Lead Scoring Assignment

SUBMITTED BY:

Shreshth Vyas

Sourav Dutta

Imran khan

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- 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’.
- Our goal is to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- We have to build an efficient model to identify the hot leads so the that the conversion rate must reach to more than 80%.

Our Way-out For the Solution

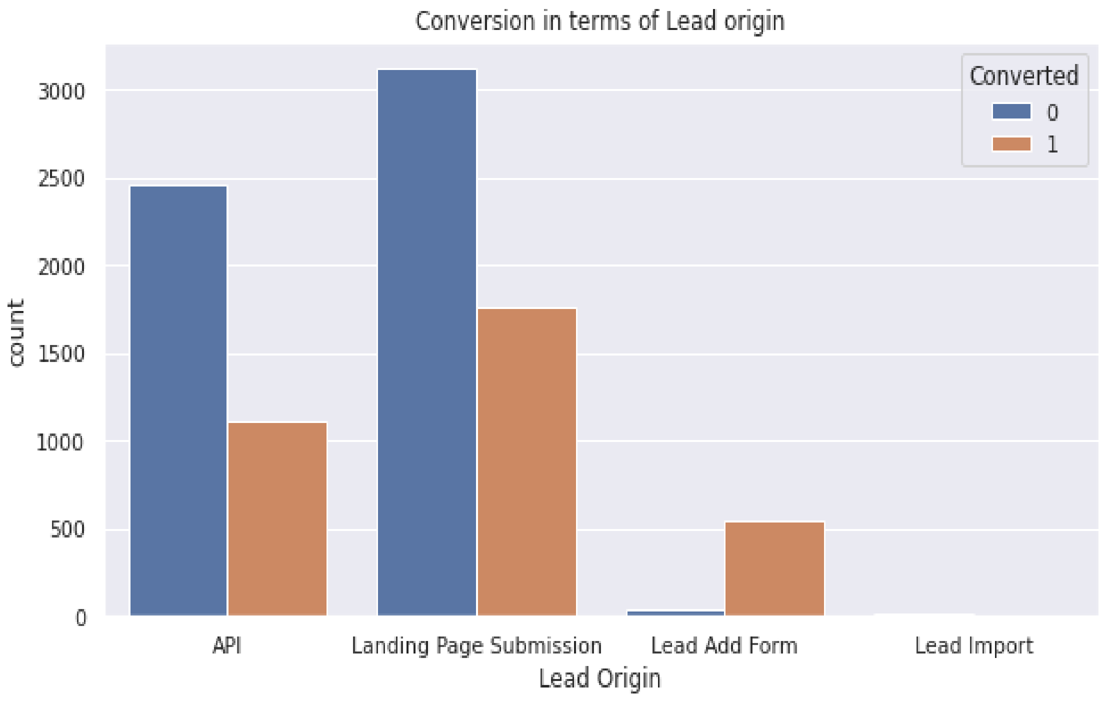
- Reading, Understanding and Cleaning the Data.
- Univariate Analysis and Bi-variate Analysis.
- Data Preparation.
- Train-Test Split.
- Feature Scaling And Feature Selection Using RFE.
- Model Building and Evaluation.
- Plotting the ROC Curve and Finding Optimal Cutoff Point
- Metrics - Precision and Recall.
- Making predictions on the test set.
- Final Observation And Recommendations.

Reading, Understanding and Cleaning the Data.

- The leads dataset had 9240 rows and 37 columns.
- There were 7 numerical columns and 30 categorical columns
- Null values were identified and the columns having null values more than 45% were removed,
- Approximately 58% of the data in city column was Mumbai so we replaced the missing values with 'Mumbai'.
- Similarly 36% null values were there in the column 'Specialization', we replaced those null values with 'Others' since 'NaN' values had the highest percentage.
- Around 85% values in the column 'What is your current occupation' was Unemployed so we imputed the missing values with 'Unemployed' .
- Columns like 'Tags', 'What matters most to you in choosing a course', 'Country' were dropped as these were not affecting our model

Univariate Analysis and Bi-variate Analysis.

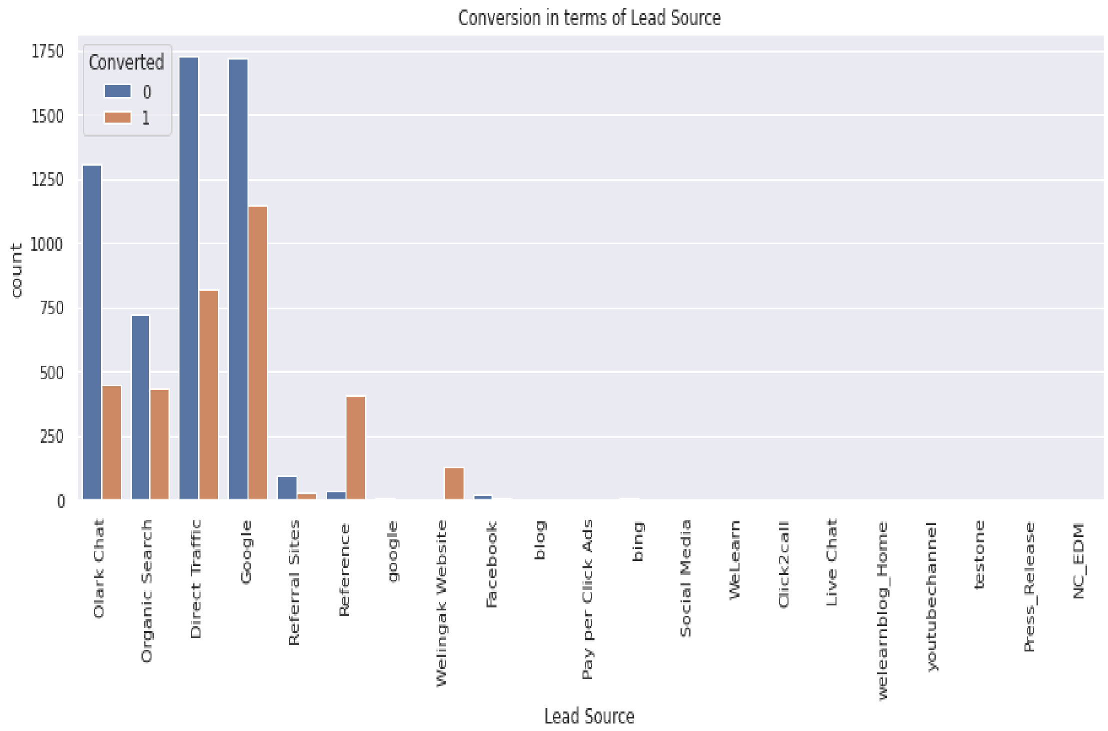
- Based on univariate and Bi-variate analysis following are the some visualization And Conclusion

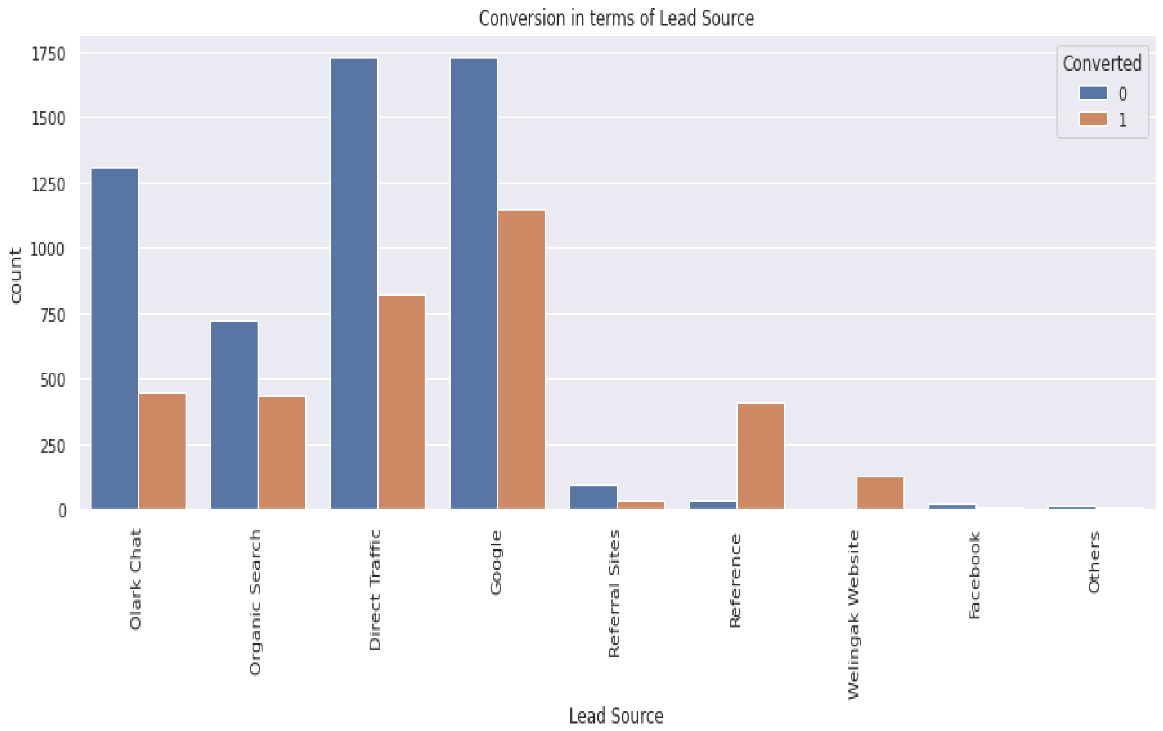


- The highest conversion rate is of 'Lead Add Form', at 94%
- Landing Page Submission' and 'API' have 36% and 31% conversion rate, respectively, but they generate maximum leads counts.
- 'Lead Import' has both the least amount of conversions and leads count.
- In order to improve overall lead conversion rate, the focus should be on improving the rate of 'API' and 'Landing Page Submission', also, on to generate more leads from 'Lead Add form' since they have a very good conversion rate.

Univariate Analysis and Bi-variate Analysis.

- Based on univariate and Bi-variate analysis following are the some visualization And Conclusion

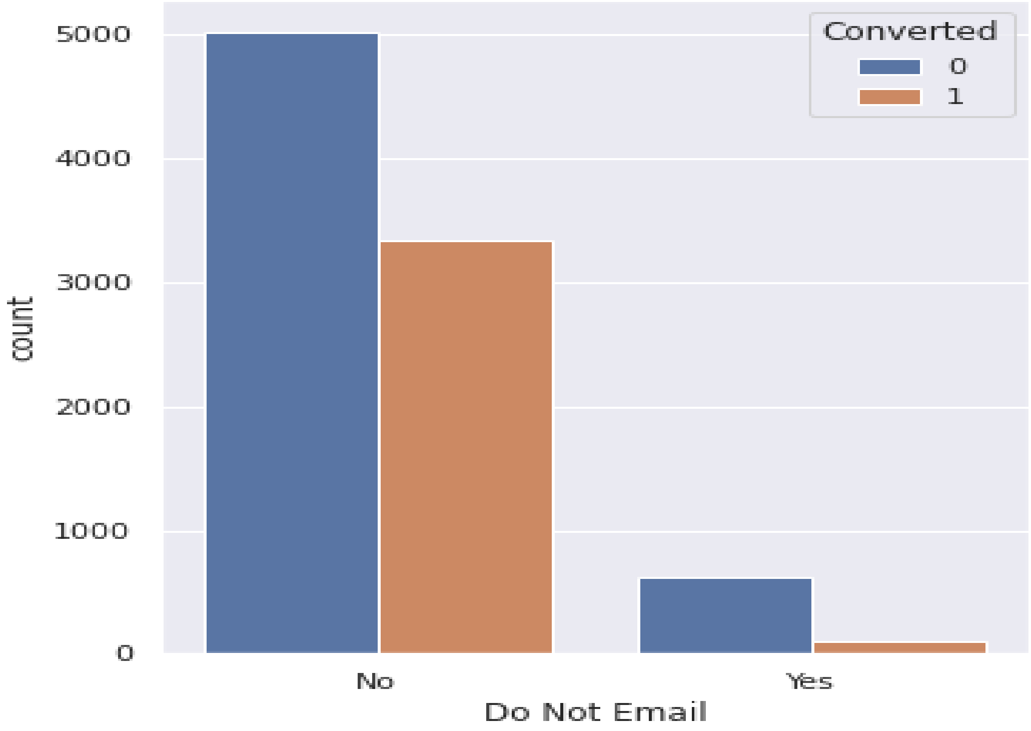

Conversion in terms of Lead Source

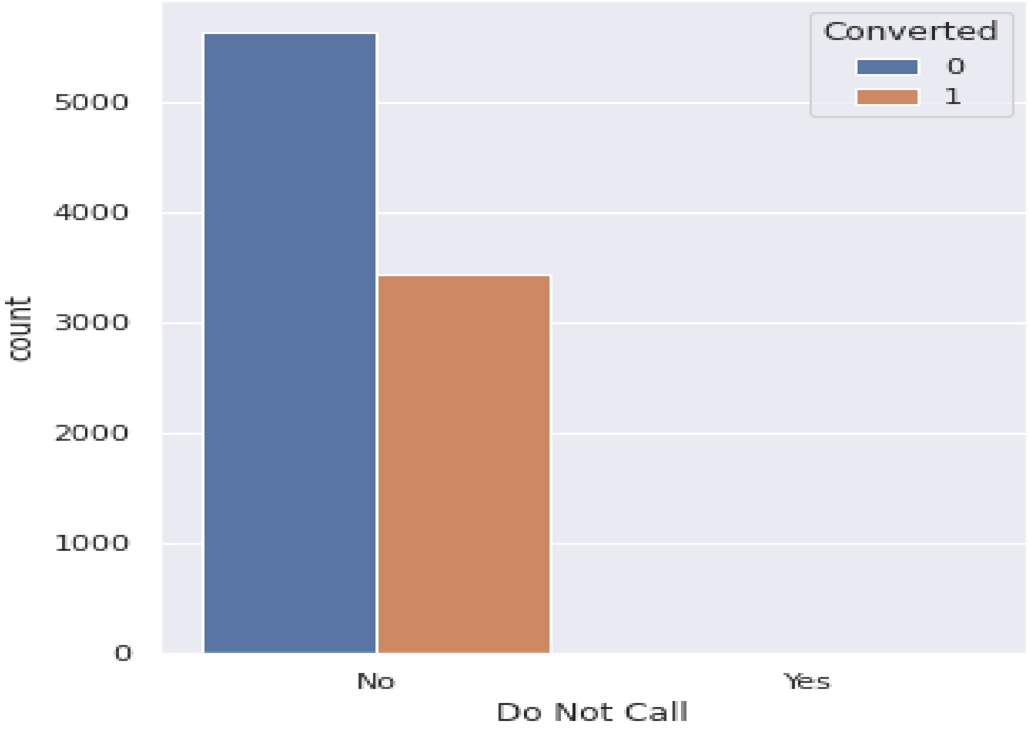

Conversion in terms of Lead Source

- Some Lead sources had very low count .So, we merged them into a common category 'Others'

Univariate Analysis and Bi-variate Analysis.

- Based on univariate and Bi-variate analysis following are the some visualization And Conclusion

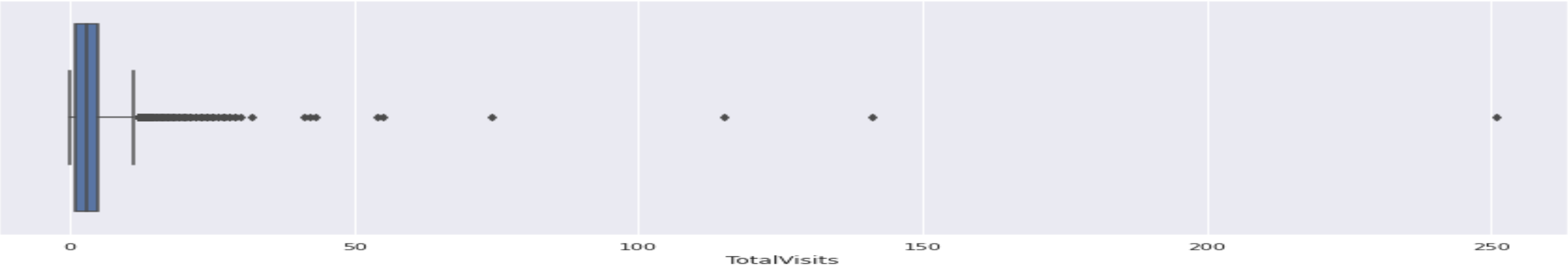


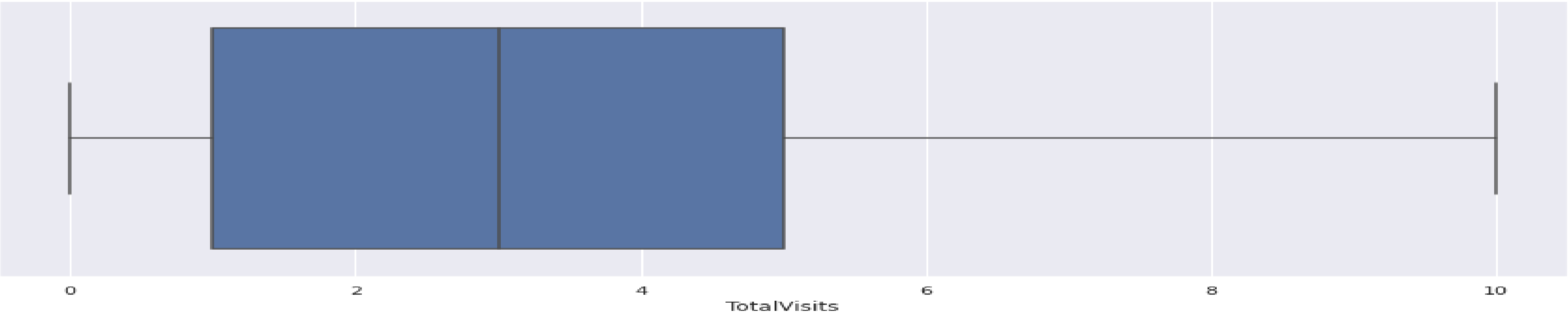


- Based on the above plot and conversion summary, we can infer that around 99% of customers do not like to be called or receive emails about the course.

Univariate Analysis and Bi-variate Analysis.

- Based on univariate and Bi-variate analysis following are the some visualization And Conclusion

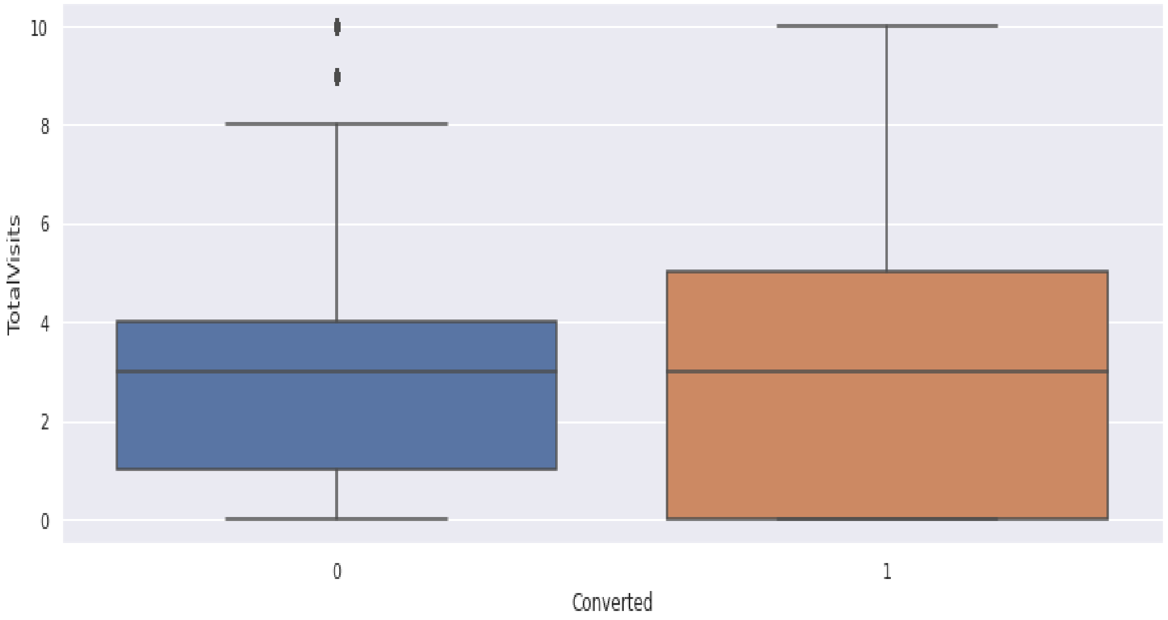


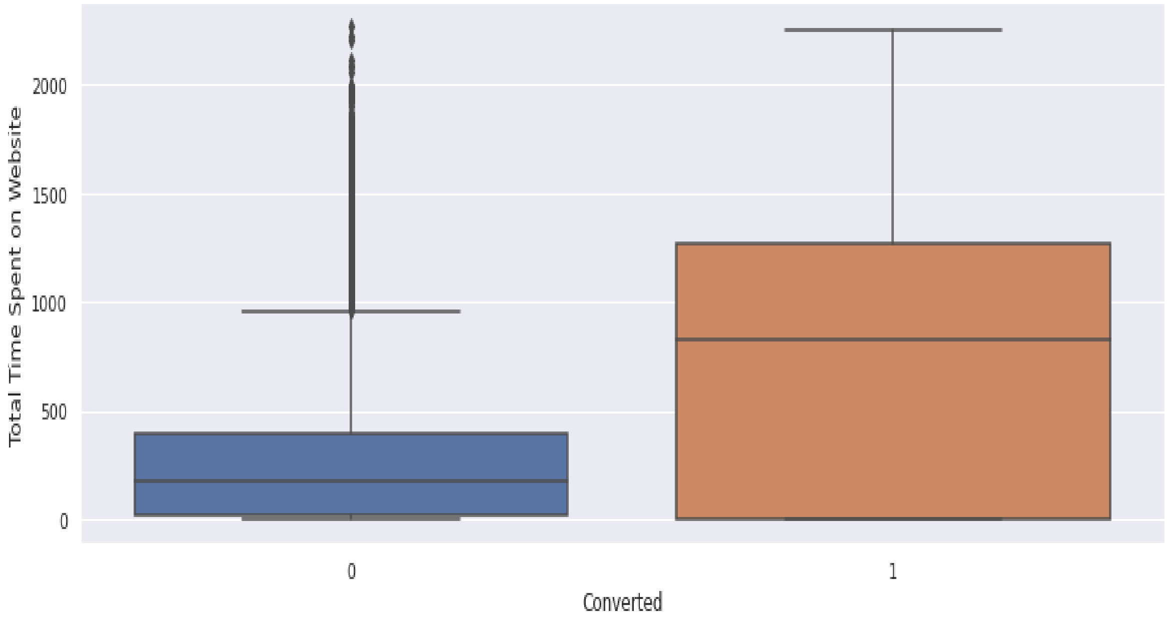


- There were number of outliers in Total Visits column. We capped the outliers to 95%.

Univariate Analysis and Bi-variate Analysis.

- Based on univariate and Bi-variate analysis following are the some visualization And Conclusion

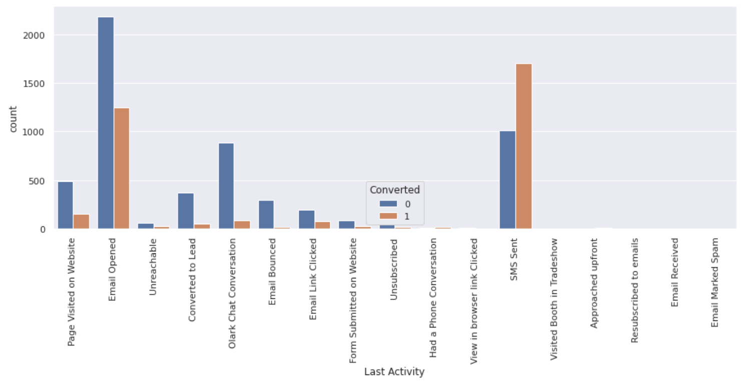


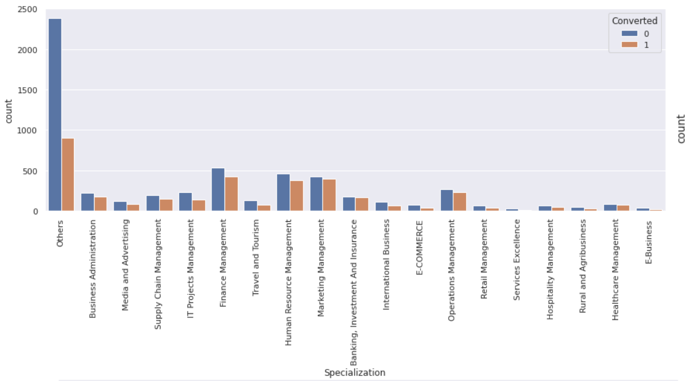


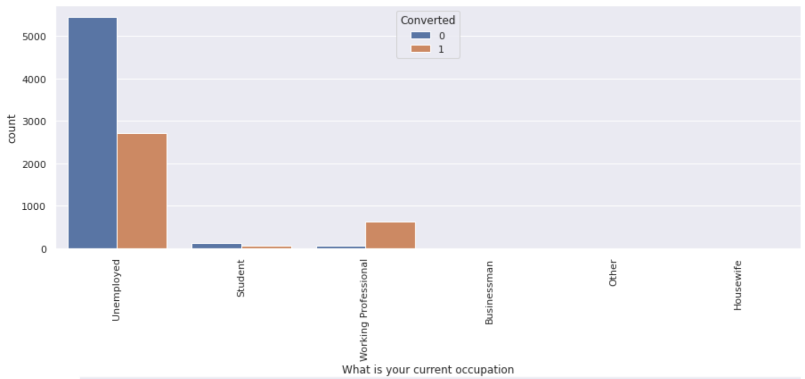
- People who visits the platform have equal chances of applying and not applying for the course, it is 50-50.
- More time people spend on the website, more chance there is of them opting for a course

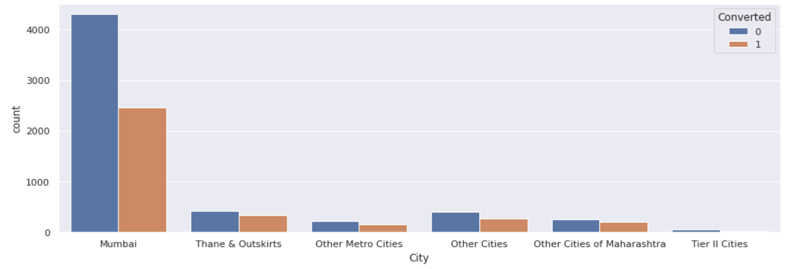
Univariate Analysis and Bi-variate Analysis.

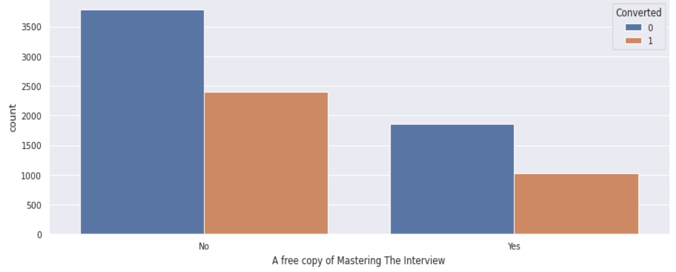
- Based on univariate and Bi-variate analysis following are the some visualization And Conclusion

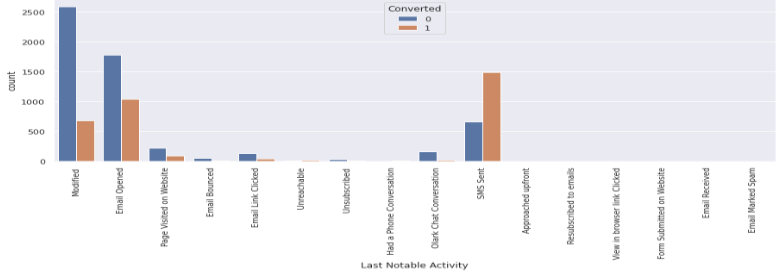








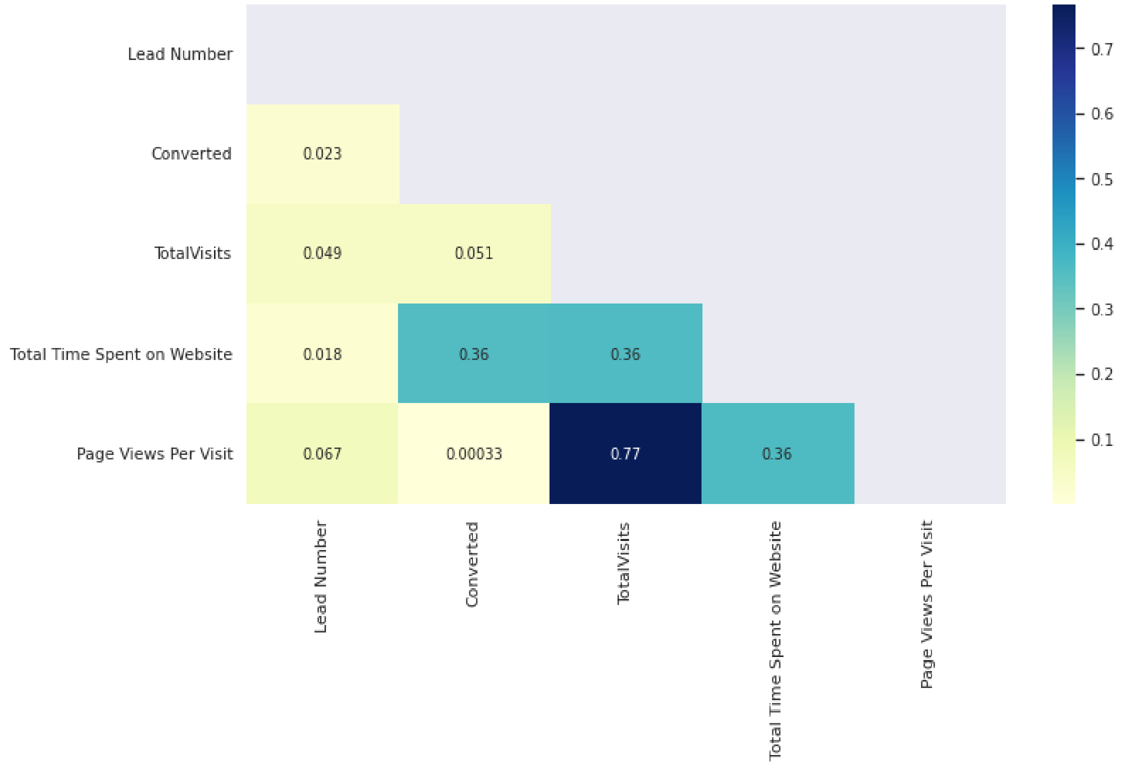




- Maximum leads are being generated from the city of Mumbai, with conversion rate of around 36%. Hence focus should me more on increasing conversion rate of Mumbai city.
- Working Professionals and Unemployed people generates maximum leads.
- Conversion rate for Working Professionals is high around 92% and Conversion rate for Unemployed is around 33%

Univariate Analysis and Bi-variate Analysis.

- Based on univariate and Bi-variate analysis following are the some visualization And Conclusion



- 'Total visits' and 'Page views per visit' columns are correlated.
- Hence we should have either of this column in our model to avoid multi-collinearity.
- Based on our data analysis, we conclude that many variables are not significant to the model. Hence we can drop them for further analysis
- Columns 'Lead Number', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations' are dropped for further analysis.

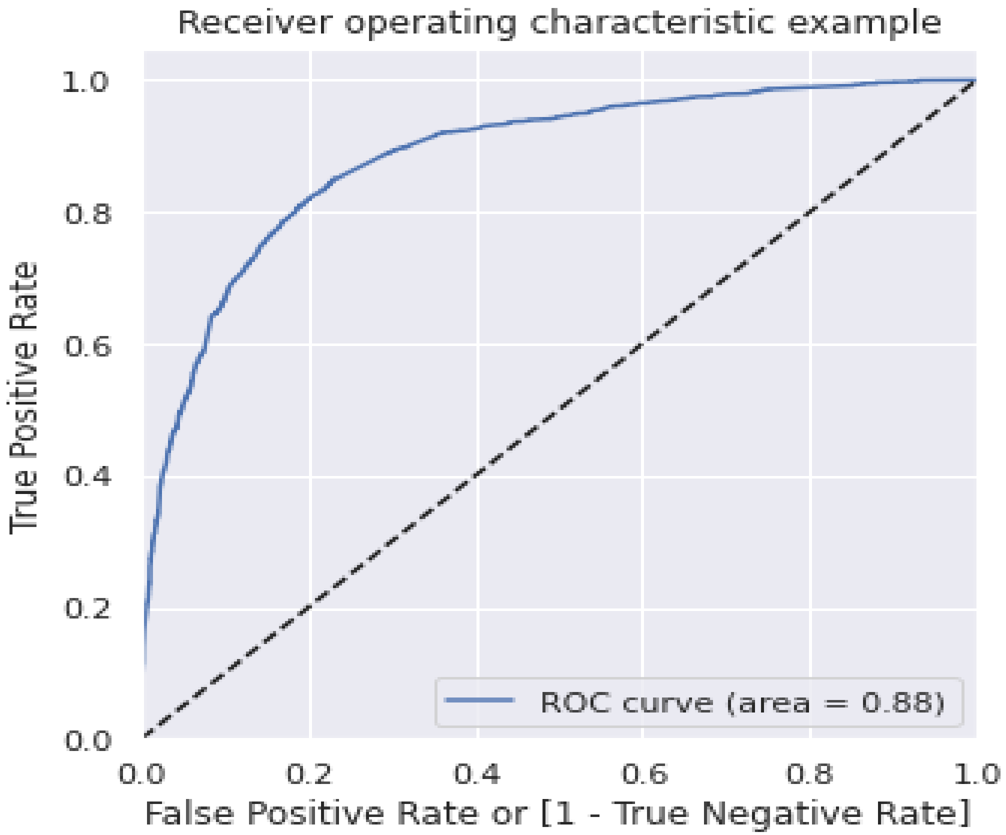
Data Preparation

- Binary variables (Yes/No) converted to (1/0) for the column 'A free copy of Mastering The Interview’, 'Do Not Email’ and 'Do Not Call’.
- Dummy variables created for the categorical variables for the column 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization’, 'What is your current occupation’, 'City’ and 'Last Notable Activity’.
- Dummy data is combined with the original dataset.
- After that original columns are dropped leading to final shape (9074, 75) of the data frame.
- In this way the data is prepared for the Train data –Test data Split

Train-Test Split, Feature Scaling Using RFE And Model Building

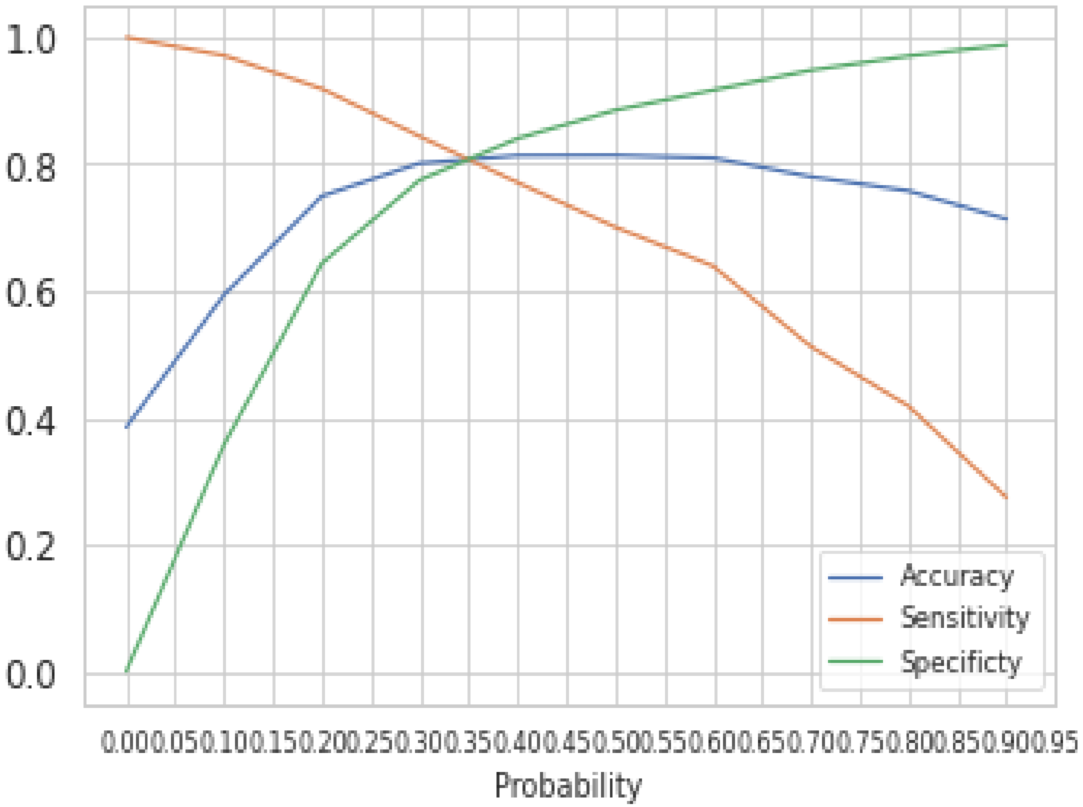
- Feature variable is set to X
- Response variable is set to y
- Splitting the data into train and test
- Logistic regression run.
- Four Model is build and evaluated with help of train and test data.
  - Overall accuracy 0.81
  - Sensitivity is 0.7
  - Specificity is 0.89
  - false positive rate is 0.115
  - positive predictive value is 0.79
  - Negative predictive value 0.83

Plotting the ROC Curve



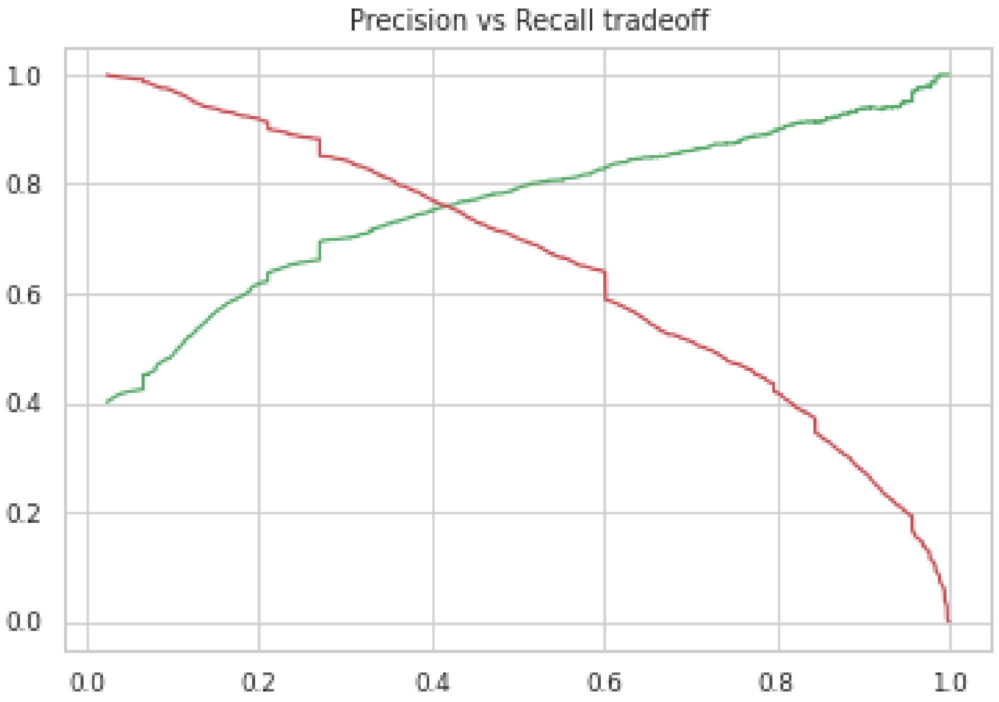
- ROC shows the tradeoff between sensitivity and specificity.
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test will be deemed.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test will be deemed.

Finding Optimal Cut-off Point



- Based on the adjacent curve we can see that the optimal cutoff is at 0.35. That is the point where all the parameters - Accuracy, Sensitivity, Specificity are equally balanced.
- When the optimal cutoff is selected to be 0.35, the various performance parameters, that is Accuracy, Sensitivity & Specificity, are all 80%.

Metrics - Precision and Recall



- In accordance to our business objective, the recall percentage is significant as we do not want to leave out any hot leads which are willing to get converted. Hence Recall value 81% suggest a good model.
- As seen from above plot, there is a tradeoff between Precision and Recall. Precision and Recall are inversely related, so if one increases other will decrease.

Making predictions on the test set.

Final Observation:

- Comparing the Model Performance parameters obtained for Train & Test data:

Train Data:

- Accuracy : 80.96%
- Sensitivity : 80.98%
- Specificity : 80.94%
- Precision : 72.69%
- Recall : 80.98%

Test Data:

- Accuracy : 80.35%
- Sensitivity : 79.37%
- Specificity : 80.91%
- Precision : 70.34%
- Recall : 79.37%

- There is around 1% difference on train and test data's performance metrics. This means that the final model did not overfit training data and is performing well.
- High Sensitivity will make sure that all possible leads who are likely to convert are correctly predicted, where as high Specificity will ensure that the leads that are on the brink of the probability of getting converted or not are not selected.
- Based on the business requirement, we can increase or decrease the probability threshold value which in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model as required.

Final Recommendations:

- The X-Education should focus on the leads having
  - lead origin - lead add form
  - occupation - Working Professional
  - Lead source - Wellingak website
- Sales Team of the company should first focus on the 'Hot Leads' which are identified with having Lead Score above 35.
- Once the Sales Team is done with the 'Hot Leads' only after that they should focus on the 'Cold Leads'(Customer having lead score <= 35).
- Team can make some important variables mandatory to enter, such as city, specialization , occupation which can potentially explain Conversion better. It could be used in our model and build important decisions for the business.
- The model has a high recall score than precision score. Hence it can adjust with the company's requirements in coming future.
- High Sensitivity in the model will ensure that almost all leads who are likely to Convert are correctly predicted.
- High Specificity in the model will ensure that leads that are on the brink of the probability of getting Converted or not, are not selected.
- Team can focus least on customers who do not want to be called about the course.
- If a customer's the Last Notable Activity is Modified, they may not be the potential lead.