Lead Score Case Study

Submitted by:

Karina Sarkar

Aparna Salunkhe

Problem Statement:

X Education sells online courses to industry professionals. The business advertises its classes via a number of websites and search engines, including Google. Upon accessing the website, these individuals may pursue the available courses, complete the course registration form, or view some videos.

When someone fills out a form with their phone number or email address, they are categorized as leads. Additionally, the business receives leads from previous recommendations.

After obtaining these leads, sales team members begin calling, emailing, and so on. While most leads do not convert during this process, others do. At X education, the lead conversion rate is typically 30%.

Business Goal:

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 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%.

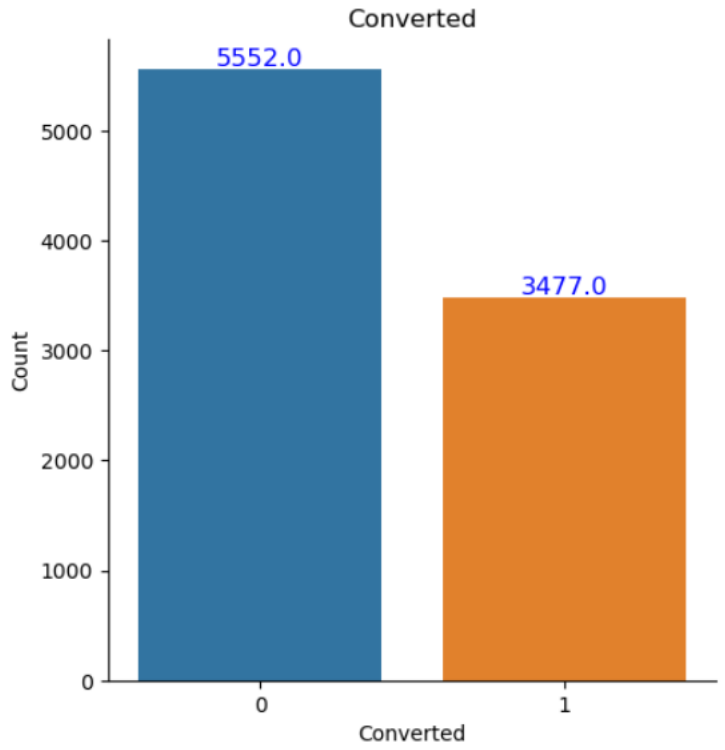
Strategy:

- Sourcing the data for analysis
- Cleaning the data
- Data Preparation
- Exploratory data analysis (EDA)
- Feature Scaling
- Splitting into train-test data
- Building Logistic Regression model
- Calculating Lead Score
- Model Evaluation by calculating Specificity, Sensitivity, Precision and recall
- Applying the Final model to test data

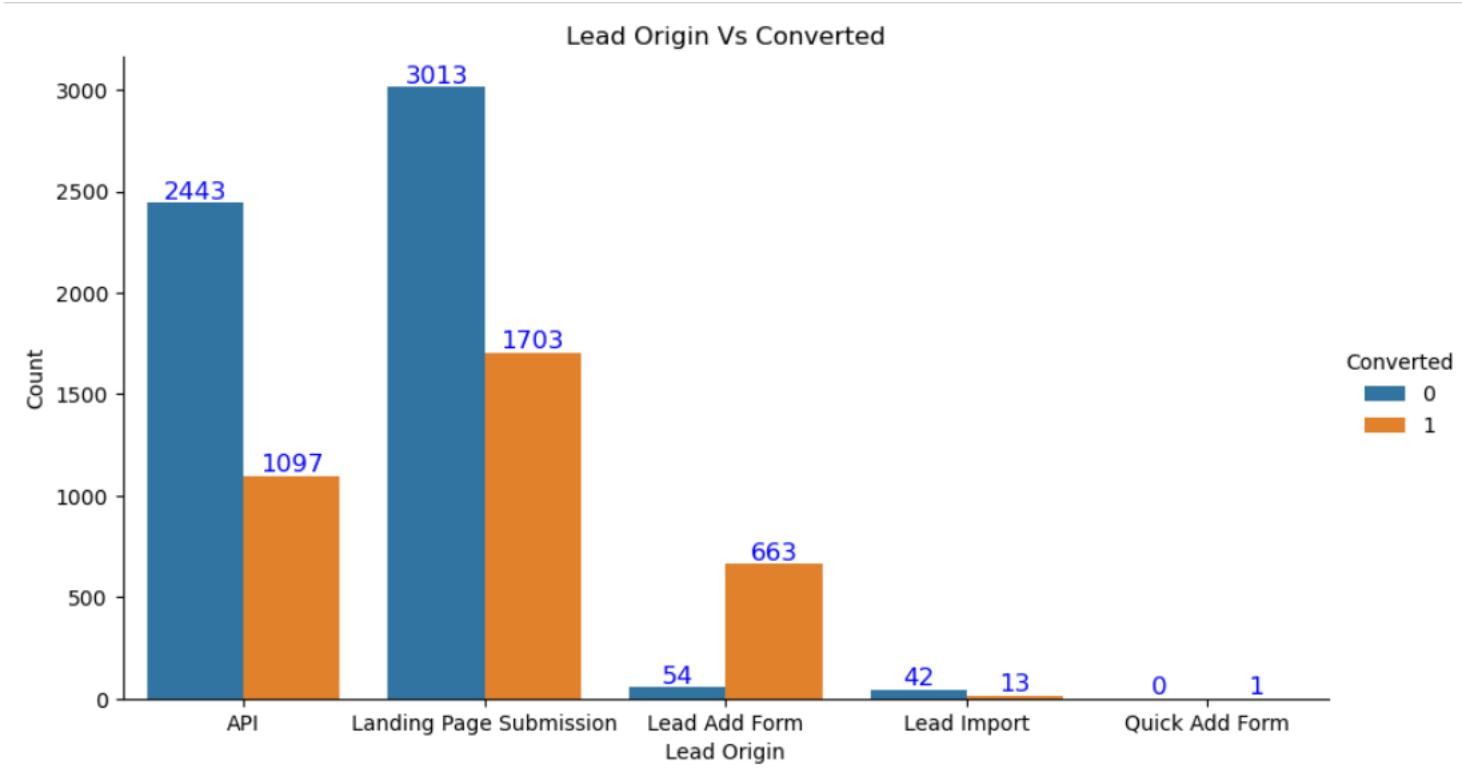
Exploratory data Analysis:

After thorough data analysis, we observed that:

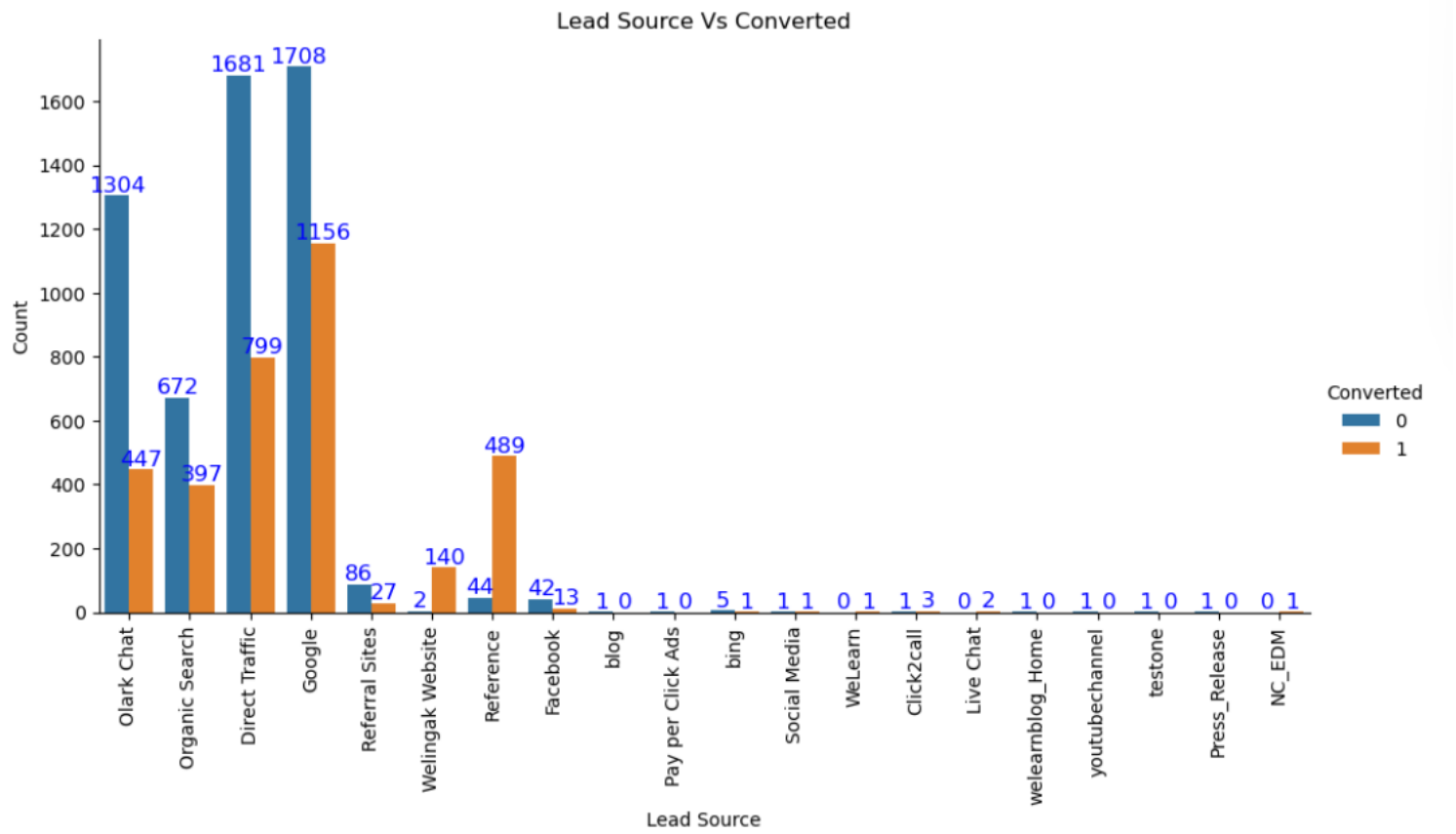
The initial conversion rate is around 38%



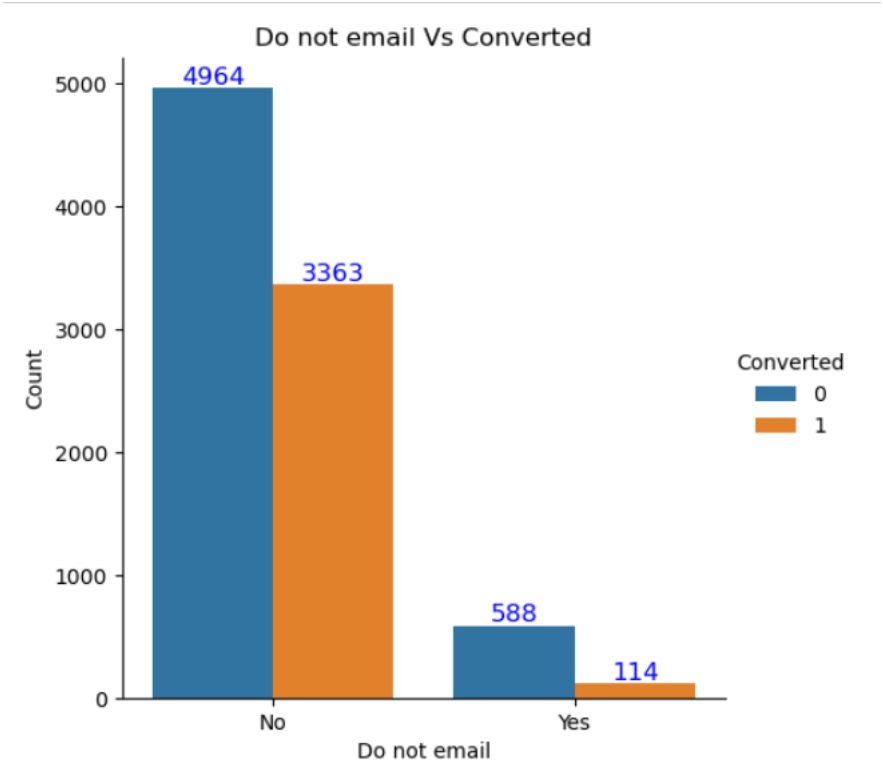
Maximum conversion happened from Landing Page Submission. Also there was only one request from quick add form which got converted.

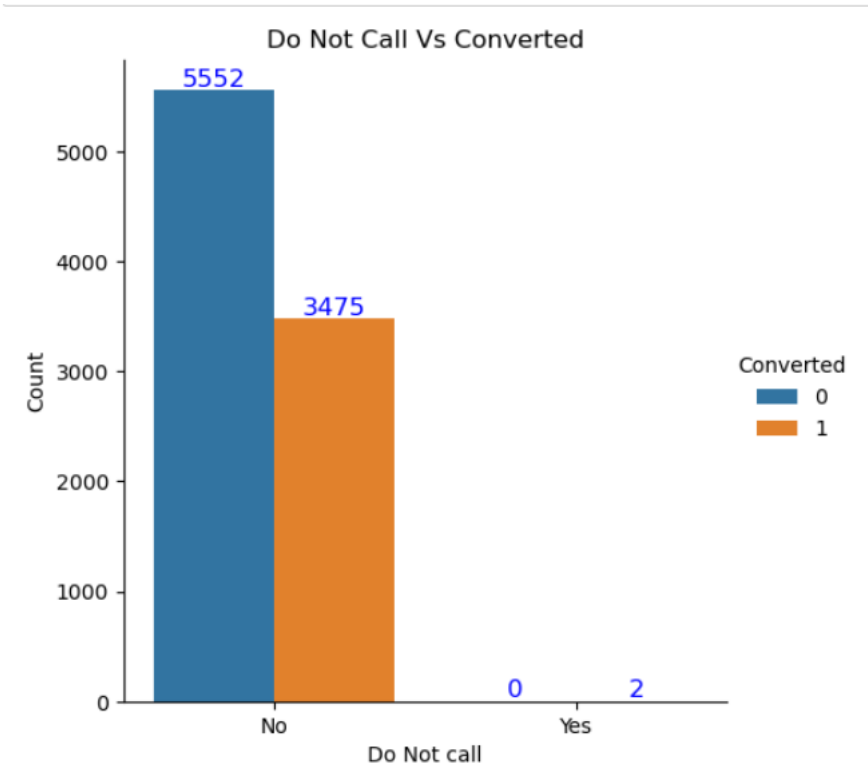


We received major lead conversions from Google

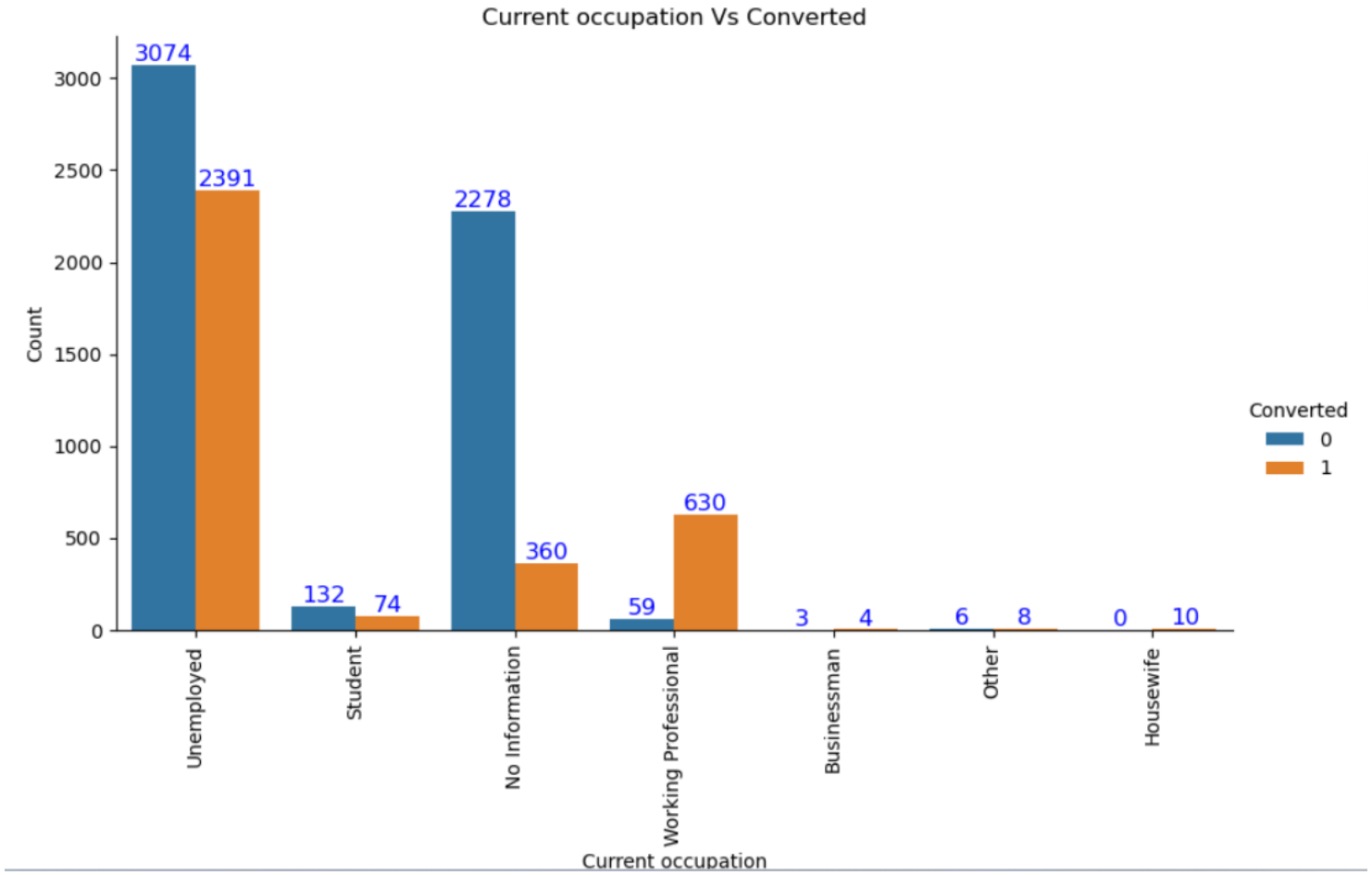


Maximum lead conversions occurred when emails were sent and calls were made

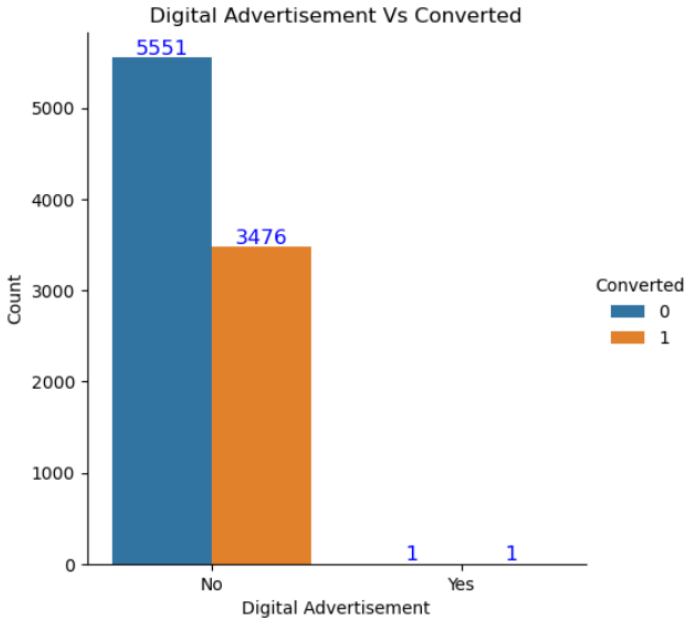


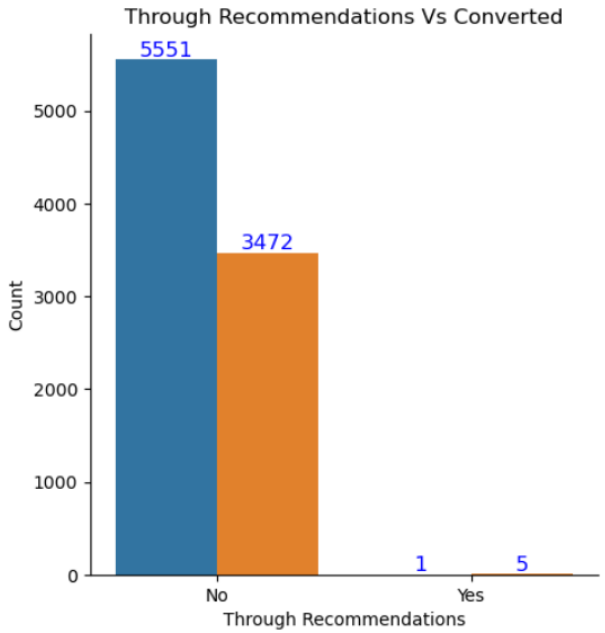


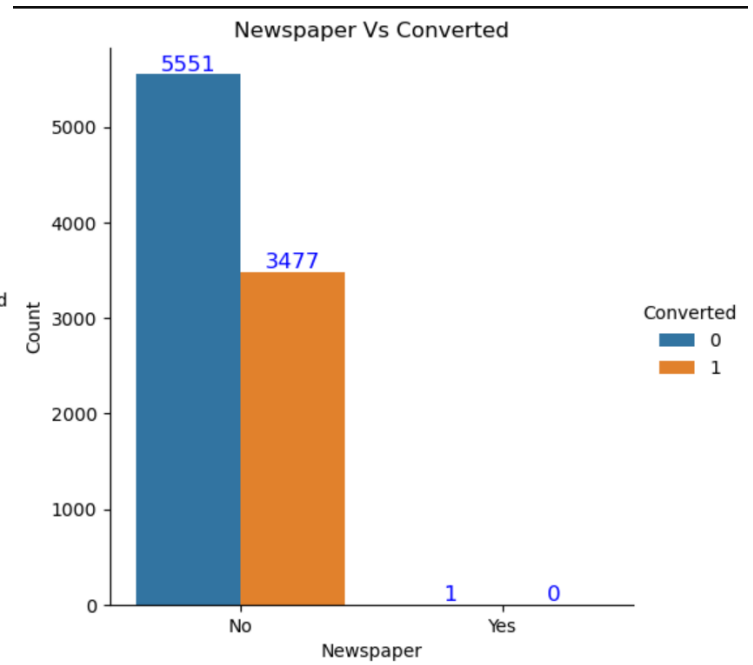
Most people who were converted were unemployed



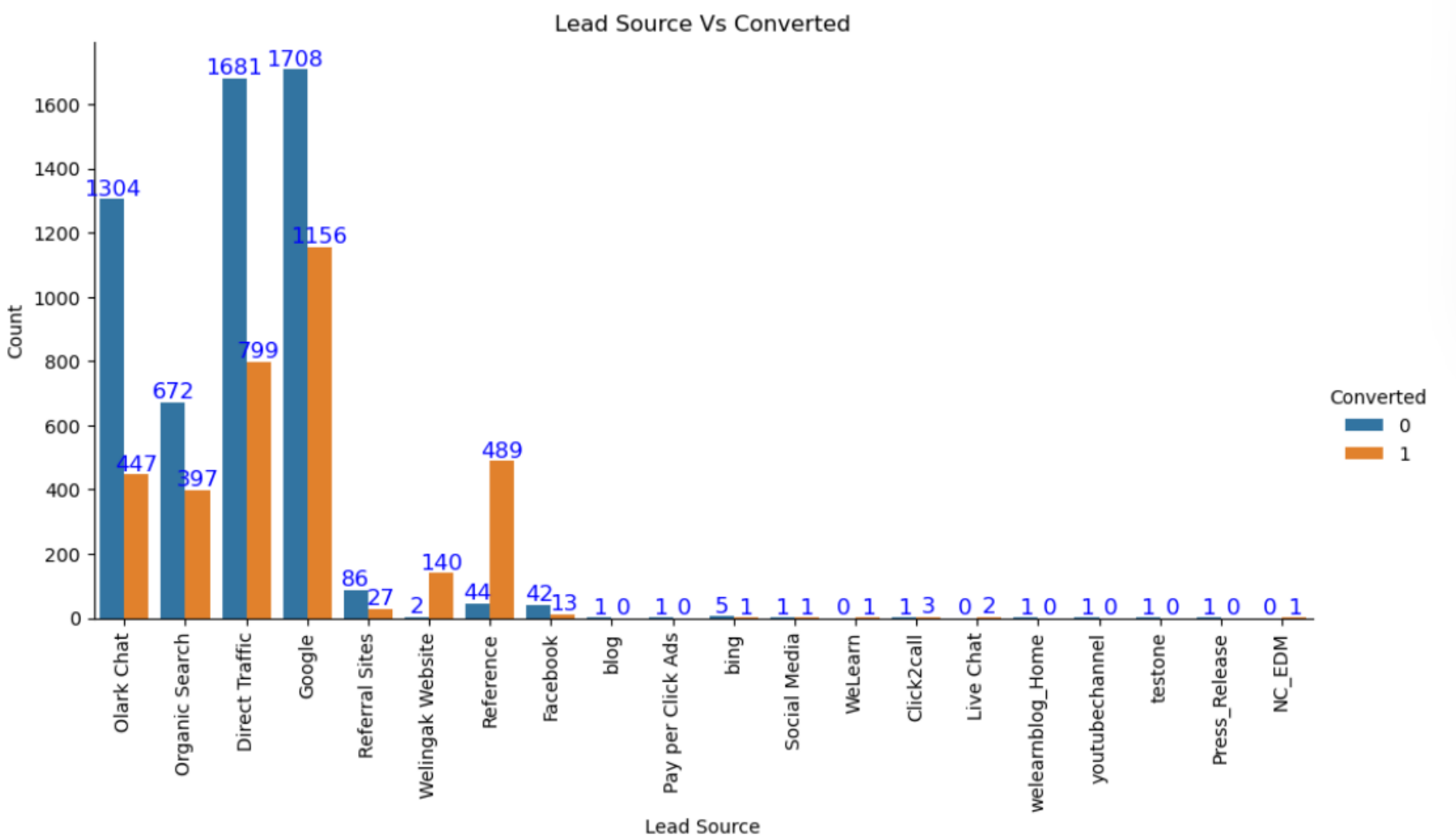
Not much impact happened through newspapers, digital advertisements or recommendations







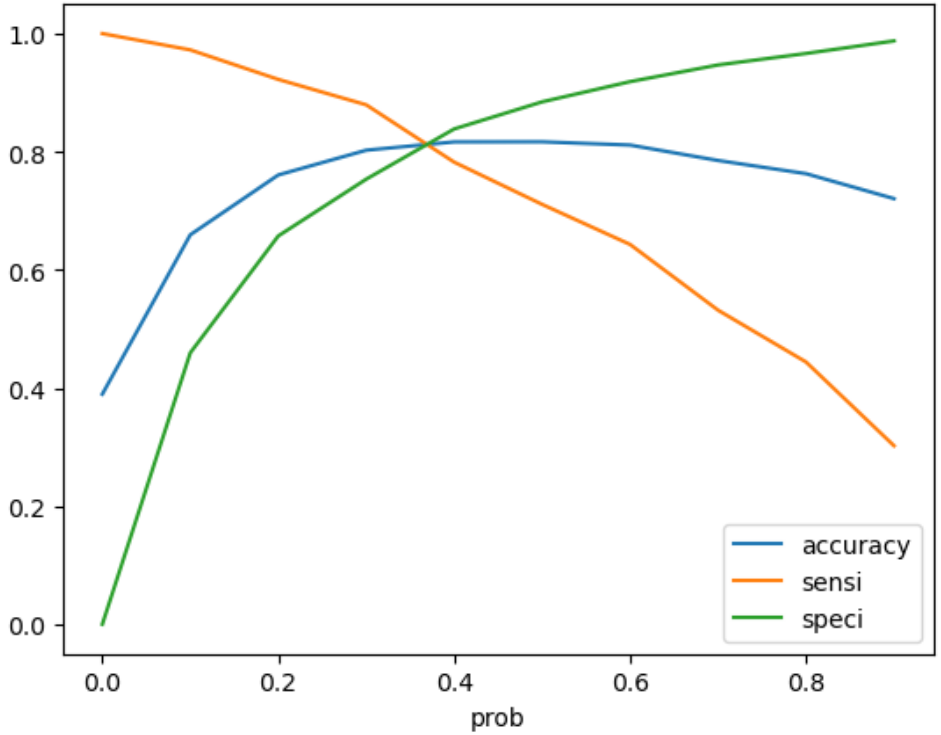
Last Activity value of SMS Sent had more conversion



Variables having impact on conversion rate:

Do Not Email
TotalVisits
Total Time Spent on Website
Page Views Per Visit
LeadOrigin_Lead Add Form
LeadSource_Olark Chat
LeadSource_Welingak Website
LastActivity_Email Opened
LastActivity_Olark Chat Conversation
LastActivity_SMS Sent
CurrentOccupation_No Information
CurrentOccupation_Working Professional
LastNotableActivity_Had a Phone Conversation
LastNotableActivity_Modified
LastNotableActivity_Unreachable

Model Evaluation - Sensitivity and Specificity on Train Data Set



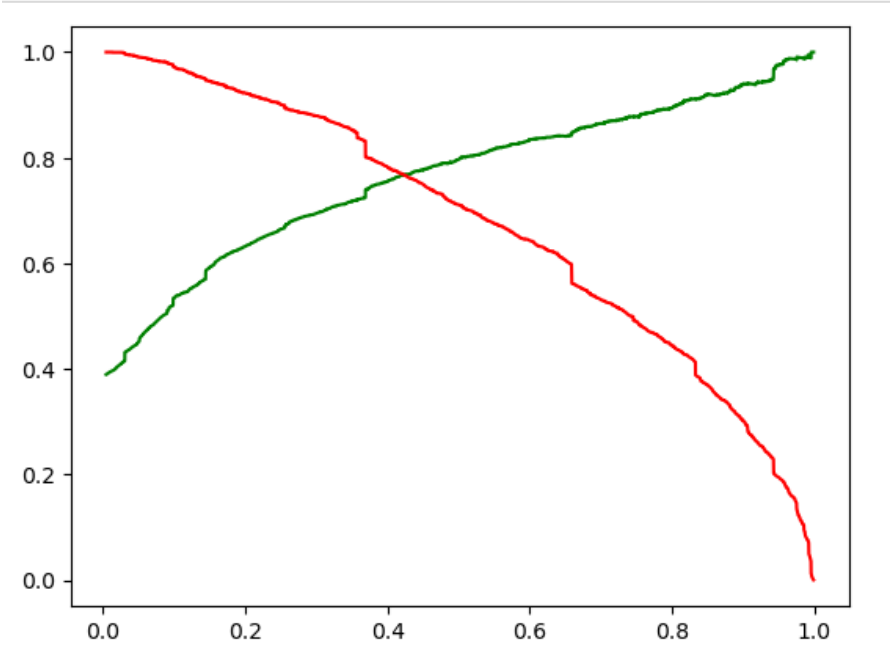
The graph depicts an optimal cut off of 0.37 based on accuracy, sensitivity and specificity

Confusion metrics:

3166	692
491	1971

Accuracy - 81%
Sensitivity - 80 %
Specificity - 82 %
False Positive Rate - 17 %
Positive Predictive Value - 74 %
Positive Predictive Value – 86%

Model Evaluation – Precision and Recall on Train Data Set



The graph depicts an optimal cut off of 0.42 based on precision and recall

Model Evaluation - Sensitivity and Specificity on Test Data Set

Confusion metrics:

1393	301
203	812

Accuracy - 81%
Sensitivity - 80 %
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.
- Sensitivity and Specificity values of test set are around 80% 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.