DATA MINING

ASSIGNMENT-02

Course Instructor : Sir Zain Mirza

LAIBA SHEIKH

20B-051-SE

1) Which dataset you've selected?

I have selected **E-commerce Price Prediction** (which will accurately predict the price the price of product based on the given factors) to perform .

2) What analysis you've done in the starter code?

I have imported the required libraries that are :

```
# Basic libraries

import numpy as np        # linear algebra
import pandas as pd       # data processing, CSV file I/O (e.g. pd.read_csv)
import os
```

```
# Plot related libraries

import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Linear Regression Model

from sklearn.linear_model import LinearRegression, RidgeCV
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.compose import TransformedTargetRegressor
from sklearn.utils import shuffle
```

After importing I've loaded the dataset into a variable called train_ecomm_df and test_ecomm_df . After that I looked the summary of the dataset and examined the data type of each column in my Dataset.
Then I performed my Preprocessing Part. First, I examined the number of rows and columns. To manage noisy data in the data set I have dropped the rows that have any Null Values now time to start EDA (Exploratory Data Analysis). The Analysis done by me contains:

1) I have check the that is data contain any null value or not because the training set seems to have no null data .
2) After that , prepare the data for model building . In which I do the few following things :
   2.1) Merge train and test data
   2.2) Impute the unknown values with mode.
   2.3) Get the categorial columns .
   2.4) Get back the Tran and test data .
3) After this , I have define the X and Y

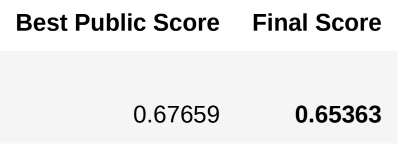
4) Build Linear Regression Model , using transformed target regressor model.

5) Ridge CV implementation (Initialize Linear Regression algorithm with Ridge regularizer (K-fold with 10 folds)).

3) What information you got?

The information I have derived from my Analysis is given further below:

- The data is a mix of categorical, ordinal, numeric and date values
- The **Y-Target** attribute **Selling Price** has got a skewed data when we visualize its distribution
- We need to apply the transformation method to make it normal.
  Here, **np.log1p** method is used
- It is always good to start with linear model rather than ensembles or neural network.
- The indention was to get exposure to real time data not the leaderboard (pun indented)
- First tried with LinearRegressor model with RidgeCV .
- During the iteration, applied the data with QuantileTransformer of 300 estimators but the result was not converging towards 0.5, hence switched to Log transformer.
- The final submission score is as follows :



Best Public Score	Final Score
0.67659	**0.65363**

- These scores stood **38th** position. The challenge was quite tough, solely because of the data.
- Although the feature scaling and engineering parts were not done extensively here, the **Linear Regressor** with RidgeCV seemed to have done pretty good job.