Lead Score Case Study

Submitted
Surbhi Mathur

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 h ave 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 D ata Analysis.
- Feature Scaling
- Splitting t he data into Test and Train dataset.
- Building a logistic Regression model and calculate L ead S core.
- Evaluating the model by using different m etrics - 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.

- Read the Data source.
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis Feature Standardization
- Convert data into clean
- format suitable for analysis

Feature Scaling and Splitting Train and Test Sets

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

Model Building

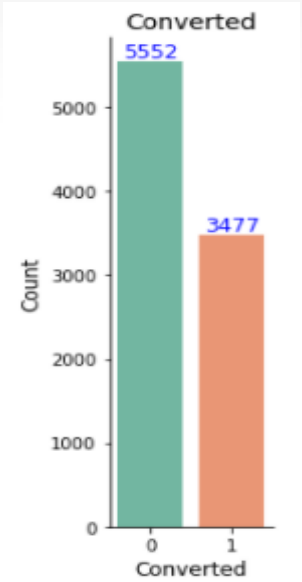
- Feature Selection using RFE
- Determine the optimal model
- using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity,
- precision and recall
- evaluate the model.

Result

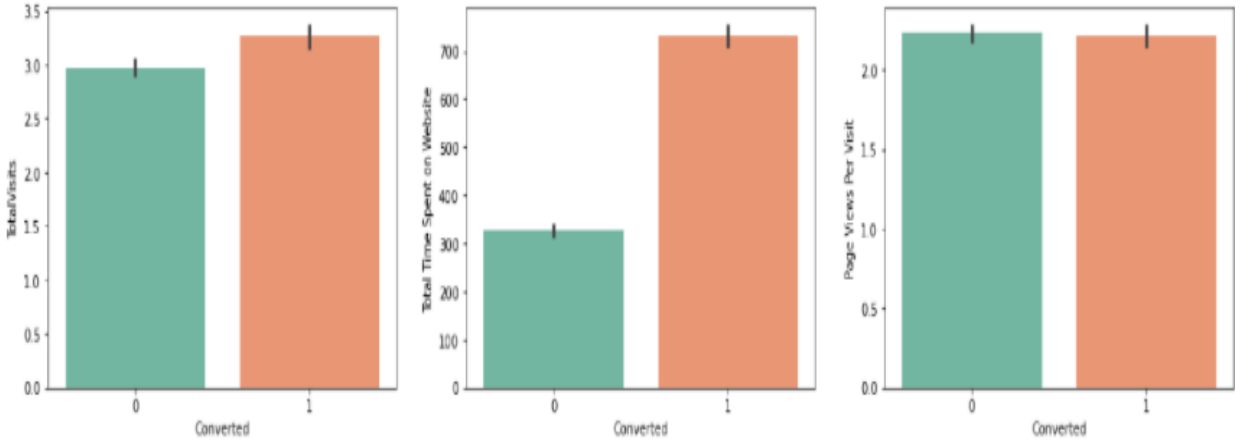
- Determine the lead score a nd check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the testset using cut off threshold from sensitivity and specificity metrics

Exploratory DataAnalysis

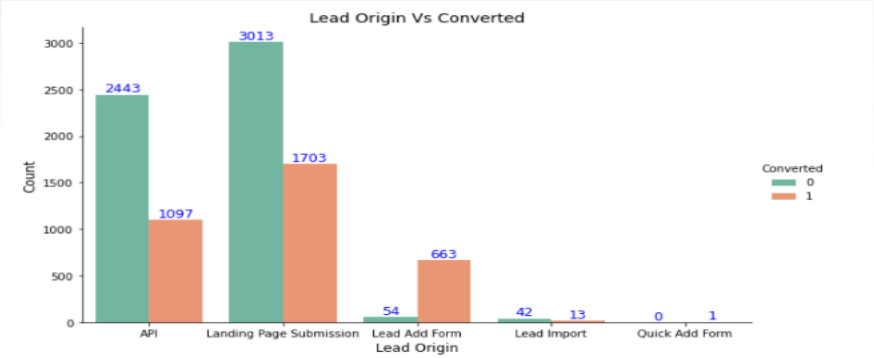
We have around 39% Conversion rate in Total



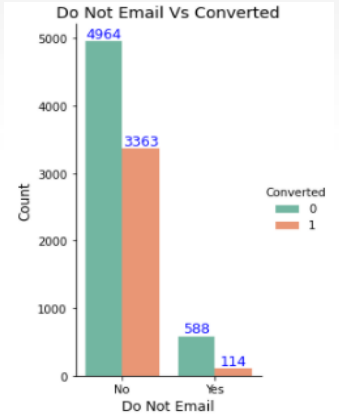
The conversion rates were high f or Total Visits, Total Time Spent on Website and Page Views Per Visit

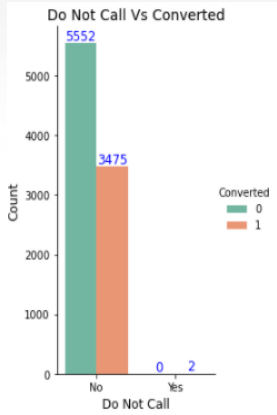


In Lead Origin, maximum conversion happened from Landing Page Submission

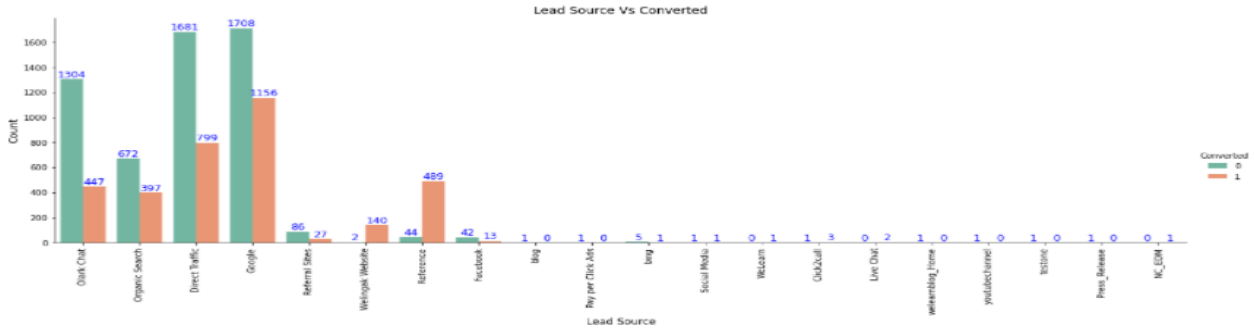


Major conversion has happened from Emails sent and Calls made

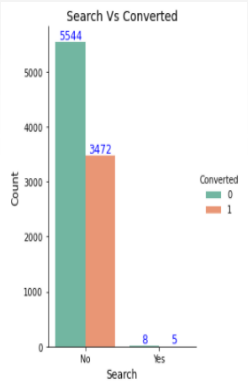


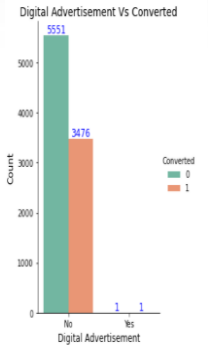


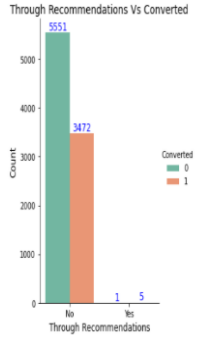
Major conversion in the lead source is from Google



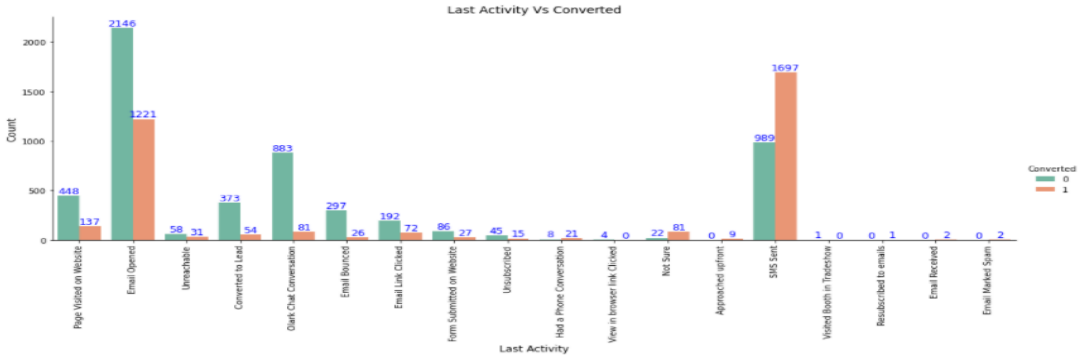
Not much impact on conversion rates through Search, digital advertisements and through recommendations



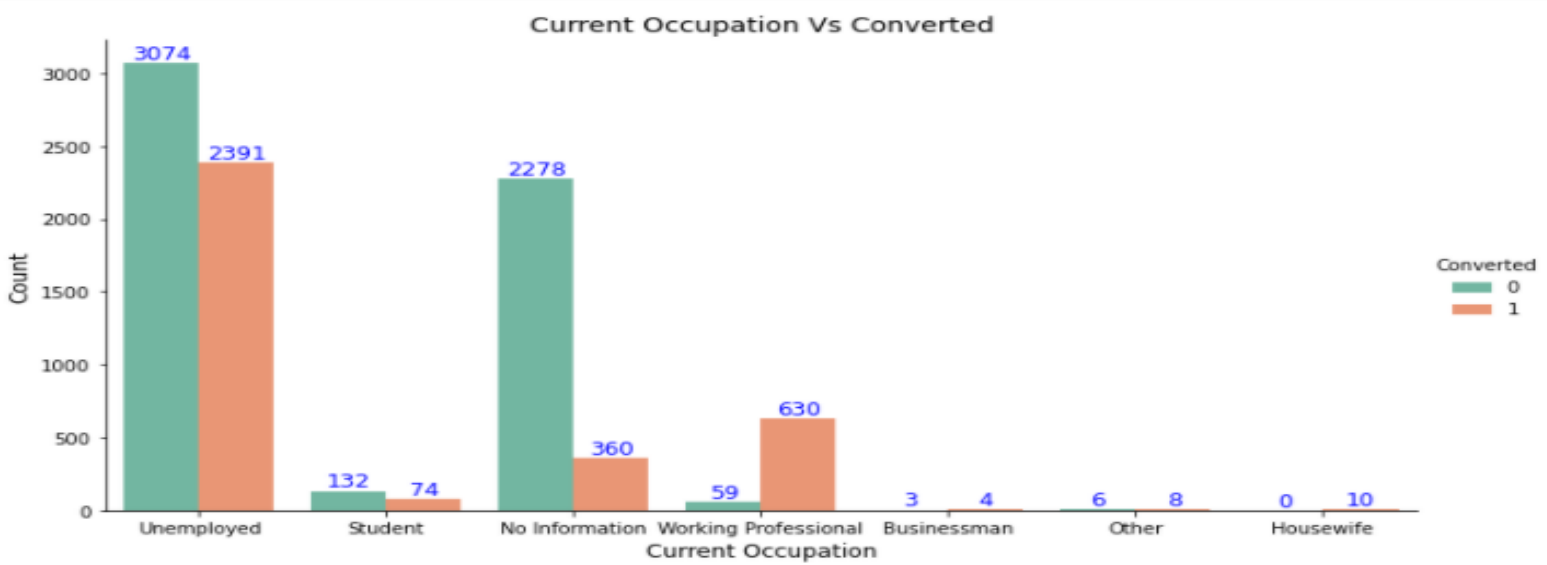




Last A ctivity value of SMS Sent' had more conversion.



More conversion happened with people who are unemployed

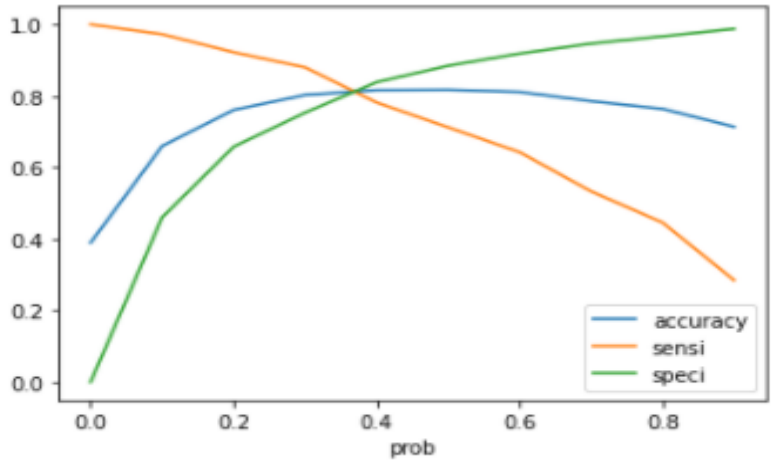


Variables Impacting the Conversion Rate

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead O rigin – Lead Add Form
- Lead S ource - Olark Chat
- Last S ource – Welingak Website
- Last Activity – Email Bounced
- Last A ctivity – Not Sure
- Last A ctivity – 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



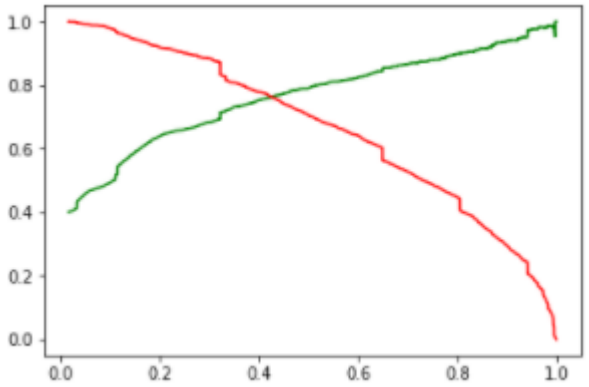
Confusion Matrix

[3166, 692]
[491, 1971]

- 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 Dataset

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



Confusion Matrix

[3412, 446]
[709, 1753]

- Precision - 79 %
- Recall - 71 %

Model Evaluation –Sensitivity and Specificity on Test Dataset

Confusion Matrix

```
[1393,  301]
[ 203,  812]
```

Accuracy- 81%
Sensitivity- 79%
Specificity - 82 %

Conclusion

- While w e 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 t he lead score calculated s hows 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
- LeadAdd Form from Lead Origin
- Had a Phone Conversation from Last NotableActivity Hence overall this model seems to be good.