: Status de pago realizado en junio 2005 * PAY\_5: Status de pago realizado en mayo 2005 * PAY\_6: Status de pago realizado en abril 2005 | -2 - Sin consume  -1 - Pago complete  0 - Usó crédito rotativo (Tarjeta de crédito)  1 - Pago atrasado en 1 mes  2 - Pago atrasado en 2 meses  Así sucesivamente…  n - Pago atrasado en n meses | category (Categoría) | Variable de entrada (Predictor) |
| 12-14 | BILL\_AMT1  BILL\_AMT3  BILL\_AMT5 | Estado de cuenta mensual:   * BILL\_AMT1: Estado de cuenta en septiembre 2005 * BILL\_AMT3: Estado de cuenta en julio 2005 * BILL\_AMT5: Estado de cuenta en mayo 2005 | Monto del estado de cuenta mensual en NTD (NT dollar - nuevo dólar taiwanés) | int64 (Número entero) | Variable de entrada (Predictor) |
| 15-20 | PAY\_AMT1  PAY\_AMT2  PAY\_AMT3  PAY\_AMT4  PAY\_AMT5  PAY\_AMT6 | Monto de pagos anteriores:   * PAY\_AMT1: Monto pagado en septiembre 2005 * PAY\_AMT2: Monto pagado en agosto 2005 * PAY\_AMT3: Monto pagado en julio 2005 * PAY\_AMT4: Monto pagado en junio 2005 * PAY\_AMT5: Monto pagado en mayo 2005 * PAY\_AMT6: Monto pagado en abril 2005 | Monto del pago anterior mensual en NTD (NT dollar - nuevo dólar taiwanés) | int64 (Número entero) | Variable de entrada (Predictor) |
| 21 | DefaultPaymentNextMonth | Comportamiento del cliente | 0 - No Incumple  1 - Incumple | category (Categoría) | Variable de salida (Variable a predecir) |

# Labores generales para el Entrenamiento de los Modelos:

## Valor de aleatorización:

Se utiliza un valor de aleatorización (random\_state) de 42. Este valor permite que todas las particiones de datos que se realizan de manera aleatoria, siempre se dividan de la misma manera, lo cual permite la reproducibilidad de los resultados.

## Algoritmos a utilizar:

Para la variable que se requiere predecir (**DefaultPaymentNextMonth**) se entrenarán varios algoritmos de clasificación para evaluar cual genera resultados con mejor calidad de predicción. Los 6 algoritmos que se evaluarán en presentan en la sección 4, junto a los resultados de Accuracy para el set de Test.

## Métricas de evaluación de la calidad de predicción de los modelos:

Para poder cuantificar el desempeño de la calidad de predicción de los modelos se utiliza la siguiente métrica de calidad:

* **Accuracy (Exactitud):** Cantidad de registros correctamente clasificados con respecto al total de registros.

El valor máximo de la **Accuracy** es de 1, que quiere decir que la totalidad de registros fueron clasificados correctamente. Por lo tanto, entre más alto es el valor de **Accuracy**, quiere decir que mejor es la calidad de predicción del modelo.

* **Kappa**: Tiene en cuenta la Exactitud que se generaría simplemente por casualidad.

Donde:

* + O -> Exactitud esperada
  + E -> Exactitud marginal basada en los totales de la matriz de confusión

El **Kappa** puede tomar valores entre -1 y 1; un valor de 0 significa que no hay acuerdo entre las clases observadas y predichas, mientras que un valor de 1 indica una concordancia perfecta de la predicción del modelo y las clases observadas. Los valores negativos indican que la predicción está en la dirección opuesta a la verdad, pero los valores negativos grandes rara vez ocurren, si es que se producen.

## Partición de los datos en el Set de Train y Test:

El dataset a entrenar se dividió en un subset de Train y uno de Test. A continuación se detallan los porcentajes y cantidad de registros para cada subset.

Figura 2 Detalle de la partición de los datos en los sets de Training y Testing

|  |  |  |
| --- | --- | --- |
| **Set** | **Porcentaje de datos** | **Cantidad de registros del dataset** |
| **Training** | 80% | 24000 |
| **Testing** | 30% | 6000 |
| **TOTAL** | **100%** | **30000** |

## Parámetros utilizados en los modelos:

Los modelos se entrenan de modo que el algoritmo automáticamente seleccione los valores óptimos para generar el modelo de predicción.

# Entrenamiento y optimización de los Modelo de Predicción

Una vez realizado el preprocesamiento y la partición de los datos según lo indicado anteriormente, se procede a entrenar el set de datos de entrenamiento (**train**) con los 6 algoritmos indicados anteriormente en la variable que se requieren predecir en el dataset**. A continuación el análisis detallado de los resultados de los parámetros utilizados en el entrenamiento y los resultados de predicción en el set de prueba (test) para cada algoritmo.**

Figura 3 Resultados de Accuracy y Kappa en el set de Test para los 6 modelos entrenados

|  |  |  |
| --- | --- | --- |
| **Modelo** | **Accuracy** | **Kappa** |
| LR: Logistic Regression | 0,7812 | 0,0000 |
| RF: Random Forest Classifier | 0,8155 | 0,3623 |
| SVC: Support Vector Classifier | 0,7812 | 0,0000 |
| GBC: Gradient Boosting Classifier | 0,8207 | 0,3693 |
| KNC: K Neighbors Classifier | 0,7498 | 0,1040 |
| GNB: Gaussian Naïve Bayes | 0,3802 | 0,0650 |

# Selección del modelo de predicción a utilizar:

Con base en los resultados presentados anteriormente, se puede ver que el modelo que obtuvo mejores resultados de **Accuracy** y **Kappa** es el generado con el algoritmo GBC (Gradient Boosting Classifier), por lo tanto este será el algoritmo a utilizar para generar las predicciones.

## Afinamiento del modelo

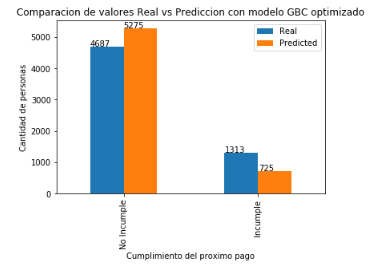
Con el fin de buscar optimizar los niveles de predicción del modelo, se procede a realizar diversas iteraciones de entrenamiento variando los valores de los siguientes parámetros del algoritmo GBC: n estimators, learning rate, max features, max depth. A continuación se presentan los valores de los parámetros optimizados y los resultados de los indicadores de calidad de predicción.





# Resultados y análisis de las predicciones

Figura 4 Gráfico de comparación de valores Reales vs Predichos con el modelo GBC optimizado



Con el gráfico anterior y basándose en el resultado de Accuracy se podría decir que el modelo tiene un buen nivel de predicción, pues logró predecir correctamente el 82% de los datos en el set de Test.

Sin embargo, para este problema en particular lo realmente no es el nivel de predicción general, sino más bien la capacidad de predicción en cada registro en particular, para que a la hora de que se está haciendo el estudio de aprobación de crédito de un cliente, tomar sus características socioeconómicas para poder predecir si el cliente llegará a predecir o no en el pago de su préstamo y tomar en cuenta este criterio para ver si se le aprueba o no el crédito. Para analizar la calidad de predicción individual del modelo es mejor tomar en cuenta el resultado del indicador Kappa.

Con el modelo optimizado, el resultado del indicador Kappa es de 0,378. Esto quiere decir que aunque el modelo tiene una buena capacidad de predicción general, su capacidad de predicción a nivel individual es muy deficiente, pues solo en el 37% de los clientes se predijo correctamente. Por lo que basándose en este modelo, se podría aprobar el crédito a muchos clientes que el modelo haya dicho que iban a cumplir, cuando realmente el cliente iba a incumplir.

Debido a lo anterior, se recomienda utilizar algoritmos de predicción más complejos como Neural Networks, las cuales requieren un nivel de parametrización y procesamiento mayor, y por ende requieren mayor conocimiento y experiencia, pero permiten generar resultados con un nivel de predicción más alto.