# Understanding Dataset with the help of EDA

1. Import the dataset after importing required libraries
2. Now we can observe there are 3 continuous columns and 4 discrete columns.
3. From 7 columns Purchased\_ABC\_product is dependent column and remaining are independent columns.
4. We can observe that there are no NA values in the dataset.
5. Now we start the visualization of data so that we can know how the data is distributed.
6. For continuous columns like Var1and Var2 i used a distribution plot and i neglected customer\_id column because it has no importance.

* For Var1 column data is mostly distributed between 0 to 500
* For Var1 column data is mostly distributed between 0 to 50

1. Now I used a bar plot for the Group column because the pie is overlapping. From plot we can say that data Group feature has most values as G1 and G5
2. Now for the next 3 categories I used pie charts.

* For Category column data contains most values as C1
* For Ratings Column data contains most values as Bronze
* For Purchased\_ABC\_product Column data is equally distributed between classes.

# Model Building and Evaluation

1. Firstly before model building i have converted categorical values into numerical using label encoding.
2. Now I have selected logistic regression which gives the predicted values in the form of probabilities.
3. Now I have added these probabilities to the test data as asked.
4. Now I have evaluated the model using a Confusion matrix and ROC AUC test. In confusion matrix the model achieved

True Positive = 3836

True Negative = 3100

False Positive = 1557

False Negative = 756

Accuracy = 75%

# Finding out the variable Influencing the output

1. First I used correlation table to know how the features are related. Then I found Var1 and Var2 highly correlated. These are input features so i have dropped one feature and again checked and i found out Var2 is most influencing on Output.
2. From the filter method using information gain also i have found out Var2 is influencing feature.
3. SHAP values are calculated for each feature, for each value present, and approximate the contribution towards the output given by that data point. It is used to know the feature which shows influence on output. From using this method also the result is the same as the above two methods.

# Identifying the group of customers to increase sales

G1 9230

G2 98

G3 1480

G4 228

G5 4336

G6 135

G7 12

G8 5

From above result we can say that G1 group is highly favorable to our products but G2,G4,G6,G7 and G8 are not so keen to our products so we should contact G3 and G5 groups to increase our sales ao they also reach sales similar to G1 and by seeing this the remaining unfavourable groups may consider using our product.

# 