Lead Scoring Case Study

Presented by:
Bharath Konda
Surya Verma
Treesa John

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

The typical lead conversion rate at X education is around 30%.

The company wants to increase the lead conversion rate to 80% or more by identifying the most potential leads (Hot Leads)

The lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
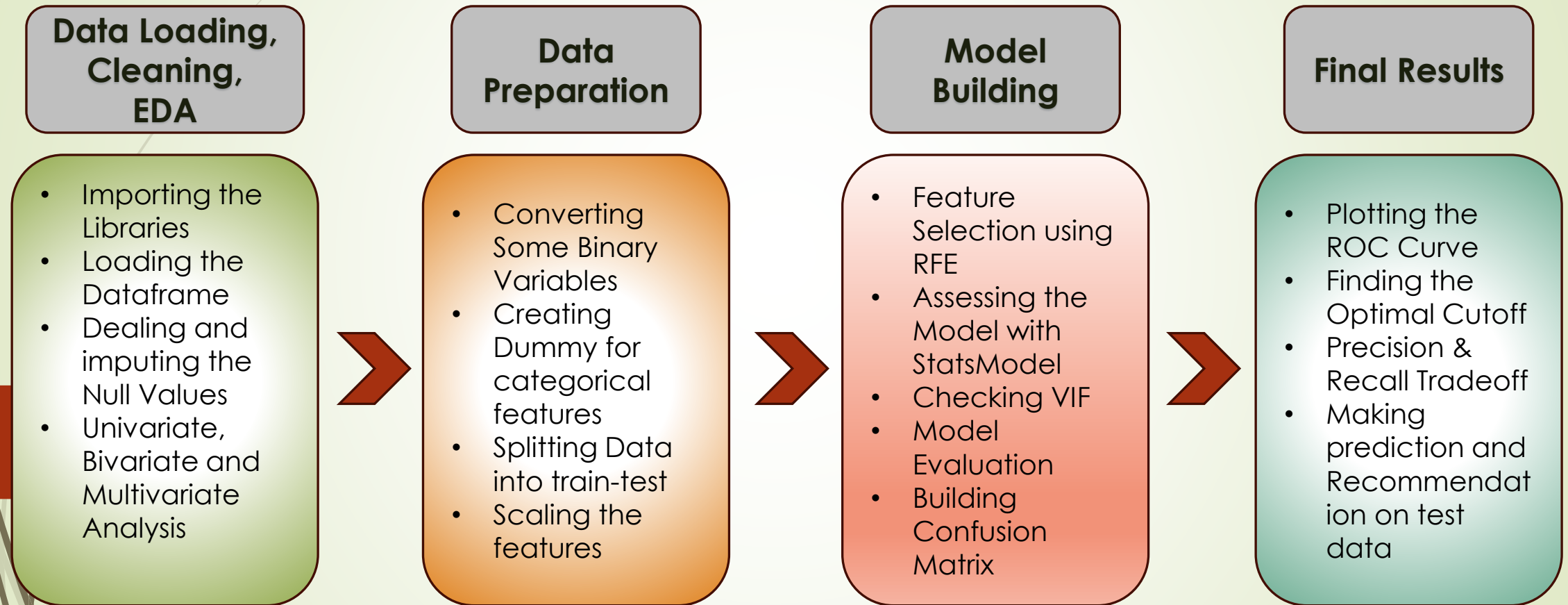
Business Goal

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
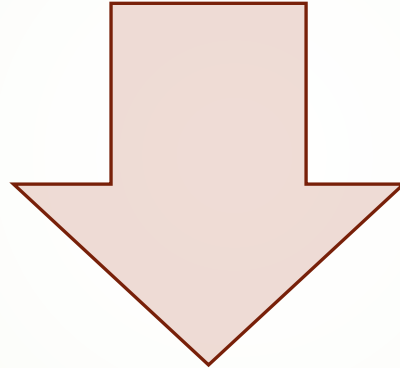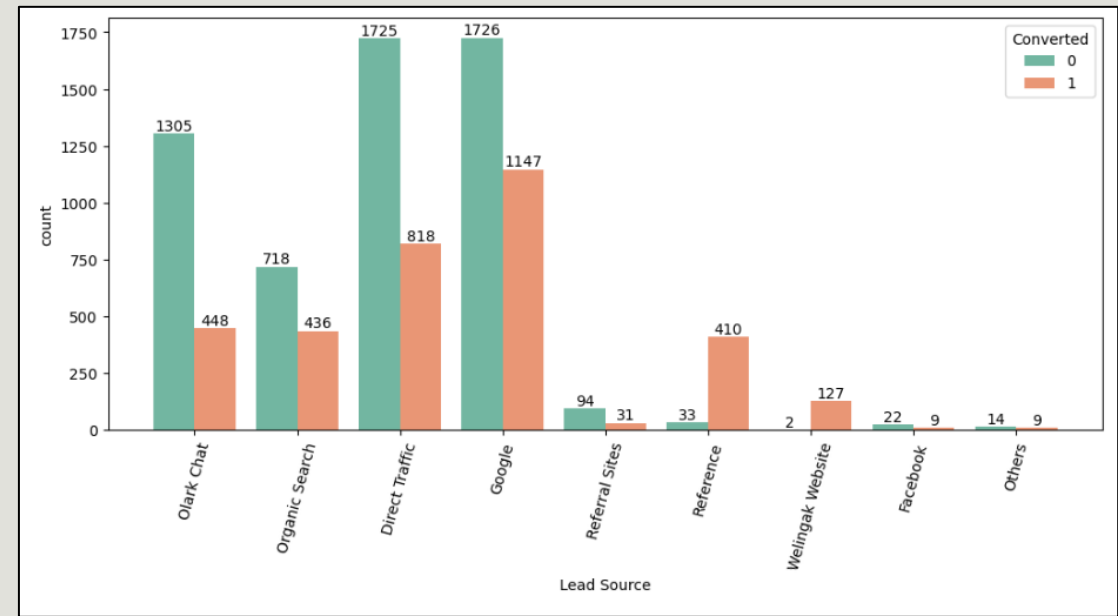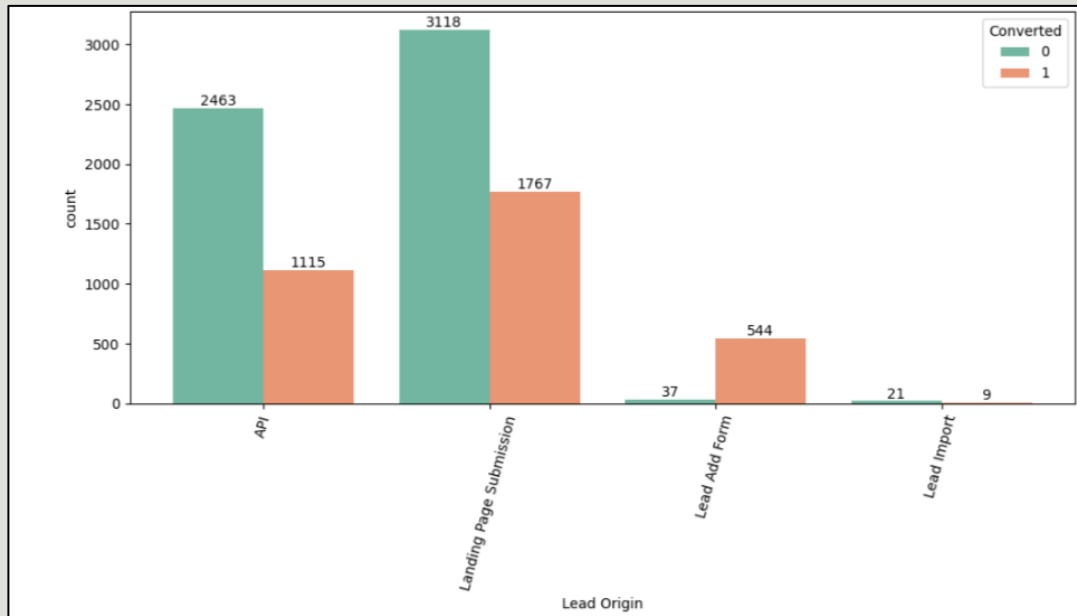
A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Problem Solving Methodology

Data Loading, Cleaning, EDA

- Importing the Libraries
- Loading the Dataframe
- Dealing and imputing the Null Values
- Univariate, Bivariate and Multivariate Analysis

Data Preparation

- Converting Some Binary Variables
- Creating Dummy for categorical features
- Splitting Data into train-test
- Scaling the features

Model Building

- Feature Selection using RFE
- Assessing the Model with StatsModel
- Checking VIF
- Model Evaluation
- Building Confusion Matrix

Final Results

- Plotting the ROC Curve
- Finding the Optimal Cutoff
- Precision & Recall Tradeoff
- Making prediction and Recommendation on test data

Data Analysis: Univariate, Bivariate, Multivariate

- Leads originating from API and Landing Page Submission are considerable high but their conversion rate is low.
- Although the number of leads from 'Lead Add Form' are less, their conversion rate is high.
- Lead Import are very less in count.

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.
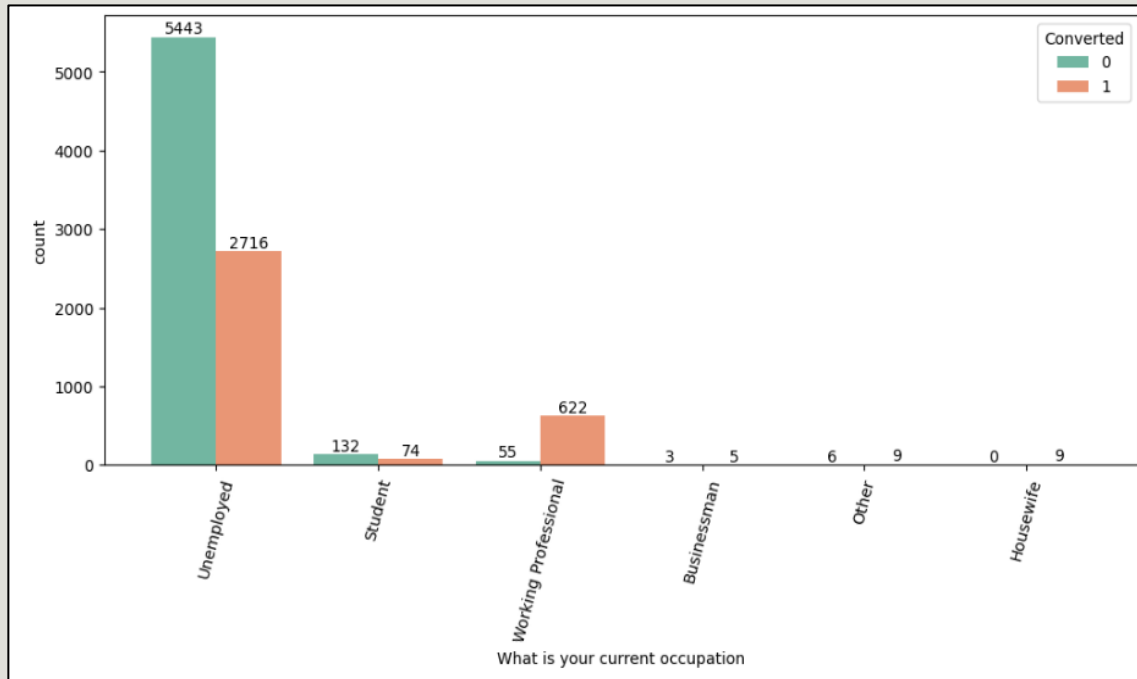
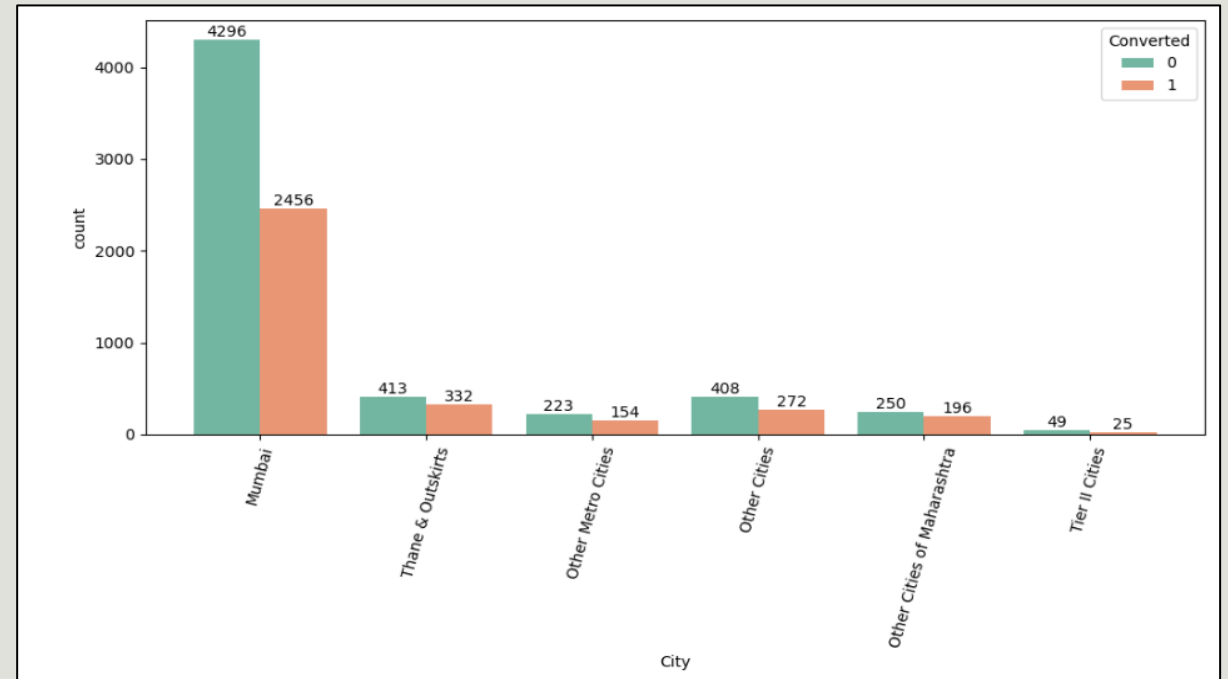# Data Analysis: Univariate, Bivariate, Multivariate
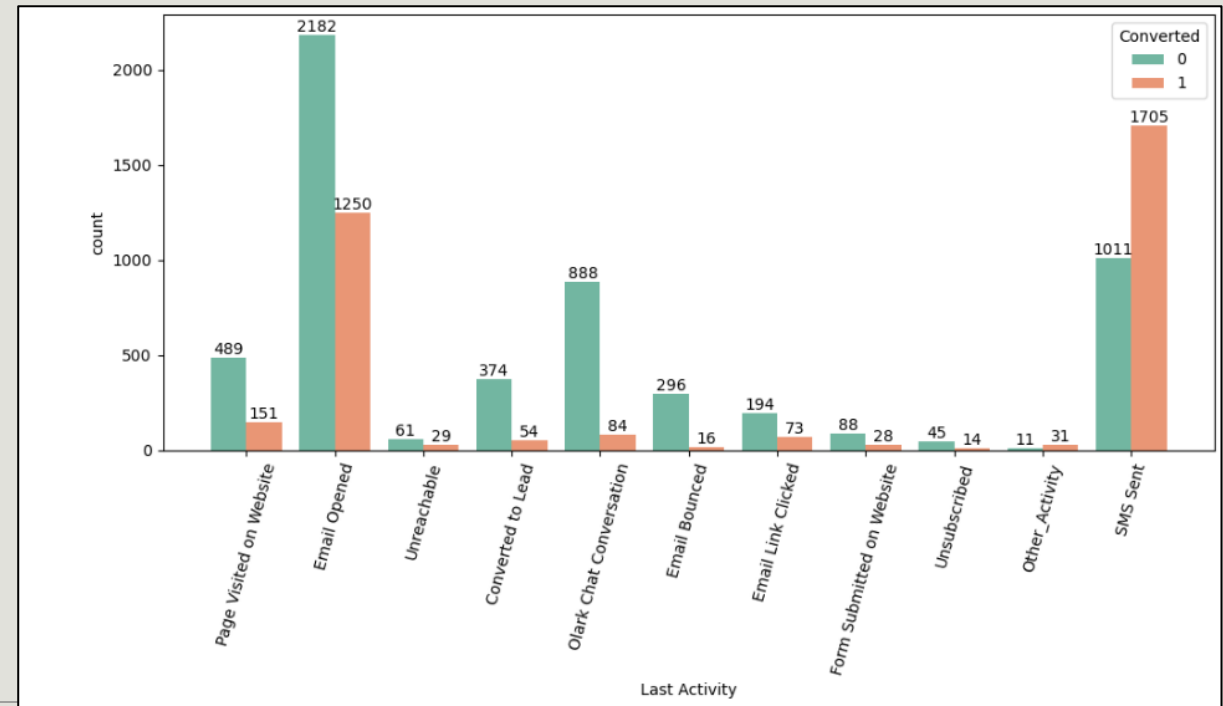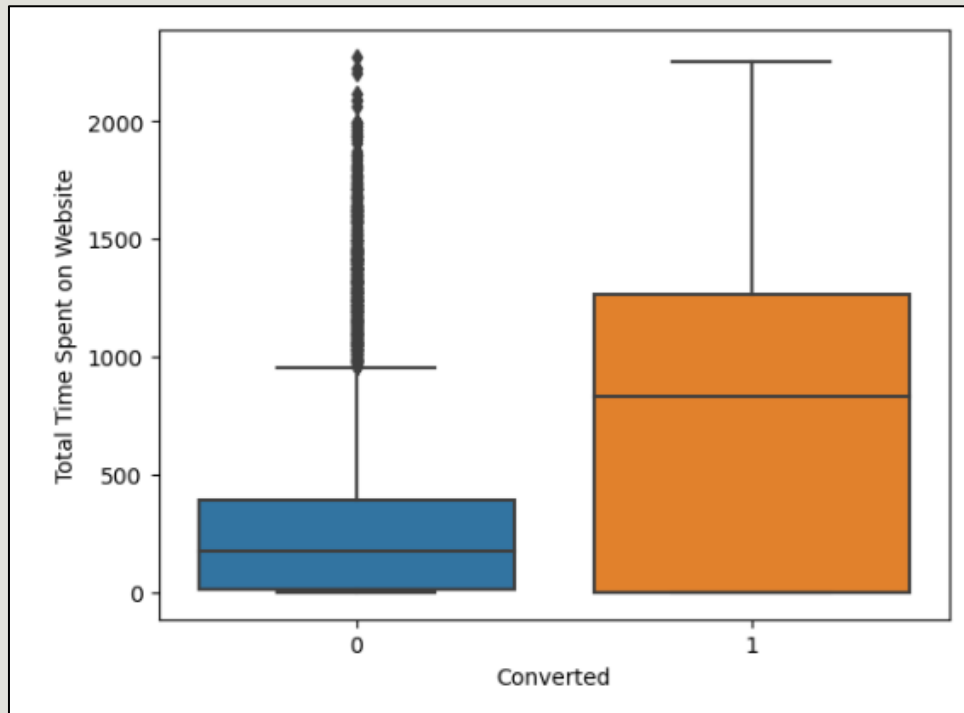
- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.

- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is the most.

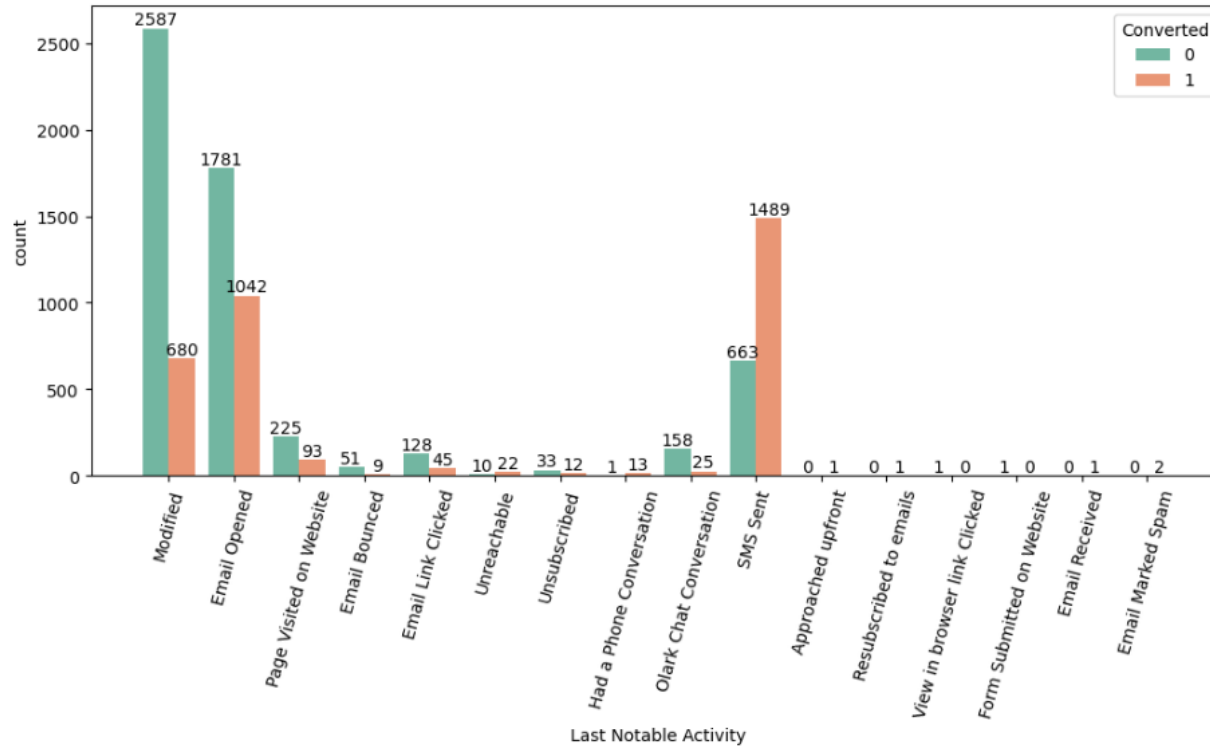# Data Analysis: Univariate, Bivariate, Multivariate

- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in numbers but has around 30-35% conversion rate.

- Most leads are from Mumbai with around 50% conversion rate.

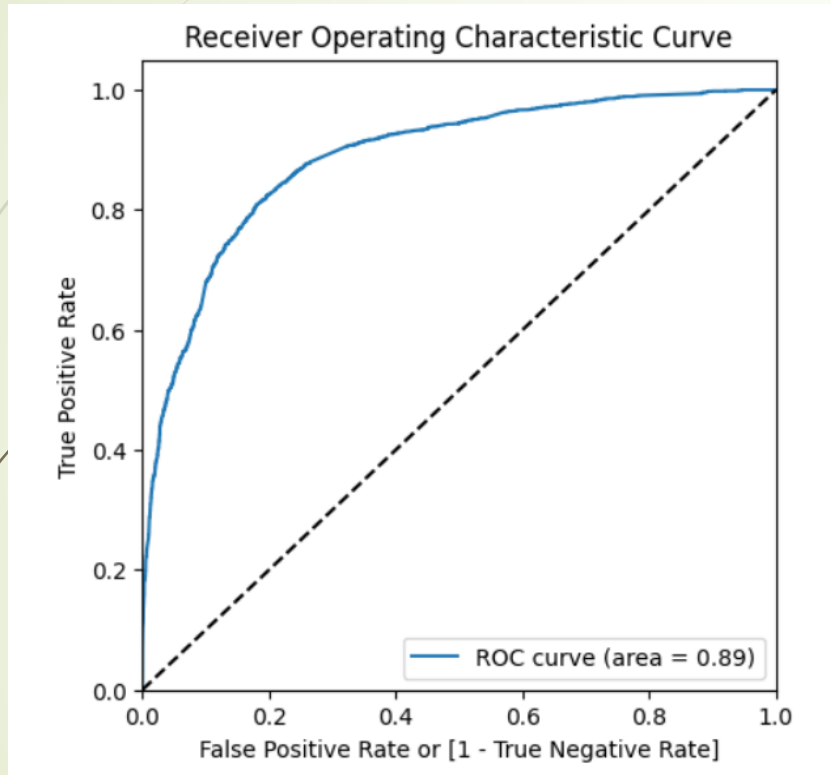# Data Analysis: Univariate, Bivariate, Multivariate

- Conversion rate for leads whose last notable Activity was 'SMS Sent' seems to be high

- Most leads are from Mumbai with around Correlation between 'Total Visits' and 'Page Views Per Visit' is 0.77

ROC Curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.
An ROC curve is a popular method for measuring the accuracy of a classification model.

ROC Curve show tradeoff between sensitivity and specificity.

We have achieved higher (0.89) area under the ROC curve which means better overall classification performance

Optimal Cut-off Point for a Hot Lead
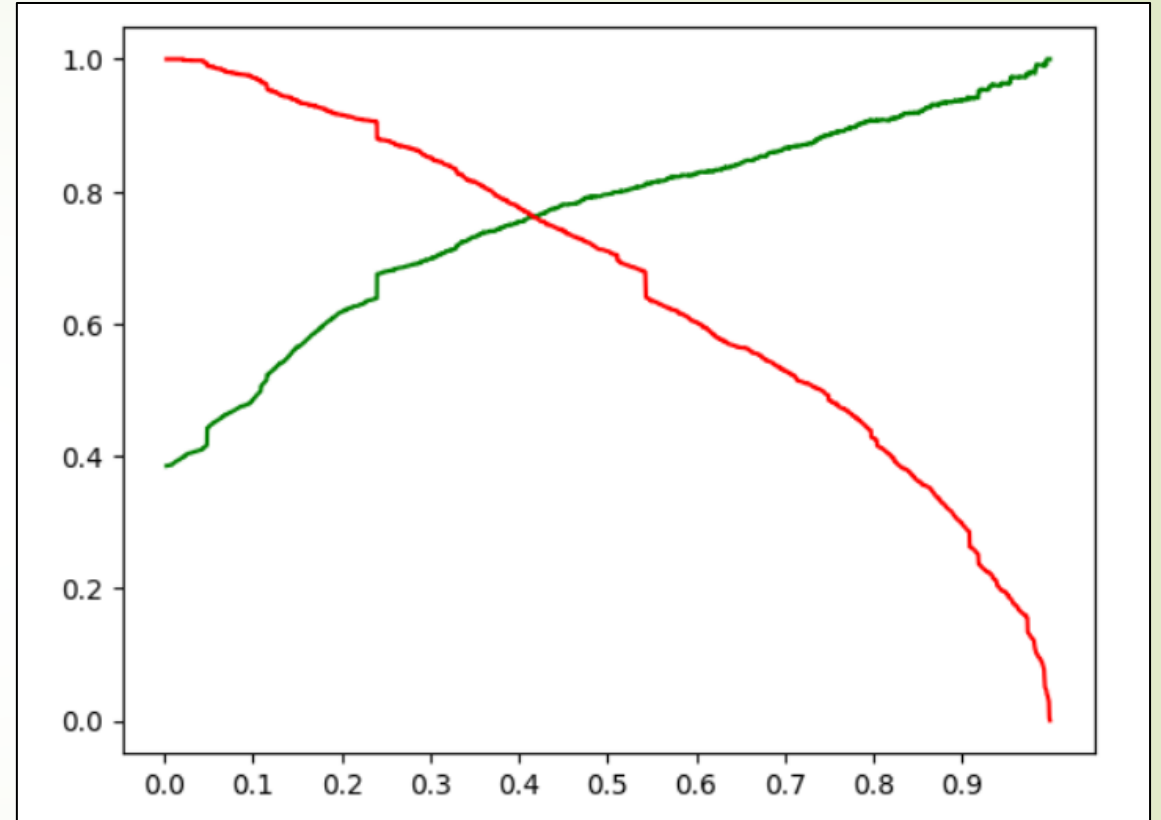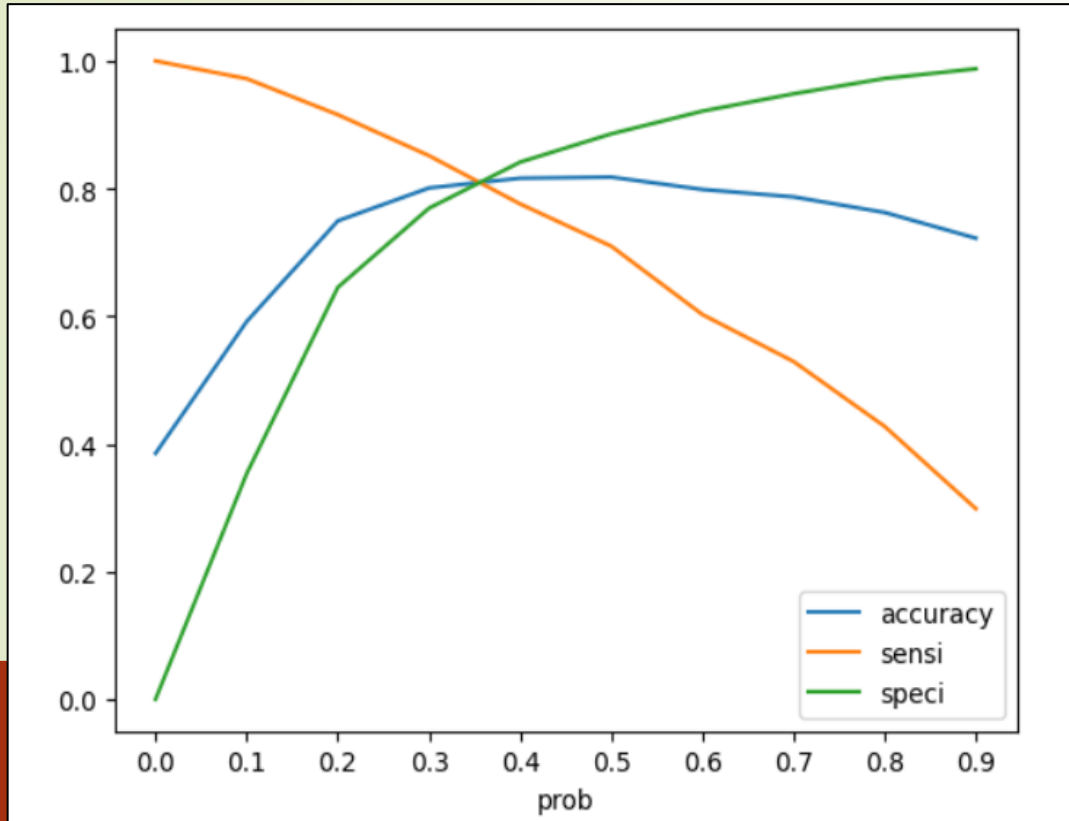
Cutoff point as per Sensitivity, Specificity and Accuracy is 0.34 and Cutoff point as per Precision and Recall is 0.4. We have proceeded with 0.34 Cutoff as it gives the Higher Recall value

Final Results

Final Important Features

Train Data
Accuracy : 81.04%
Sensitivity : 81.93%
Specificity : 80.49%

Test Data
Accuracy : 80.39%
Sensitivity : 80.49%
Specificity : 80.33%

- ➢ Lead Source_Welingak Website
- ➢ Lead Source_Reference
- ➢ What is your current occupation_Working Professional
- ➢ Last Activity_Other_Activity
- ➢ Last Activity_Unsubscribed
- ➢ Last Activity_SMS Sent
- ➢ Total Time Spent on Website
- ➢ Lead Source_Olark Chat
- ➢ Last Notable Activity_Modified
- ➢ Last Activity_Olark Chat Conversation
- ➢ Lead Origin_Landing Page Submission
- ➢ Specialization_Others
- ➢ Do Not Email

Conclusion

For higher conversion of leads, calls should be made to

- Leads coming from Welingak Websites and Reference
- Working Professionals
- Leads who spent more time on website
- Leads whose last activity was SMS sent & Olark Chat Conversation

Calls can be ignored to leads with

- Lead origin is Landing Page Submission
- Lead specialization mentioned as Others
- Lead who have chosen Yes for Do not Email