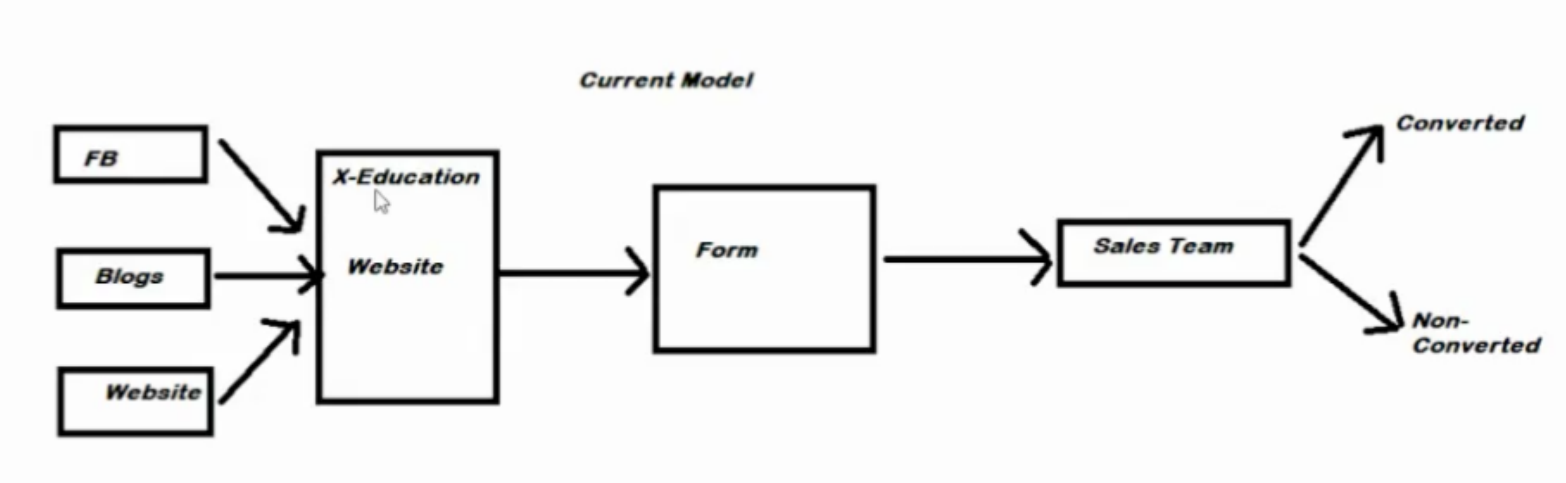
Lead Scoring Case Study

By
Sourav Acharya
Kartikey Pandey

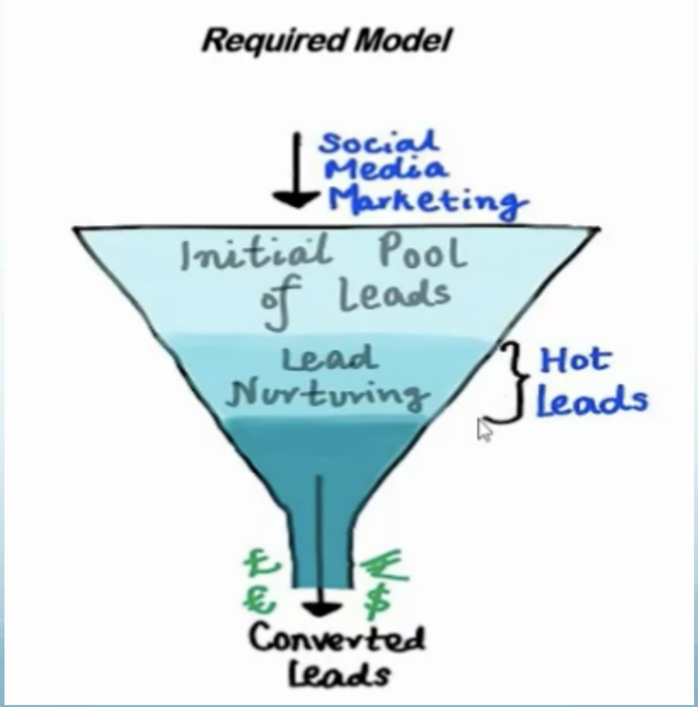
Problem Statement Background

- An education company named X Education sells online courses to industry professionals.
- It gets its Leads from –
  - through forms filled by people on website
  - Referrals
- X Education gets a lot of leads, but its lead conversion rate is very poor



Problem Statement

- Find potential candidate and Follow-up with them and convert them
- X Education would like to select the most promising leads.
- The company requires a model wherein 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.
- Ballpark of the target lead conversion rate to be around 80%.



Expectation

- 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.
- Find out areas of focus and categories which can help increasing lead conversion

Plan of Action

- Read and understand the data
- Clean the data
- Prepare the data for modeling
  - create dummies for all categorical variable
  - prepare test and train split
  - scaling
- Modeling
  - use RFE
  - build logistic regression
  - check p-value and vif
  - find optimal probability cutoff
  - check performance of model on test data
  - generate score variable
- Final analysis and recommendation

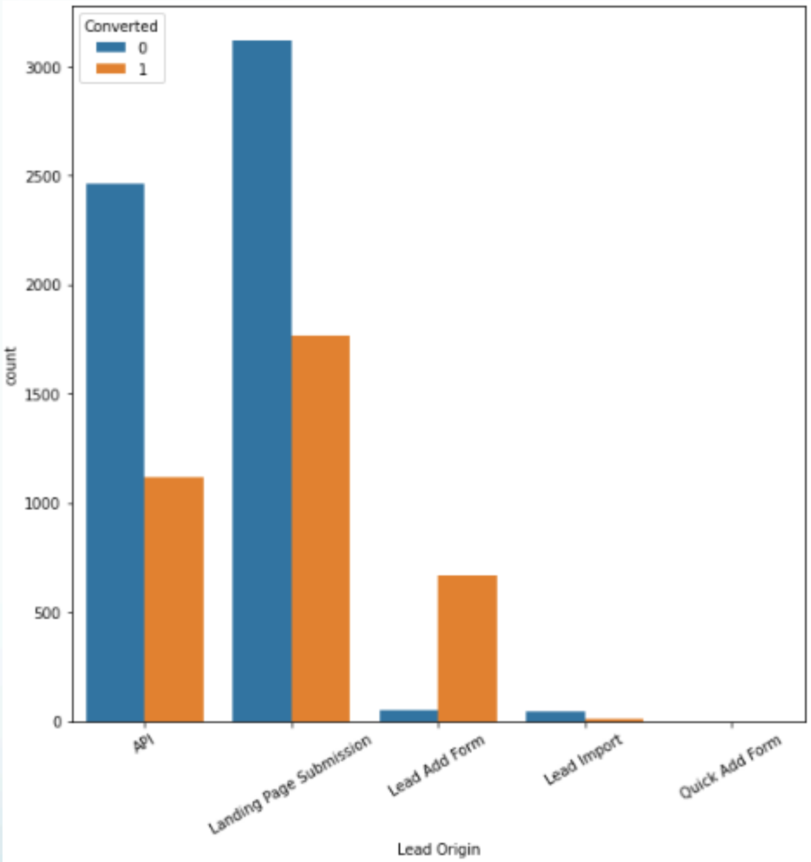
Reading Data

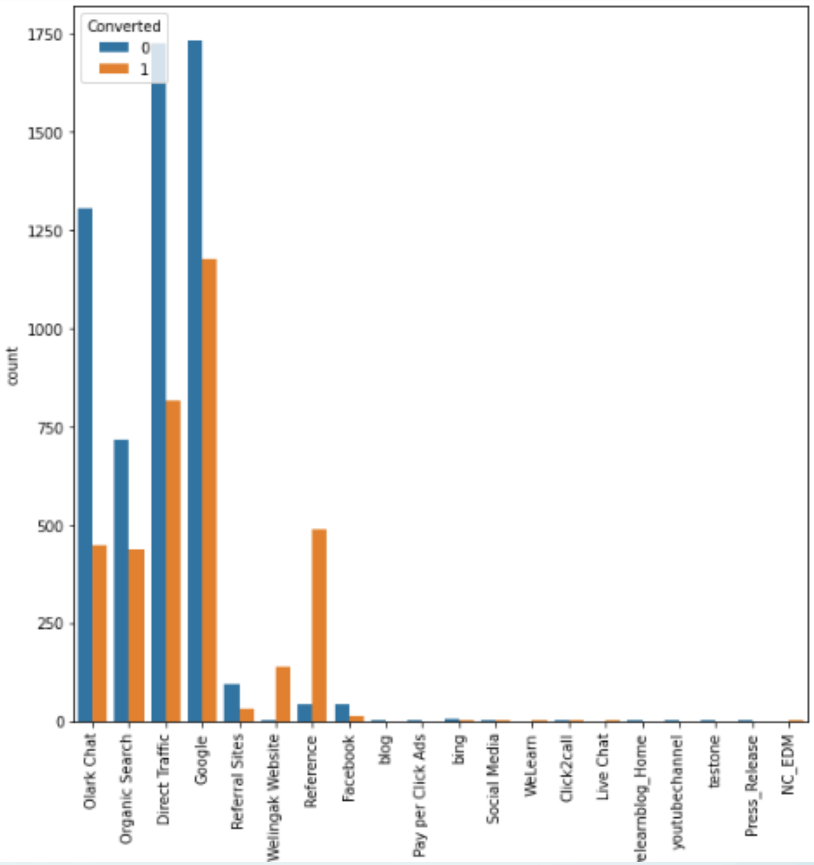
- Csv provided - Leads.csv
- Data numbers –
  - No. of rows: 9240
  - No. of columns: 37

Cleaning of Data

- There were missing data for 17 columns
- Some Columns like Specialization contained value ‘Select’. This got populated as user did not select any option from the list under that column. So, these were replaced with NAN
- Any column with more than 75% of missing data where dropped.
- Columns with duplicate or redundant data were dropped.
  - Prospect ID and Lead Number both contained unique value so one of them was dropped
- Null values in Column were filled values after analyzing data for that column
  - Column like Lead Source, missing value was replaced by mode for that column
  - Columns like TotalVisits, missing value was replaced by mean of that column
  - Columns like Country, missing value was replaced by mode but was kept for dropping in future as it was imbalance data
  - Columns like ‘What matters most to you in choosing a course’ which were highly imbalance were dropped.
- Final data after cleanup – 9240 rows and 32 columns

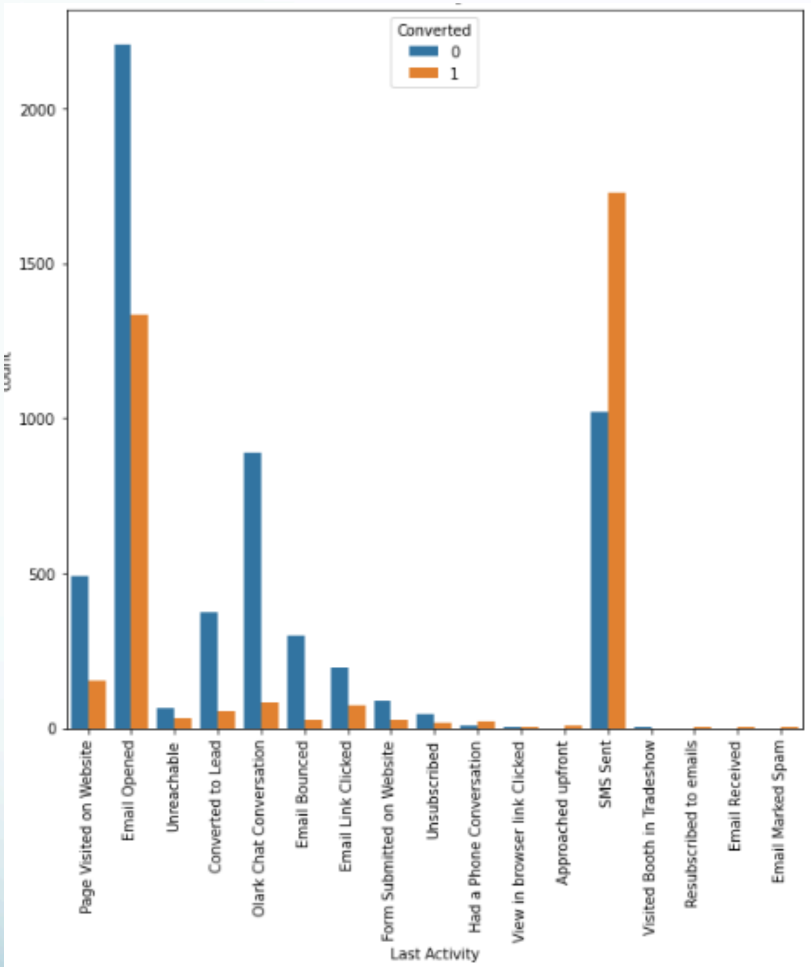
Data Analysis

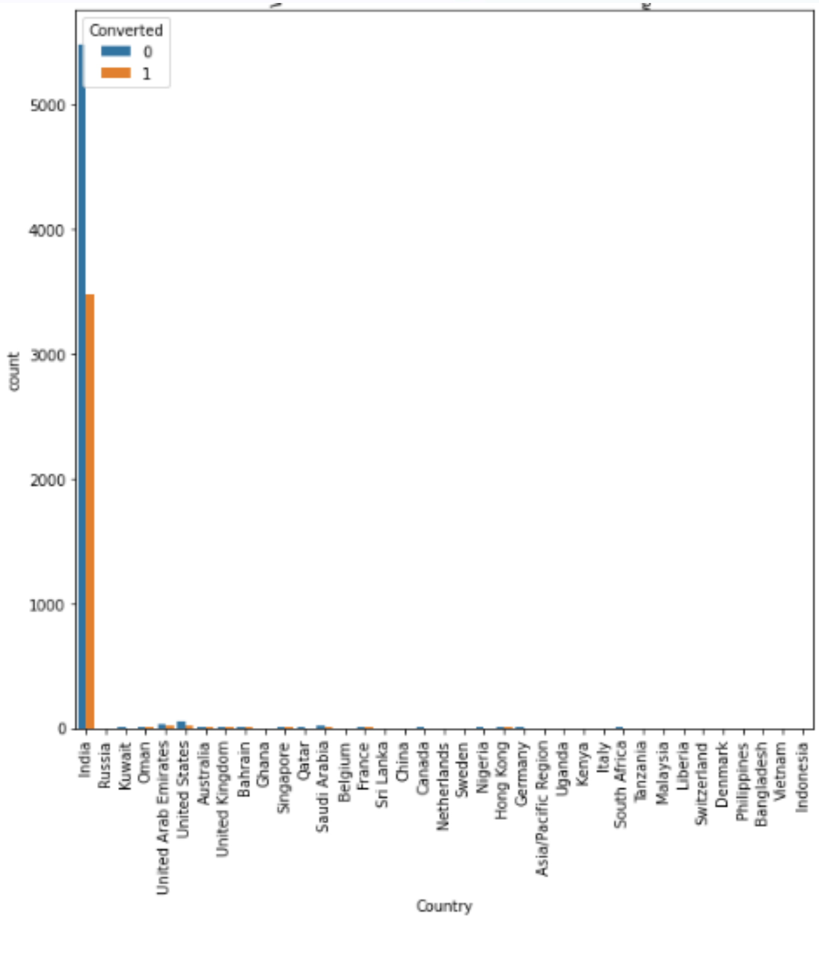




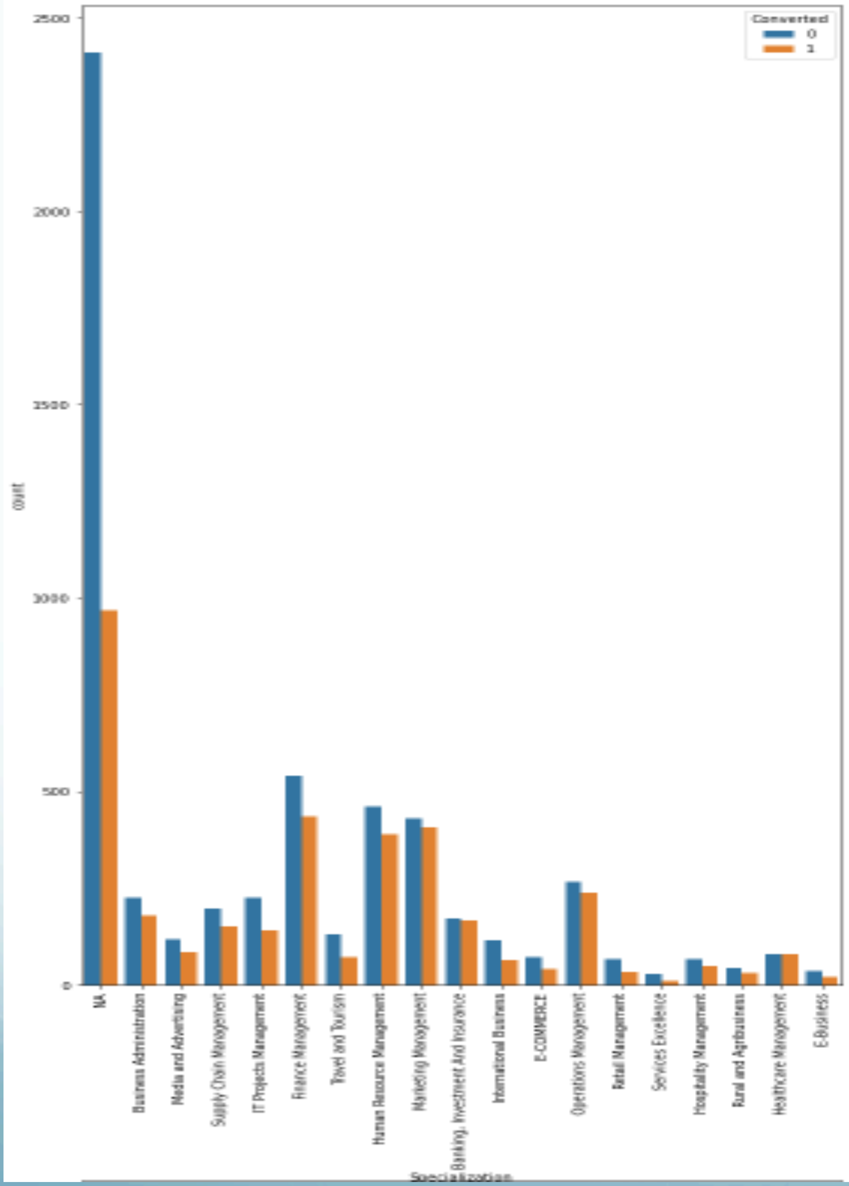
- API

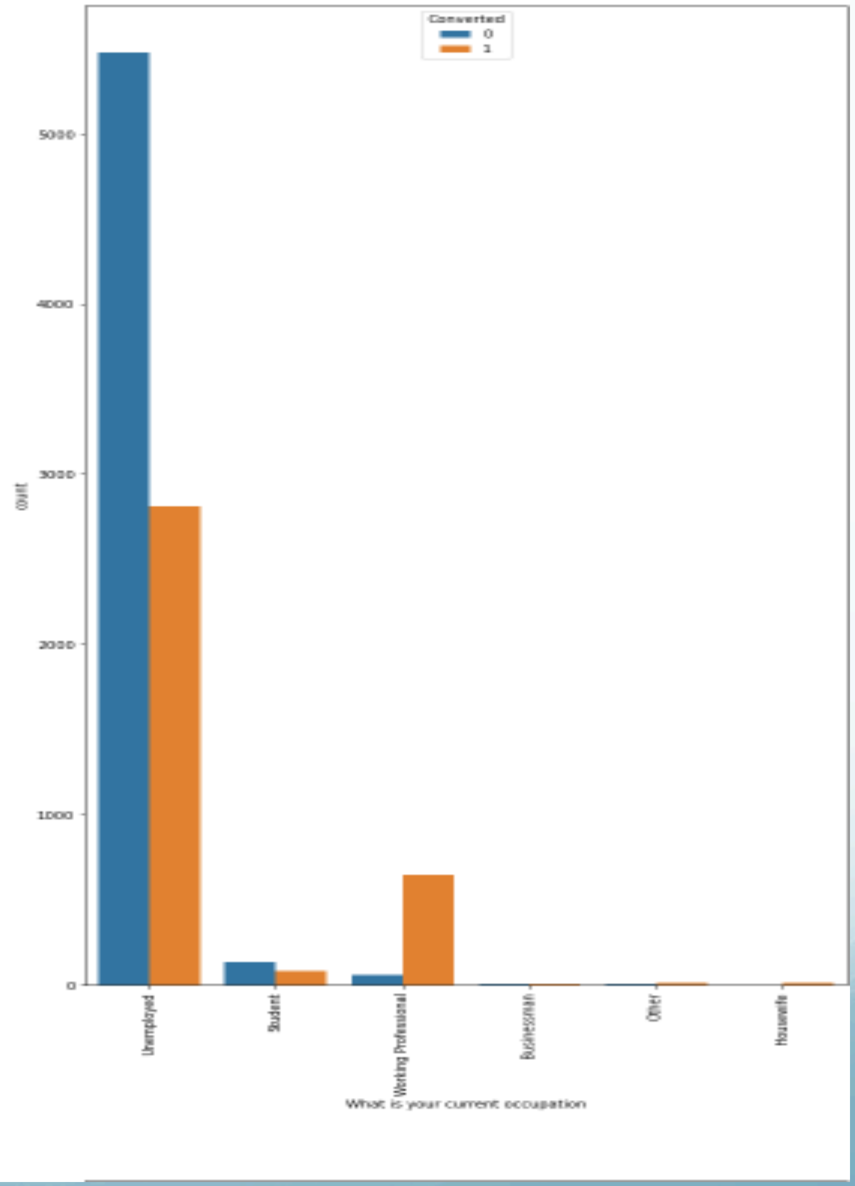
Data Analysis



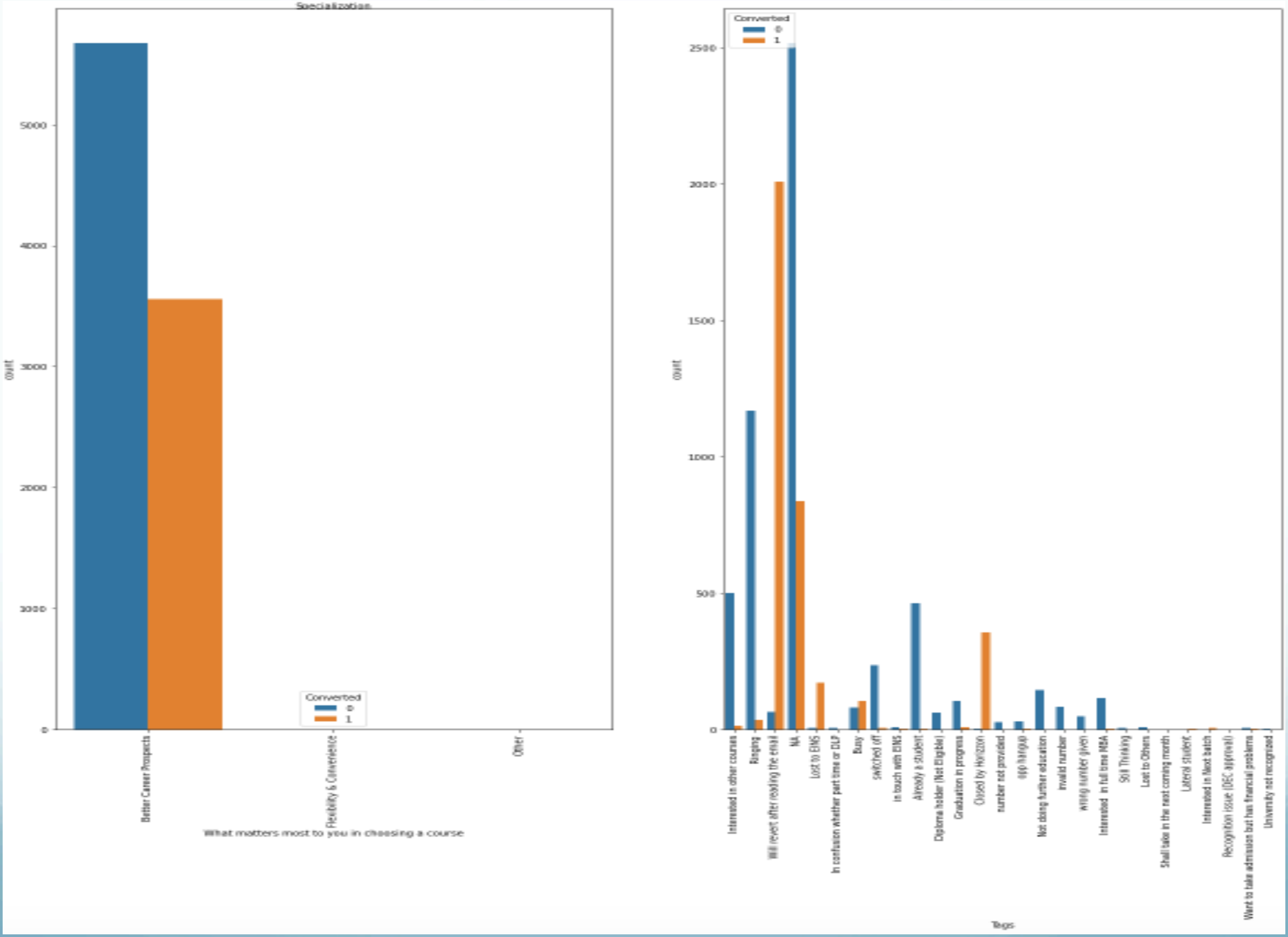


Data Analysis

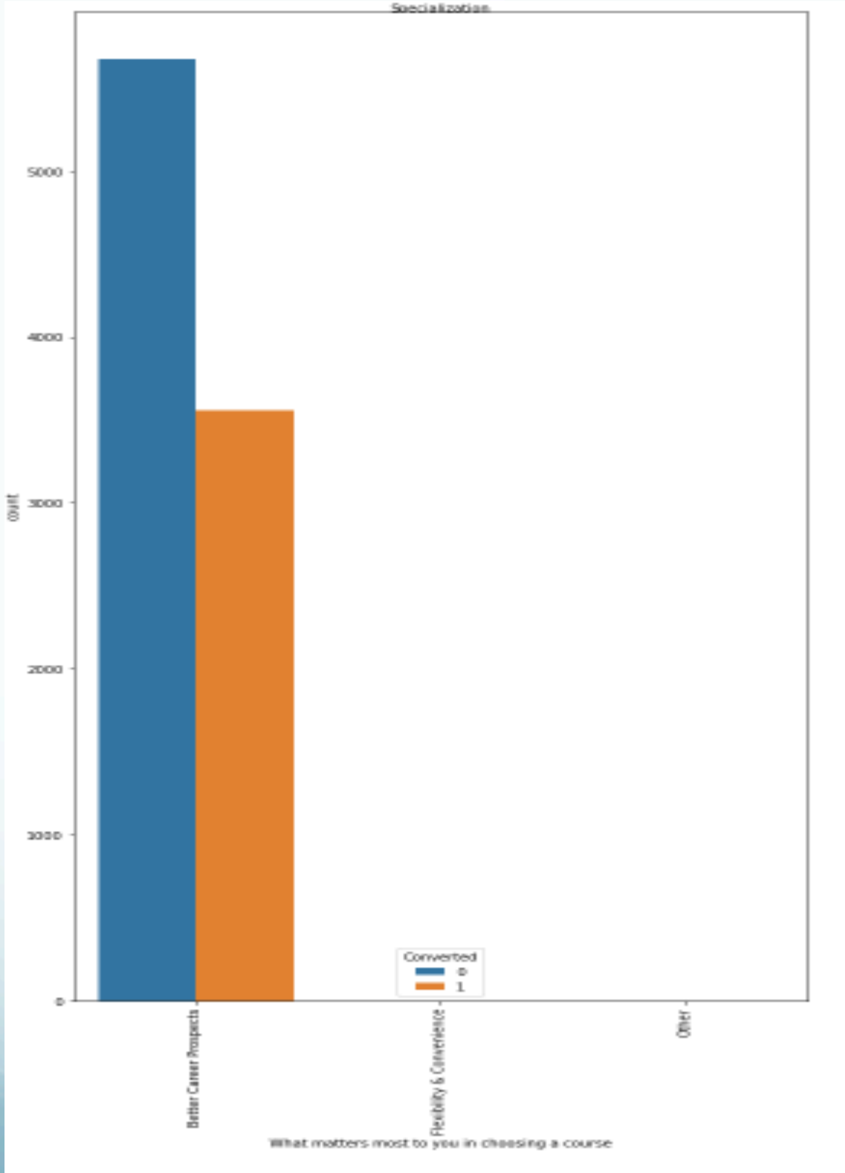


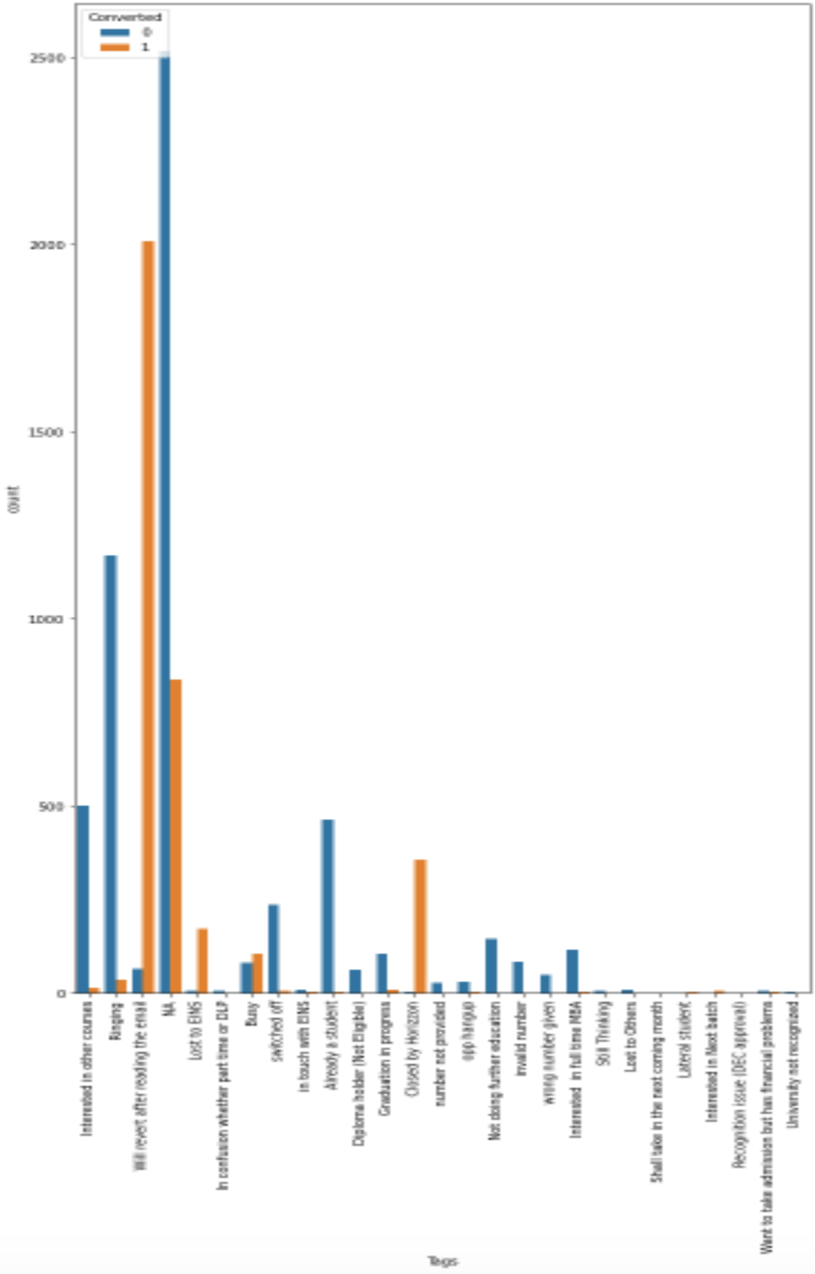


Data Analysis

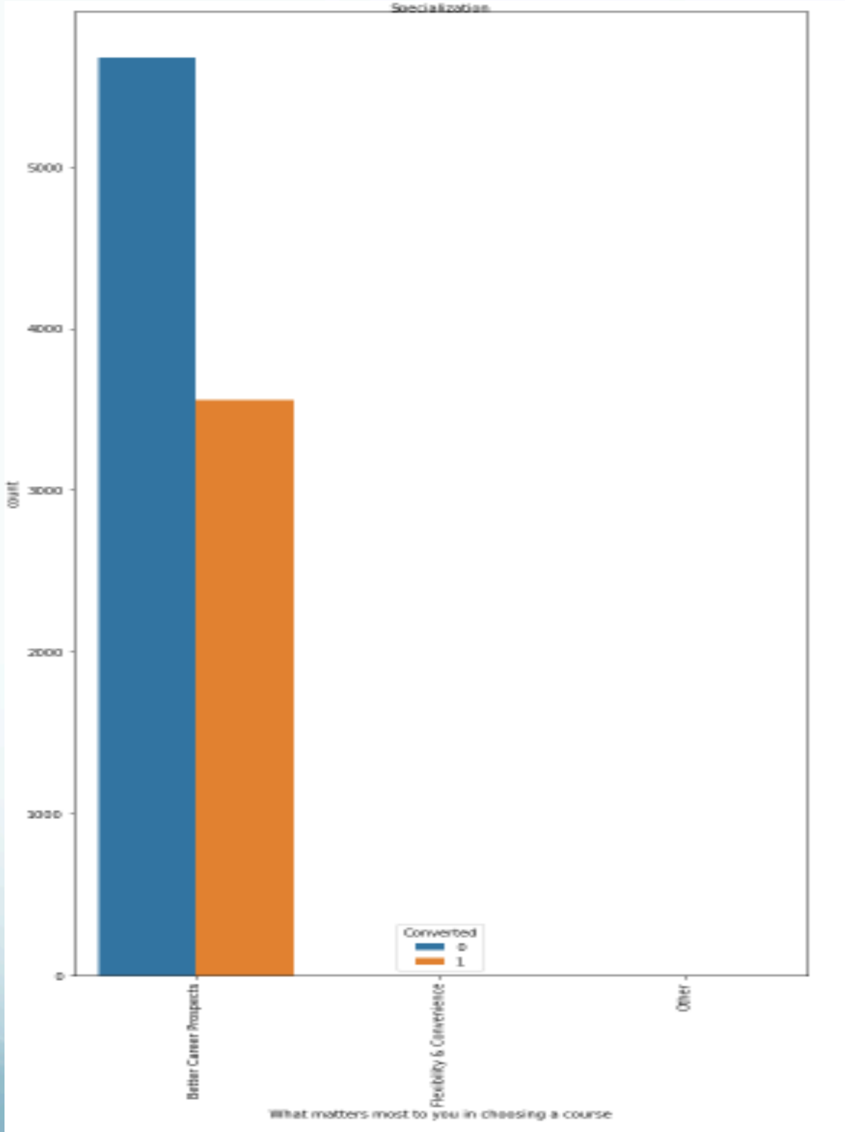


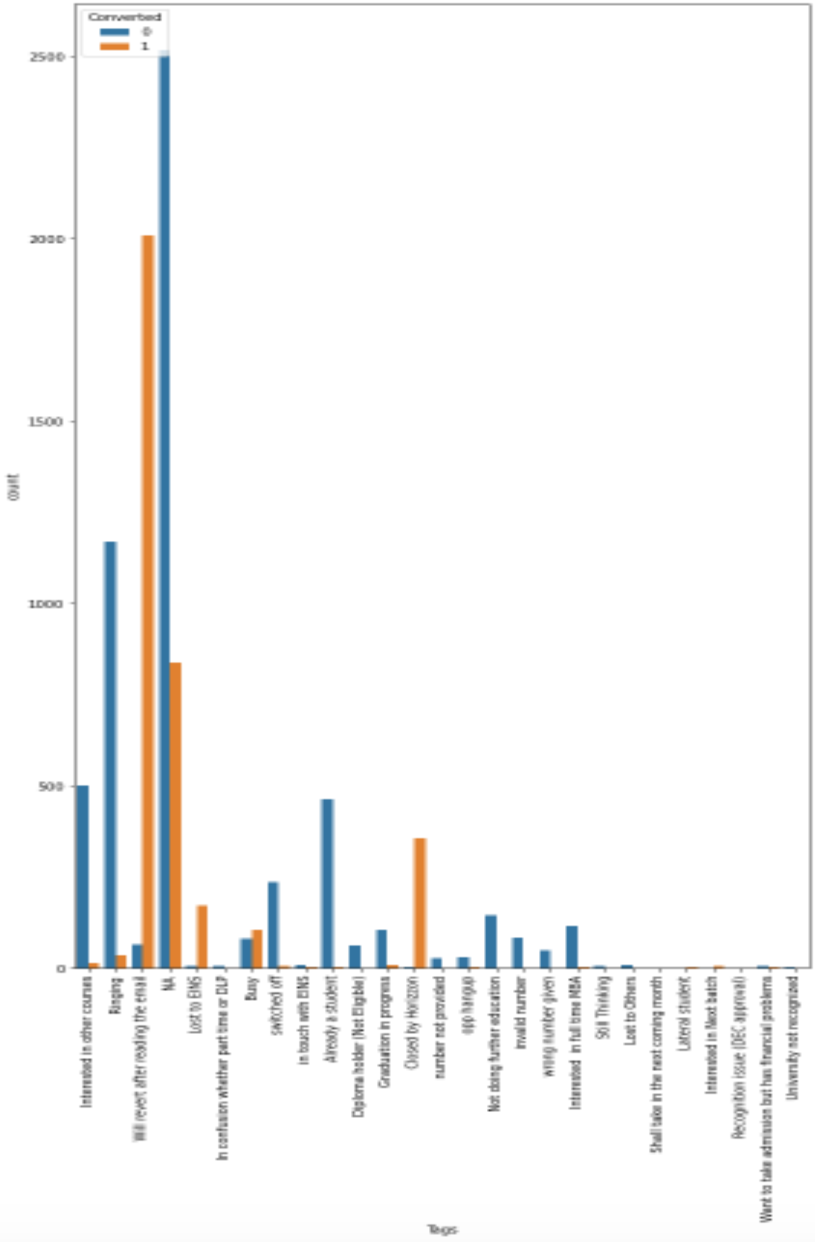
Data Analysis





Data Analysis





Data Analysis

- Lead Orgin
  - Increase conversion of API and Landing Page Submission. Increase Lead Add Form counts
- - Lead Source
  - Increase wellingak Website and reference count and increase conversion of olrak chat, organic search, direct traffic, Google
- - Last Activity
  - Focus on increasing SMS sent and email opened
- - Country
  - India has most traffic which is significantly high
- Specialization
  - There is a huge count of missing data. Need to focus on that
- What is current occupation
  - Focus on increasing unemployed conversion rate. Focus on working profression and increase the count

Data Analysis

- What matters most to you in choosing a course
  - Better career Prospects is the only option. Imbalance data
- Tags
  - will revert after reading the email high conversion rate. Focus on finding NA and getting them converted
- Lead Quality
  - Focus on increasing might be and high revelance count and find missing data and increase those conversion count
- City
  - Mumbai has high count and conversion. Focus on other cities too.

Data Analysis

- Final data after EDA:
  - Columns: 16
  - Rows: 9240

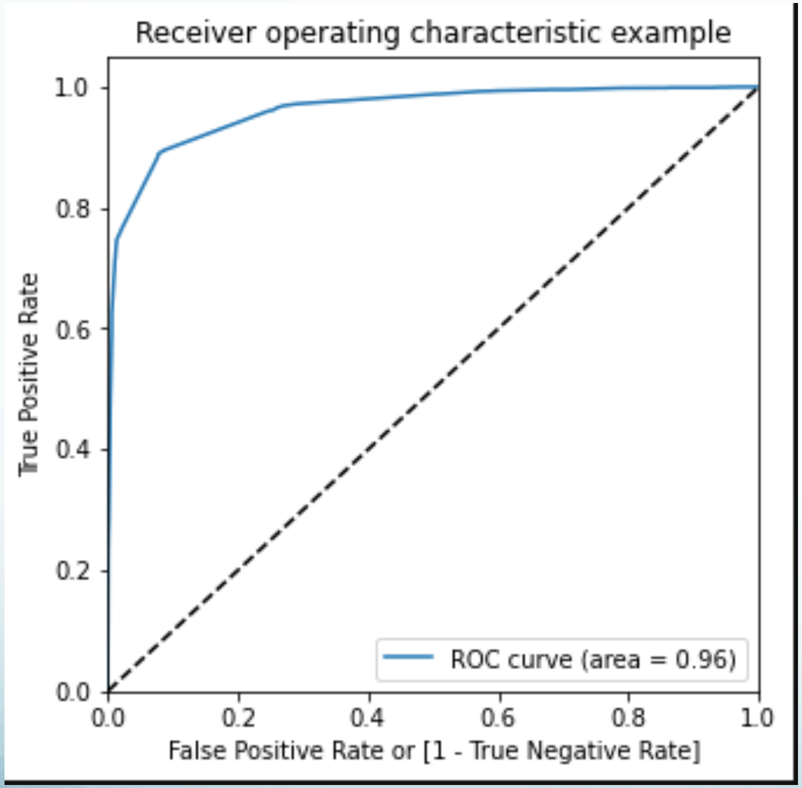
Data Preparation for Modeling

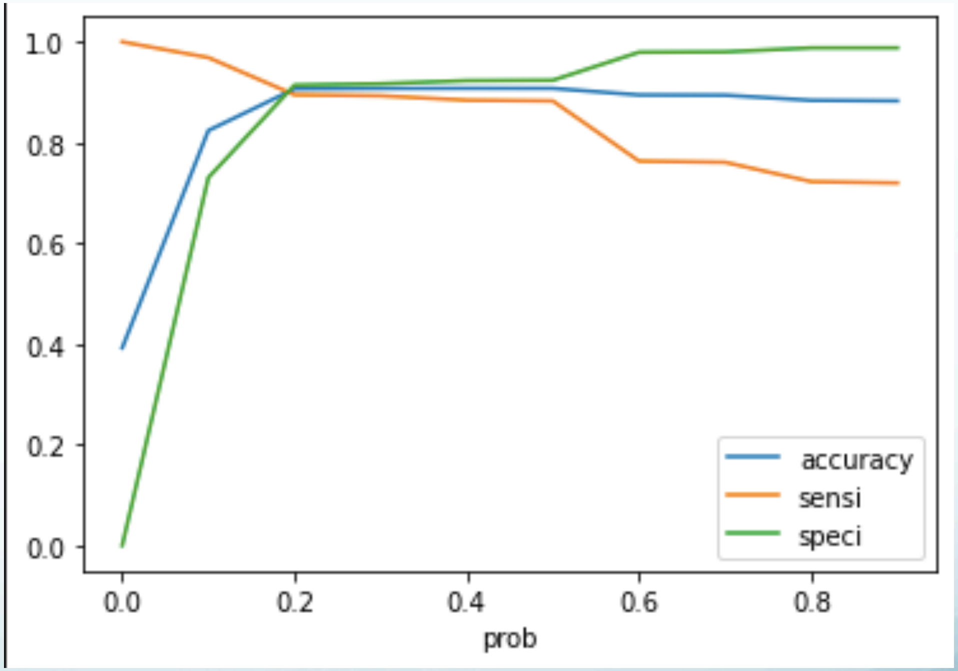
- Converting some binary variables (Yes/No) to 0/1
- Creating a dummy variable for some of the categorical variables.
- Checking and removing outliers
- Finding correlated features and removing them
- Final Data set after data preparation for modeling:
  - No. of rows: 9214
  - No. of columns: 86

Data Modeling

- Splitting the data in training and test set
- Select Features using RFE
- Accessing the model to make it better using statsmodel
- Building model
- Applying model on test data

ROC and Optimal Cut-off Curve





Final Data Model

- Training data:
  - accuracy : 90%
  - sensitivity: 89%
  - specificity: 91%
- - Test data:
  - accuracy : 90%
  - sensitivity: 88%
  - specificity: 91%
- Both training and test results match so we can say our model is good

Recommendations

- Using model look Lead Score and higher lead score, probability of conversion is high
- Top Features to Focus on:
  - Tags which were Busy, Switched Off, Lost to EINS
  - Lead Source from Welingak Website
  - Last Activity was Email Opened and SMS Sent
- Tags like which have not been identified needs to be identified too