Lead Scoring Case Study

X Education
Online Courses

SUBMITTED BY

ANIL THAKRE
ATUL MANI
YAMUNA

Business Case

PROBLEM STATEMENT

- X Education sells online courses to industry professionals
- It has Poor lead conversion ratio of around 30%
- Target lead conversion rate is around 80%

OBJECTIVE

- To build Logistic Regression model to select the leads that are most likely to convert into paying customers
- Model to be flexible in order to adjust to if the company's requirement changes in the future

Solution approach

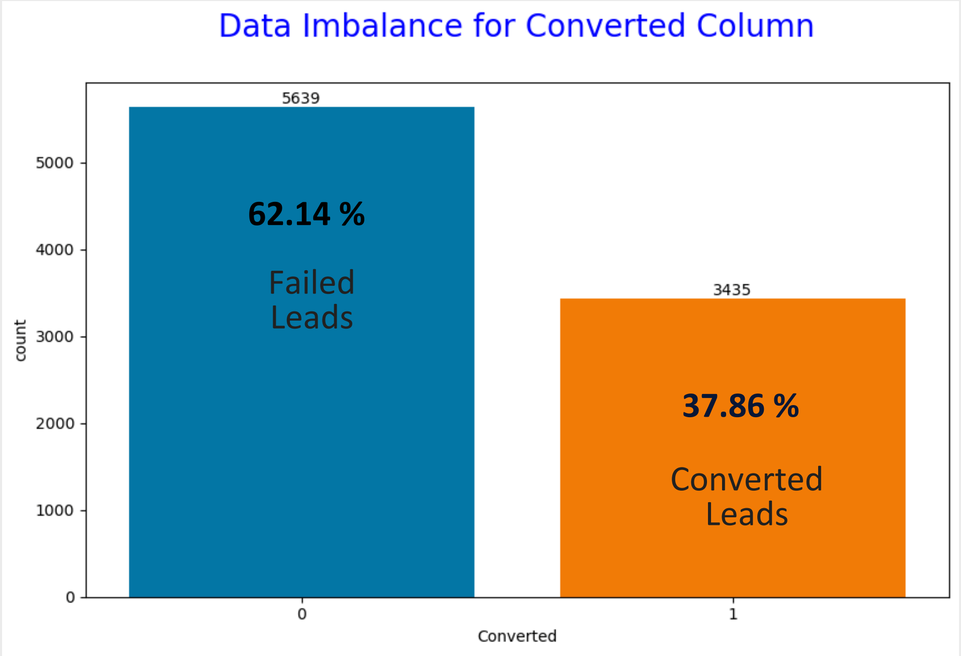
STEPS

- Understand Business need
- Load & Study data
- Handle Missing values and Outliers
- Perform Exploratory Data Analysis
- Perform Data preparation (Scaling, imputations, feature engineering, etc.)
- Build Model using RFE and Manual approach using p-values and VIF
- Find optimal probability cut-off
- Test the model performance using measures such as confusion matrix, accuracy, sensitivity, etc.
- Create additional feature for Lead Score, define threshold value of this score based on Business requirement

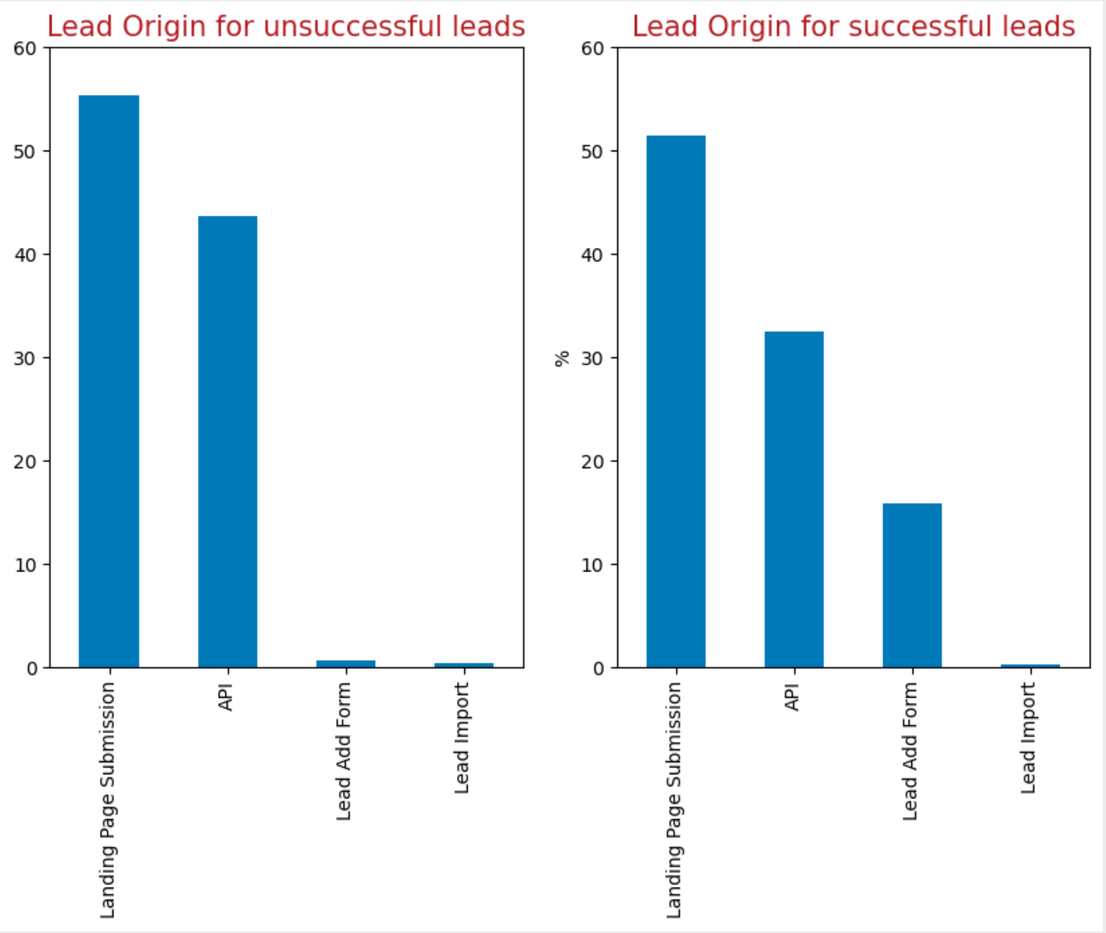
Key Observations

DATA PROPERTIES

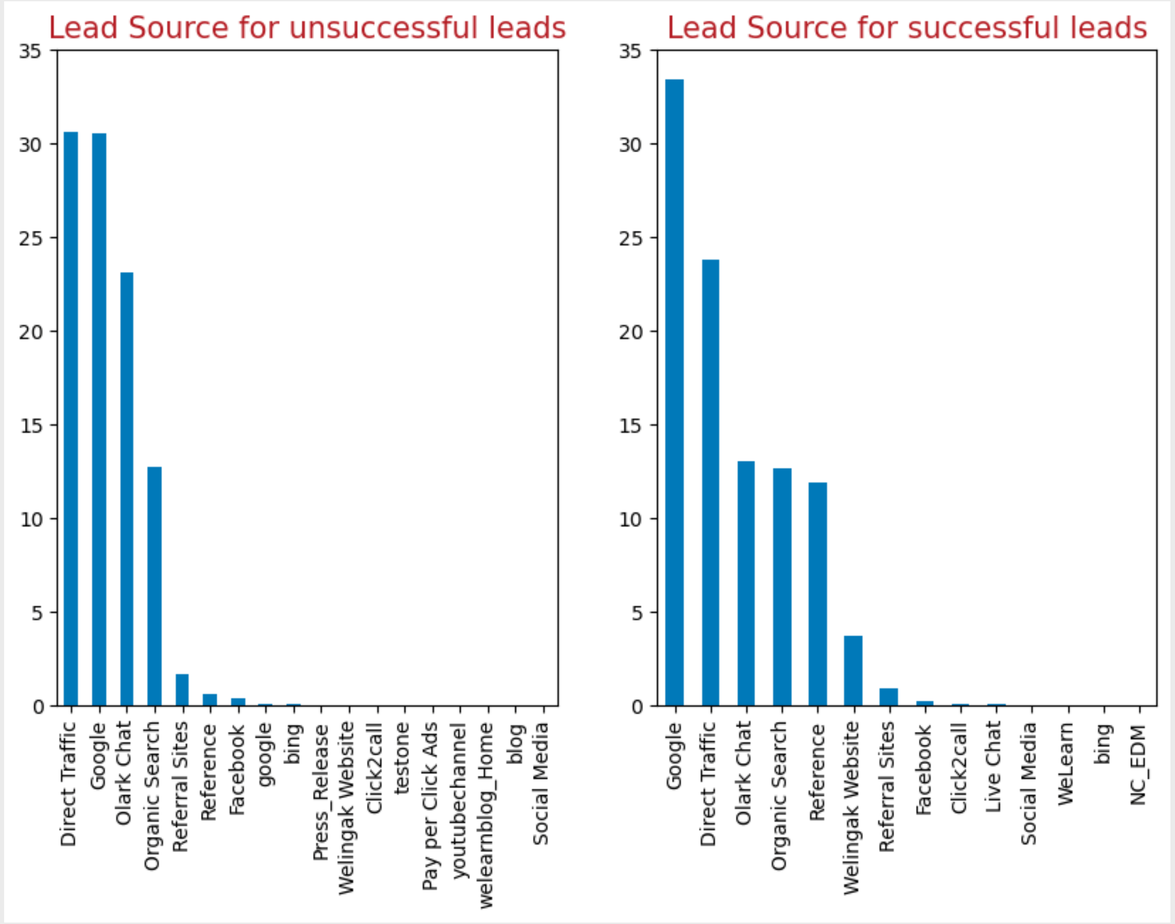
- Number of rows —> 9240
- Number of Features —> 37
- 7 features had missing values greater than 40% , hence were dropped
- 5 features had only single constant value, hence not useful for analysis
- The target column ‘Converted’ indicated 37.86 % lead conversion



Data Analysis

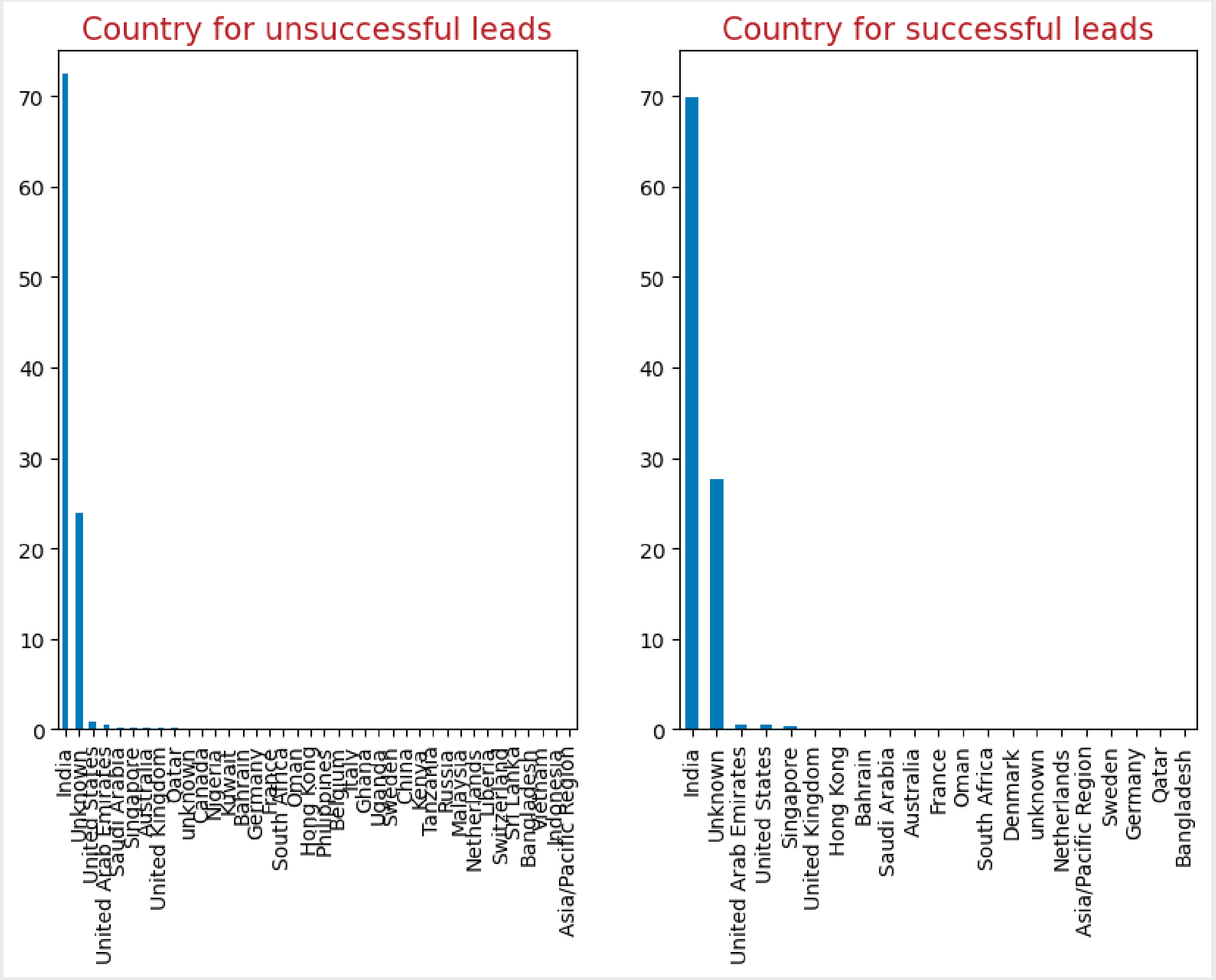


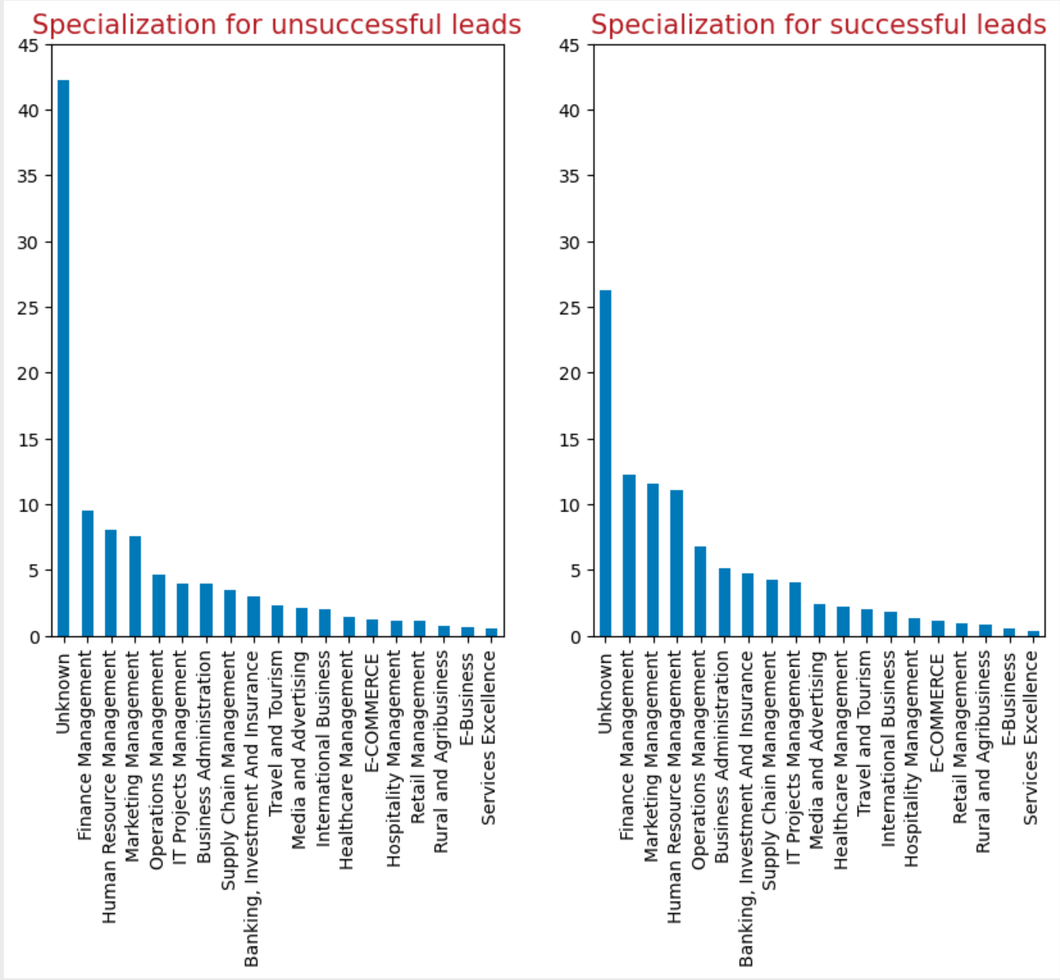
Lead Add form Entry is prominent in Successful Leads



Only few Lead sources play major role to determine Lead success

Data Analysis...

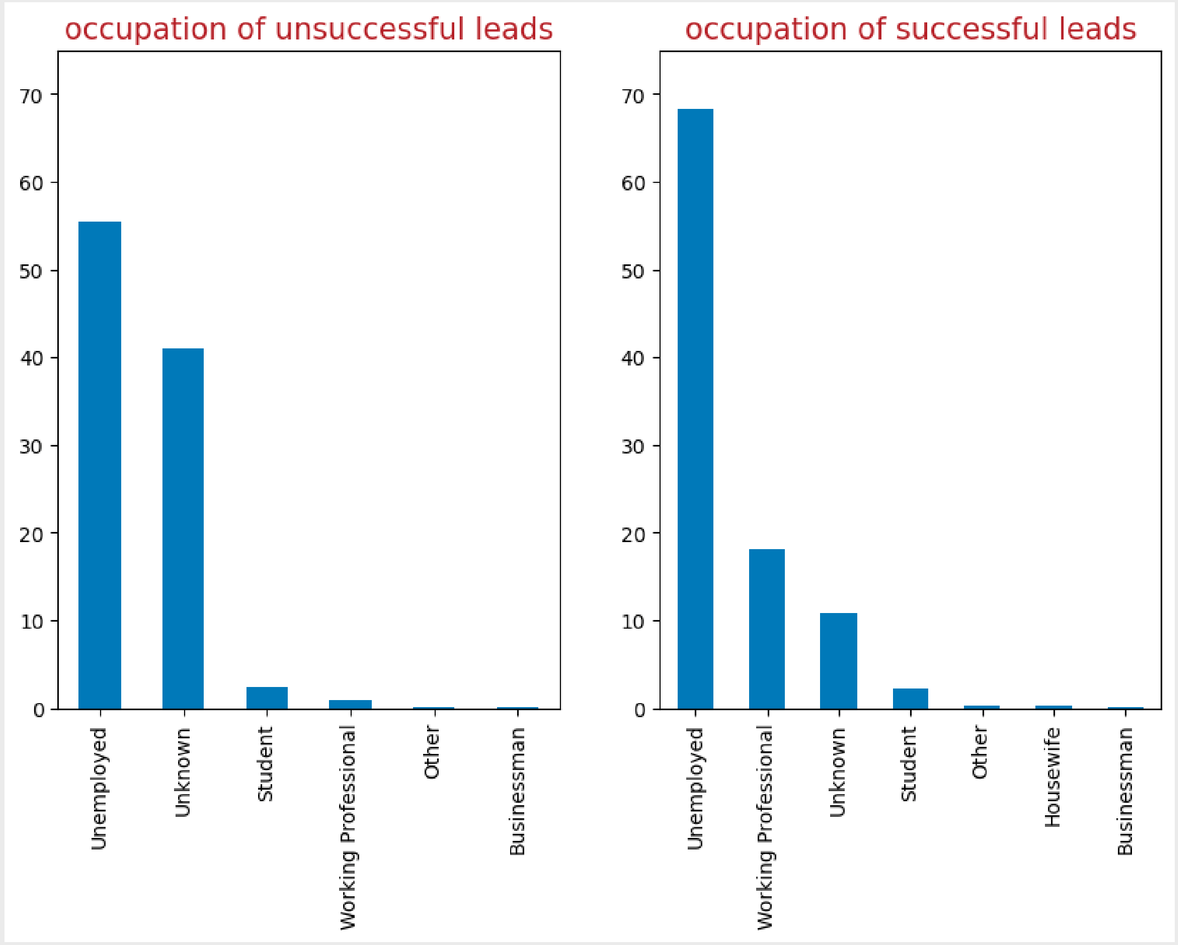


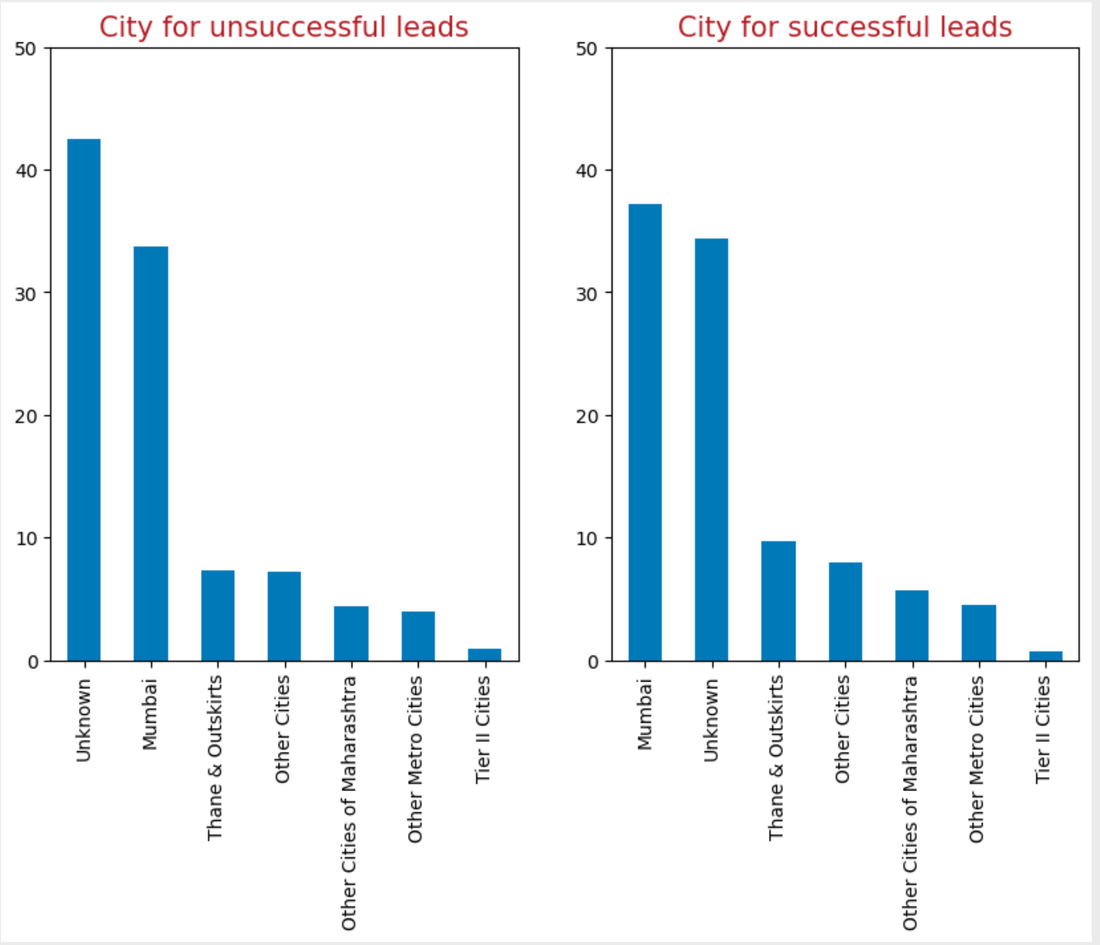


Majority of professional are from India and there are many unknown values

Most of the entries have unknown value but Finance, Marketing & HR are popular compared to others

Data Analysis...

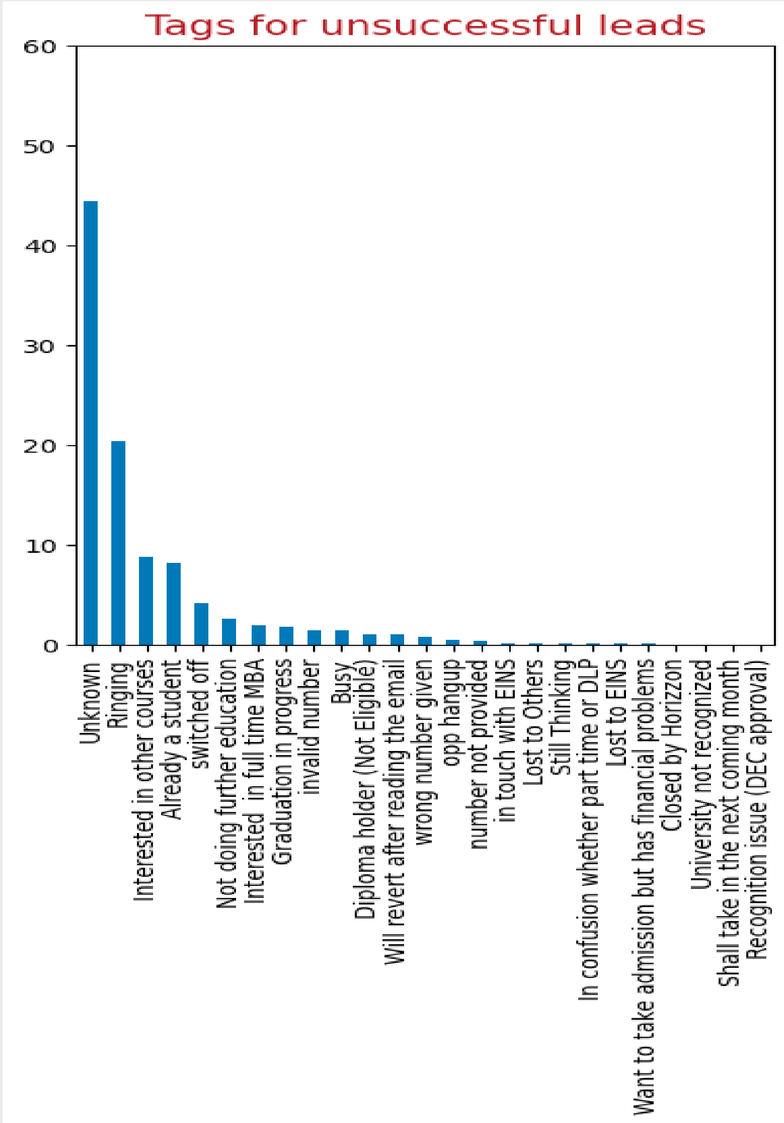


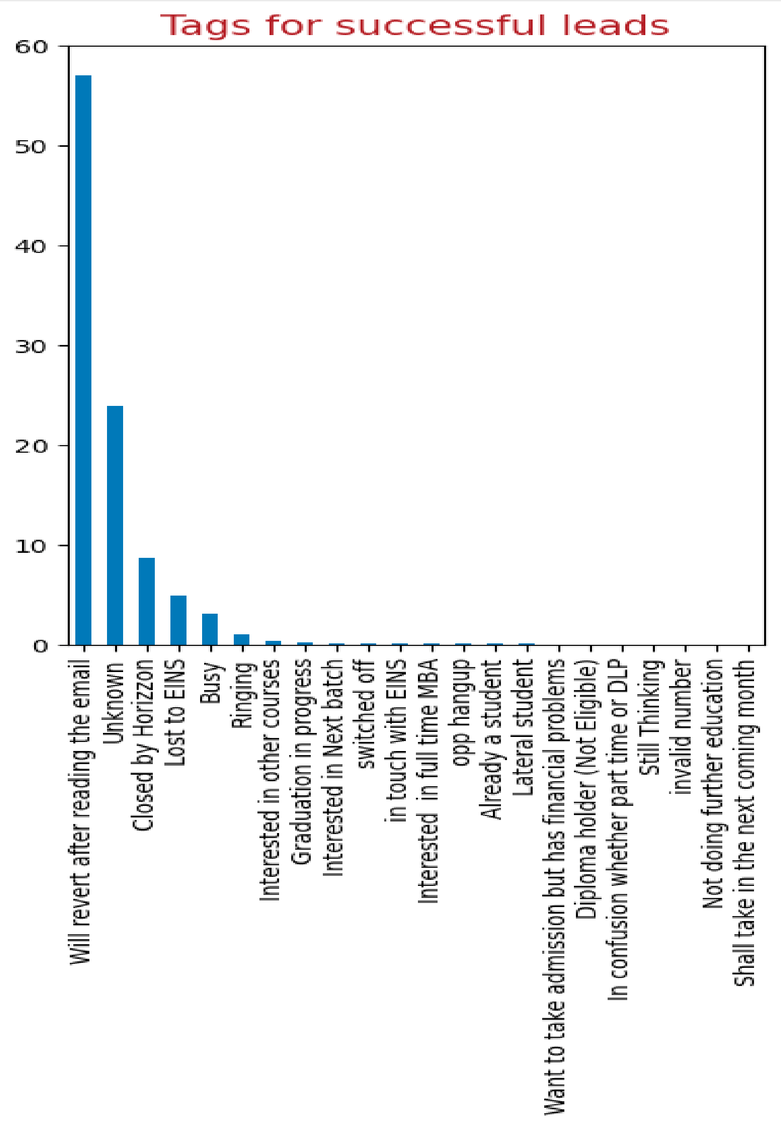


Working Professionals have highest probability of Conversion followed by Unemployed candidates

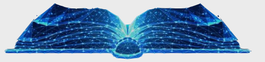
Most of the candidates are located in Mumbai & nearby area

Data Analysis...

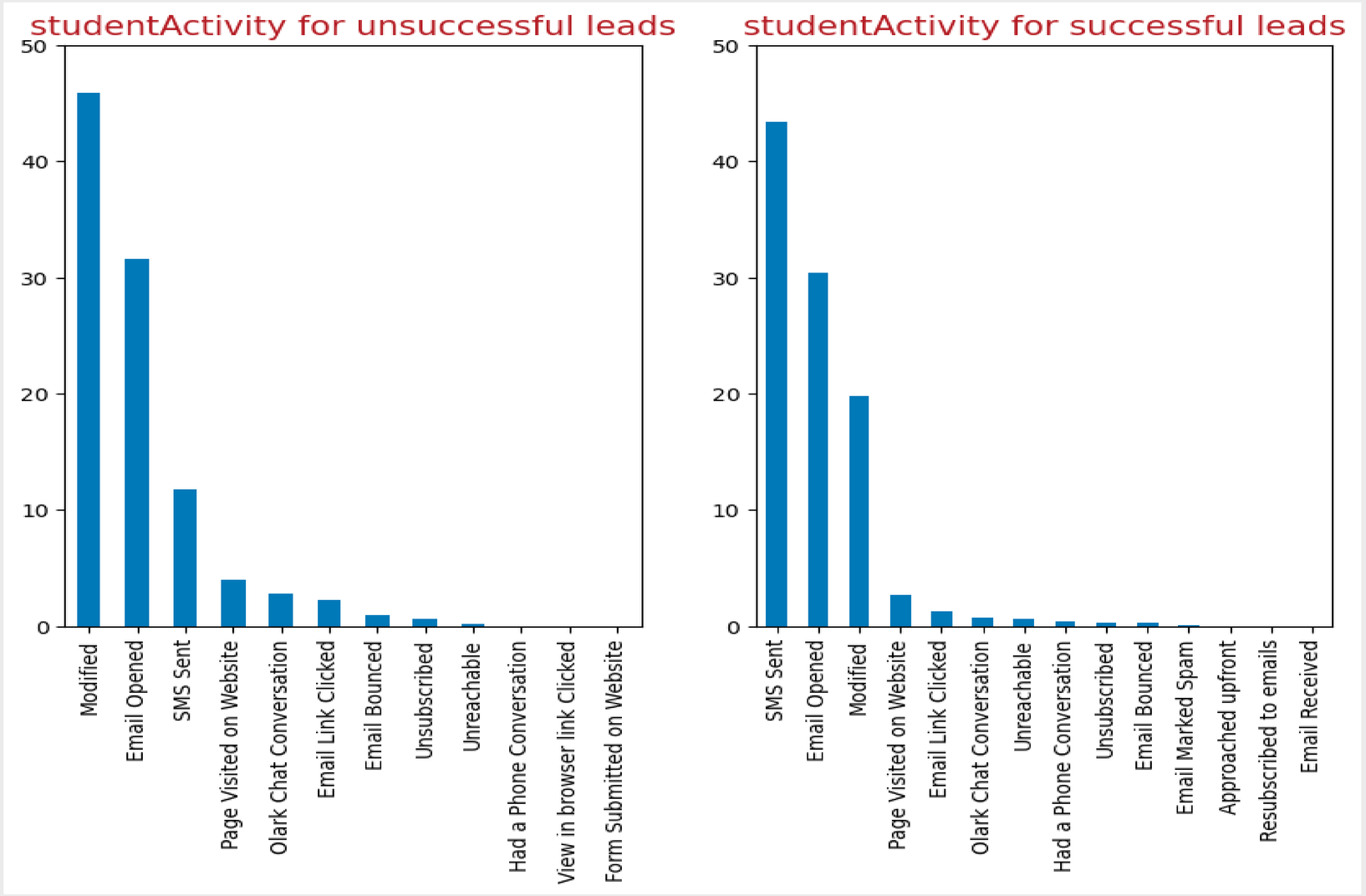




Tags assigned to customers indicating the current status of the lead had considerable changes for successful versus unsuccessful leads.



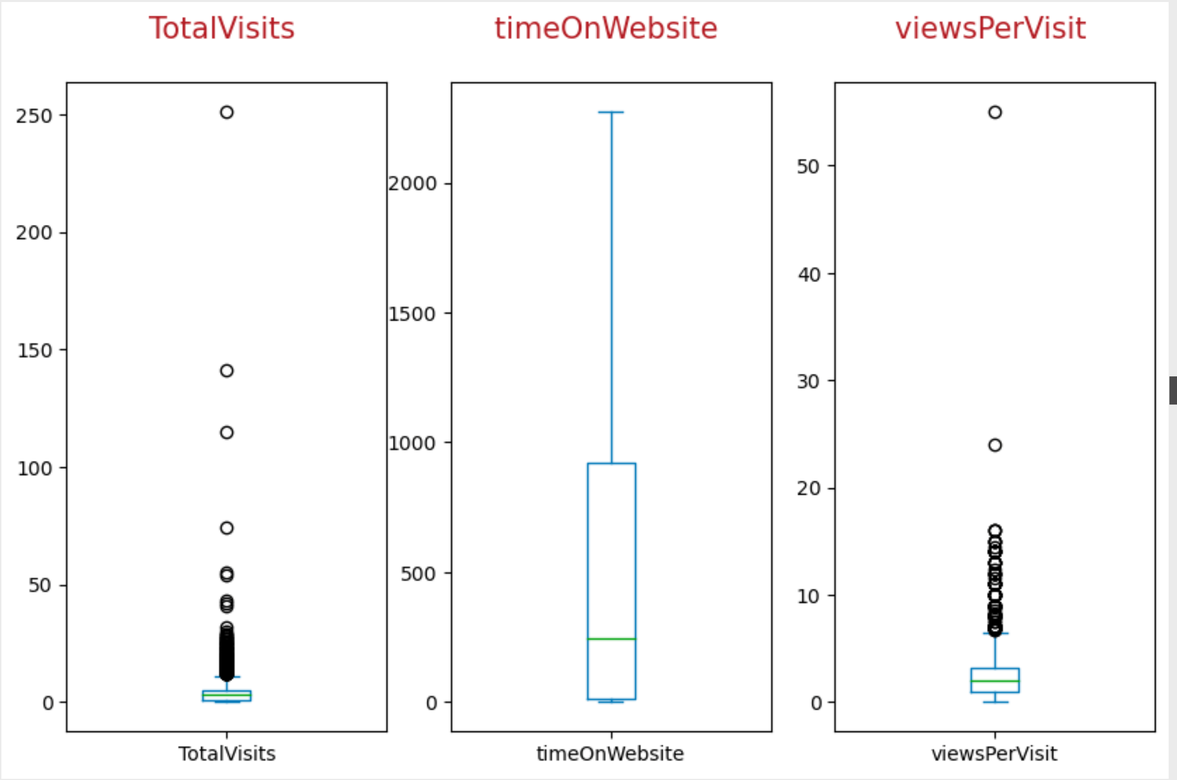
Data Analysis...



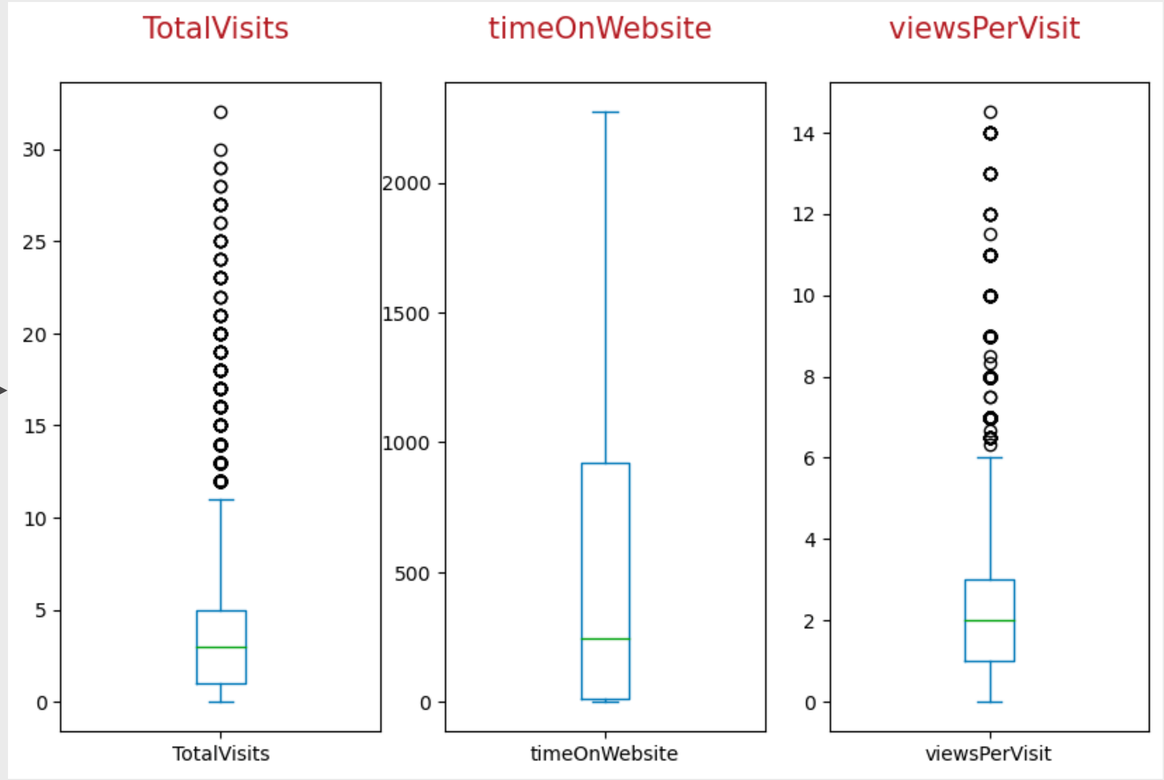
The last notable activity performed by the student can be used as significant predictor for Lead conversion success

Outlier Handling

Two out of three numerical columns had outliers, the following is result of the outlier treatment



Before



After

Model Predictors

Dummy Variables	coef
Tags_Closed by Horizzon	7.18
Tags_Lost to EINS	6.55
Lead Source_Welingak Website	4.17
Tags_Will revert after reading the email	4.14
Last Activity_SMS Sent	2.20
occupation_Working Professional	1.48
occupation_Unemployed	1.40
studentActivity_Olark Chat Conversation	-0.90
Tags_Graduation in progress	-1.59
studentActivity_Modified	-1.60
Tags_Interested in other courses	-2.18
Tags_Interested in full time MBA	-2.39
Tags_Not doing further education	-3.41
Tags_opp hangup	-3.63
Tags_Ringing	-3.98
Tags_switched off	-4.41
Tags_invalid number	-4.66

As can be seen from the table, the following features are part of model dummy features

✦ Tags

✦ Lead Source

✦ Last Activity

✦Last Notable Activity

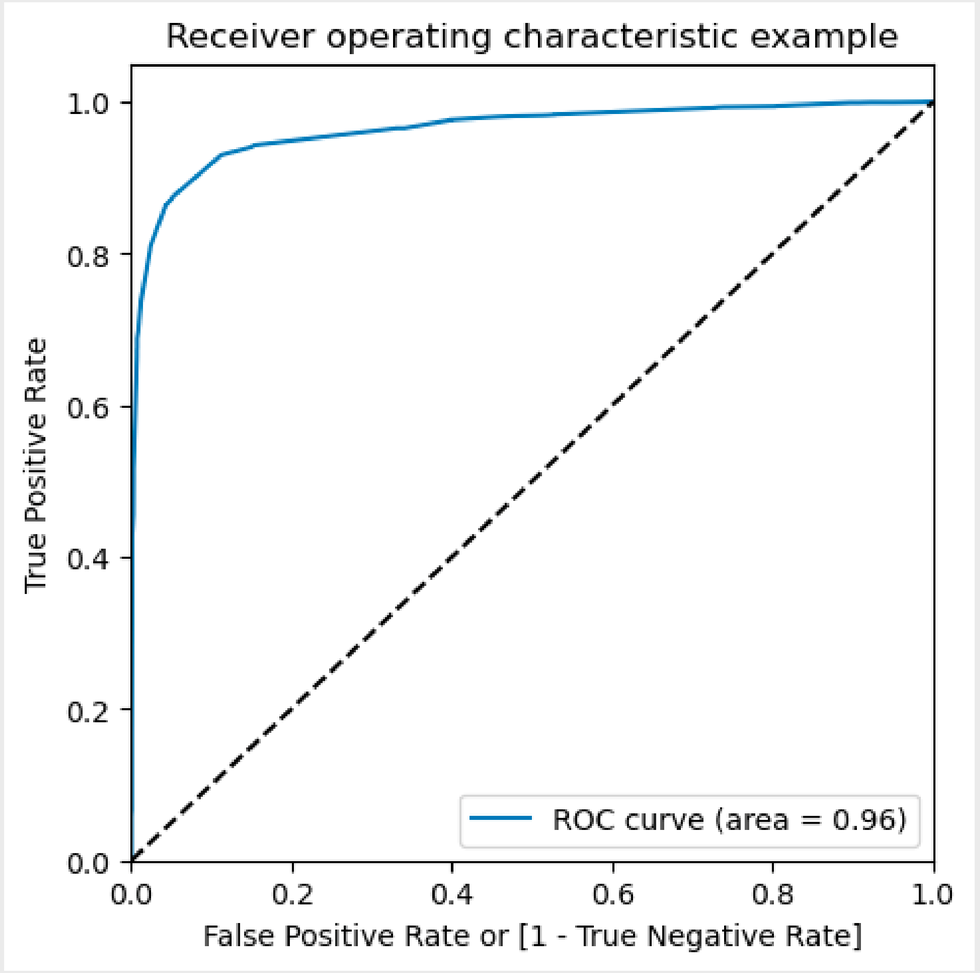
✦Occupation

❖ For the customers whose current status of the lead is indicated as ‘Closed by Horizon’ , ’Lost to EINS’ & ‘Will revert after reading the email’ have high chance of conversion, but there are multiple Tags which negatively impact chances of conversion

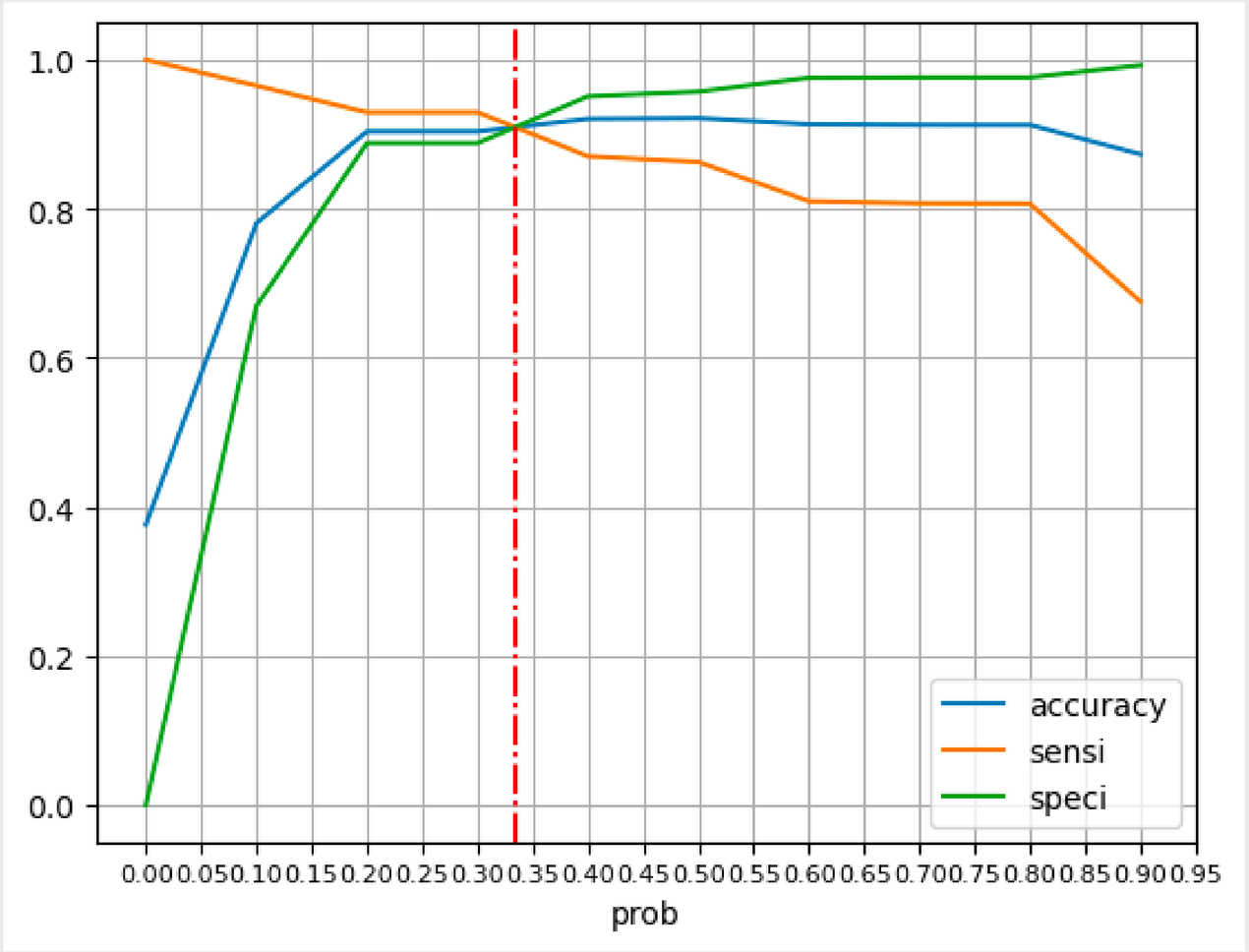
❖Leads obtained through Welingak Website have great conversion ratio.

❖We shall focus more on Working Professional & Unemployed categories of employment for better conversion

Model Evaluation



ROC Curve area is very high hence model is strong



Optimal cutoff probability is 0.335 where we get balanced sensitivity and specificity

Model Evaluation

Results of Evaluations on Test data

Recall : 0.86

TP / (TP+FN)

Precision : 0.92

TP / (TP+FP)

Sensitivity : 0.88

TP / (TP+FN)

Specificity : 0.94

TN /(TN+FP)

The Recall is a critical measurement here and it is well above 0.8

Recommendations

- The Lead score of 35 shall be considered as hot Leads to obtain 80% Lead conversion
- The following field are critical for prediction hence missing values to be reduced for these fields
  - ✦Tags
  - ✦ Lead Source
  - ✦ Last Activity
  - ✦Last Notable Activity
  - ✦Occupation
- In case of any changes the optimal cutoff score can be changed from 35 to higher side (>35) in case efficiency is important and resources are less
- For the Business requirement where high lead conversion is more important than cost of resources the cutoff can be reduced below 35 (<35)

Thank You !