LEAD SCORE CASE STUDY

Presented By:

ACHINTYA KUMAR DUTTA
SANDEEP KUMAR PALIT

PROBLEM STATEMENT

- 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’.
- 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.

- **Business Objective:**
- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

SOLUTION METHODOLOGY

- **Data Cleaning and manipulations:**
  - Check and handle the duplicate data
  - Check and handle upper case and lower case issue
  - Handling the columns containing large number of Nulls
  - Handling the columns which have no/least variance of data
  - Imputation of the data wherever necessary
  - Check and handle the outliers
- **EDA**
  - Data binning
  - Plotting of categorical and numerical variable in Univariate analysis
  - Bivariate analysis to understand patterns within the data

SOLUTION METHODOLOGY

- **Data Preparation for modelling:**
  - Conversion of the categorical to numeric features like Do Not Email converted to 0 or 1.
  - Creation of Dummy variables for categorical variables
  - Drop the original features after dummy creation
  - Feature scaling
  - Correlation check and dropping of highly correlated features
- **Test Train Split:**
  - Splitting the data into train and test sets
- **Model Training:**
  - Recursive Feature Elimination, here we do used RFE to use a course number of features as a start and then for each execution, removed the features based on high p value and high VIFs.

SOLUTION METHODOLOGY

- **Model Training (continue…)**
  - We have used rfe score to decide the features and then reduced from there
  - In each run we have checked the accuracy score of the model and compared with the previous score.
  - We have continued this recursion till the p value of the feature has been been reduced to negligible value and the vif’s are reduced to near 1.
  - Finally after deciding the features through multiple iterations we have used confusion matrix to calculate the True Positive, True Negative, False Positive and False Negative scores to give decide the Sensitivity and Specificity.
  - We have plotted the ROC curve to see how the model performs for predicting the data or in other words it shows the trade-off between the Specificity and Sensitivity, where more the crve towards the upper and left corner and more the AUC the better the model is.

SOLUTION METHODOLOGY

- **Model Training (continue…)**
  - We have plotted the graph for the sensitivity, specificity and the accuracy to get the optimal cut-off point.
  - Based on this cut-off we have applied our probability and checked the accuracy score again.
  - Finally we have also checked the Precision and the Recall.
- **Applying the model on Test Data set:**
  - Prepared the Test Data set with the set of rfe columns
  - Predicted on the test data set
  - Add the leadID from index
  - Concatenated the test and predicted dataset
  - Rearranged the columns and mapped the final predicted values based on cutoff value

SOLUTION METHODOLOGY

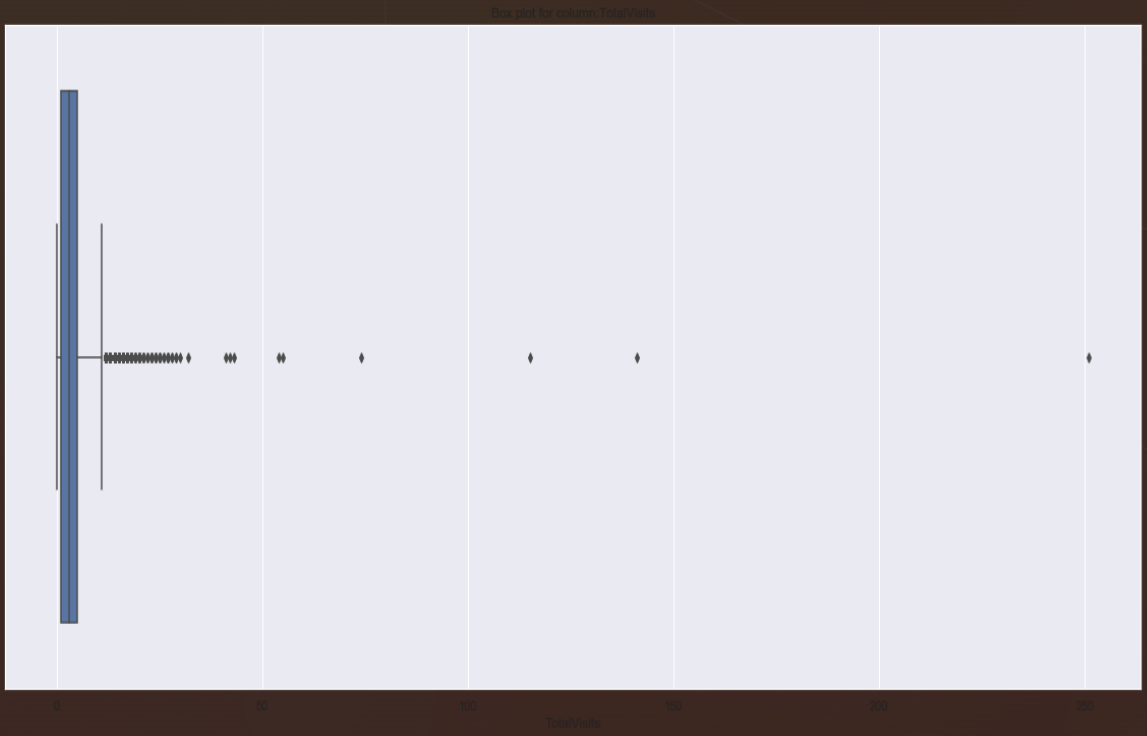
- **Applying the model on Test Data set (continue …)**
  - Calculated the accuracy, sensitivity , specificity of the model execution on the test dataset
  - Plot the ROC Curve
  - Generated the classification report
  - Calculated the Cross Validation Score to see how our model performs.
- **Other activities based on the business needs**
  - Calculated the Lead Score on the dataset
  - Determined the feature importance
  - Checked the percentage of the conversion based on the management requirement.

DATA MANIPULATION

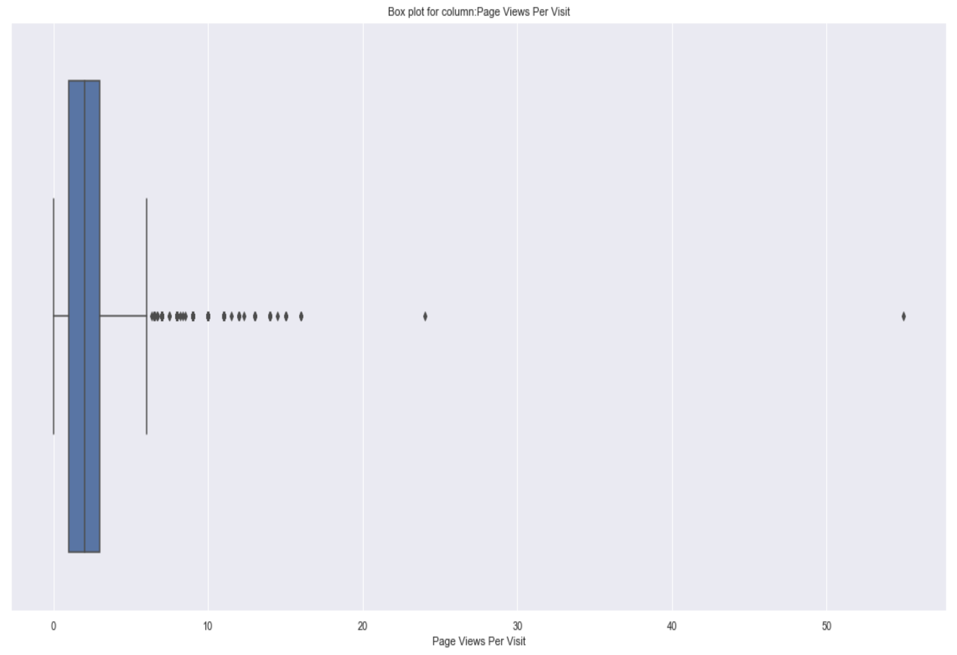
- Total Number of rows- 9240 and 37 columns.
- Checked the percentage of the Nulls for each columns and dropped those which has more than 30% of Nulls, e.g. - Lead Quality, Asymmetrique Activity Index, Asymmetrique Profile Score, Asymmetrique Activity Score, Tags, Lead Profile, What matters most to you in choosing a course.
- Dropped few columns which have least variations where either all the records or negligible amount of records contains a different value.
- Handled the columns which have Nulls as well as ‘Select’ as a value. Removed those where accumulated % of Nulls and Select values are more than 30%.
- Imputed some columns like ‘Page Views Per Visit’ which have Nulls with mean value.

OUTLIER ANALYSIS

Box Plot for TotalVisits

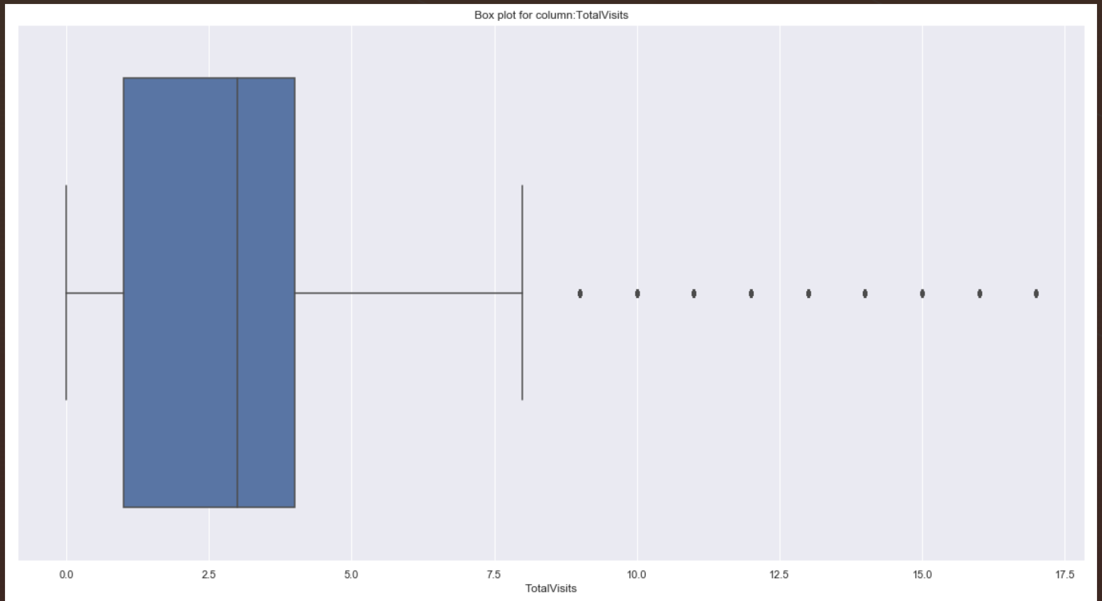


Box plot for Page Views per visit

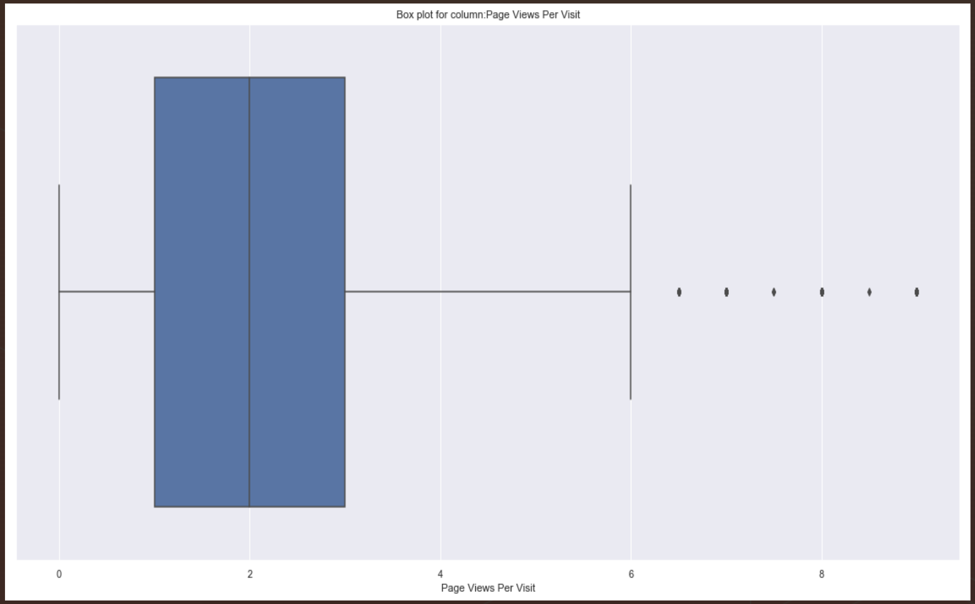


OUTLIER ANALYSIS

Totalvisits after Outlier handling

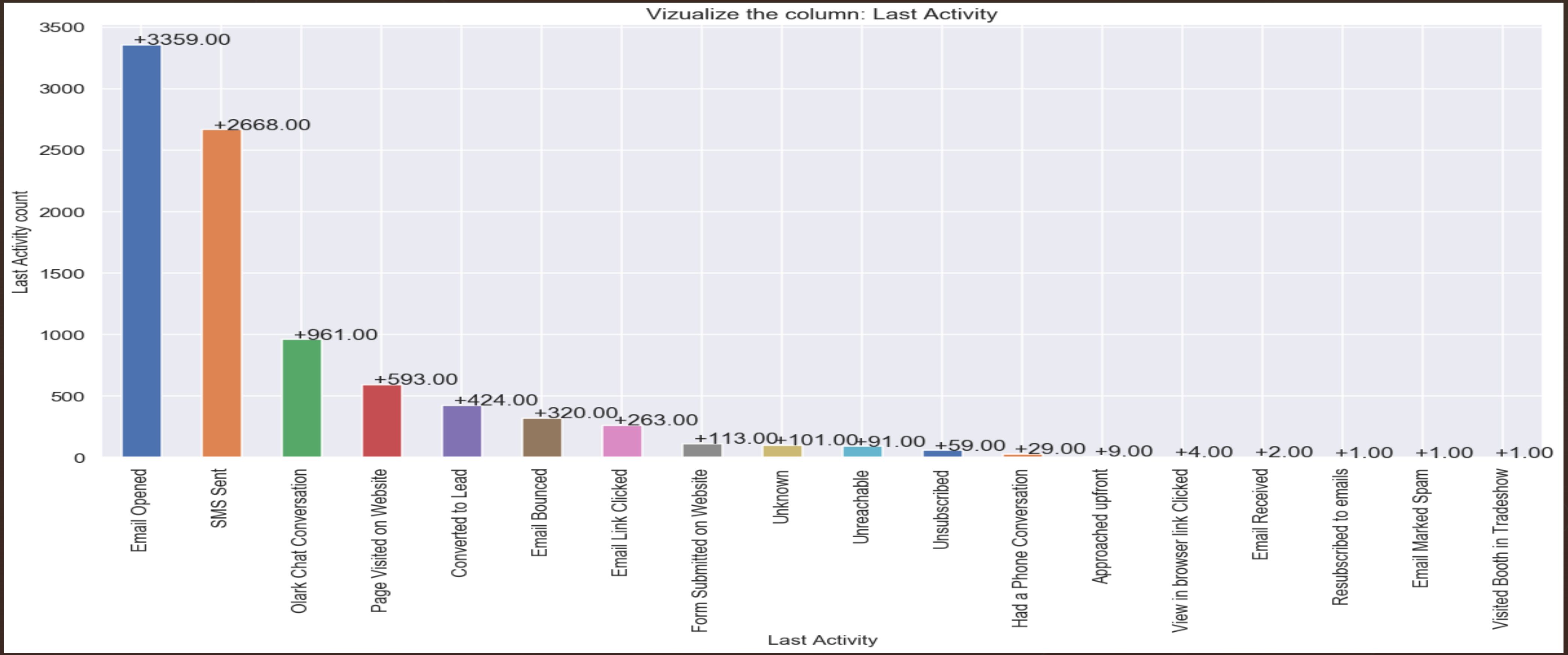


Page Views per visit after outlier handling



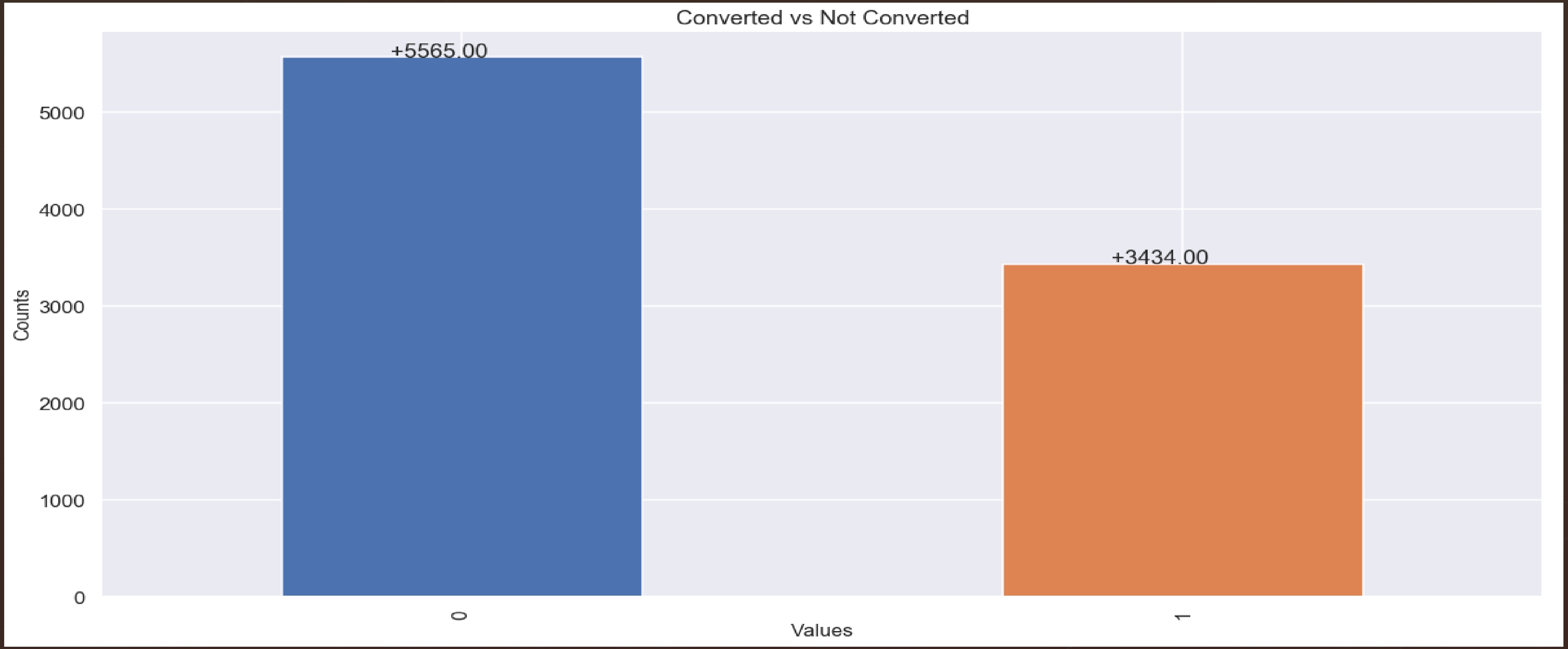
EDA

Visualization of Last Activity



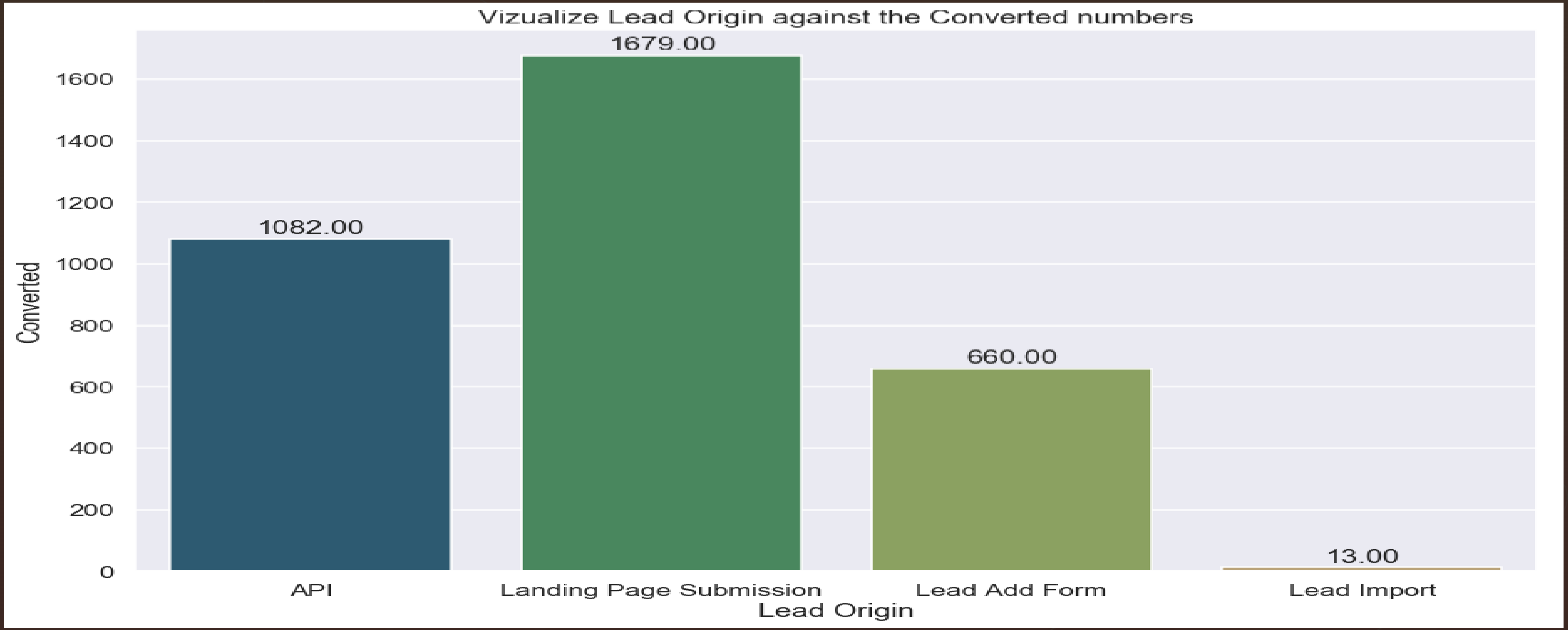
EDA

Converted/Not Converted



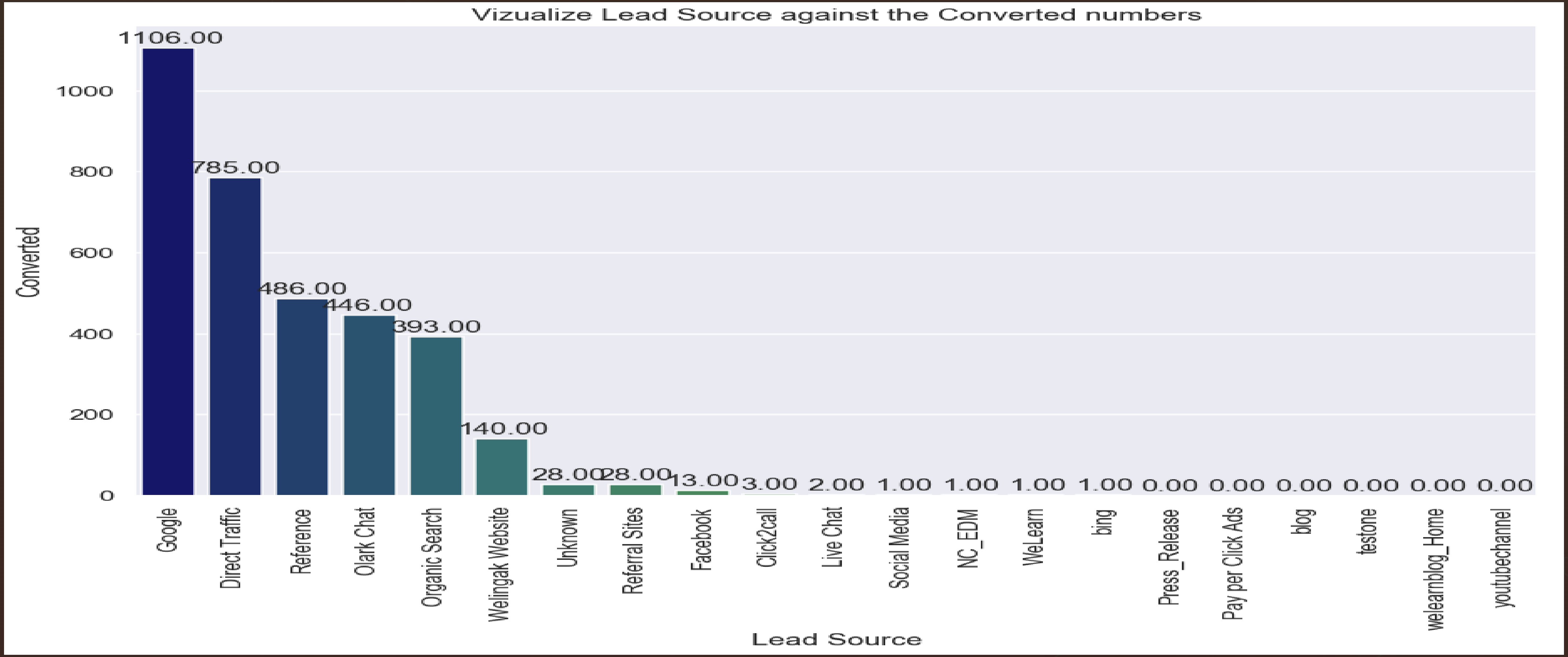
EDA

Lead Origin against Converted



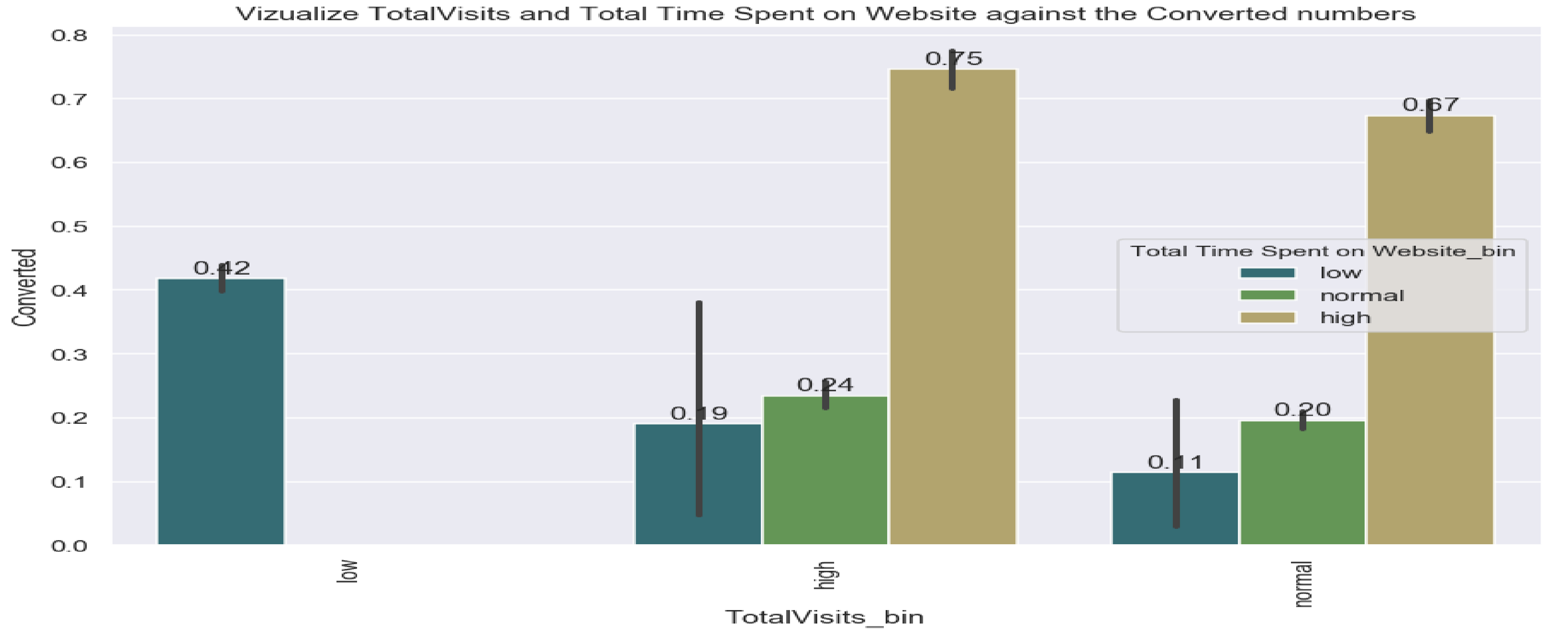
EDA

Lead Source against Converted



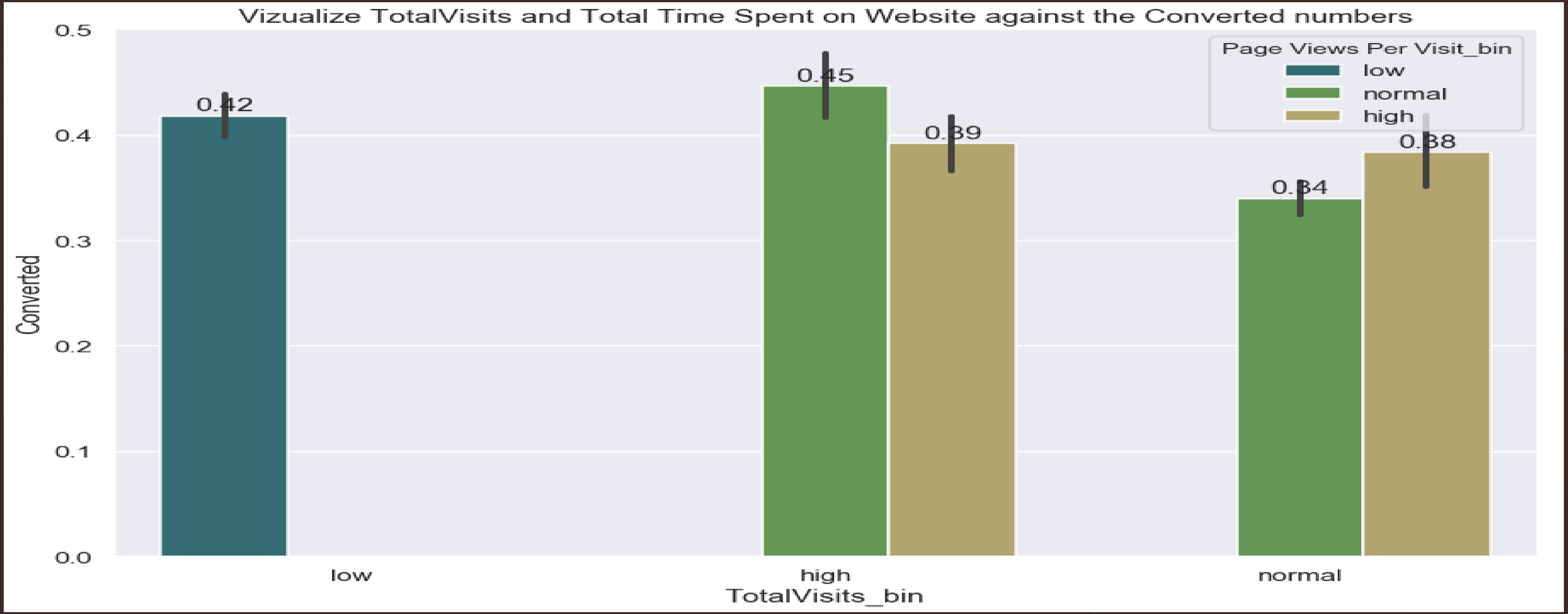
EDA

Total Visits, Total Time spent against the Converted numbers



EDA

Total Visits, Page views per visits against the Converted numbers



MODEL BUILDING

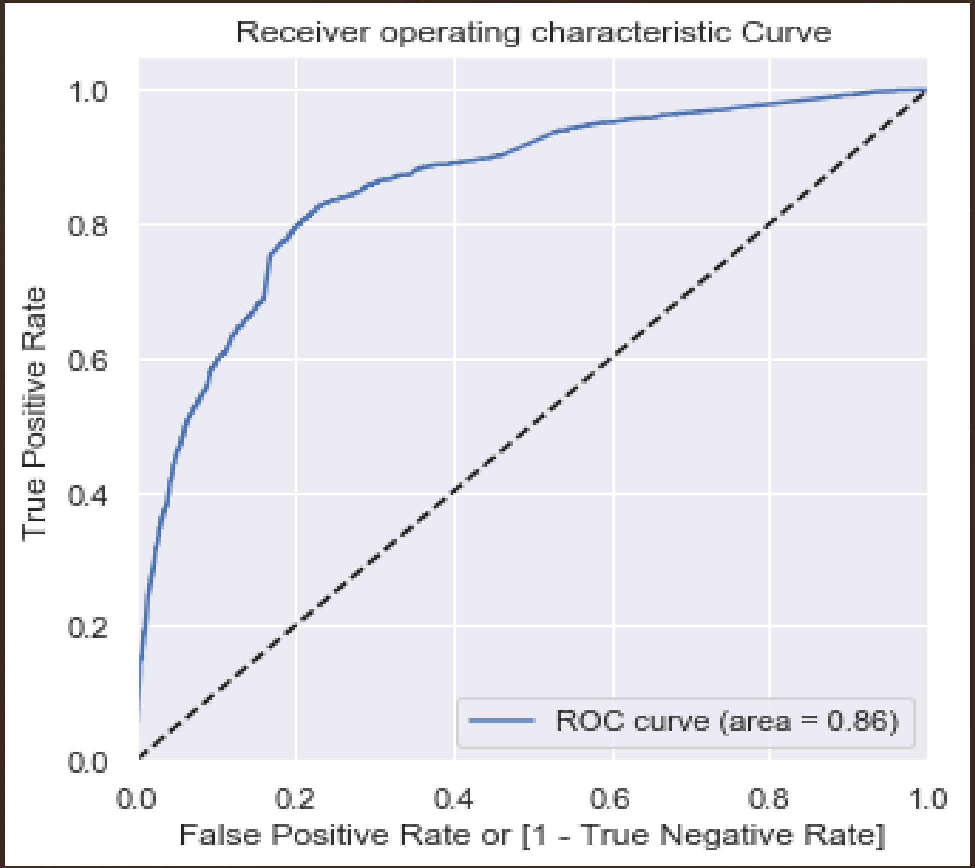
- Splitting the Data into Training and Testing Sets
- As you know, the first basic step for regression is performing a train-test split, we have chosen 70:30 ratio. Use RFE for Feature Selection
- Running RFE with 13 variables as output
- Building Model
- Removing the variable whose p- value is greater than 0.05 and vif value is greater than 2
  PREDICTIONS ON TEST DATA SET
  Model accuracy on test data - 78%, after cross validation we found that the mean accuracy is 80%.

MODEL RESULT

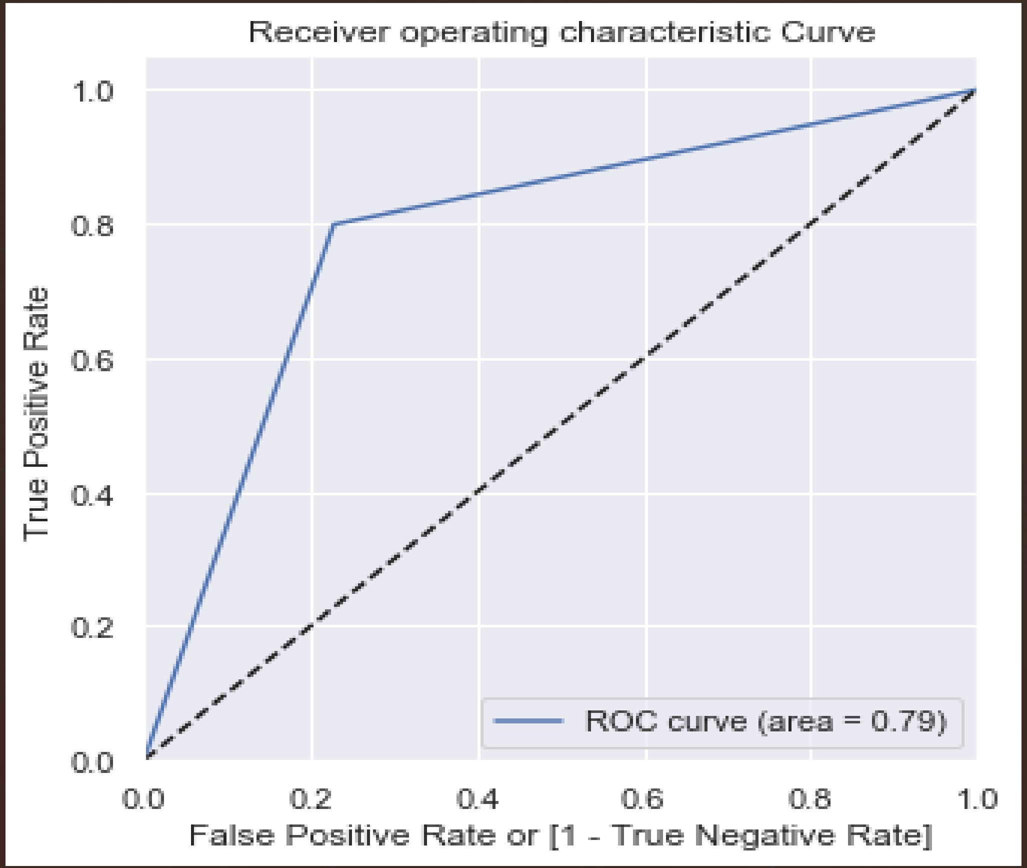
- **Accuracy of the Train dataset:** 0.79
- **Sensitivity of the Train dataset:** 0.8
- **Specificity of the Train dataset:** 0.77
- **AUC of the Train dataset:** 0.86
- **Accuracy of the Test dataset:** 0.78
- **Sensitivity of the Test dataset:** 0.8
- **Specificity of the Test dataset:** 0.77
- **AUC of the Test dataset:** 0.79

ROC CURVES

ROC Curve for Train dataset

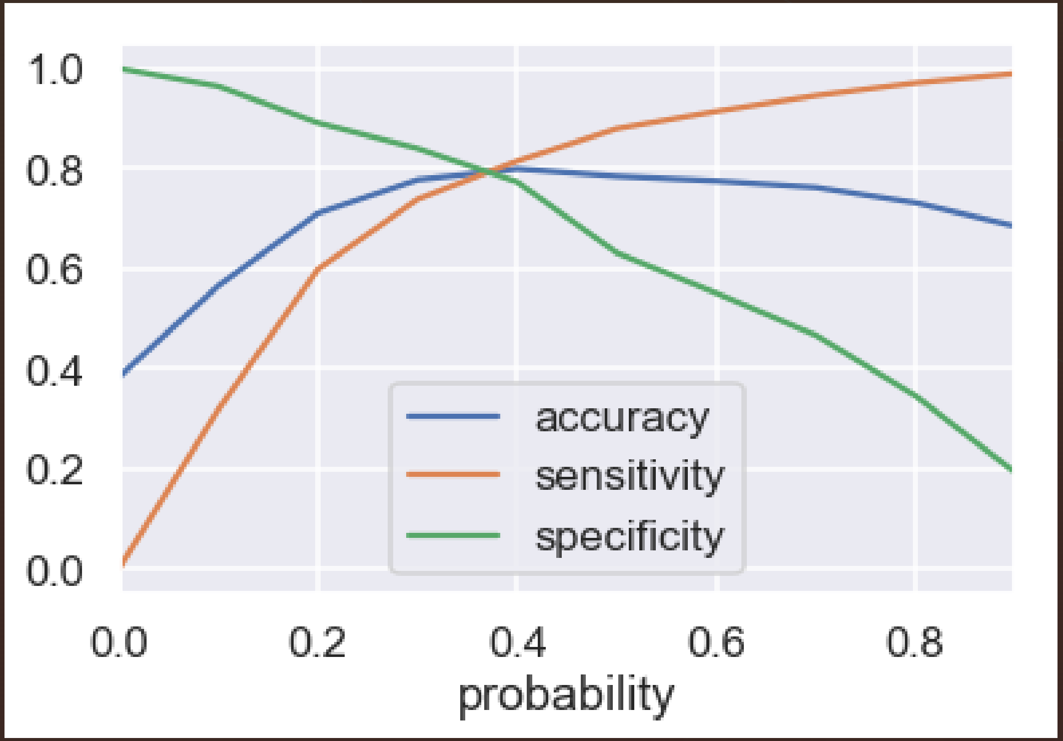


ROC Curve for Test dataset

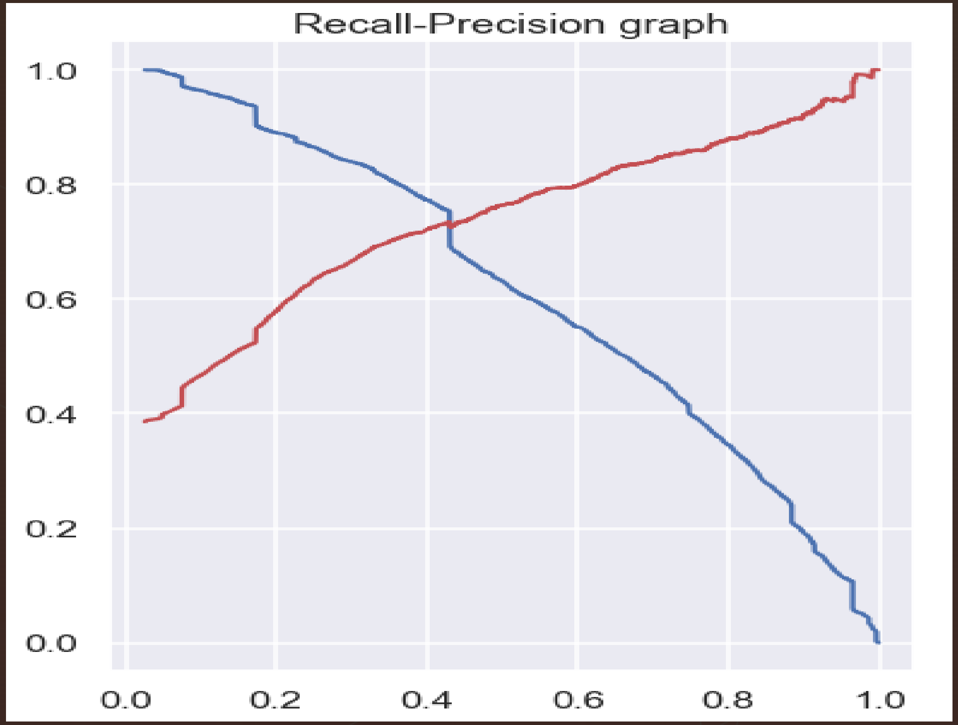


ROC CURVES

Optimal Cut-off using Accuracy, Specificity and Sensitivity



Optimal Cut-off using Precision and Recall



CONCLUSION

- All the identified variables have p values < 0.05
- All the identified variables VIF very low, so there is no multicollinerity among them < 0.05
- The overall accuracy is 0.78, which can be though as a accuracy good score
- Accuracy of the Train dataset: 0.79
- Sensitivity of the Train dataset: 0.8
- Specificity of the Train dataset: 0.77
- AUC of the Train dataset: 0.86
- Accuracy of the Test dataset: 0.78, after cross validation we found that the mean accuracy is 0.8, which can be though of a good score
- Sensitivity of the Test dataset: 0.8
- Specificity of the Test dataset: 0.77
- AUC of the Test dataset: 0.79
- Identified top 3 features with feature scores:
- Lead Source_Welingak Website - 100.00
- Lead Source_Reference - 61.55
- Last Activity_Had a Phone Conversation - 28.51

CONCLUSION

- Another point to note here is that, depending on the business requirement, we can increase or decrease the probability threshold value with in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.
- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.
- The model seems to perform well with ~80% accuracy and ~ 80% cross validation score, which can be thought a good performance for a model.