LEAD SCORING CASE STUDY

Purpose of the case study

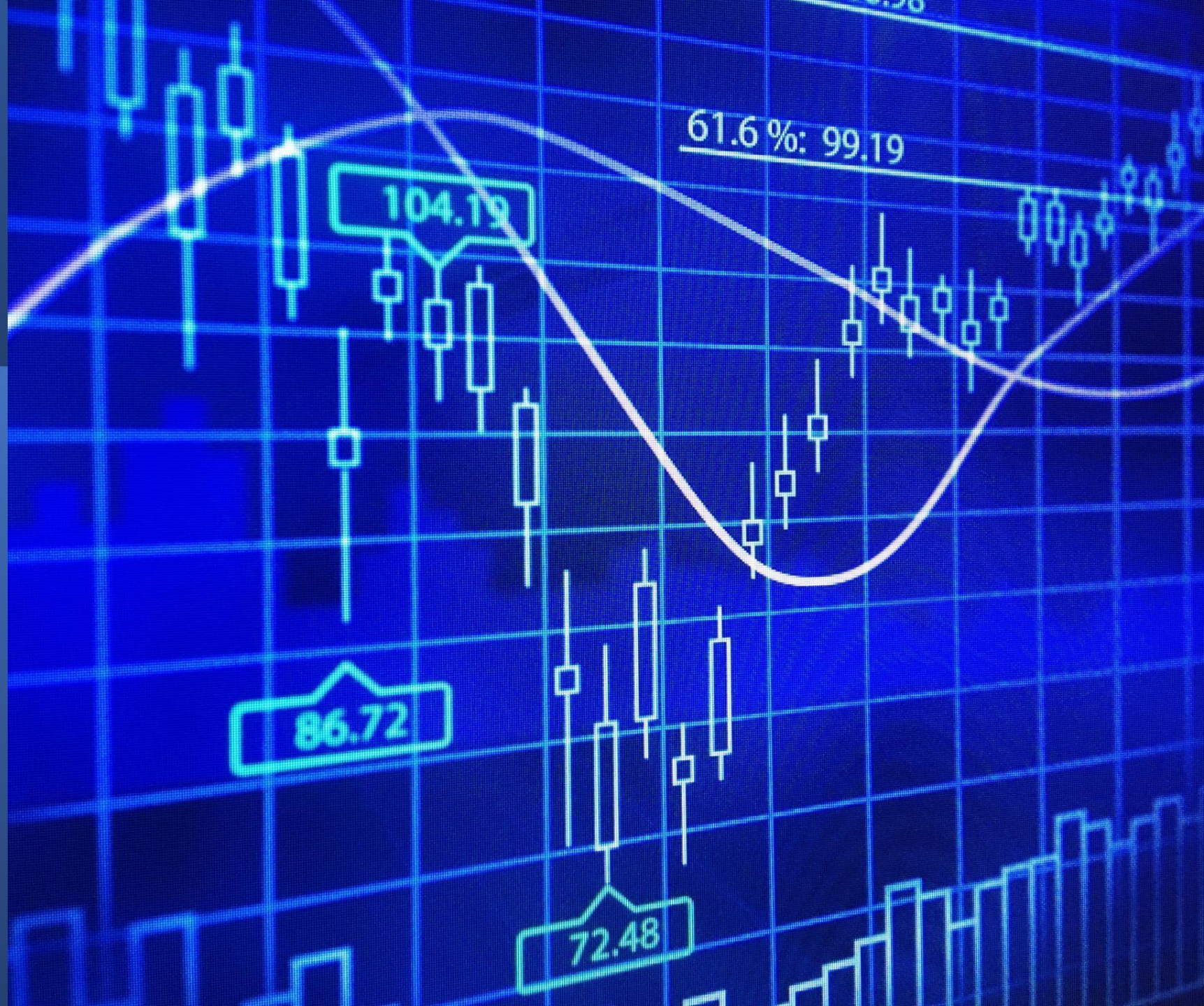
An education company named X Education sells online courses to industry professionals.

X Education wishes to identify the most potential leads, also known as ‘Hot Leads’ so that the sales team can focus more on communicating with the potential leads rather than making calls to everyone.

This Machine Learning model is to help identify the potential leads who are most likely to convert into paying customer based on certain factors learnt by the model.

This model also helps assign 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 can be identified easily.

Data Handling



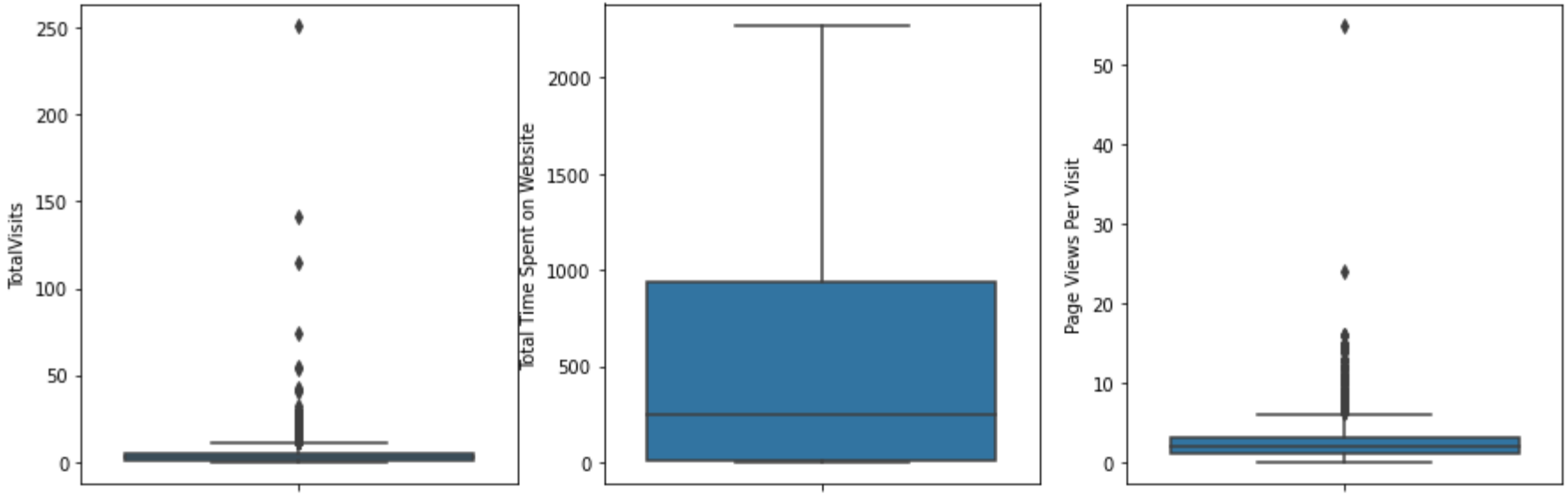
Null Value Handling

- ‘Select’ values in the columns were replaced with Null values for imputation.
- Columns with more than 40% null values are dropped after ensuring there are no columns that would impact the inferences if dropped.
- Columns that were highly skewed were dropped since they won’t be helpful in drawing inferences.
- Columns created by the Sales team after the Lead was acquired were dropped so the data that was provided by the customer could be used to identify the potential Hot Leads.
- Null value handling for remaining columns is summarized below.

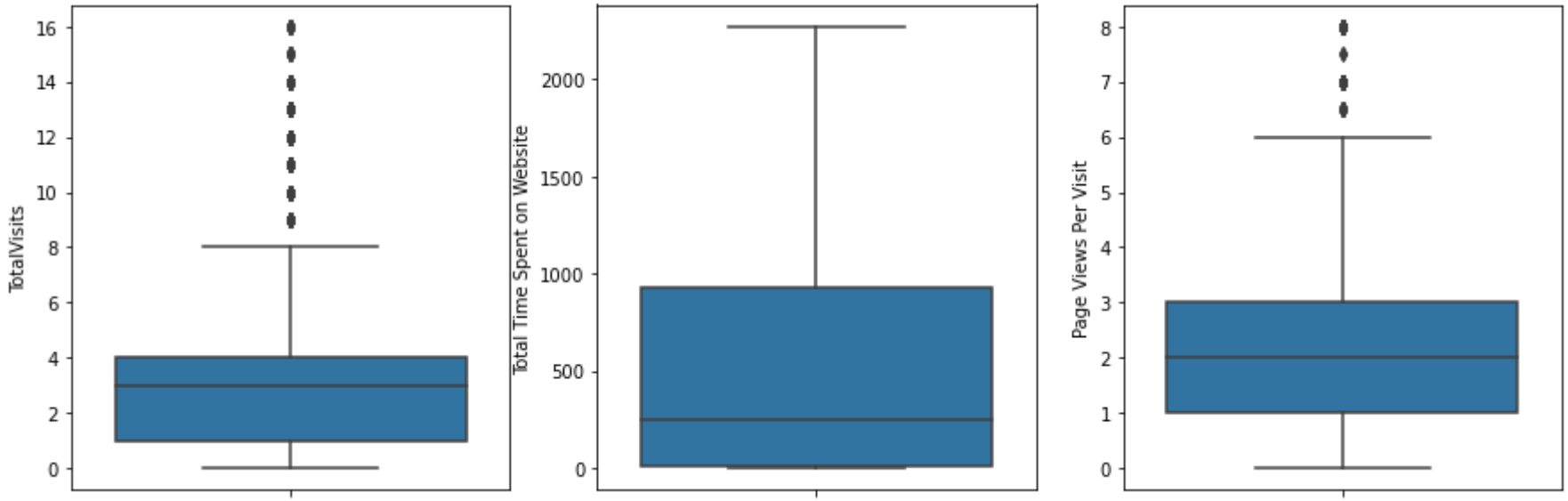
TotalVisits	1.482684 %	Impute the missing values with the median based on the skewness of the numerical variable
Page Views Per Visit	1.482684 %	Impute the missing values with the median based on the skewness of the numerical variable
Lead Source	0.389610 %	Impute the missing values with Mode since it is a categorical variable
Specialization	36.580087 %	Null values were renamed to another category called ‘Not Specified’ instead of imputing the value
What is your current occupation	0.420148 %	Null values were renamed to another category called ‘Unknown’ instead of imputing the value
How did you hear about X Education	78.463203 %	Dropped as column with more than 40% Null Values
Lead Quality	51.590909 %	Dropped as column with more than 40% Null Values
Lead Profile	74.188312 %	Dropped as column with more than 40% Null Values
Asymmetrique Activity Score Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score	45.649351 %	Dropped as column with more than 40% Null Values
Country, What matters most to you in choosing a course, City	26.634199 % 29.318182 % 39.707792 %	Dropped as highly skewed variable after null value imputation
Tags Last Activity	36.287879 % 1.114719 %	Drop the columns as these pertain to the Sales Team and would not be helpful in Model Predictions.

Outlier Handling

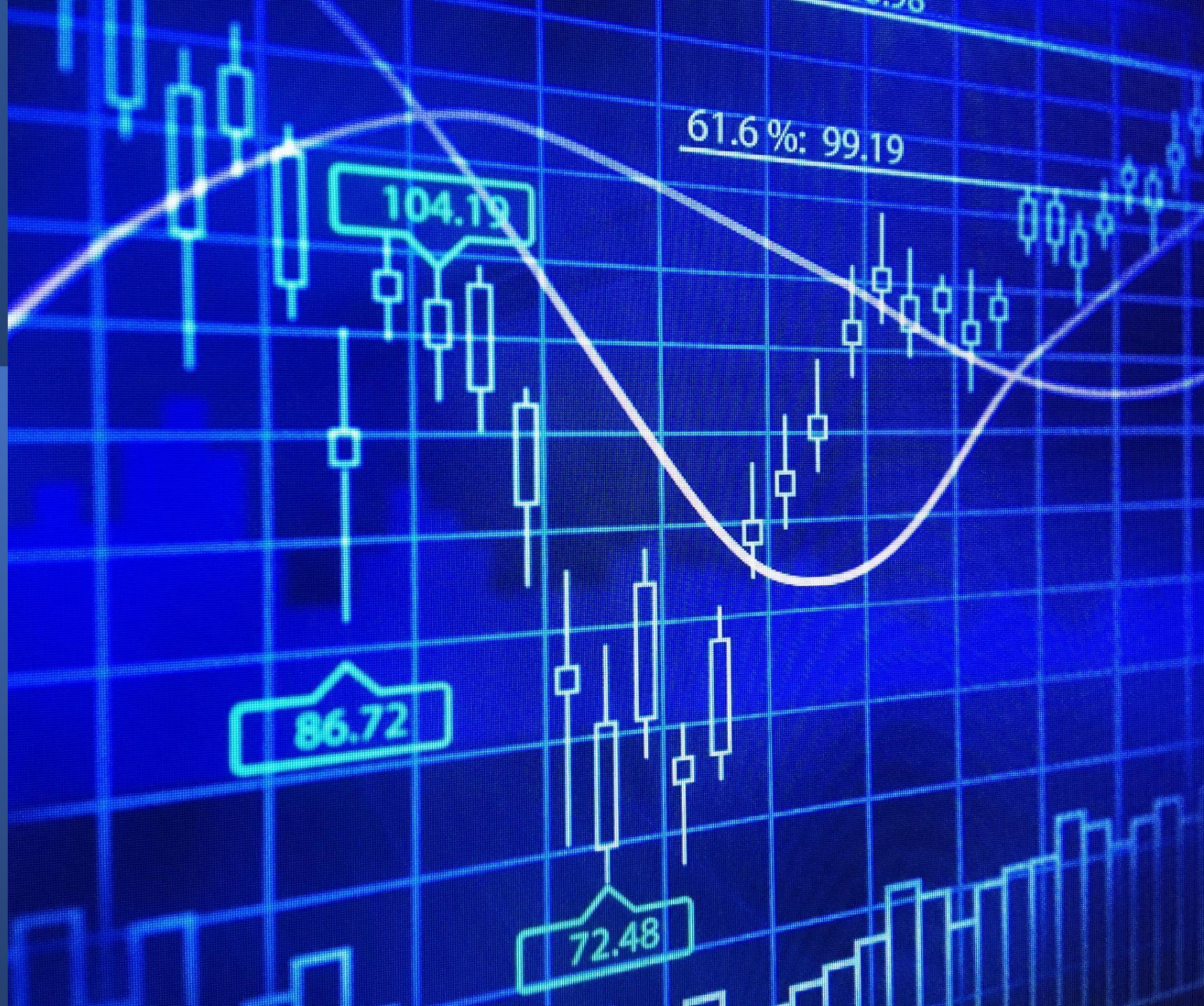
Outlier Detection (TotalVisits, Page Views Per Visit)



Outlier Handled (TotalVisits, Page Views Per Visit)

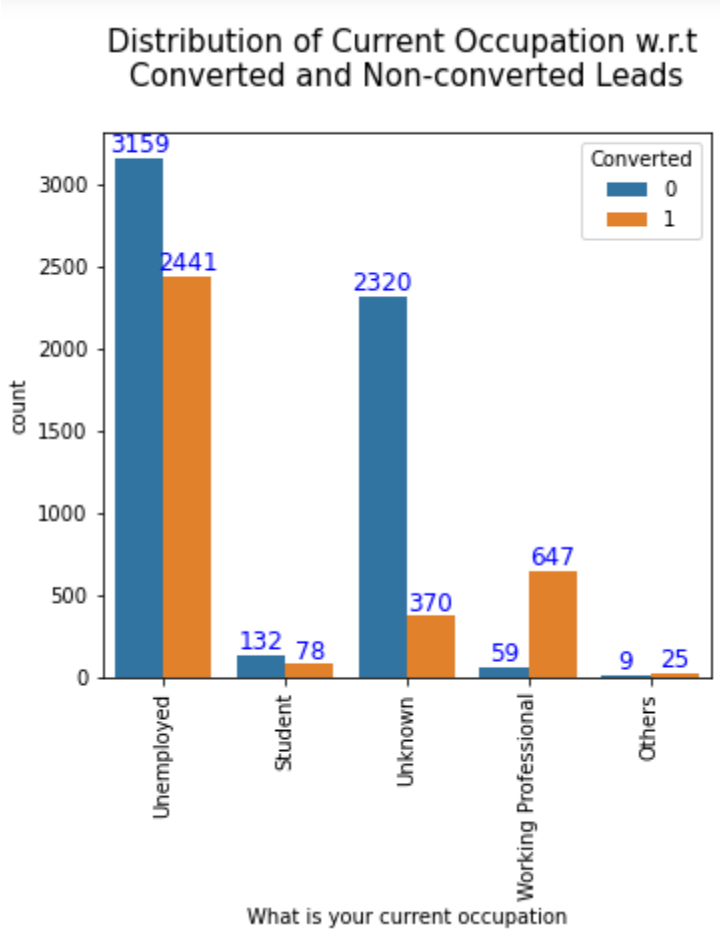


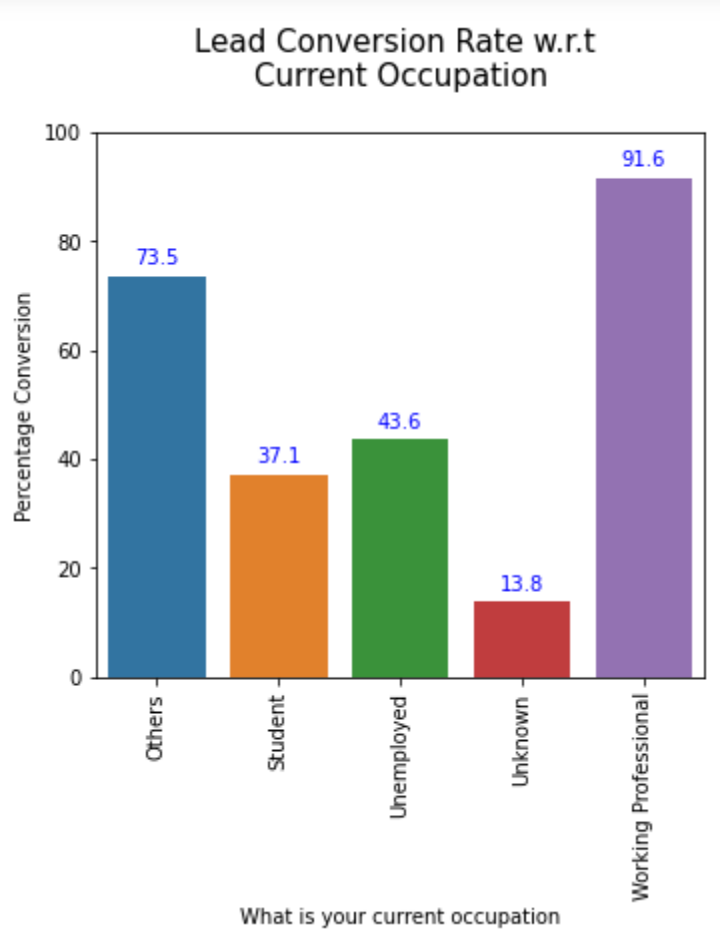
Data Analysis



***W**orking Professionals can be potential Hot Leads compared to Leads from other occupations as they demonstrate high conversion rate.*

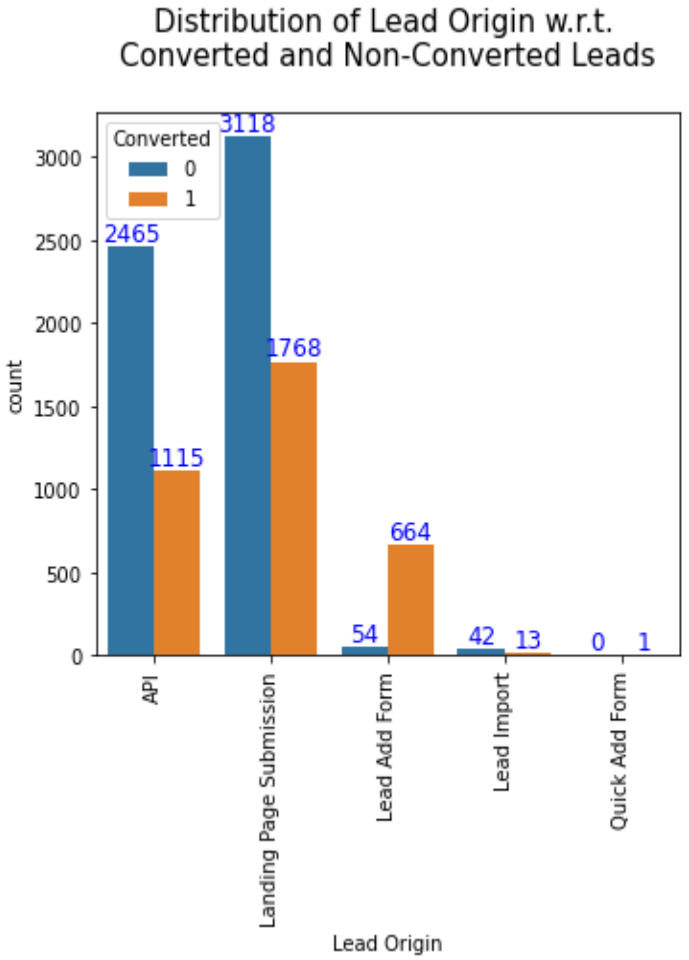
Leads who do not provide their current occupation details demonstrate least conversion rate.

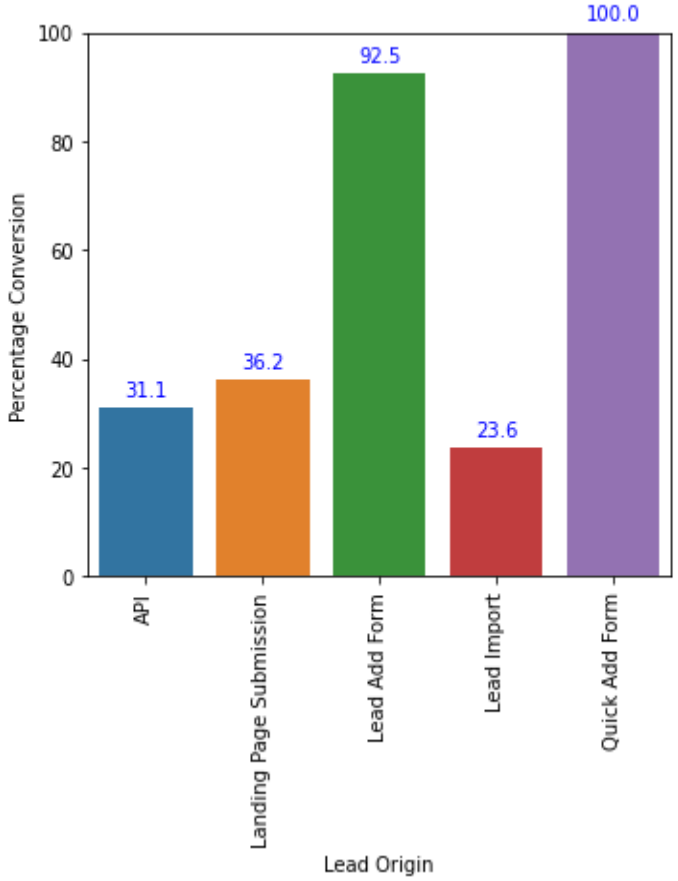




- *Leads originated from Lead Add Form demonstrate high conversion rate compared to every other Lead Origin.*

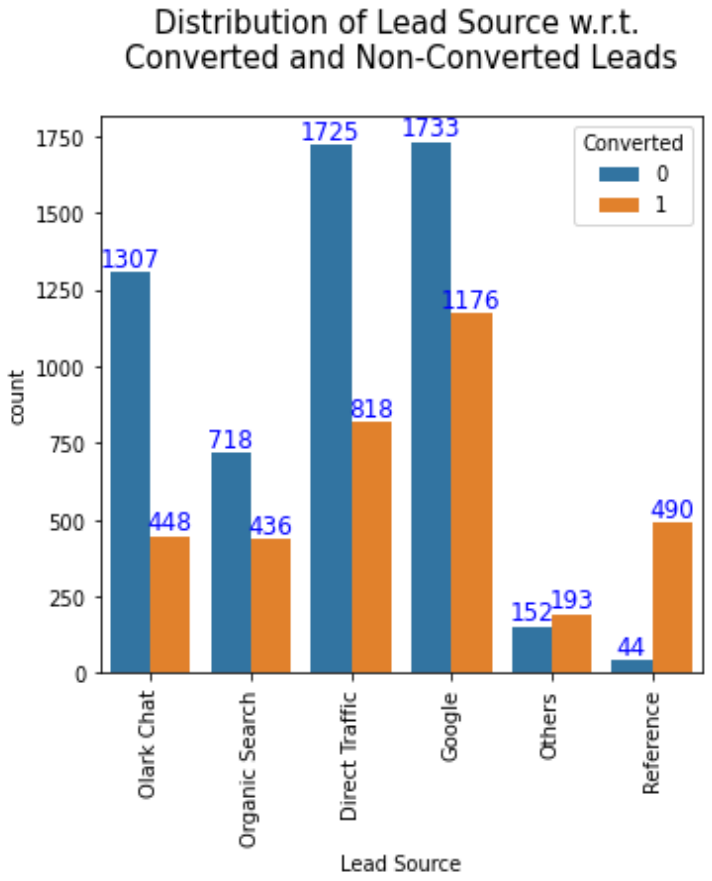
- *Lead Import has the lowest lead conversion rate.*

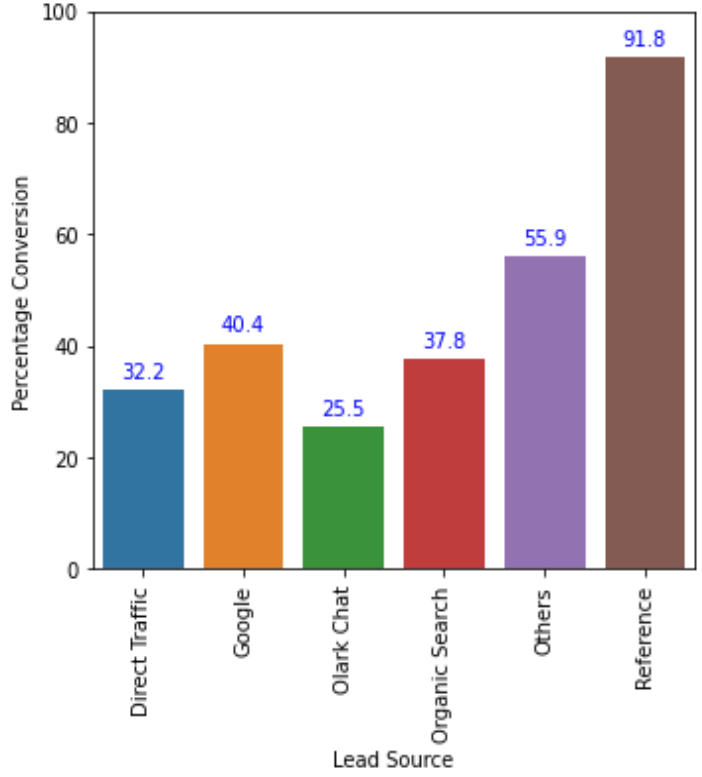




Although not the major Lead source, the conversion rate of Leads originating from Reference category looks quite promising

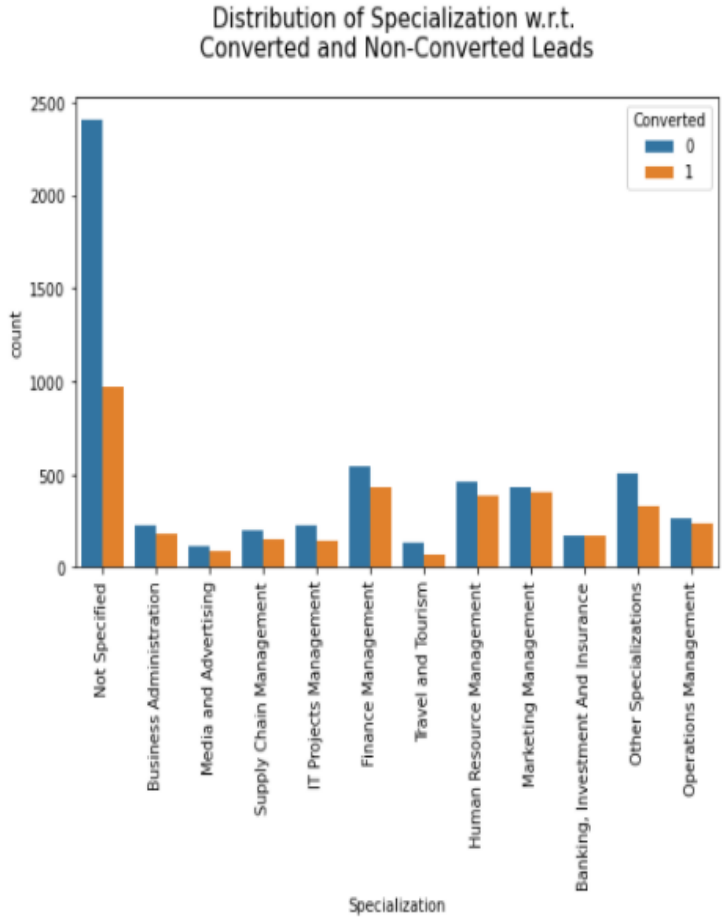
Leads originating from Olark Chat shows very minimal conversion rate compared to other sources.

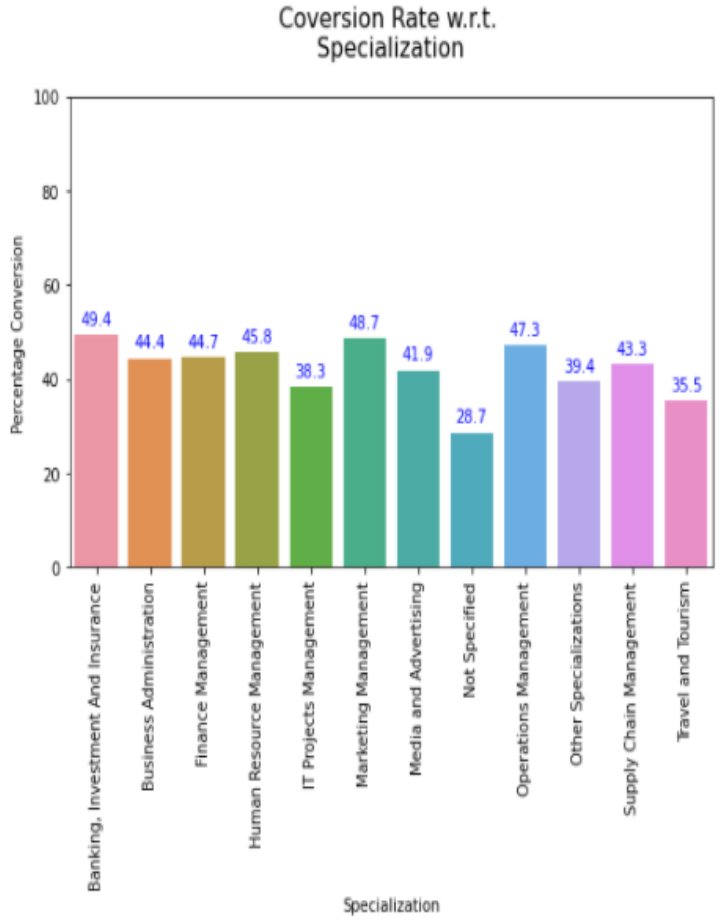




Leads who do no specify their specialization demonstrate least conversion rate.

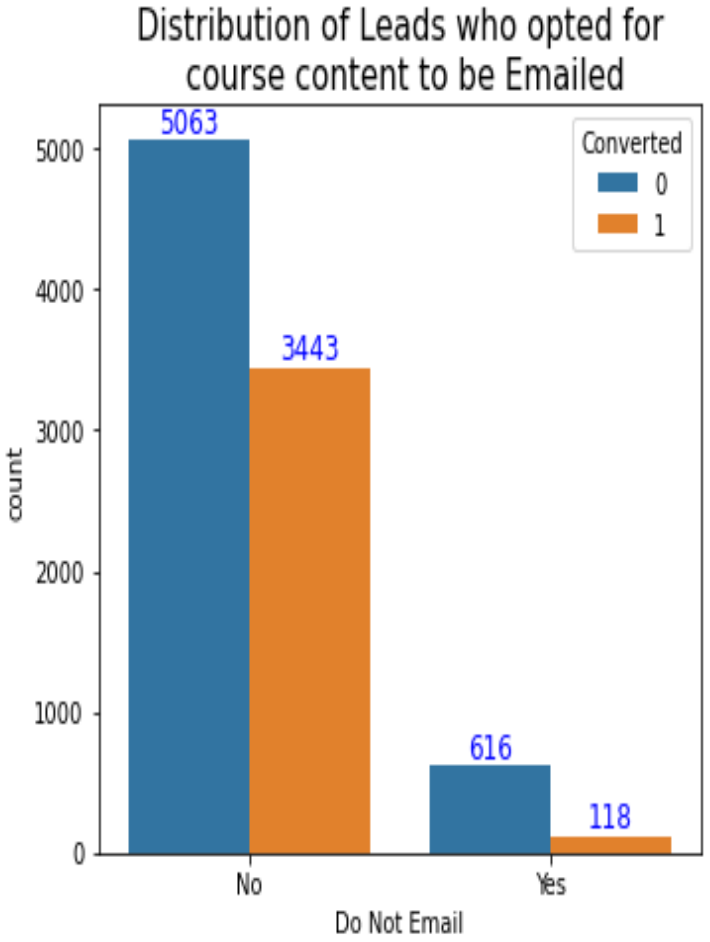
Leads opting for Specializations like 'Banking, Investment And Insurance', 'Marketing Management' & 'Operations Management' have high conversion rate in comparison to the rest.

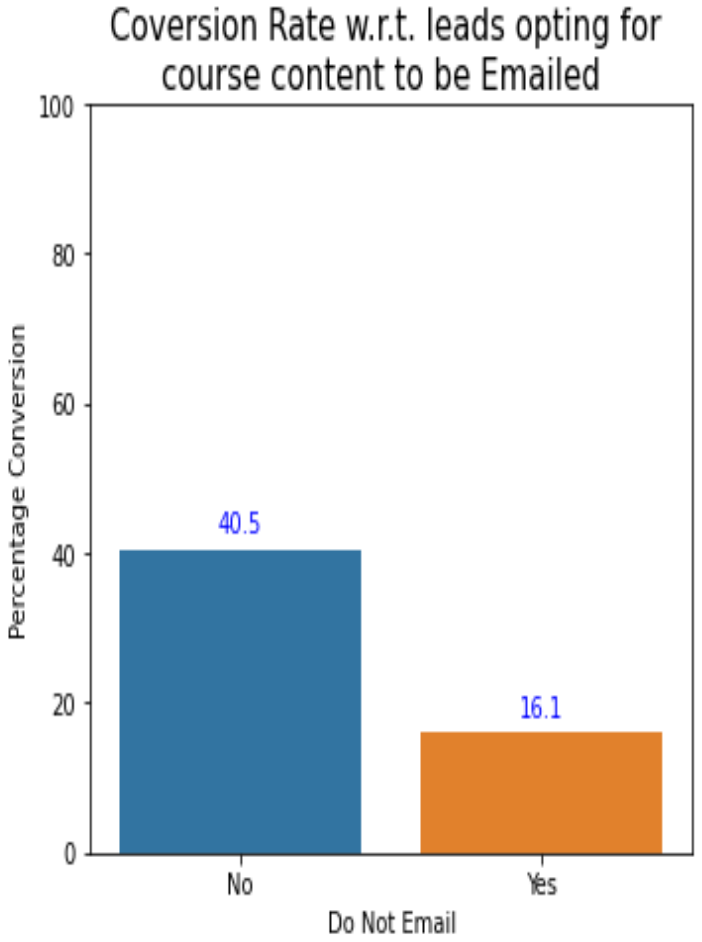




Leads who prefer to be emailed with course content can be highly focused as their conversion rate is comparatively higher than the Leads who preferred the course content not to be Emailed.

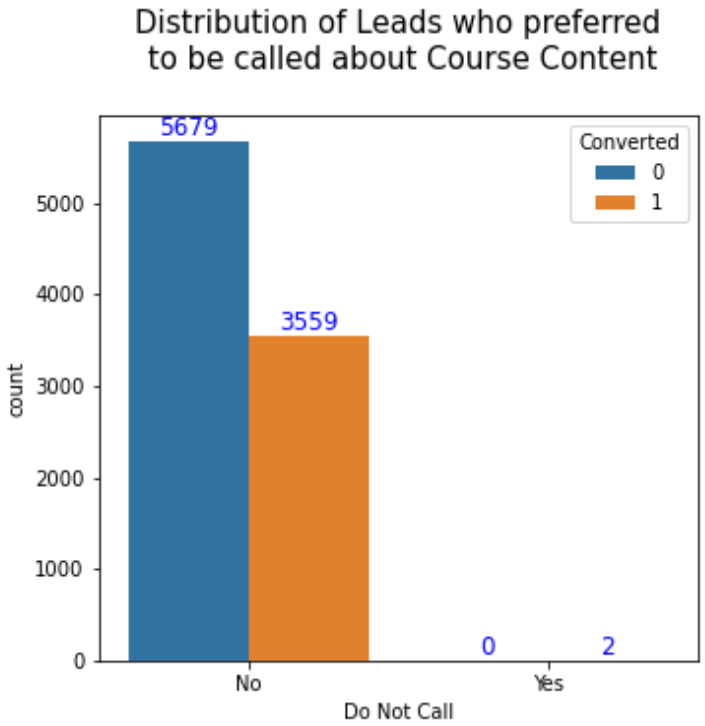
Also note that the percentage of Leads who want the course to be emailed are more compared to the percentage of people who do not want the course to be emailed.

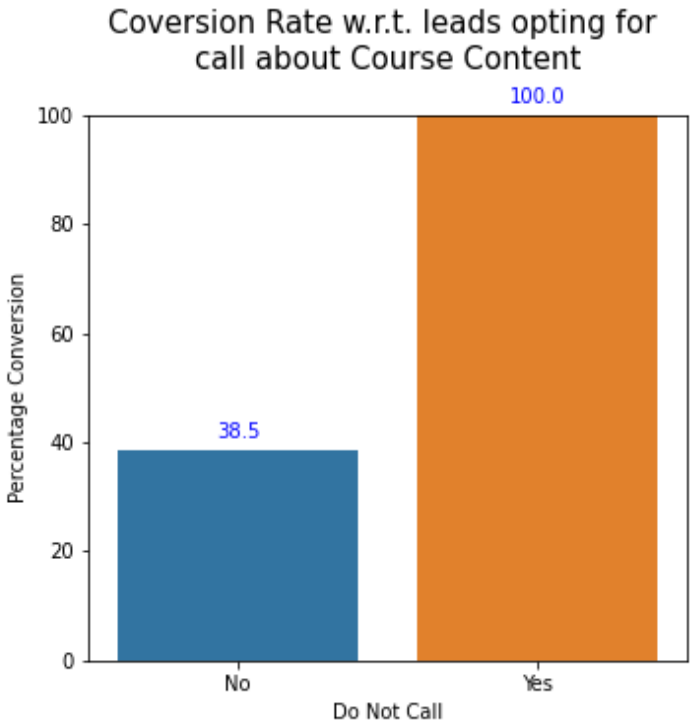




The percentage of Leads who do not prefer to be called are very less. However, their conversion rate is quite high.

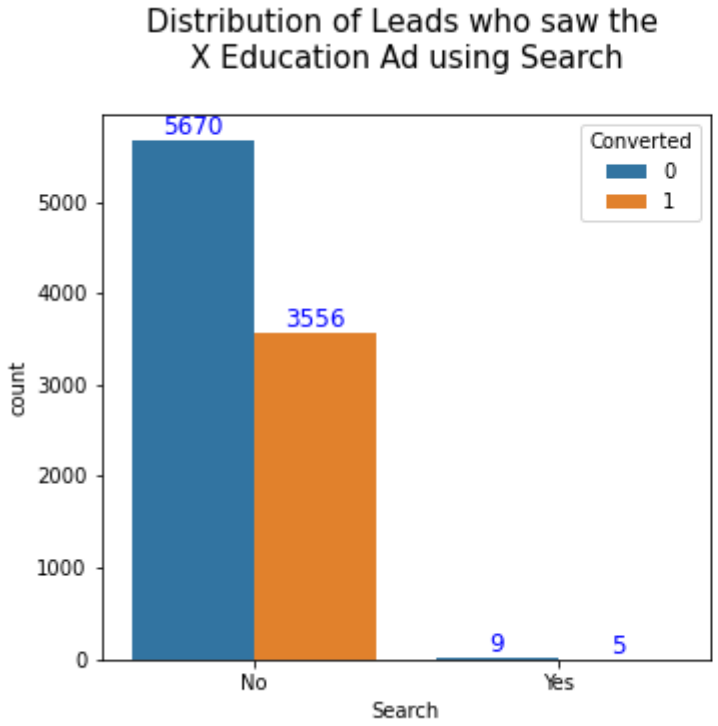
Among Leads who prefer to be called, the lead conversion rate is approximately 39%.

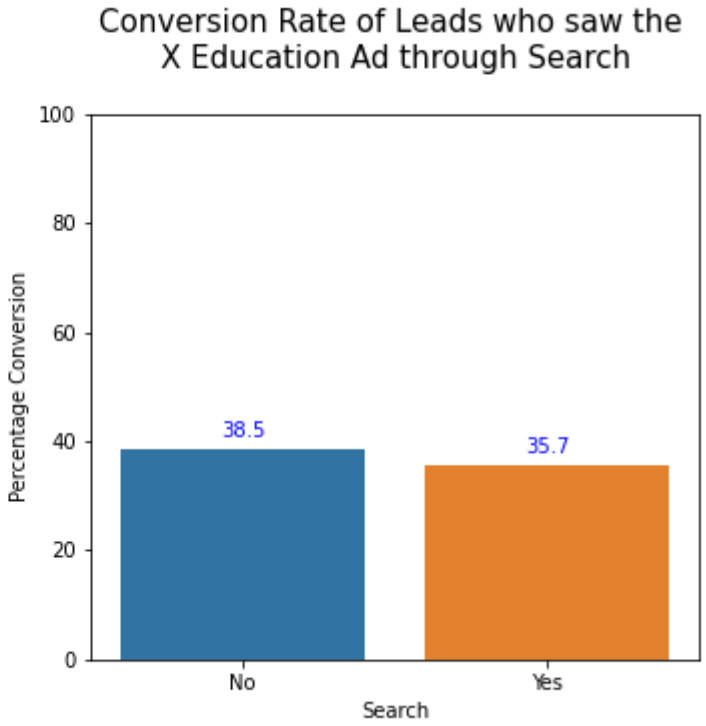




The conversion rate of Leads who did not see the Ad in Search is marginally high compared to those who did see the Ad.

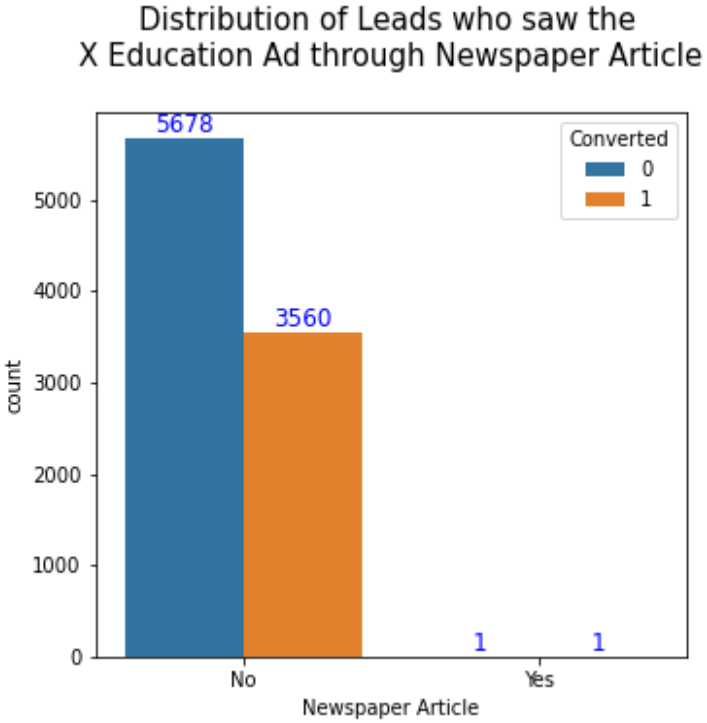
Also, the ratio of the leads who saw the Ad in Search is low compared to the ratio of leads who did not.

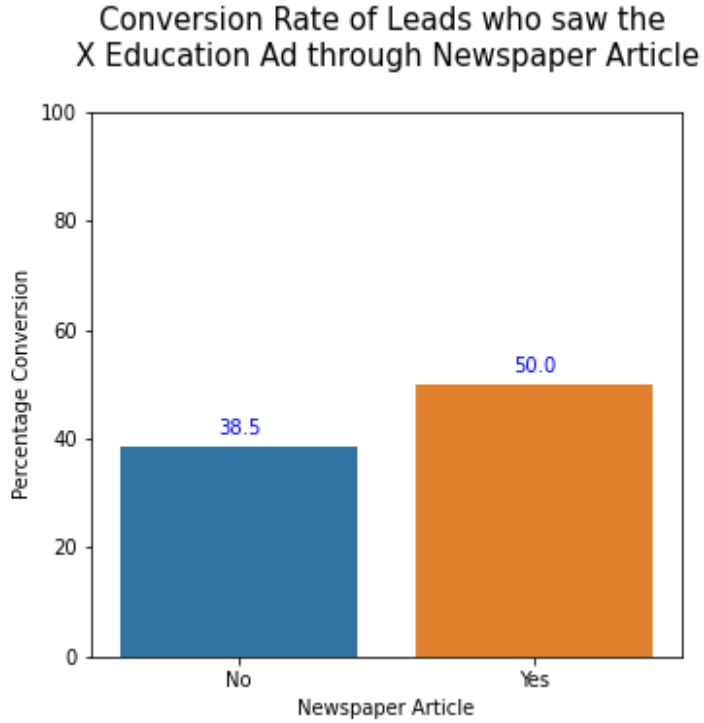




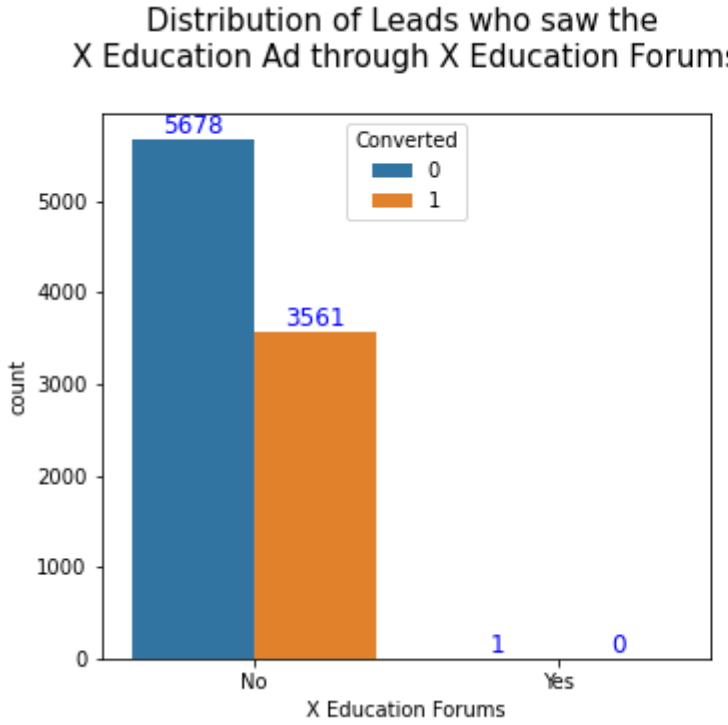
Leads who saw the X Education Ad in the Newspaper Article demonstrated high conversion rate.

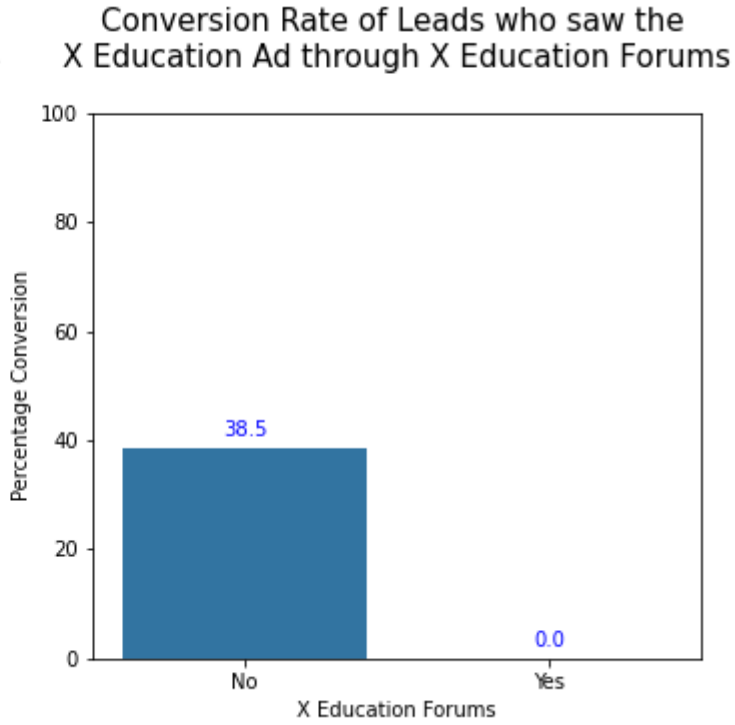
However, majority of the Leads did not come across the X Education Ad in a Newspaper Article, but their conversion is approximately 39%



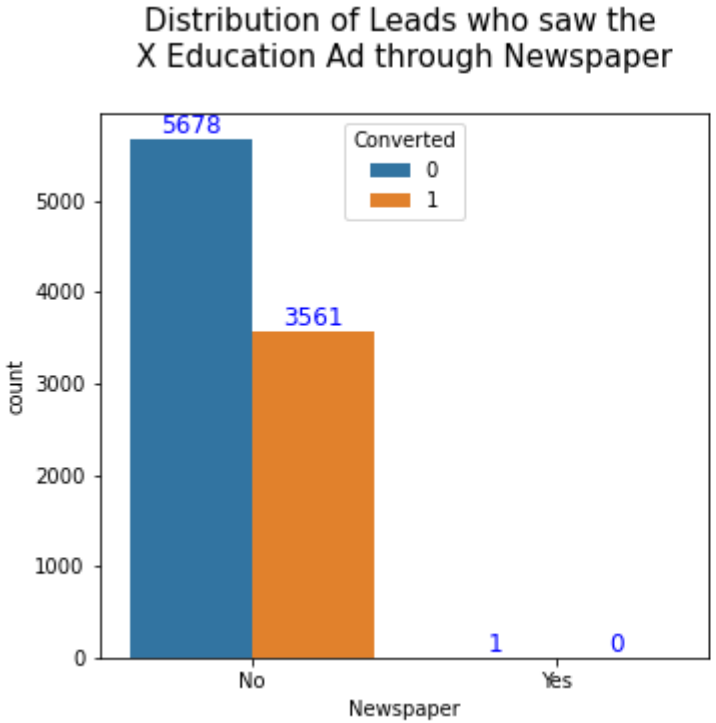


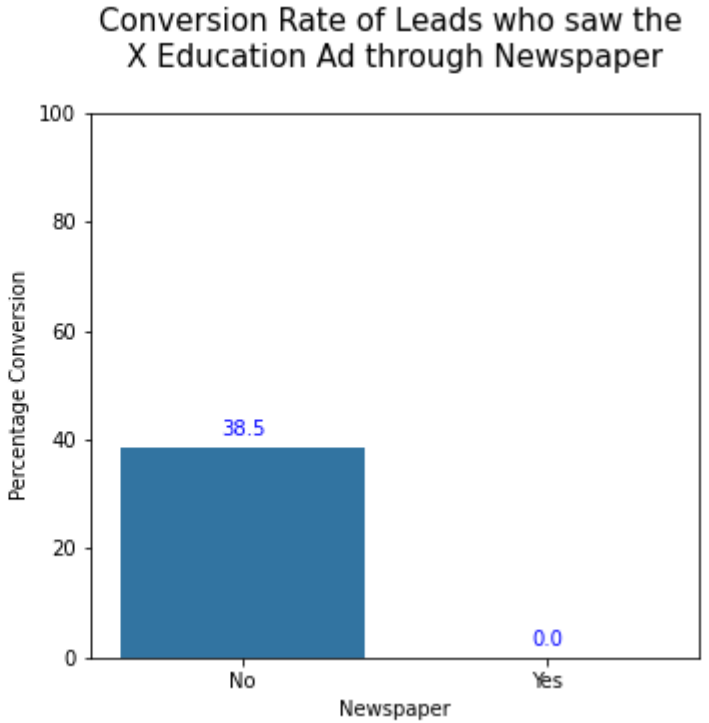
Ad in X Education Forums need not be the driving factor for improving the Lead conversion rate as majority of the Leads acquired did not seem to come across the Ad in X Education Forum but demonstrated nearly 39% Conversion rate.





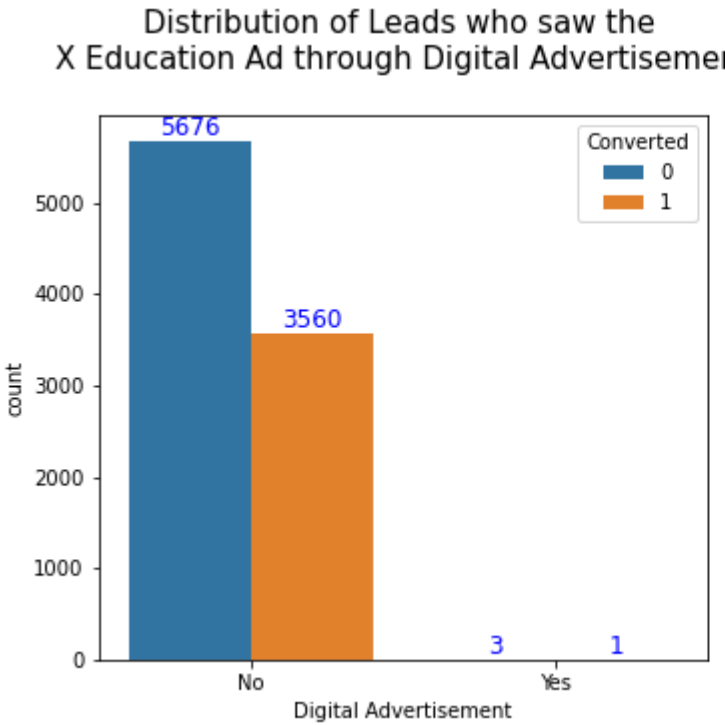
Like X Education Forum, majority of the Leads acquired did not come across the X Education Ad in Newspaper. However, they still comprise 40% of the Lead conversion rate.

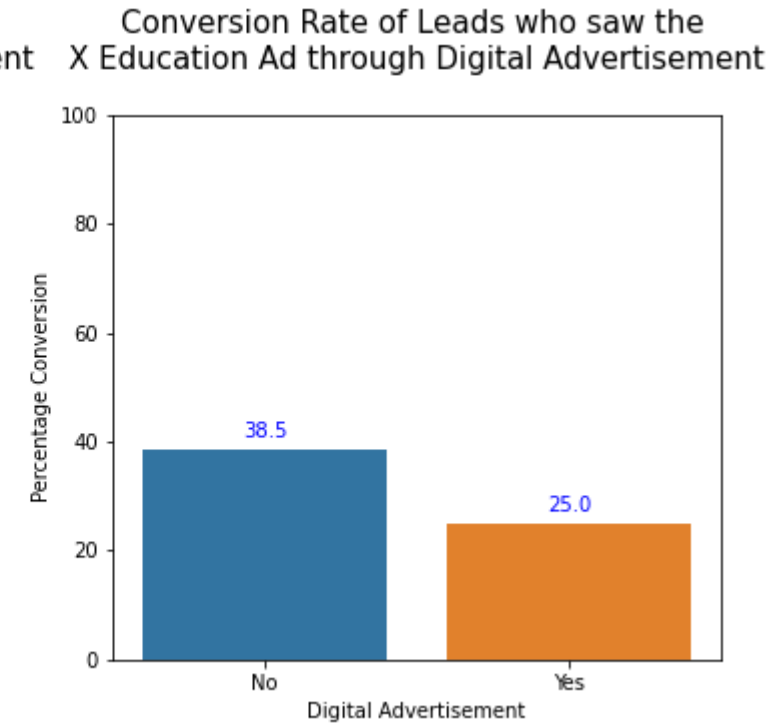




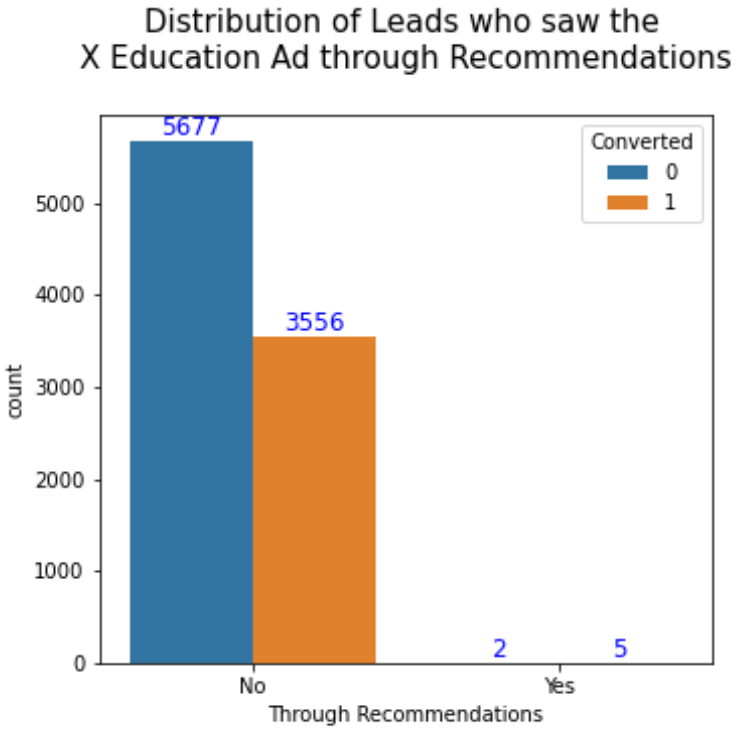
Although major number of Leads did not come across the Digital Advertisement of X Education, they demonstrated better conversion rate compared to conversion rate among people who saw the Digital Advertisement.

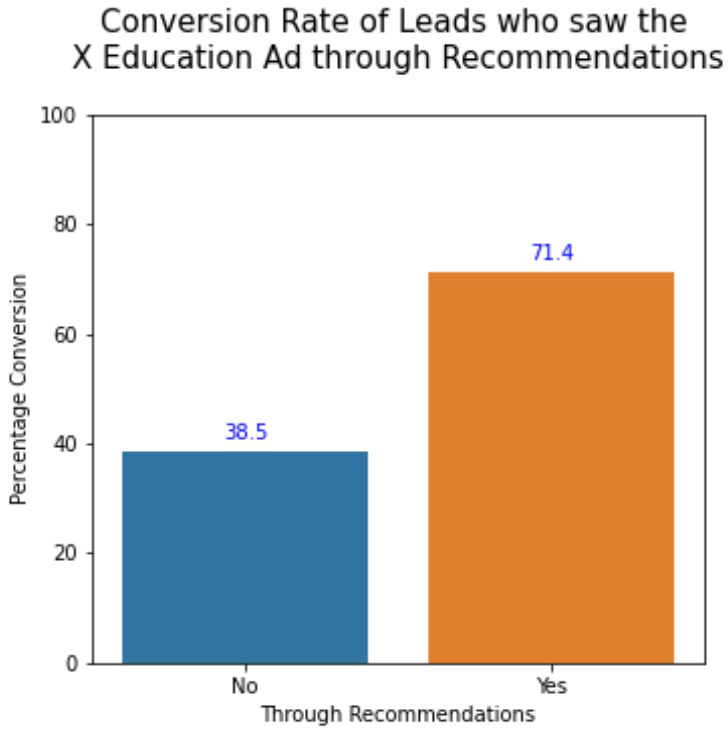
Therefore, Digital Advertisement is not a major driving factor for determining the Hot Leads although it displayed better response compared to other forms of Advertisements.





Through Recommendation can be an important factor to focus in determining Hot Leads since the conversion rate of Leads who came through recommendation is quite high compared to the people who did not come through recommendations.

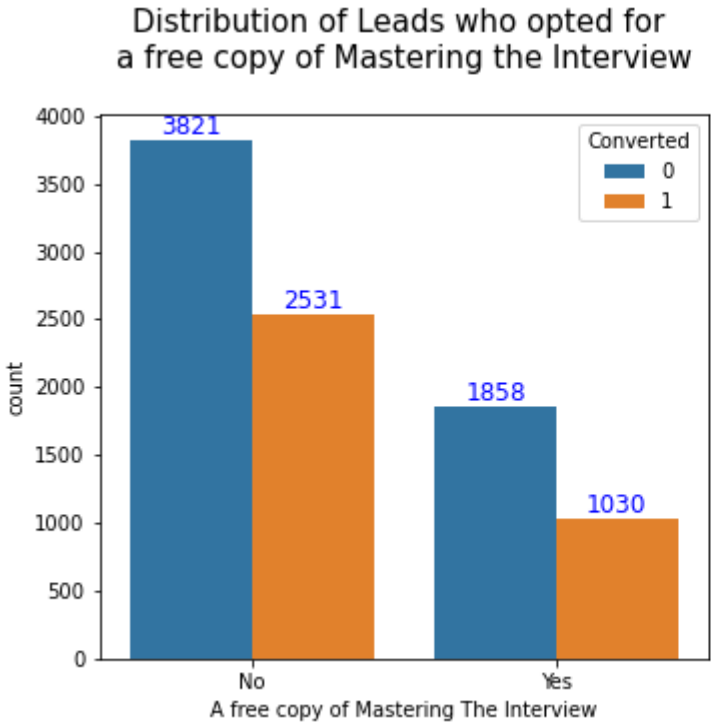


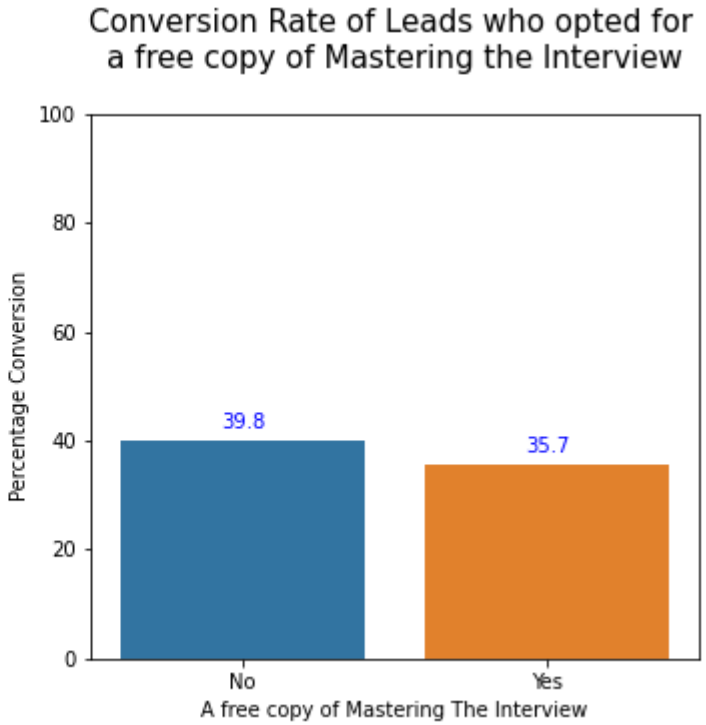


Conversion rate among Leads who opted for 'A free copy of Mastering the Interview' is higher than Leads who did not opt for it.

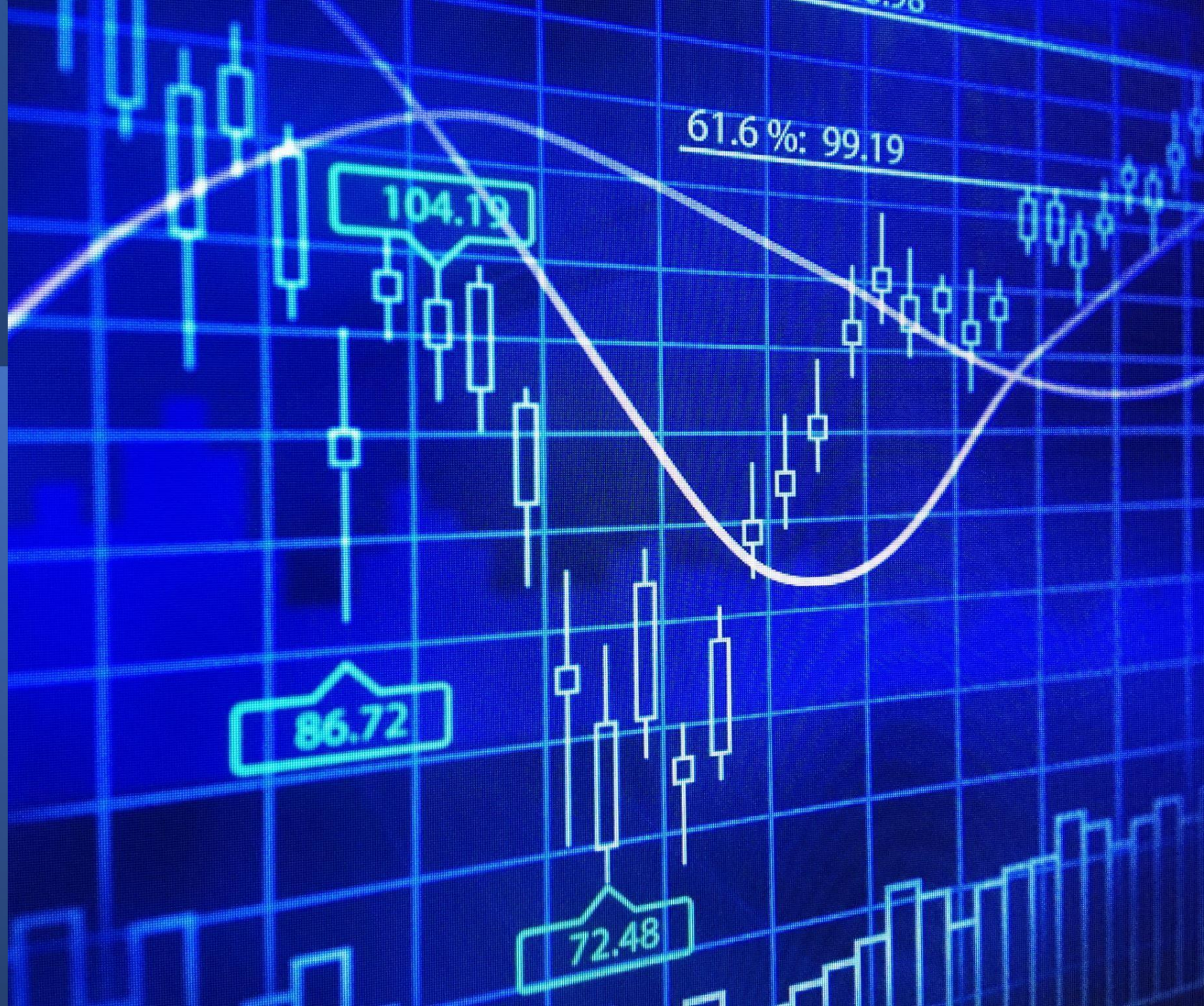
31% of the total Leads opted for a free copy of Mastering The Interview. But approximately 36% of them have converted.

Among 69% of Leads who did not opt for a free copy of Mastering the Interview, the conversion rate is close to 40%





Data Preparation



Dummy Variables

Splitting Data

Scaling Data

Dummy Variables

We need Dummy Variables for converting the categorical variables into Numeric format to use for Model Building.

- Variables ‘Do Not Email’ and ‘A free copy of Mastering the Interview’ contained only two values in them which were converted into Binary format.
- Variables ‘Lead Origin’, ‘Lead Source’, ‘Specialization’, ‘What is your current occupation’ contain more than 2 values for which Dummy variables were created using the function pd.get_dummies().

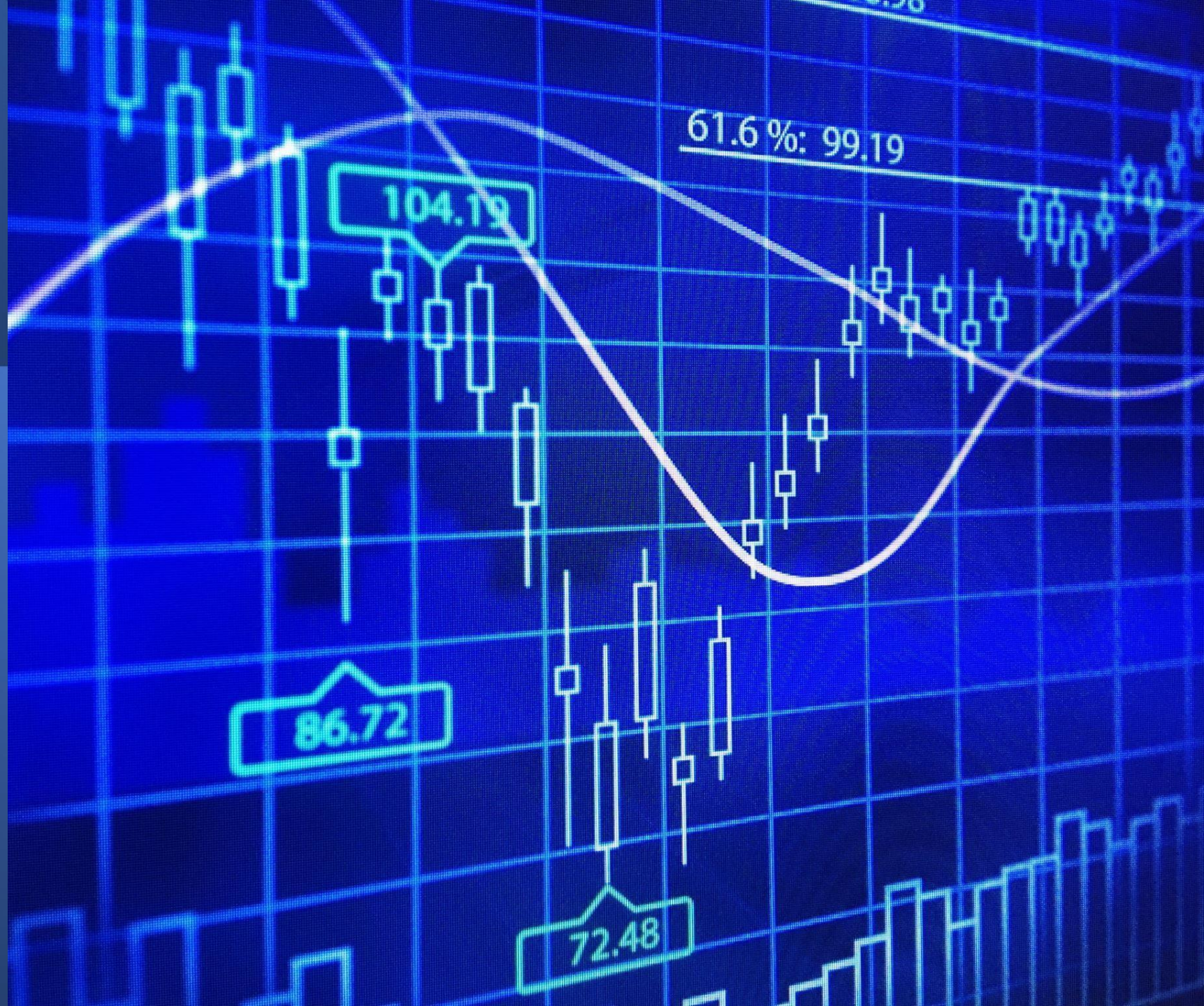
Train & Test Split

The data obtained after converting all categorical columns to numeric must be spilt into Training Data and Test Data using a 70:30 Ratio.

Scaling the Data

The Numeric Variables ‘TotalVisits’, ‘Total Time Spent on Website’, & ‘Page Views Per Visit’ are scaled using Standard Scaler.

Model Building



Approach & Summary

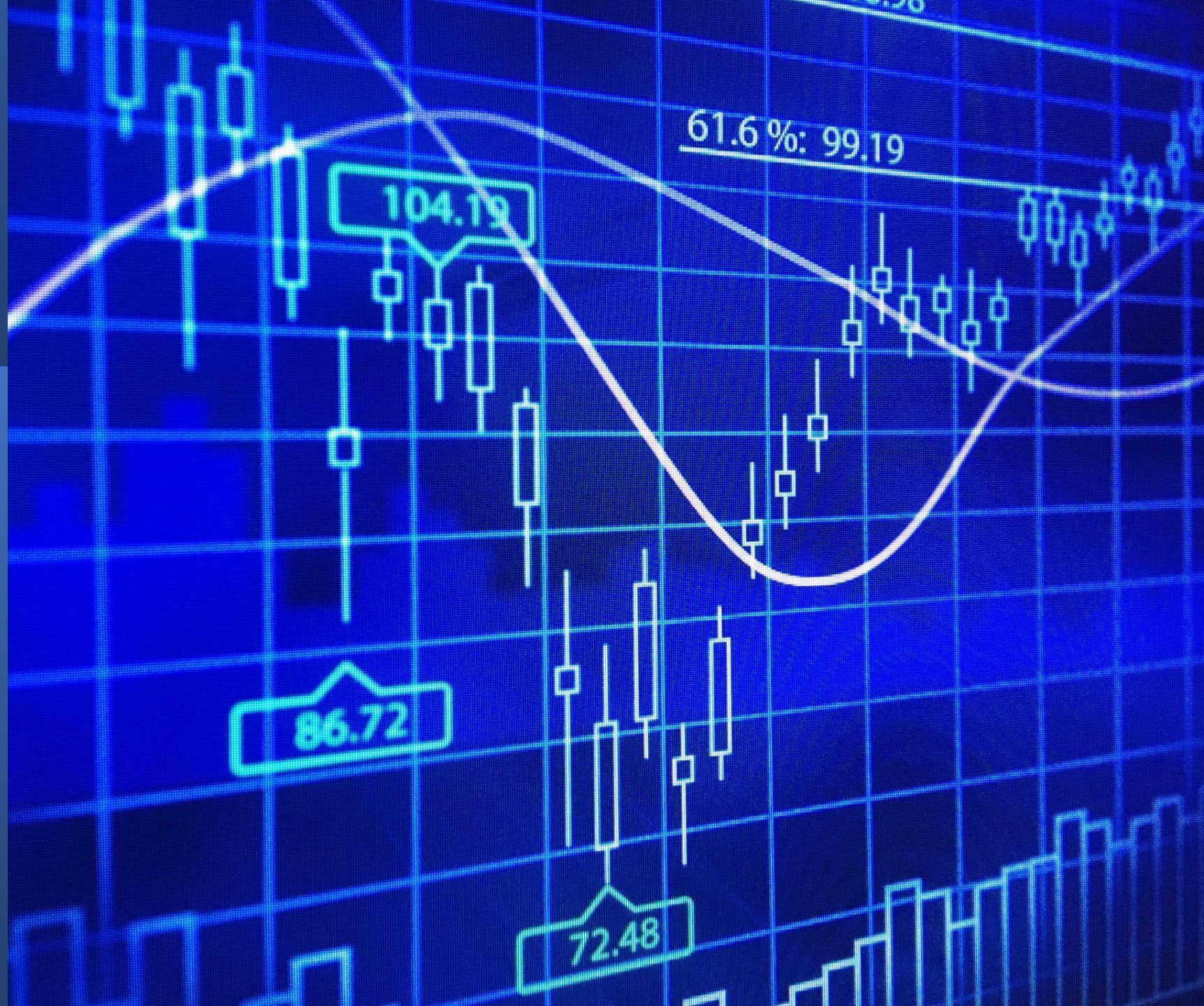
Model Building Approach

The model is built using Mixed approach. Where RFE feature elimination reduces the data to 20 essential variables. We then manually reduced the features/variables one at a time based on their respective p-value and VIF score.

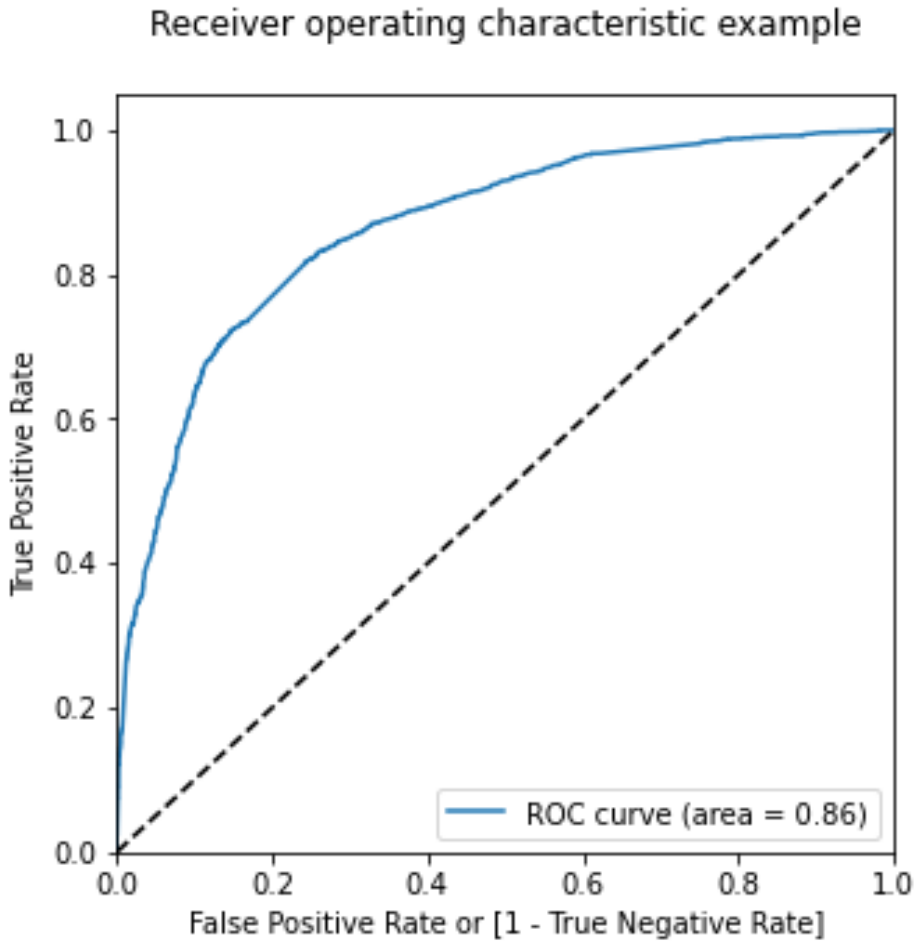
Final Model Summary

1. With unit increase in "**LeadOrigin_Lead Add Form**", the conversion rate will increase by 3.9816 units.
2. With unit increase in "**CurrentOccupation_Working Professional**", the conversion rate will increase by 3.9147 units.
3. With unit increase in "**CurrentOccupation_Others**", the conversion rate will increase by 2.1206 units.
4. With unit increase in "**CurrentOccupation_Unemployed**", the conversion rate will increase by 1.4149 units.
5. With unit increase in "**Total Time Spent on Website**", the conversion rate will increase by 1.0849 units.
6. With unit increase in "**CurrentOccupation_Student**", the conversion rate will increase by 1.0066 units.
7. With unit increase in "**LeadSource_Olark Chat**", the conversion rate will increase by 0.9618 units.
8. With unit increase in "**Banking, Investment And Insurance**", the conversion rate will increase by 0.3463 units.
9. With unit increase in "**Do Not Email**", the conversion rate will decrease by 0.3463 units.

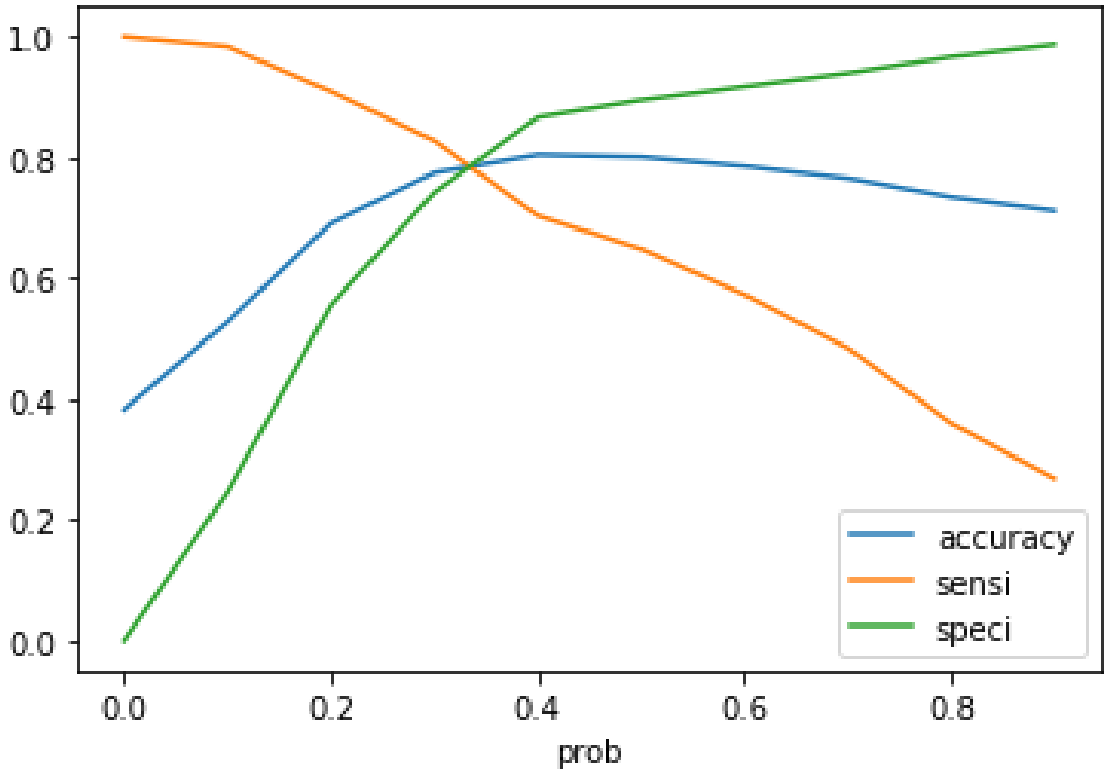
Model Evaluation



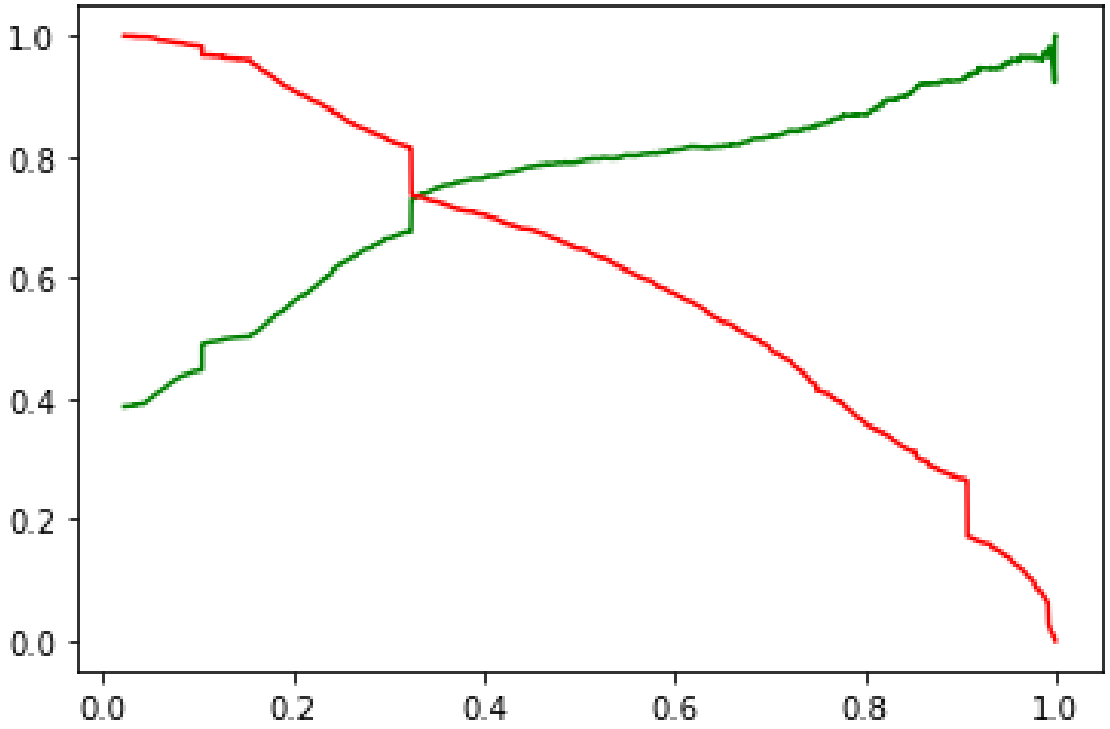
ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

Optimal Cutoff Point



Precision Recall Tradeoff



Final Scores

Evaluation Metrics

The Final Evaluation Metrics for the train Dataset:

- The Accuracy is : 77.50%
- The Sensitivity is : 82.79%
- The Specificity is : 74.22%

The Final Evaluation Metrics for the test Dataset:

- The Accuracy is : 77.30%
- The Sensitivity is : 81.70%
- The Specificity is : 74.47%

Cut-Off Value : 0.30