Lead Scoring Case Study

Logistic Regression

- By Ramya Patnayakuni, Priya Nayak and Pradeep Sharma

Problem Statement

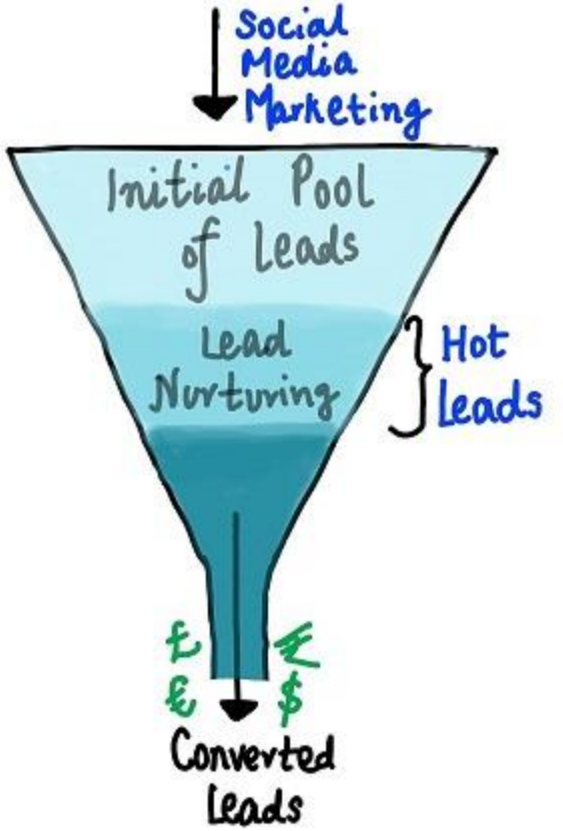
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of 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.

Problem Statement

A typical lead conversion process can be represented using the following funnel:



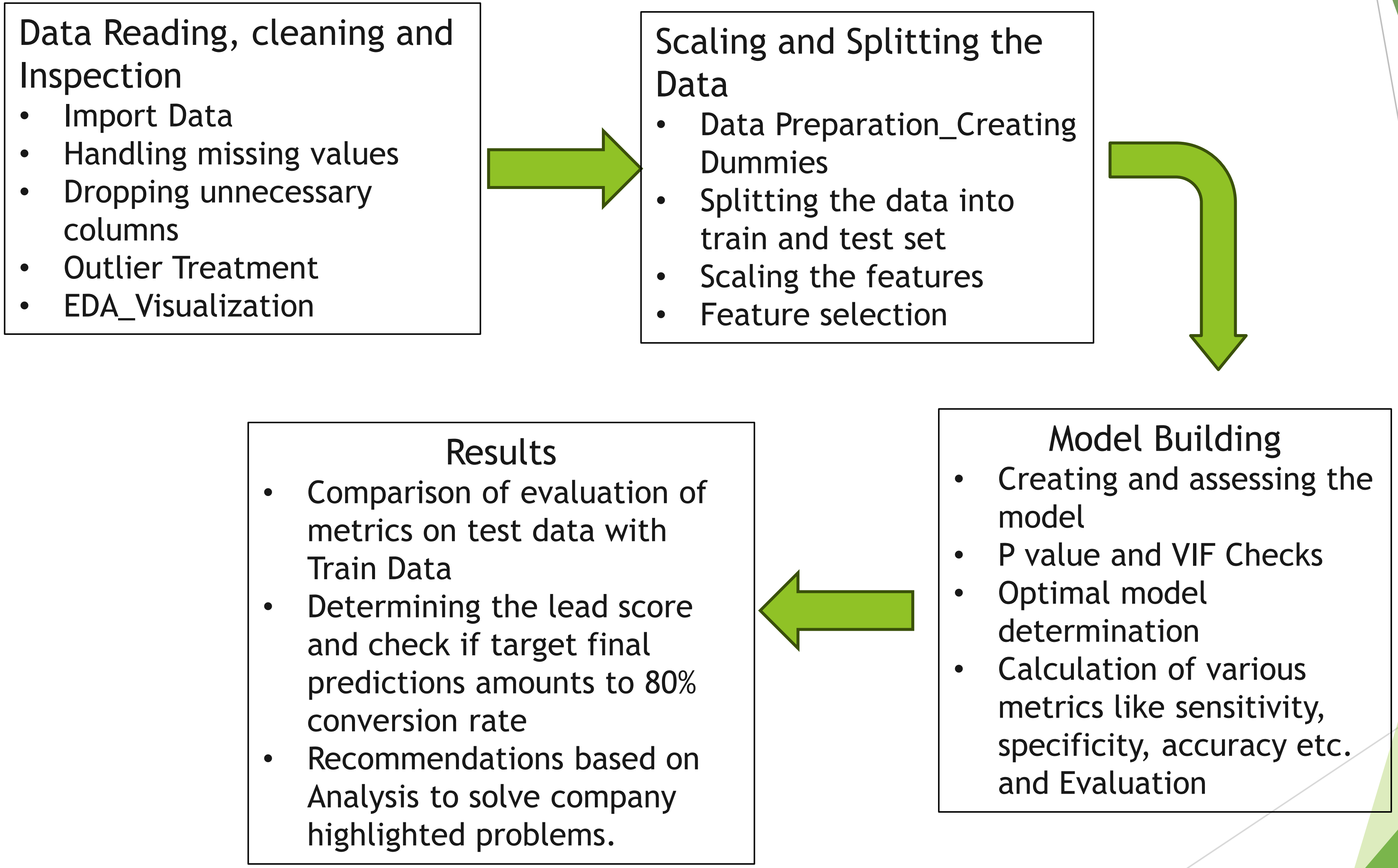
X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you **need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.**

Goals of the Case Study

1.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.

2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Analysis Approach



Data Reading, cleaning and Inspection

- Import Data
- Handling missing values
- Dropping unnecessary columns
- Outlier Treatment
- EDA_Visualization

Scaling and Splitting the Data

- Data Preparation_Creating Dummies
- Splitting the data into train and test set
- Scaling the features
- Feature selection

Model Building

- Creating and assessing the model
- P value and VIF Checks
- Optimal model determination
- Calculation of various metrics like sensitivity, specificity, accuracy etc. and Evaluation

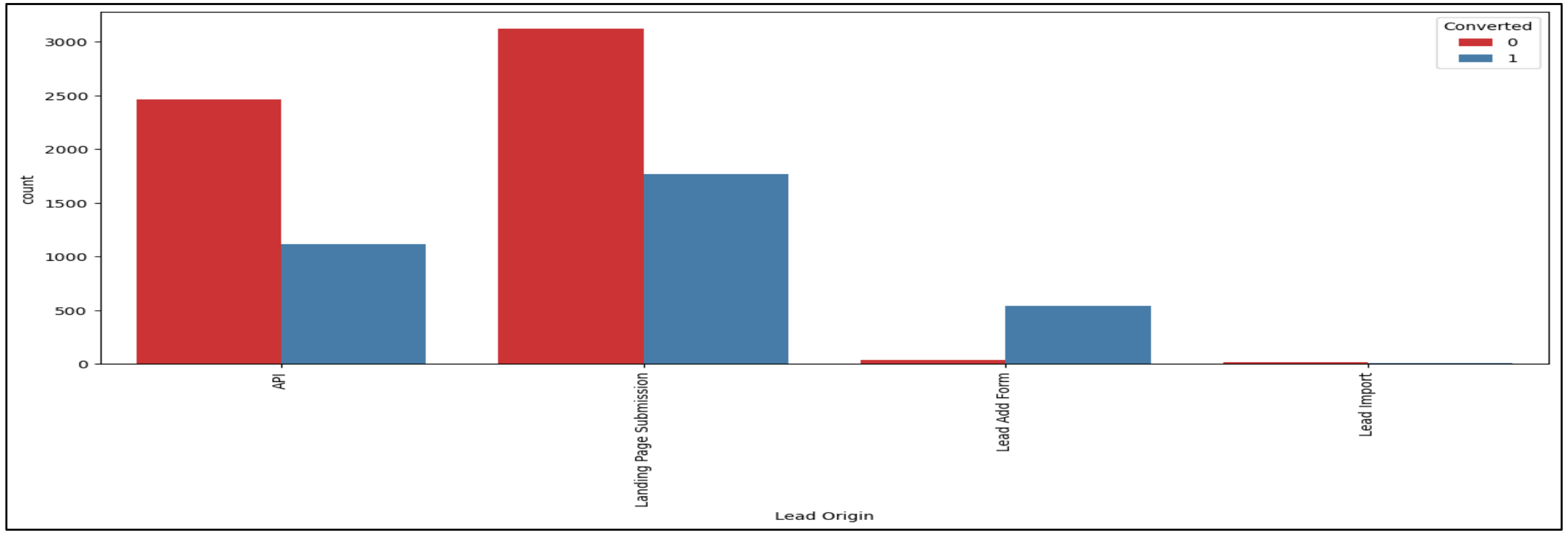
Results

- Comparison of evaluation of metrics on test data with Train Data
- Determining the lead score and check if target final predictions amounts to 80% conversion rate
- Recommendations based on Analysis to solve company highlighted problems.

Data Cleaning

- Initially Data contains 9240 rows and 37 columns
- Checked for nulls in all the columns and removed the the columns where null percentage is greater than 40%
- There is 37% missing values present in the Specialization column.It may be possible that the lead may leave this column blank if he may be a student or not having any specialization or his specialization is not there in the options given. So, created a another category 'Others' for this.
- Imputed Missing values in columns like “Tags column”, “What is your current occupation”, 'Country’, 'City’,.
- There were a few columns in which only one value was majorly present for all the data points. These include Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, 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, What matters most to you in choosing a course. Since practically all of the values for these variables are No, it's best that we drop these columns as they won't help with our analysis.
- After Data cleaning, the dataset contains 9074 rows and 13 columns

Analysing Categorical Variables w.r.t Count variable

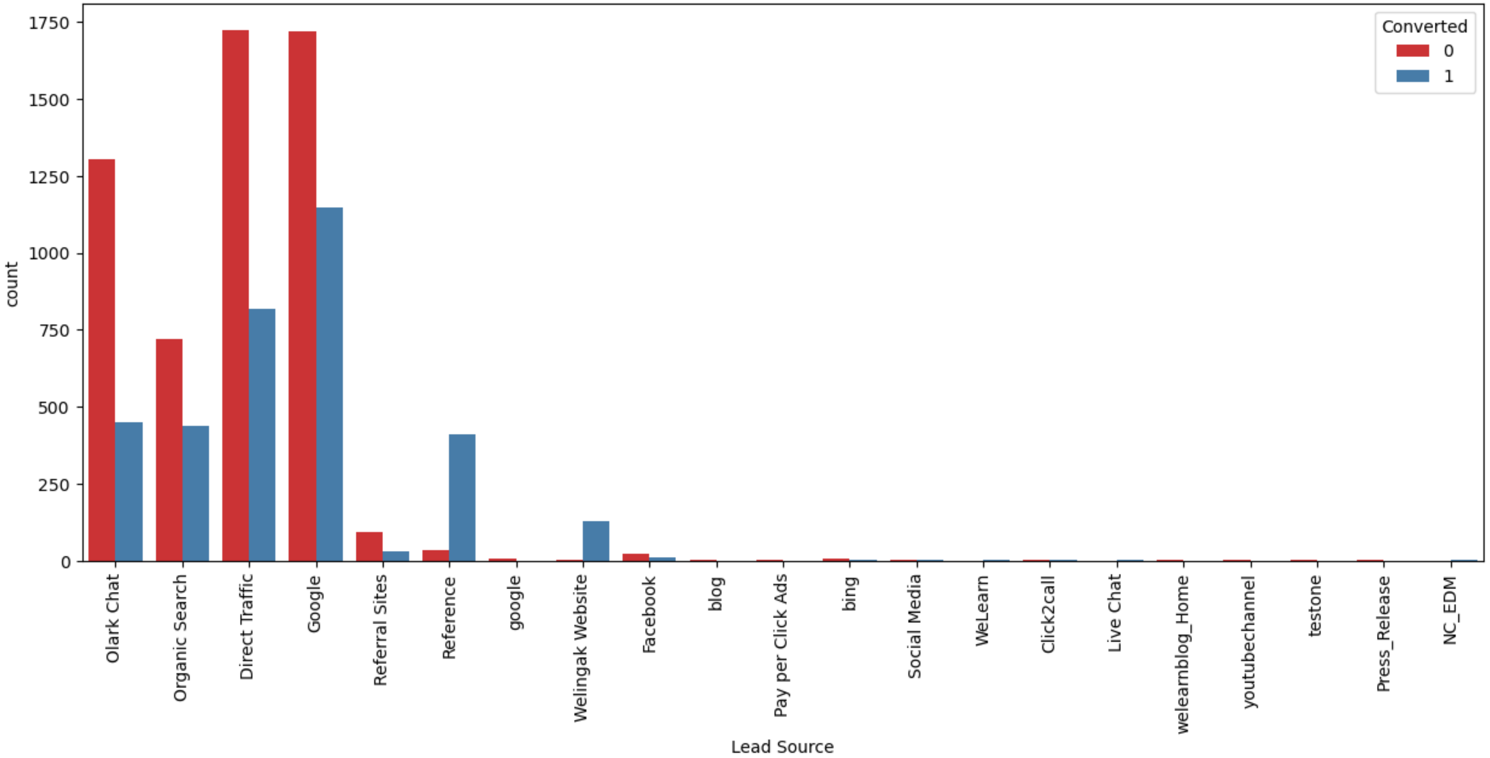


Inferences:

- API and Landing Page Submission have 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 very high.
- Lead Import are very less in count

To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

Analysing Categorical Variables w.r.t Count variable

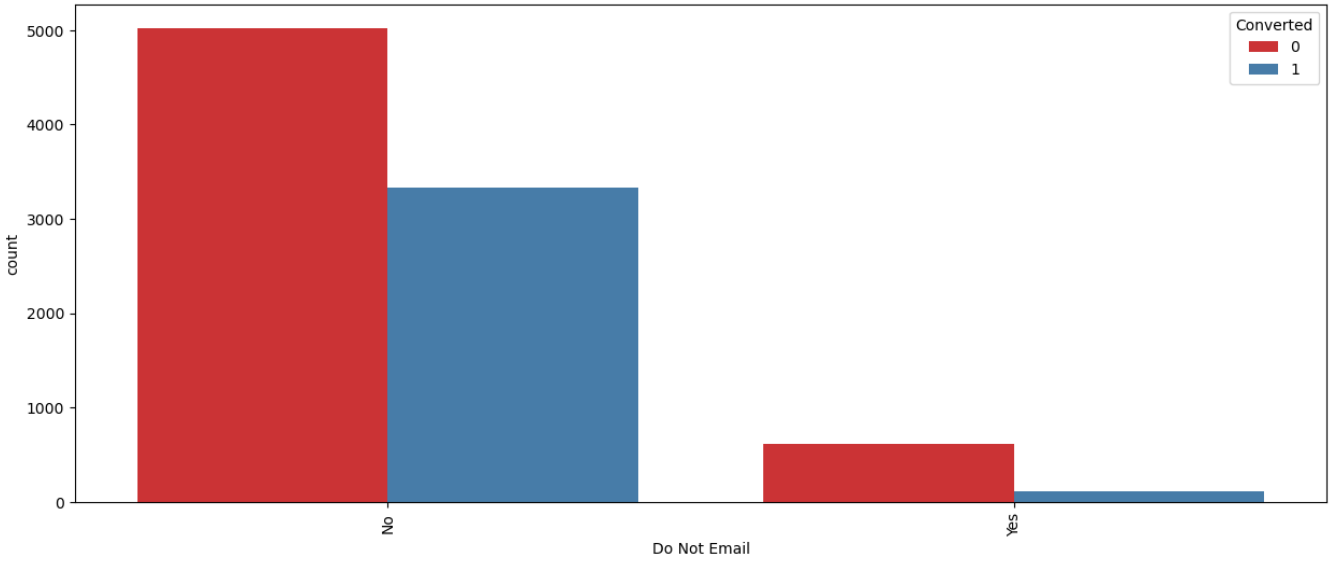


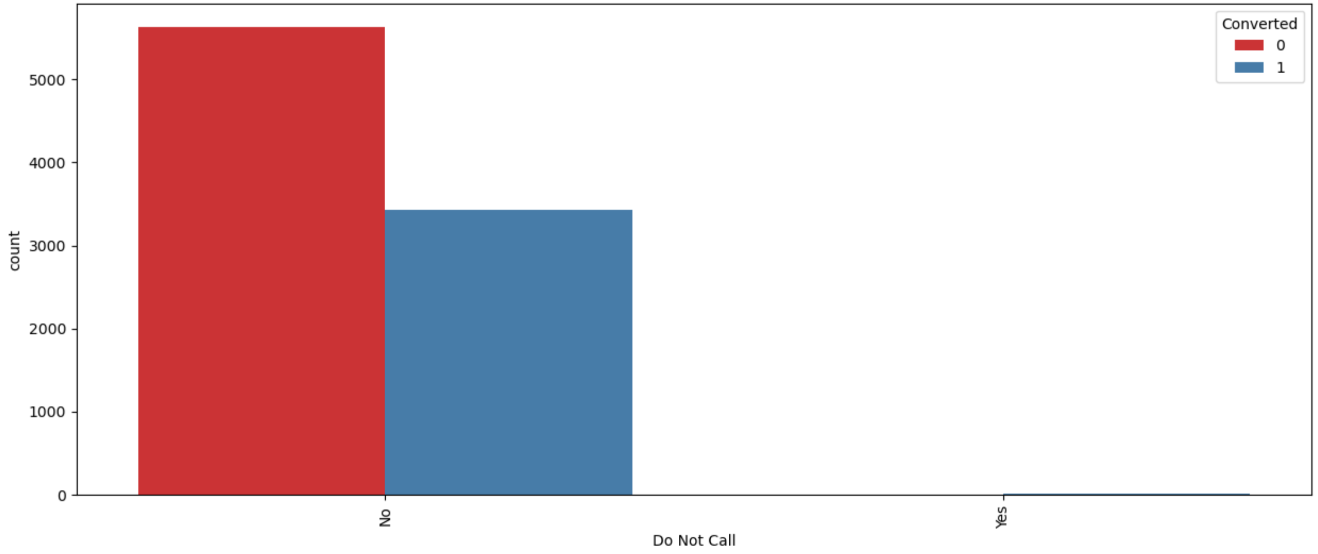
Inferences:

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.

To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

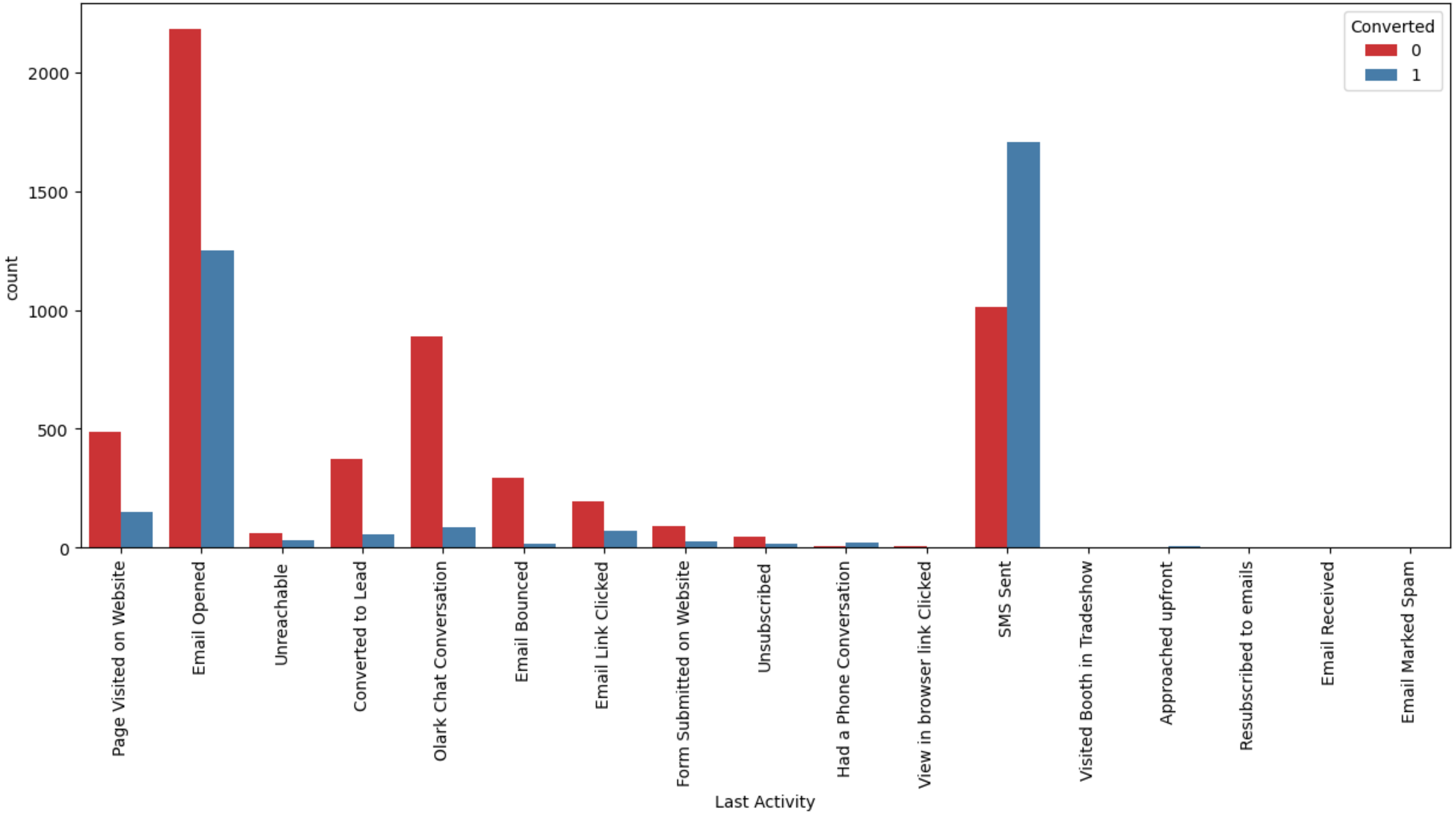
Analysing Categorical Variables w.r.t Count variable





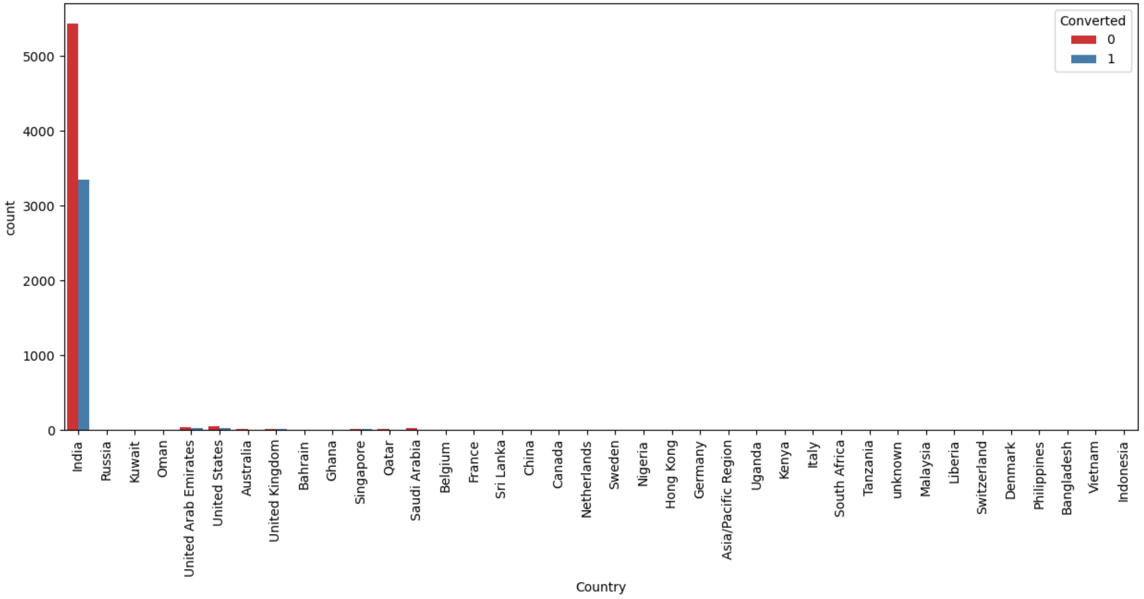
Most of the customers don’t want to receive mails/calls Irrespective of their interest to join the course or not.

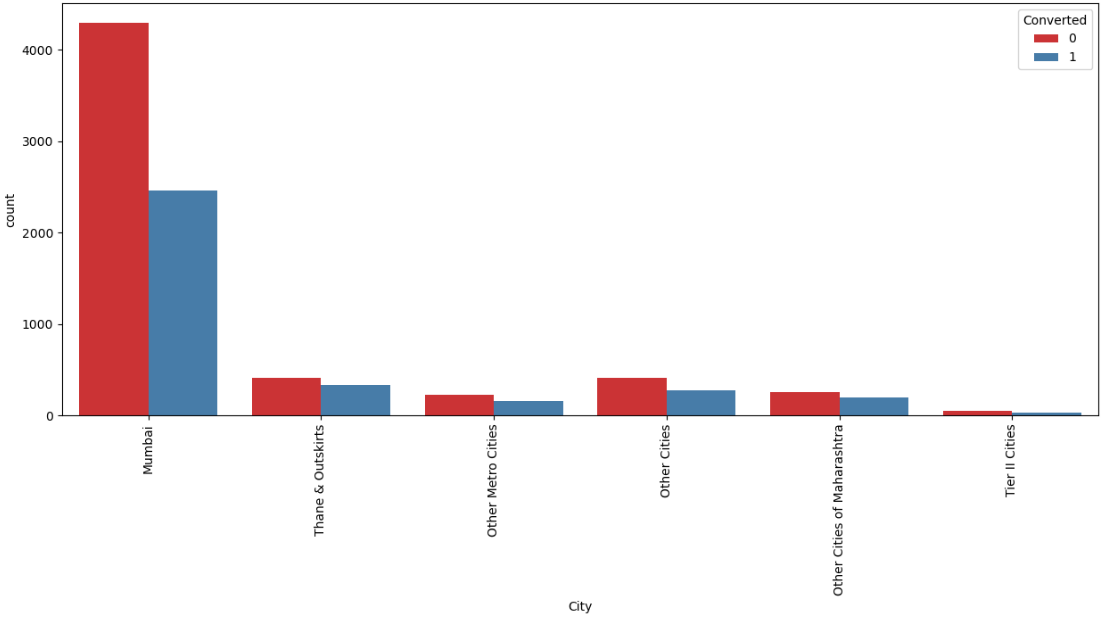
Analysing Categorical Variables w.r.t Count variable



- Most of the leads have their Email opened as their last activity and an sms is sent for most.

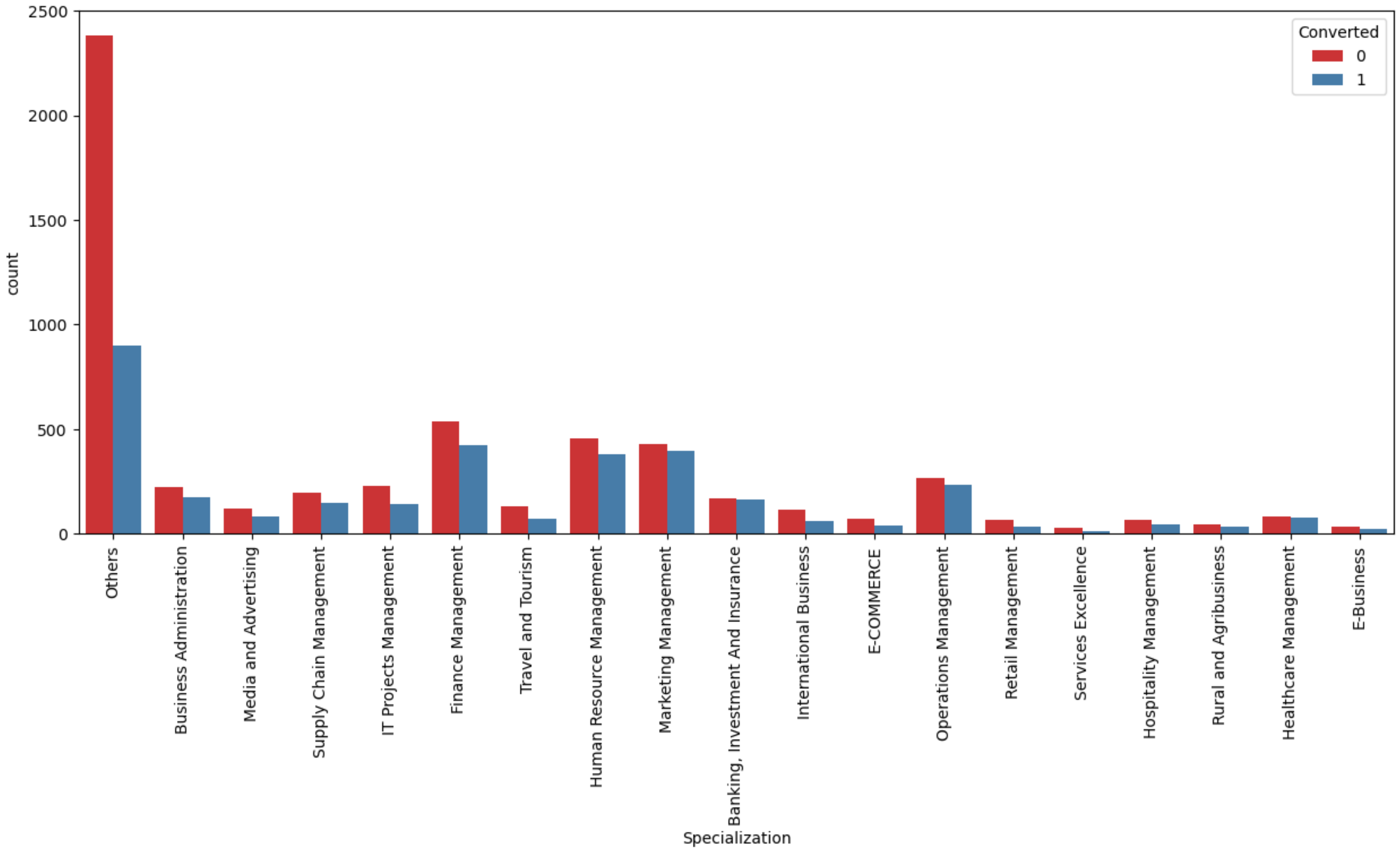
Analysing Categorical Variables w.r.t Count variable





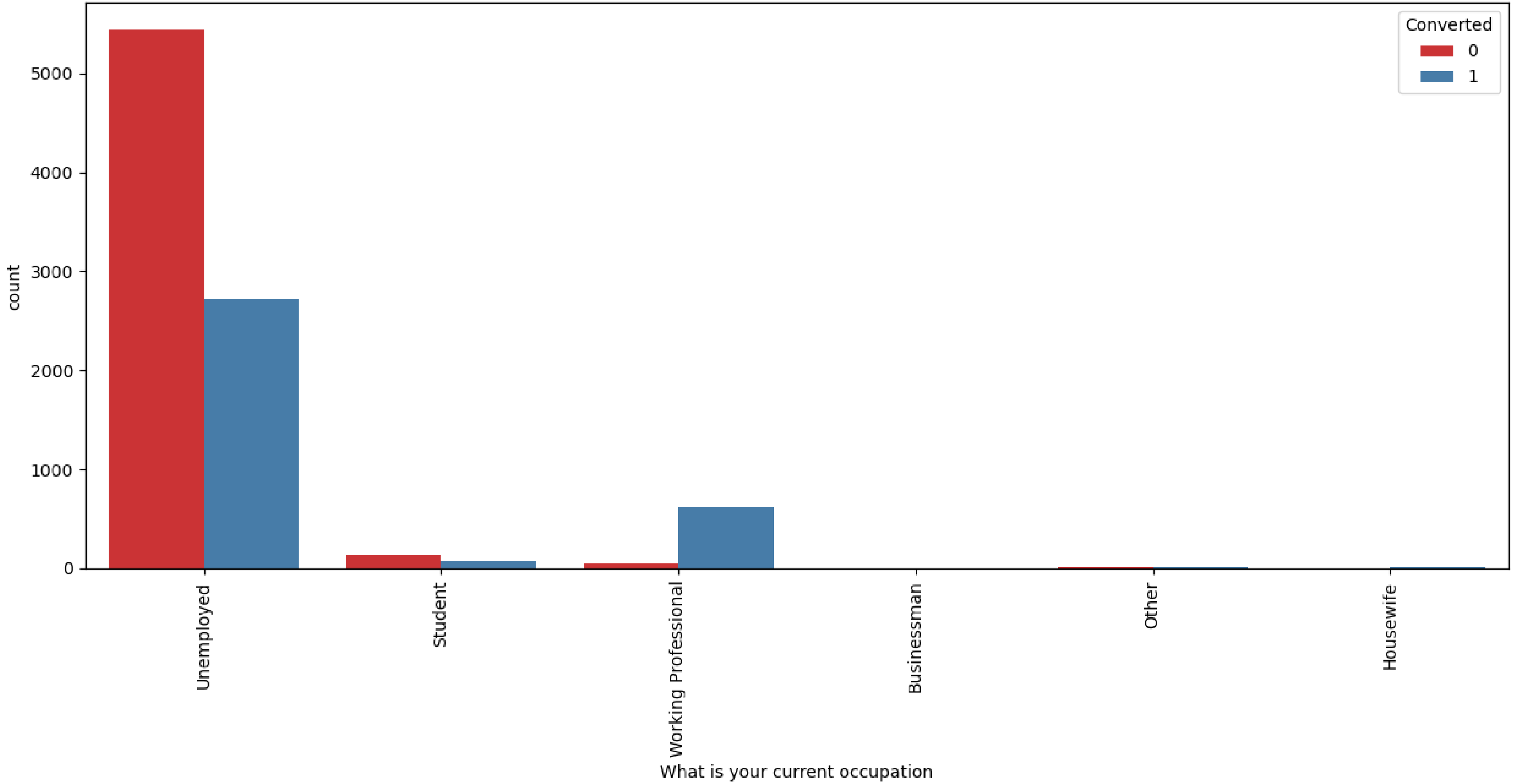
- Country and city are not that helpful in analysis but we can infer that most of the leads are from India and Mumbai city

Analysing Categorical Variables w.r.t Count variable



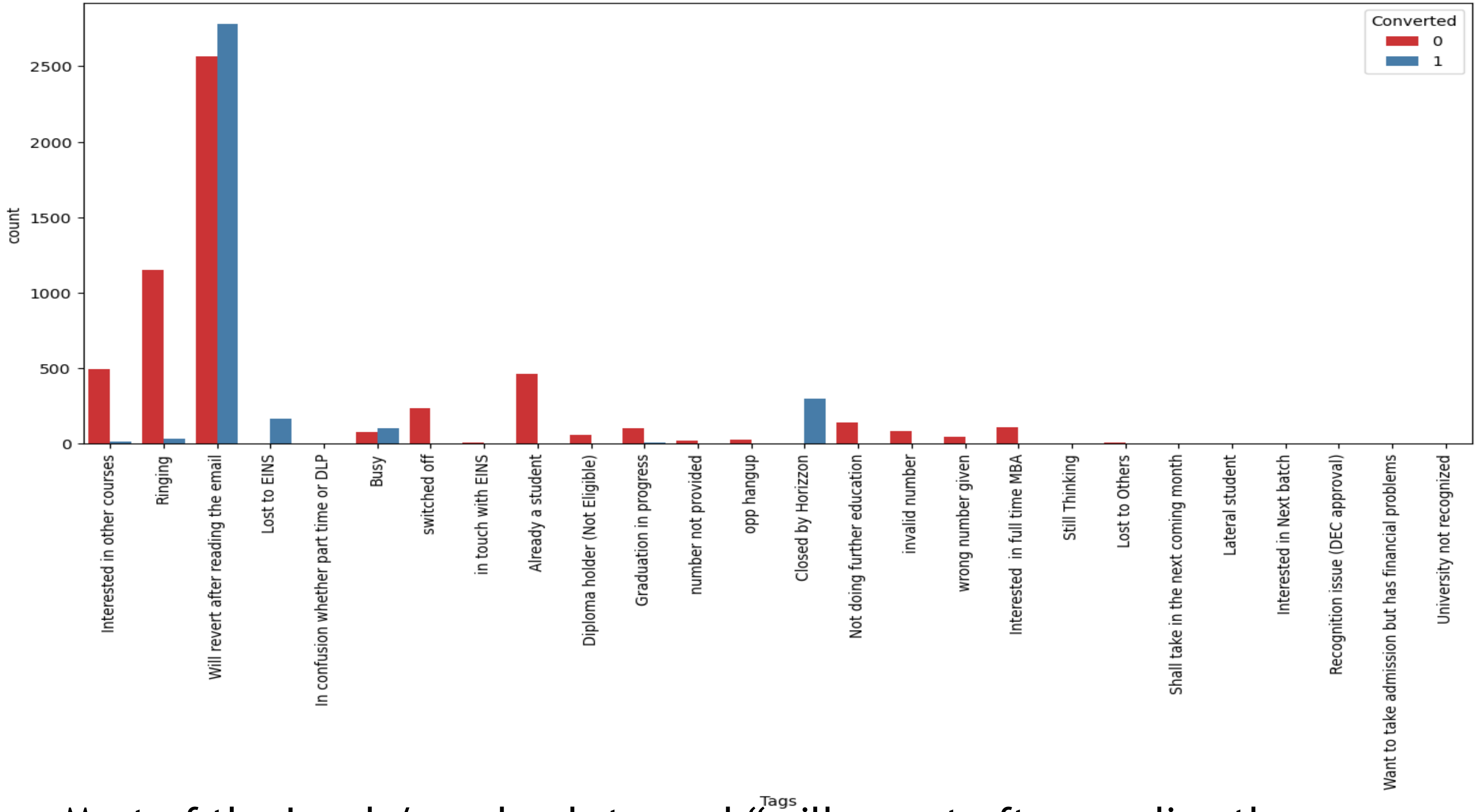
Though conversion rate is less, the others category have more conversions. As assumed, he/she may be a student or not having any specialization or his specialization is not there in the options given.

Analysing Categorical Variables w.r.t Count variable



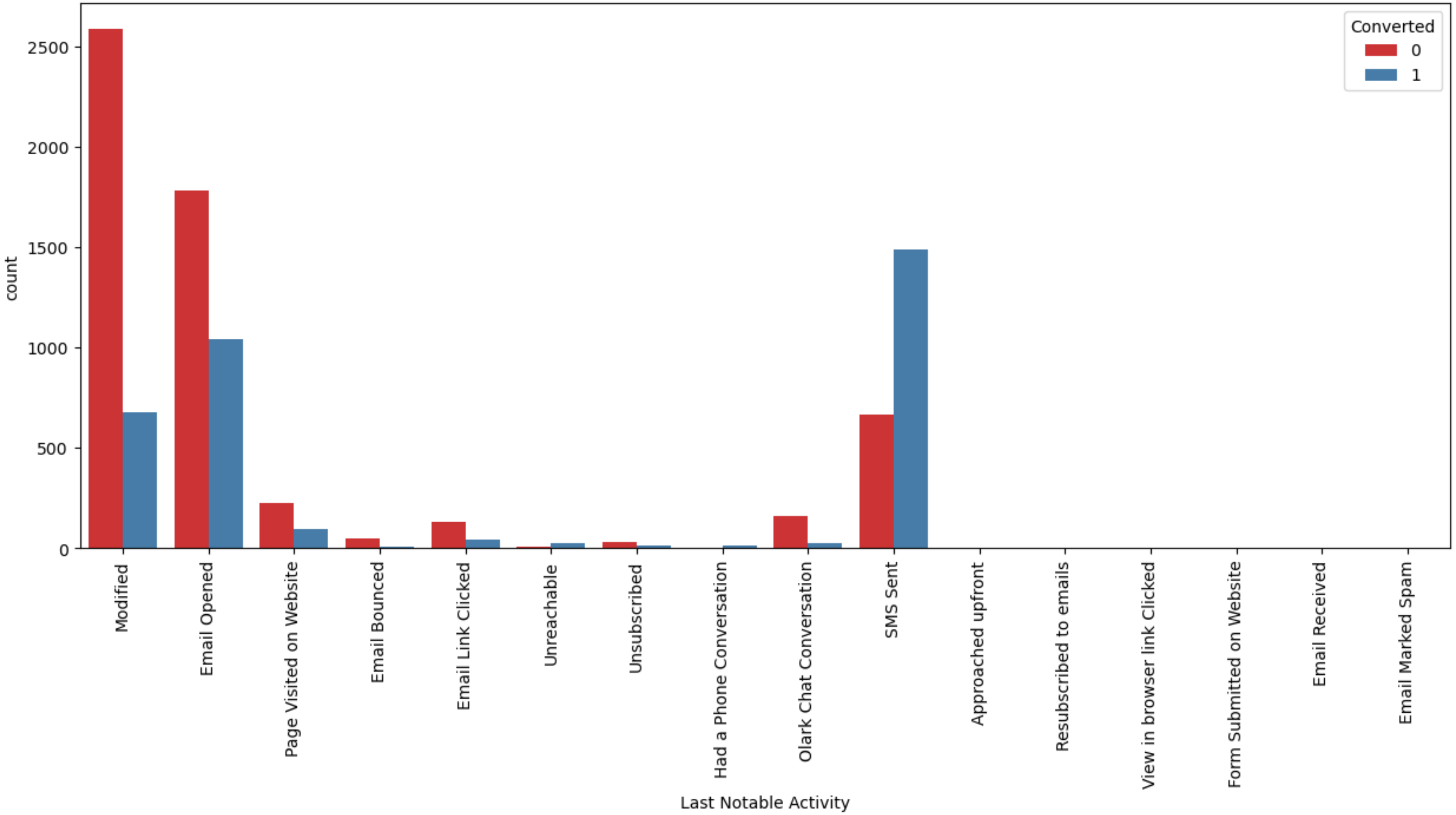
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in numbers

Analysing Categorical Variables w.r.t Count variable



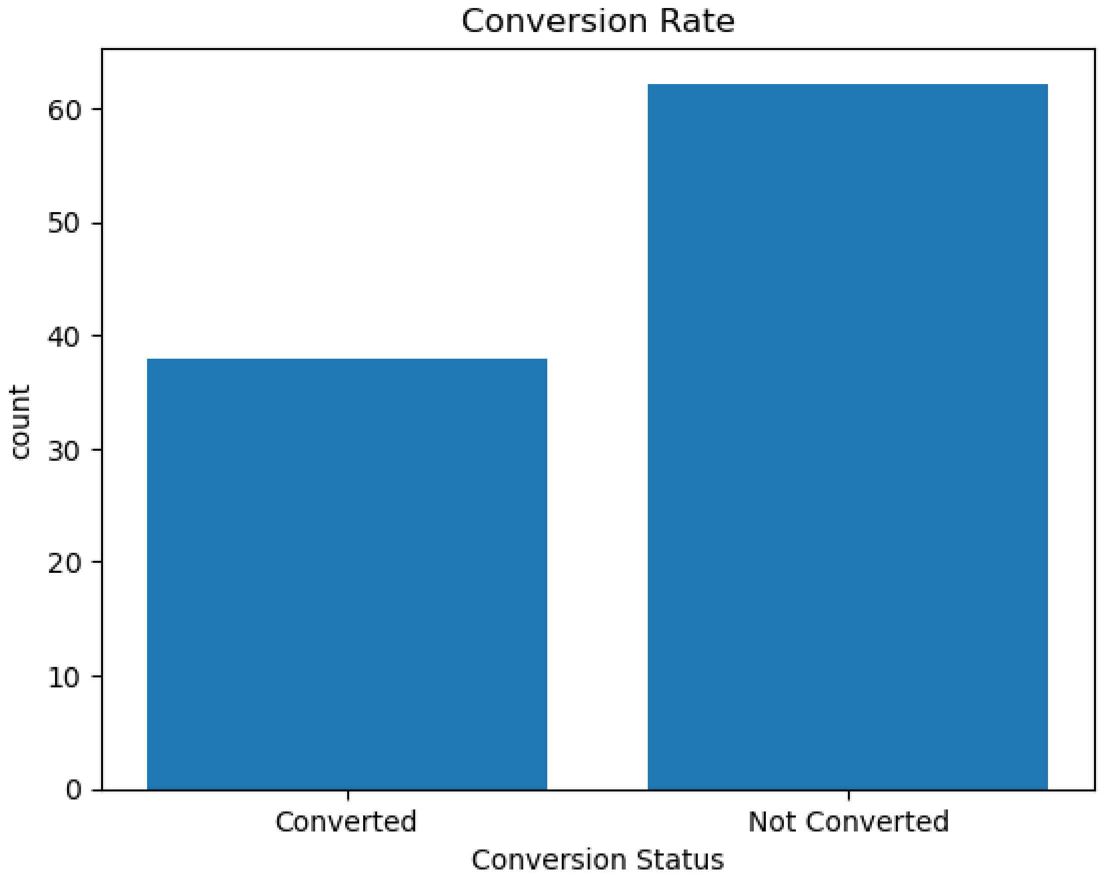
- Most of the Leads/non-leads tagged “will revert after reading the mail”

Analysing Categorical Variables w.r.t Count variable



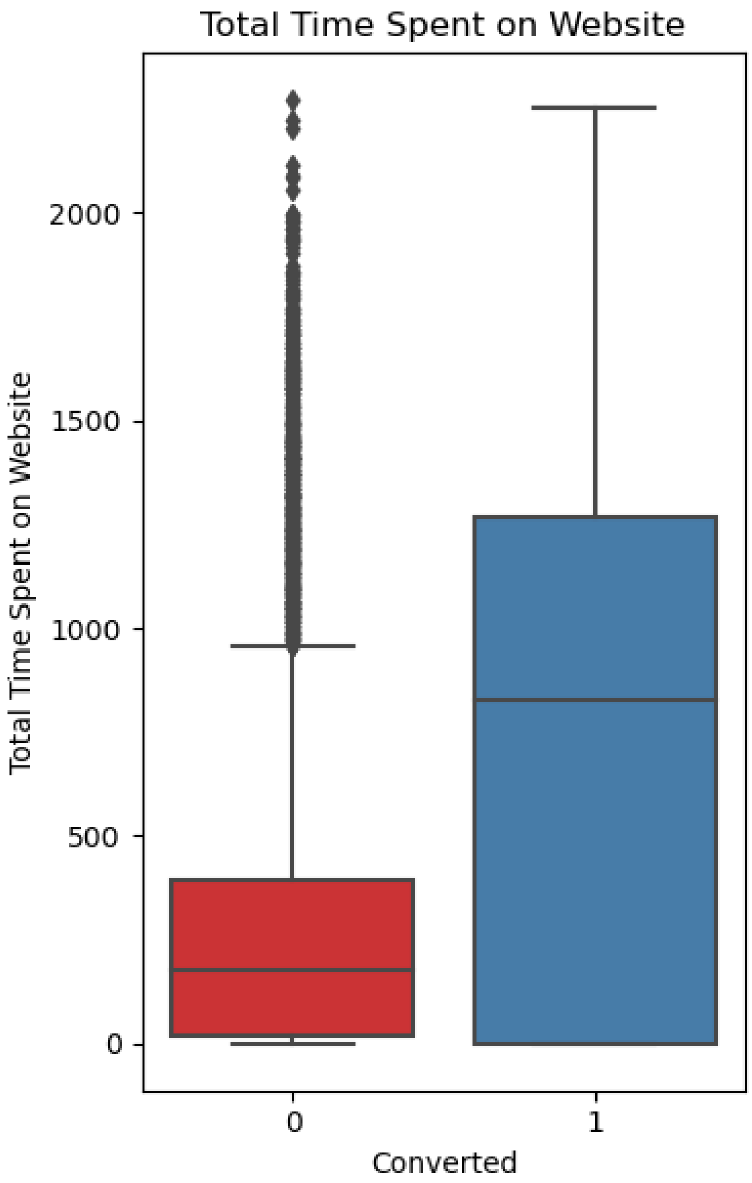
- Most of the lead have their Email opened as their last activity.

Analysing Numerical Variables w.r.t Count variable



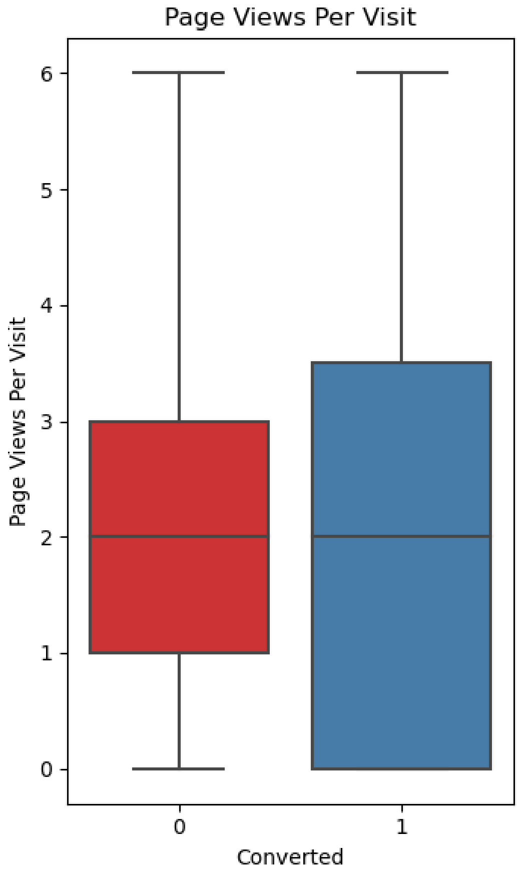
Lead conversion percentage is around 38%

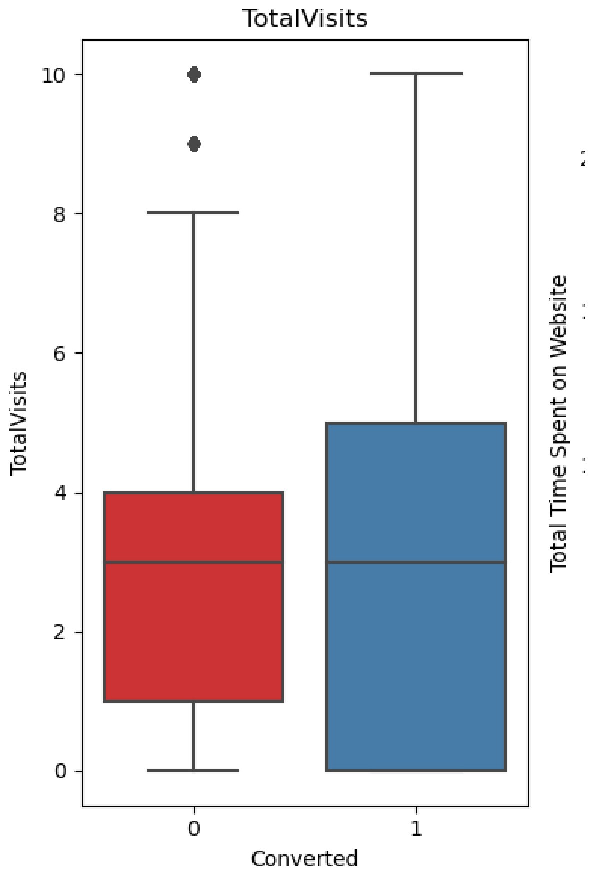
Analysing Numerical Variables w.r.t Count variable



Inference:
1.Leads spending more time on the website are more likely to be converted.
2.Website should be made more engaging to make leads spend more time.

Analysing Numerical Variables w.r.t Count variable



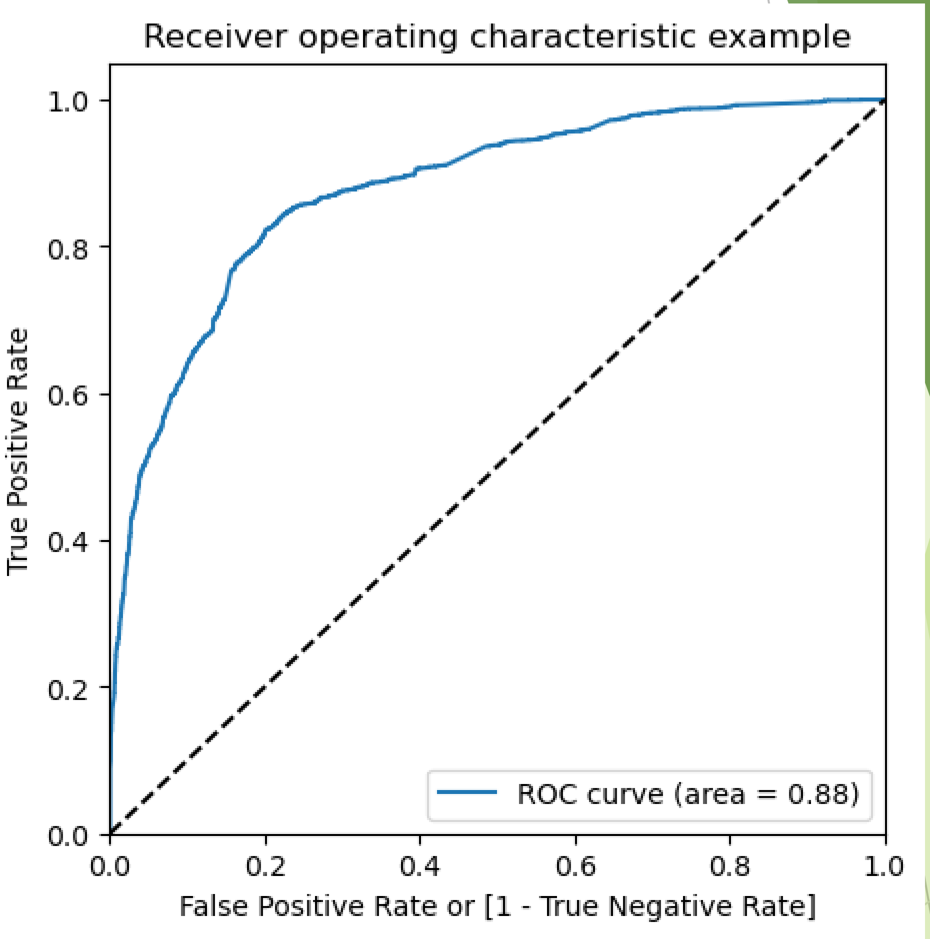


Inference:
1.Median for converted and unconverted leads is the same.
2.Nothing can be said specifically for lead conversion from Page Views Per Visit

Inference:
1.Median for converted and not converted leads are the close.
2.Nothing can be concluded on the basis of Total Visits

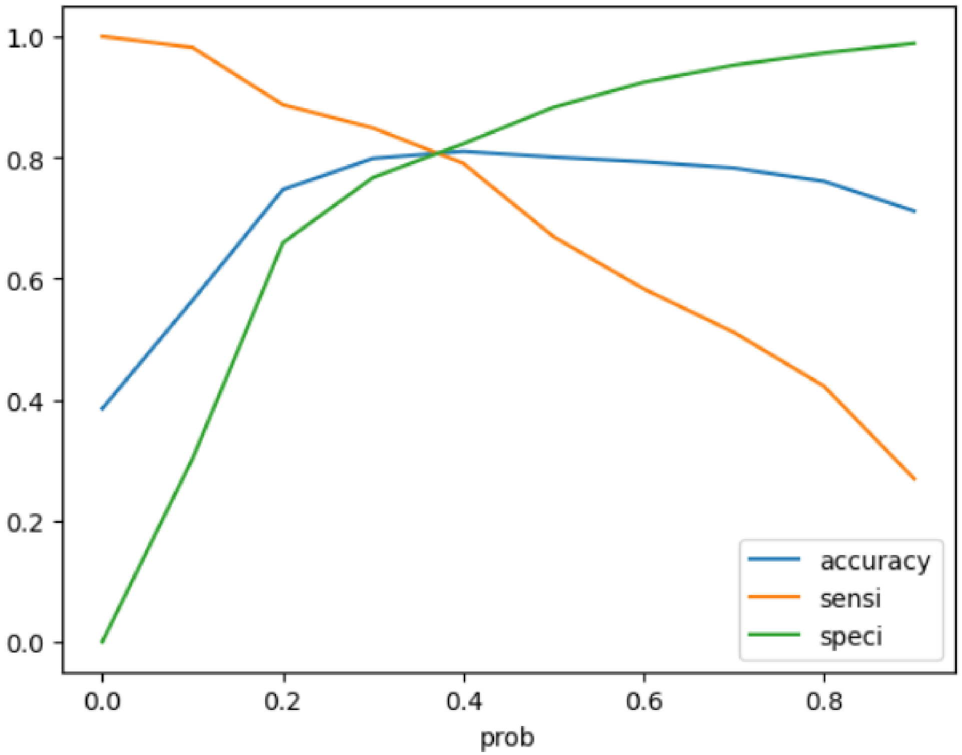
Data Preparation for Modelling & Model Evaluation

- Converted some binary variables (Yes/No) to 1/0
- Dummy variable created for the categorical variables and dropping the first one.
- Concatenated the dummy data to the lead data data-frame.
- The dataset contains 9074 rows and 13 columns
- Split the data into train and test data sets in 70:30 ratio.
- Scaled the numerical features using standard scaler.
- RFE to select the columns and build Logistic regression Model.
- Ensuring all p values tending to 0 and VIF values are low, arriving at the final model with n number of variables.
- Making predictions on Train data set based on assumed probability 0.5.
- Calculated the sensitivity, specificity, accuracy of model on train data set with predicted prob of 0.5
- Found optimal cutoff point using ROC curve, 0.36 is the optimum point to take it as a cutoff probability
- Recalculating the sensitivity, specificity, accuracy of model on train data set with predicted prob of 0.36

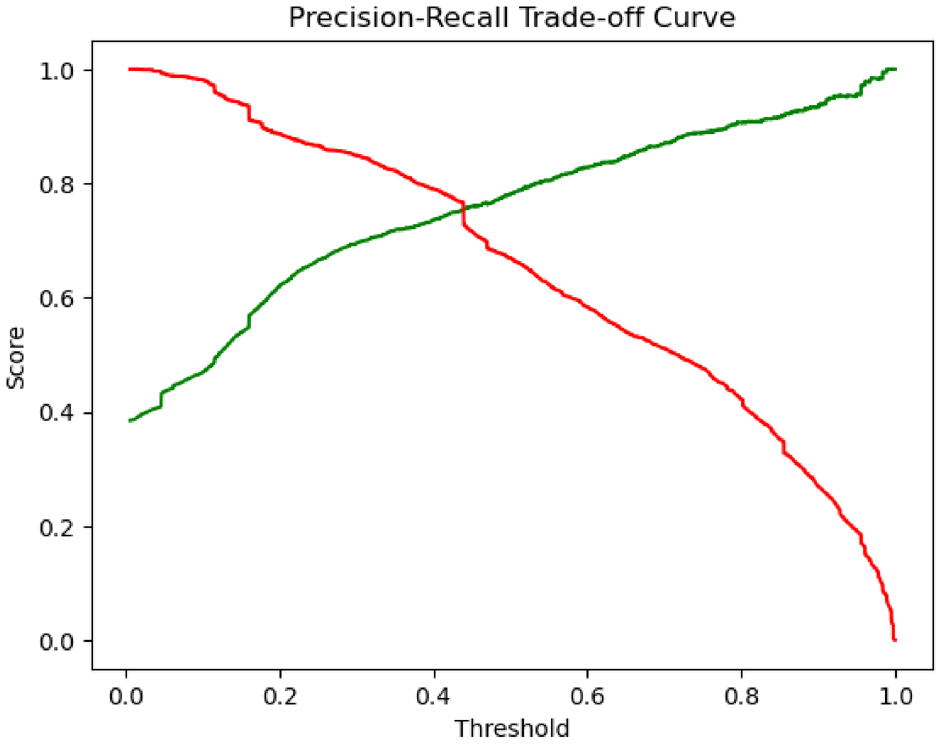


Since we have higher (0.88) area under the ROC curve , therefore our model is a good one.

Model Evaluation



0.36 is the optimum point to take it as a cutoff probability based on metrics



The graph depicts an optimal cutoff of 0.42 based on precision and recall

Model Evaluation using Train & Test Data sets

Train Dataset

Confusion Matrix: [[3129, 776],
[452, 1994]],

Trained Data Results
Accuracy : 80.66 %
Sensitivity : 81.52 %
Specificity : 80.12 %

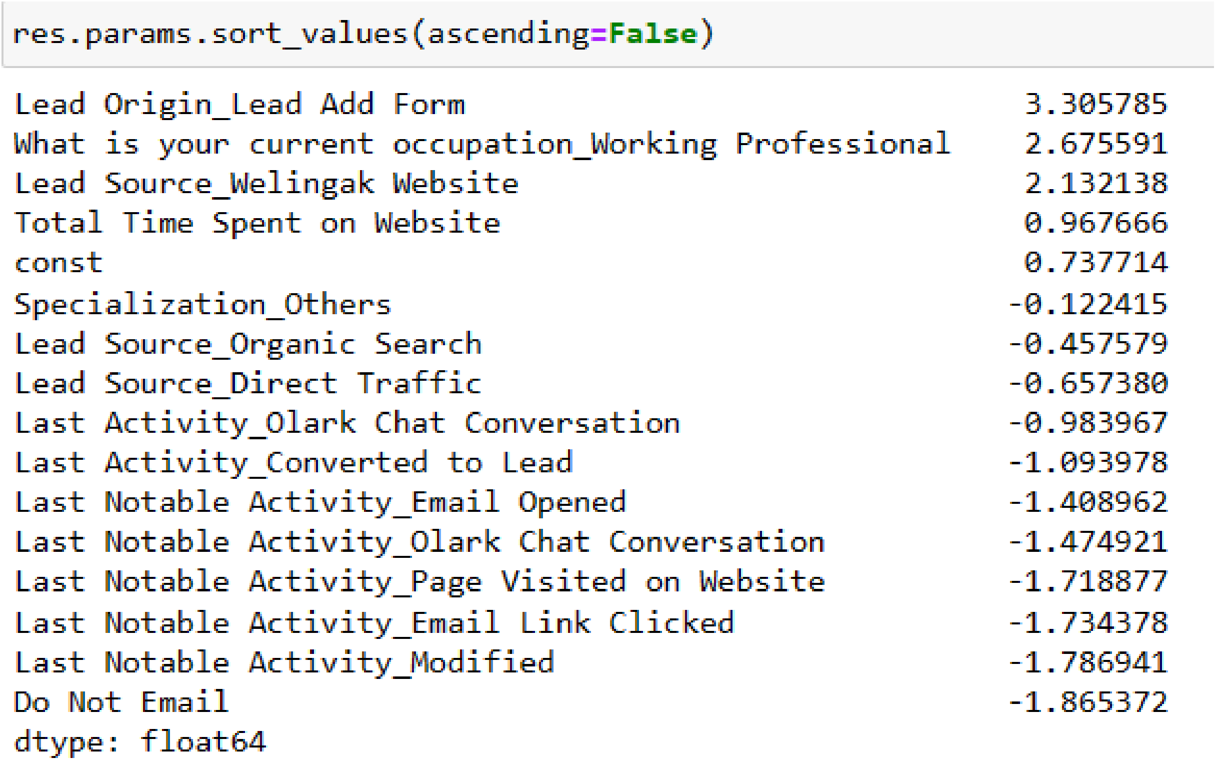
Test Dataset

Confusion Matrix: [[1393, 341],
[198, 791]],

Test Data Results
Accuracy : 80.20 %
Sensitivity : 79.98 %
Specificity : 80.33 %

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%.

Important Features from final Model



```
res.params.sort_values(ascending=False)
```

```
Lead Origin_Lead Add Form                               3.305785
What is your current occupation_Working Professional    2.675591
Lead Source_Welingak Website                            2.132138
Total Time Spent on Website                             0.967666
const                                                   0.737714
Specialization_Others                                  -0.122415
Lead Source_Organic Search                             -0.457579
Lead Source_Direct Traffic                             -0.657380
Last Activity_Olark Chat Conversation                  -0.983967
Last Activity_Converted to Lead                        -1.093978
Last Notable Activity_Email Opened                     -1.408962
Last Notable Activity_Olark Chat Conversation          -1.474921
Last Notable Activity_Page Visited on Website          -1.718877
Last Notable Activity_Email Link Clicked               -1.734378
Last Notable Activity_Modified                         -1.786941
Do Not Email                                           -1.865372
dtype: float64
```

Top 5 important features:

1. Lead Origin_Lead Add Form
2. What is your current occupation_Working Professional
3. Lead Source_Welingak Website
4. Total Time Spent on Website