SUMMARY :

An education company named **X Education** sells online courses to industry professionals. When these people fill up a form providing their email address or phone number, they are classified to be a lead.The typical lead conversion rate in course at X education is around 30%. To make this process more efficient, the company wishes to identify the most potential leads.So, we need to build model who can give the conversion rate around 80%.

Importing dataset to the python file and also basic python libraries numpy,pandas,matplot and seaborn.Then we convert the columns which Have '**Yes** & **No'** to '**1** & **0** '. Then we check the '**Null Values'** present in variables in percentage and those variable whi has null values **above 35%** we drop the column.

Then we check the distribution of datase into the **variables** among the **categories** and the **imbalance category** which have values **saturated** in **one category** and in ' **Select** ' category. We will drop those variables.

After that we will impute the null values, but first we will **drop** the null values **rows** from dataset variables column which have less than **2%.** Then we will **impute** the variables **null values** with variables **mode value** or **assign special category**.

Now , we have clean dataset. So, we can do the **EDA** on dataset. First, we will do **the bivariate analysis** with **categorical variables** by using **countplot** from **seaborn**.Then we will perform **univariate analysis** on **continuous variables.** Here, on **univariate analysis** we can see there is **outliers** are present, which can affect on model evaluation and confusion matrix. So, we will **remove** the outliers using **quantiles**.

We will start model building.We will create the dummy variables for categorical variables.Then split the dataset between train datase and test dataset.Then we will scale the continous variable by using StandardScaler().Then we will check the correlation between variables and **drop the high correlation variables**.Then we will use **GLM method** for model building.But the variables are high in number due to dummy variables. So, we will use automatic method for elimination using **RFE selection** method and get **top 15 variables**.

After this we remove the feature by **manual elimination** on basis **P-value** and **VIF in iterations.** Then we will create **confusion matrix** and from that we will calculate **sensitivity,specificity,precision,recall** and **F1-score.**

Now we will plot the **ROC curve** and calculate the cut-offs using **different probabilities** and also **acuuracy,sensitivity and specificity table** and also plot their **graph**.From that we will choose the **proper threshold** which is **0.35** .and again **calculate confusion matrix, sensitivity, specificity, precision,recall and F1-score**. Also **use same threshold** for **test set** and calculate the above **parameters** again.

Parameters	Train Dataset	Test Dataset
Accuracy	81.32	82.25
Sensitivity	80.78	82.18
Specificity	81.65	82.29
Precision	73.30	72.24
Recall	80.78	82.18
F1-score	0.76	0.77

Conclusion on Model :

Top Positive Correlation variables for customer conversion :

1. Lead Origin_Lead Add Form
2. Lead Source_Welingak Website
3. Last Activity_SMS Sent

Top Negative Correlation variables for customer conversion :

1. Do not email_Yes
2. Lead Profile_Student of SomeSchool
3. Last Activity_Olark Chat Conversation