



Lead Scoring Case Study

Presented By:
Koushik Potnuru
Anjali Mishra
Nikitha R



Lead Scoring 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 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%.
- 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.



Lead Scoring Case Study for X Education

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:

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



Problem solving methodology

Feature Scaling and Splitting Train and Test Sets:

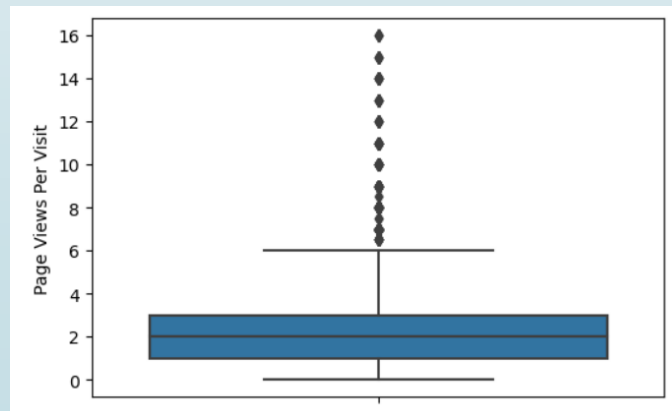
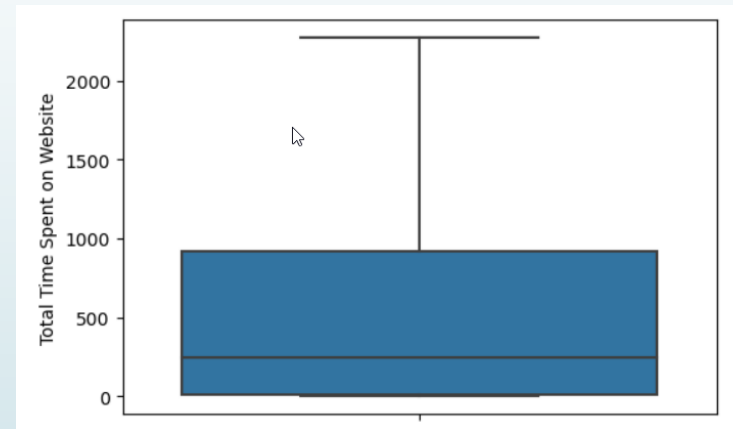
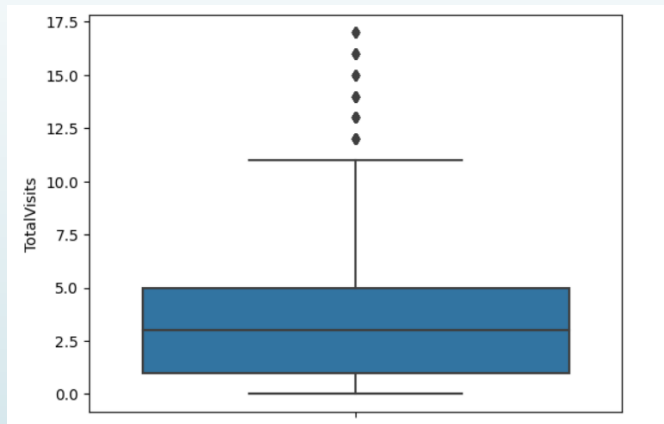
- Feature Scaling of Numeric Data.
- Splitting data into train and test set.

Model Building:

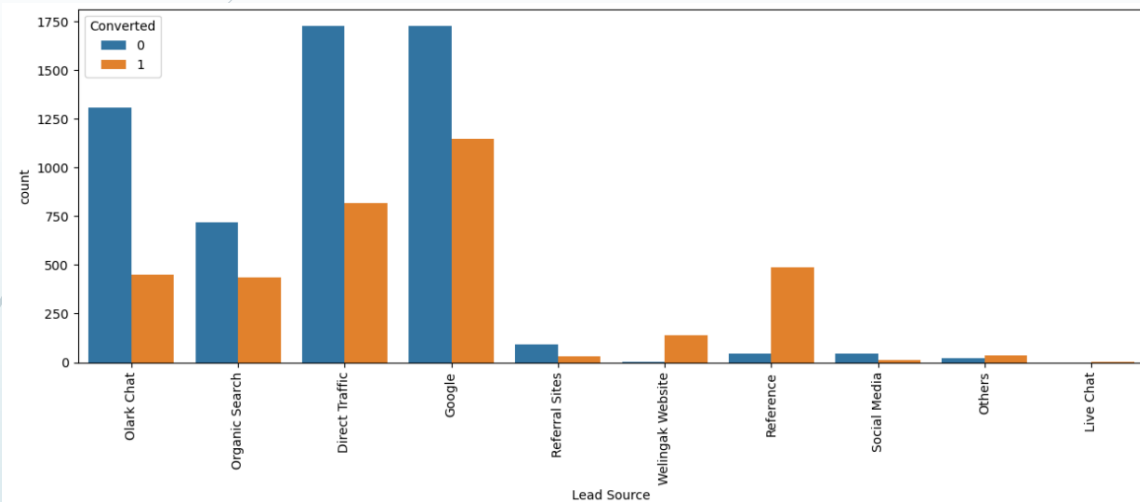
- Feature Selection using RFE.
- Determine the optimal model using Logistic Regression.
- Calculate various metrics like accuracy, sensitivity, specificity precision and recall.
- evaluate the model.

Outliers :

Total Visits, Total Time Spent on Website, Page Views Per Visit.

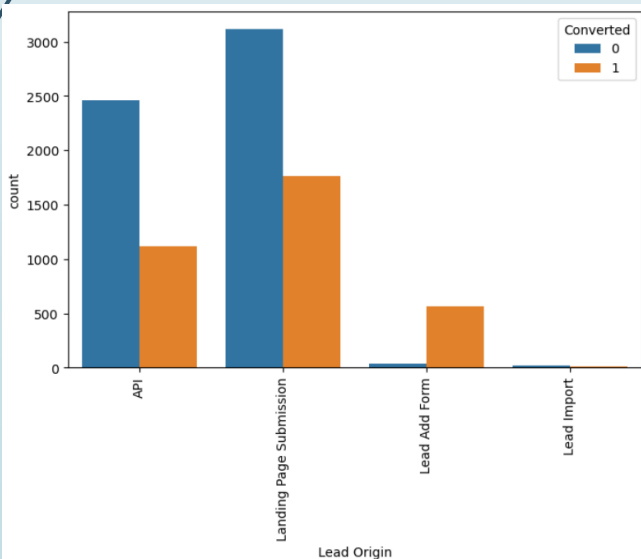


Exploratory Data Analysis



Lead Source vs Converted

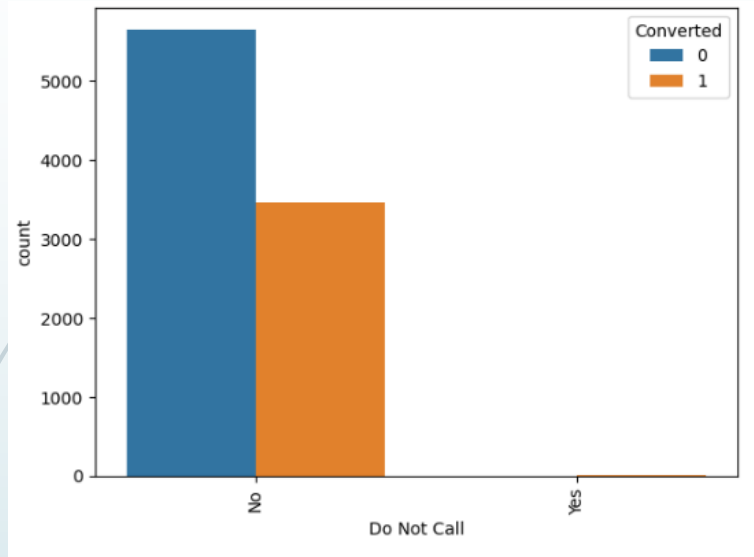
- The largest number of leads are generated by Google and Direct traffic.
- The conversion rates of reference leads and leads from the Welingak website are high.



Lead Origin vs Converted

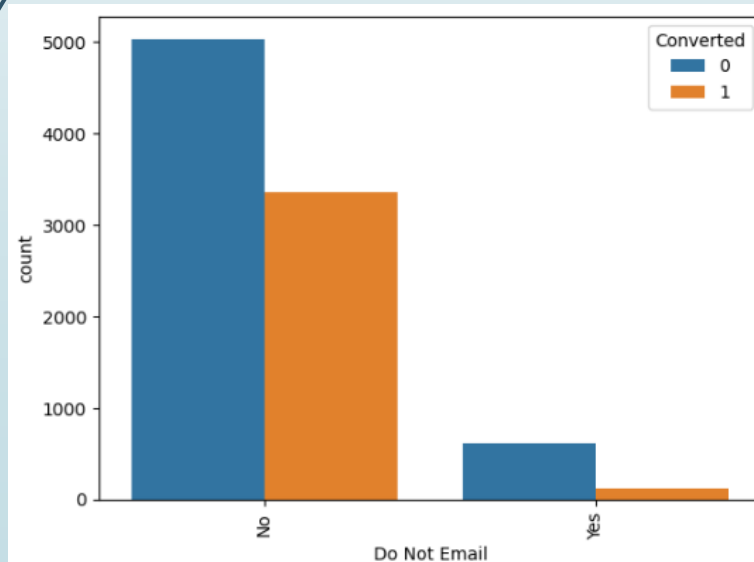
- The API and Landing Page Submission sources tend to generate a higher number of leads and conversions.
- The Lead Add Form has a very high conversion rate, although the number of leads generated is relatively low.

Exploratory Data Analysis



Do Not Call vs Converted

- Most leads not to be performed through the phone.

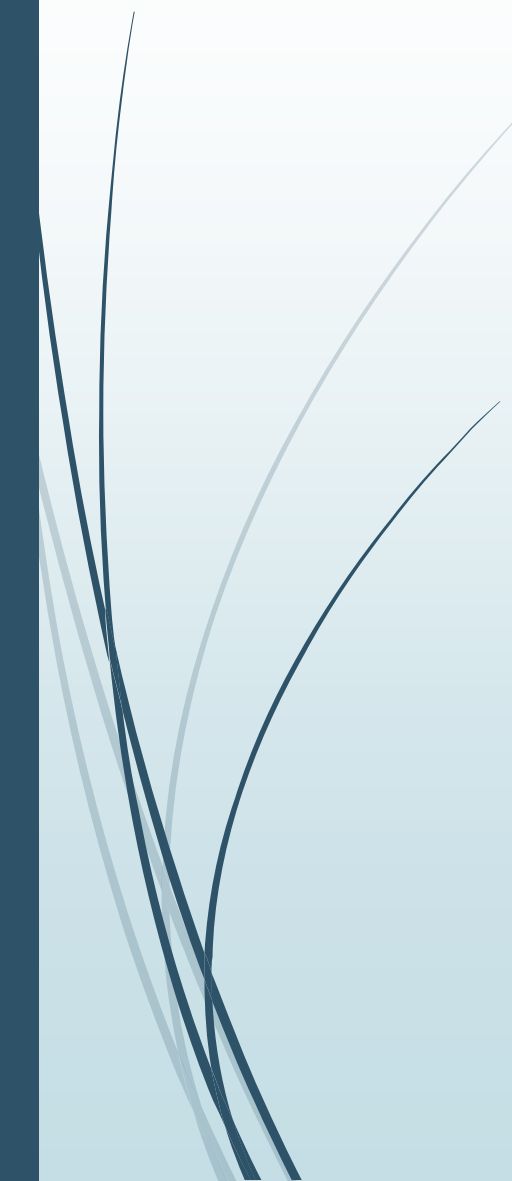


Do not Email vs Converted

- Most leads not to be performed through the mails.

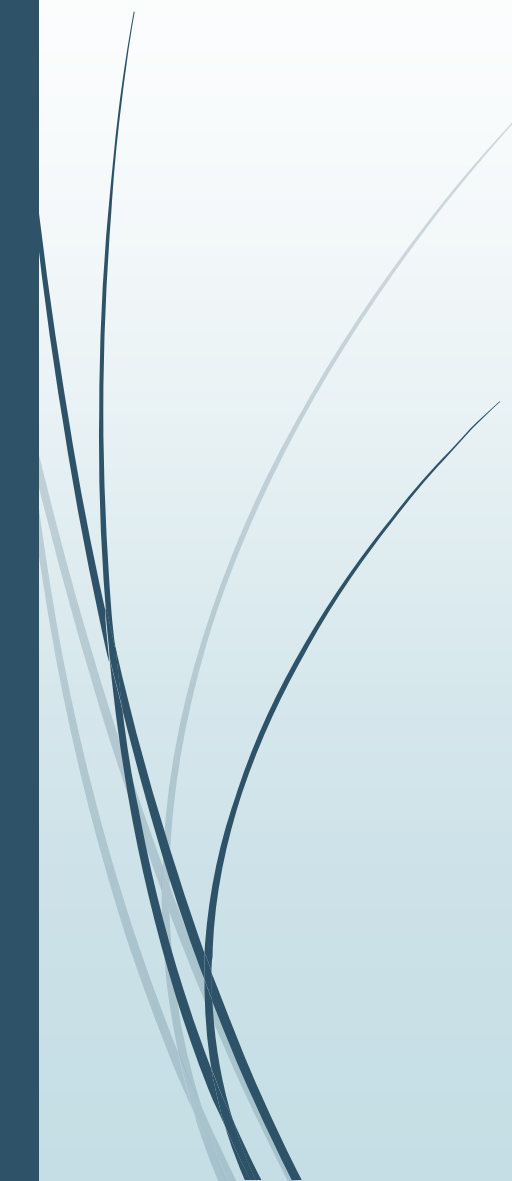


Feature Scaling & Splitting Train & Test Sets

- Feature scaling of Numeric Data.
 - Splitting data into Train & Test set.
- 



Model Building

- Feature Selection using RFE
 - Determined Optimal Model using Logistic Regression.
 - Calculated accuracy, sensitivity, specificity, precision, recall.
- 

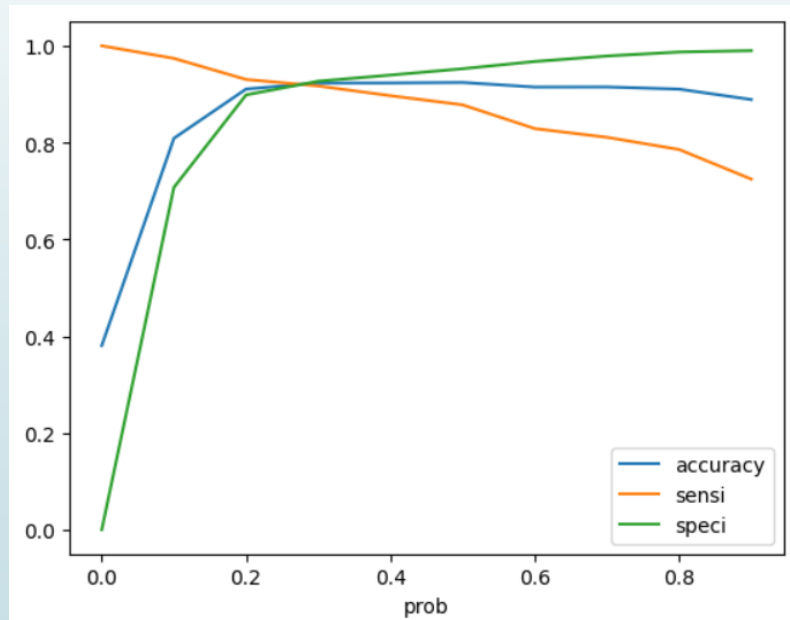


Variables Impacting the Conversion Rate

- 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

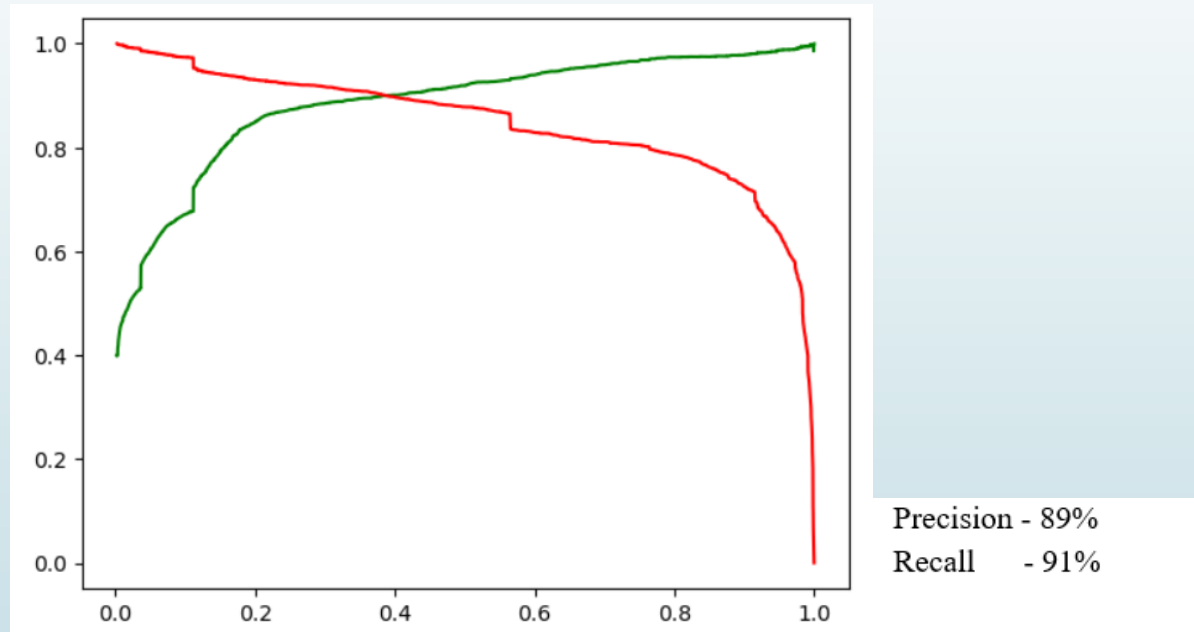
The graph depicts an optimal cutoff of 0.3 based on Accuracy, Sensitivity, Specificity.



- a) Accuracy : 92.30%
- b) Sensitivity : 91.69%
- c) Specificity : 92.68%

Model Evaluation- Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.41 based on Precision.





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 92%, 91% and 93% which are approximately closer to the respective values calculated using trained 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.