PRIMERA ENTREGA

PROYECTO ANALITICA DE DATOS

POR

AURA LUZ MORENO DÍAZ

JUAN JOSE MOLINA OCAMPO

UNIVERSIDAD DE ANTIOQUIA

2022

1. PROBLEMA PREDICTIVO A RESOLVER

El banco Santander de origen español desea predecir la cantidad de clientes que podrían realizar una transacción conociendo que la cantidad monetaria no posee importancia; dichos resultados se presentarán en datos binarios, donde 1 es el resultado de una transacción exitosa y 0 el cliente no realizó una transacción sin importar el monto de la transacción realizada. Los resultados de este análisis permitirán obtener mucha información para resolver otros problemas como pueden serlo: cantidad de colaboradores para atender la demanda, dinero disponible, espacio de las instalaciones, satisfacción del cliente… entre otros aspectos relativos a la realización de una transacción.

Se puede revisar la competencia aqui:
https://www.kaggle.com/c/santander-customer-transaction-prediction

2. DATASET

El dataset proviene de una competencia en Kaggle en el cual se proveen datos sobre las posibles transacciones de los usuarios sin importar el monto.

El dataset esta compuesto por un conjunto de archivos .csv que proporcionan la información requerida con los individuos y su probabilidad de elección.

Existe un primer archivo llamado train.csv que tiene el set de datos.
ID_code
#Target
#var_0 hasta la 200

Otro de los archivos es nombrado como test.csv que contiene los datos de prueba y con algunos datos que no están incluidos en el scoring.
ID_code
#var_0 hasta la 200

El archivo que tiene los datos organizados es sample_submission.csv
target_ID es binario
ID_Code es un string
#Var que va de 0 a 200 para el cálculo de probabilidades del modelo entrenado

Se deberá predecir el valor de la columna target

3. **METRICAS DE DESEMPEÑO**

https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/#

Presentemos primero las principales métricas de clasificación usadas en Machine Learning:

- **Matriz de confusión o error**
- **Precisión**
- **Recall o sensibilidad o TPR (Tasa positiva real)**
- **Precisión**
- **Especificidad o TNR (Tasa negativa real)**
- **F1-Score**
- **Área bajo la curva de funcionamiento del receptor (ROC) (AUC)**
- **Pérdida logarítmica**
- **Cohen's Kappa**

Las métricas del negocio es saber con cuanta probabilidad un usuario hará una transacción nuevamente sin importar el monto. Con esta información podrán realizar análisis financieros para ofrecer otro tipo de productos a los usuarios según sus elecciones.

4. DESEMPEÑO

Lo que se esperaría saber con este challenge es identificar cuáles clientes realizarían una transacción específica en el futuro, sin importar el monto de la transacción. De esta manera se podrían ofrecer productos y servicios específicos para cada tipo de cliente.