Lead Scoring Case Study

-Submitted by
Riya Mehta
Saiteja Gundeti

Problem statement

X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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%.

Steps to Problem Solving

STEP 1- Data Cleaning and Preparation

- Read and convert data to suitable format for analysis
- Remove duplicate data and Outlier treatment
- EDA

STEP 2- Feature Scaling And Splitting Train & Test sets

- Feature Scaling of Numeric data
- Spilliting data into Train test sets

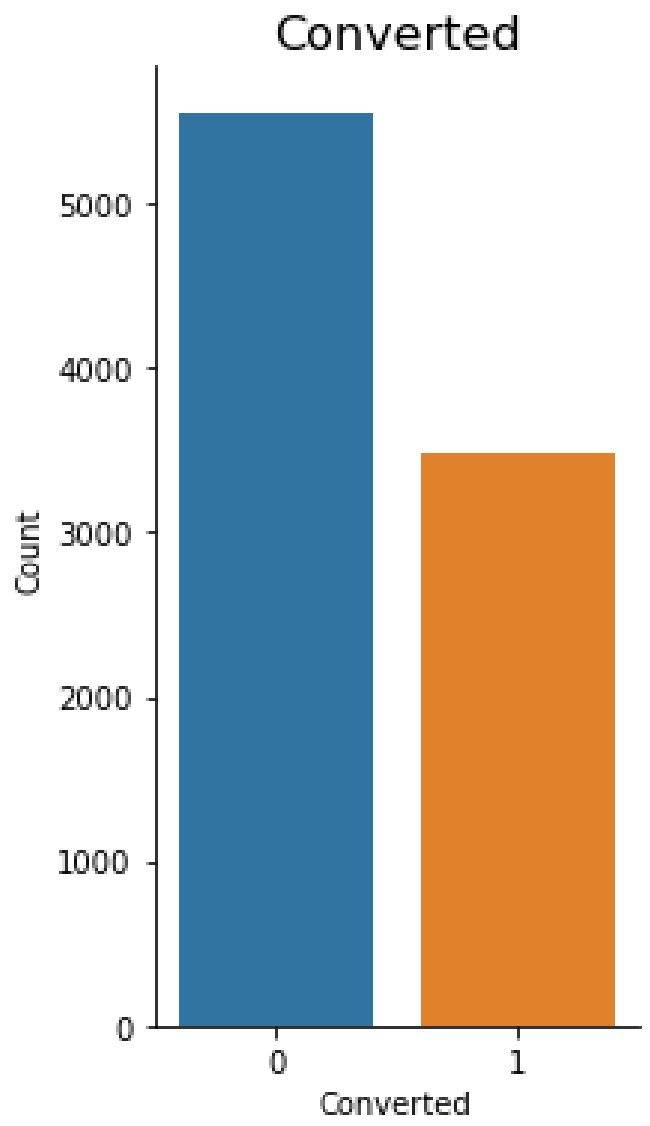
STEP 3- Model Building

- Feature Selection using RFE
- Determine optimal model using logistic regression
- Calculate Accuracy, sensitivity, specificity, precision and recall and evaluate model

STEP 4- Result

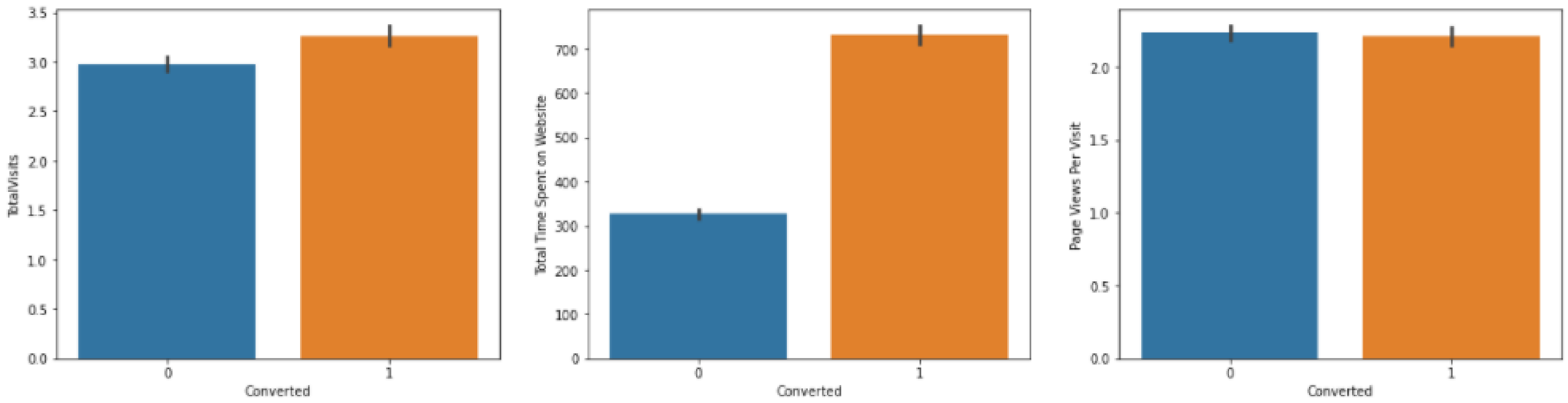
- Determine lead score & check if final prediction amounts to 80% conversion rate
- Evaluate final prediction on test set using metrics

Exploratory Data Analysis

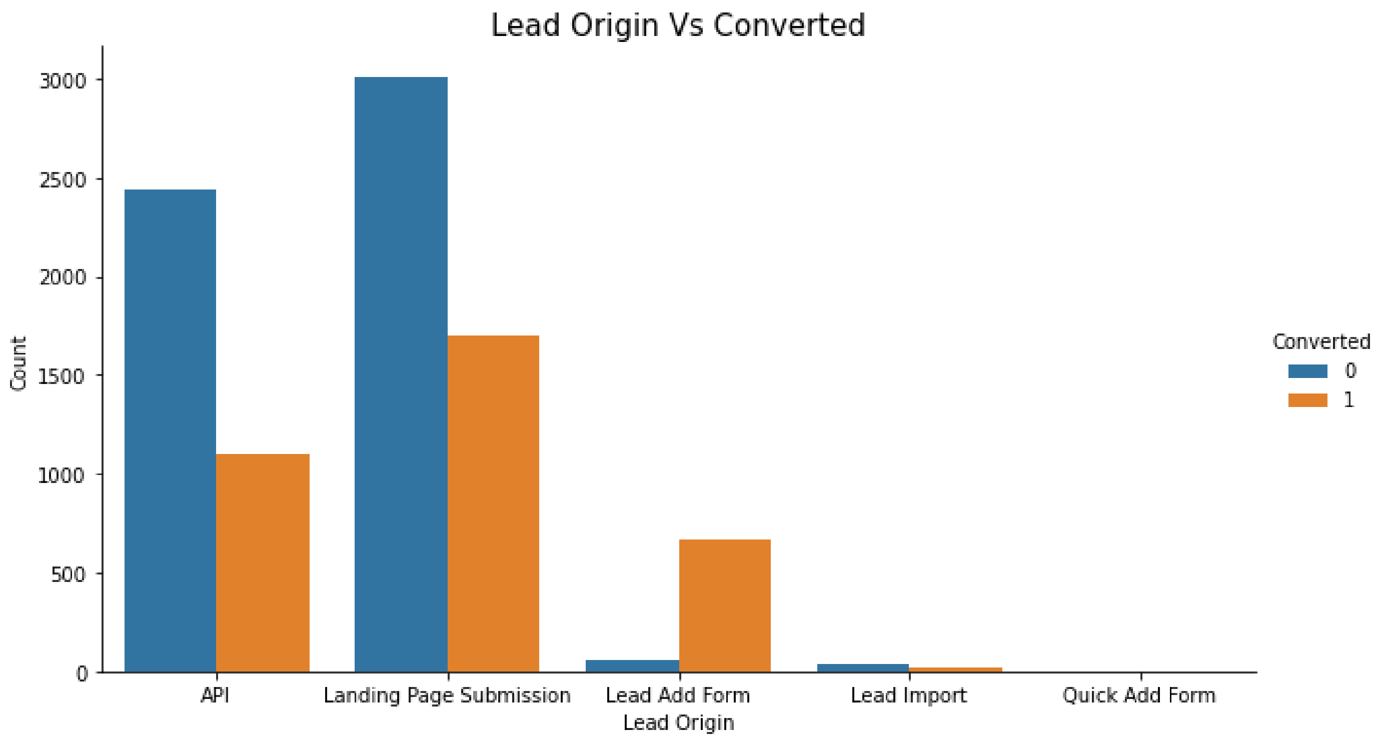


We have conversion rate of around 39%

Exploratory Data Analysis

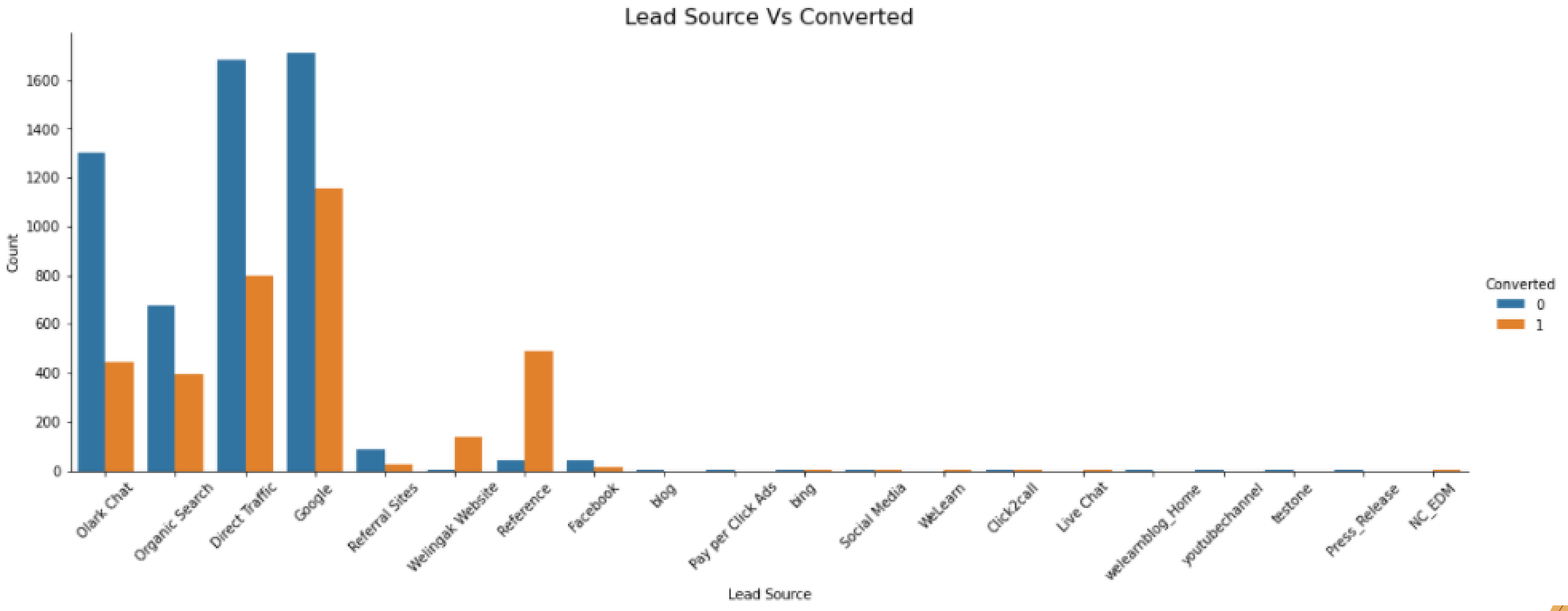


The Conversion rates are high for Total visits, Page views per visit and Total time spent

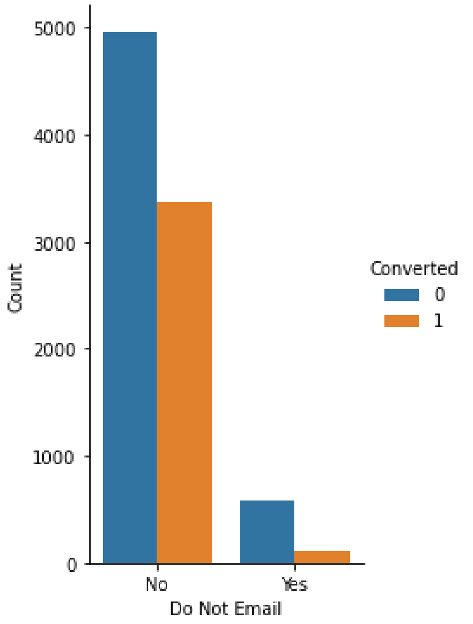


In Lead Origin, maximum conversion happened from Landing Page Submission

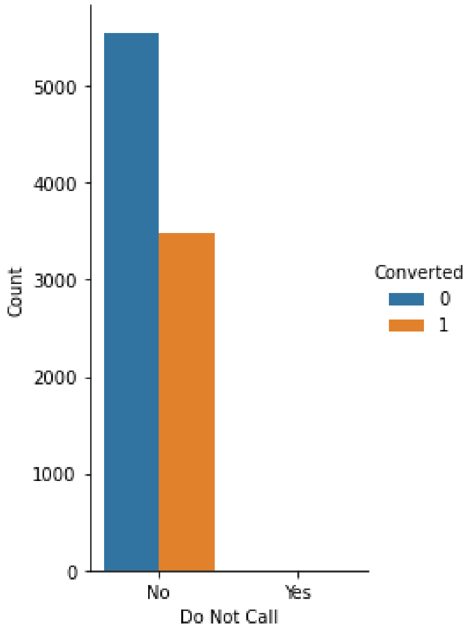
Major conversion in the lead source is from Google



Do Not Email Vs Converted



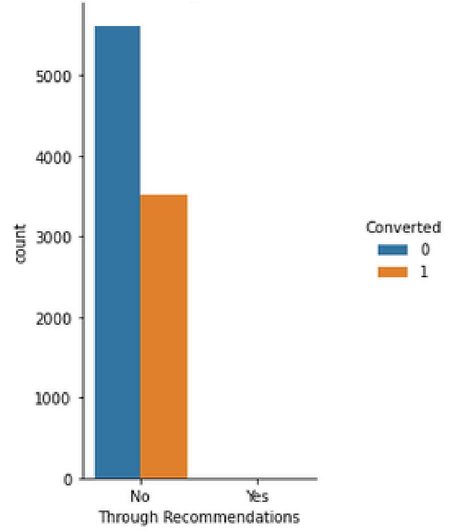
Do Not Call Vs Converted



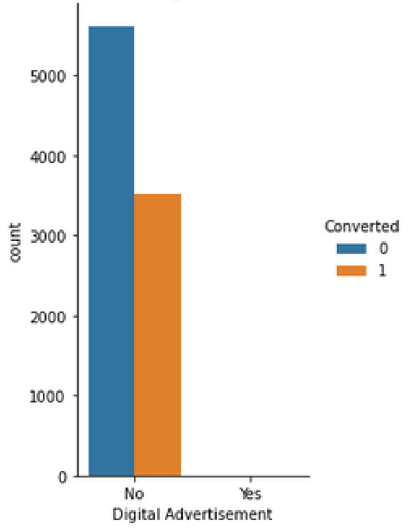
Major conversion has happened from Emails & calls

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

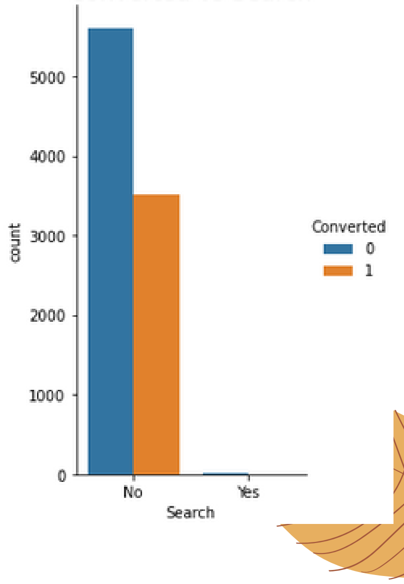
Converted vs Through Recommendations

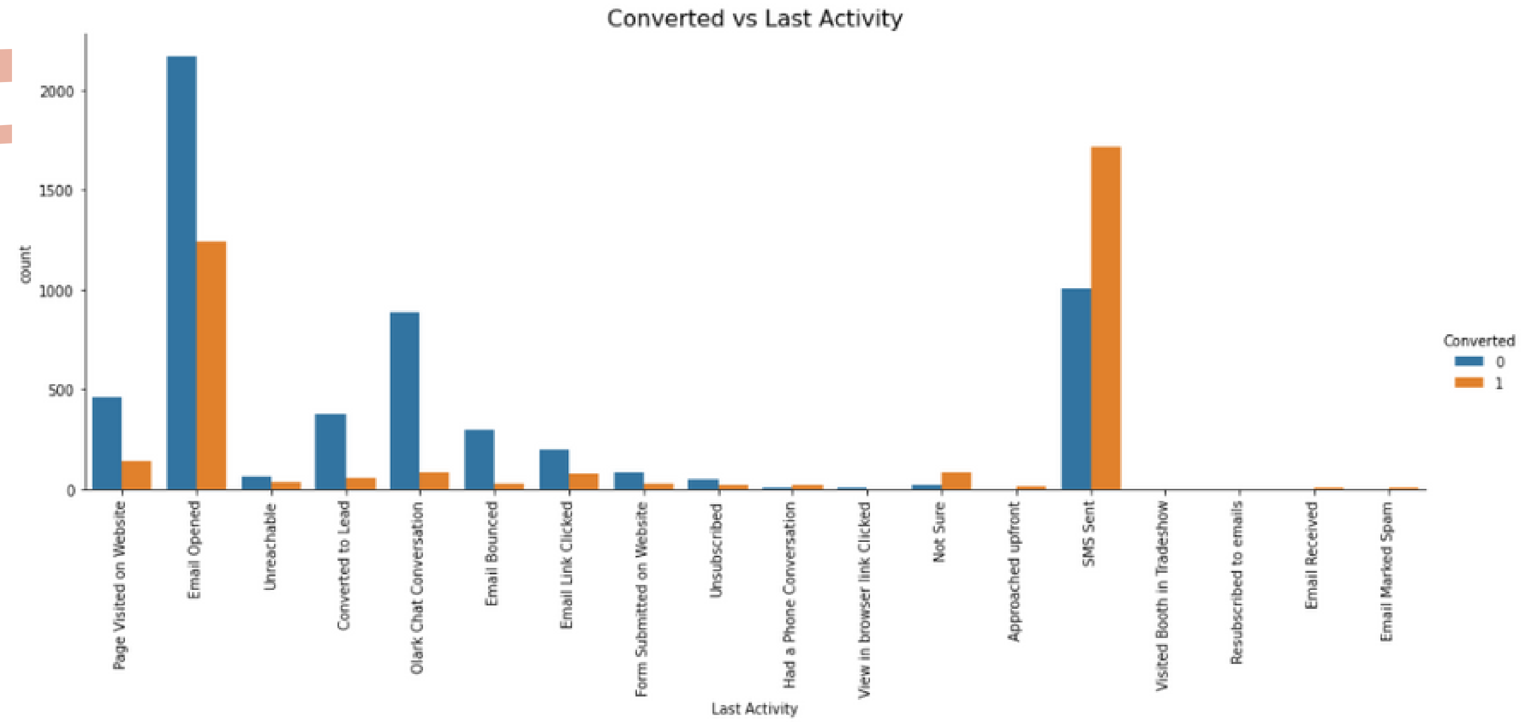


Converted vs Digital Advertisement



Converted vs Search

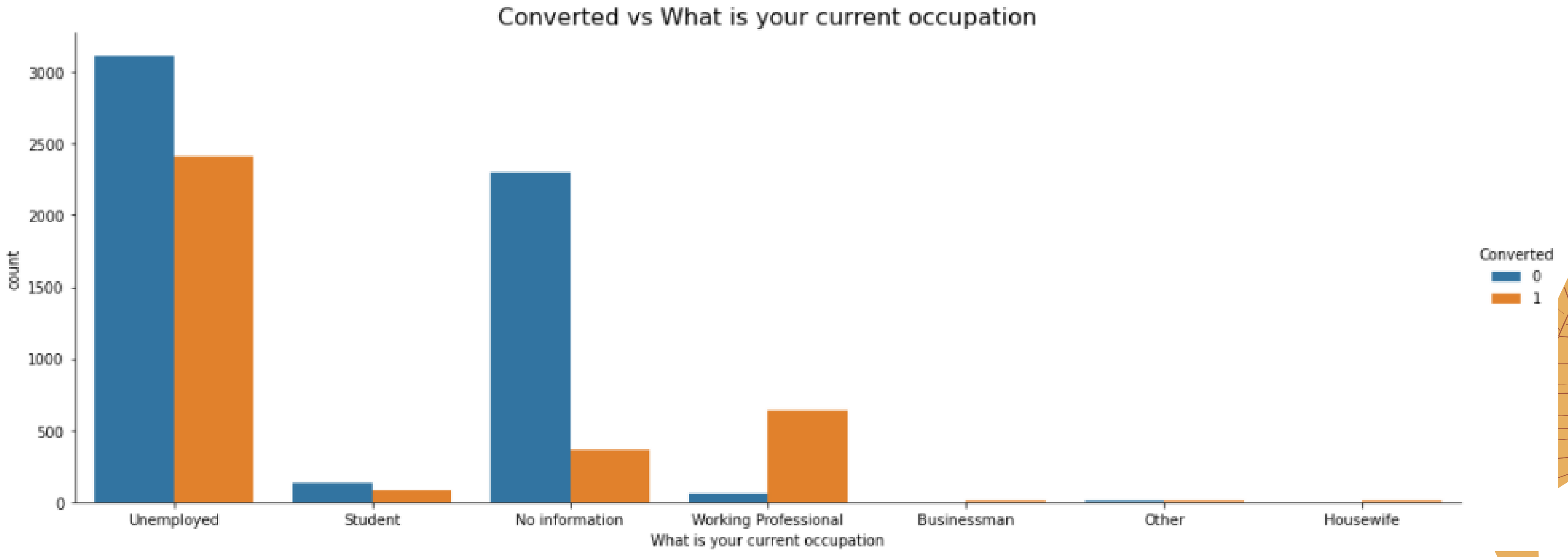




Converted vs Last Activity

Last Activity value of SMS Sent' had more conversion.

More conversion happened with people who are unemployed



Converted vs What is your current occupation

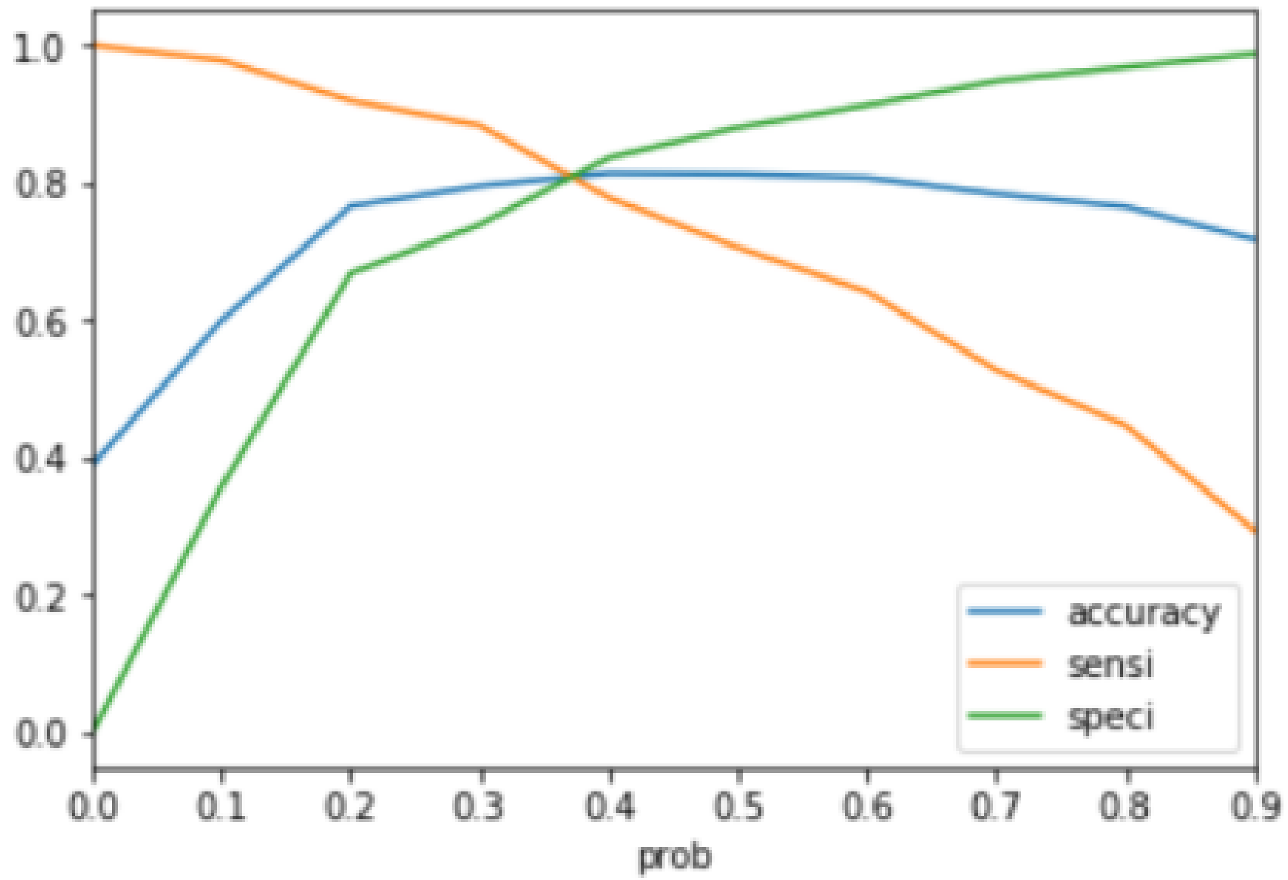
Variables Impacting the Conversion Rate

- 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



Confusion Matrix

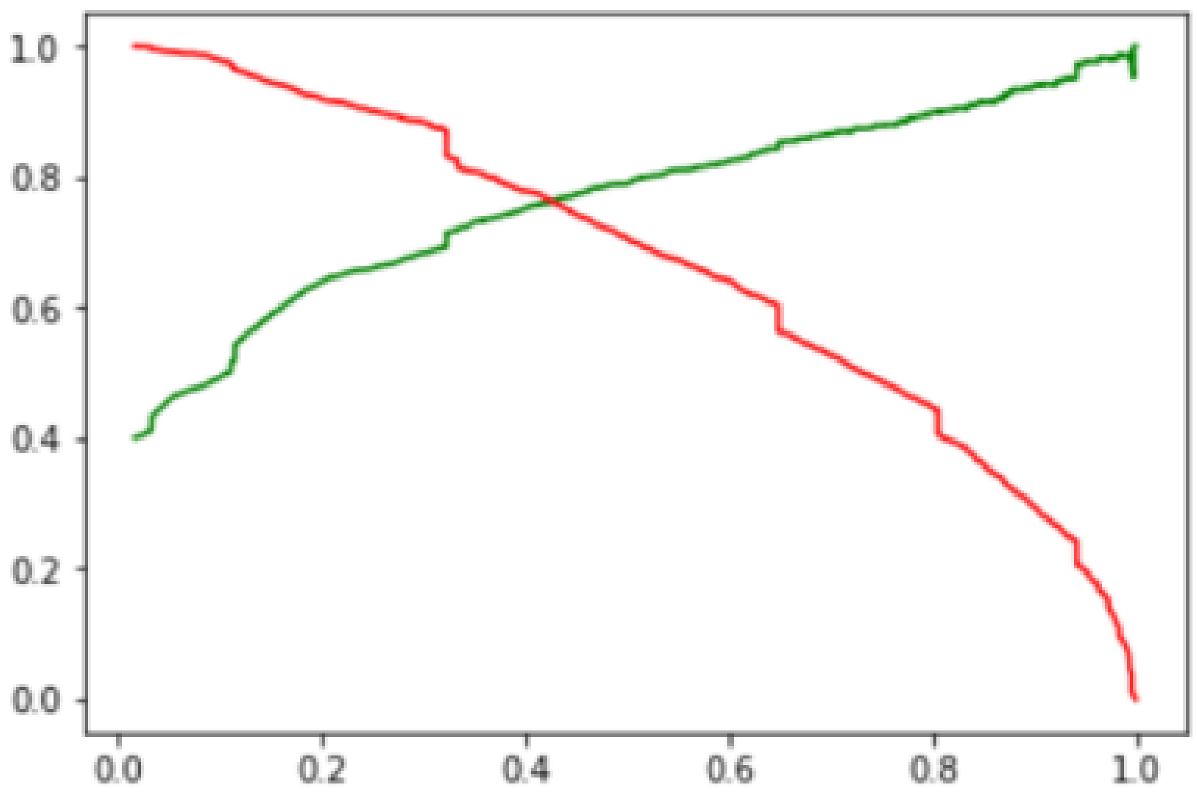
3161	697
974	1965

- Accuracy - 81%
- Sensitivity - 80 %
- Specificity - 82 %
- False Positive Rate - 18 %
- Positive Predictive Value - 74 %
- Positive Predictive Value – 86%

Model Evaluation

Precision and Recall on Train Datase

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



Confusion Matrix

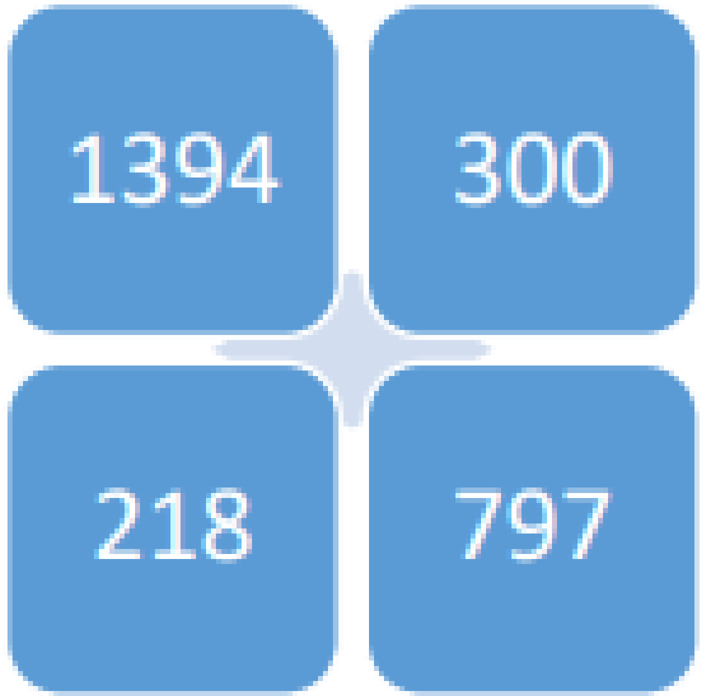
3397	461
725	1737

- Precision - 79 %
- Recall - 71 %

Model Evaluation

Sensitivity and Specificity on Test Dataset

Confusion Matrix



1394	300
218	797

- Accuracy - 81 %
- Sensitivity - 79 %
- 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 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
- Had a Phone Conversation from Last Notable Activity
- Hence overall this model seems to be good.