MACHINE LEARNING FOR BUSINESS INTELLIGENCE: TAREA 1

Joaquín Cárdenas Liebethal

7 de Mayo 2018

1. Limpieza y Preparación de los Datos

Primero se le hecho una mirada a los datos para saber con que se estaba tratando, cantidad de columnas y sus valores, si esque acaso existian valores nulos o mal formateados. Posteriormente se pasaron las variables categoricas a vectores de 0 o 1, indicando si la fila se entontraba o no en el atributo. Los atributos aumentaron de 18 columnas a 66, con 1379 filas.

2. Problema 1

2.1. Objetivo

Para este problema se requeria estimar la variable Y1, usando como modelo una regresión lineal. Ademas se solicito ajustar parametros para alcanzar el mejor rendimiento del modelo.

2.2. Metodologia

- Se definio un arreglo y que contenia las variables de objetivo entregadas por la columna Y1 del dataset.
- Se uso el resto de los atributos como predictores de Y1, menos Y2, YTOT, YTOT2. Por ultimo se agregó otra columna de puros 1.
- Se utilizo el modelo Ridge de sklearn junto con la tecnica de GridSearch para encontrar el mejor modelo.

2.3. Resultados

Para evaluar el modelo se utilizo la metrica de "Mean Square Error"(MSE). El modelo arrojo el sgte valor: 663482294919.599. El valor es altisimo y escapa de la racionalidad del problema y su variable objetivo que era Y1. Se asume una incorrecta implementación del algoritmo.

3. Problema 2

3.1. Objetivo

Para este problema se requeria determinar que personas poseen ingresos menores a $200.000, usando como modelo una regresión logistica. Ademas se solicito mostrar los resultados en una matriz de confusión y utilizar el f1-score como metrica de evaluación del modelo.

3.2. Metodologia

- Como se tuvo que usar Logistic Regression como modelo, se creo una nueva columna 'sub200' como objetivo. Esta nueva columna/atributo indica si la persona tiene un ingreso menor a $200.000 como 1 o 0 de lo contrario.
- Para el entrenamiento del modelo, se usaron los mismos predictores y como objetivo esta nueva columna 'sub200'.
- Se separo los datos en datos de entrenamiento y testeo usando los metodos de la libreria de sklearn train_test_split(), con un tamaño de datos de entrenamiento del 25 %.

3.3. Resultados

A continuación el FScore para las etiqueas 1 y 0. En la Figura 1 se muestra la matriz de confusión. Se utilizaron 1034 datos de entrenamiento y 345 datos para el testeo. De los 345, 30 personas tenian un ingreso real menor a $200.000 de los cuales el clasificador predijo 4 correctamente.

- FScore de 0.95398773 0.21052632 para 1 y 0 respectivamente

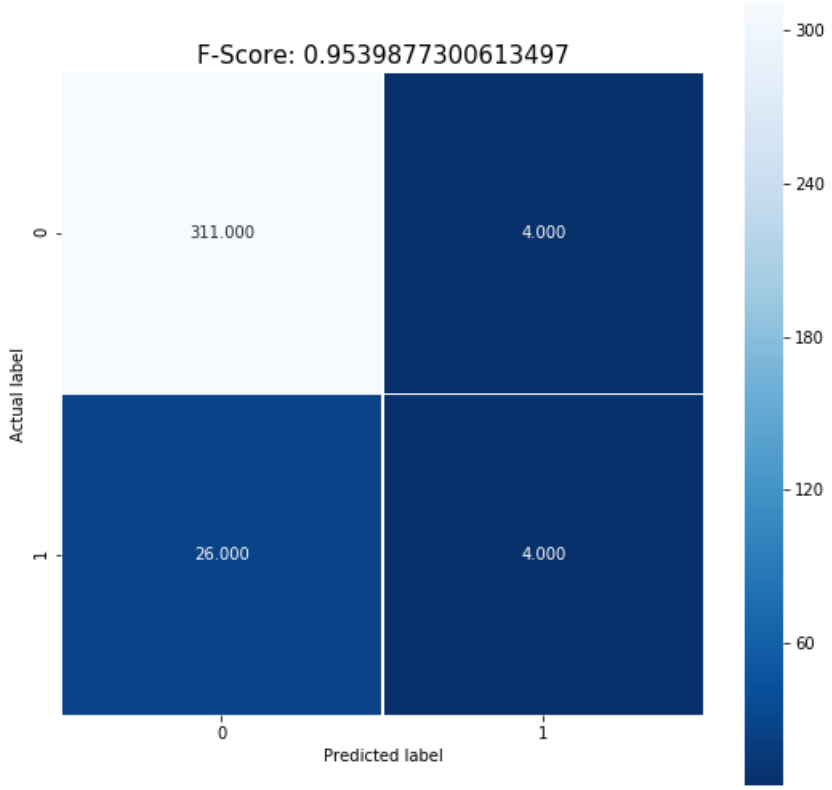

Figura 1: Matriz de confusion de los datos de prueba