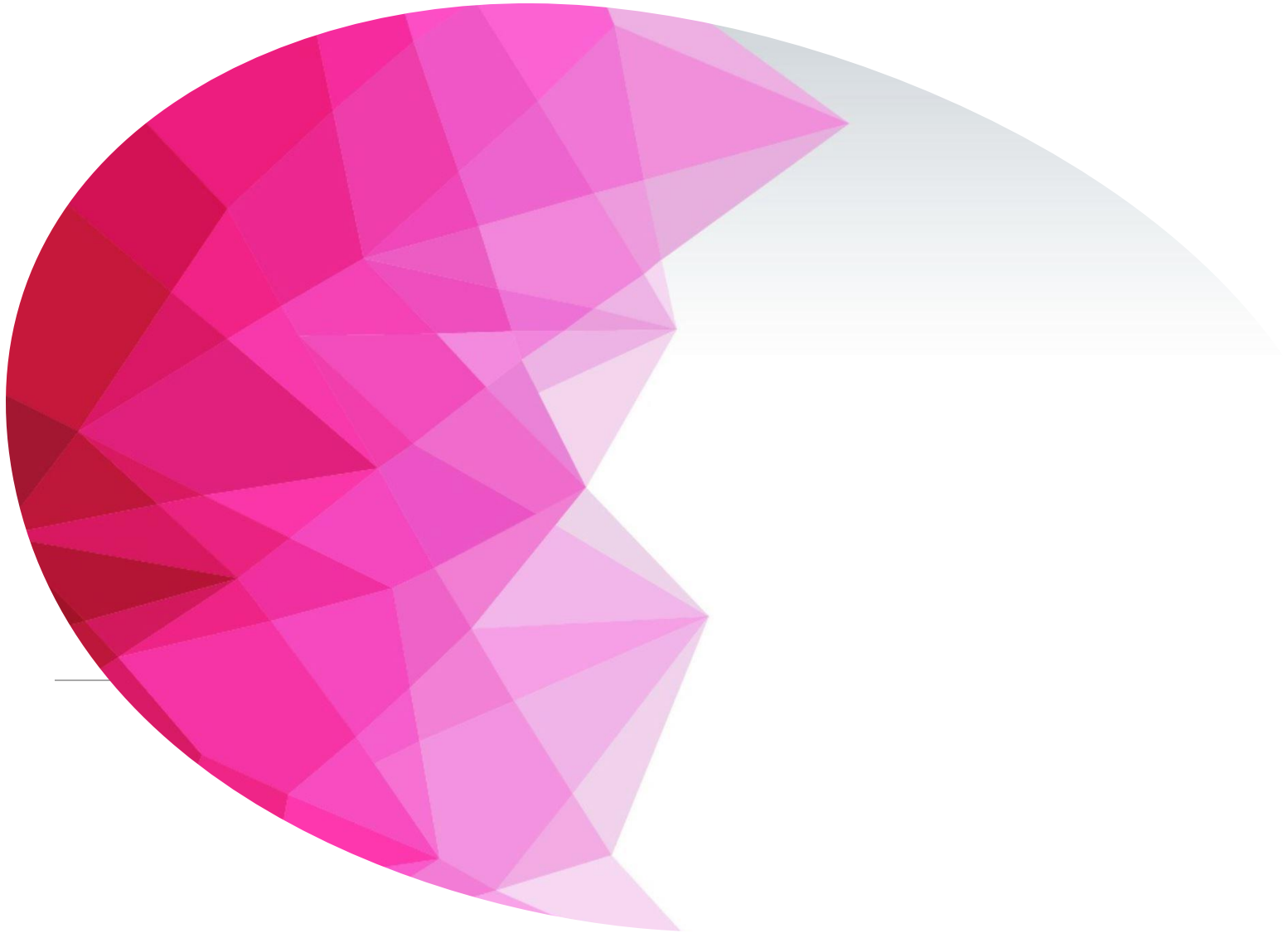


LEAD SCORE CASE STUDY

LOGISTIC REGRESSION

BISWAJEET SINGH

Goal

Build a Logistic Regression Model to assign a lead Score between 0 and 100 to each 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 Statement

Problem Statement An online education company, X Education has a low lead conversion rate. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. The goal is to build a model which assigns a lead score 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 target is to achieve a lead conversion rate of 80%.

Analysis Approach

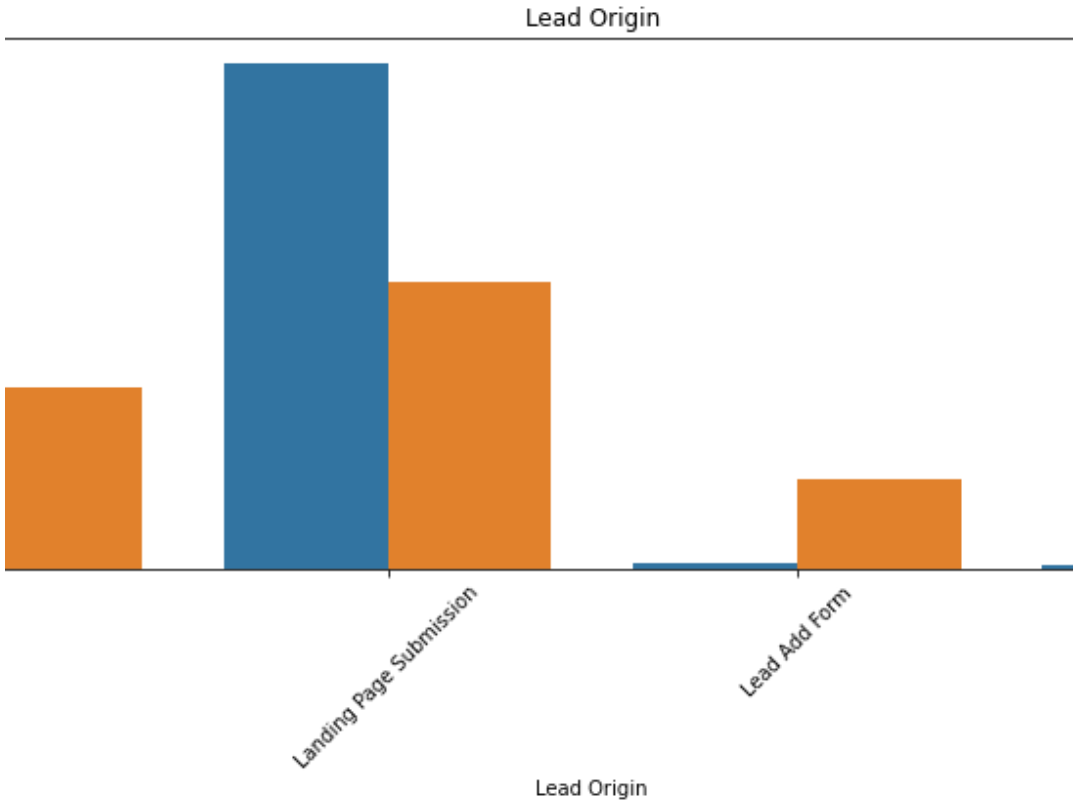
Import and Inspection of data

Necessary Cleaning of Data i.e. removing duplicates, unnecessary columns, columns with high null values etc.

Exploratory Data Analysis(Univariate and Bivariate)

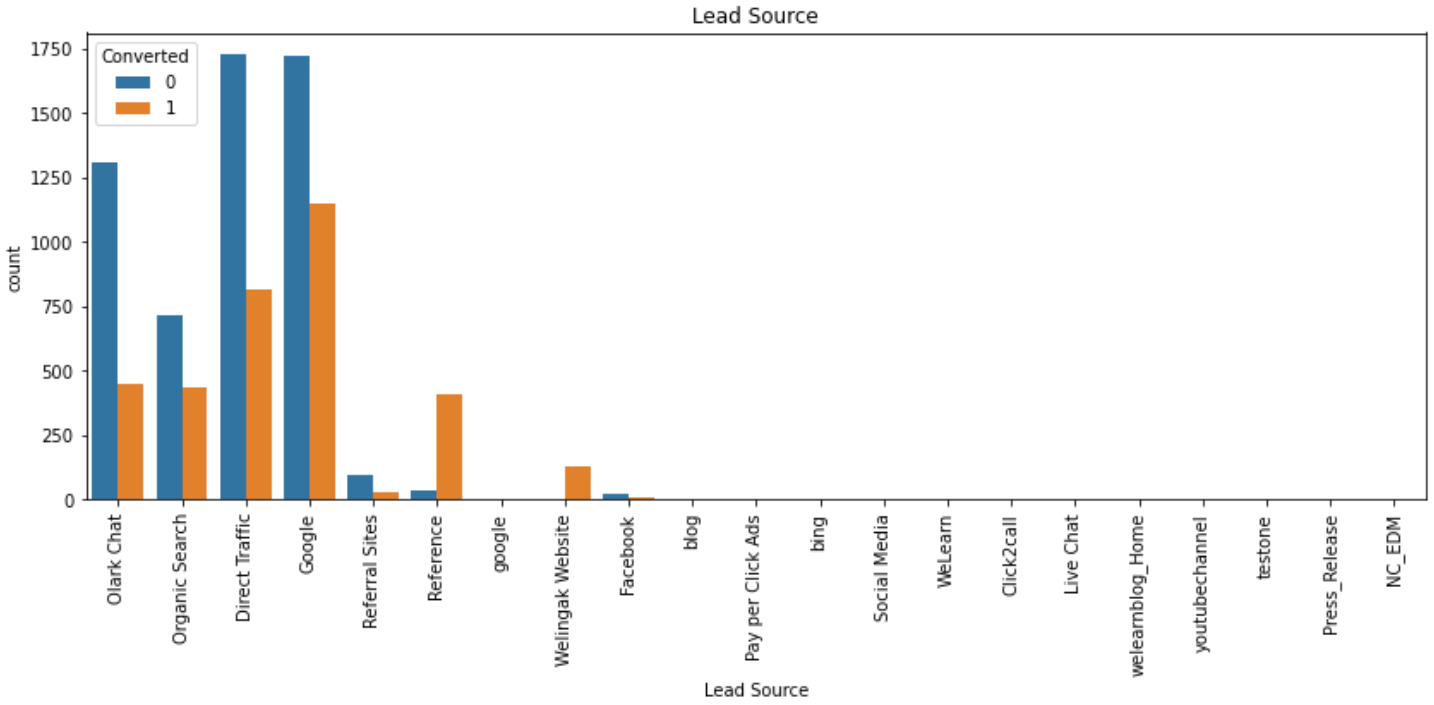
- Outlier Treatment and Standardization
- Optimal Logistic Regression Model Building
- Evaluation of Key Metrics such as Accuracy, Specificity, Sensitivity, Precision and Recall

EDA (EXPLoRATORY DATA ANALYSIS)



LEAD ORIGIN VS Converted

- API and landing page submission has 30-35% conversion rate but count of lead originated from them are considerable
- Lead add form has more than 90% conversion rate but count of lead are not high
- lead import has very less in count

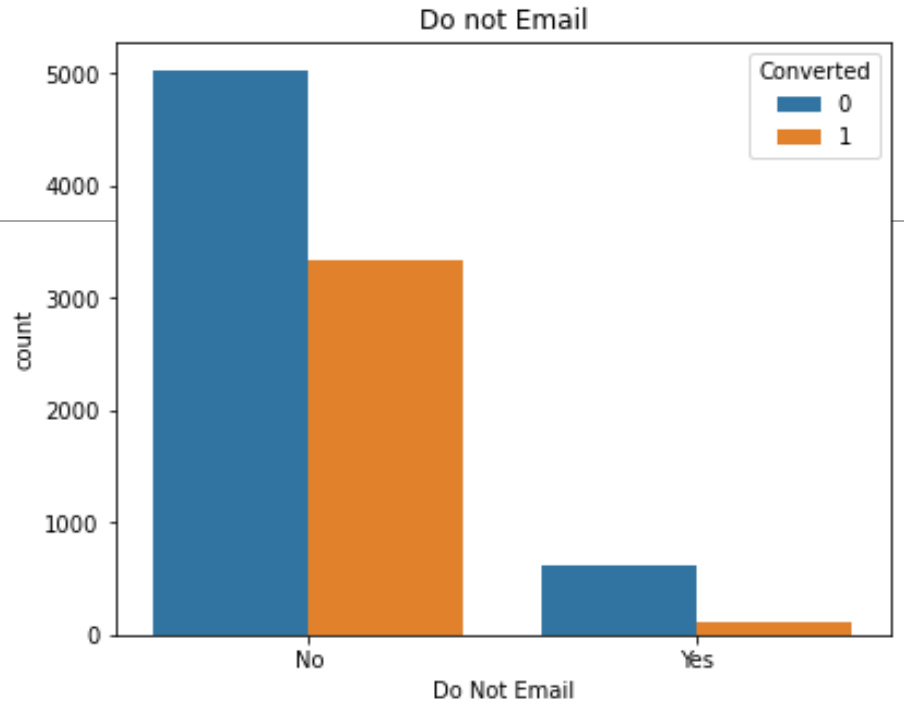


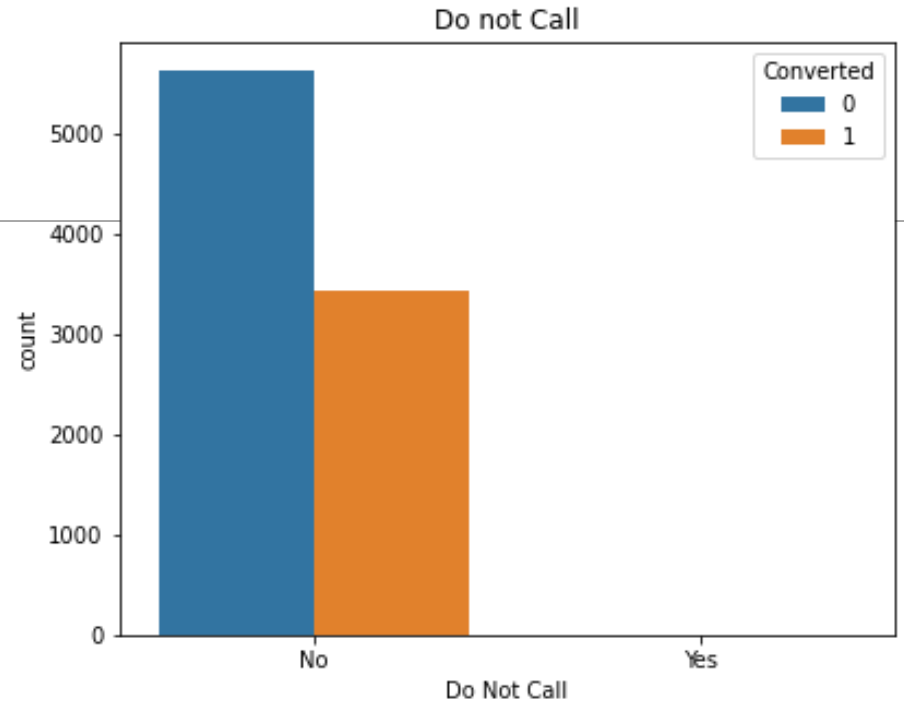
LEAD source

We see there Google is mentioned 2 times in the data so we are replacing the google with Google

Google search has had high conversions compared to other modes, whilst reference has had high conversion rate

Conversion rate of reference and Welingak website is high



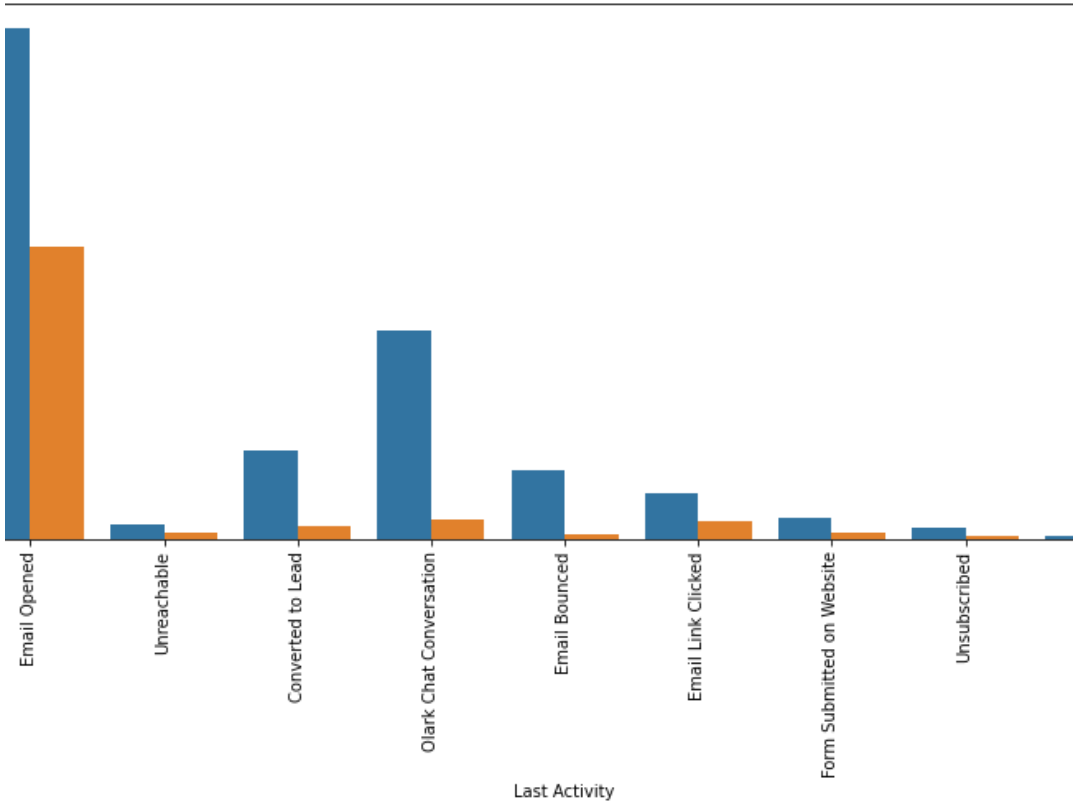


Do not call and dop not email

Google search has had high conversions compared to other modes, whilst reference has had high conversion rate

Most leads prefer not to informed through phone

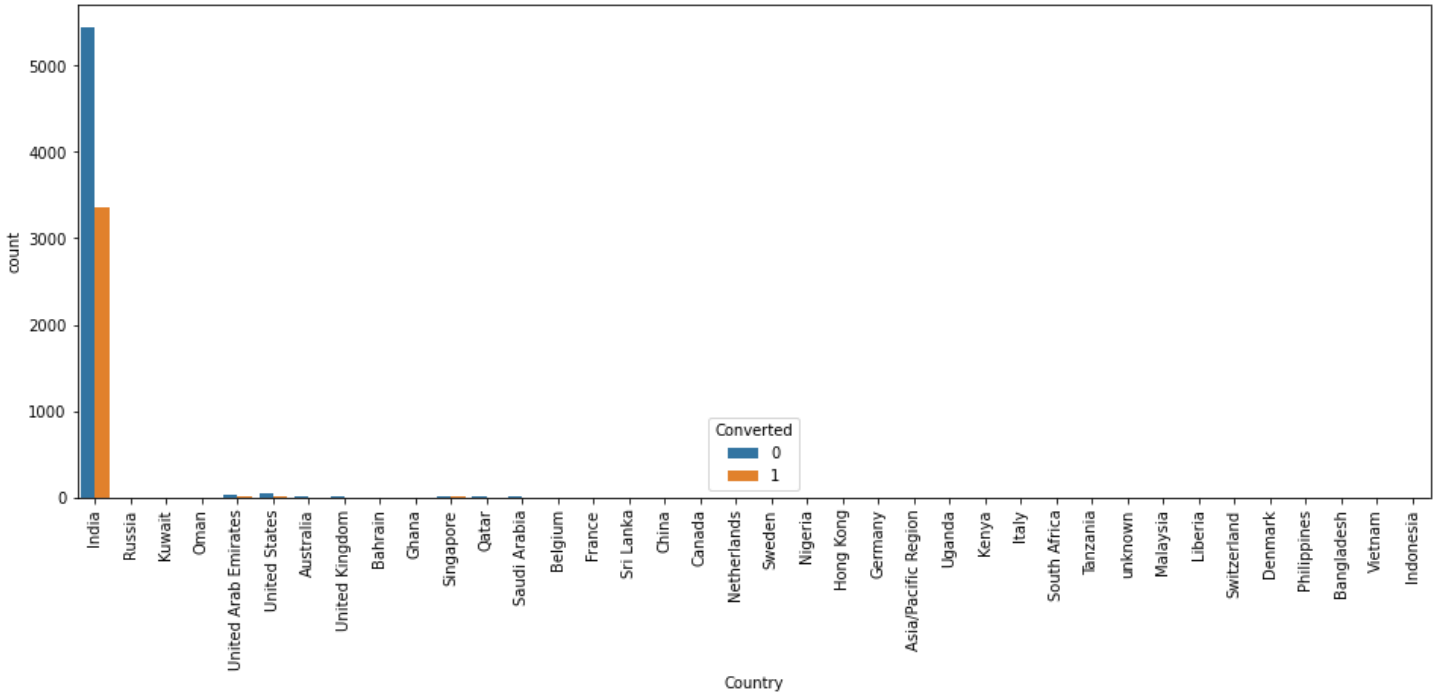
Last Activity



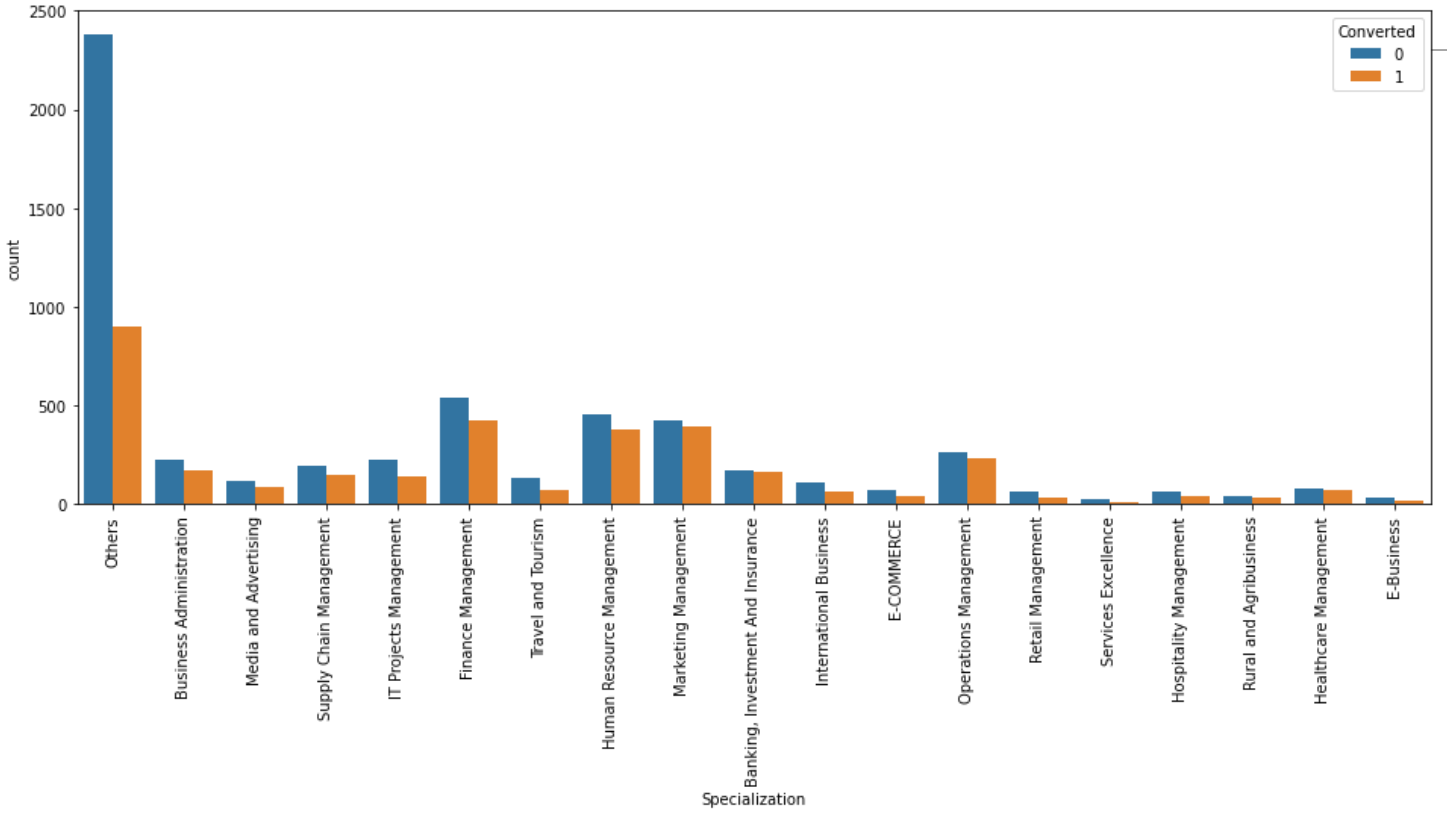
Most of the lead have their Email Opened as thier last activity

Consversion rate for leads with last activity as SAMS sent is almost 60%.

Country



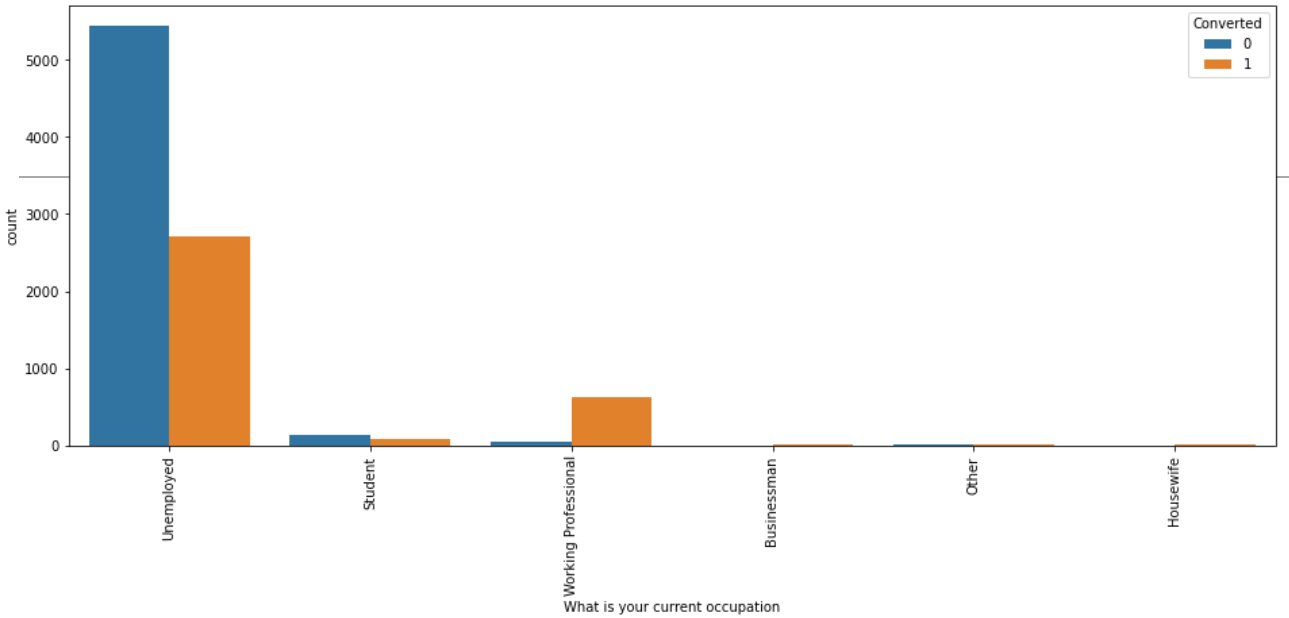
Most Value showing in India



Specialization

Focus should be more on the Specialization with high conversion rate

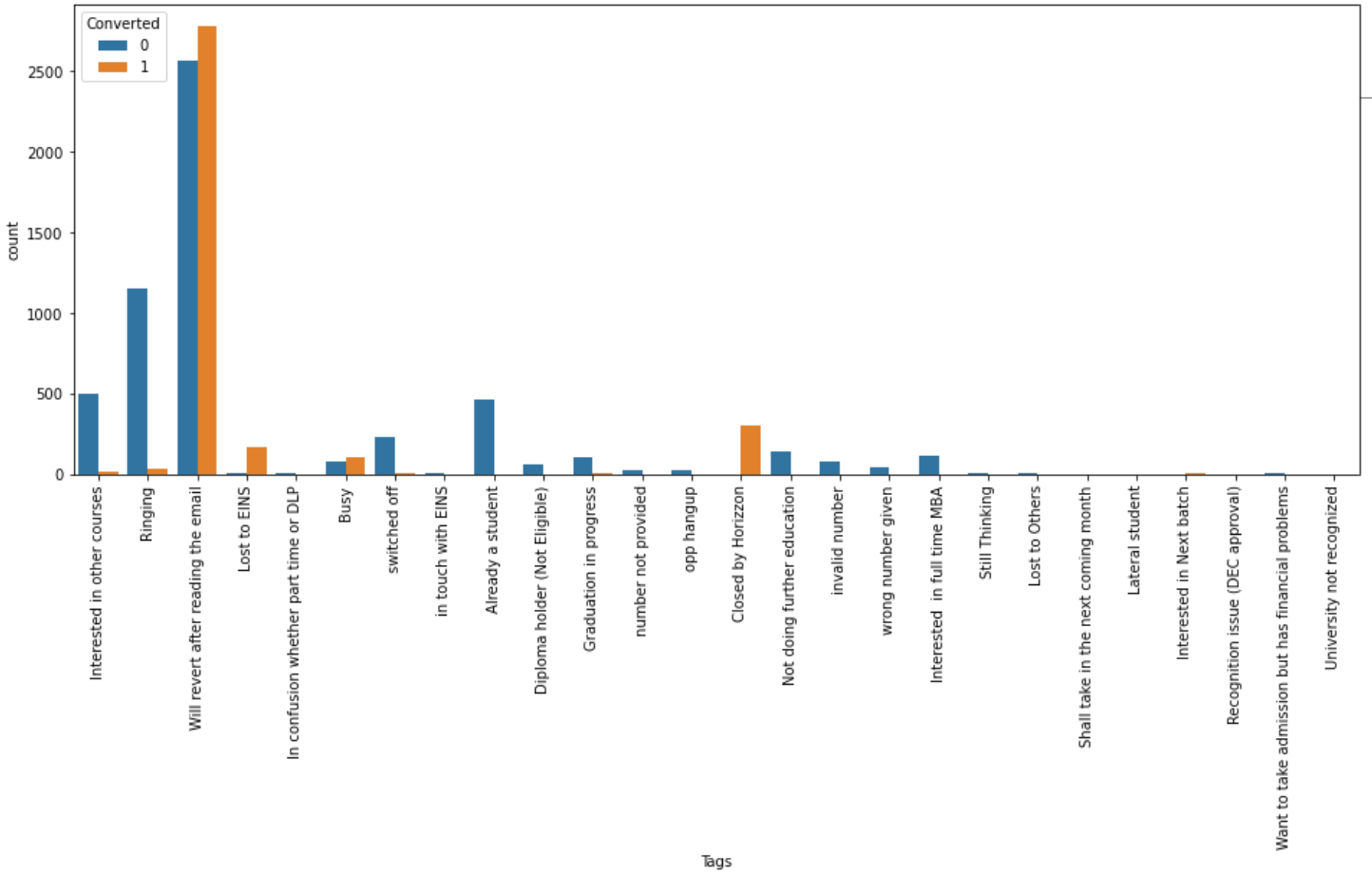
On the other hand, marketing management, human resources, has high converstion rates people from these specialization can be promisin leads



Unemployed percentage is more in numbers.

working professionals has a high chances of joining it

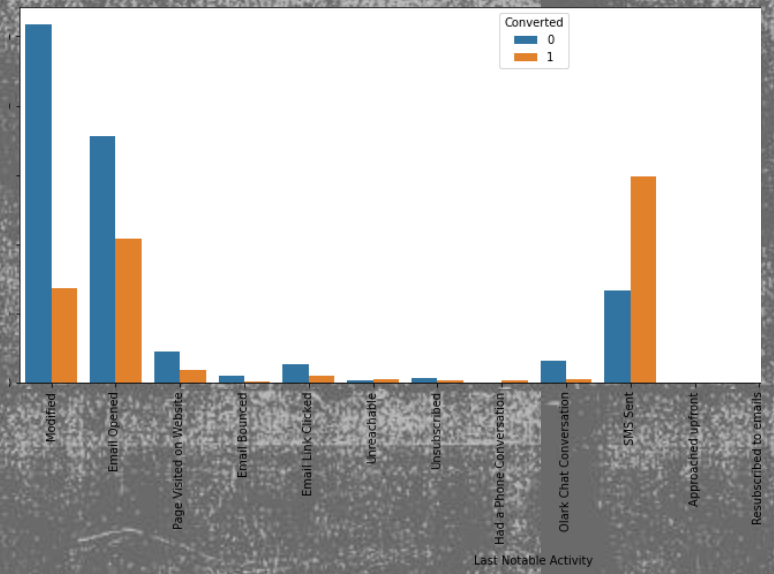
What is your current occupation



Tags

we see that 'We revert after reading the email' has the higher conversion rate

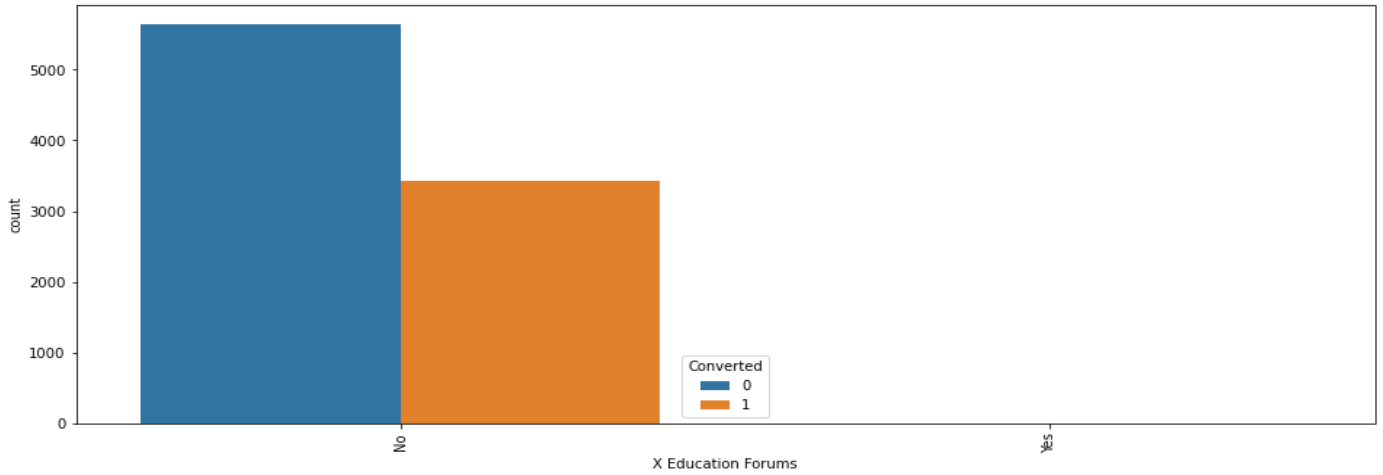
Last Notable Activity

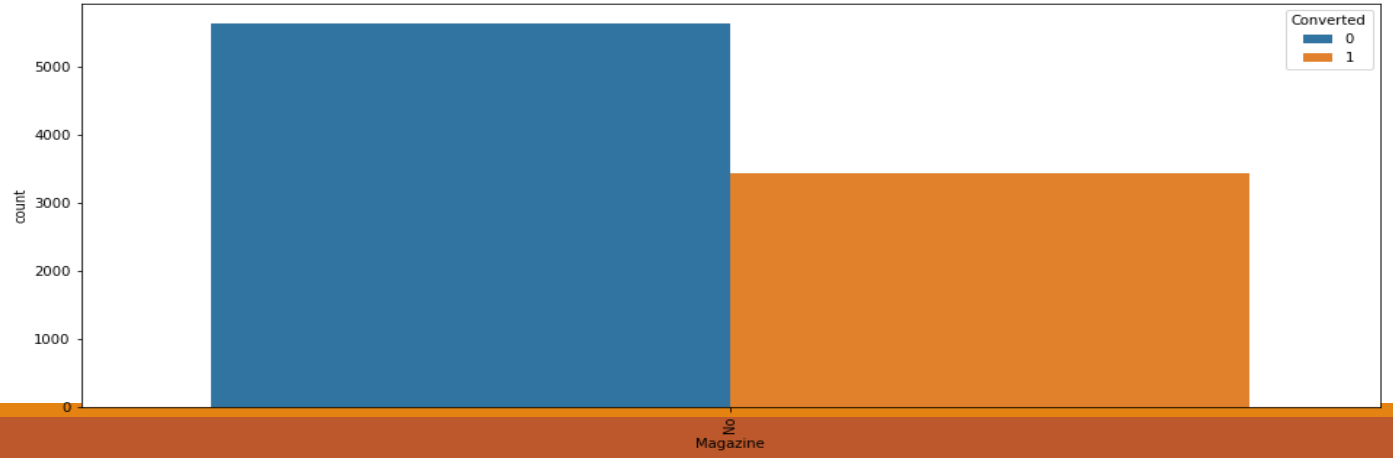


Most of the Columns in the data atre not adding any values to the model, dropping those values which are not required

It is understandable from the above EDA that there are many elements that have very little data and so will be of less relevance to our analysis.

MAgazine and Newspaper article





Entries are No only. No conclusion can be drawn with this parameter

Model Building

Splitting into train and test set

Scale variable in train set

Build first model

Use RFE to eliminate less relevant variables

Eliminate the variables based on P values

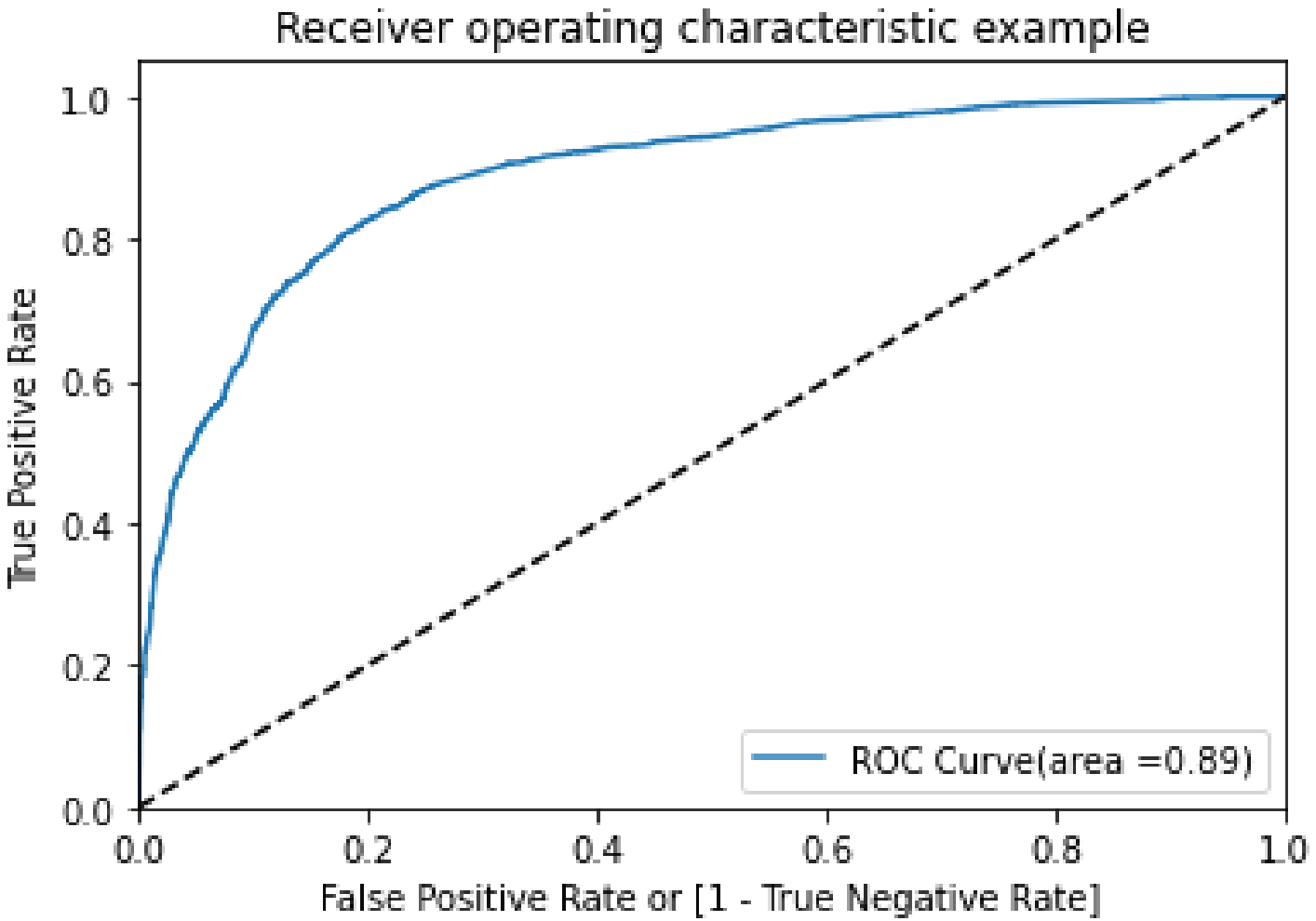
Check VIF values for all the existing columns

Predict using train set

Evaluate accuracy and other metric

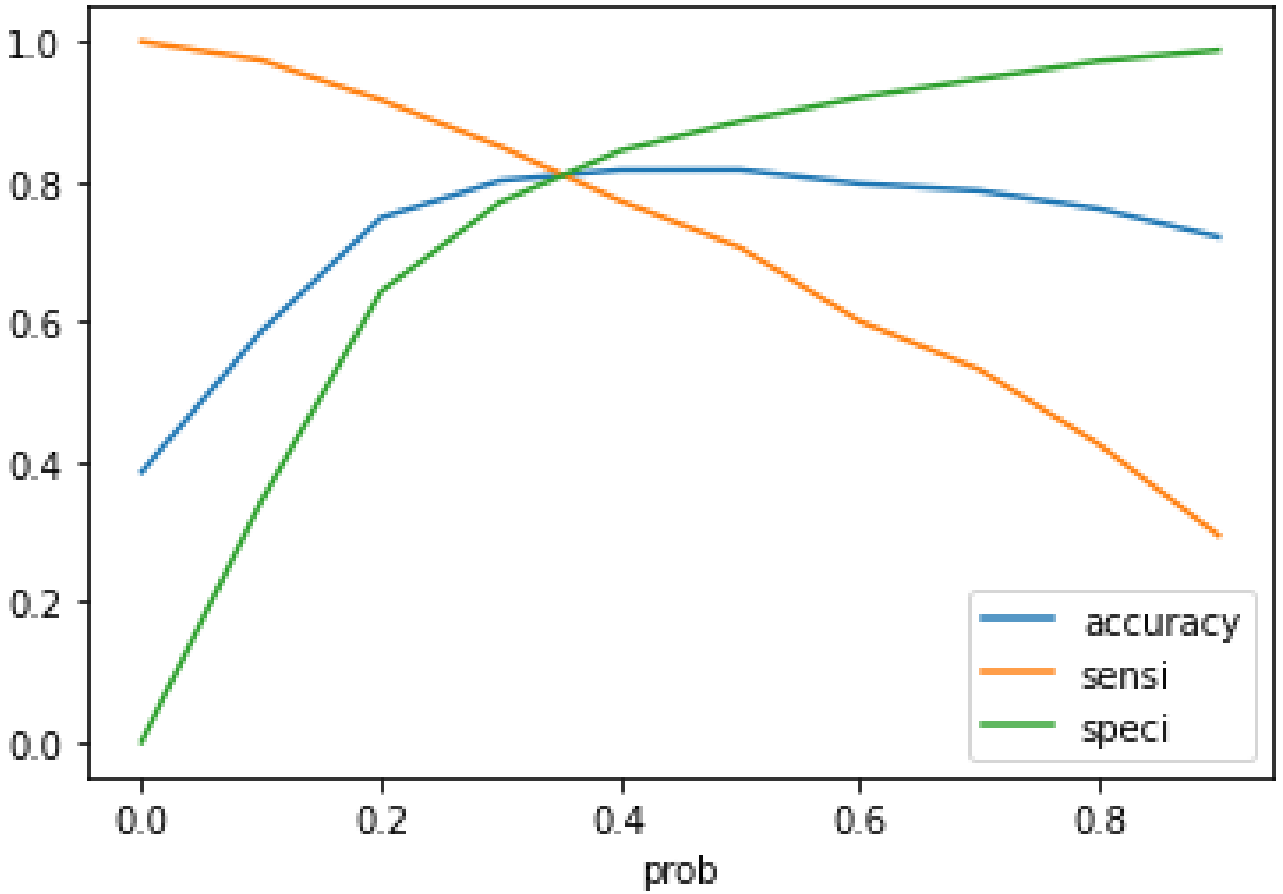
Predict using test set

Precision and recall analysis on test prediction



Optimize Cut off (ROC Curve)

The area under ROC curve is 0.89 which is a very good value.



Model Evaluation(Train)

With the current cut off as 0.34 we have accuracy, sensitivity and specificity of around 80%

Train data:

Accuracy – 81%

Sensitivity – 81.7%

Specificity – 80.6%

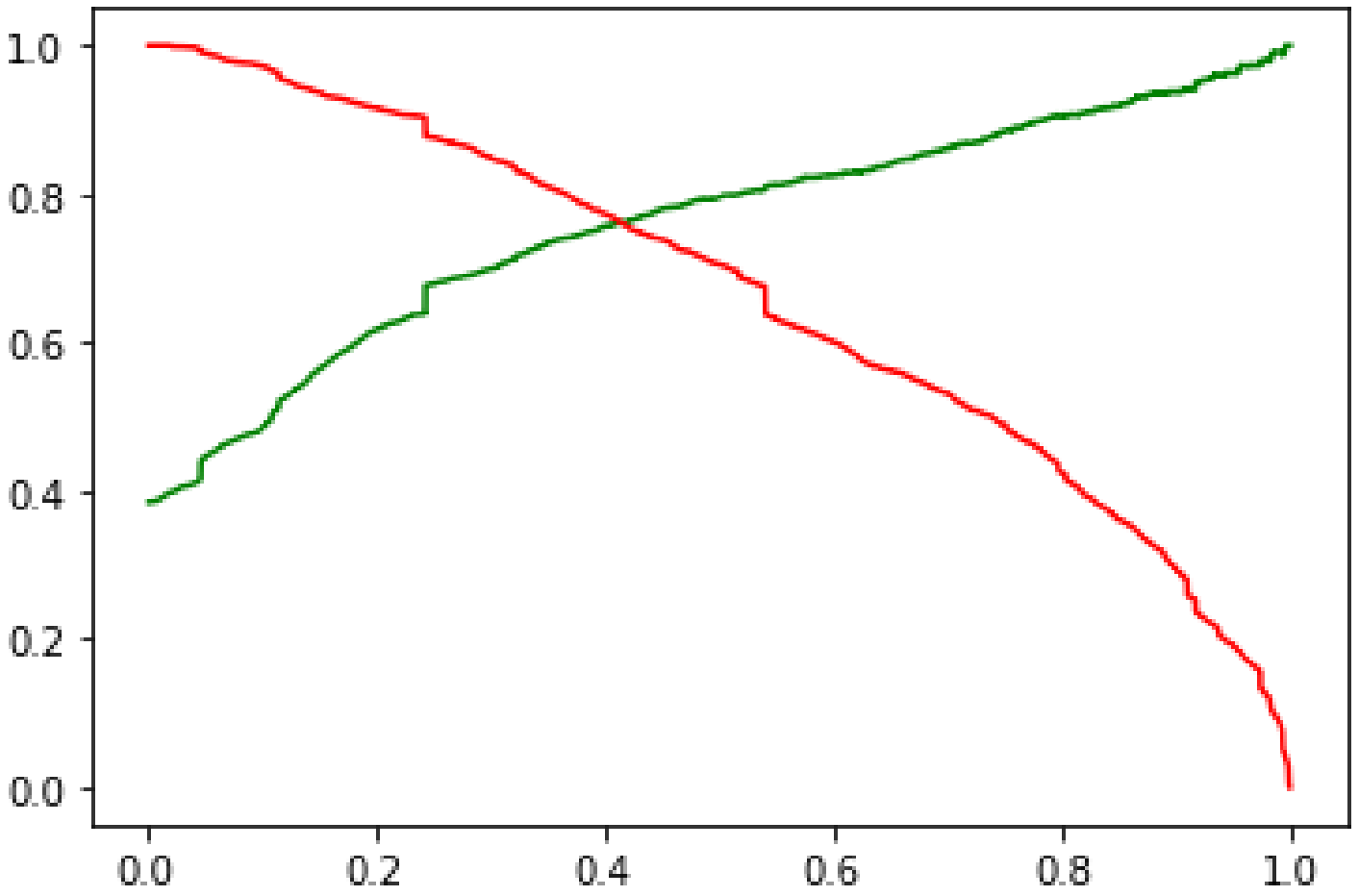
Test data:

Accuracy – 80.4%

Sensitivity – 80.4%

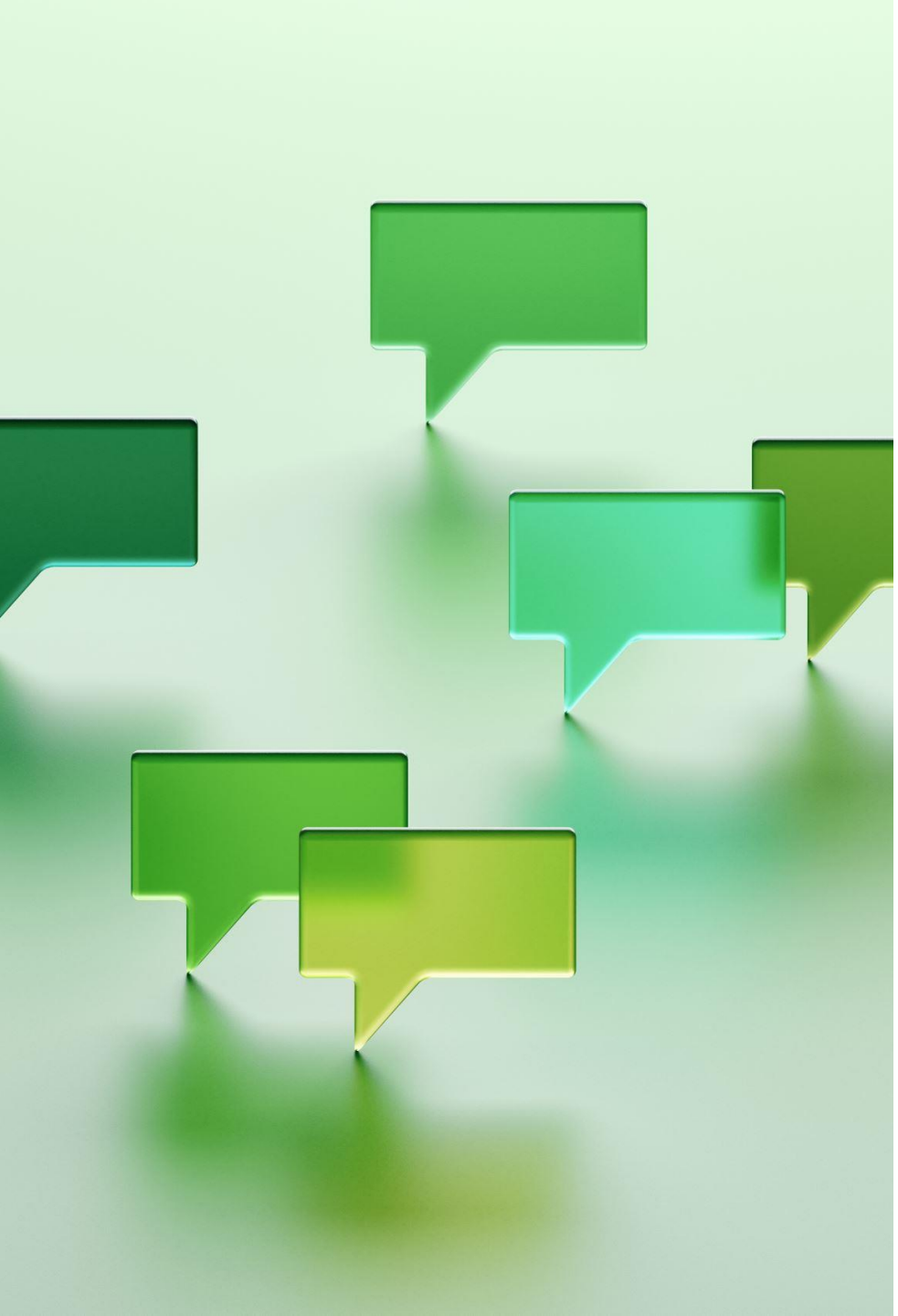
Specificity – 80.5%

Thus we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% . 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 to get a higher lead conversion rate of 80%.



Precision Recall

With the current cut off as 0.34 we have Precision around 72% and Recall around 81%.



Conclusion:

It was found that the variables that mattered the most is the potential buyers are:

1. the total time spend on website

2. total number of visits.

3. the working professionals as current occupations

4. the last Activity of SMS sent, Olark chat conversations.

5. When the lead origin is Lead add format. 6. When their current occupation is as a working professional.
Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses