Lead Scoring Case Study

SUBMITTED BY :

VIKASH KUMAR

SINDHU P

SIBA PRASAD SABAT

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:

- X Education needs help in selecting the most promising leads, i.e. the leads that are 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 Test data based on the Sensitivity and Specificity Metrics.

Solution Methodology:

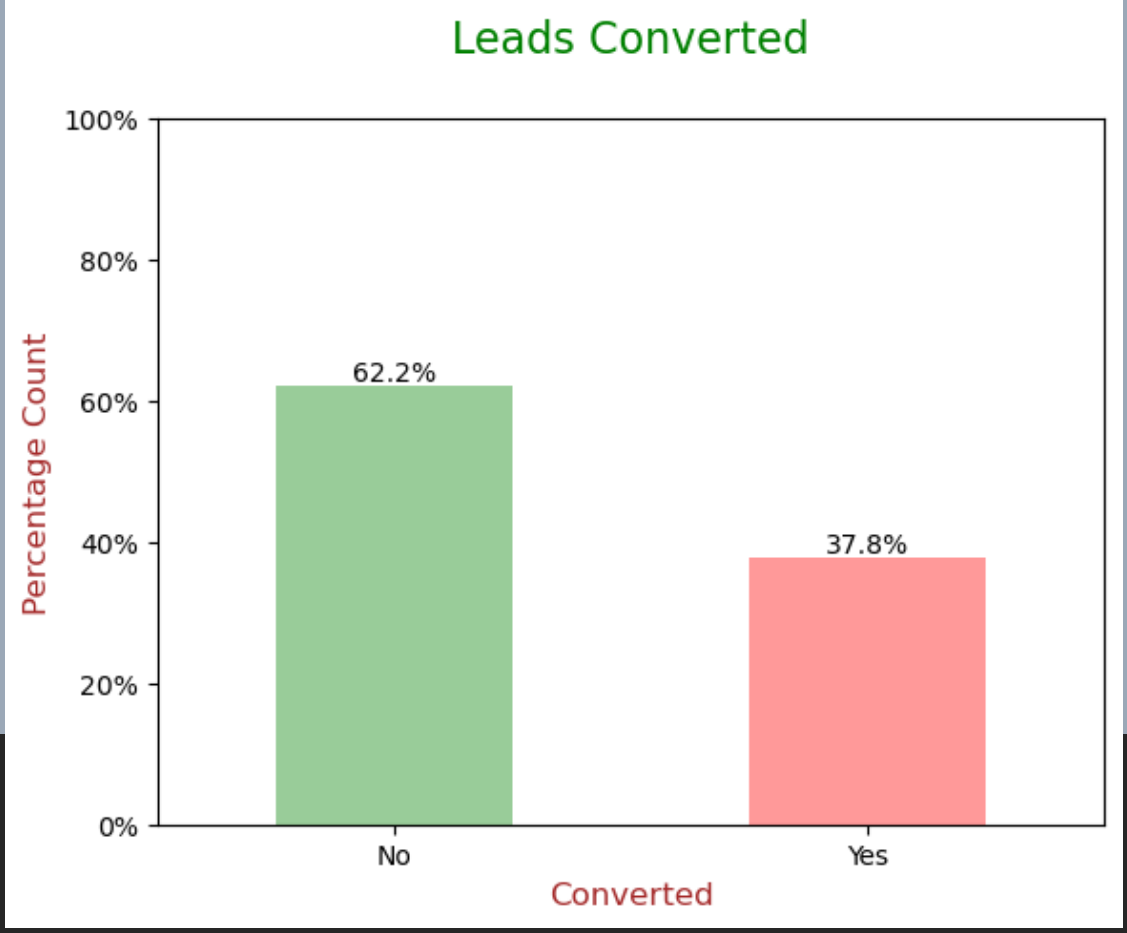
Data Cleaning, Preparation & Sourcing

Feature Scaling and Splitting Train & Test Sets

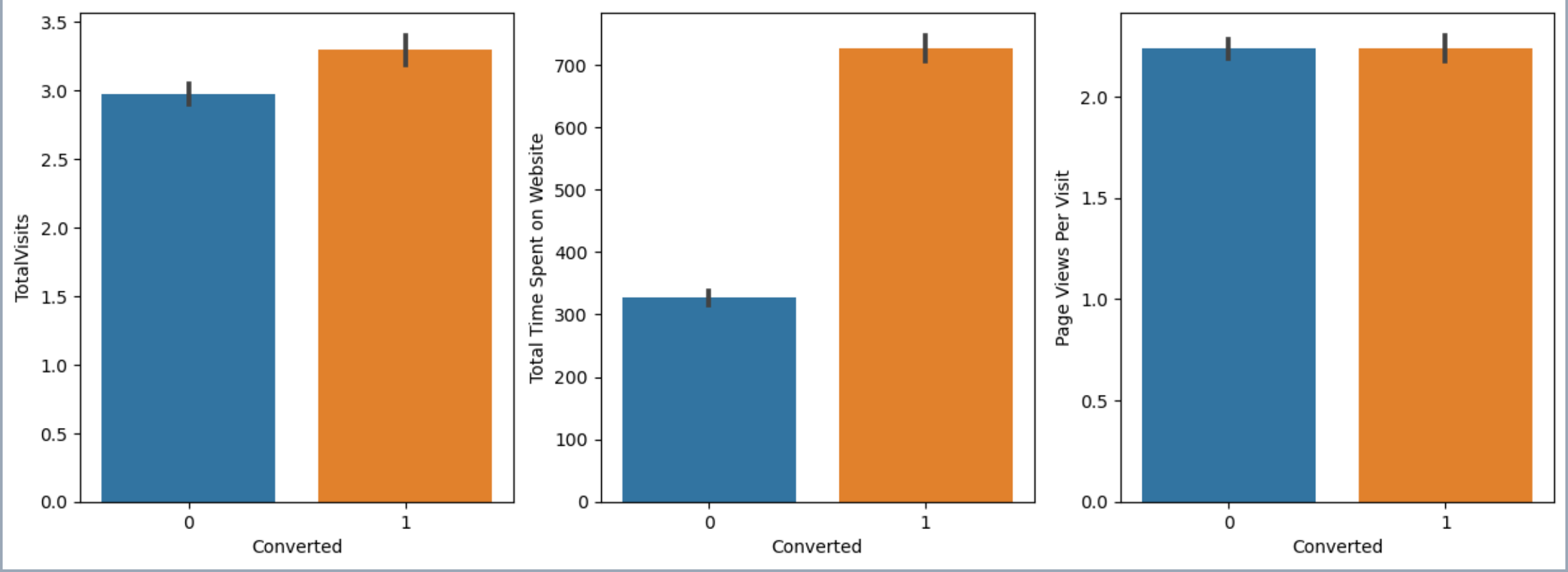
Model Building

Results

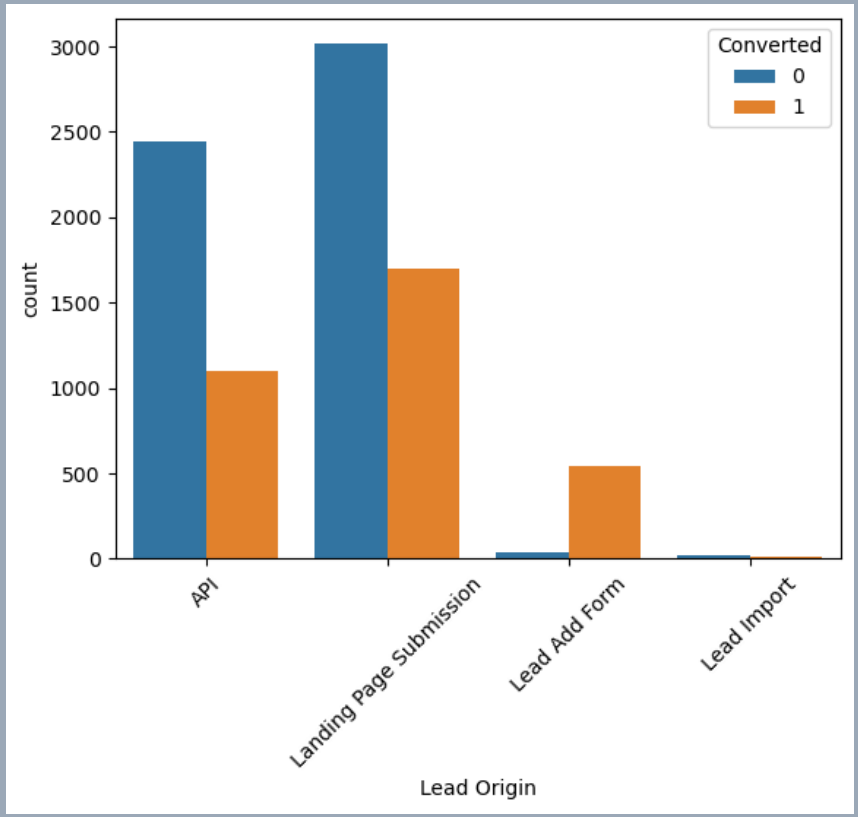
Exploratory Data Analysis



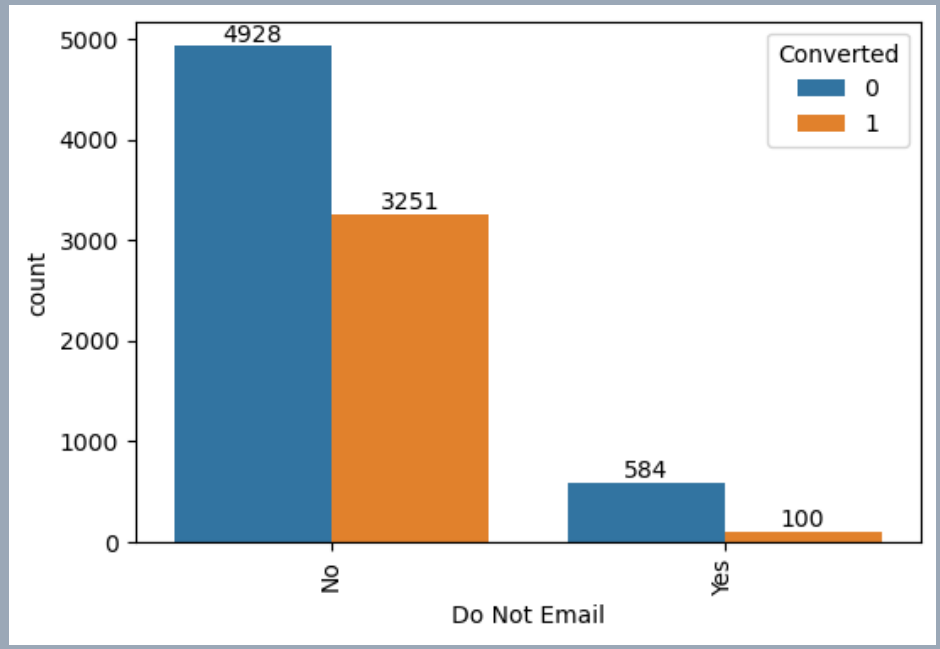
We have around 38% lead conversion rate.

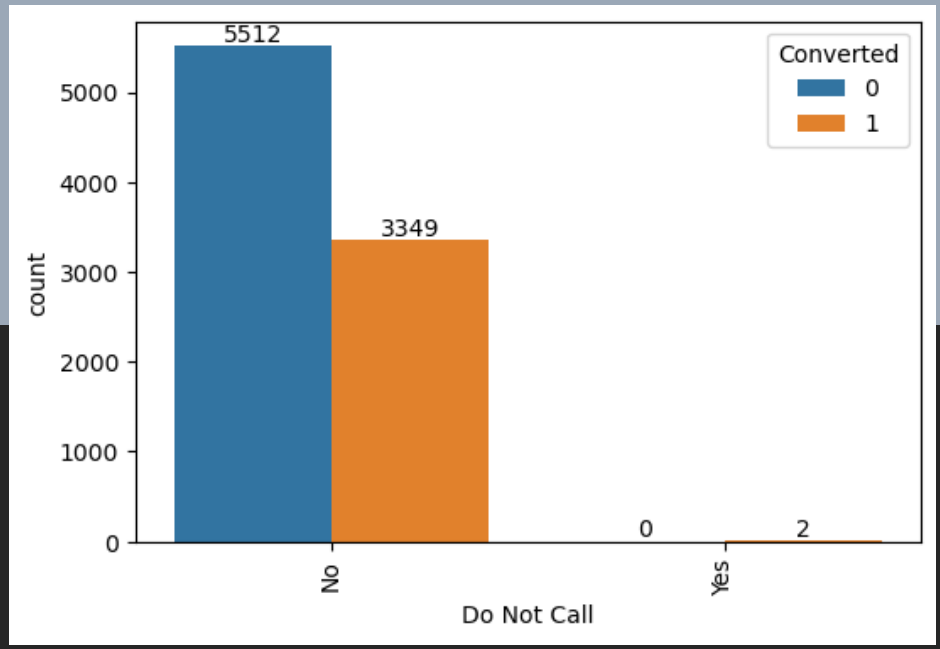


The Conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit.

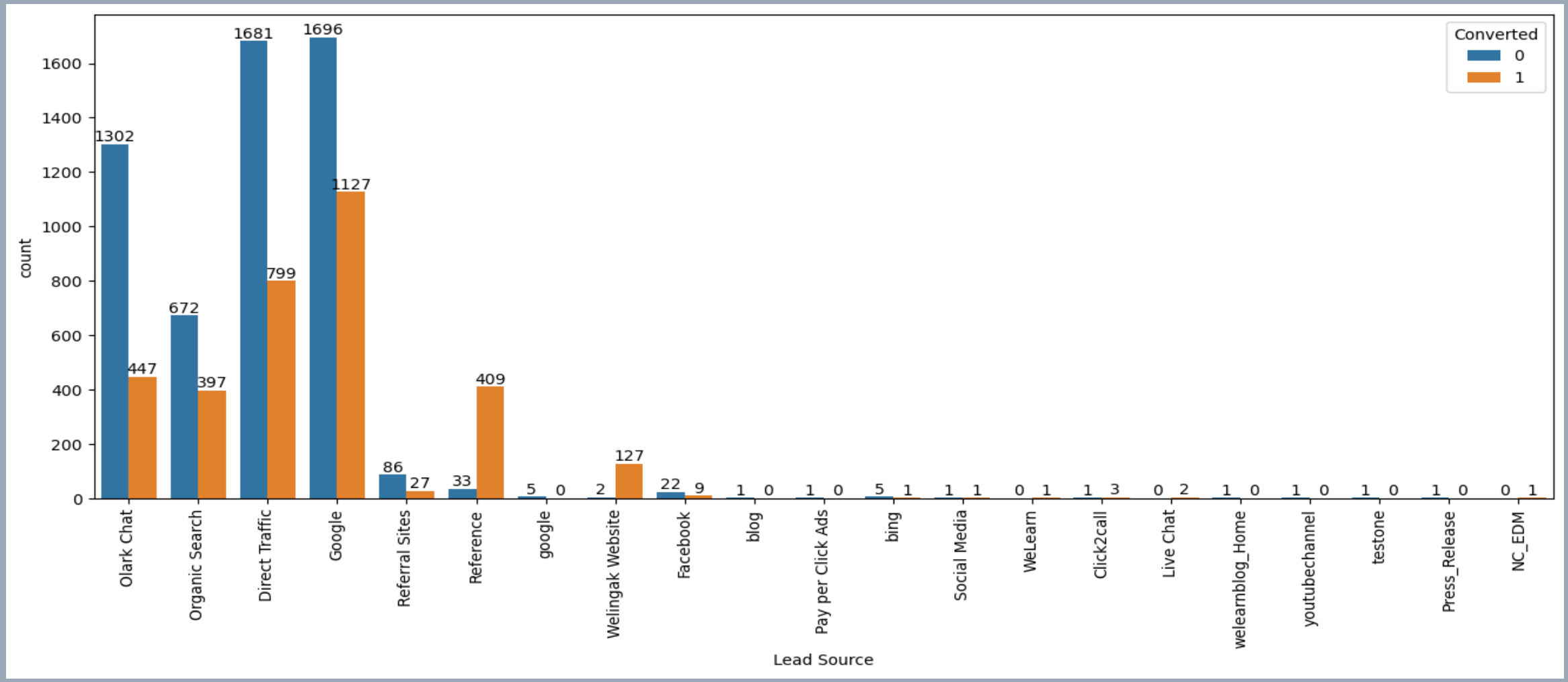


Major conversion has happened from Emails sent and Calls made

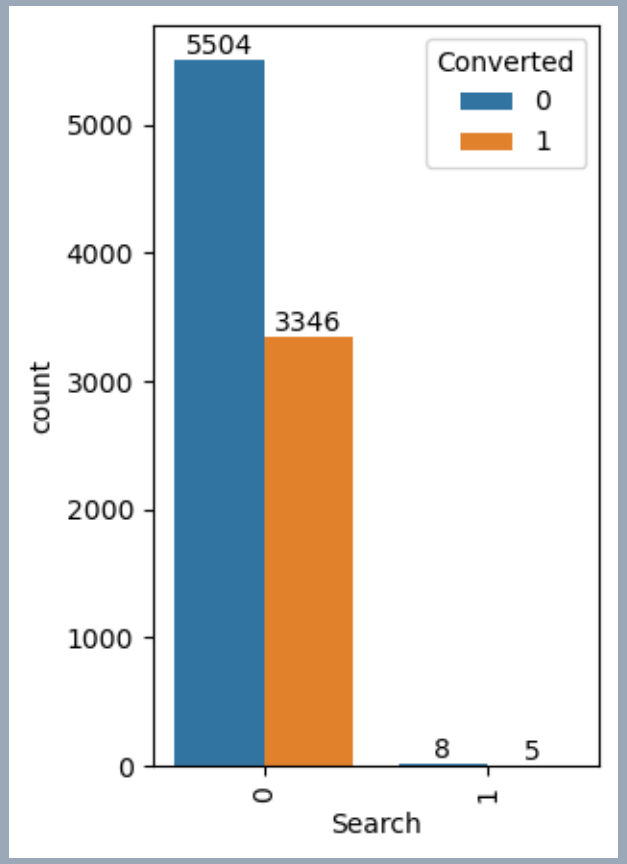


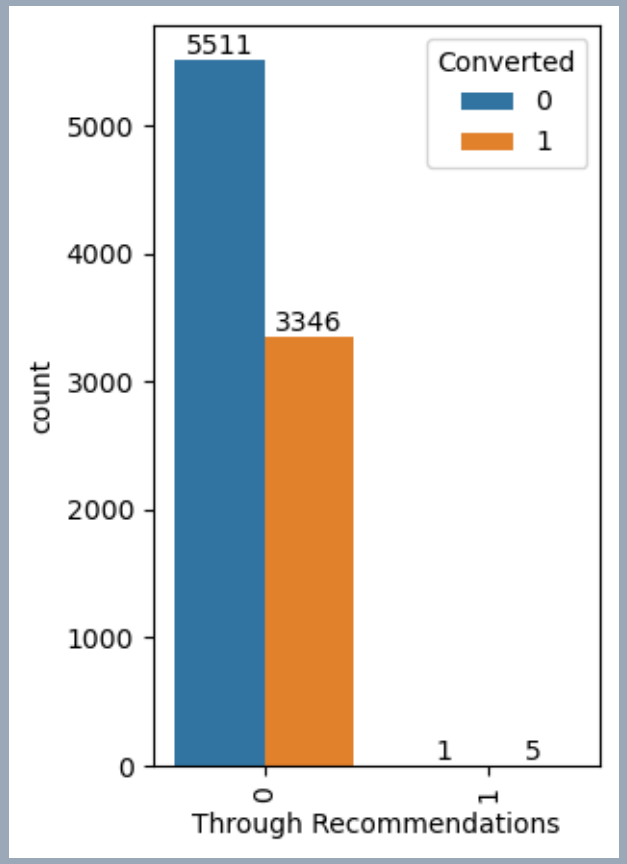


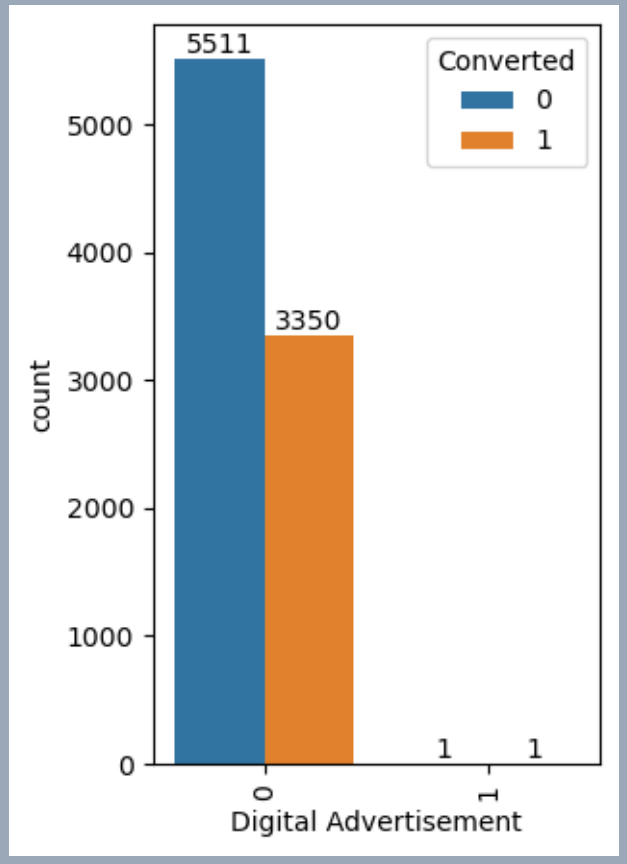
In Lead Origin, maximum conversion happened from Landing Page Submission



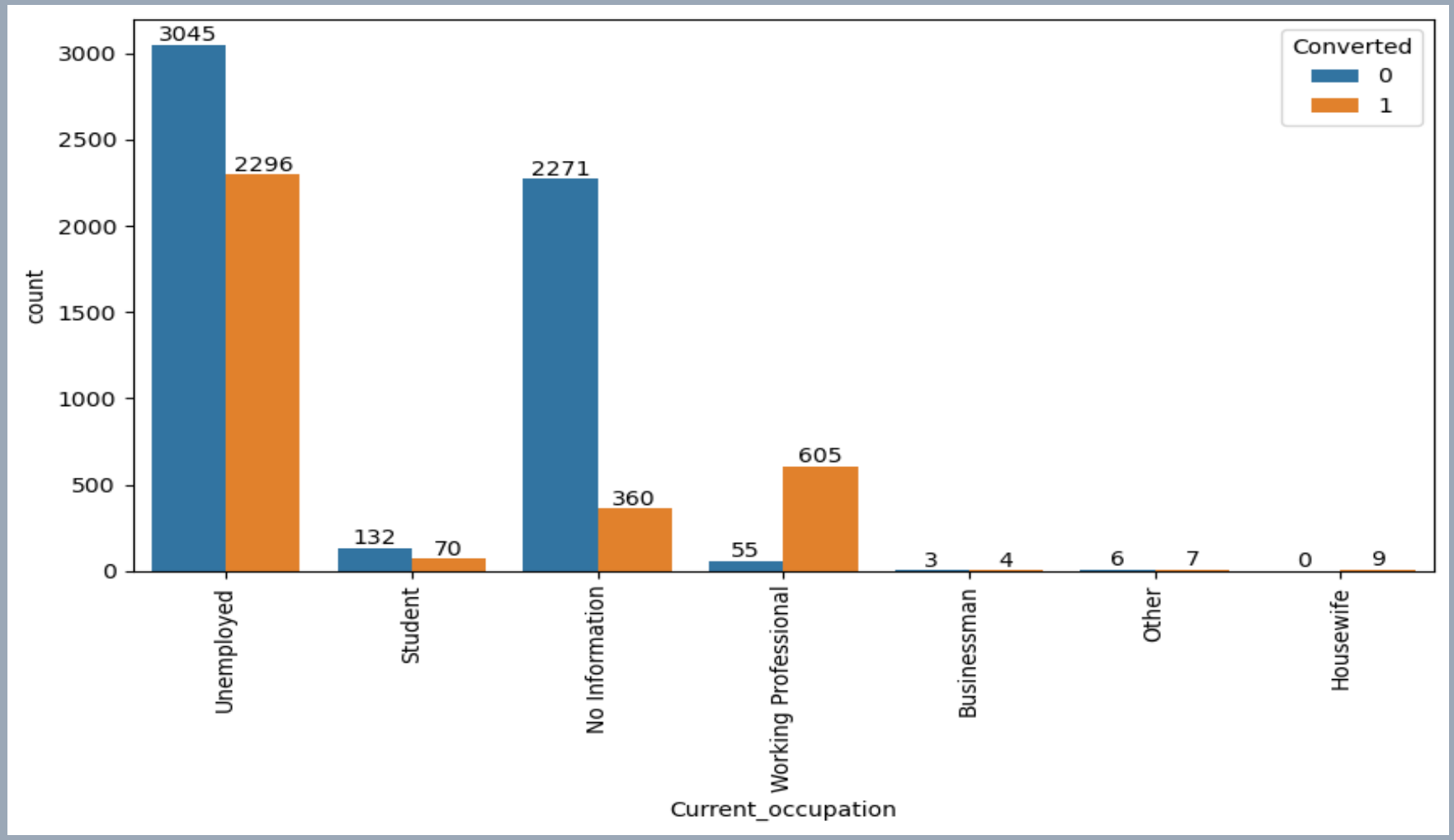
Major conversion in the lead source is from Google



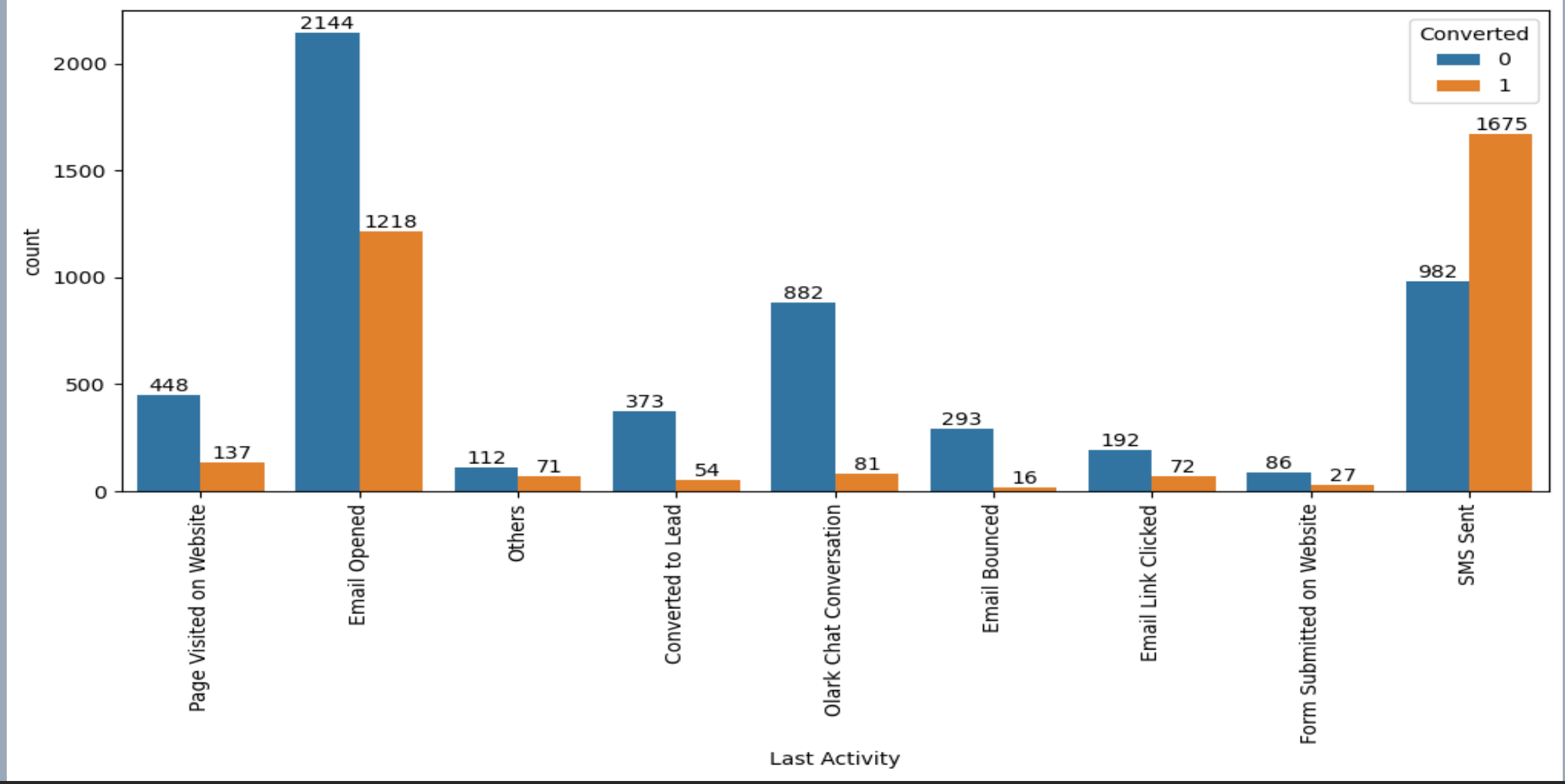




Not much impact on conversion rates through Search, digital advertisements and through recommendations.



More conversion happened with people who are unemployed



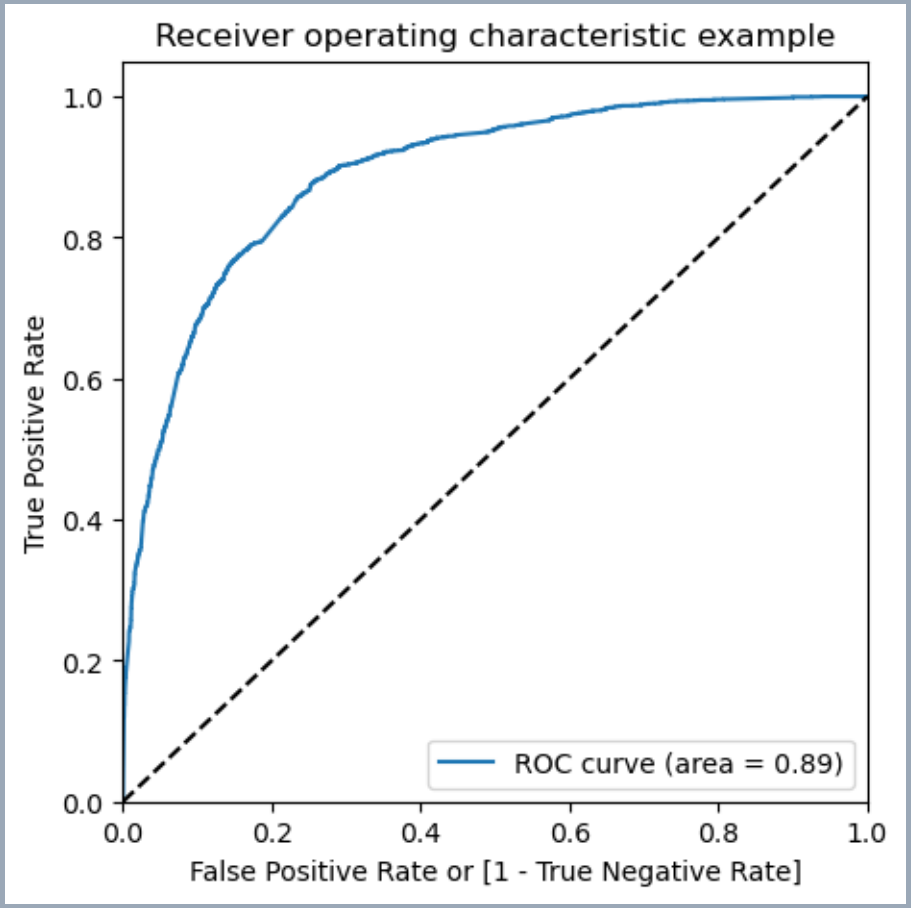
Last Activity value of SMS Sent' had more conversion.

Variables impacting the conversion rate:

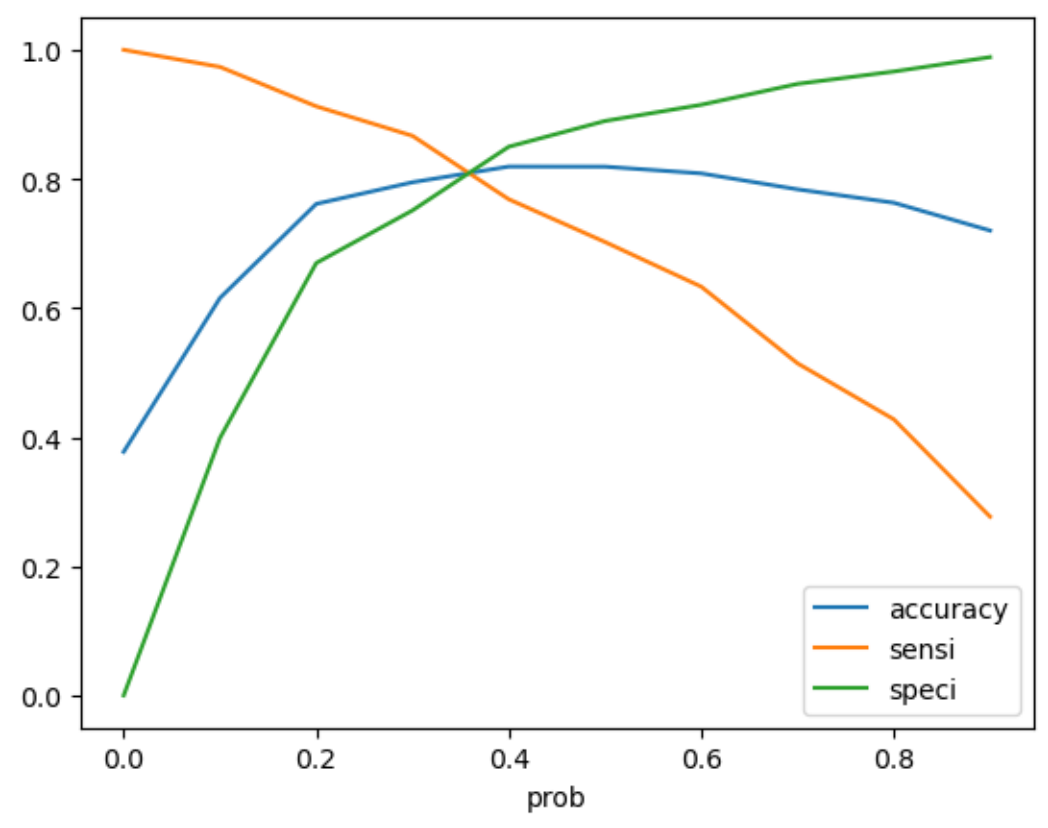
- Do not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity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

ROC Curve

The model has 0.89 area under the curve.
Hence it’s a good model



Model Evaluation - Sensitivity and Specificity on Train Data Set



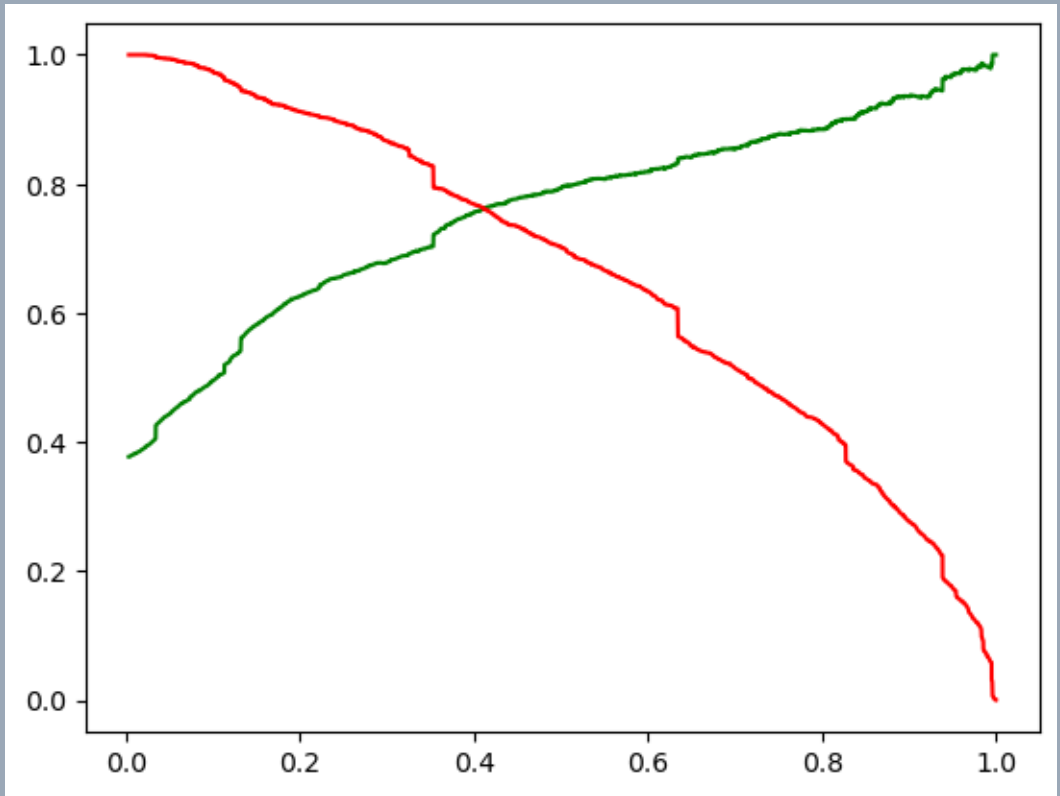
Confusion Matrix

3040	823
400	1941

- Accuracy – 80.3%
- Sensitivity – 82.9 %
- Specificity – 78.7 %
- False Positive Rate - 21 %
- Positive Predictive Value - 74 %
- Negative Predictive Value – 88%

The graph depicts an optimal cut off of 0.35 based on Accuracy, Sensitivity and Specificity

Model Evaluation – Precision and Recall on Train Data Set



Confusion Matrix

3437	426
697	1644

- Precision - 79%
- Recall – 70 %

The graph depicts an optimal cut off of 0.40 based on the precision and recall

Model Evaluation – Sensitivity and Specificity on Test Dataset

Confusion Matrix

1301	348
193	817

- Accuracy – 79.7 %
- Sensitivity – 70.2 %
- Specificity – 88.9 %

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 79.7%, 70.2% and 88.9% 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 80.2% (in train set) and 80.9% (in test set)
- The top 3 variables that contribute for lead getting converted in the model are:
  - Lead source from Welingak Website
  - Lead source from Reference
  - Total time spent on the website
- Hence overall this model seems to be good.

Thank You