LEAD SCORING CASE STUDY

Submitted By Svetlana, Sathwikava and Pikasa

PROBLEM STATEMENT

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%

BUSINESS OBJECTIVE

- Lead X wants us to build a model to give every lead a lead score between 0 - 100 .This is so that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They also want the model to be able to handle future constraints as well as peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

PROBLEM SOLVING APPROACH OVERVIEW

Data Cleaning and EDA

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

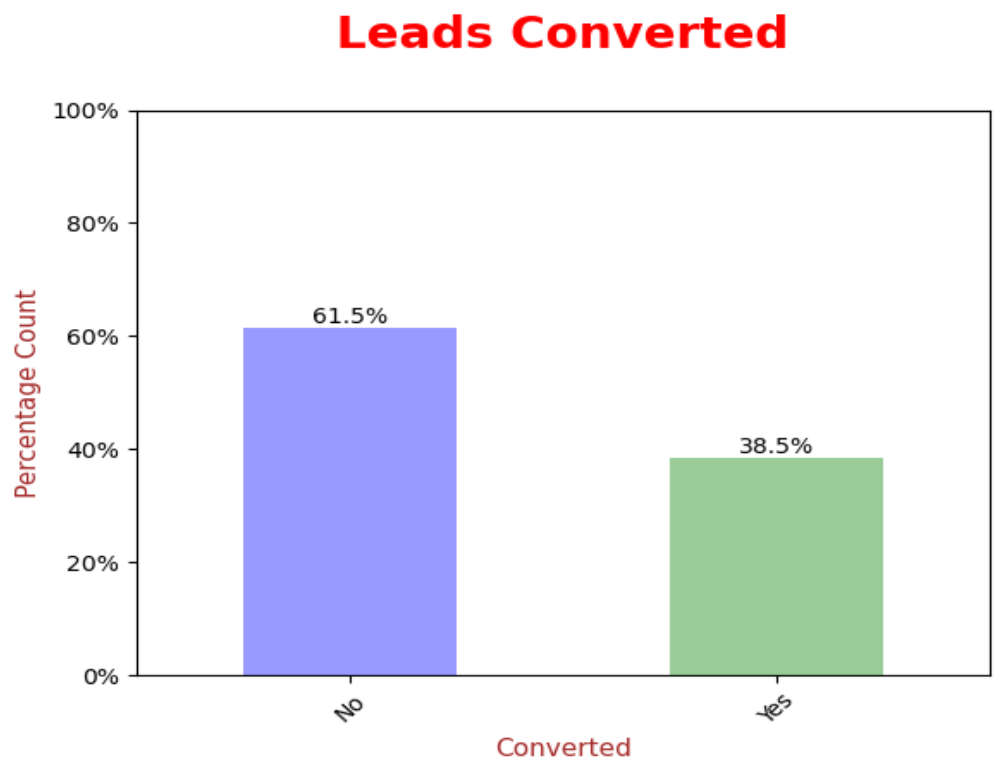
Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Checking Performance Metrics/Results

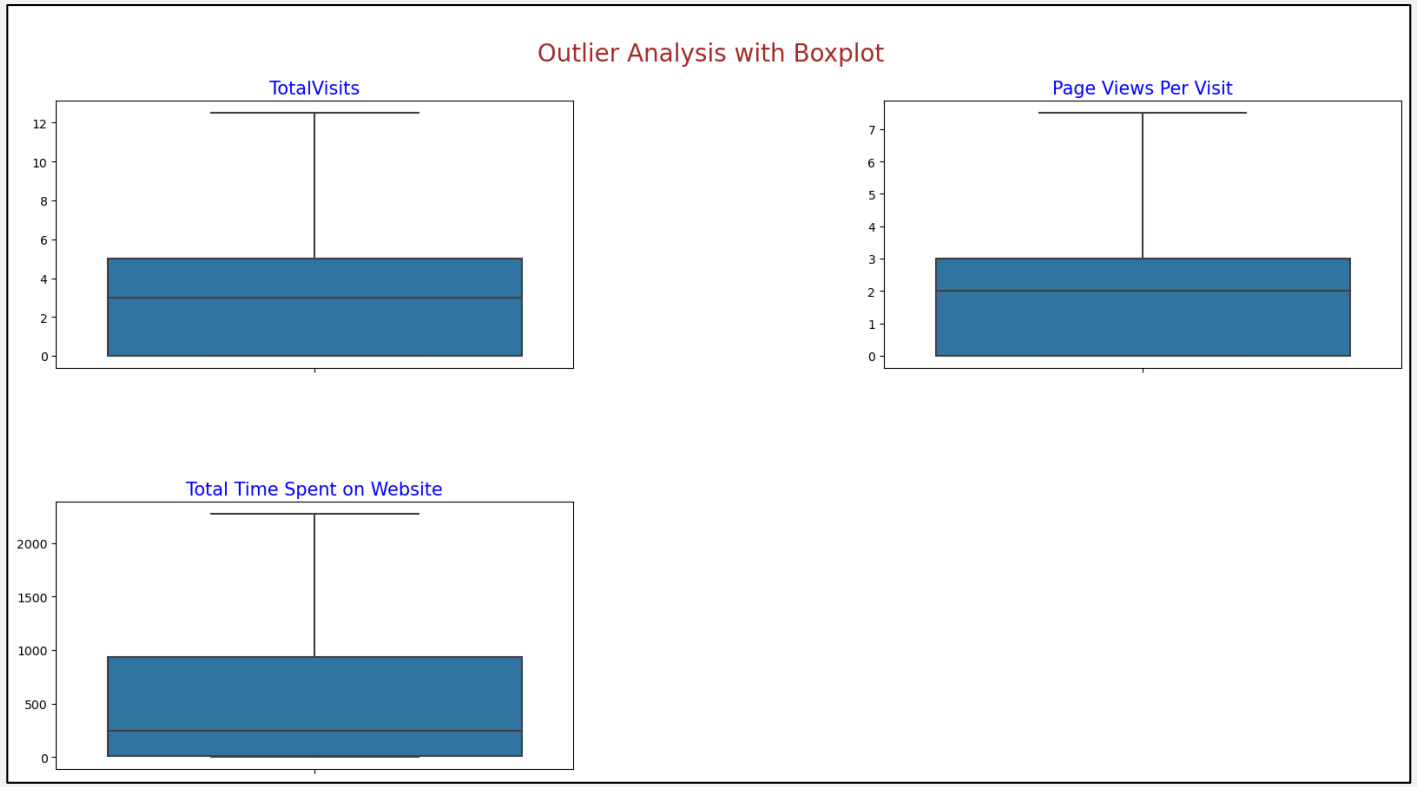
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

CHECKING DATA DISTRIBUTION



As we can see, the conversion rate of leads is around 39%

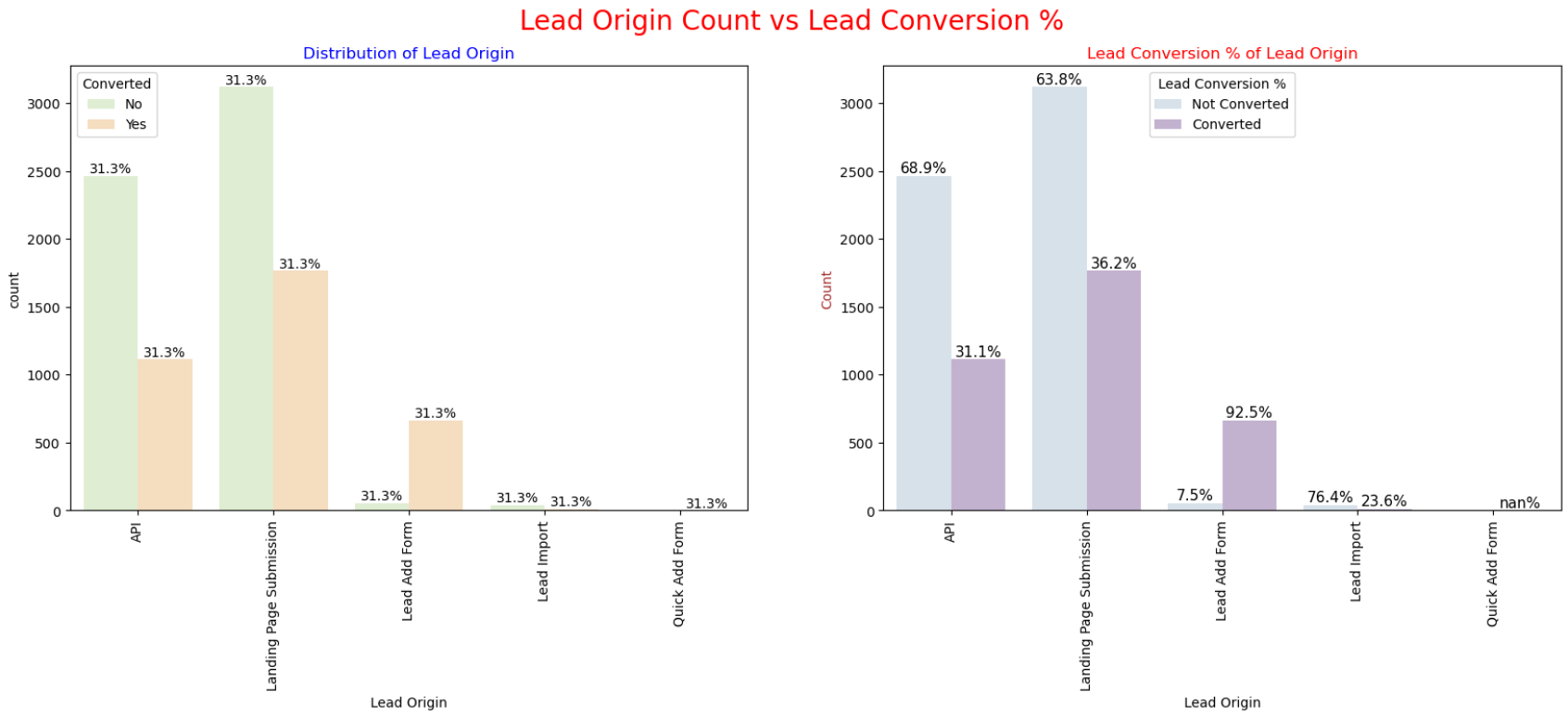
OUTLIER TREATMENT



After removing the outliers, we obtained the distributions as seen here.

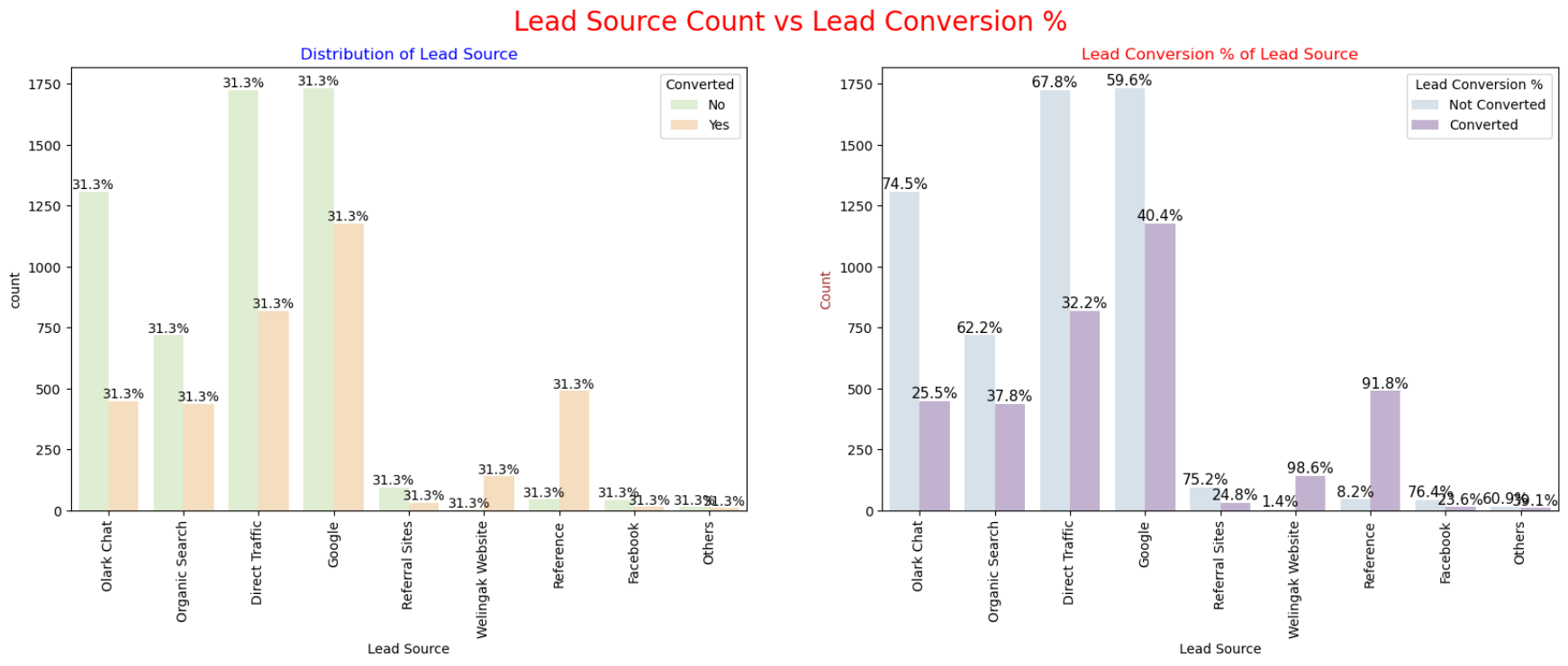
RELATIONSHIP BETWEEN DATA COLUMNS AND TARGET

LEAD ORIGIN VS. LEAD CONVERSION



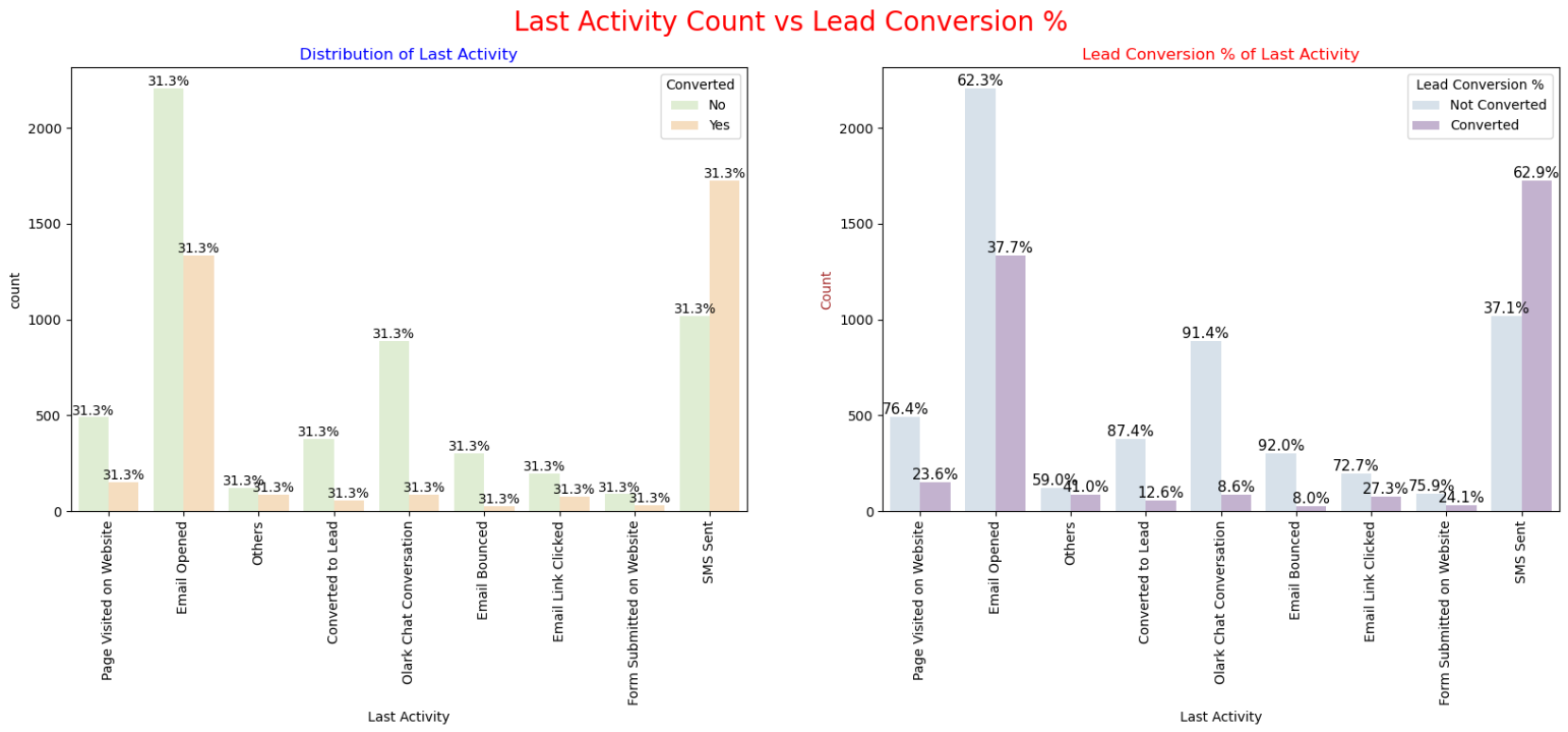
- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.

LEAD SOURCE VS. LEAD CONVERSION



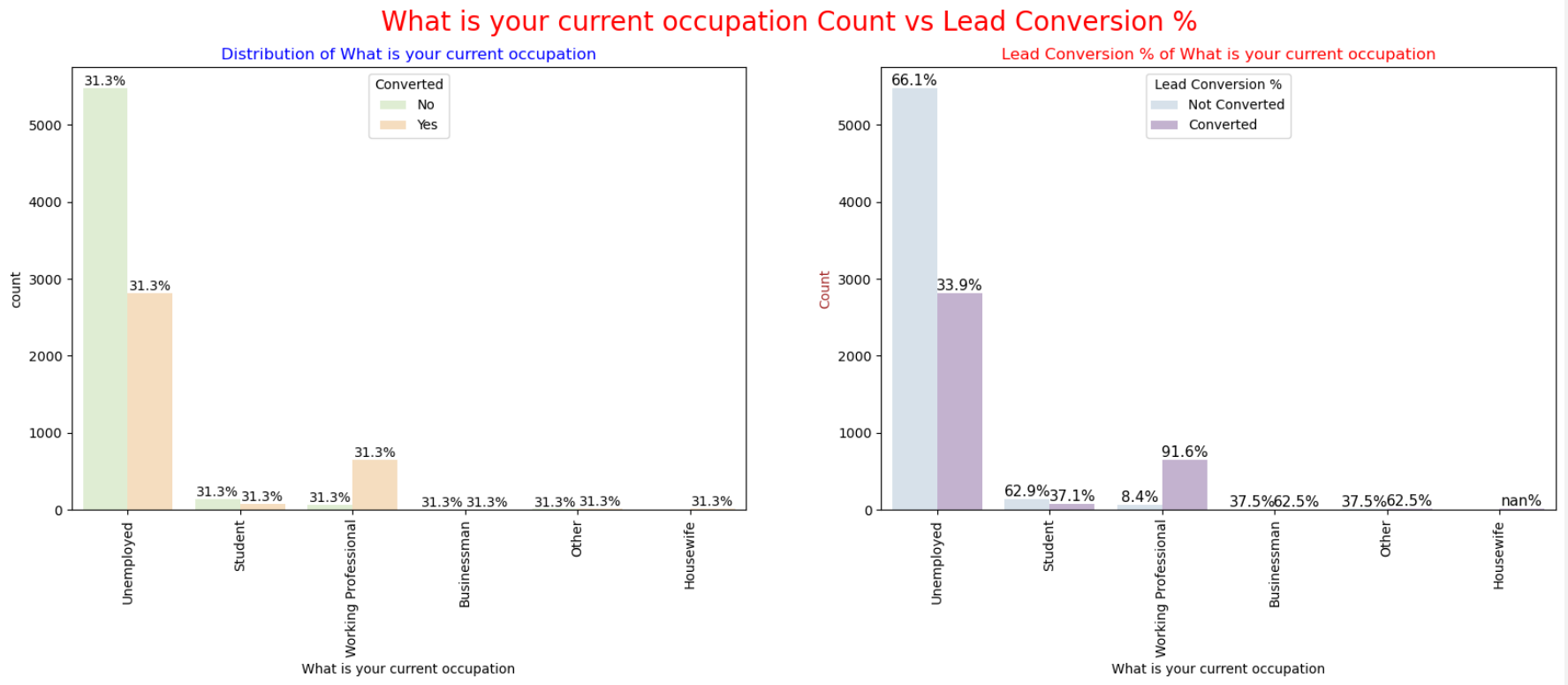
Focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads, as well as generate more leads from reference and welingak website

LAST ACTIVITY COUNT VS. LEAD CONVERSION



Most of the lead have their Email opened as their last activity. Lead conversion from SMS sent is also high, more than 60%.

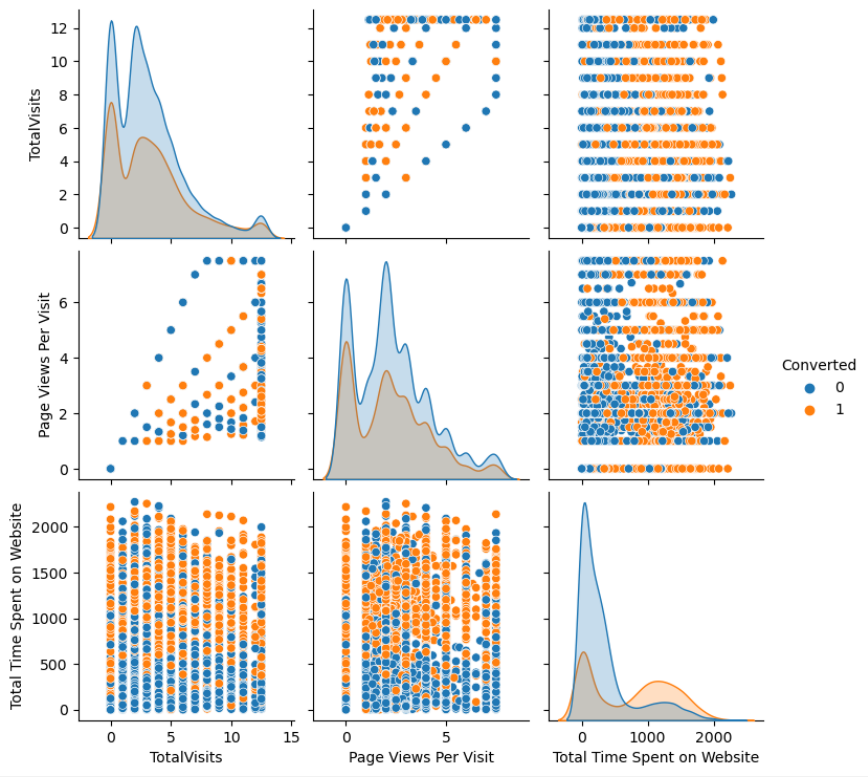
CURRENT OCCUPATION VS. LEAD CONVERSION



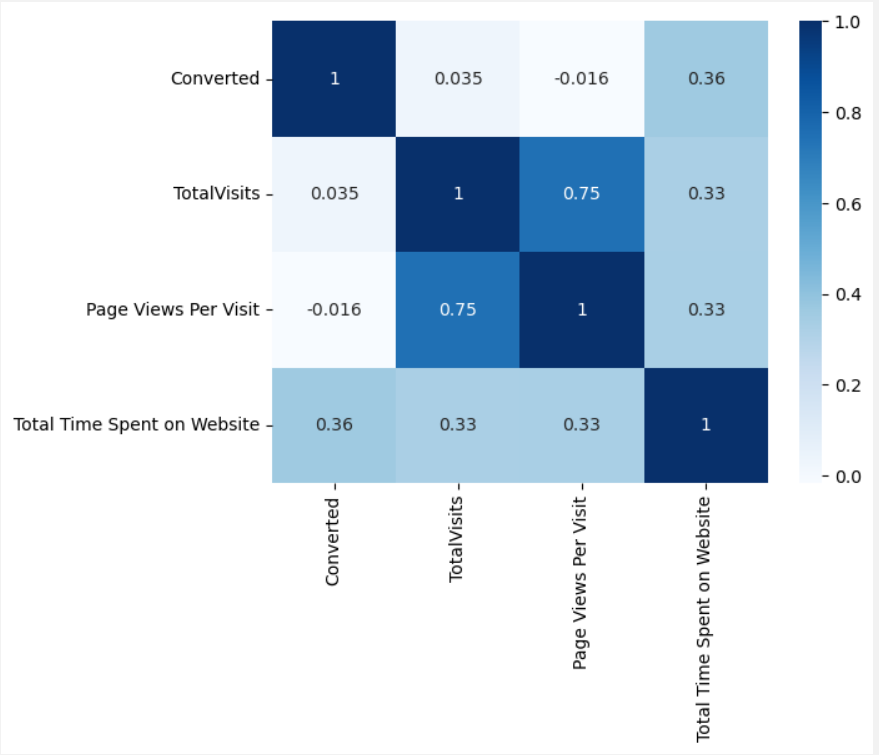
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in numbers but has around 30-35% conversion rate.

Thus, focus should be on working professionals to improve lead conversion rate.

DATA PREPARATION



Bivariate Analysis of Numerical Variables

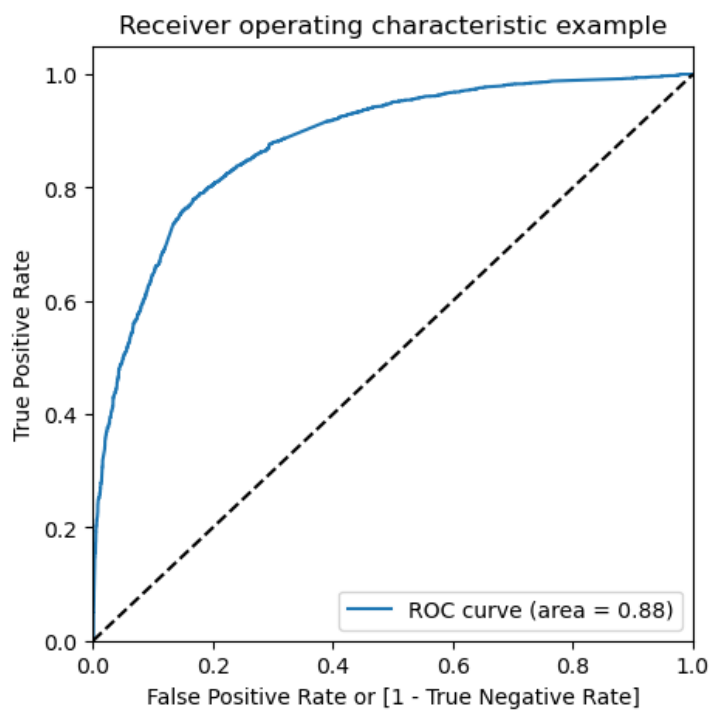


Heatmap Generated

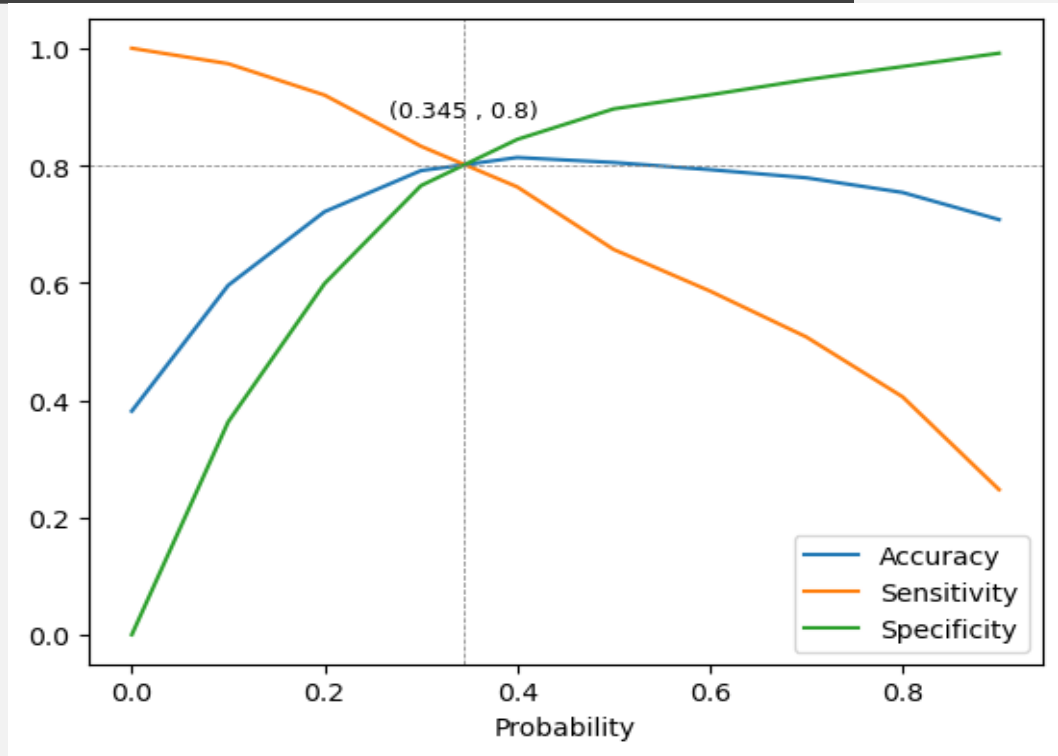
DATA PREPARATION FOR MODELLING

- Dummy variables were created and concatenated with lead data
- Splitting of data for training and testing was done
- Scaling was done with standardization
- RFE Feature Selection was performed

MAKING & OPTIMIZING MODEL



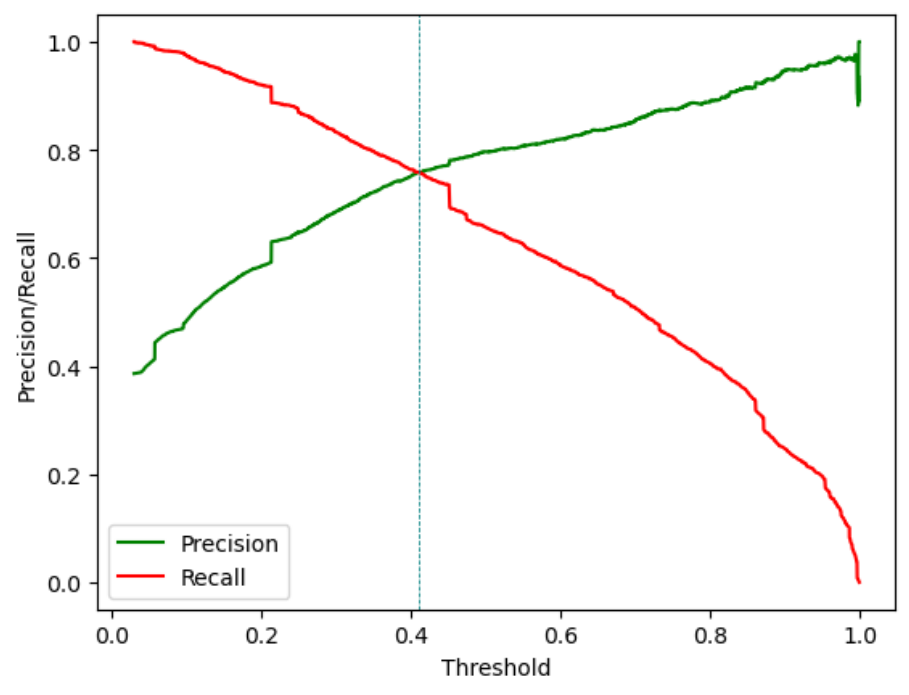
Finding ROC of test data



Evaluating accuracy, precision and sensitivity of test data set.

EVALUATION METRICS

- Precision and Recall Trade Off



	Converted	Converted_Prob	Prospect ID	final_predicted	precision_recall_prediction
0	0	0.474082	1871	1	1
1	0	0.073252	6795	0	0
2	0	0.249087	3516	0	0
3	0	0.768973	8105	1	1
4	0	0.212973	3934	0	0

OBSERVATIONS

****

Confusion Matrix
[[3406 596]
 [596 1870]]

****

True Negative : 3406
True Positive : 1870
False Negative : 596
False Positve : 596
Model Accuracy : 0.8157
Model Sensitivity : 0.7583
Model Specificity : 0.8511
Model Precision : 0.7583
Model Recall : 0.7583
Model True Positive Rate (TPR) : 0.7583
Model False Positive Rate (FPR) : 0.1489

****

Test Data

****

Confusion Matrix
[[1353 324]
 [221 874]]

****

True Negative : 1353
True Positive : 874
False Negative : 221
False Positive : 324
Model Accuracy : 0.8034
Model Sensitivity : 0.7982
Model Specificity : 0.8068
Model Precision : 0.7295
Model Recall : 0.7982
Model True Positive Rate (TPR) : 0.7982
Model False Positive Rate (FPR) : 0.1932

****

Train Data

INFERENCES

- The training and test data are similar in accuracy, precision and sensitivity.
- The following are the top features that present us with high probability of lead conversion rate:

- Lead Source_Welingak Website: 5.388662
- Lead Source_Reference: 2.925326
- Current_Occupation_Working Professional: 2.669665
- Last Activity_SMS Sent: 2.051879