Lead Score Case Study

By:

Abhik Gupta
Mayur Singhal
Amruta Kapse

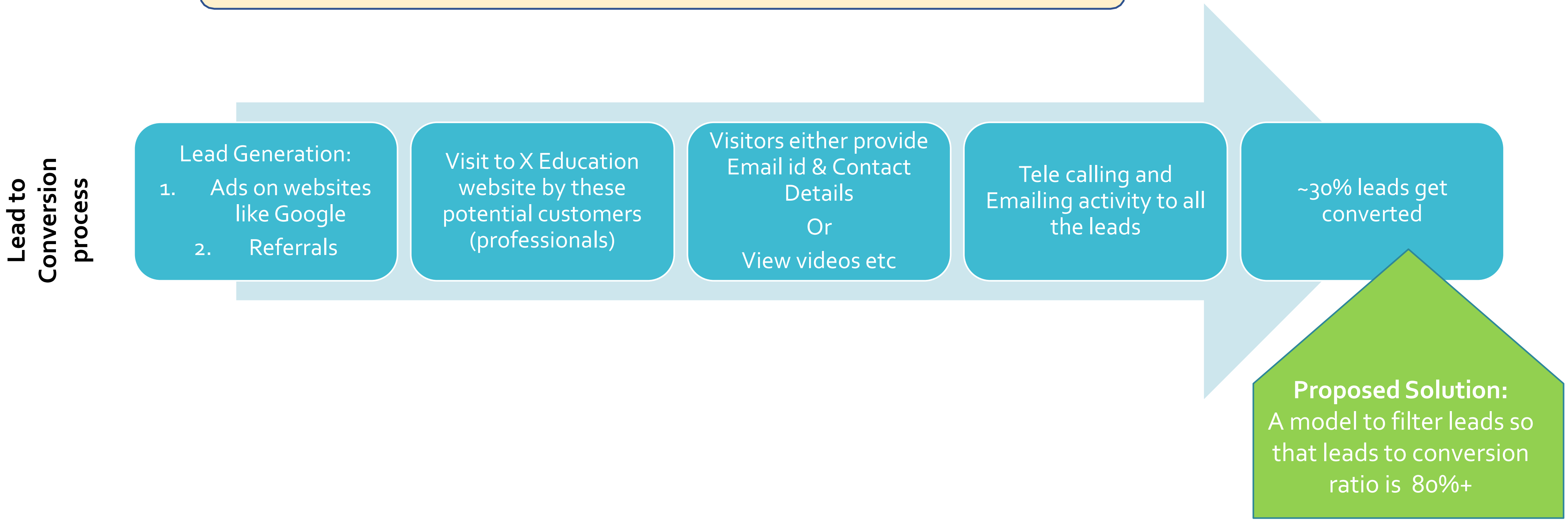
Agenda

- Problem Statement & Business Understanding
- Proposed Solution
- Strategy and Problem Solving Methodology
- Data Cleaning and Preparation
- Data Imbalance
- Exploratory Data Analysis
- Variables Impacting the Conversion Rate
- Model Evaluation - Sensitivity and Specificity on Train and Test Data Set
- Model Evaluation - Precision and Recall on Train and Test Data Set
- Conclusions and Recommendations

Problem Statement

- To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher lead score have a higher conversion chance and the leads with lower lead score have a lower conversion chance.

Business Understanding



Lead to Conversion process

Lead Generation:

1. Ads on websites like Google
2. Referrals

Visit to X Education website by these potential customers (professionals)

Visitors either provide Email id & Contact Details

Or

View videos etc

Tele calling and Emailing activity to all the leads

~30% leads get converted

Proposed Solution:

A model to filter leads so that leads to conversion ratio is 80%+

Proposed Solution

Selection of Hot Leads

Communicating with Hot Leads

Conversion of Hot Leads

Leads Clustering

We can categorize the leads based on their Lead Score or probability to convert, resulting in a smaller group of hot leads to focus on.

Focus Communication

We might have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

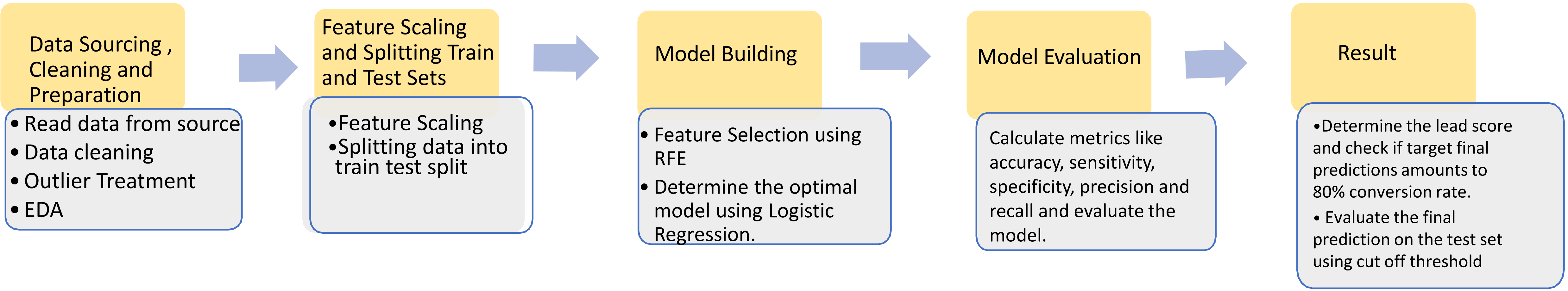
Increase conversion

We could achieve the 80% target if we focused on hot leads that were more likely to convert.

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.

Problem Solving Methodology



Data Sourcing , Cleaning and Preparation

- Read data from source
- Data cleaning
- Outlier Treatment
- EDA

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling
- Splitting data into train test split

Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression.

Model Evaluation

Calculate 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

Data Cleaning and Preparation

- Following columns contain more than 45% null values initially:
  - How did you hear about X education
  - Lead Profile
  - Lead Quality
  - Asymmetrique Activity Index
  - Asymmetrique Profile Index
  - Asymmetrique Activity Score
  - Asymmetrique Profile Score

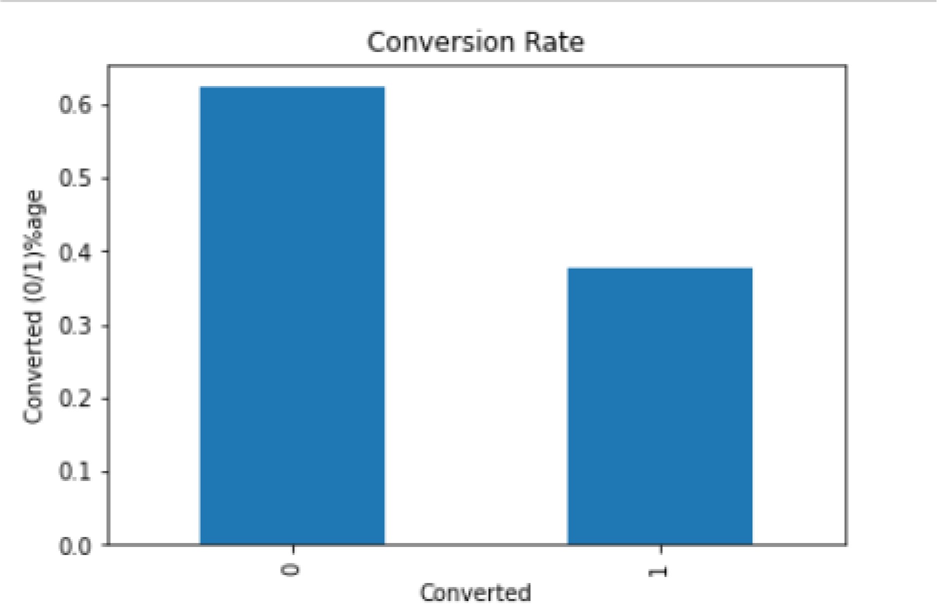
- Following columns have default value of ‘select’ as a dominating value which is same as null value. So, we have converted ‘select’ to ‘NA’.
  - Specialization
  - How did you hear about X Education
  - Lead Profile
  - City

- Following columns have less than 2% nan values. We can afford to drop their rows altogether
  - Last Activity
  - Lead Source
  - Page Views Per Visit

Data Cleaning and Preparation

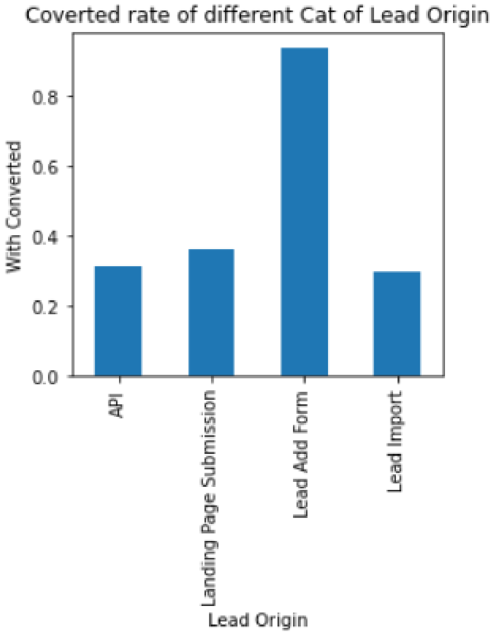
- Following columns have been dropped which contain single value as their contribution is insignificant and causing data imbalance:
  - Magazine
  - Receive More Updates About Our Courses
  - Update me on Supply Chain Content
  - Get updates on DM Content
  - I agree to pay the amount through cheque

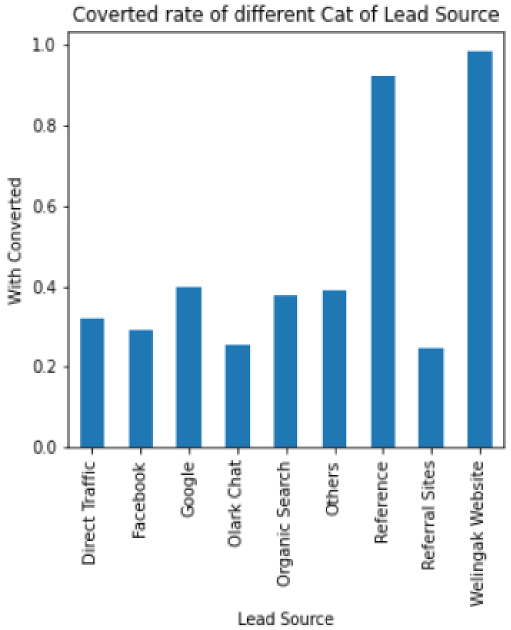
Data Imbalance

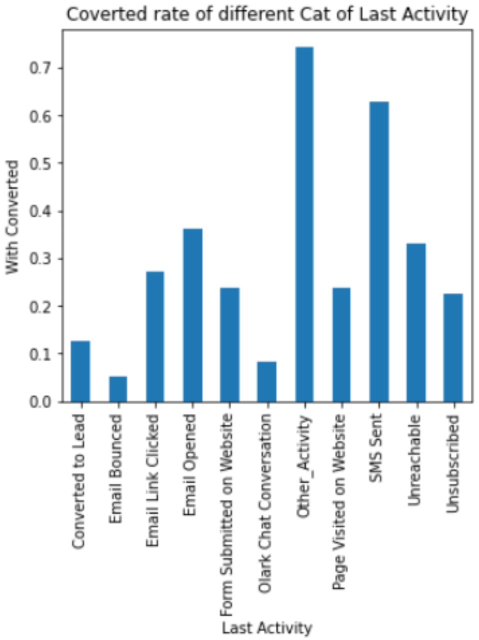


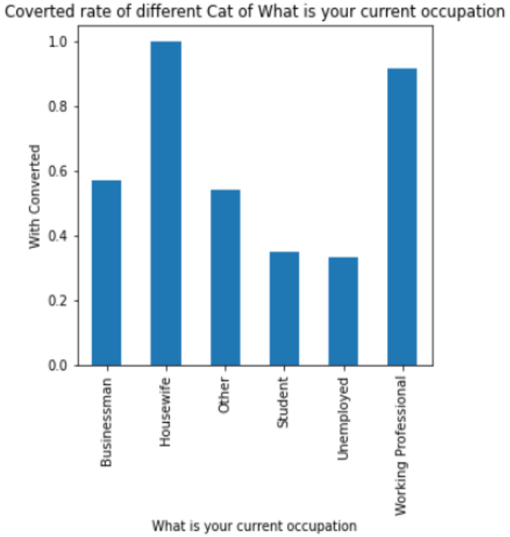
- Inference
  - We have around 38% Conversion rate in total
  - However after model building, the conversion rate is expected to be increased by 80%

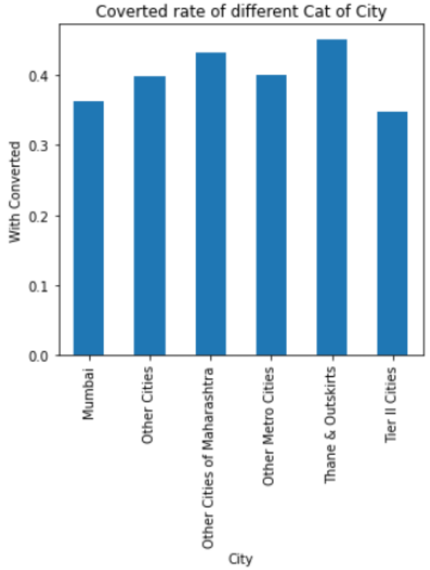
Exploratory Data Analysis

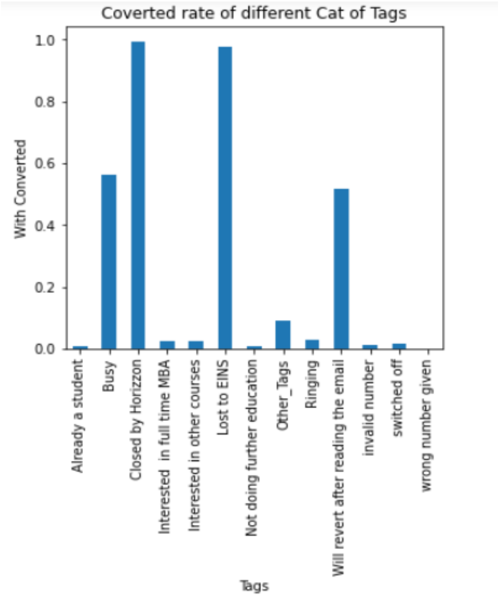

Coverted rate of different Cat of Lead Origin


Coverted rate of different Cat of Lead Source


Coverted rate of different Cat of Last Activity


Coverted rate of different Cat of What is your current occupation

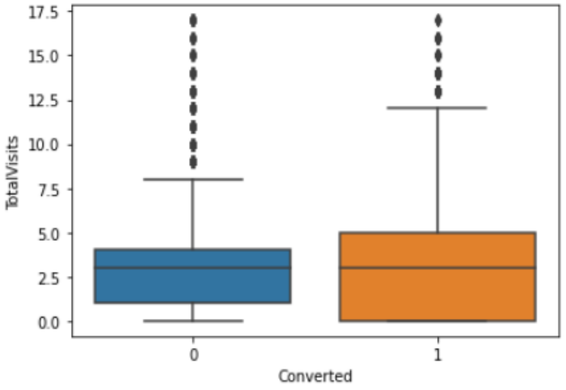

Coverted rate of different Cat of City

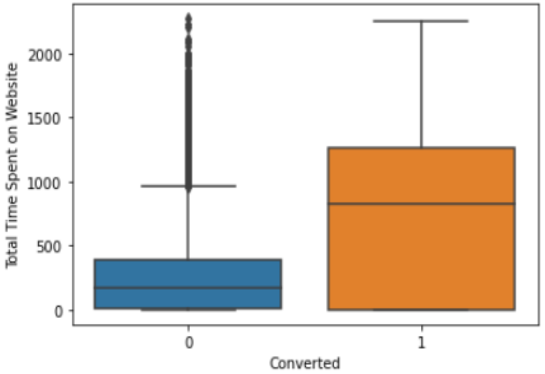

Coverted rate of different Cat of Tags

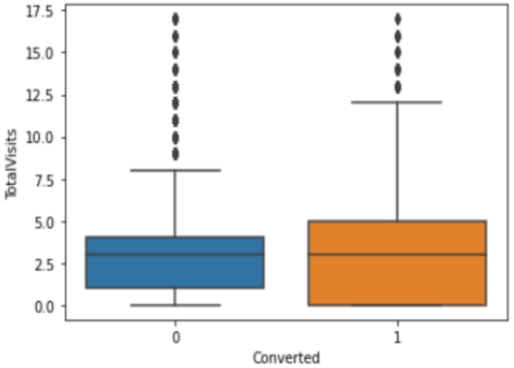
Key Takeaways

- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable
- Conversion Rate of reference leads and leads through welingak website is high.
- Google and Direct traffic generates maximum number of leads
- Most of the lead have their Email opened as their last activity
- Conversion rate for leads with last activity as SMS Sent is almost 60%
- Working Professional have high conversion rate
- Unemployed leads are high in number but has 30-35% conversion rate
- Most leads are from Mumbai with around 30% conversion rate
- In Tags closed by horizon has high conversion rate of 99%

Exploratory Data Analysis – Outlier Treatment







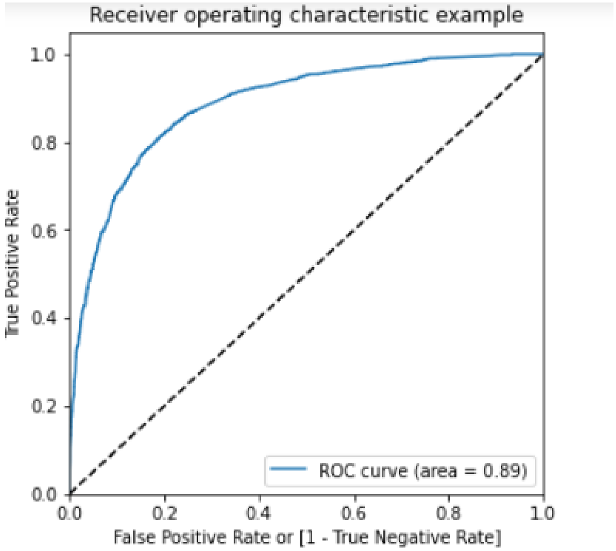
Key Takeaways

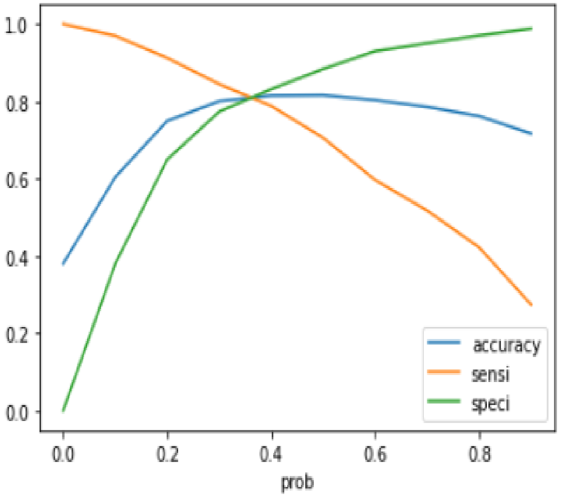
- Median for converted and not converted leads are the same for Total Visits
- Leads spending more time on the website are more likely to be converted.
- Median for converted and non-converted leads is the same for
- Inter Quantile Range (IQR) method has been used to treat outliers in the data.
- The higher values has been capped to 99%

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 – Working Professional
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity - Unreachable

Model Evaluation - Sensitivity and Specificity on Train and Test Data Set





- On plotting ROC curve , we see AOC is around 0.89
- Having probability 0.35 indicates , any conversion probability >35% is said to be converted to lead

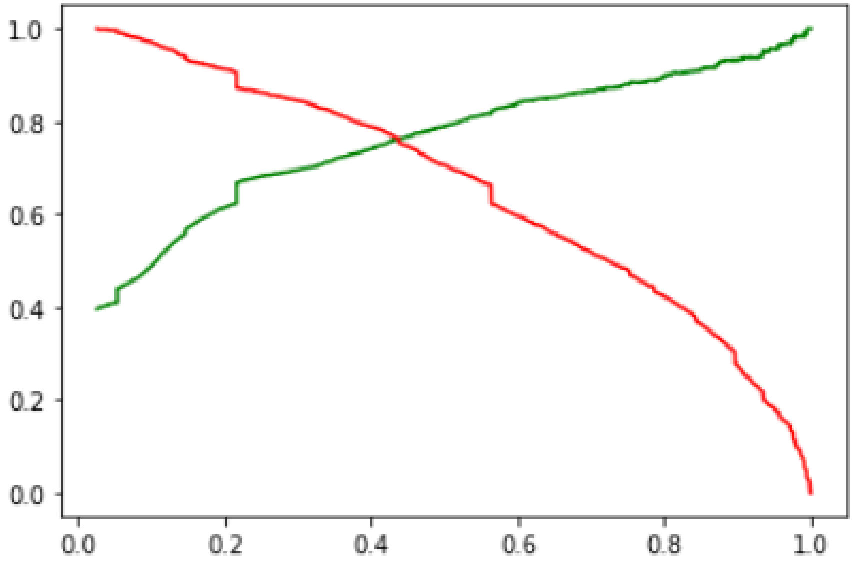
Model Performance

- Train Set
  - Accuracy - 80.88%
  - Sensitivity - 81.76%
  - Specificity- 80.34%.

- Test Set
  - Accuracy- 80.05%
  - Sensitivity- 80.38%
  - Specificity- 79.86%

Model Evaluation - Precision and Recall on Train and Test Data Set

- The graph depicts an optimal cut off of 0.44 based on Precision and Confusion Matrix Recall



Model Performance

- Train Set
  - Accuracy - 81.60%
  - Precision - 76.19%
  - Recall - 75.07%

- Test Set
  - Accuracy- 80.58%
  - Precision 74.58%
  - Recall- 73.65%

Conclusion and Recommendations

- The calculated lead score indicates that the conversion rate on the final predicted model on the test set is around 78% (near to target 80%)
- The top variables that contribute towards lead conversion are:
    - Lead Source_Welingak Website
    - Total Time Spent on Website
    - What is your current occupation_Working Professional
    - Last Notable Activity_SMS Sent
- We may note the following points to consider an individual a potential lead:
    - If they spend a lot of time on website and this can be done by making the website interesting and thus bringing them back to site
    - If their last activity is through SMS or through Olark chat conversation
    - If they are working professionals
- It’s good to collect data often and run the model and get updated with the potential leads. Since best time to call your potential lead is within few hours after the lead shows interest in courses
- Also based on the business scenario, you can modify your lead score threshold. If you want to aggressively target customers you can consider threshold of 35-40%. On the other hand If targets are already met you can keep threshold of 75-80%
- However, Focusing on hot leads will increase the chances of obtaining more value to the business as number of people we contact are less but the conversion rate is high