Lead Scoring Case Study

Logistic Regression Assignment

Lead Score Case Study for X Education

Pawan Dixit & Vaibhav Mudholkar

Index

Overview

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%.

Approach

Data Collection, Cleaning & Preparation

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

Scaling & Splitting Train and Test Sets

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

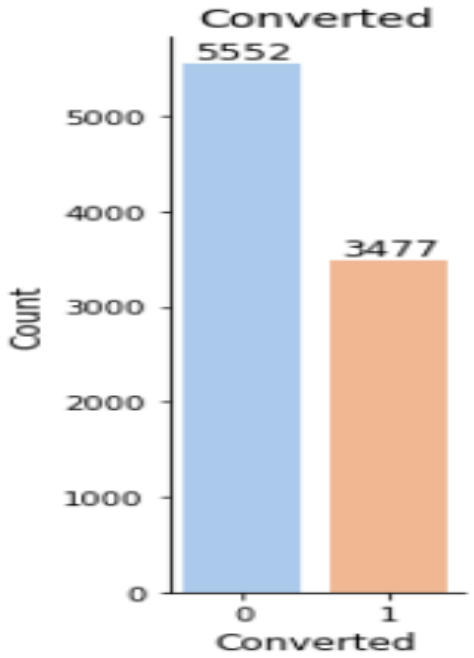
Model Building

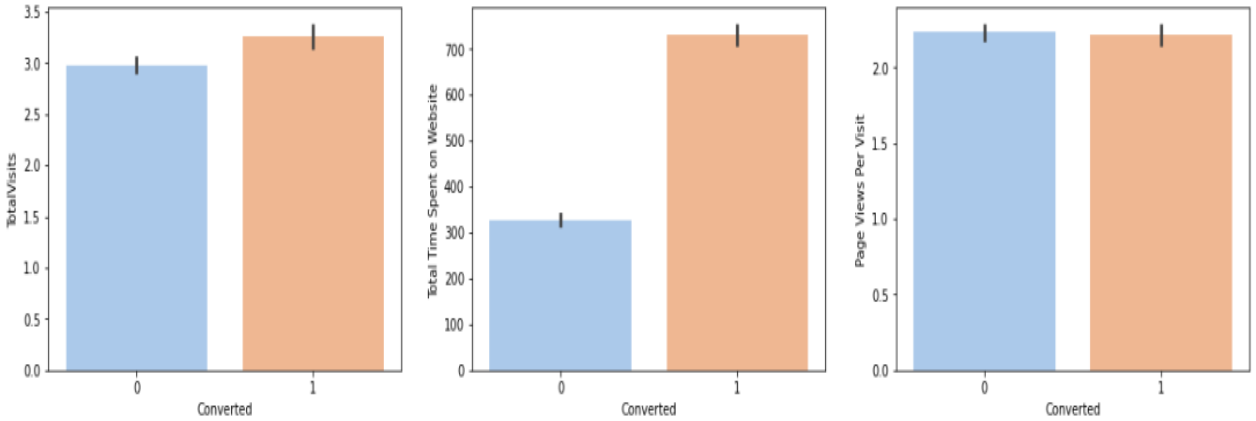
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Output

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

Exploratory Data Analysis (EDA)



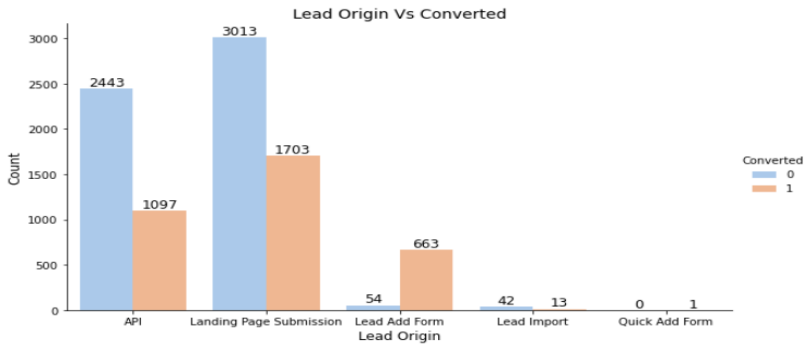


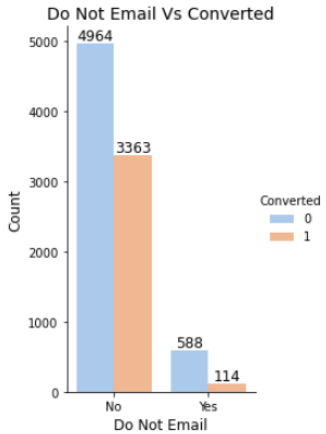
Conversions for all numeric values

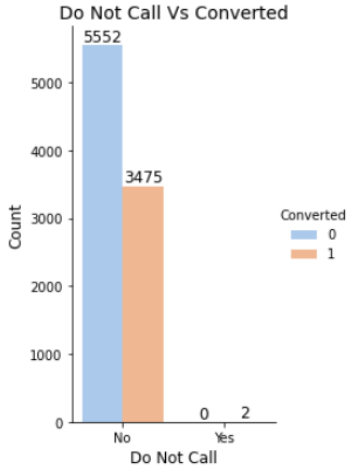
Converted: We have around 39% Conversion rate in Total

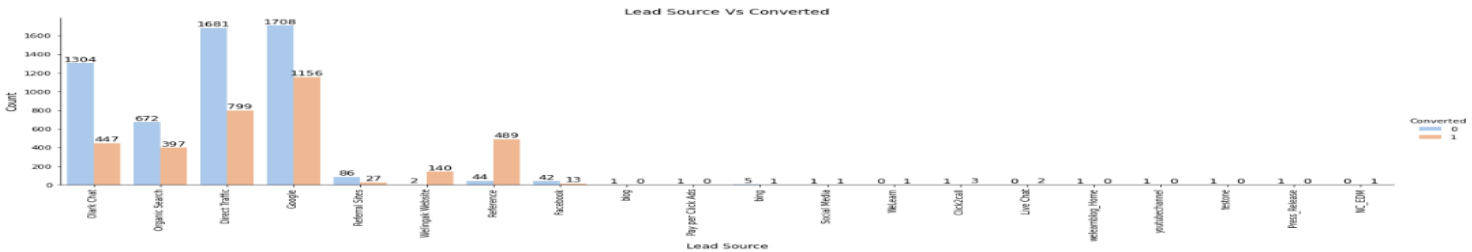
Conversions for all numeric values: The conversion rated were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

Exploratory Data Analysis (EDA)









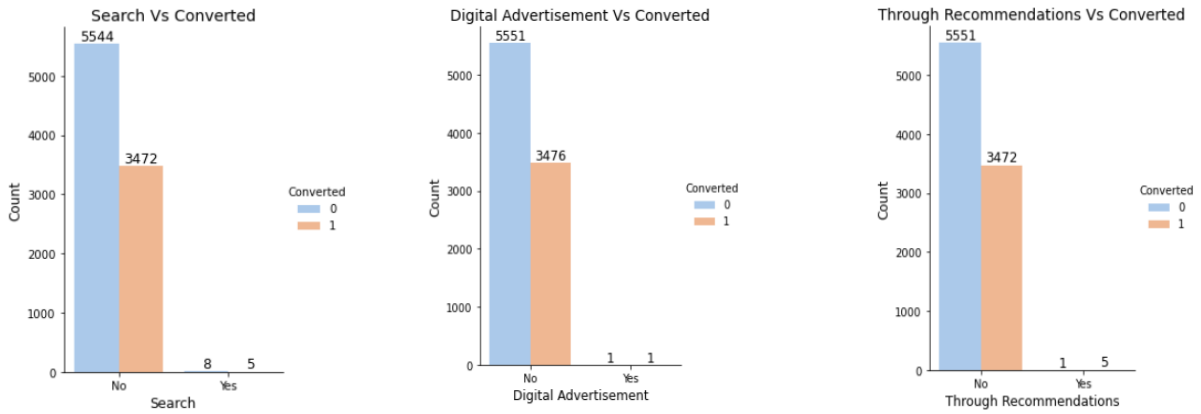
Lead Origin Vs Converted: In Lead Origin, maximum conversion happened from Landing Page Submission

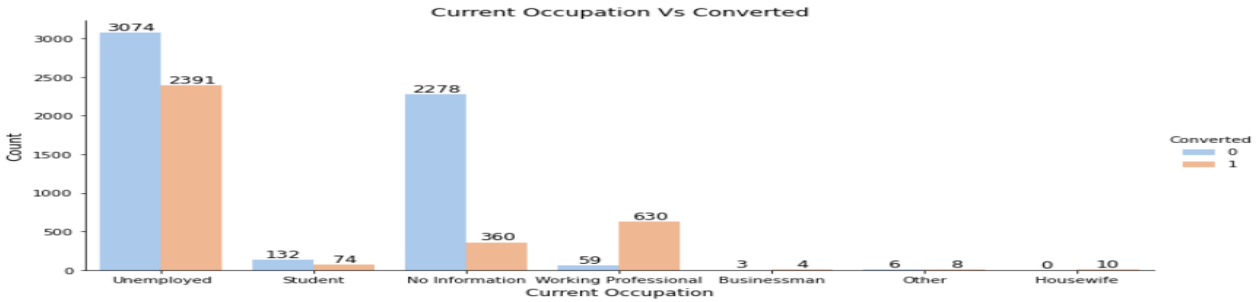
Do not Email & Do not Call Vs Converted:

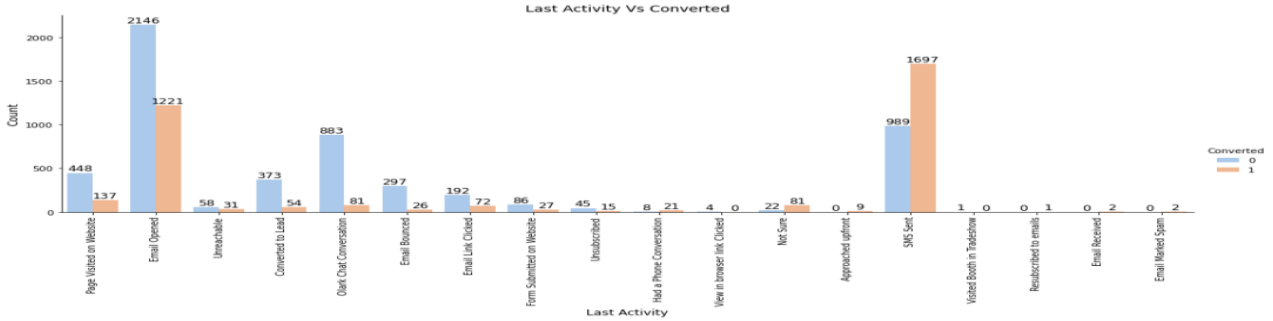
Major conversion has happened from Emails sent and Calls made

Lead Source Vs Converted: Major conversion in the lead source is from Google

Exploratory Data Analysis (EDA)







Search, Through Recommendation, Digital Advertisement Vs Converted :

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

Current Occupation Vs Converted :

More conversion happened with people who are unemployed

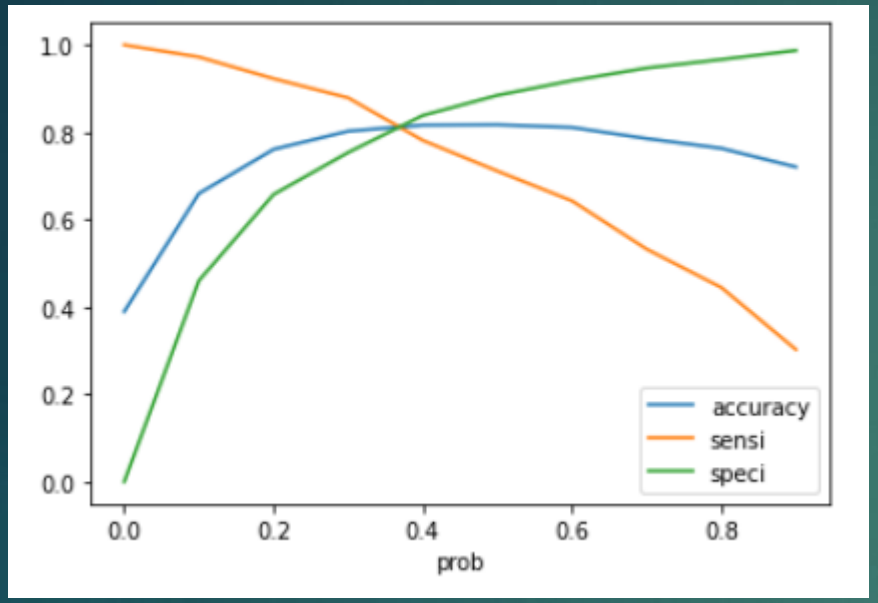
Last Activity Vs Converted

Last Activity value of SMS Sent' had more conversion

Variable Impacting the Conversion Rate

- TotalVisits
- Total Time Spent on Website
- Search
- 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 on Train Data Set : Sensitivity & Specificity



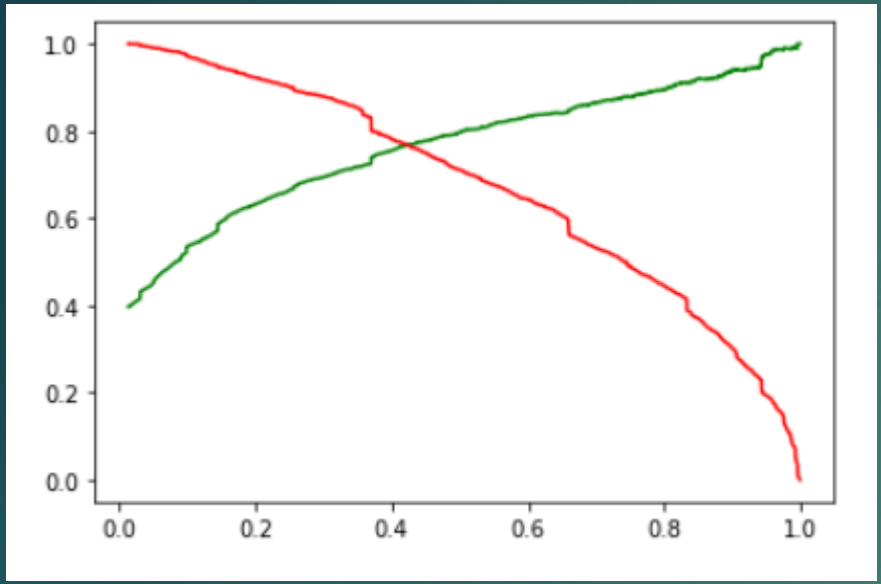
The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity

Confusion Matrix

3165	693
493	1969

- Accuracy-81%
- Sensitivity-79.9%
- Specificity-82%
- FalsePositiveRate-17.9%
- PositivePredictiveValue-73.9%
- NegativePredictiveValue–86.5%

Model Evaluation on Train Data Set: Precision & Recall



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

Confusion Matrix

3413	445
711	1751

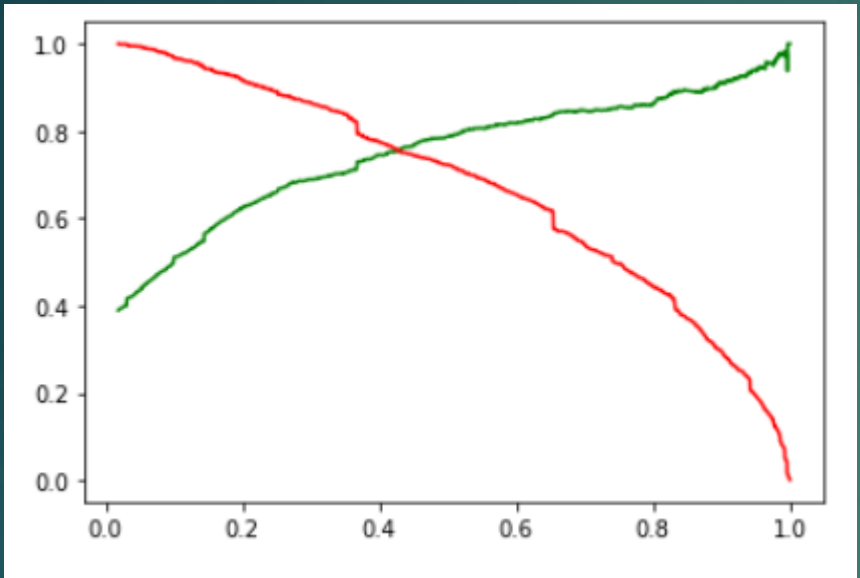
- Precision -79.7%
- Recall-71.1%

Model Evaluation on Test Data Set : Sensitivity & Specificity

Confusion Matrix

- Accuracy-81%
- Sensitivity-79.4%
- Specificity-82.4%

Model Evaluation on Test Data Set: Precision & Recall



- Precision -73%
- Recall-79.4%

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 similar 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.