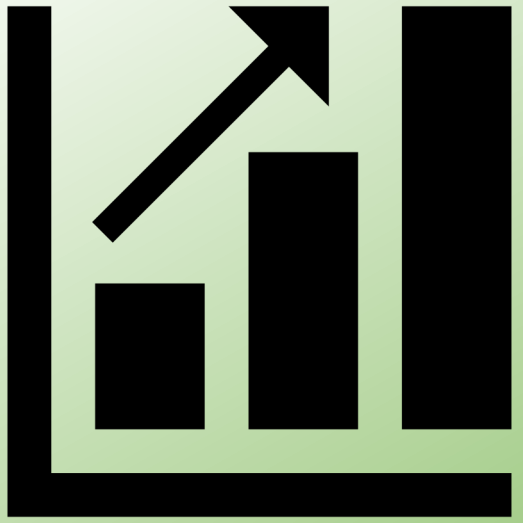
Leading Score Case Study

Submitted By
Arindam Banerjee
Chitra D Nair
Devasish Mahapatro

Problem Statement

Create a model in such a way that the customers with high lead score have higher conversation chance and low lead score have lower conversation chance. The ballpark of the target lead conversion rate is around 80%. Also the model should be able to adjust if the company's requirement changes in near future.

STRATEGY



1. Source the data for analysis and clean it.

2. Exploratory Data Analysis

3. Feature Scaling

4. **Splitting the data into test and Train Dataset**
Building a Logistic Regression model and calculate Lead Score.

5. Evaluating the model by using different metrics

6. Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Techniques Followed To Solve The Problem

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

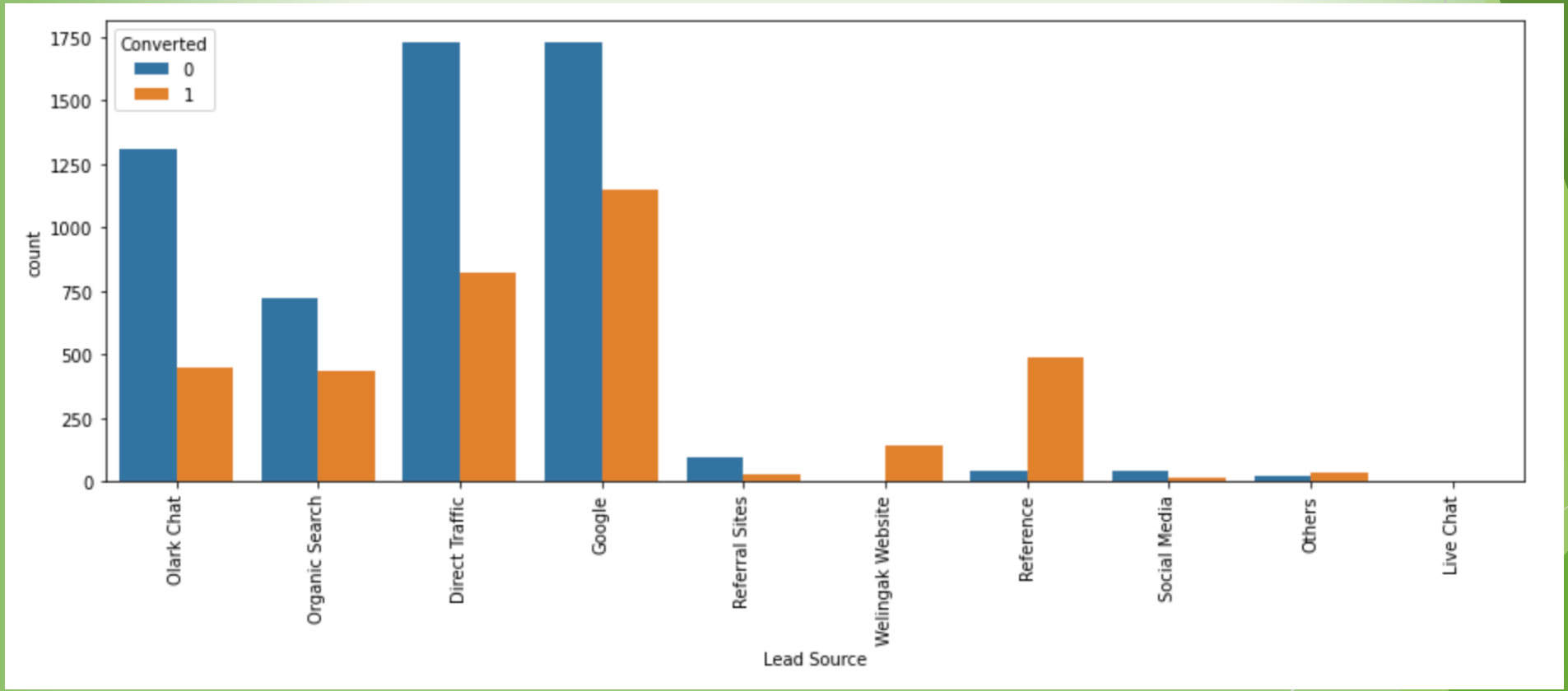
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

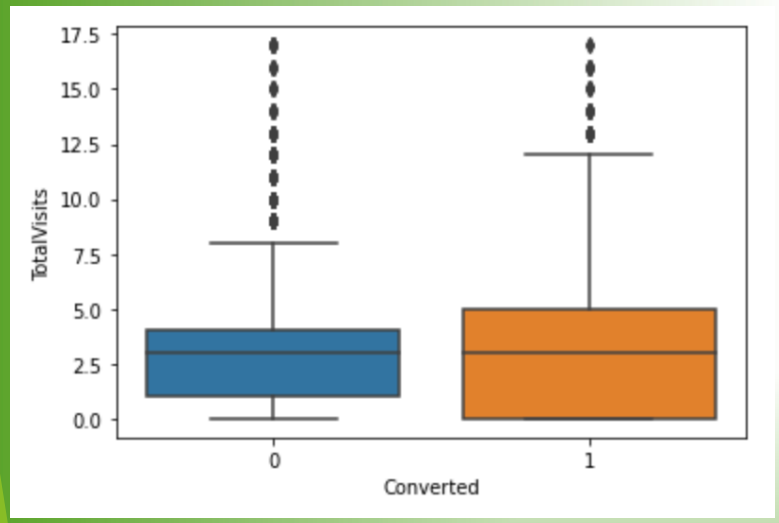
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

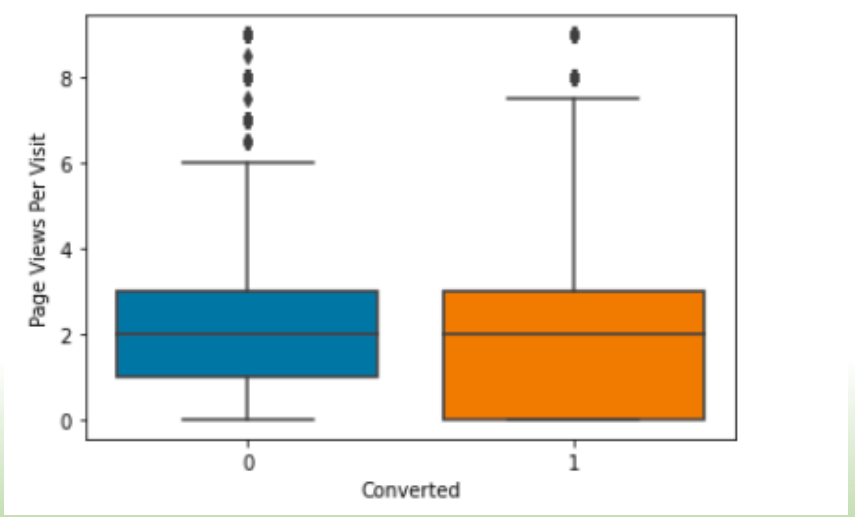
Exploratory Data Analysis

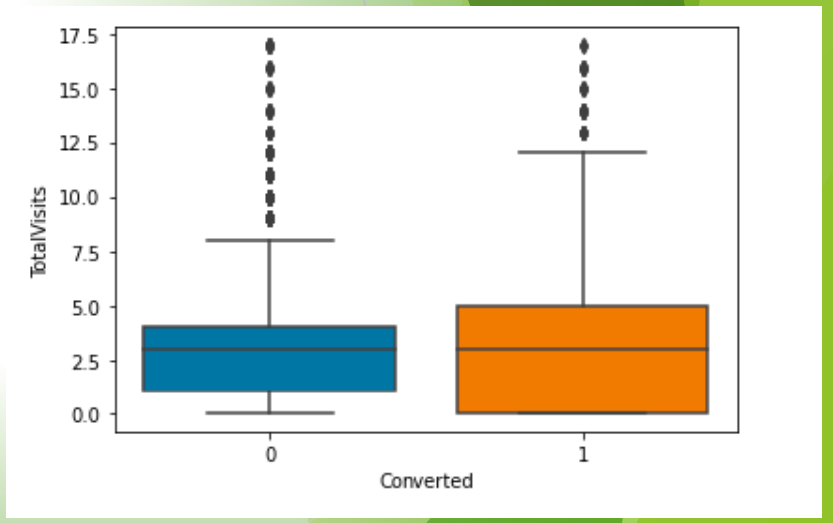
Major Conversion in the lead score is from Google



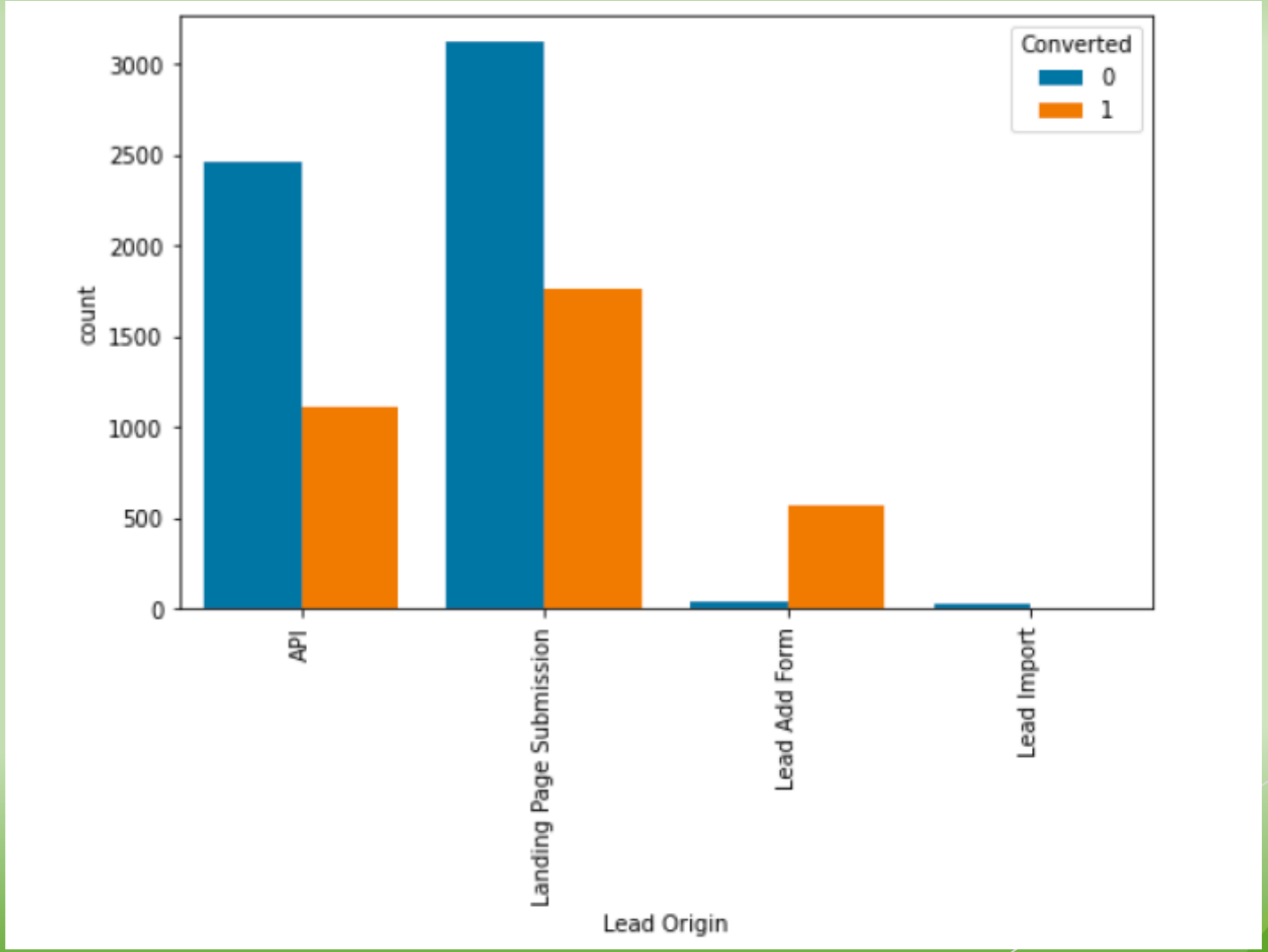
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit.



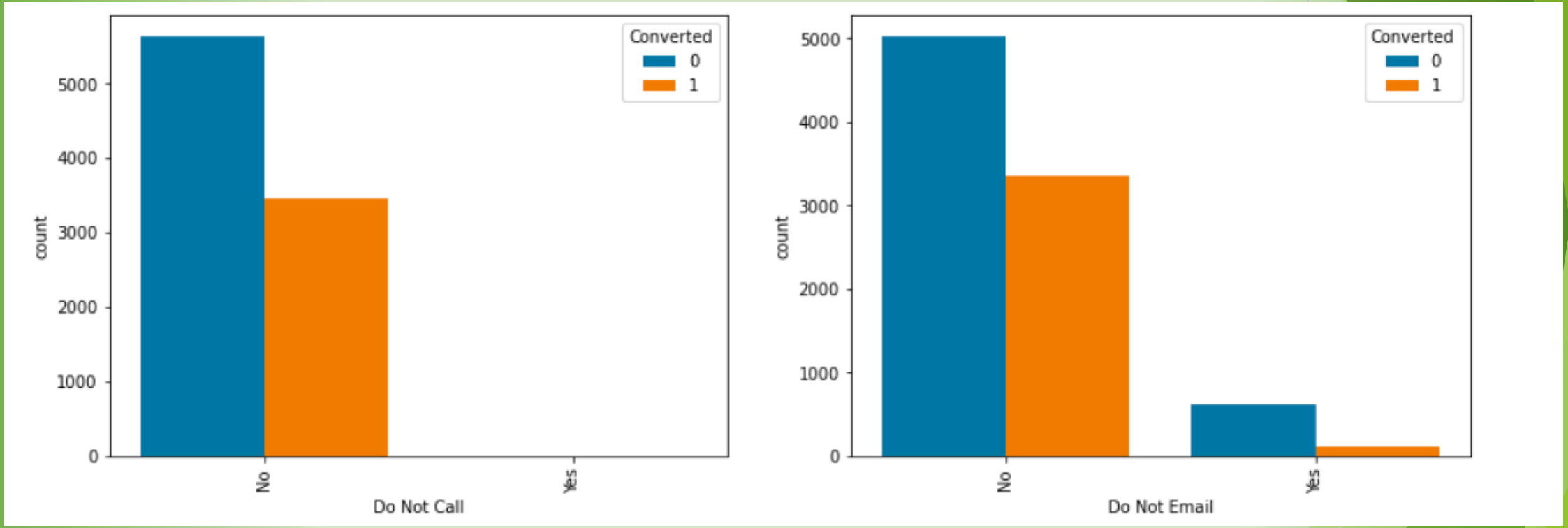




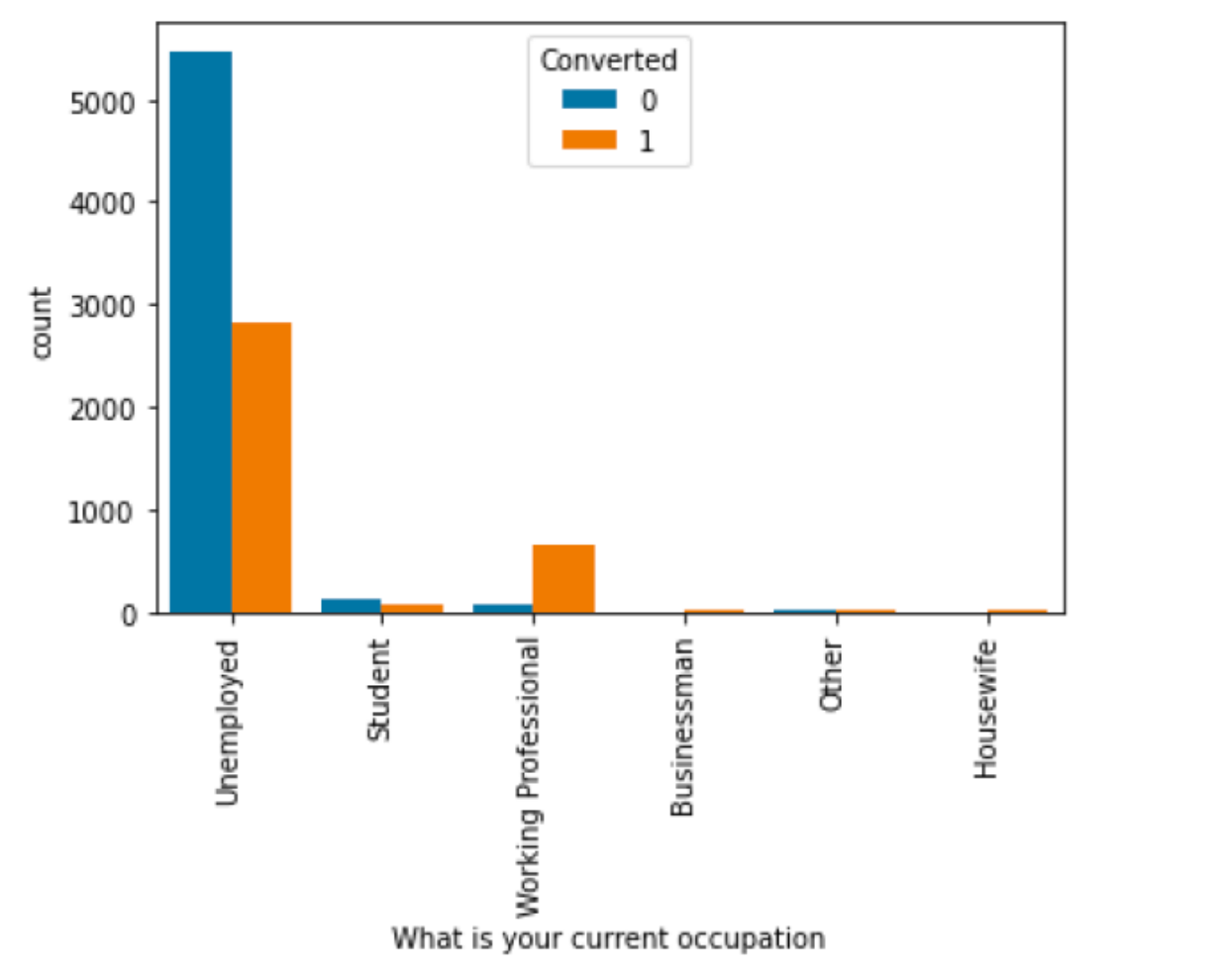
In Lead Origin, maximum conversion happened from Landing Page Submission



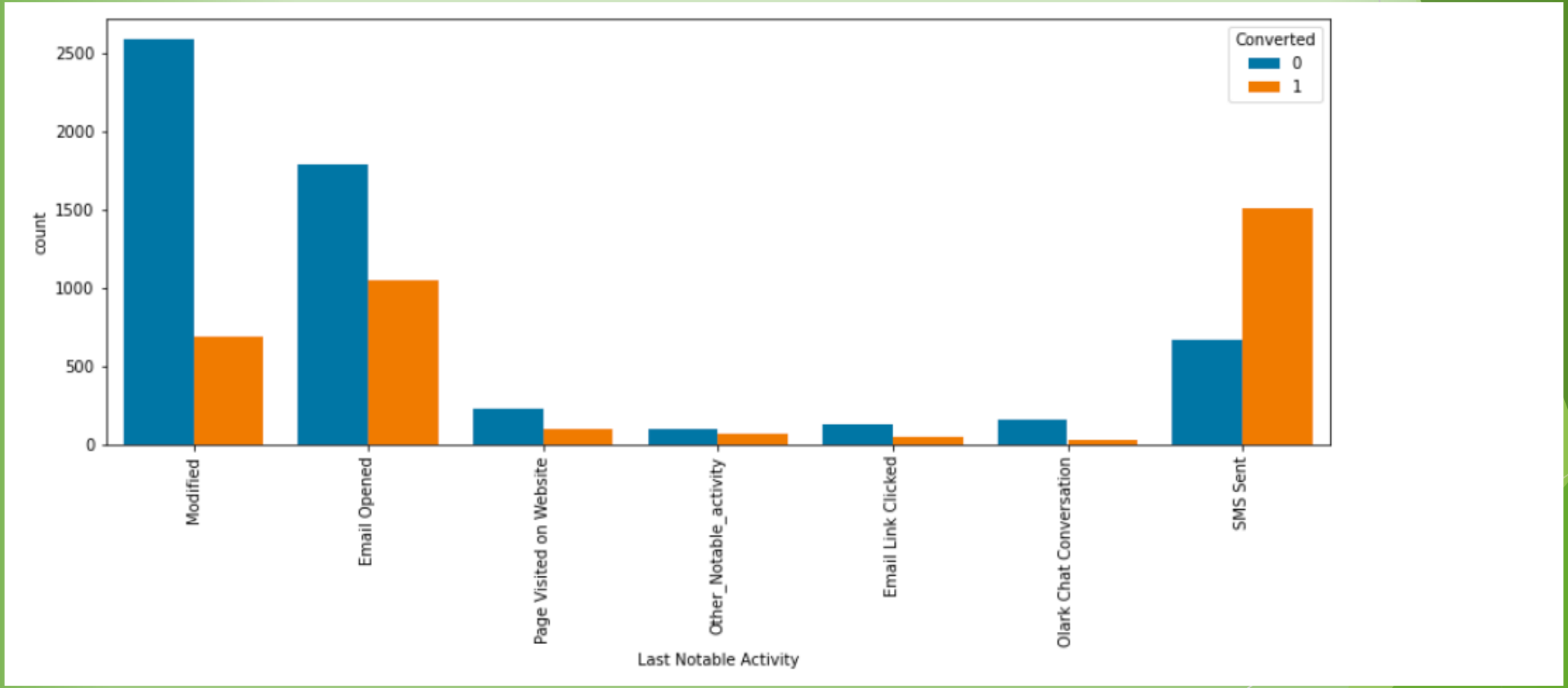
Major conversion has happened from Emails sent and Calls made



More conversion happened with people who are unemployed



Last Activity value of SMS Sent' had more conversion.

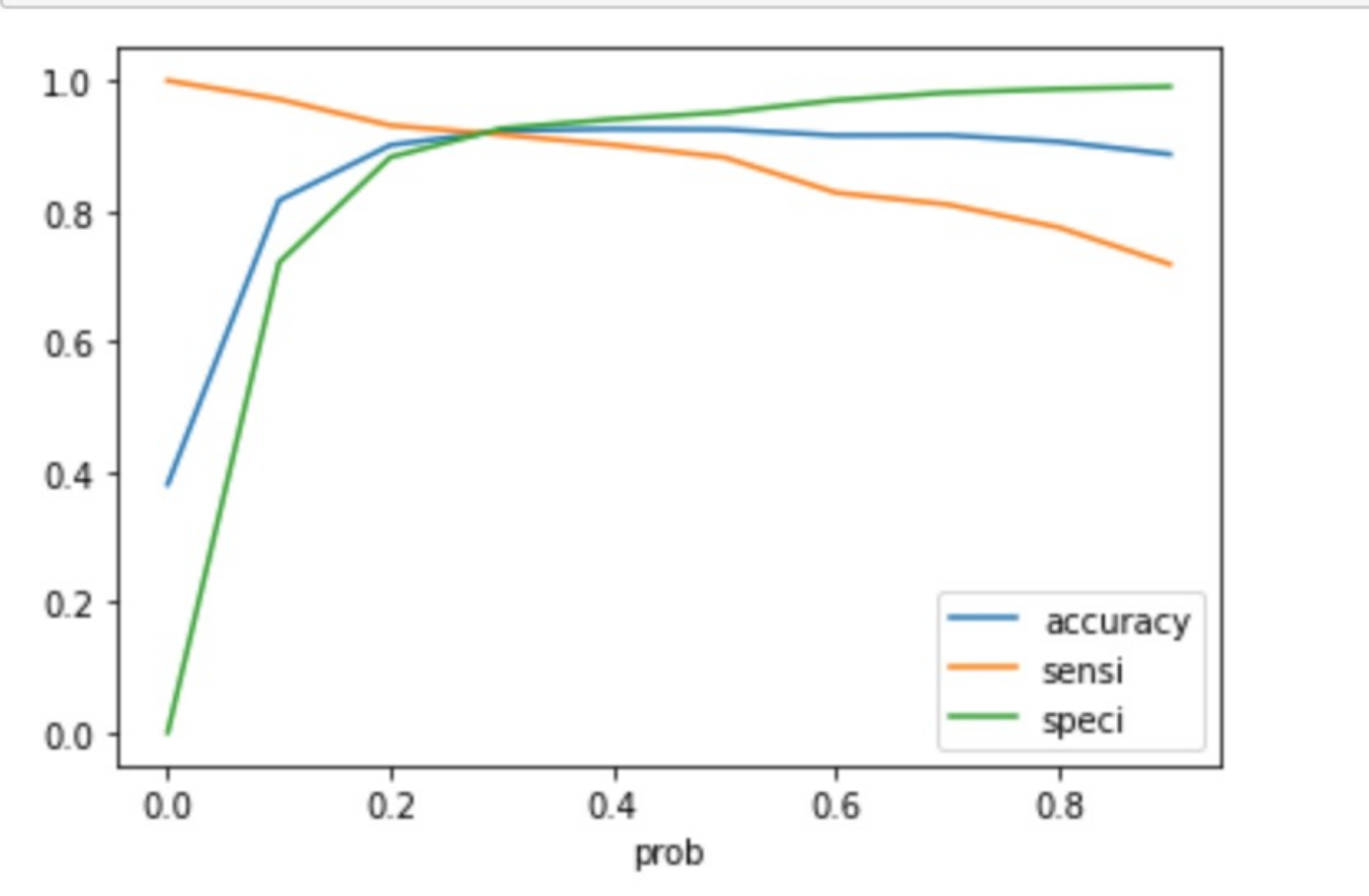


Variables Impacting the Conversion Rate

- 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

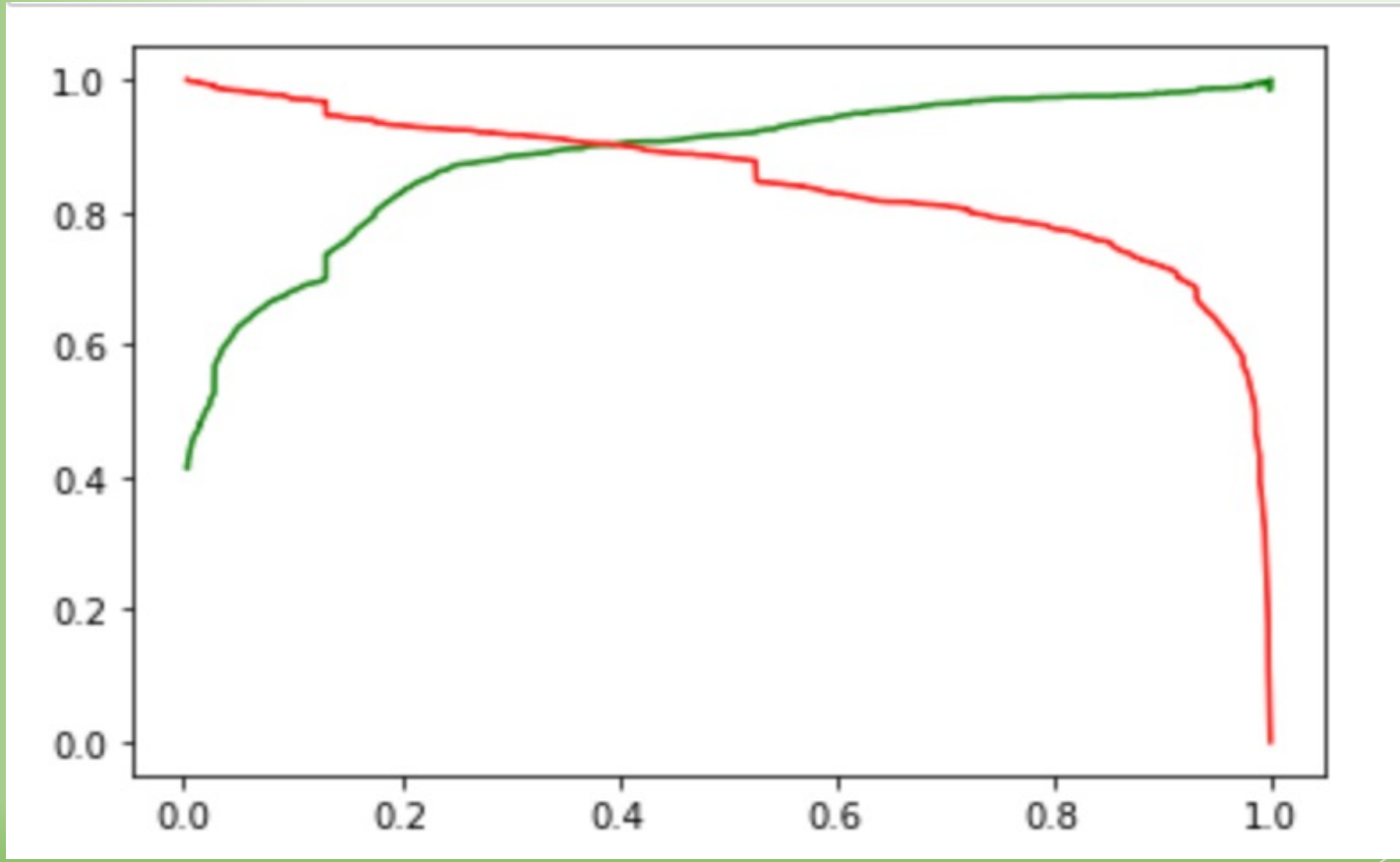
Model Evaluation - Sensitivity and Specificity on Train Data Set

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



Model Evaluation – Precision and Recall on Train Data Set

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



Model Evaluation Sensitivity and Specificity on Test Dataset

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

Confusion Matrix

1581	95
94	916

- Accuracy = 93%
- Sensitivity = 91%
- Specificity = 94%

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 92%, 91% and 94% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 79% in test set.
- The top 3 variables that contribute for lead getting converted in the model are:
  - Total time spent on website
  - Lead Add Form from Lead Origin
  - What is your current occupation-Working Professional.
- Hence overall this model seems to be good.