SUMMARY

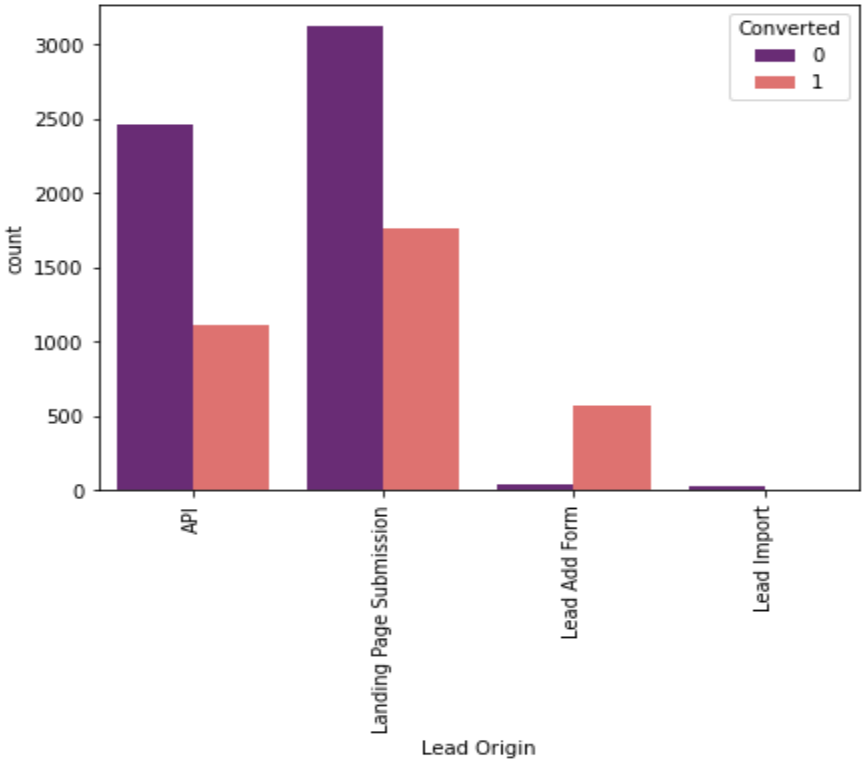
The analysis had been done for the company X Education to help them select the most promising leads, i.e. the leads that were most likely to convert into paying customers.

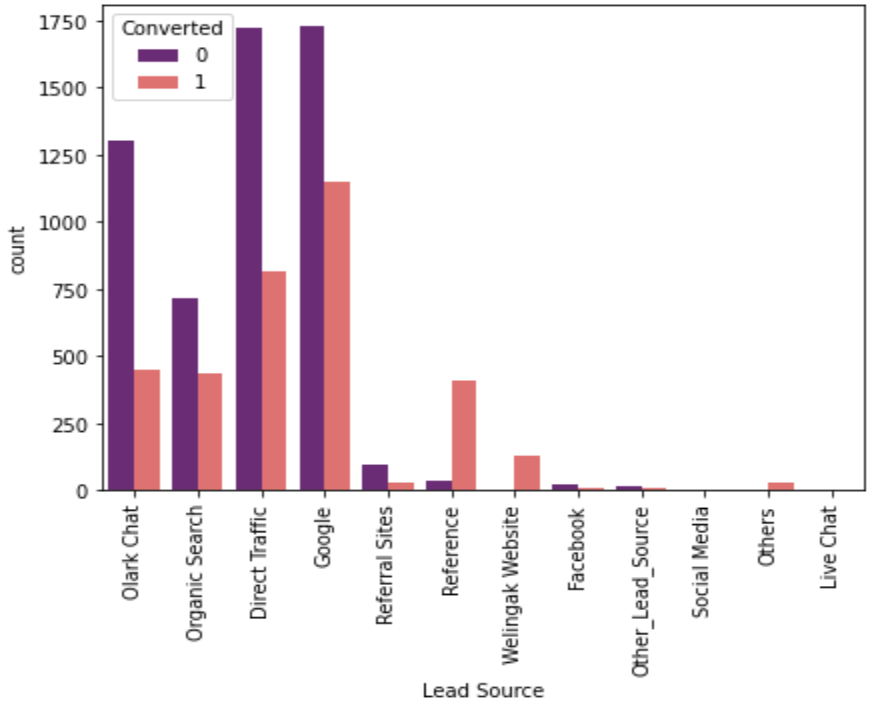
The following steps were performed to get the desired result-

1. Importing libraries and reading the data:
    - The required libraries were imported.For example the matplotlib,pandas and numpy.
    - The dataset was named as edu_lead and was read using pd.read_csv command
    - The shape,information and description of the data was checked. This DataFrame has 9240 Rows and 37 Columns
    - Duplicates were checked by using column Prospect ID and Lead Number as they are unique for each customer.

2. Data Exploration:
    - Columns which had 'Select' values were converted into NaN.
    - Replacing null values by checking the percentage null values of each columns
    - Columns greater than 45% missing values were dropped
    - Combining data in the columns that were close to each other or was representing same value
    - Imbalance variables were dropped

- Visualization of categorical and numerical data(removing outliers) done separately. Along with it the variables which affected converted column(target variable) were found and studied.



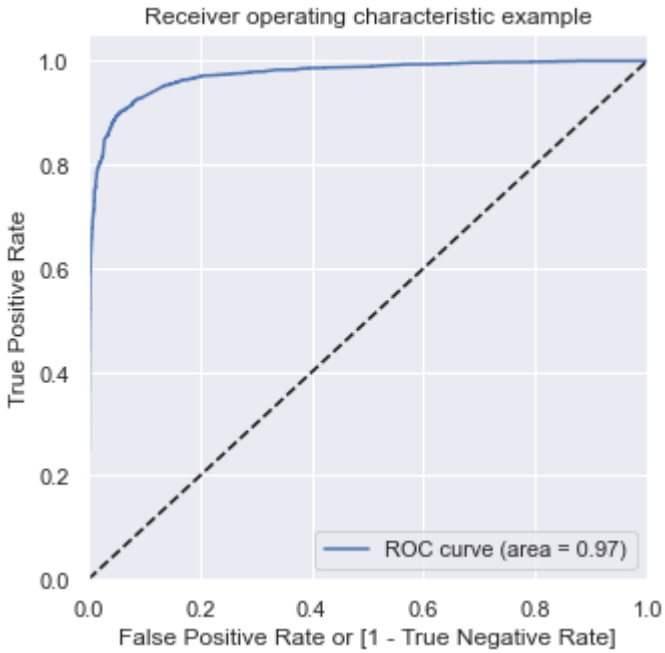


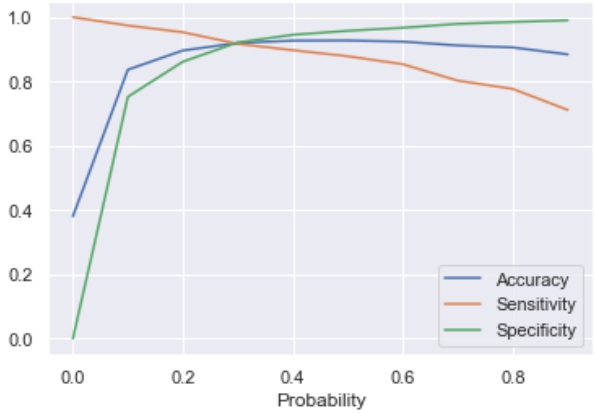
3. Preparing the data for model building:

- Dummy variables were created for categorical columns. In logistic regression models, encoding all of the independent variables as dummy variables allows easy interpretation and calculation of the odds ratios, and increases the stability and significance of the coefficients.

4. Train-Test Split & Logistic Regression Model Building:
    - The dataset divided into train and test in the ratio of 70:30
    - Scaling of the data done .Need to perform Feature Scaling when dealing with Gradient Descent Based algorithms (Logistic Regression) as these are very sensitive to the range of the data points.
    - Model Building using Stats Model & RFE-high p-values were dropped
    - VIF values checked .The values are less than 5
    - Predicted values on the train set.

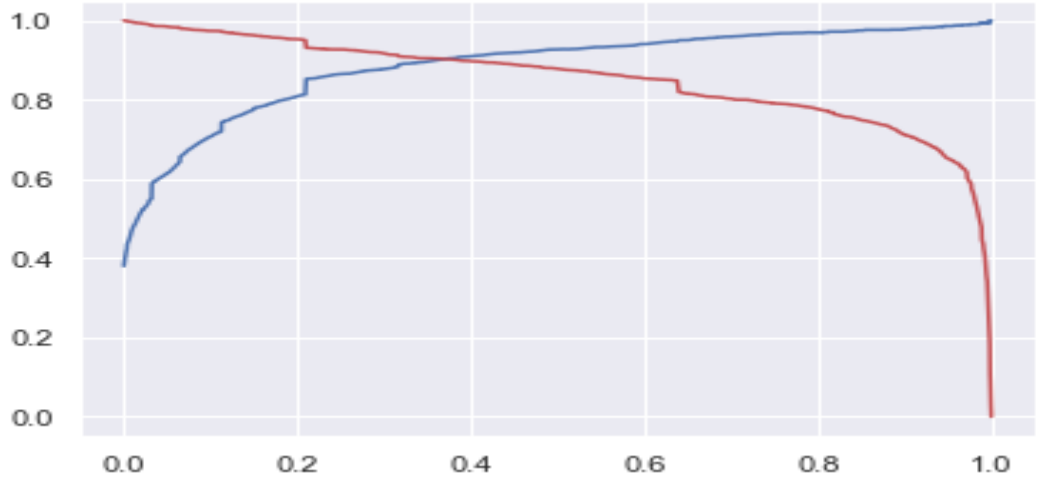
5. Model Evaluation:
    - Imported metrics from sklearn for evaluation
    - ROC CURVE-The ROC Curve should have value close to 1. Our graph shows that we are getting a good value of 0.97 indicating a good predictive model. ROC curves in logistic regression are used for determining the best cutoff value for predicting whether a new observation is a "failure" (0) i.e lead not converted or a "success" (1) i.e lead converted.





From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

- The precision and Recall seem to have trade-off at .38 . Precision measures how good our model is when the prediction is positive. Recall measures how good our model is at correctly predicting positive classes.



- Train Data:
  Observation-
  Accuracy:91.97%
  Sensitivity:91.69%
  Specificity:92.14%

6. Prediction on Test Dataset :
    - The values obtained after running model on Test Data -
      Accuracy : 92.51%

      Sensitivity : 91.88%

      Specificity : 92.89%

	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	7681	0	0.025541	3	0
1	984	0	0.015003	2	0
2	8135	0	0.692033	69	1
3	6915	0	0.002792	0	0
4	2712	1	0.963292	96	1

7. Final Observation:

- The target lead conversion rate has to be around 80%.So after running the model using the variables like Total time spend on website,Lead source –Google and Direct traffic,Last activity- SMS sent etc, the Conversion Rate is very well.