Lead Scoring Assignment Case Study

Problem Statement

X Education is a company that offers online courses to industry professionals. The company promotes its courses on a variety of popular websites, including Google. X Education wishes to identify the most promising leads, which can be converted into paying customers.

Despite the fact that the company generates a large number of leads, only a small number of them are converted into paying customers, and the company desires a higher lead conversion. Leads arrive via a variety of channels, including email, website advertisements, Google searches, and so on. The company had a 30% conversion rate throughout the entire process of converting leads into customers by approaching those leads who expressed interest in taking the course. The implementation process of lead generation attributes is inefficient in terms of assisting conversions.

Goal

The company requires the development of a model for selecting the most promising leads. Each lead should be assigned a lead score that indicates how promising the lead is.

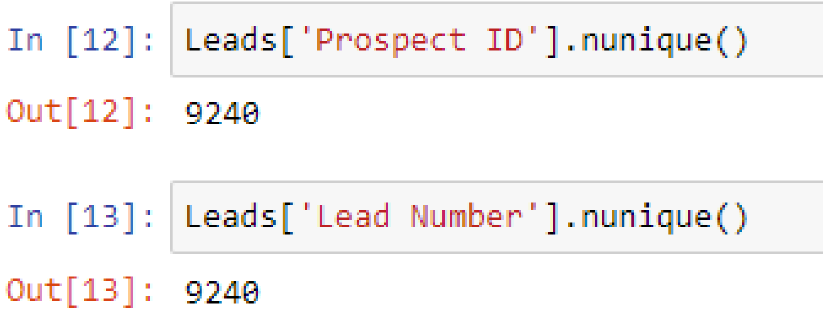
The higher the lead score, the more likely the lead is to be converted; the lower the lead score, the lower the chances of conversion. The model should have a lead conversion rate of 80% or higher.

Overall Approach

1. Import Data
2. Clean and prepare the acquired data for further analysis
3. Exploratory data analysis for figuring out most helpful attributes for conversion
4. Scaling features
5. Prepare the data for model building
6. Build a logistic regression model
7. Assign a lead score for each leads
8. Test the model on train set
9. Evaluate model by different measures and metrics Test the model on test set
10. Measure the accuracy of the model and other metrics for evaluation

Relevant Results

1. We have chosen Lead number as our unique id from Prospect ID and Lead Number.



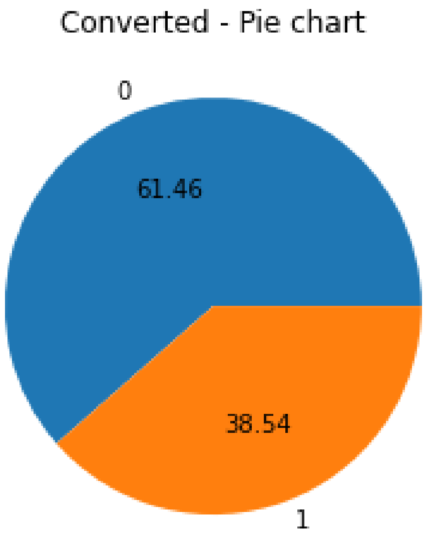
```
In [12]: Leads['Prospect ID'].nunique()

Out[12]: 9240

In [13]: Leads['Lead Number'].nunique()

Out[13]: 9240
```

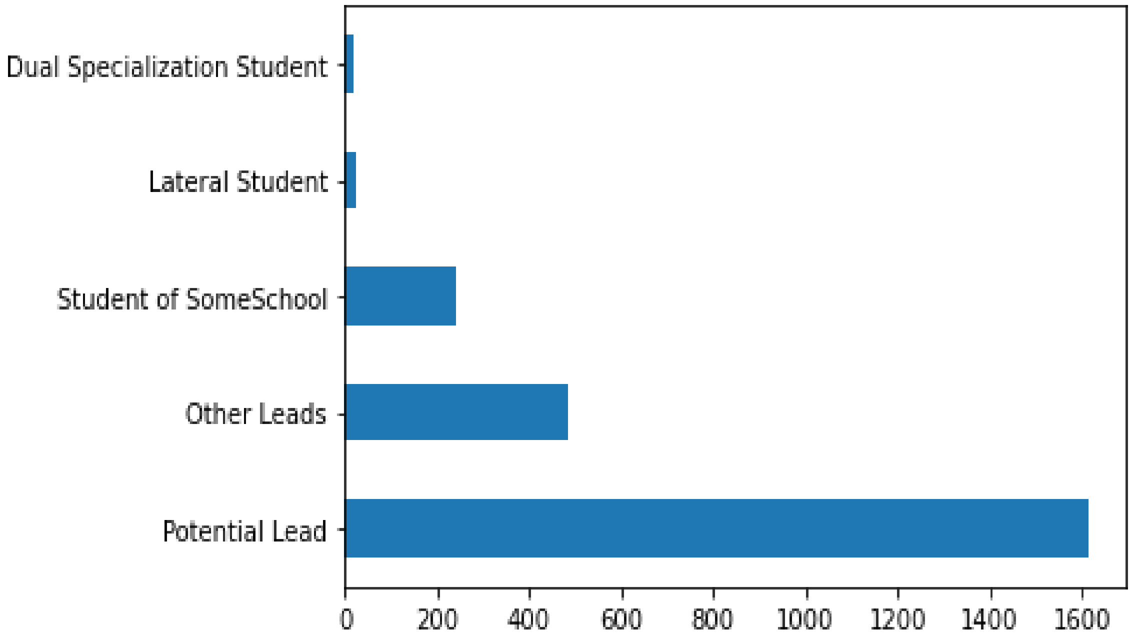
2. Doing EDA to check the values and presentation of some of the columns.
3. Data Imbalance



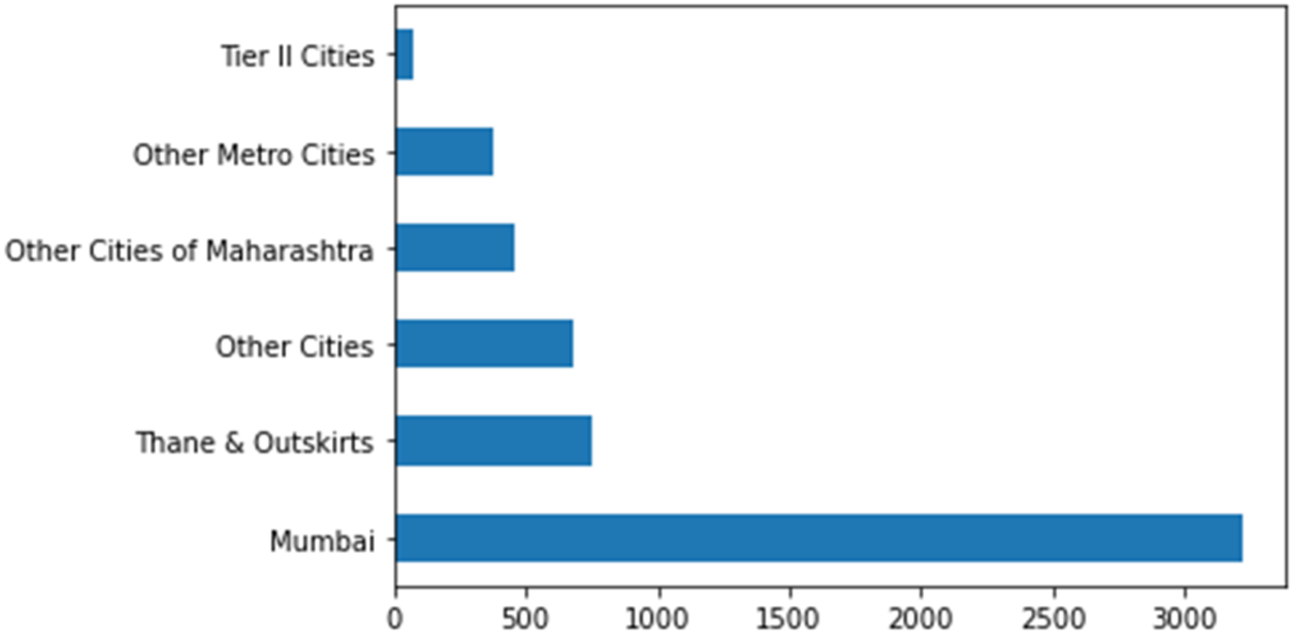
Exploratory Data Analysis

Univariate Analysis

Potential Leads are much higher as compared to other elements.

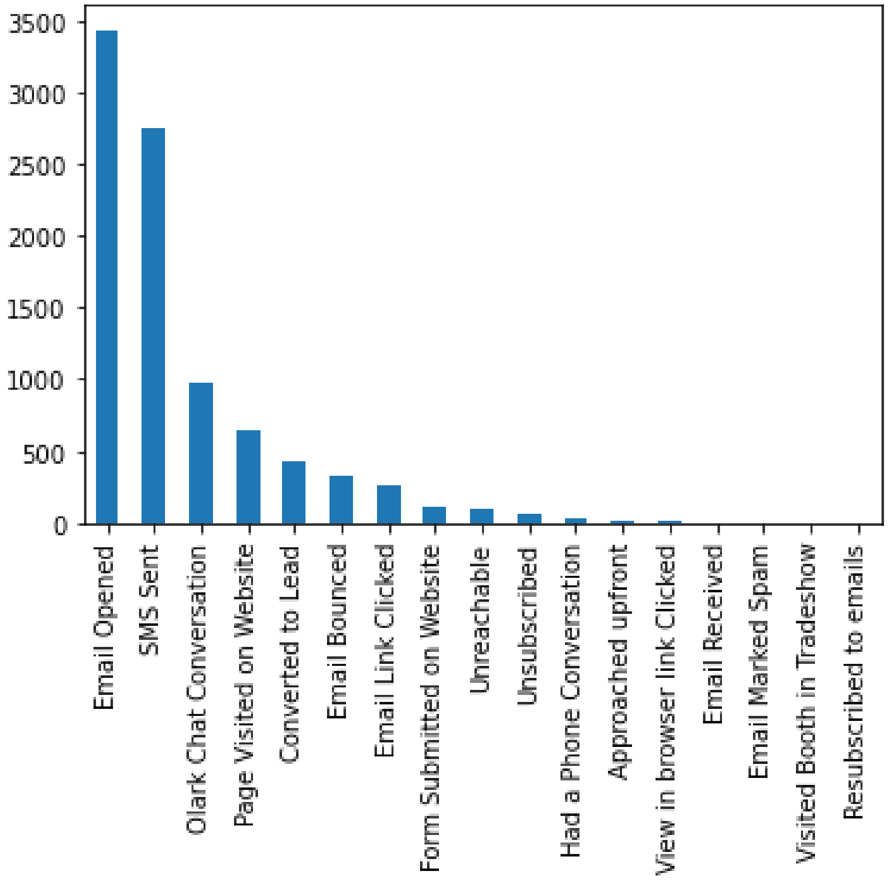


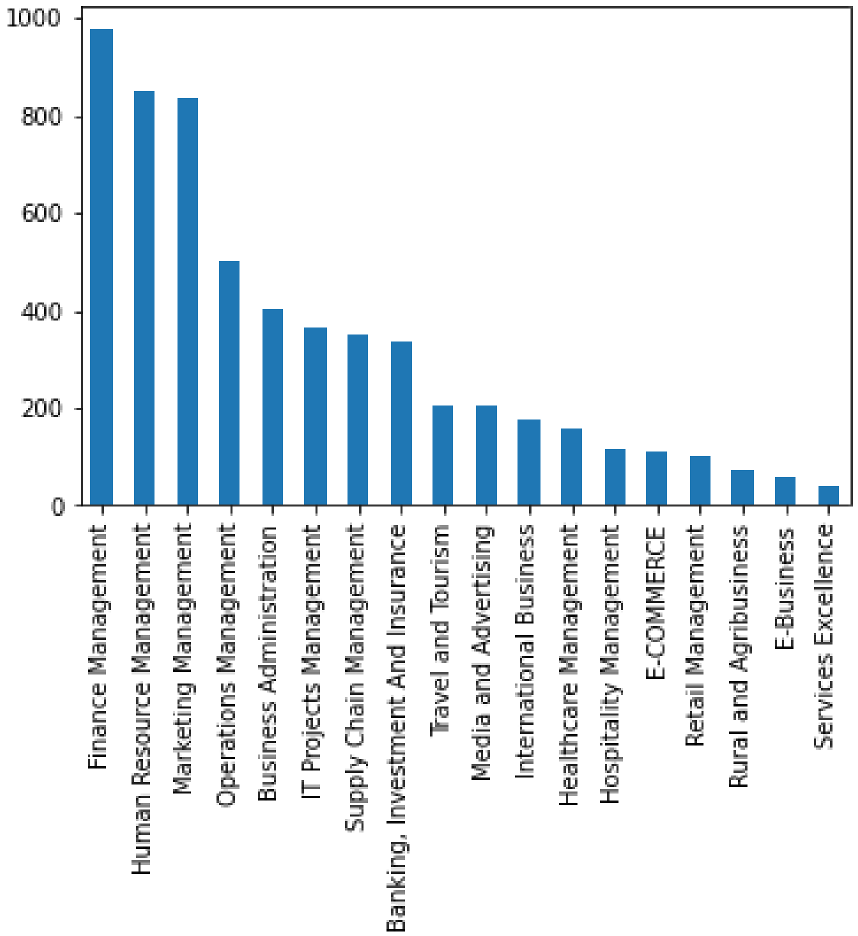
Mumbai city is having maximum no of Leads.



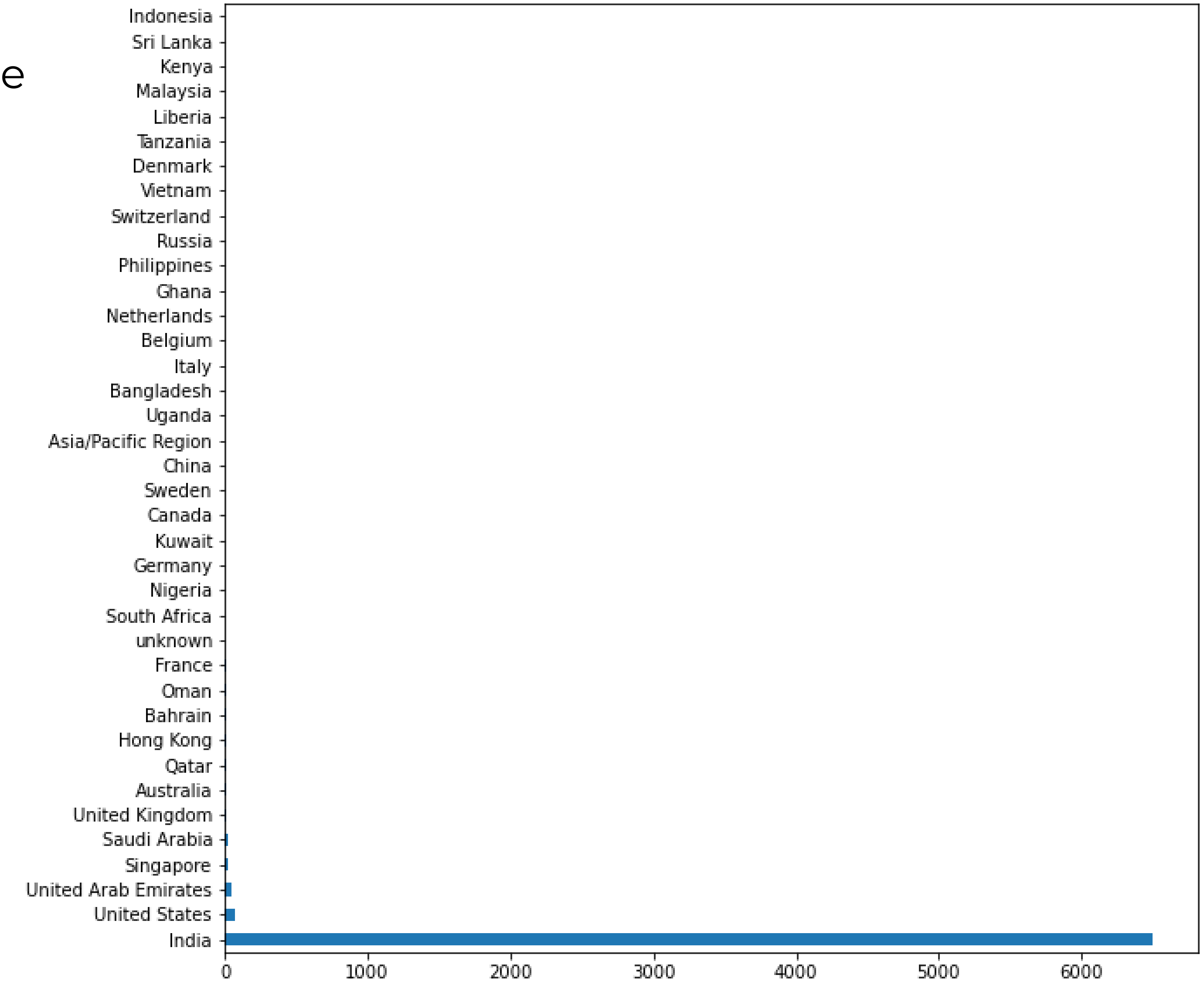
Last Activity of users is Email Opened.

Finance Management is the specialisation which has maximum count.



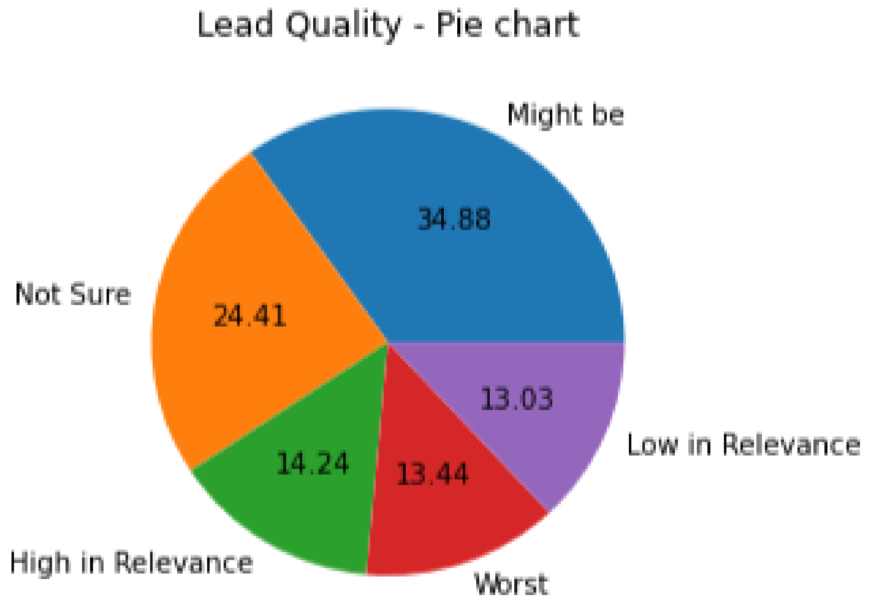


Major data is from India only. So, we have dropped this column.



Lead Quality Pie Chart

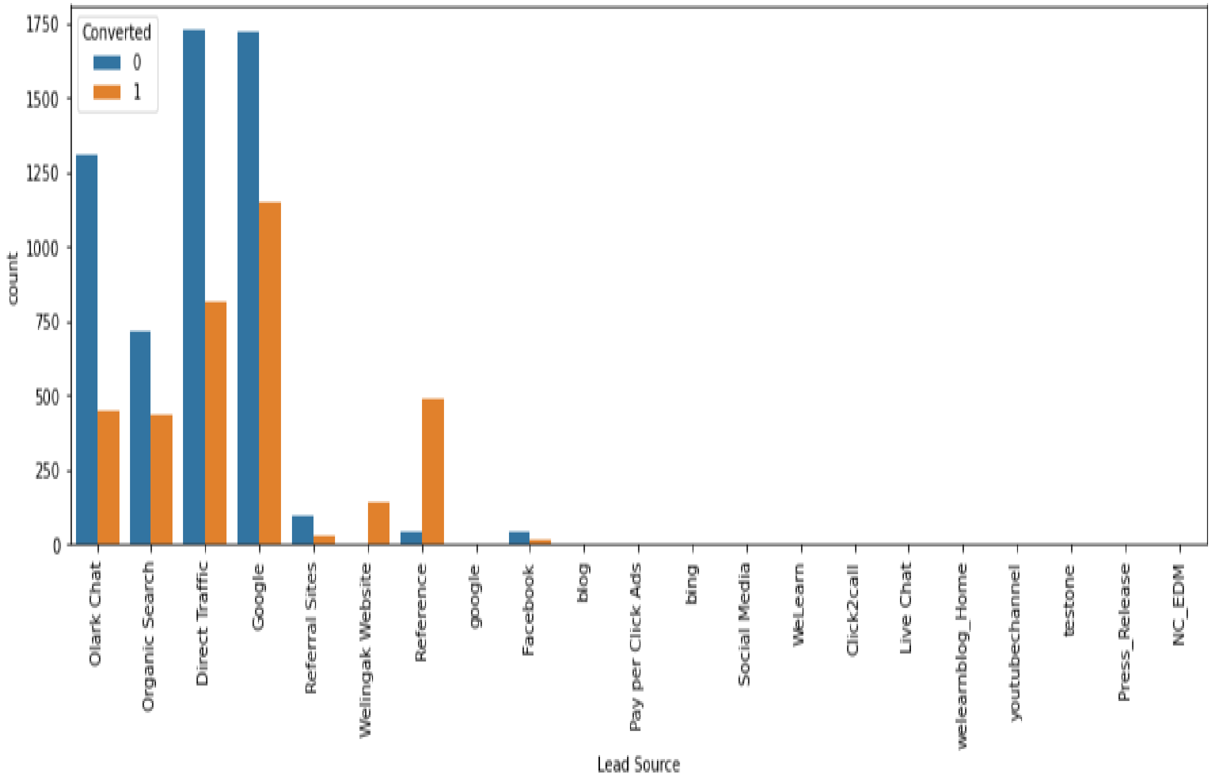
People who might want the course and not sure about it high in numbers. There is great scope to work on them in order to convert them.



- # Bivariate Analysis

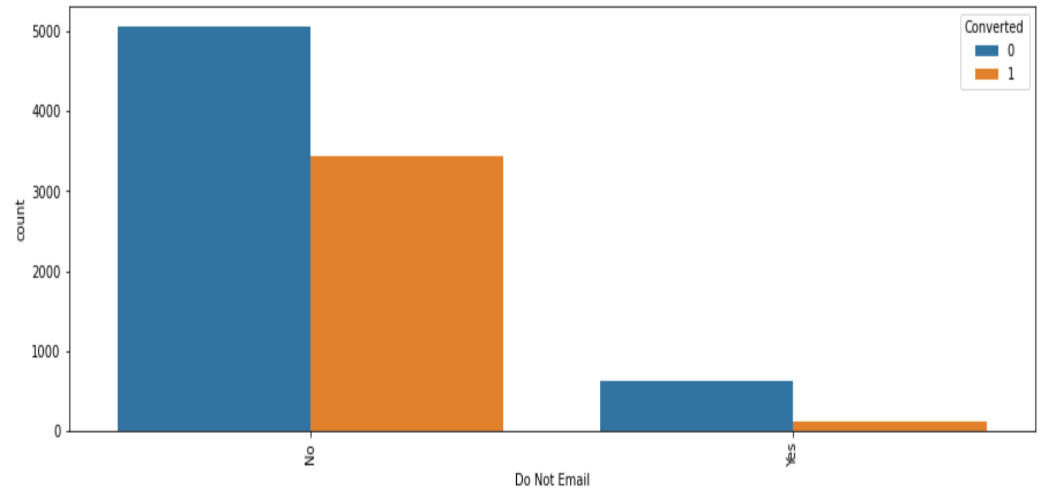
Lead Source Vs Converted

Direct traffic and Google searches had high conversion while references has high conversion rate.



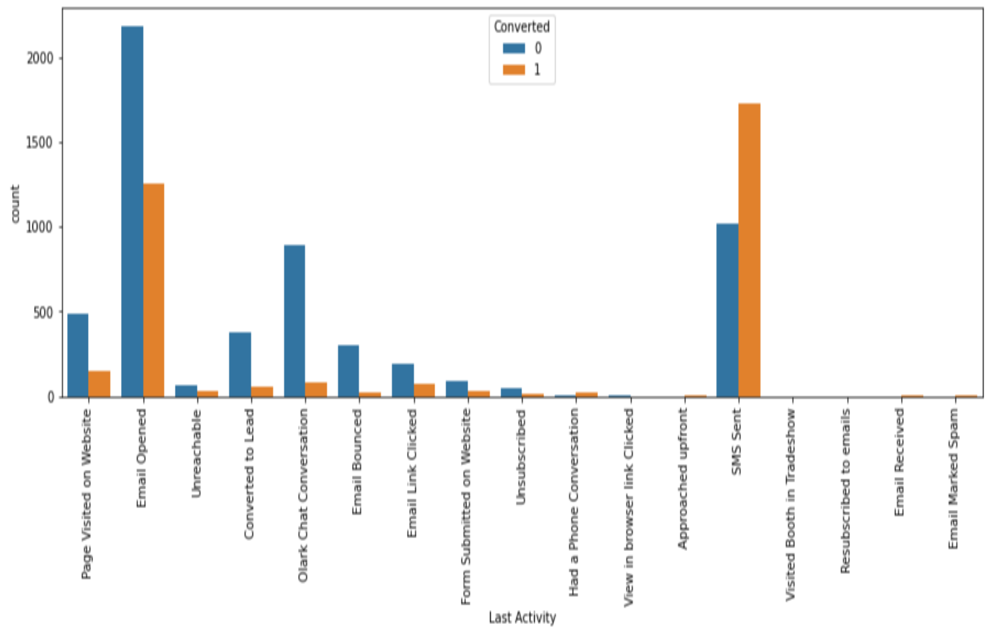
Do not Email vs Converted

Most leads prefer to get updates via Email.



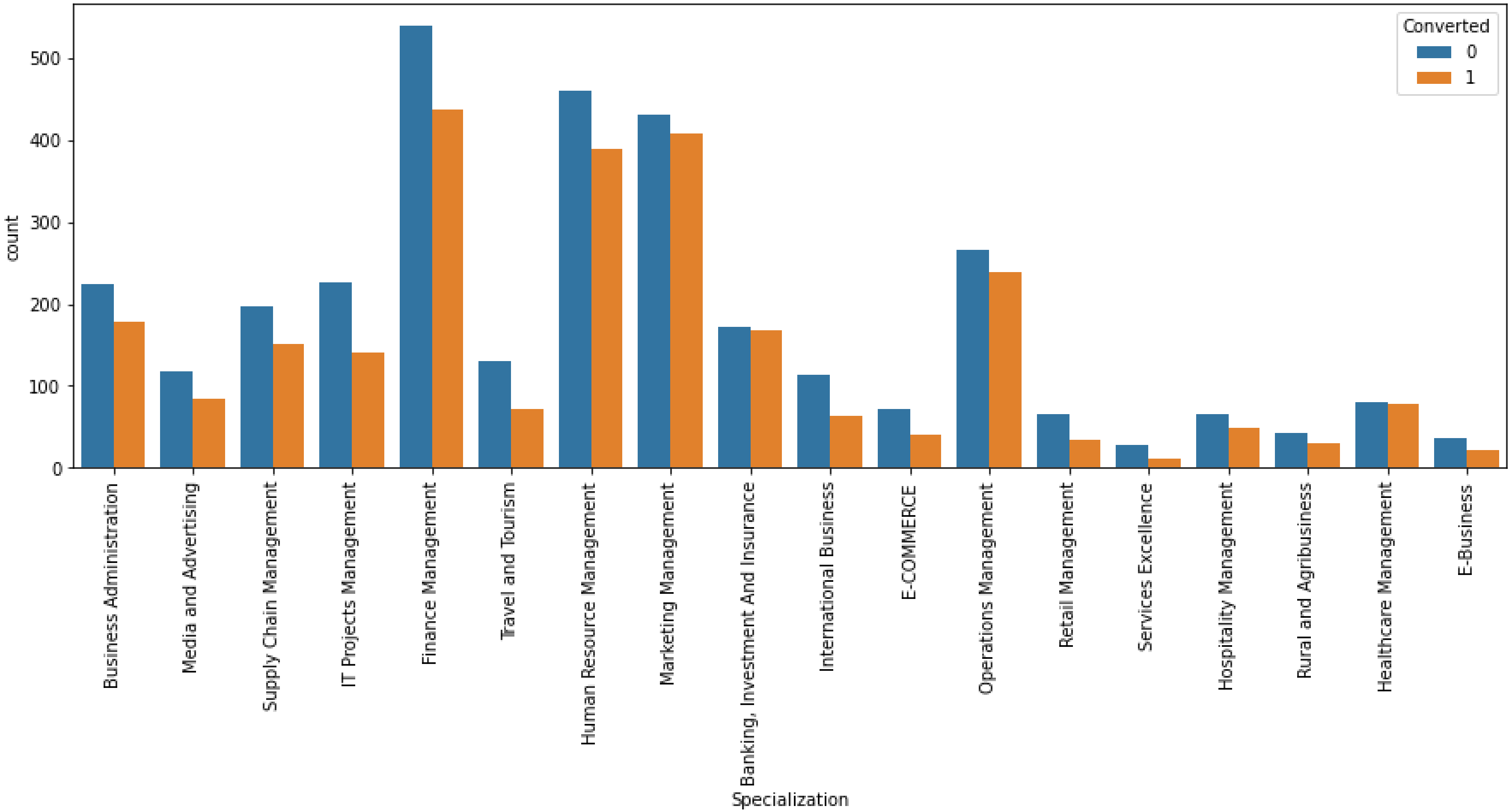
Last Activity Vs Converted

SMS shows higher conversion rate whole Email and chat shows high Conversions



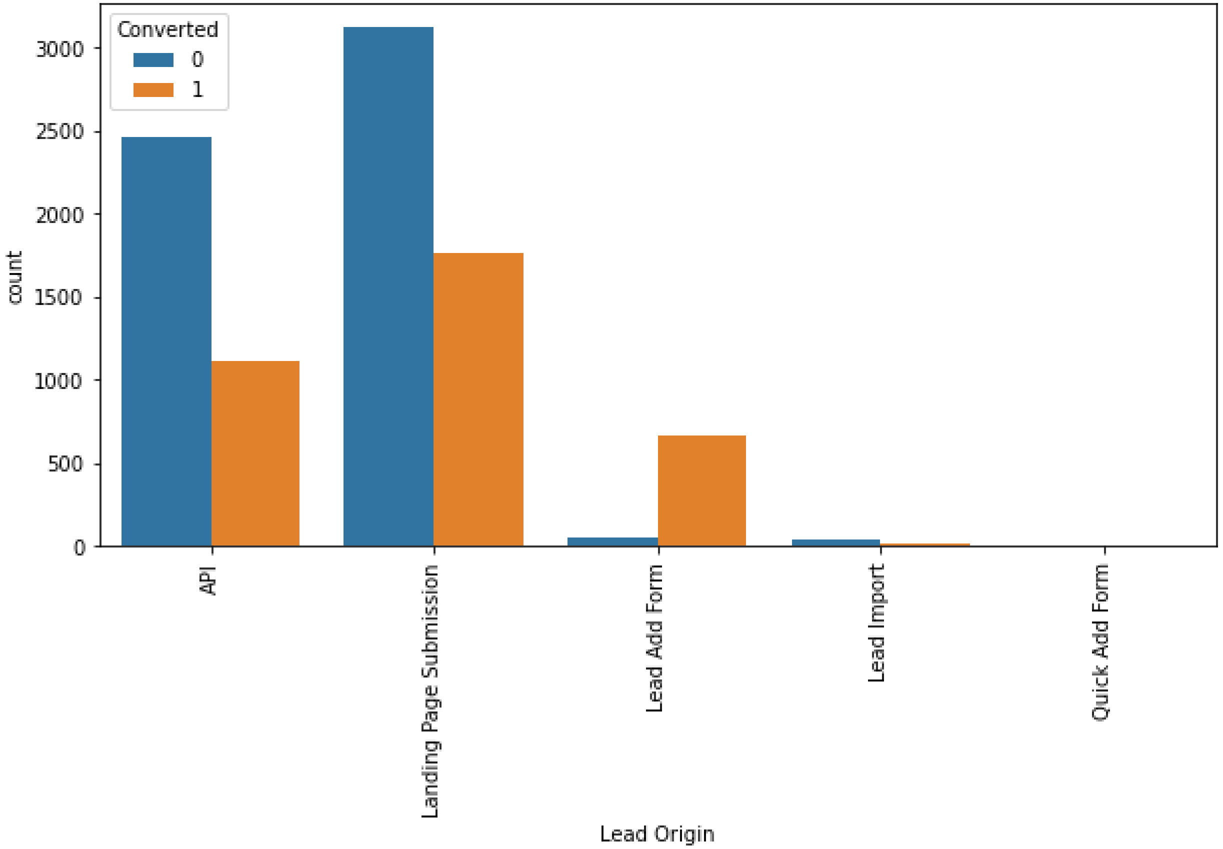
Specializations Vs Converted

Most people don’t know about specialization while people from Finance, HRD & Marketing can be promising leads.



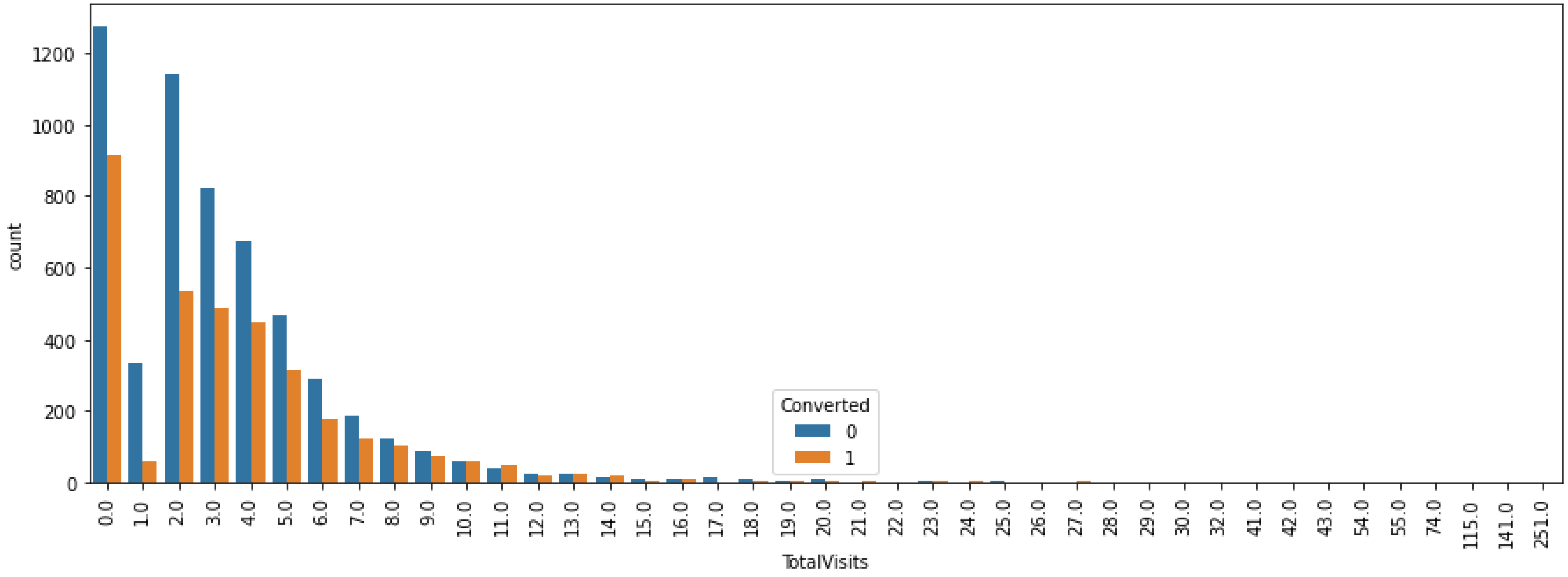
Lead Origin Vs Converted

Landing page shows high conversions



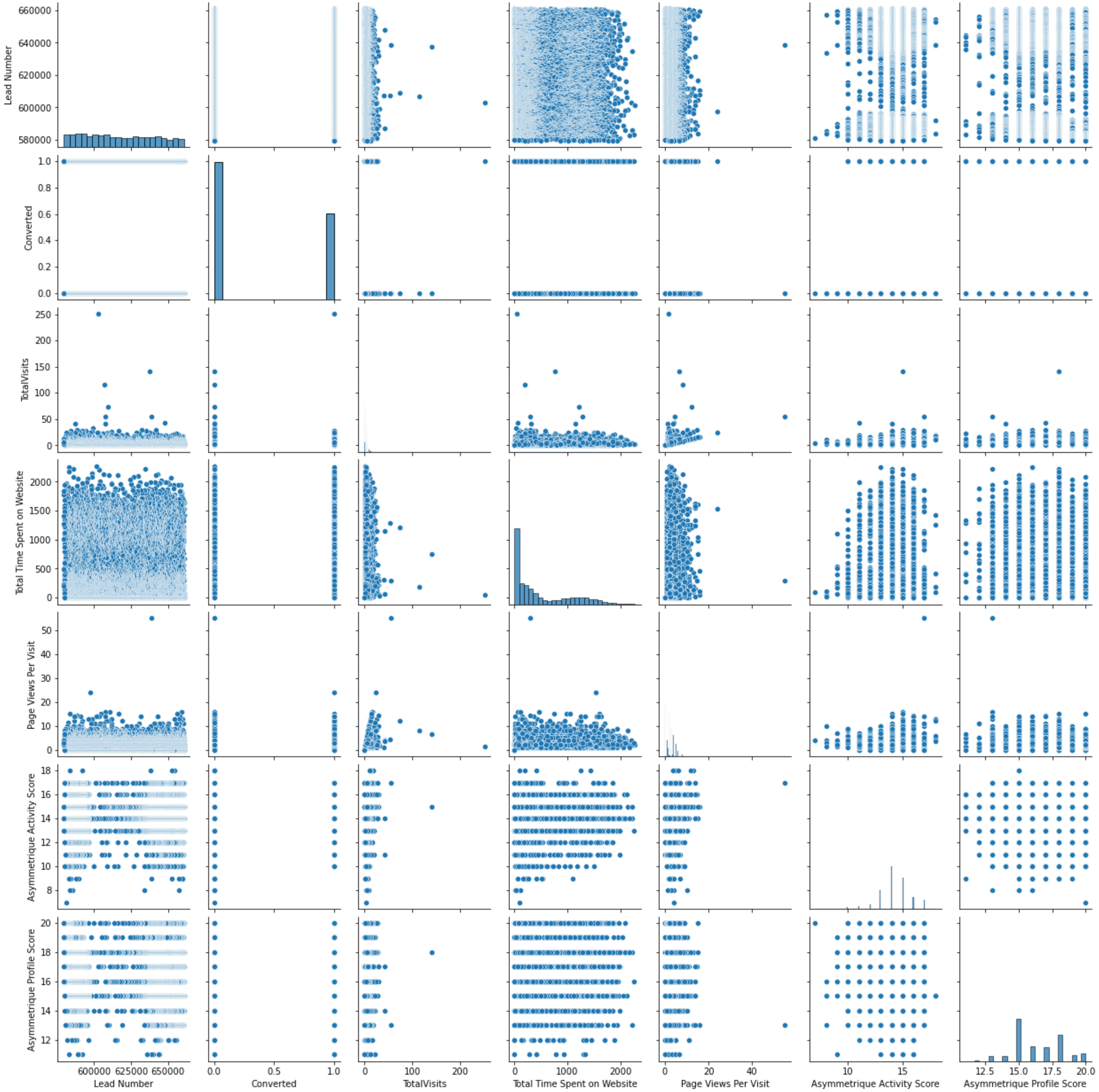
Total Visits Vs Converted

Higher the total visits higher the chances of being a hot lead or potential lead.



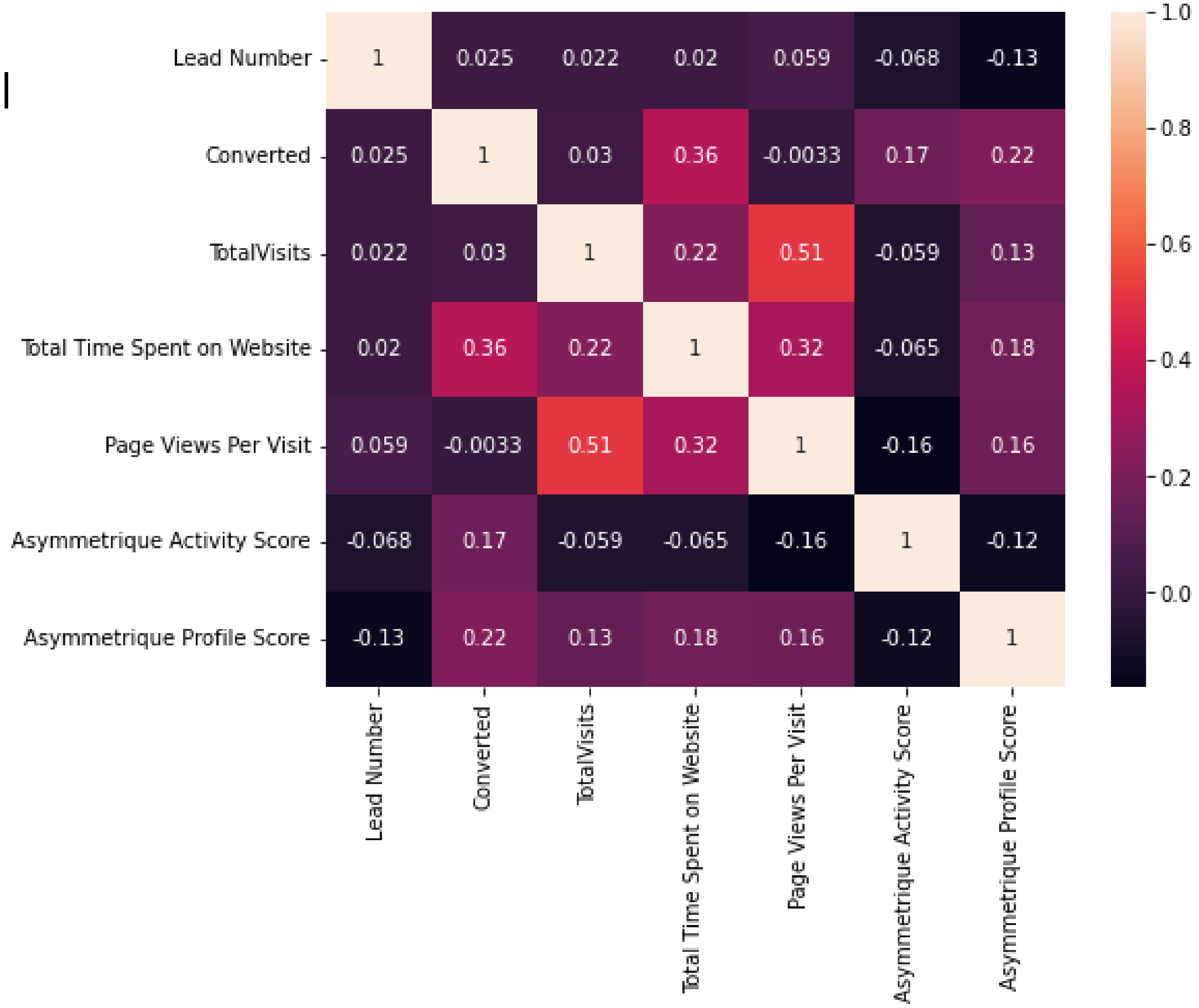
Pairplot

Plot between various numerical variables



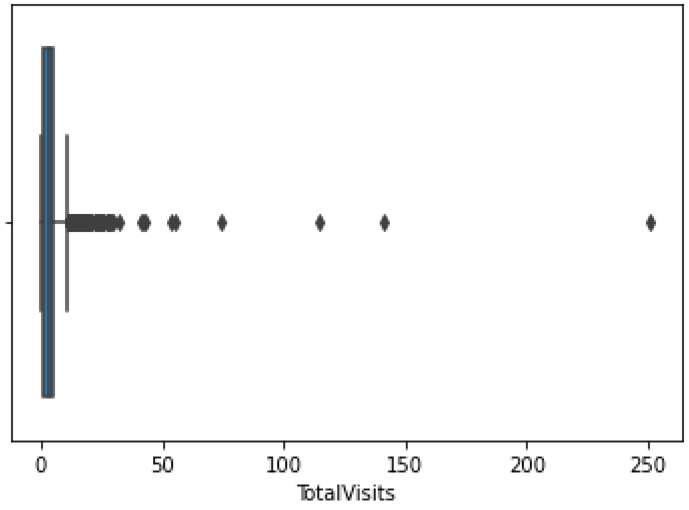
Heatmap

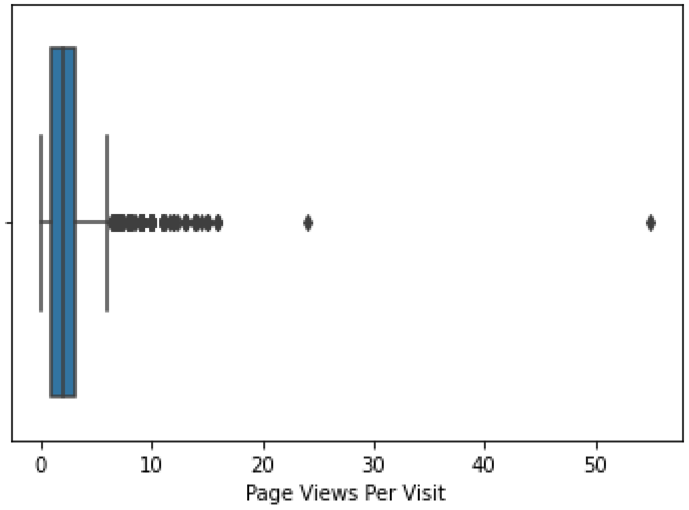
Heatmap between various numerical variables

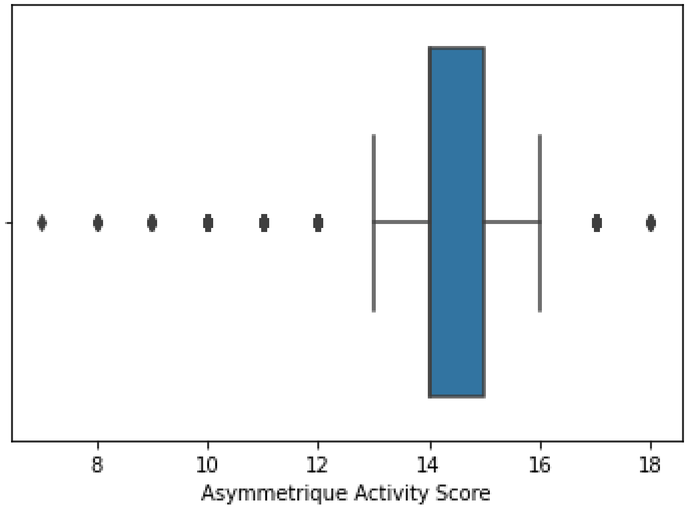


Outlier Analysis

Outliers present in Totalvisits, Page views per visit and Asymmetric Activity Score.

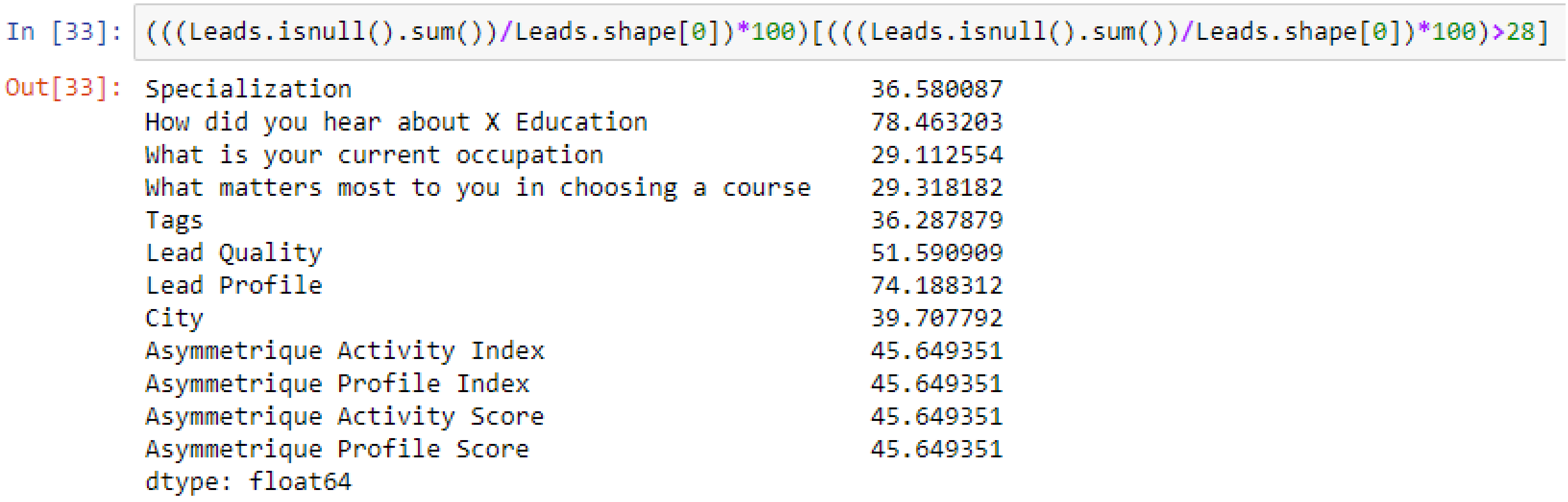






Null Value Analysis

Deleting columns that has more than 28% null values.



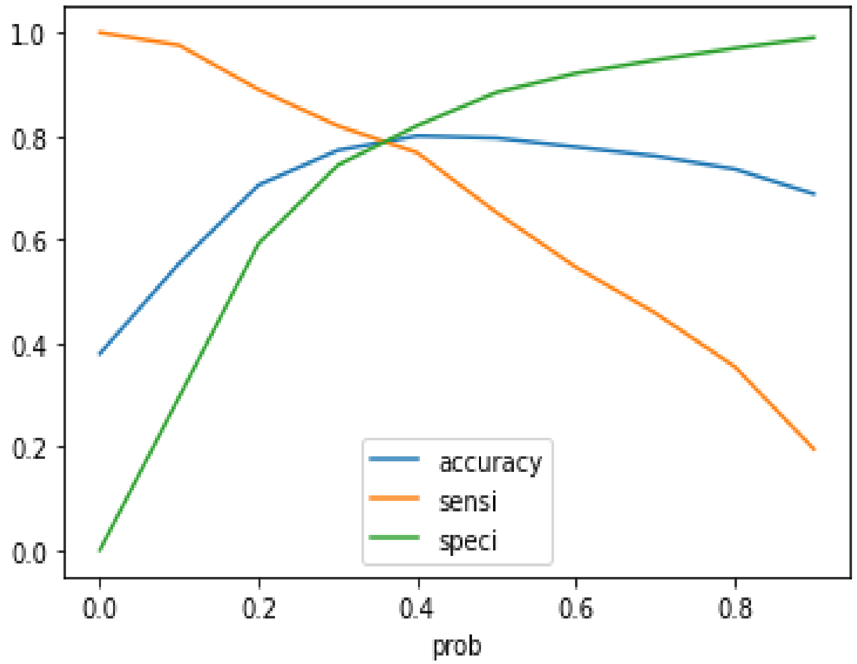
```
In [33]: (((Leads.isnull().sum())/Leads.shape[0])*100)[(((Leads.isnull().sum())/Leads.shape[0])*100)>28]

Out[33]: Specialization                                   36.580087
         How did you hear about X Education               78.463203
         What is your current occupation                  29.112554
         What matters most to you in choosing a course    29.318182
         Tags                                             36.287879
         Lead Quality                                     51.590909
         Lead Profile                                     74.188312
         City                                             39.707792
         Asymmetrique Activity Index                      45.649351
         Asymmetrique Profile Index                       45.649351
         Asymmetrique Activity Score                      45.649351
         Asymmetrique Profile Score                       45.649351
         dtype: float64
```

Model Building

- Splitting into train and test set Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables Build the next model
- Eliminate variables based on high p-values
- Check VIF value for all the existing columns
- Predict using train set
- Evaluate accuracy and other metric Predict using test set
- Precision and recall analysis on test predictions

Calculation of Optimal Cutoff



Model evaluation on (Train vs Test)

- Accuracy – 78.7
- Sensitivity – 77.6
- Specificity – 81.3

- Precision – 83
- Recall - 51

- Accuracy – 80
- Sensitivity – 79
- Specificity – 81

Conclusion

EDA

- Potential Leads are much higher as compared to other elements.
- Direct traffic and Google searches had high conversion while references has high conversion rate.
- Most leads prefer to get updates via Email.
- SMS shows higher conversion rate whole Email and chat shows high Conversions
- Higher the total visits higher the chances of being a hot lead or potential lead.
- People who might want the course and not sure about it high in numbers. There is great scope to work on them in order to convert them

Conclusion

Logistic Regression Model

- The model shows accuracy high close to 80%
- The threshold has been selected from Accuracy, Sensitivity, specificity measures and precision, recall curves.
- The model shows 79% sensitivity and 81% specificity
- The model finds correct promising leads and leads that have less chances of getting converted
- Overall this model proves to be accurate