Lead Score Case Study

Submitted by:

Ayush Chauhan

Richa Sinha

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 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 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 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.

Problem solving methodology

Data Sourcing , Cleaning and Preparation

- Import data from source
- Cleaning data and fixing null values
- Removed the duplicate data
- Outlier are treated.
- EDA - Exploratory Data Analysis
- Standardization of the feature.

Scaling the Feature and Splitting Train and Test Sets

- Feature Scaling the Numeric data
- Splitting data into train and test set.

Model Building

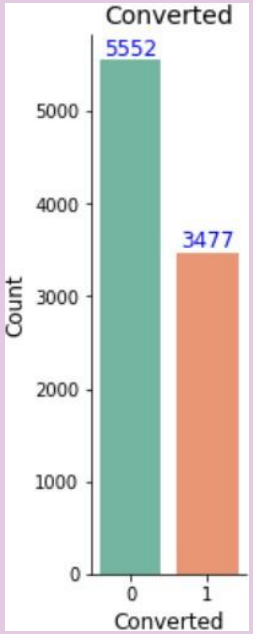
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate different metrics like, accuracy ,precision ,sensitivity, specificity and recall and evaluate the model.

Result

- Calculating lead score and checking that the final predictions results in 80% conversion rate.
- Evaluating the final prediction on the test set using the threshold from specificity and sensitivity metrics.

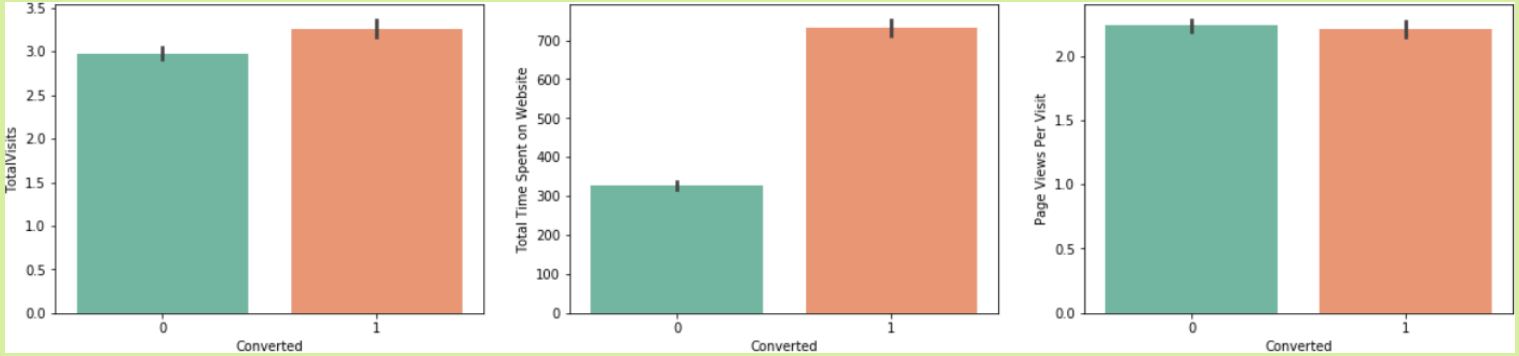
Exploratory Data Analysis (EDA)

We have around 39% Conversion rate in Total

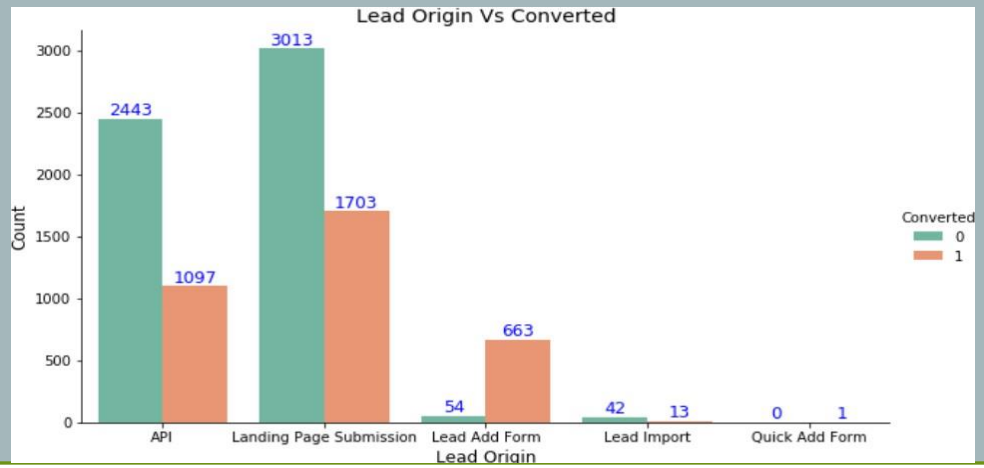


.

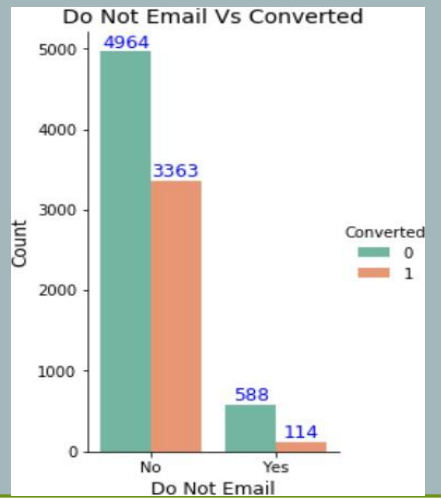
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

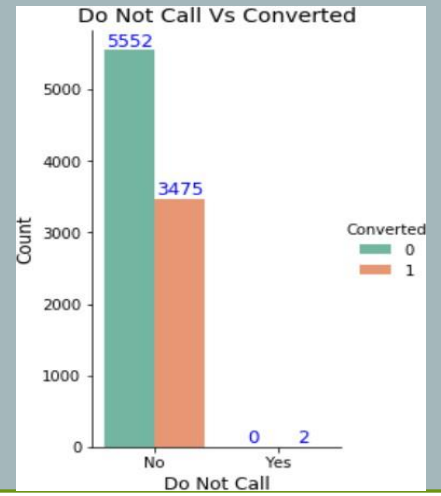


In the column ‘Lead Origin’ maximum conversion happened from ‘Landing Page Submission’

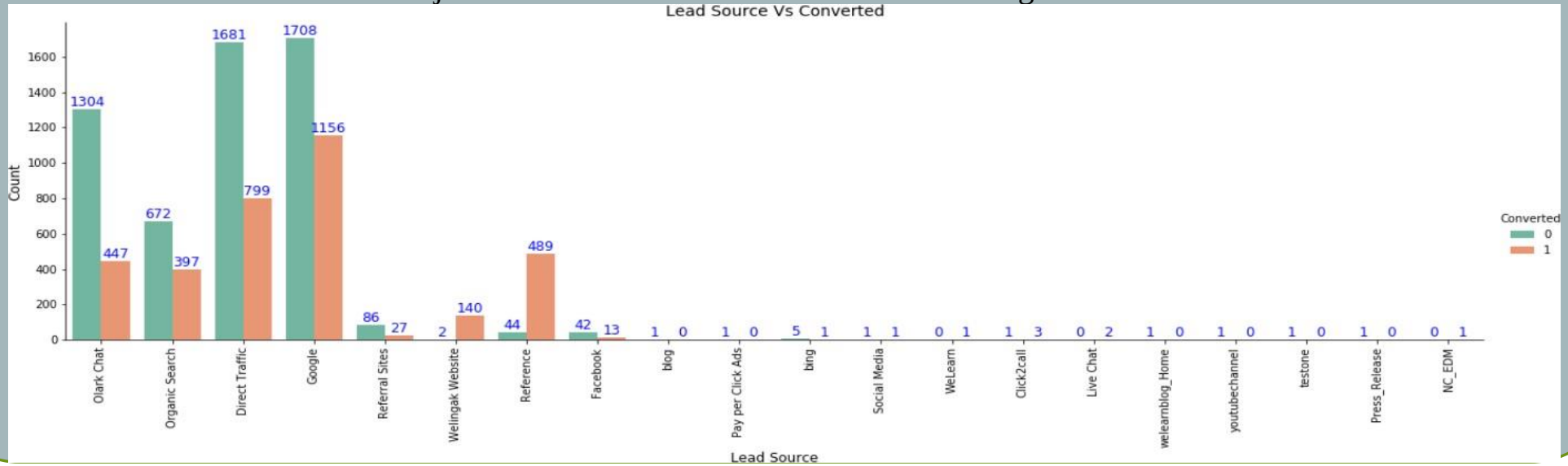


Major conversion has happened from Emails sent and Calls made

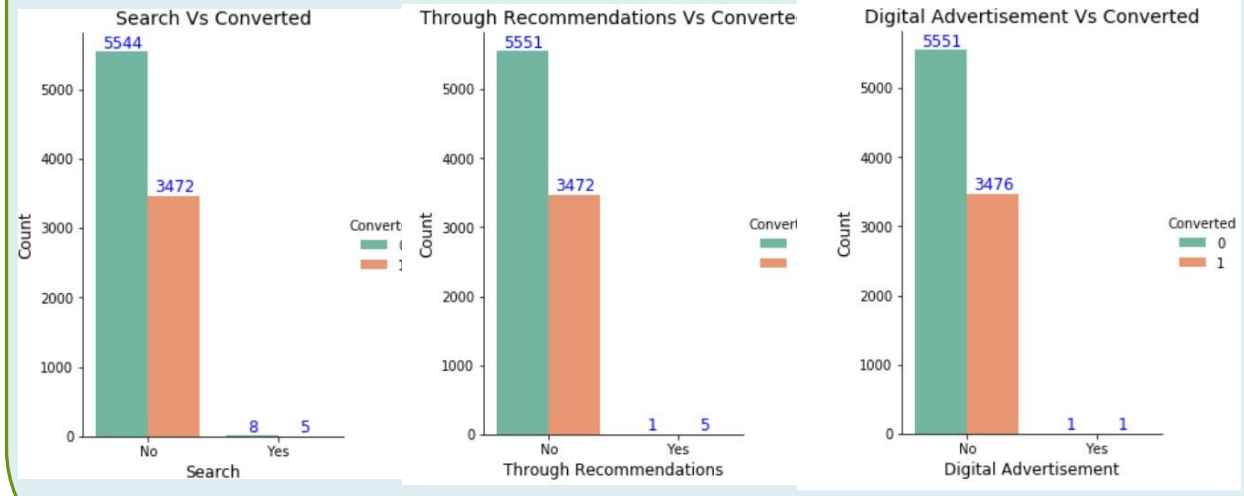




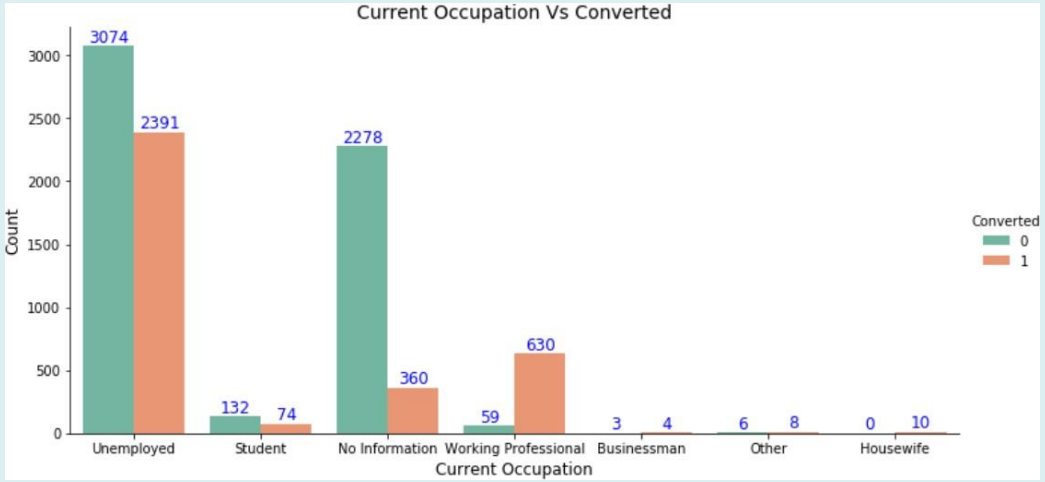
Major conversion in the lead source is from Google



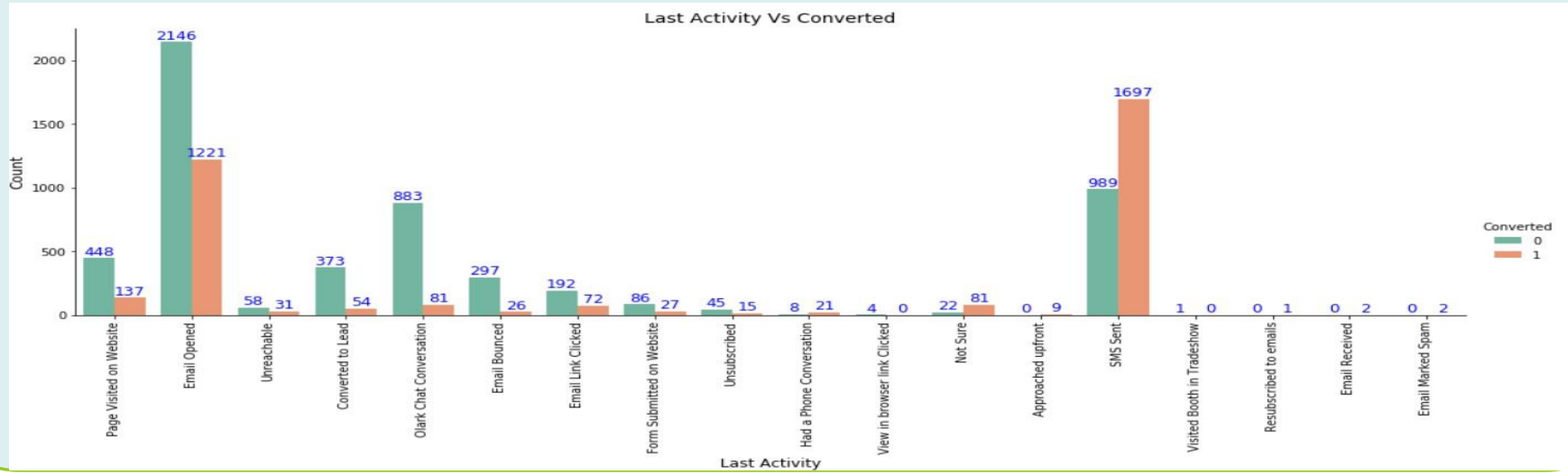
Not much impact on conversion rates through Search, digital advertisements and through recommendations



More conversion happened with people who are unemployed



Last Activity value of SMS Sent' had more conversion.

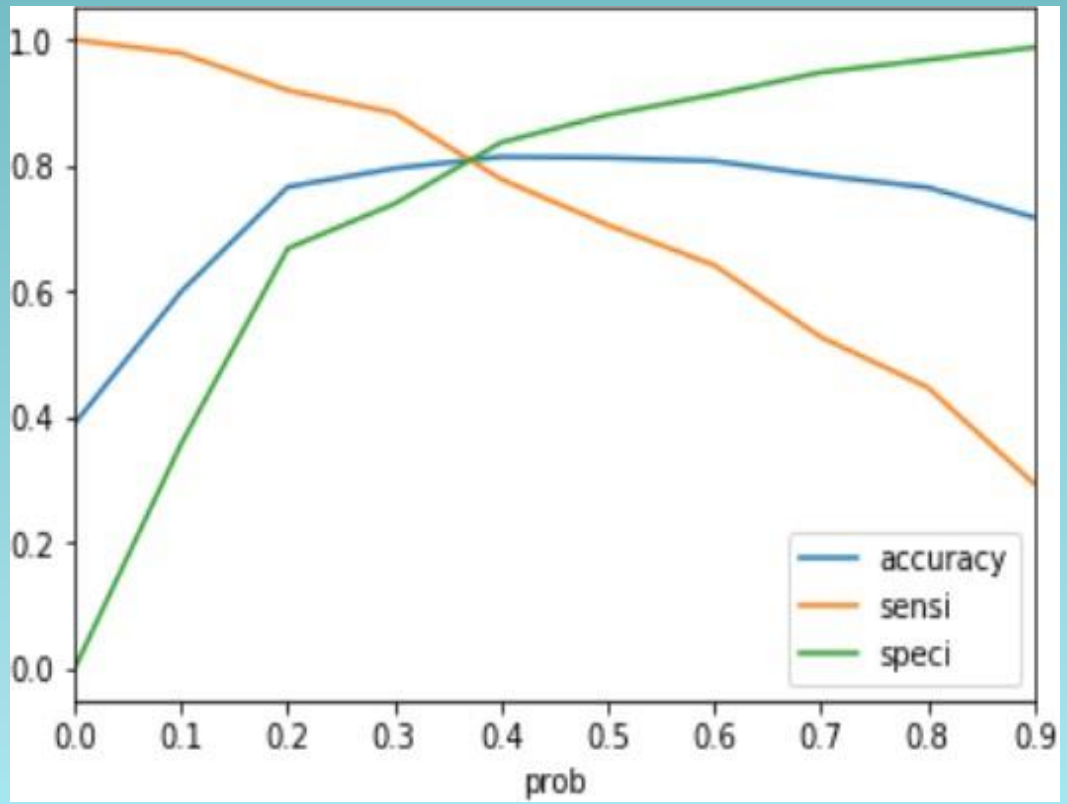


Variables Impacting the Conversion Rate

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead 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



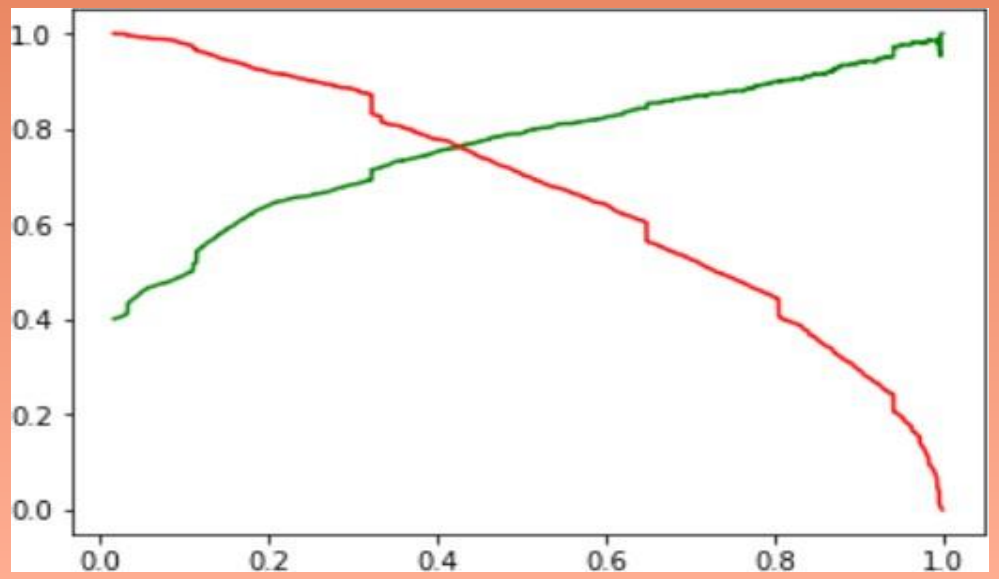
Confusion Matrix

3413	445
708	1754

- Accuracy - 81%
- Sensitivity - 71%
- Specificity - 88 %
- False Positive Rate - 11 %
- Positive Predictive Value - 79 %
- Negative Predictive Value – 82%

Model Evaluation- Precision and Recall on Train Dataset

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



Confusion Matrix

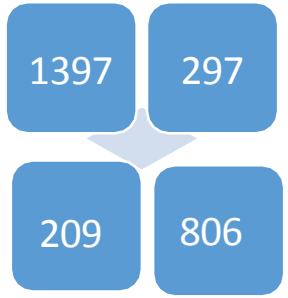
3413	445
708	1754

- Precision - 79 %
- Recall - 71 %

Model Evaluation – Sensitivity and Specificity on Test Dataset

- Accuracy - 81 %
- Sensitivity - 71 %
- Specificity - 88 %

Confusion Matrix



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%, 71% and 88% 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.