Lead Score case study for X Education

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 and address and 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 Goal :

X Education needs help in selecting the most promising leads, i.e. the leads That are the most likely to convert into paying customers

The company needs a model wherein you a lead score is assigned 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%.

Strategy

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis
- Feature Scaling
- Splitting the data into Test and Train dataset
- Building a Logistic Regression model and calculate Lead score.
- Evaluating the model by using different metrics – Specificity and Sensitivity or Precision and Recall.
- Applying the best model in the Test data based on the Sensitivity and Specificity metrics.

Data Sourcing , Cleaning and Preparation

- 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.

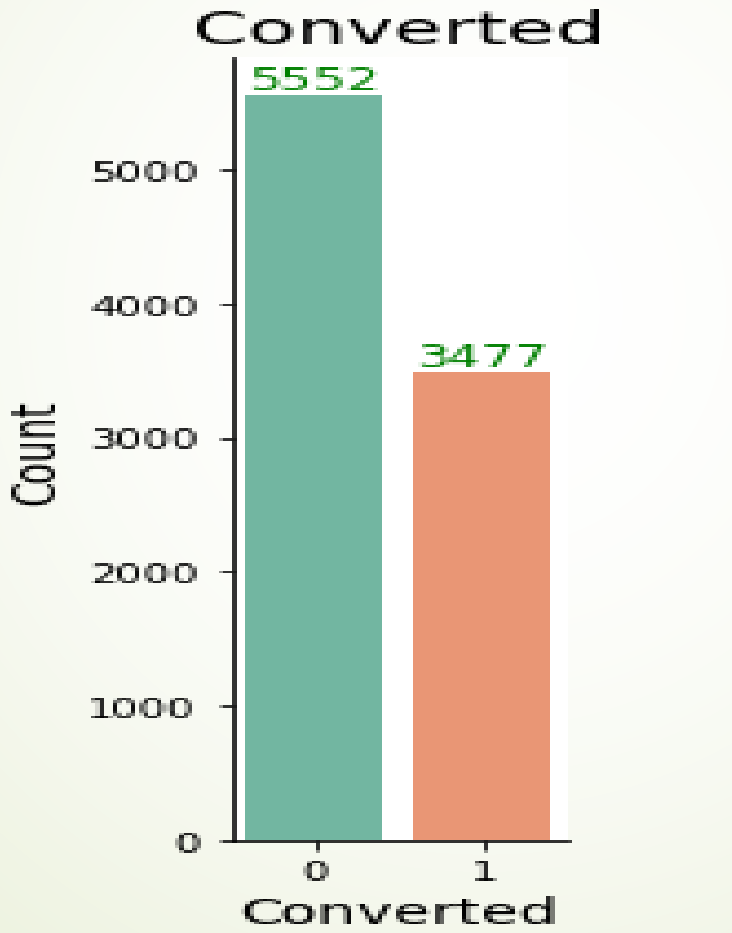
Model Building

- 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.

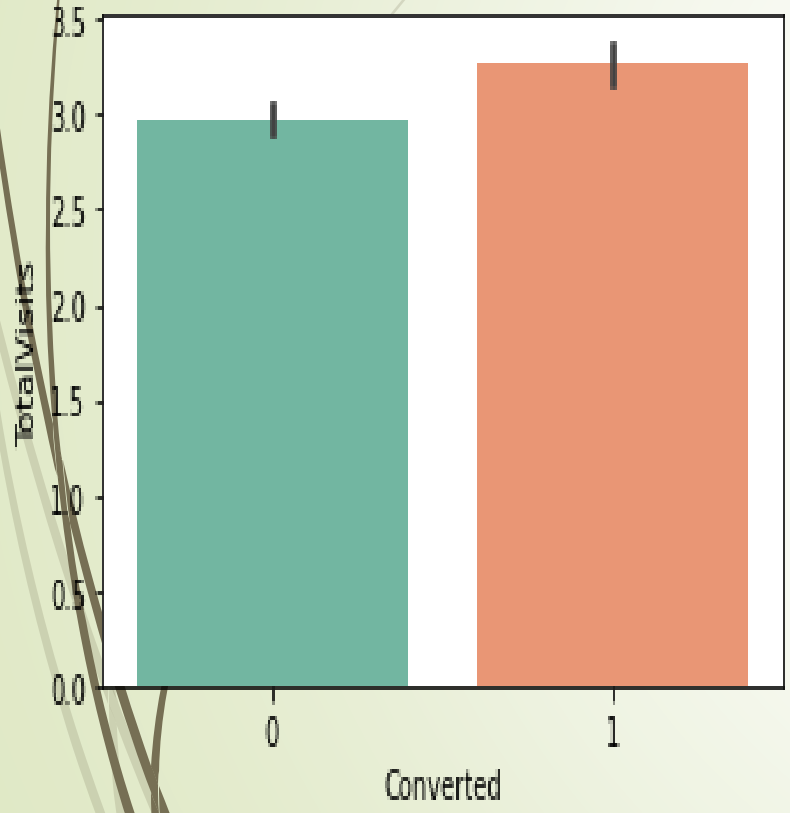
Result

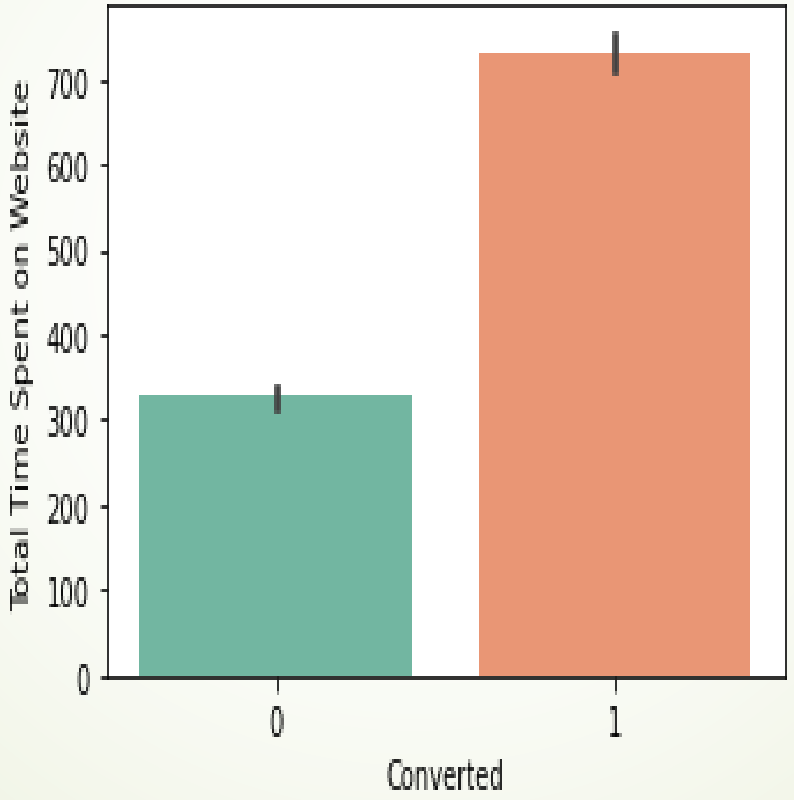
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

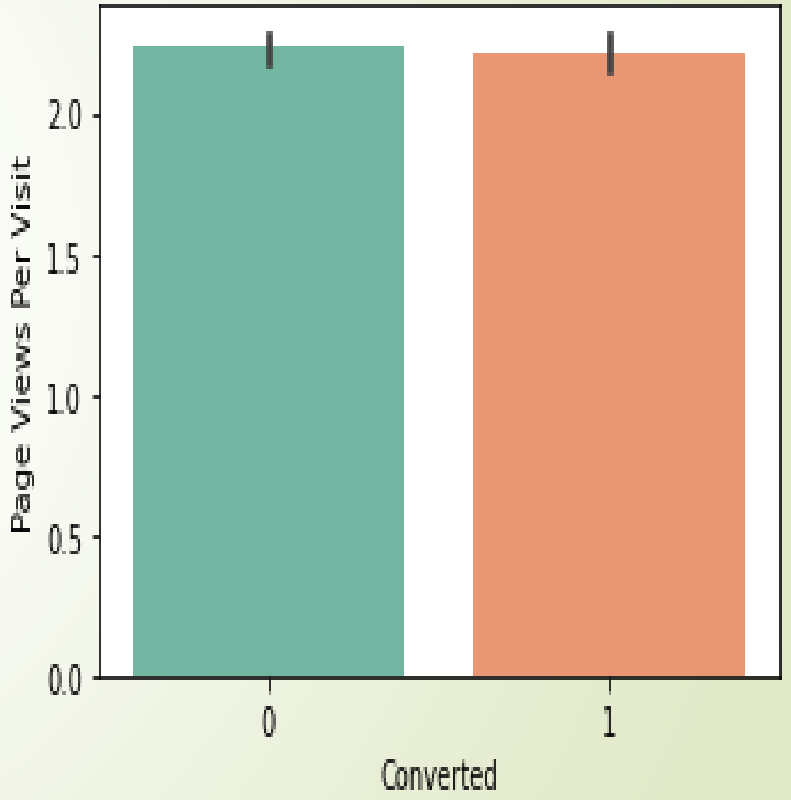
We have around 39% Conversion rate in Total



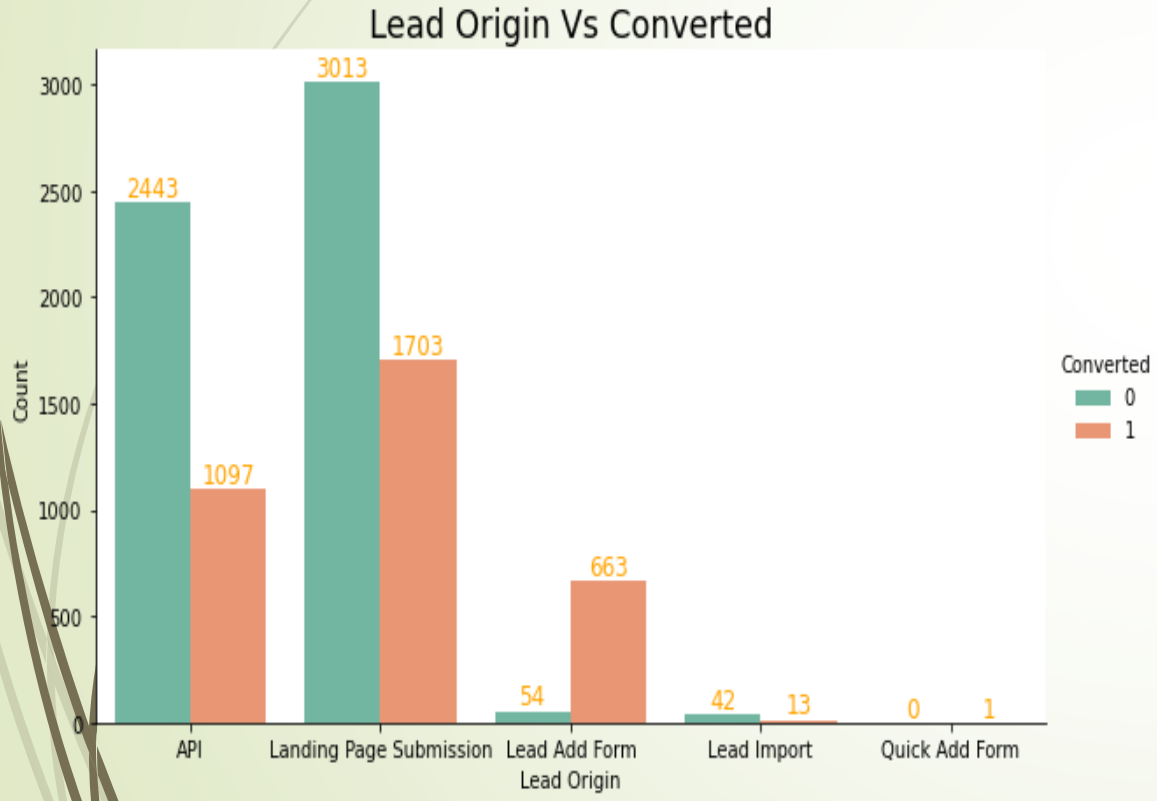
The conversion rated were high for Total Visits, Total Time Spent on Website and Page Views Per Visit



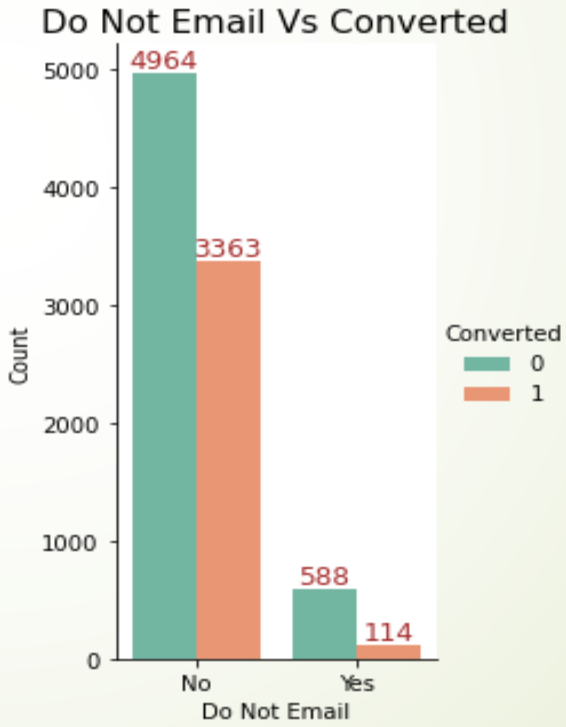


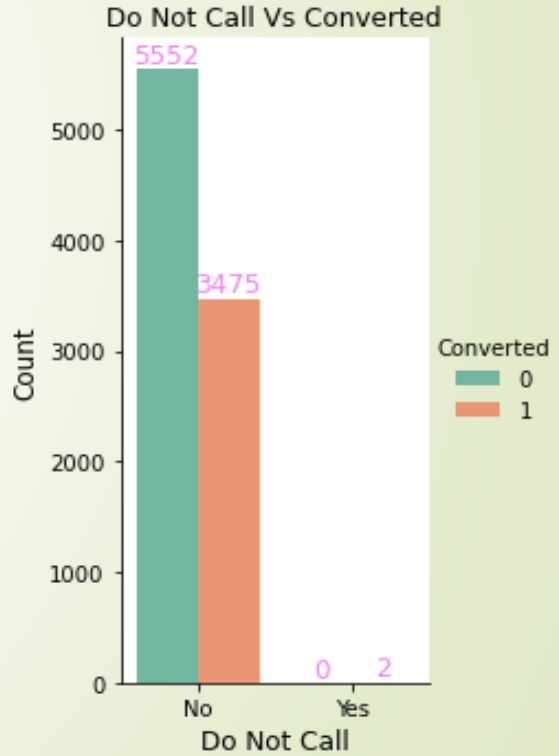


In Lead Origin, maximum conversion happened from Landing Page Submission

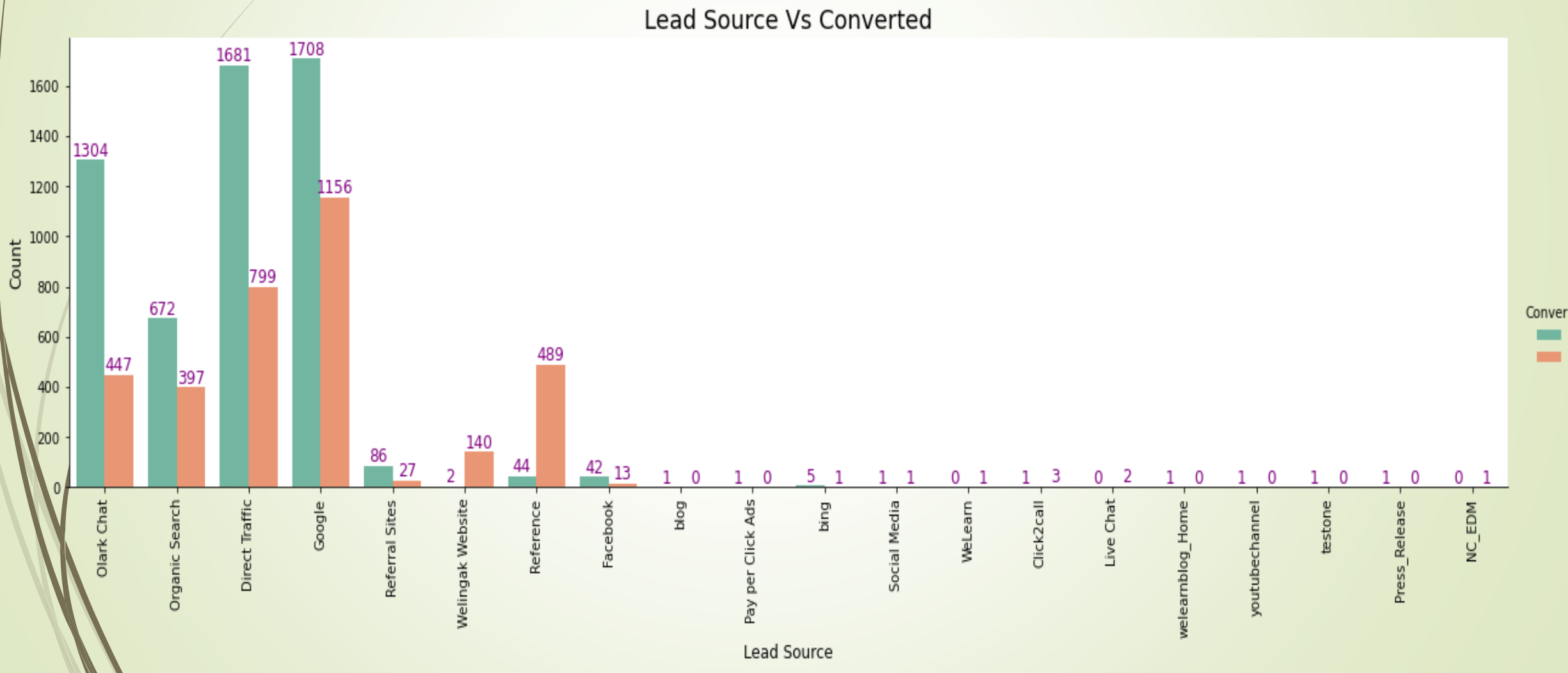


Major conversion has happened from Emails sent and Calls made

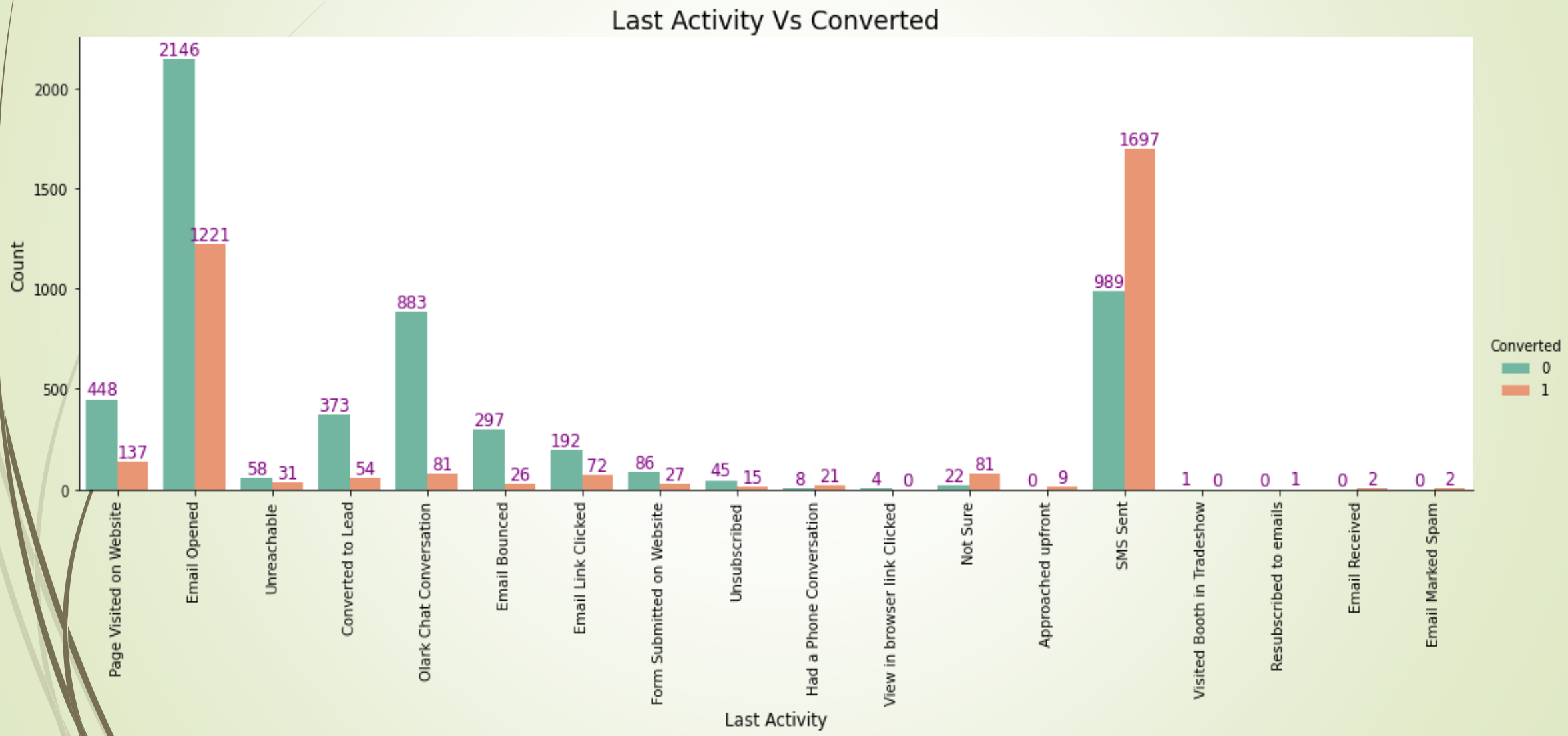




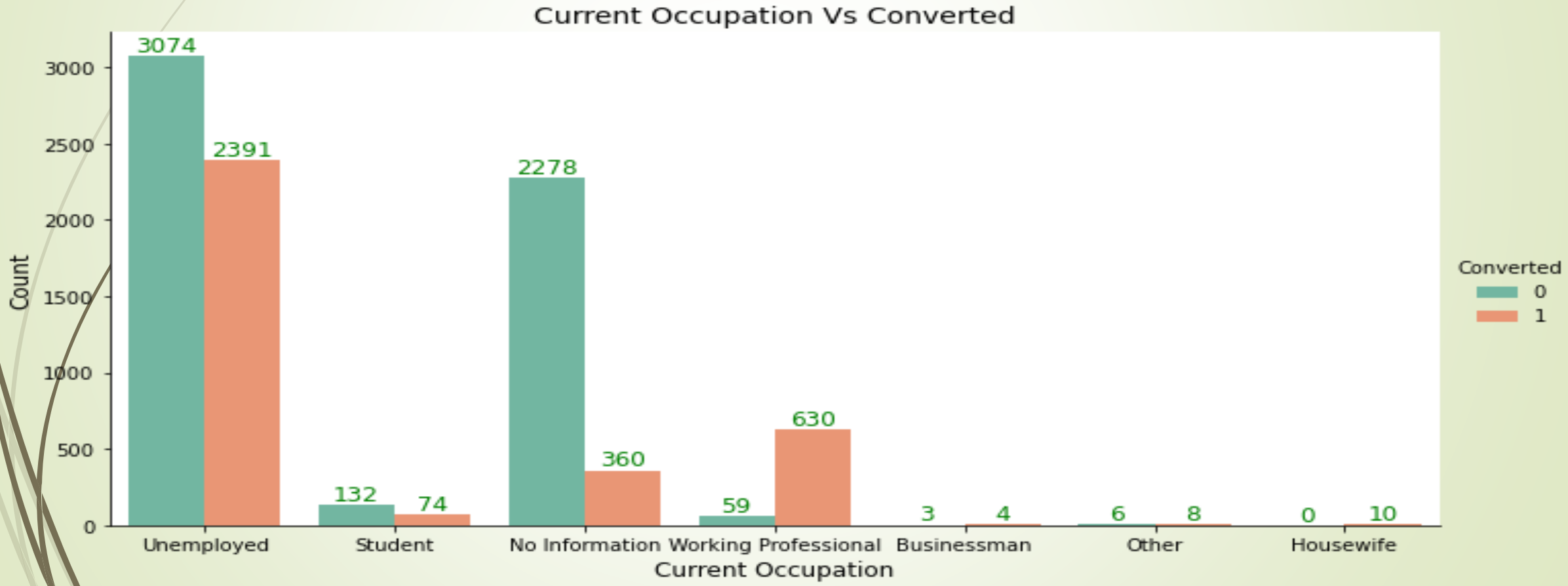
Major conversion in the lead source is from Google



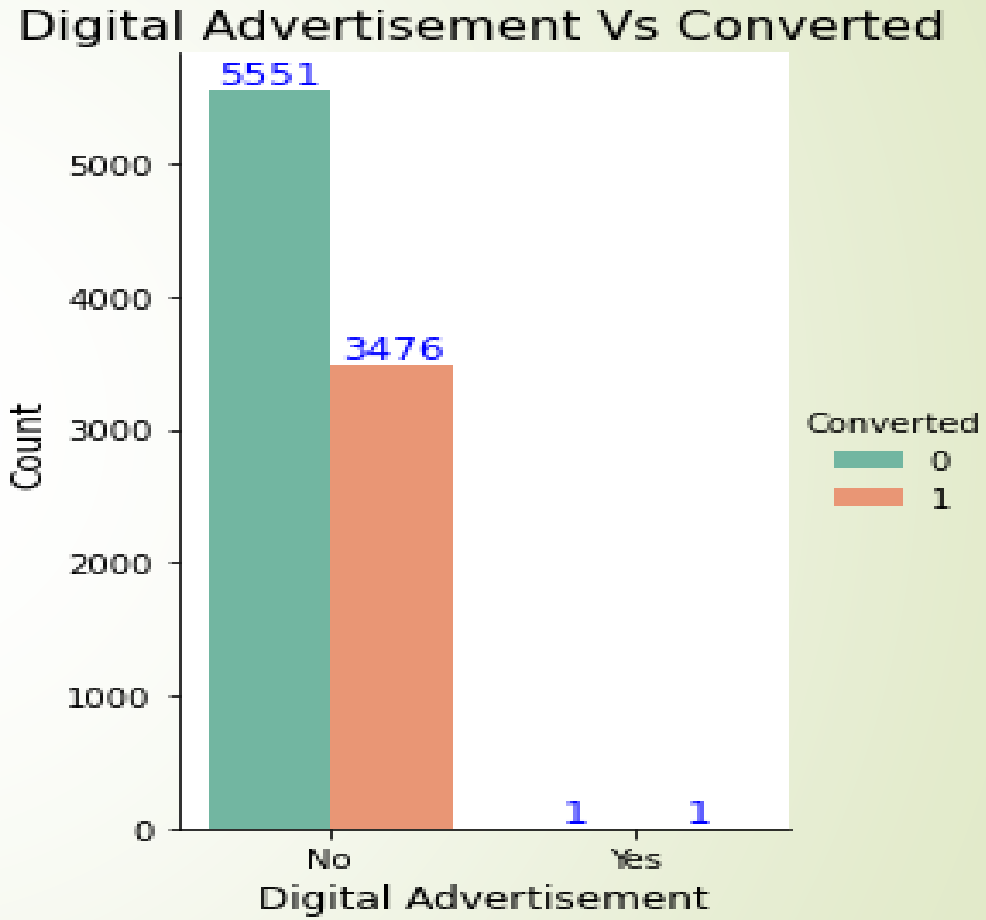
Last Activity value of SMS Sent' had more conversion



More conversion happened with people who are unemployed. It can also be noticed from the above data that - Out of 7 business men, 4 got converted - Out 10 housewives, all 10 leads got converted.



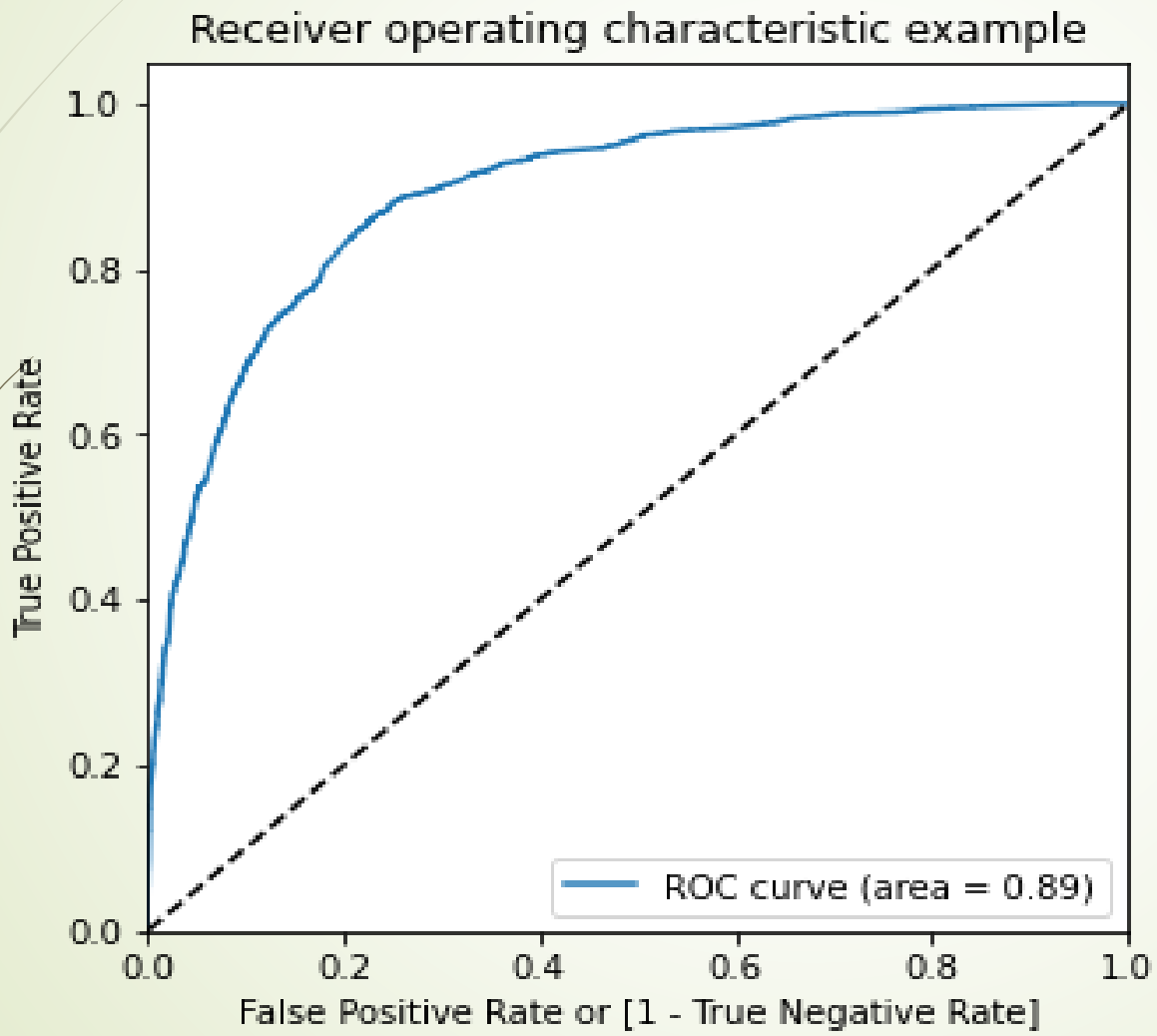
It can be noticed that there were 2 leads that came from digital advertisement of which one lead got converted



Variables Impacting the Conversion Rate

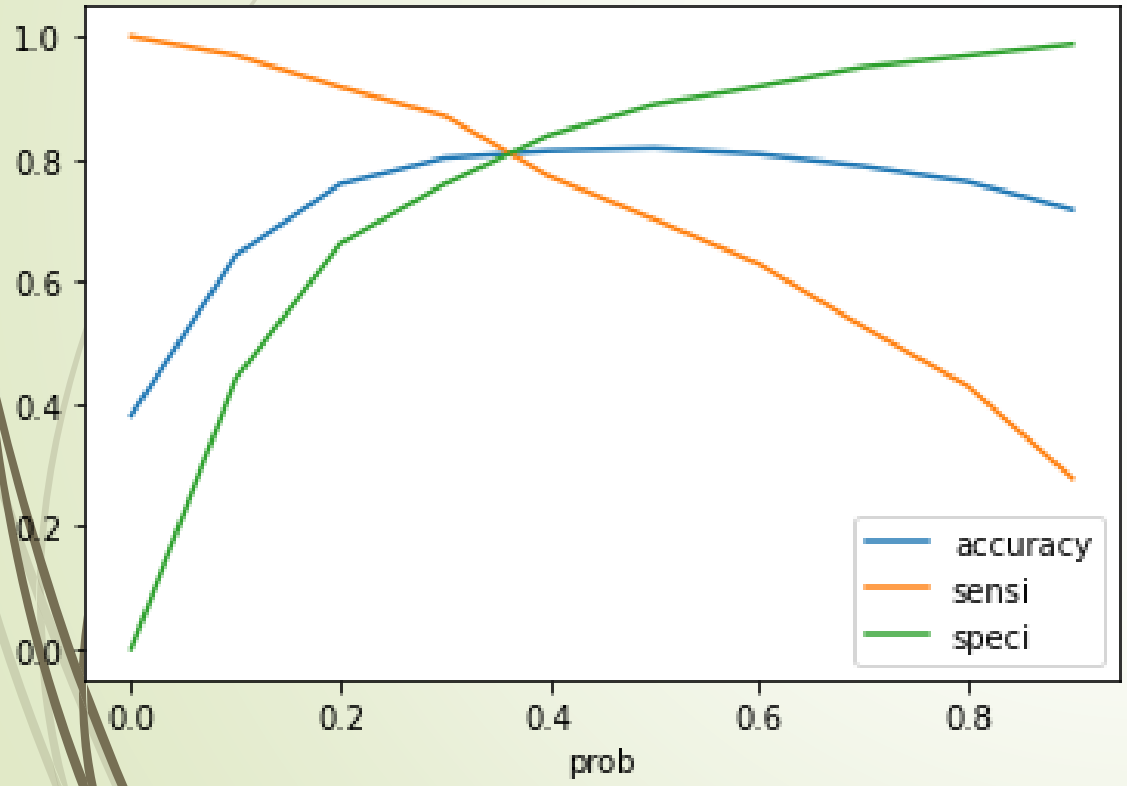
- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity – SMS Sent
- Current Occupation – No Information
- Current Occupation – Working Professional
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity - Unreachable

ROC Curve area is 0.89



Model Evaluation - Sensitivity and Specificity on Train Data Set

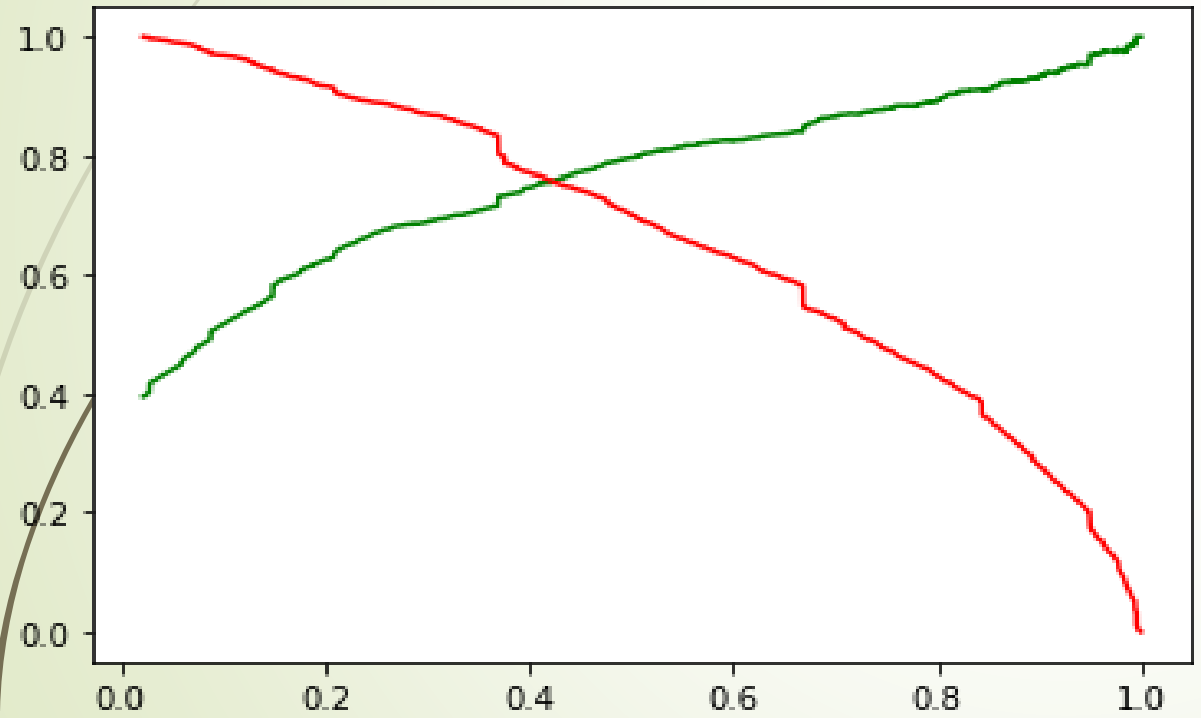
- The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity



- Confusion Matrix
- **3204 705**
- **477 1934**

- **Accuracy - 81%**
- **Sensitivity - 80 %**
- **Specificity - 82 %**
- **False Positive Rate - 18 %**
- **Positive Predictive Value - 73 %**
- **Negative Predictive Value – 87%**

The graph depicts an optimal cut off of 0.42 based on Precision and Recall



- Confusion Matrix
- **3476 433**
- **720 1691**

- **Precision - 80 %**
- **Recall - 70 %**

Model Evaluation – Sensitivity and Specificity on Test Dataset

- Confusion Matrix

- 1340 303
- 199 867

- Accuracy - 81 %
- Sensitivity - 81 %
- Specificity - 82 %

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 81%, 79% and 82% 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.