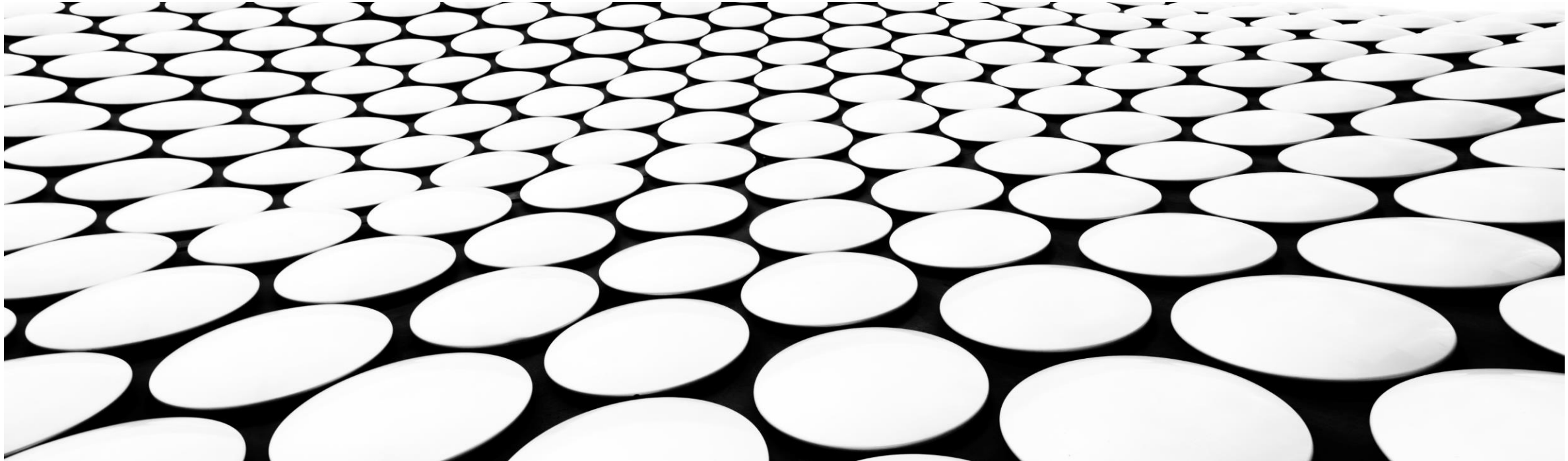
LEADS SCORING CASE STUDY

GARIMA BANSAL

MAMTA LOHANI



PROBLEM STATEMENT

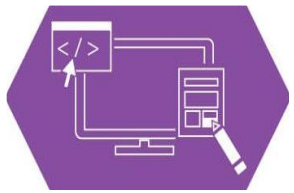
An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

BUSINESS GOAL

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

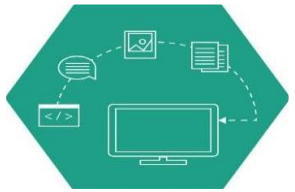
The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

MODELING PROCESS



Data Sourcing, Cleaning & Preparation

- Read Data from Source
- Clean data & Missing Value Treatment
- Outlier and Duplicate treatment
- Exploratory Data Analysis
- Feature Standardization



Feature Selection & Split

- Dummy Variable Creation
- Featuring Scaling
- Splitting Data into Test & Train Data set.

Model Training & Evaluation

- Feature Filtration using RFE method
- Evaluate Model using Model Performance Metrices.
- Determine Optimal Model

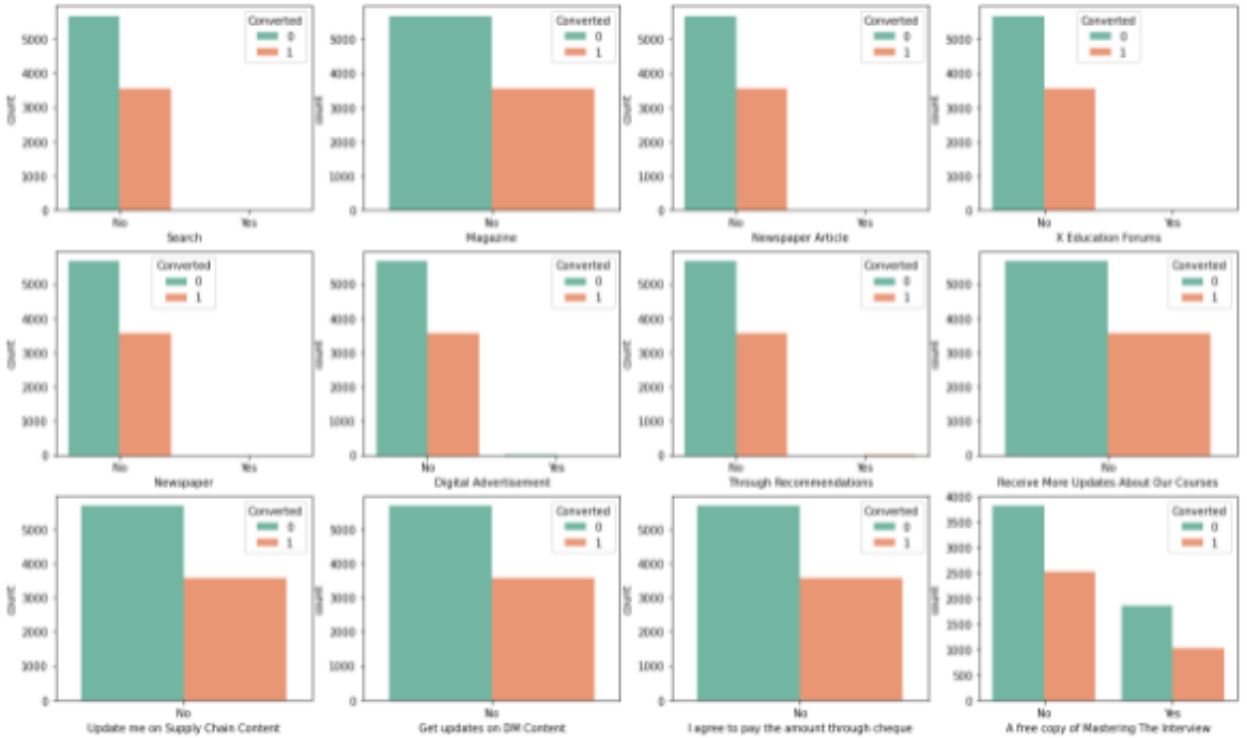
Model Result

- Predict Score using Model
- Evaluate the final prediction

EXPLORATORY DATA ANALYSIS

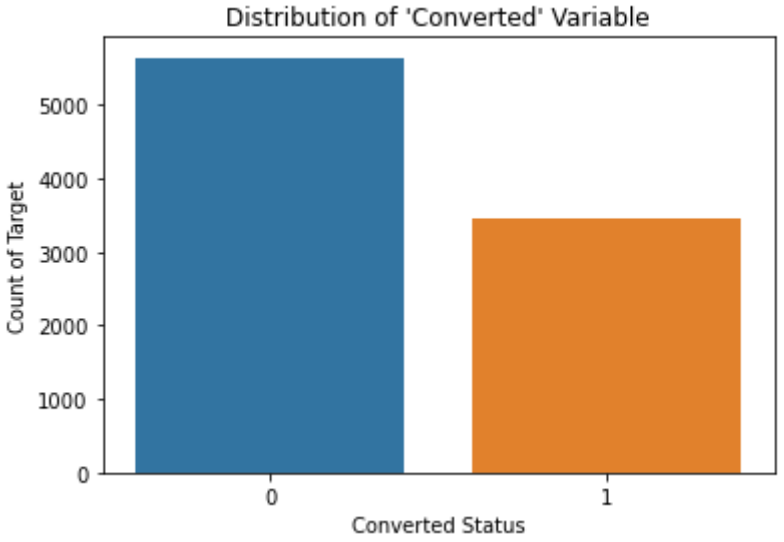
- For all these columns except 'A free copy of Mastering The Interview' data is highly imbalanced

- A free copy of Mastering The Interview" is a redundant variable

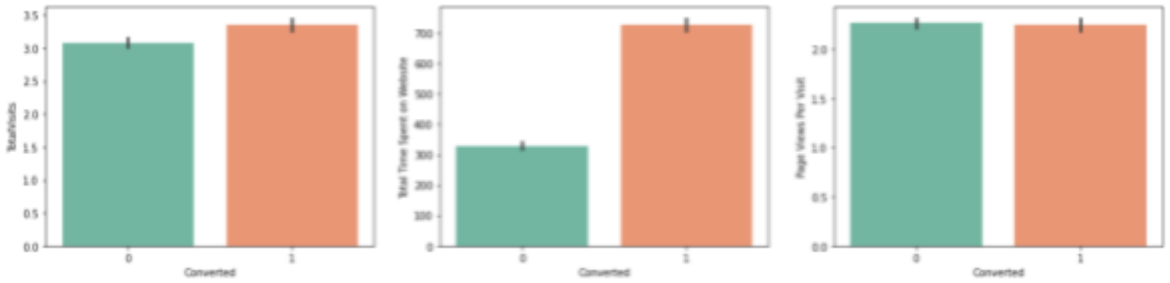


EXPLORATORY DATA ANALYSIS

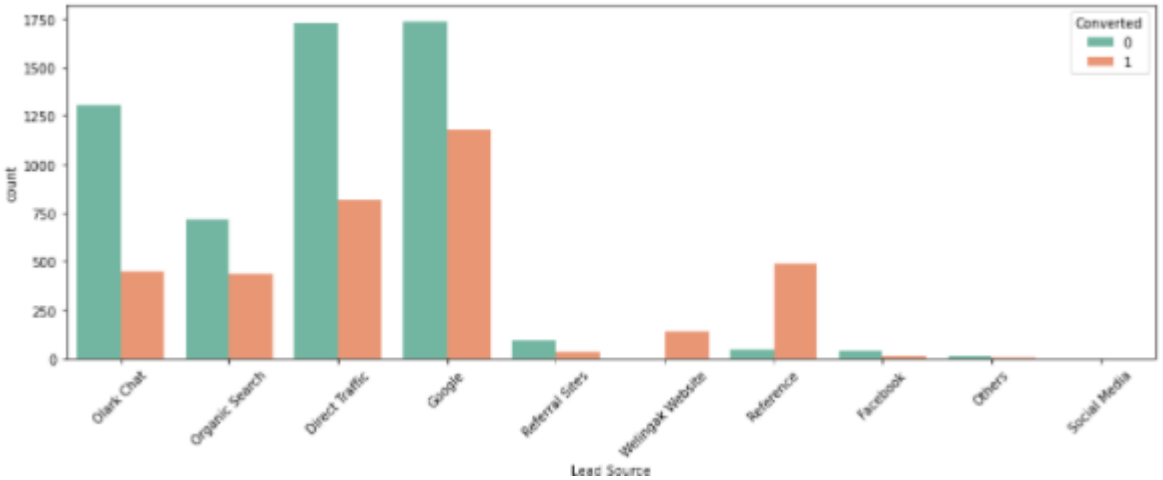
- We have around 38% Conversion rate in Total.



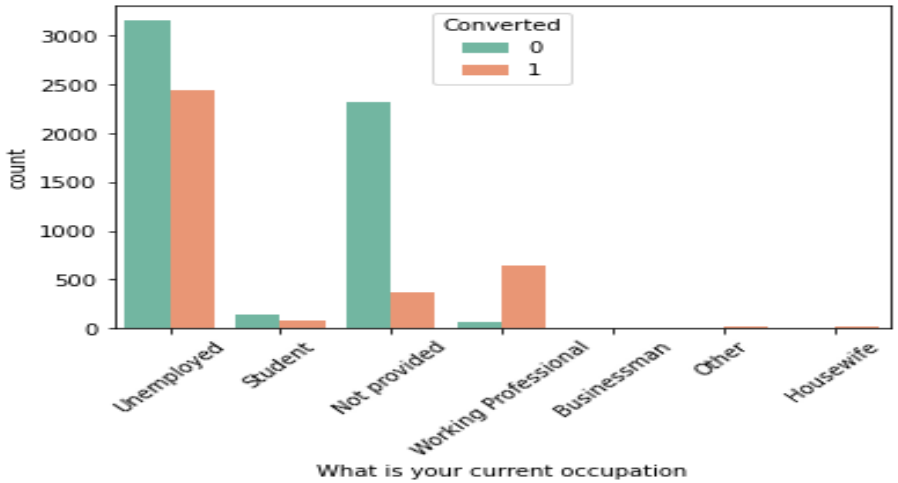
- The conversion rate is high for Total Visits, Total Time Spent on Website and Page Views Per Visit



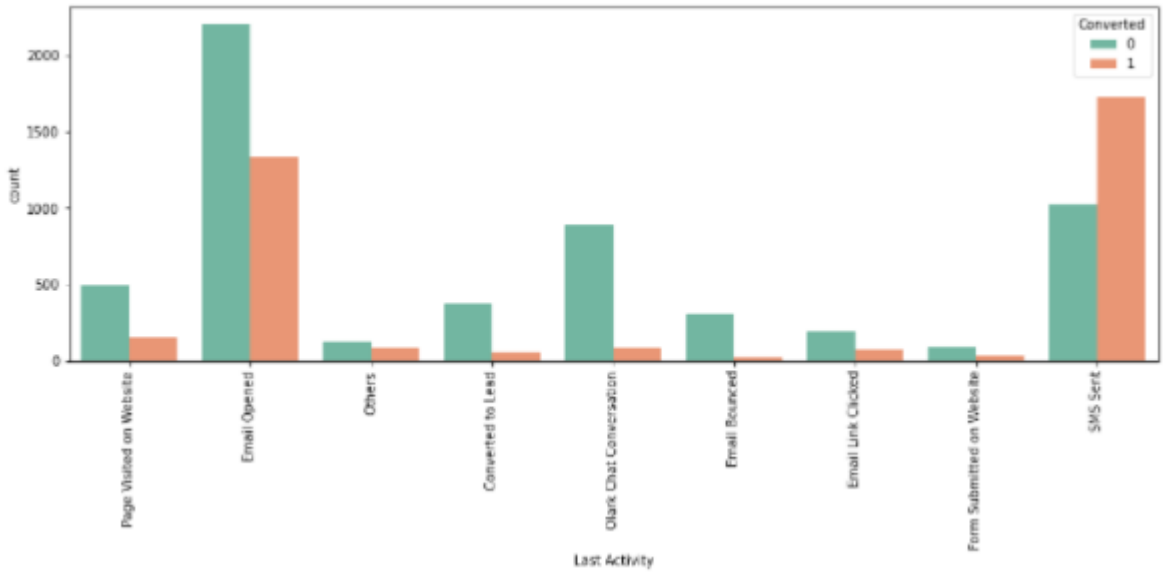
Maximum leads & Conversion are generated by Google & Direct Traffic

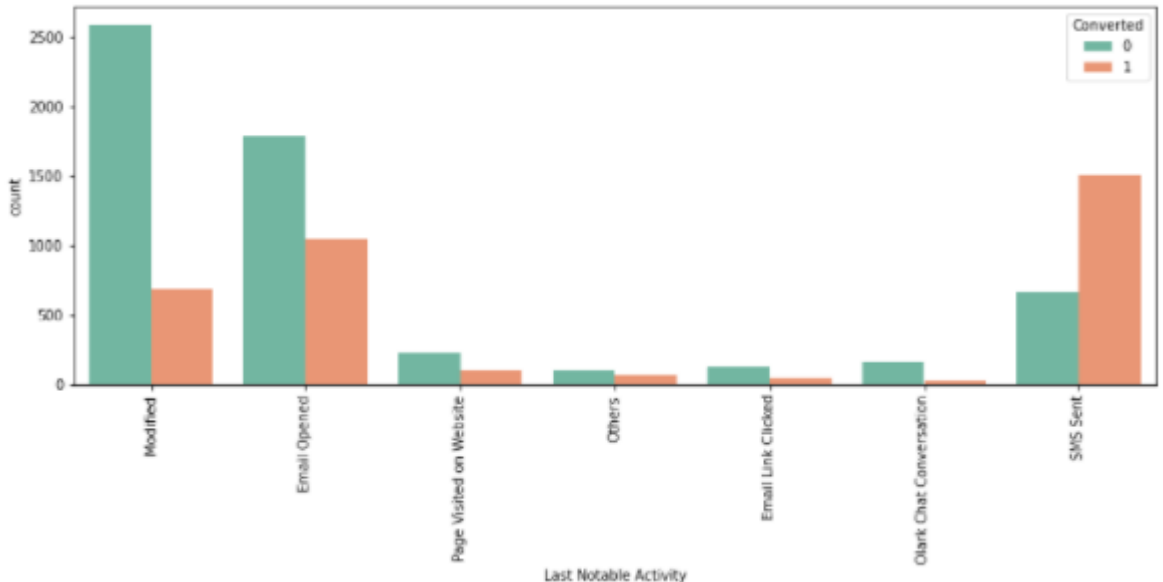


Maximum leads generated are unemployed and their conversion rate is more than 50%.



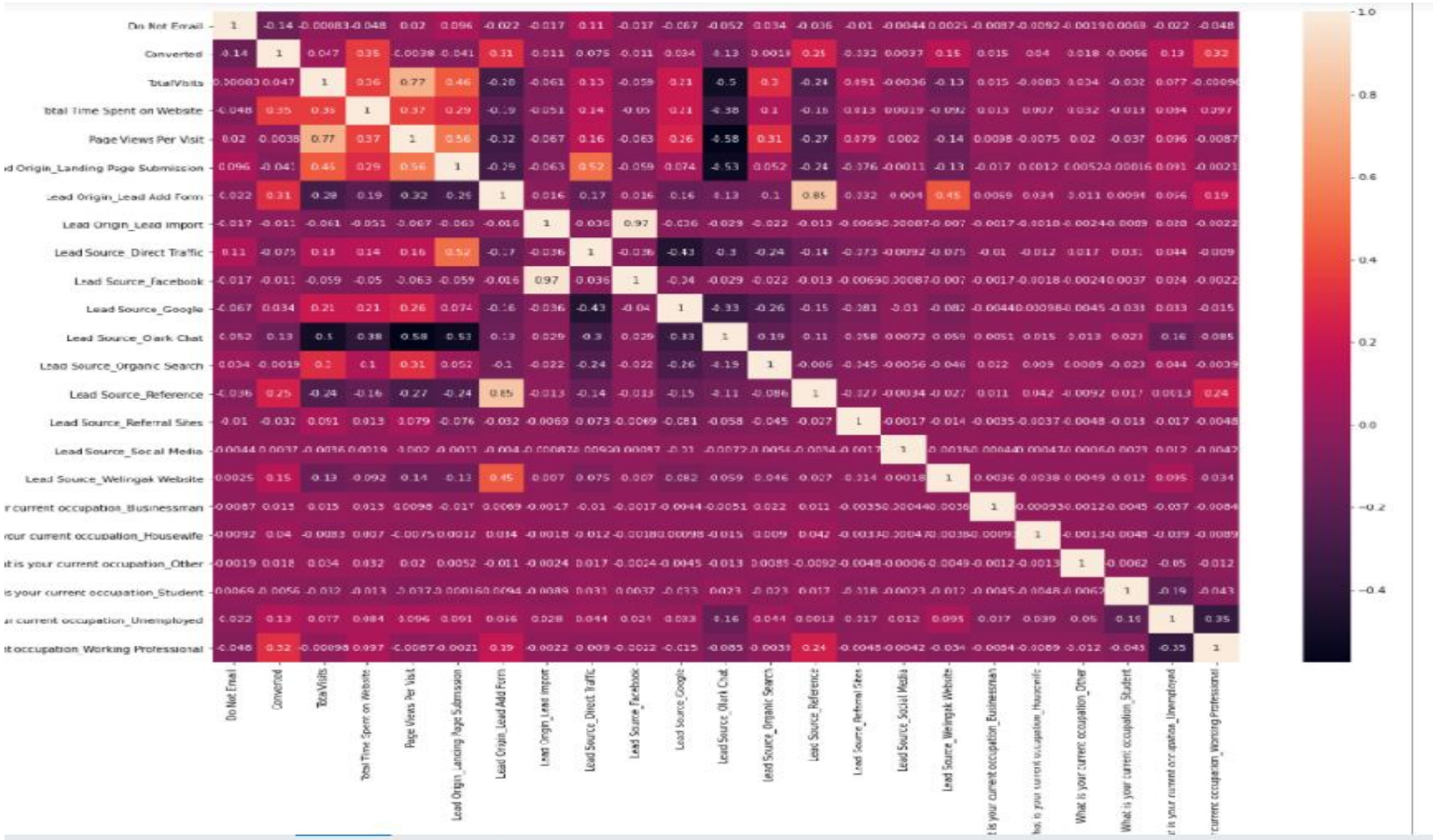
SMS sent as last activity has high conversion rate





CORRELATION

Correlation matrix after creating dummy variables. Lead Source_Olark Chat and Lead Origin_Landing Page Submission are highly correlated dummy variables.

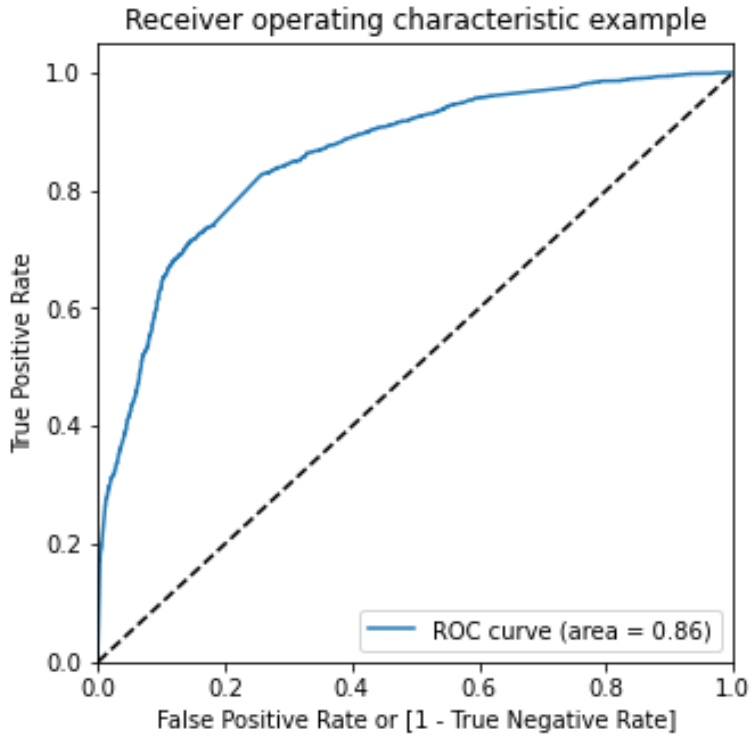


MODEL BUILDING – LOGISTIC REGRESSION

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Predictions on test data set

ROC CURVE AND CONFUSION MATRIX

After building the final model making prediction on train set, we created ROC curve to find the model stability with AUC score(area under the curve). As we can see from the graph plotted , the area score is 0.86 which is a good score.



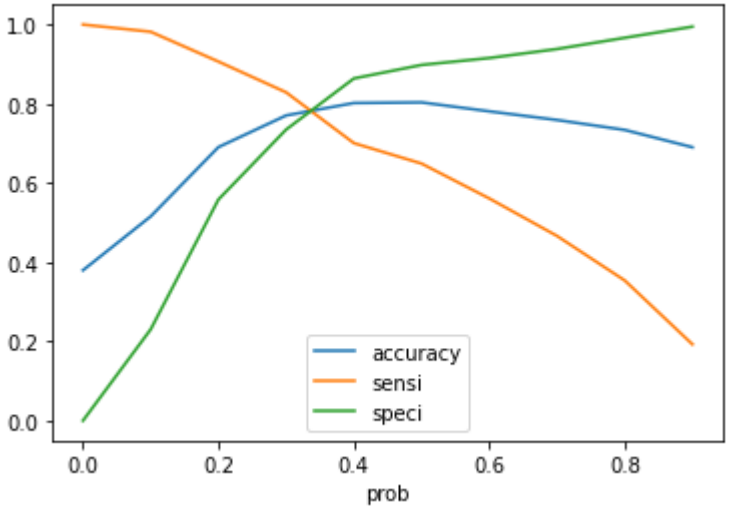
Confusion matrix

3350	403
849	1570

- Model Accuracy: 80.35%
- Sensitivity: 64.90%
- Specificity: 89.8%
- False Positive Rate: 10.19%
- Positive Predictive Value: 79.57%
- Negative Predictive Value: 80.70%

MODEL EVALUATION - SENSITIVITY AND SPECIFICITY ON TRAIN DATA SET

The graph depicts an optimal cut off of 0.3 based on Accuracy, Sensitivity and Specificity



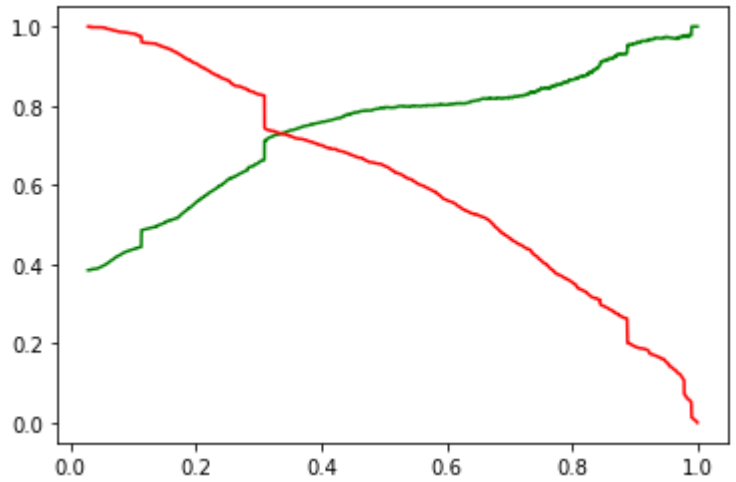
Confusion matrix

2905	1048
414	2005

- Cut-off Point: 0.3
- Model Accuracy: 77.05%
- Sensitivity: 82.89%
- Specificity: 73.49%
- False Positive Rate: 26.51%
- Positive Predictive Value: 65.67%
- Negative Predictive Value: 87.52%
- Precision: 65.67
- Recall: 82.88

MODEL EVALUATION- PRECISION AND RECALL ON TRAIN DATASET

The graph depicts an optimal cut off of 0.3 based on Precision and Confusion Matrix Recall



Confusion matrix

2905	1048
414	2005

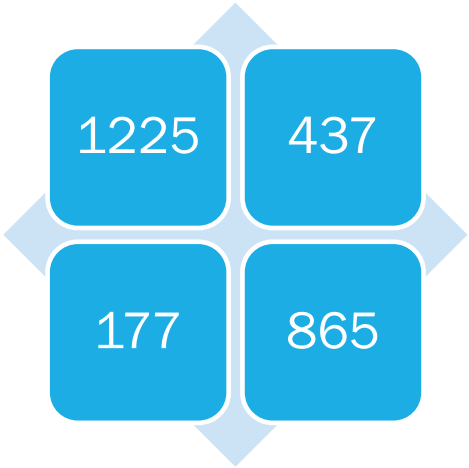
- Precision Score: 66.43
- Recall Score: 83.01

MODEL EVALUATION – SENSITIVITY AND SPECIFICITY ON TEST DATASET

After running model on test data below are the

- Accuracy: 77.52
- Sensitivity: 83.01
- Specificity: 74.13

Confusion matrix



CONCLUSION

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Lead Source_Organic Search

Thank You