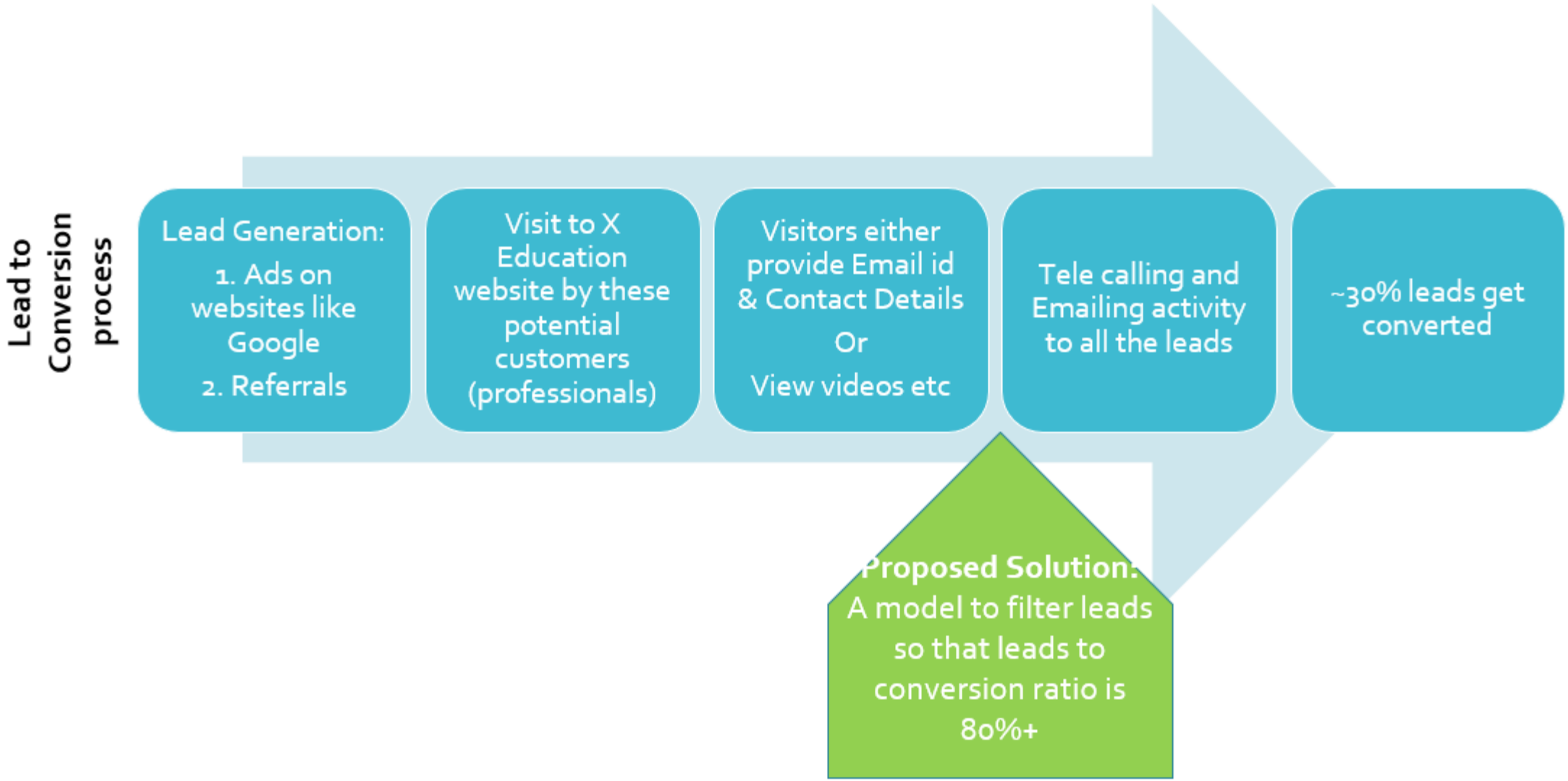
Lead Score case study

-Kamal Teja Manchu
-Daniella Brito

Lead – Conversion Process



Analysis Methodology

Data import and cleaning

- Import the data
- Inspect the data
- Identify the data quality issues and clean the data.

Outlier analysis and removal

- Identify the outliers and remove the same from the specific column.

Univariate analysis

- Visualizing the original data variables to look for any pattern or correlation.

Bivariate analysis

- Visualizing the data using heatmap and pair plots and with respect to 'Converted'.

Data preparation

- Creating the dummy variables of the categorical variables.
- Converting some binary variables (Yes/No) to 0/1

Train– test split

- Splitting the data with 70% of the data as train set.

Scaling the data

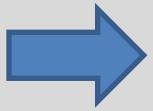
- Standardizing all the continuous variables. We used Standard Scaler method.

Model building and evaluation

- Running the stats GLM model on train dataset.
- Observing the statistical significance of the features
- Feature elimination using RFE coupled with manual feature elimination.
- Calculating accuracy and Sensitivity-Specificity metrics. Precision and Recall with tradeoff.
- Plotting the ROC curve
- Finding optimal cut off point
- Re-running the model

Making Prediction

- Making the prediction on the test set.
- Calculating accuracy and Sensitivity-Specificity metrics on the test set.



Conclusion

- Drawing a conclusion by assigning the lead score to our actual dataset such that the customers with higher lead score have a higher conversion chance.

Methodology we will be following :

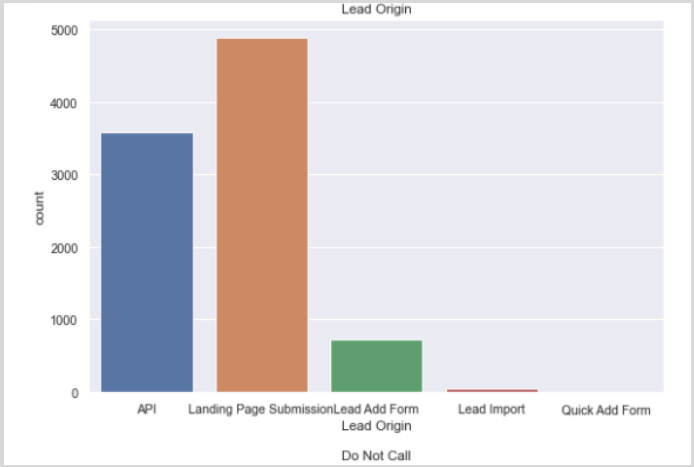
1. Data cleaning and data manipulation : Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

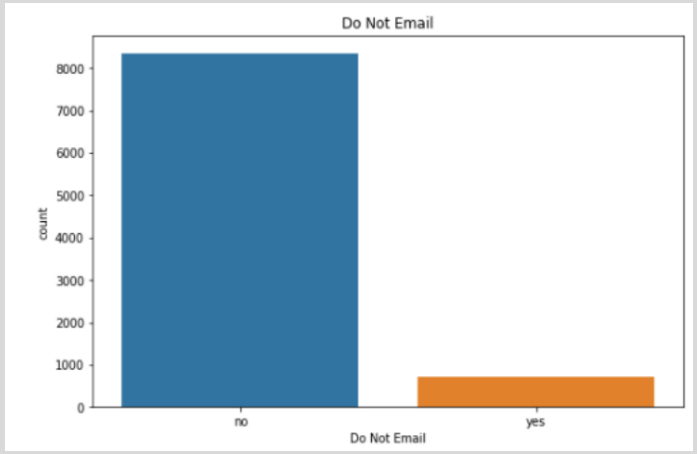
   EDA- 1. Univariate data analysis: value count, distribution of variables

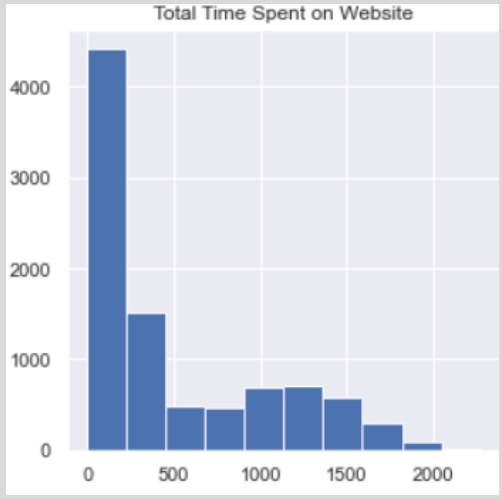
   2. Bivariate data analysis: correlation coefficients and pattern between the variables.
6. Feature Scaling & Dummy Variables and encoding of the data.
7. Classification technique: Logistic regression used for the model making and prediction.
8. Validation of the model.
9. Model presentation.
10. Conclusions and recommendations.

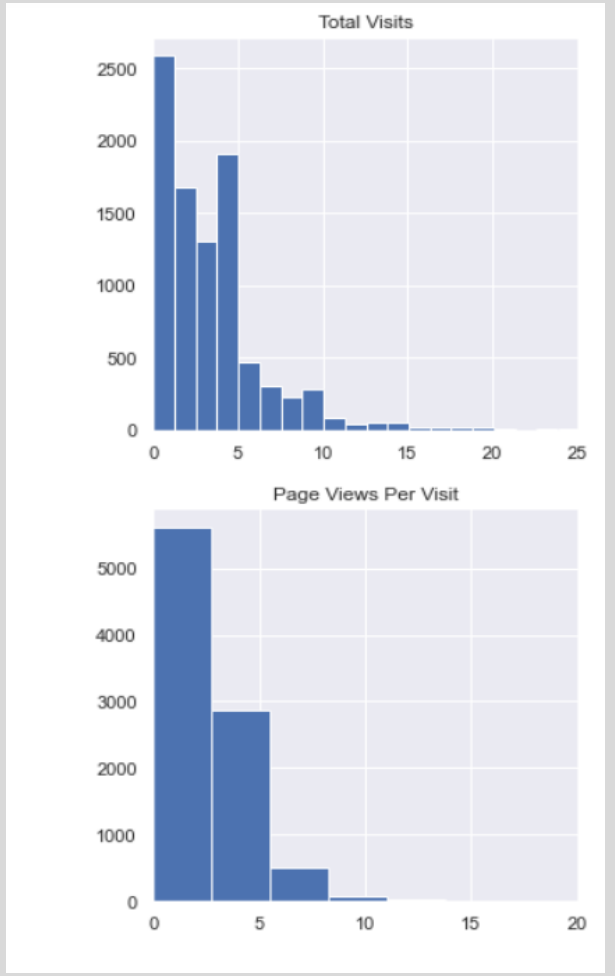
Univariate and Bi-variate Analysis

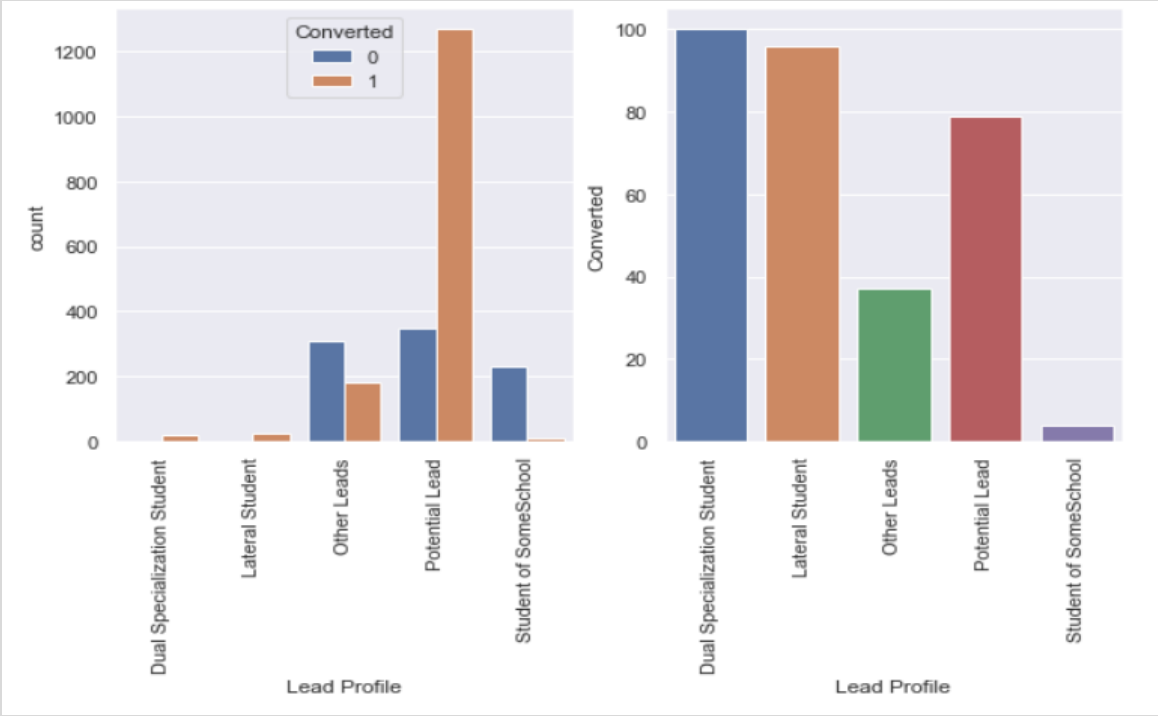
Univariate analysis- For Uni variate analysis we plot Histograms and count plots to draw inferences such as the count values .

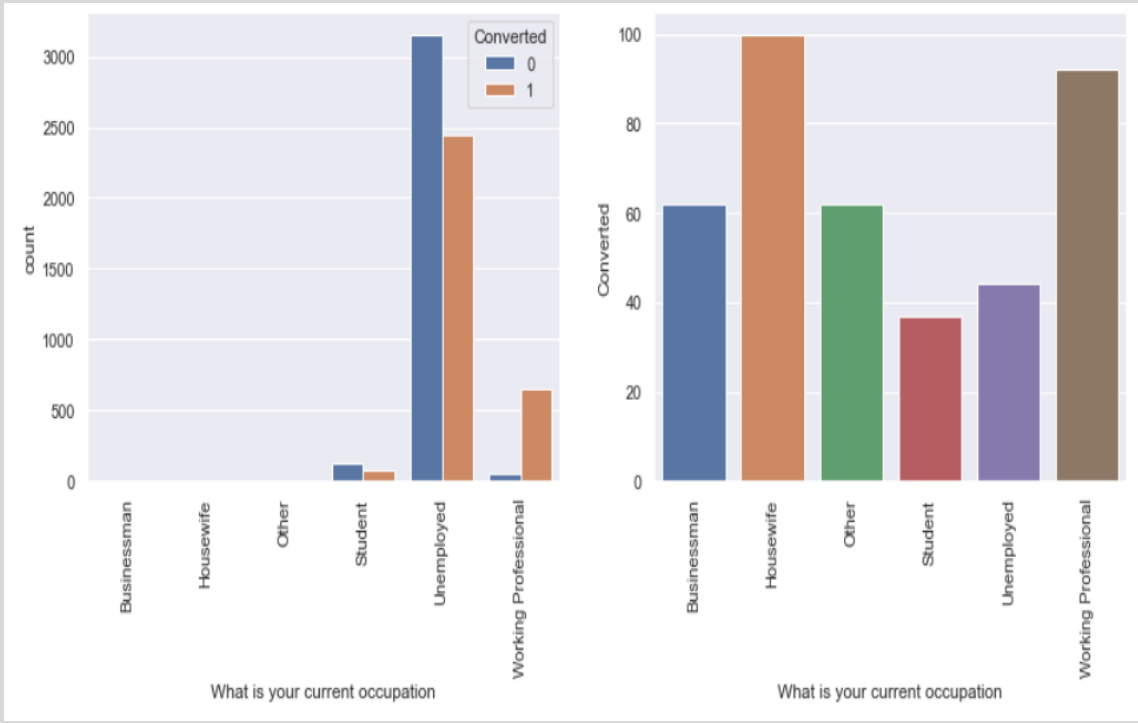






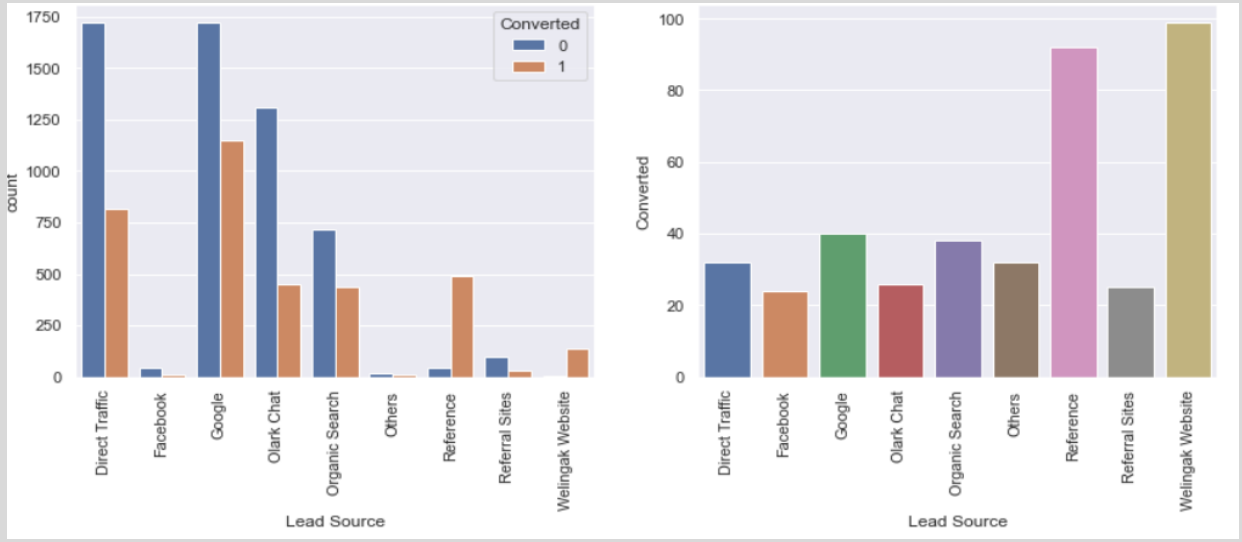






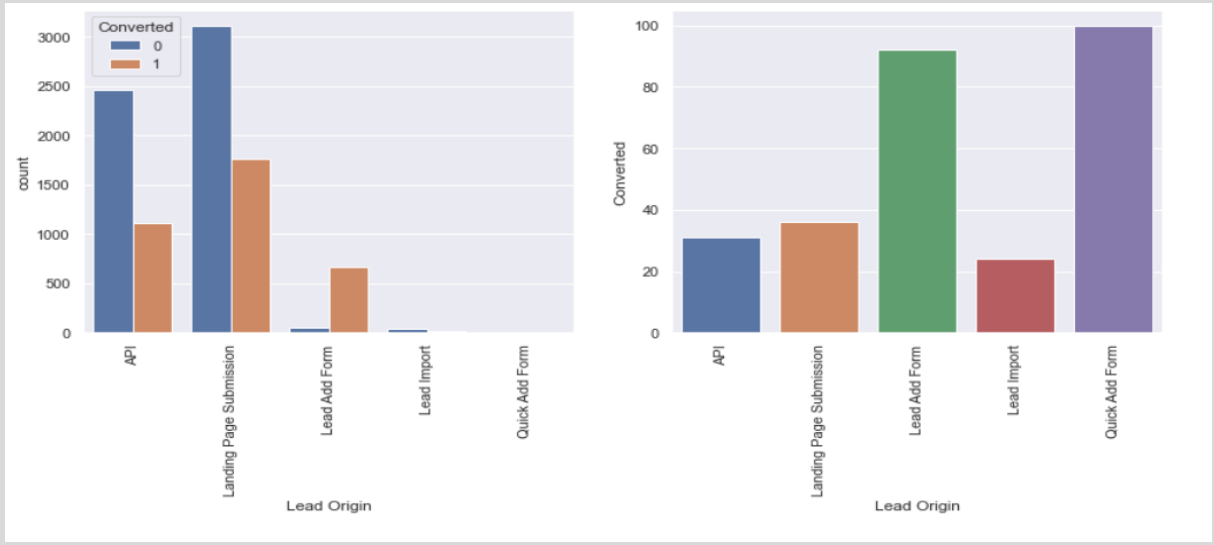
- Dual Specialization students and Lateral students have a very high conversion rate, hence emphasis should be made to acquire more such individuals
- Working Professionals have a higher conversion rate than the non-working individuals therefore efforts should be made for reaching out to more working professionals

Bi-variate Analysis-



- **Lead Source**- Google and Direct traffic generates maximum number of leads. Conversion rate of reference leads and leads through the Welingak Website is high.

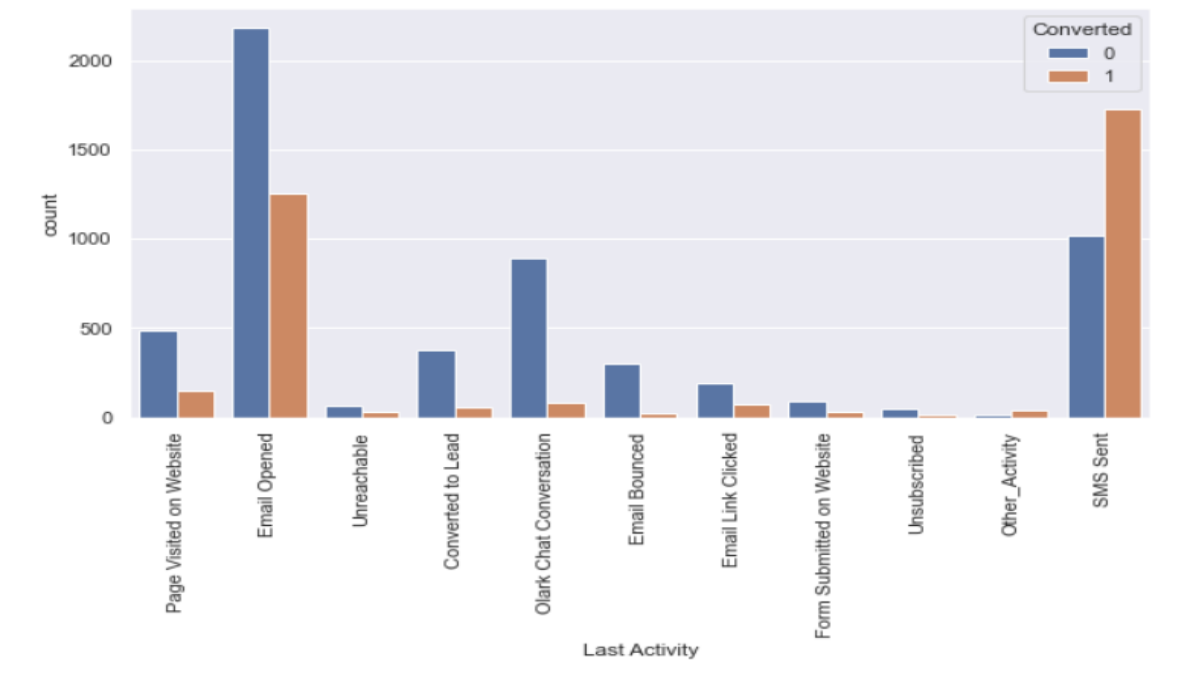
-To improve the overall lead conversion rate, focus should be on improving the lead conversion of Olark chat , organic search , direct traffic and google lead and generate more leads from reference.



Lead Origin – We can see API and Landing Page Submission have **30- 35%** conversion rate, But count of lead originated from them are considerable.

-**Lead Ad Forms –** has more than 90% conversion rate but lesser count values.

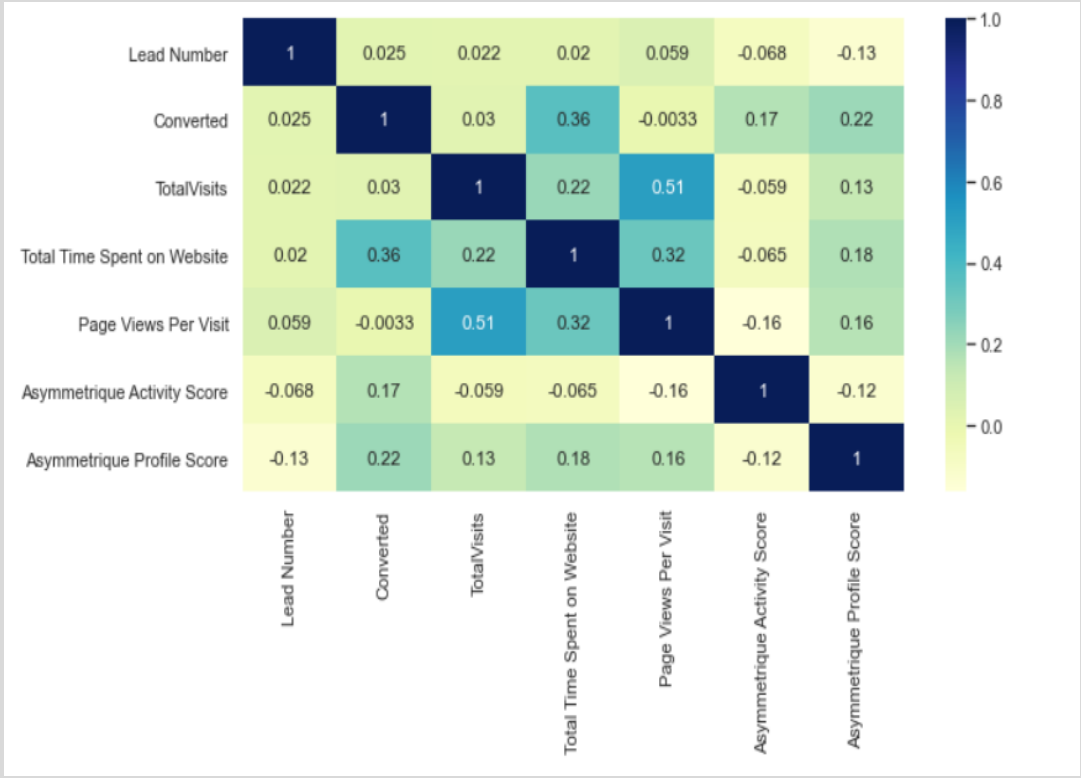
-In order to improve overall lead conversion rate we need to focus on improving lead conversion rate of API and Landing Page Submission and generated leads from Lead add form.



Lead Activity –

Most of the leads had their email opened as their last activity so a suggestion could be to send reminder emails so they can view it when they visit their inbox.

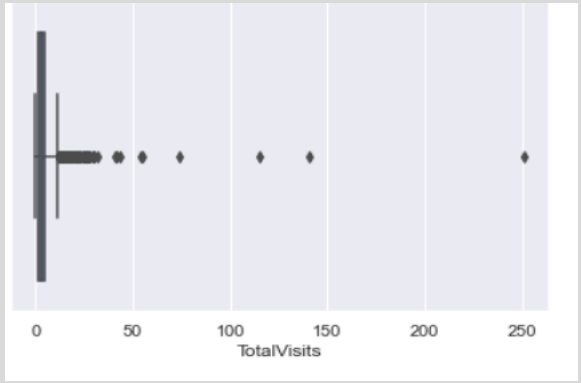
Conversion rate for leads with the activity as last SMS sent is much higher which is around 60%.

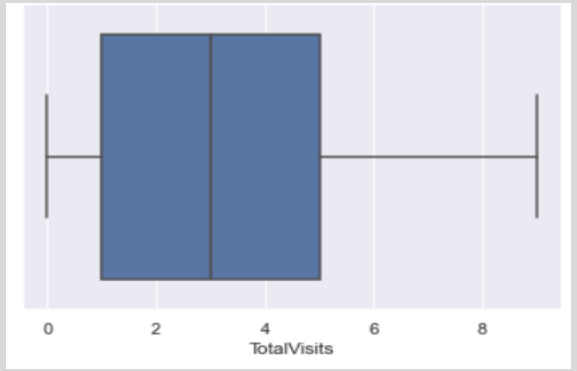


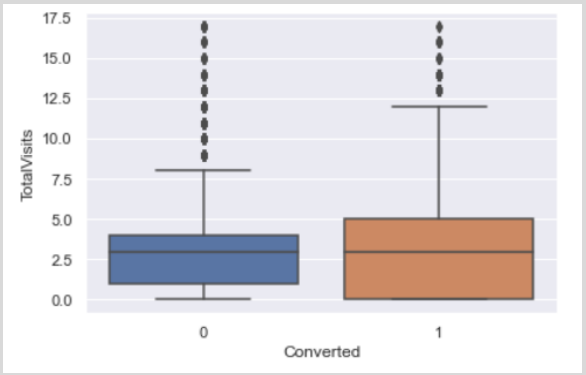
Correlation analysis(Heatmap) Conclusion- There is a positive correlation between the Total Time Spent on the Website and Conversion. We can also see correlation between Conversion and some categorical columns like Lead Origin and Lead Source .

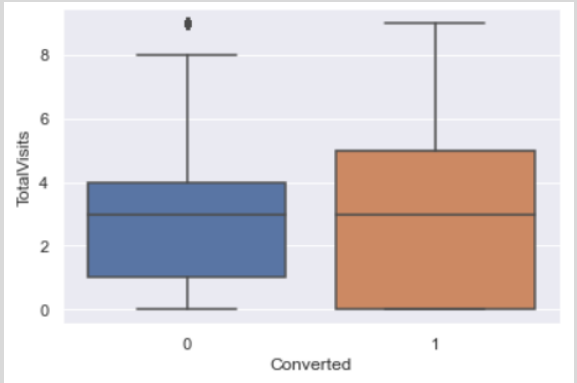
There is almost no correlation between Page Views Per Visit and Total Visits with Conversion.

Outlier Analysis -



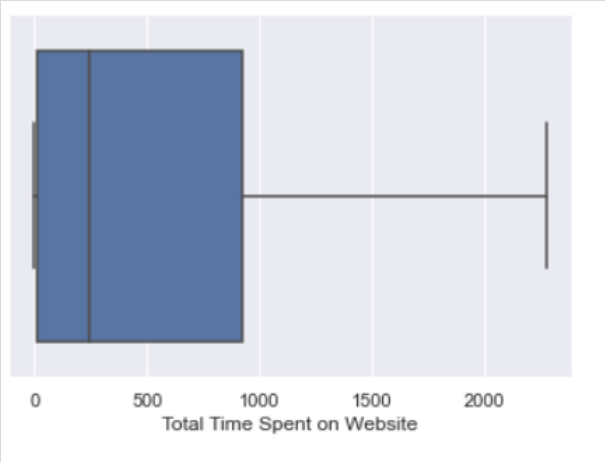


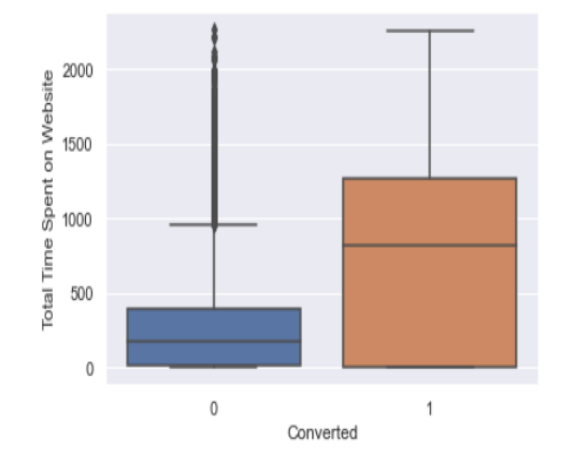


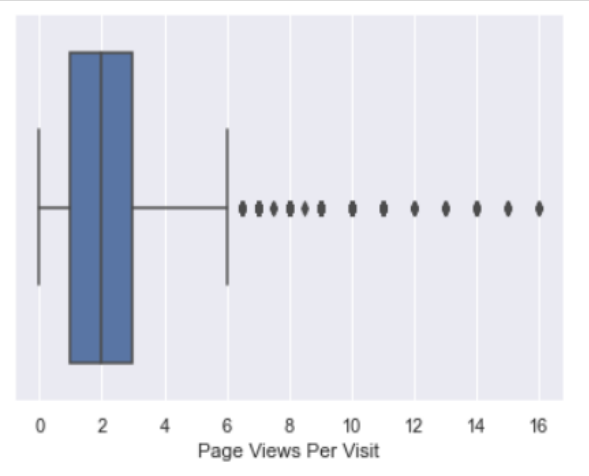


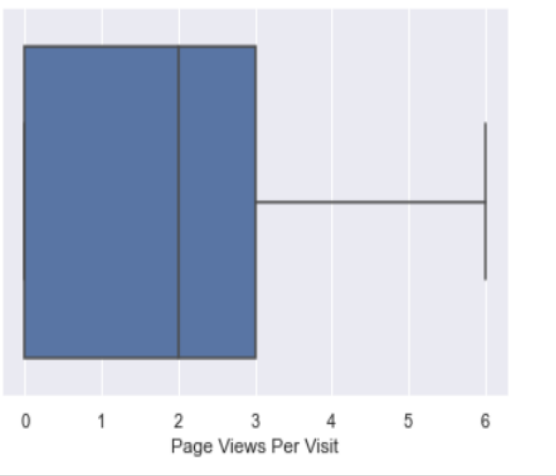
•**Total Visits, Total Time spent on website, Page per Visit**- We perform made more engaging to make these leads spend more time on the website . snippet from the assignment has been shown below – Before and after removal of the outliers.

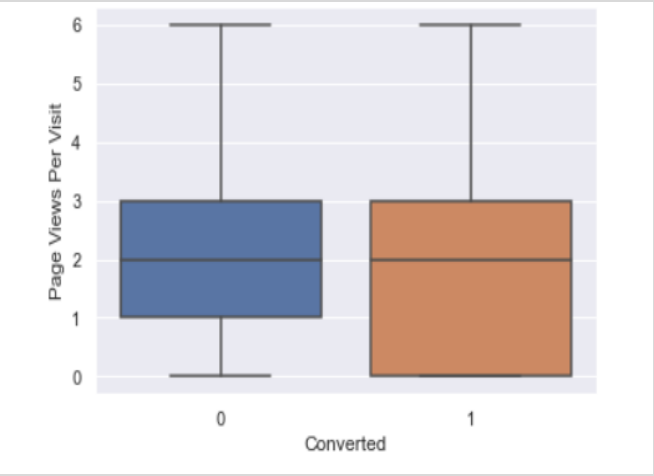
There are outliers present in the total visit columns, removing the outliers by capping the max value at 99 percentile.


0
500
1000
1500
2000
Total Time Spent on Website


Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted


0
2
4
6
8
10
12
14
16
Page Views Per Visit


0
1
2
3
4
5
6
Page Views Per Visit


Page Views Per Visit
6
5
4
3
2
1
0
0
1
Converted

Model Building –

- Splitting the Data into Training and Testing Sets .
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection - Running RFE with 15 variables as output .
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5.
- Predictions on test data set .
- Overall accuracy is then calculated .

RFE Feature Selection

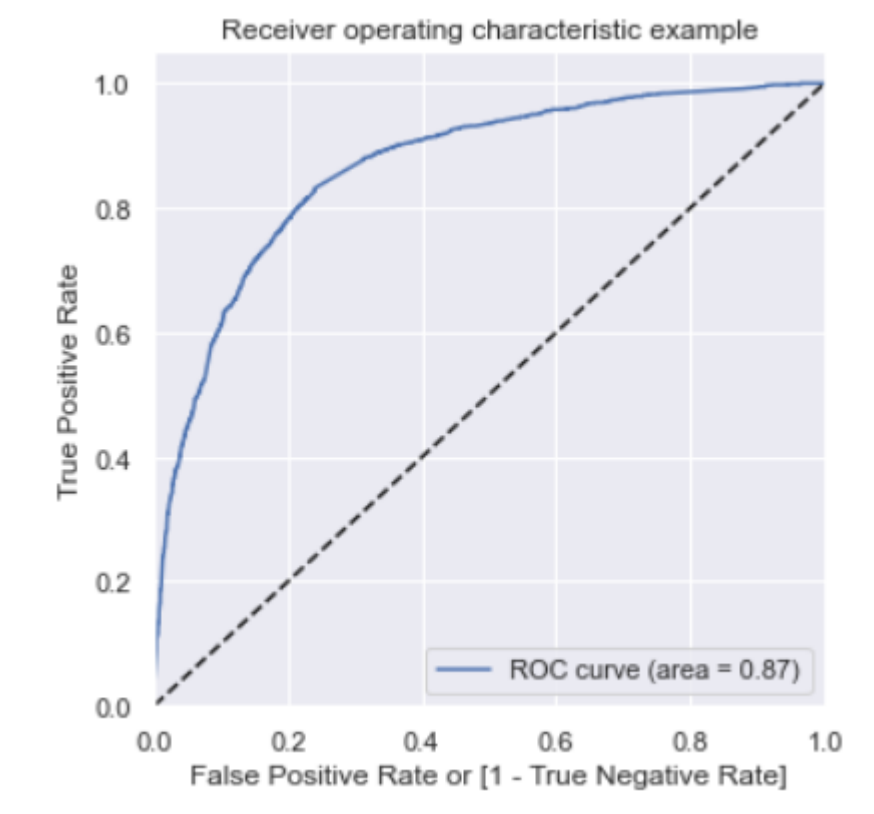
Equation for **'Converted'** from our RFE model building coupled with manual feature elimination is:

-1.9137 + 1.0914 X **Total Time Spent on Website** +1.0549 X **Lead Origin_Lead Import +**1.1609 X **Lead Source_Olark Chat +** 4.2997 X **Lead Source_Reference +** 5.6734 X **Lead Source_Welingak Website**
-1.3530 X **Last Activity_Email Bounced +**0.7422 X **Last Activity_Email Opened** -0.9704 X **Last Activity_Olark Chat Conversation +**2.3747 X **Last Activity_Other_Activity +**0.8380 X **Last Activity_SMS Sent**
+1.3305 X **Last Notable Activity_SMS Sent +**2.7214 X **Last Notable Activity_Unreachable**

Lead Source_Reference, **Lead Source_Welingak Website and Last Notable Activity_Unreachable** are the variables in the model which should be focused the most on in order to increase the probability of lead conversion.

ROC Curve –

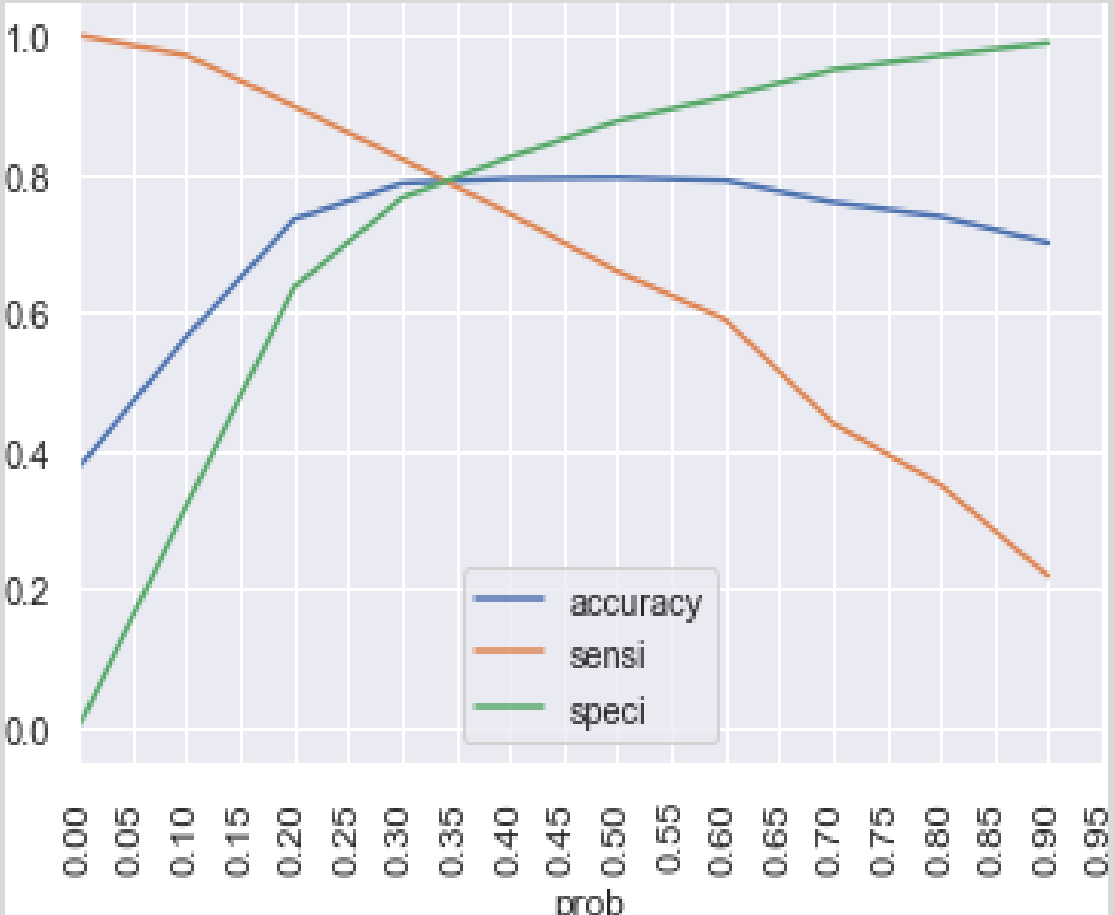
An **ROC curve** (**receiver operating characteristic curve**) is a **graph** showing the performance of a classification model at all classification thresholds. This **curve** plots two parameters: True Positive Rate. False Positive Rate.



ROC curve from shows trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

As we see. The curve looks closer to the left-hand border and the top border of the ROC space. This shows the accuracy of our model, also the area under the ROC curve is 0.87.

Optimal cut-off point–

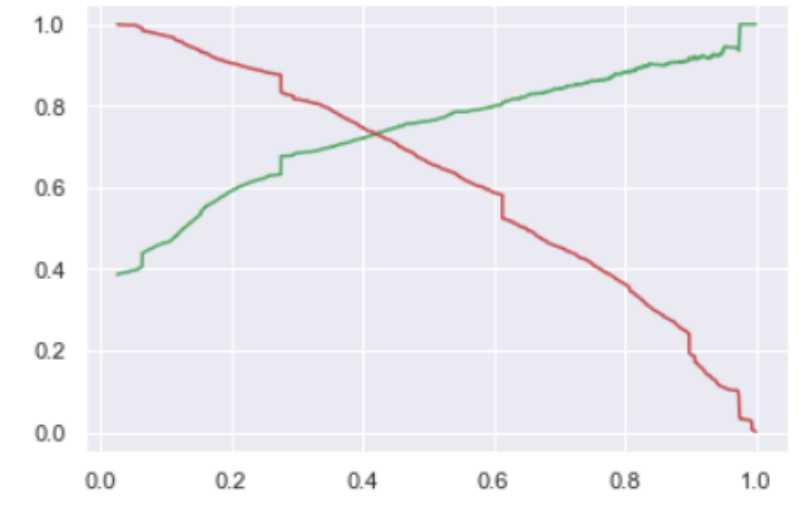


As we can see, when the probability thresholds are very low, the sensitivity is very high and specificity is very low. Similarly, for larger probability thresholds, the sensitivity values are very low but the specificity values are very high. And at about 0.34, the three metrics seem to be almost equal with decent values and hence, we choose 0.34 as the optimal cut-off point.

Precision Recall Trade-off plot–

From the graph we can see the cut-off point is taken as 0.4.

precision-**recall** tradeoff occur due to increasing one of the parameter (**precision** or **recall**) while keeping the model same.



Model Analysis

Performance of our Final Model

Overall accuracy on Test set: 0.795

Sensitivity of our logistic regression model: 0.749

Specificity of our logistic regression model: 0.832

Inferences from Model

Business Insights Derived from our Model

Top 3 variables in model, that contribute towards lead conversion are:

1. Lead Source_Welingak Website
2. Lead Source_Reference
3. Last Notable Activity_Unreachable

Recommendation -

The X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- Phone conversations also seem to get more leads .
- Increase sending SMS notifications since this helps in higher conversion.
- Get Total visits increased by advertising and campaigns since this helps in higher conversion.
- Improve the Olark Chat service since this is affecting the conversion negatively.
- Emphasis on targetting individuals whose last notable activity was visited website .
- Increase user engagement on their website since this helps in higher conversion.

Conclusion (LR Model)

Our Logistic Regression Model is decent and accurate , when compared to other models, with 79.5 % Accuracy on Test Set, 74.9 % Sensitivity and 83.2 % Specificity.
We can vary these parameters by varying the cut-off value and thus predict leads based on scenarios like availability of extra resources and vice-versa.