Lead Score Case Study

Submitted by :

Kushagra Saxena

Paras Jain

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

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

Feature Scaling and Splitting Train and Test Sets

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

Model Building

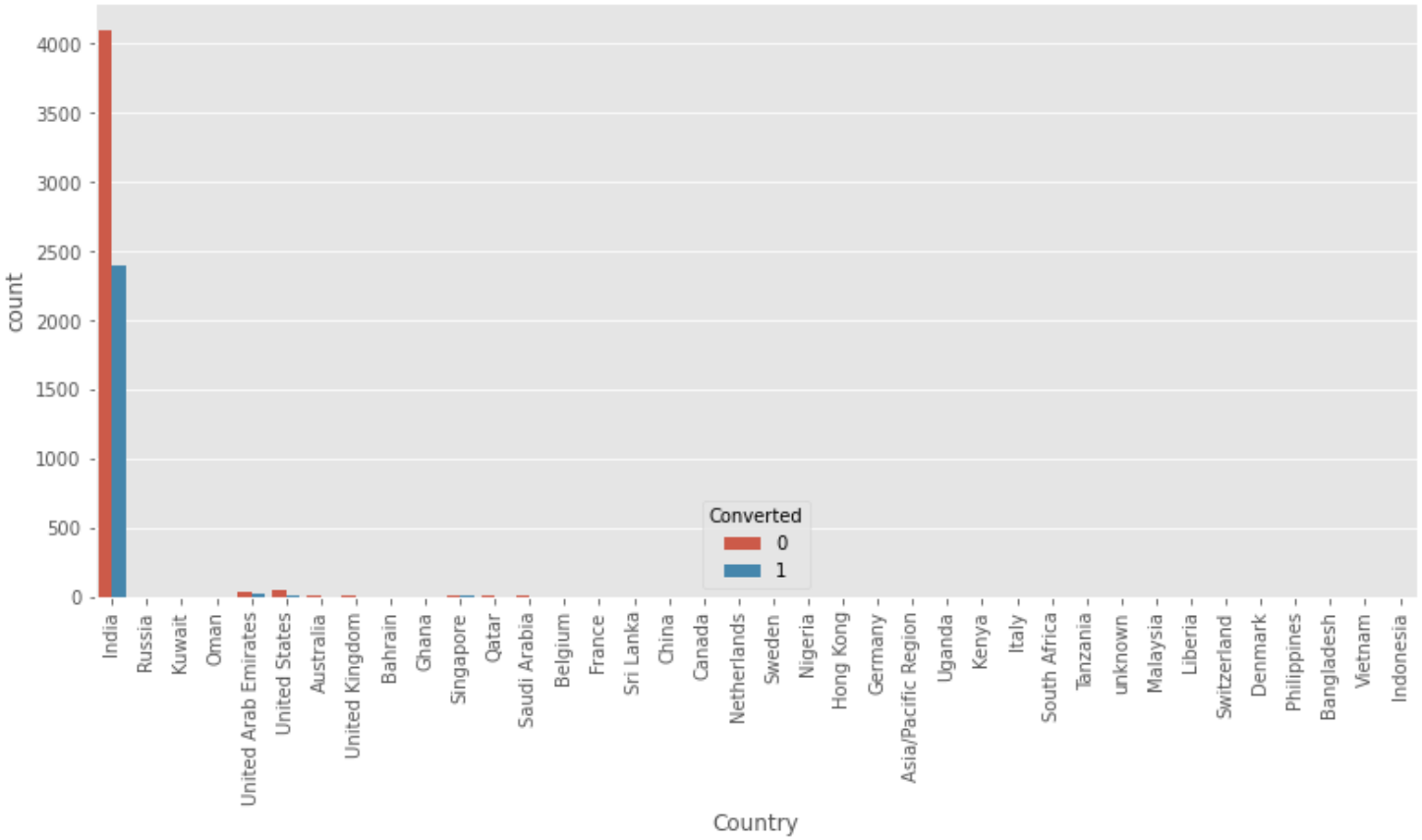
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

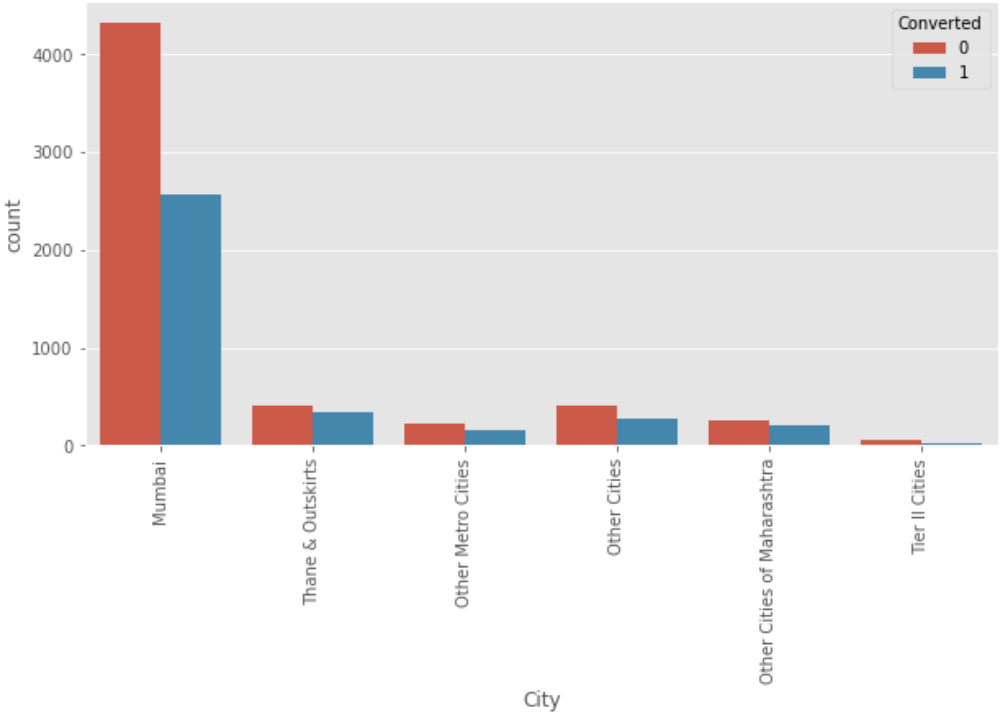
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Exploratory Data Analysis

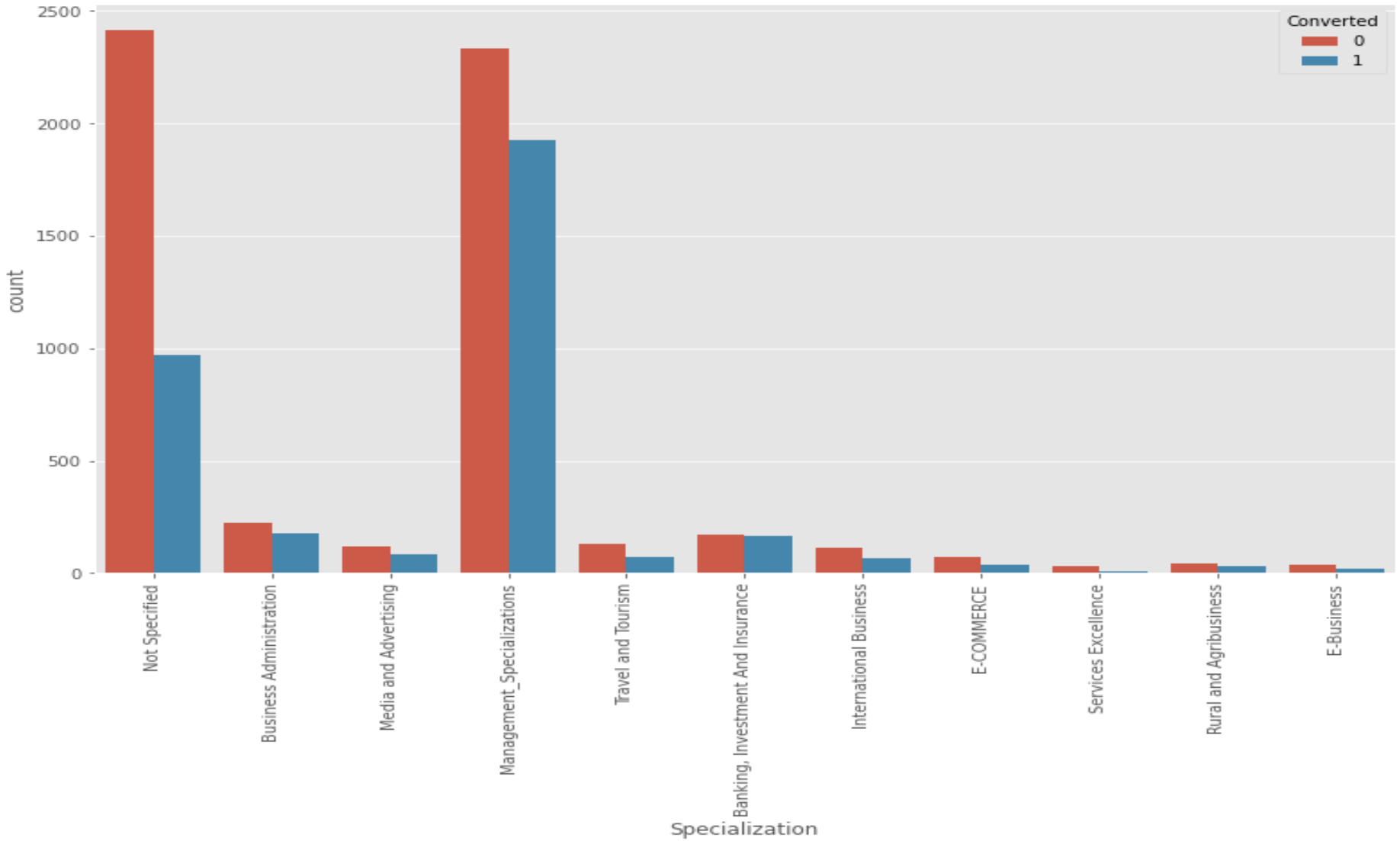
The conversion rates with respect to Country



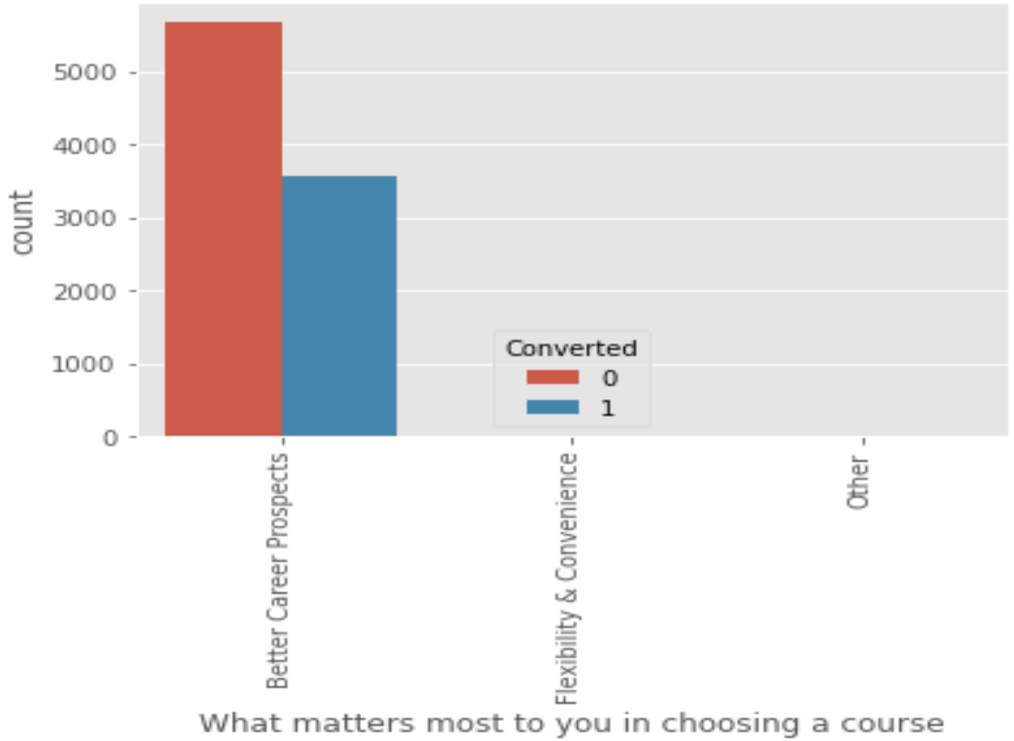
Conversion rate with respect to Cities in India



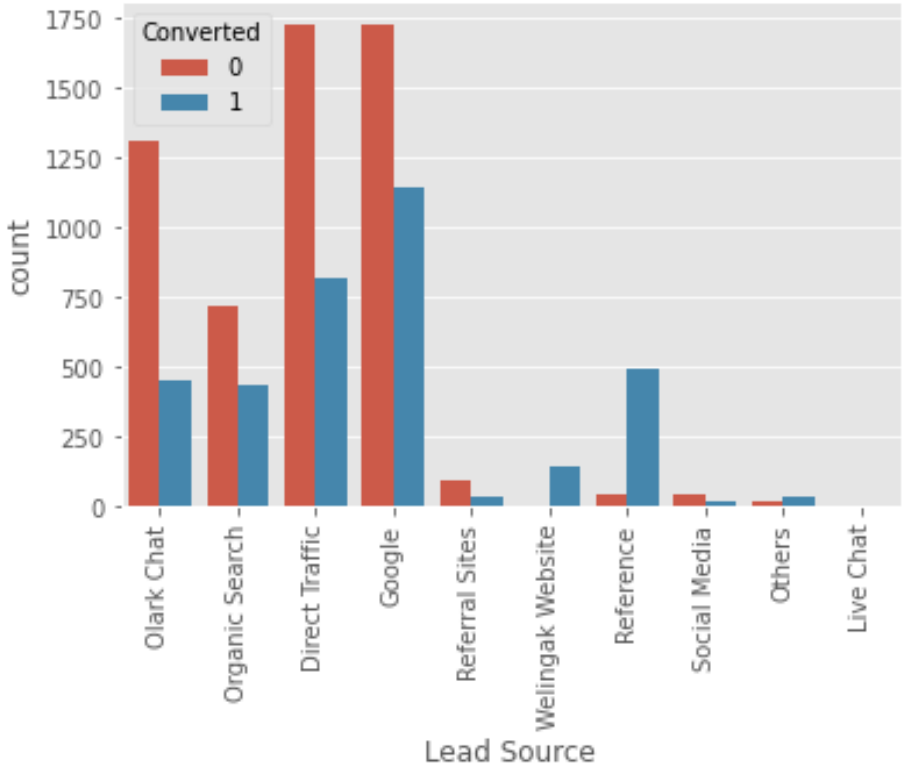
In Specialization, maximum conversion happened from Management_Specialization



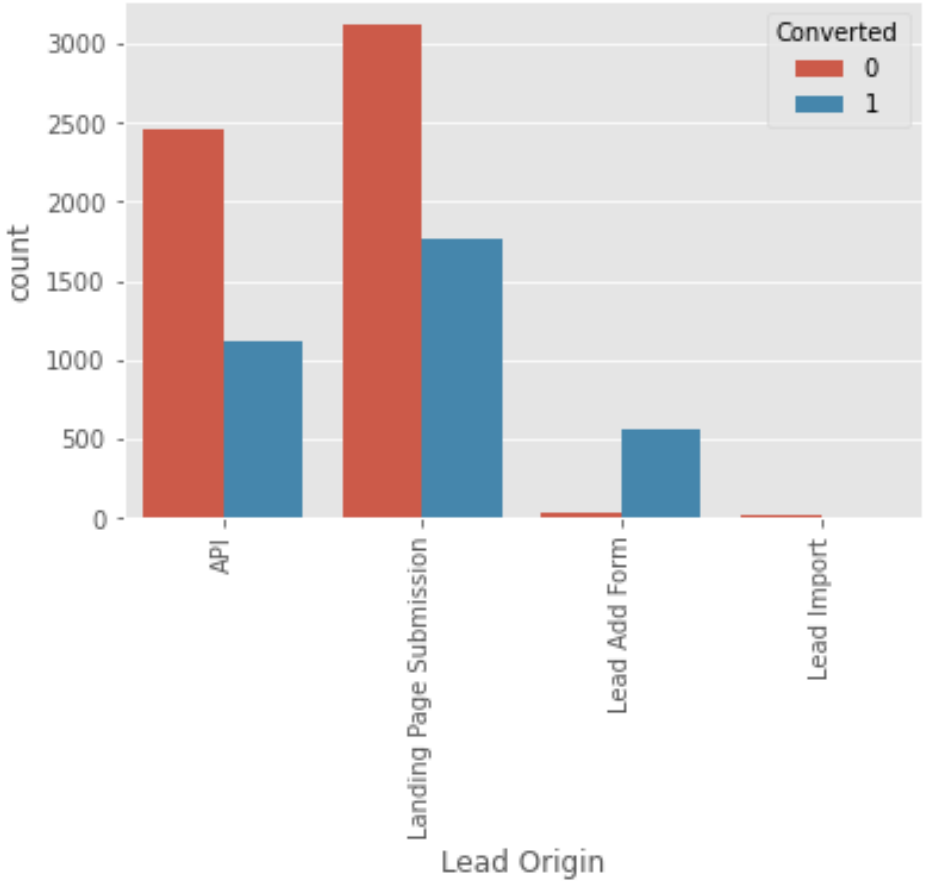
Very high conversion rate for people who aspire for ‘Better Career Prospects’



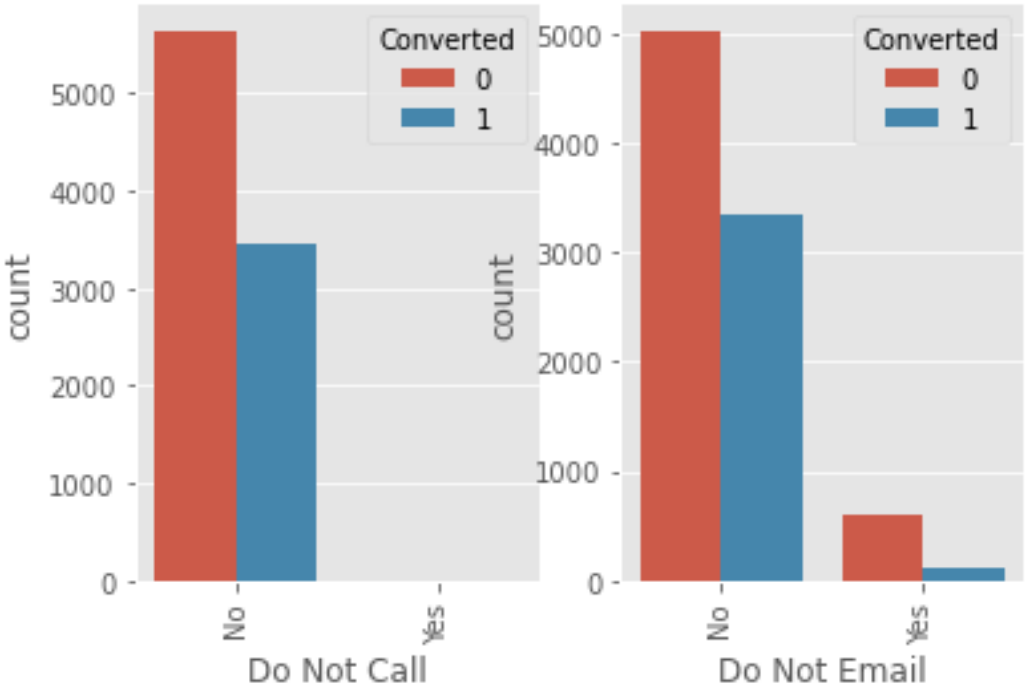
In Lead Source, Very high conversion rate for ‘Google’



In Lead Origin, Very high conversion rate reached and targets should be those who reached the Landing Page Submission



Major conversions happened where Emails and Calls are made.

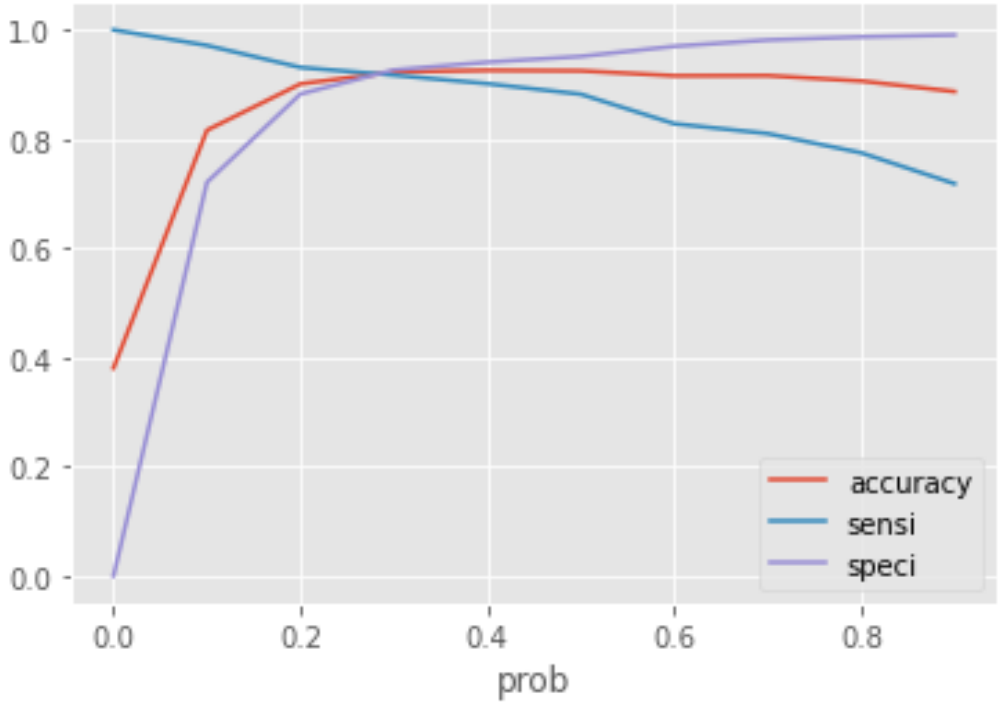


Variables Impacting the Conversion Rate

- 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 0.28 based on Accuracy, Sensitivity and Specificity



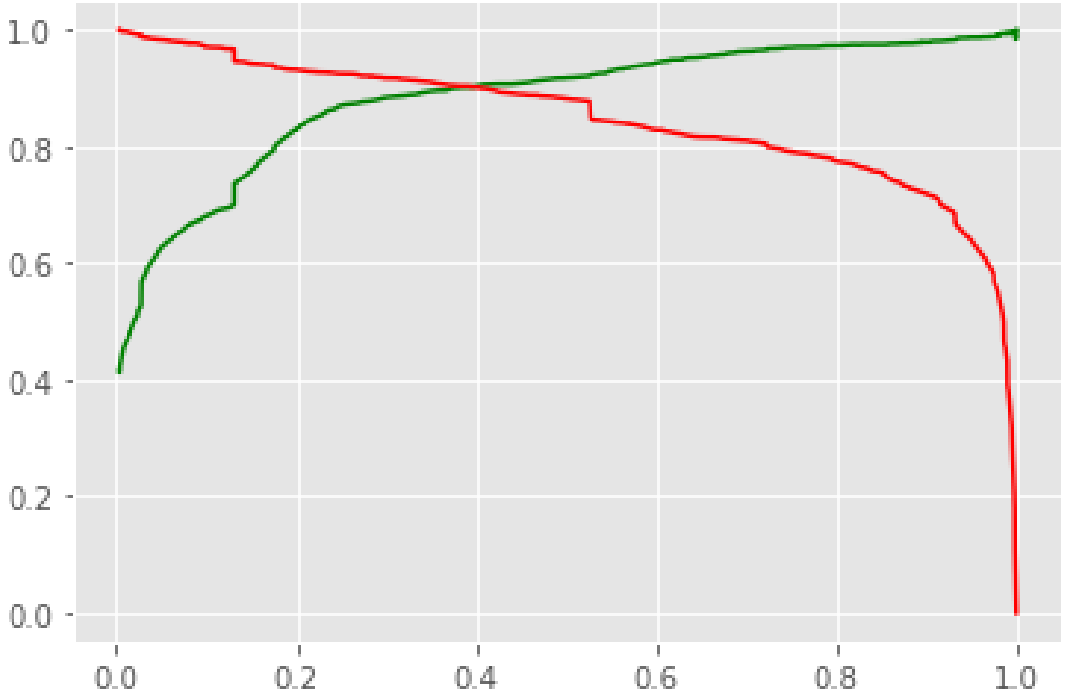
Confusion Matrix

3597	285
198	2187

- Accuracy – 92.29%
- Sensitivity – 91.70 %
- Specificity – 92.66 %
- False Positive Rate – 7.34 %
- Positive Predictive Value – 88.47 %
- Positive Predictive Value – 94.78%

Model Evaluation- Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.39 based on Precision and Recall



Confusion Matrix

1563	113
81	929

- Precision – 89.1 %
- Recall – 91.9 %
- Accuracy – 92.78 %
- Sensitivity – 91.98 %
- Specificity – 93.26 %

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. –
- The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model
- Train Data: Accuracy : 92.29% Sensitivity : 91.70% Specificity : 92.66%
- Accuracy, Sensitivity and Specificity values of Test Data: Accuracy : 92.78% Sensitivity : 91.98% Specificity : 93.26%which are approximately closer to the respective values calculated using trained set.
- The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are:
  - What is your current occupation_unemployed
  - Lead Origin _Landing Page Submission
  - Specialization_Management_Specialization
- The top three variables in our model which contribute most towards the probability of a lead getting converted are:
  - Specialization
    - Management_Specialization
    - Banking, Investment and Insurance
  - What is your current occupation
    - Unemployed
  - What matters most to you in choosing a course
    - Better Career Prospects

Hence overall this model seems to be good.