Lead Scoring Case Study

Problem Statement

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.
- Depending on the technical and business analysis we have to specify about the lead score assigned to each of the representative and also comment on the most useful features that were responsible in building the model.

Importing the libraries and dataset

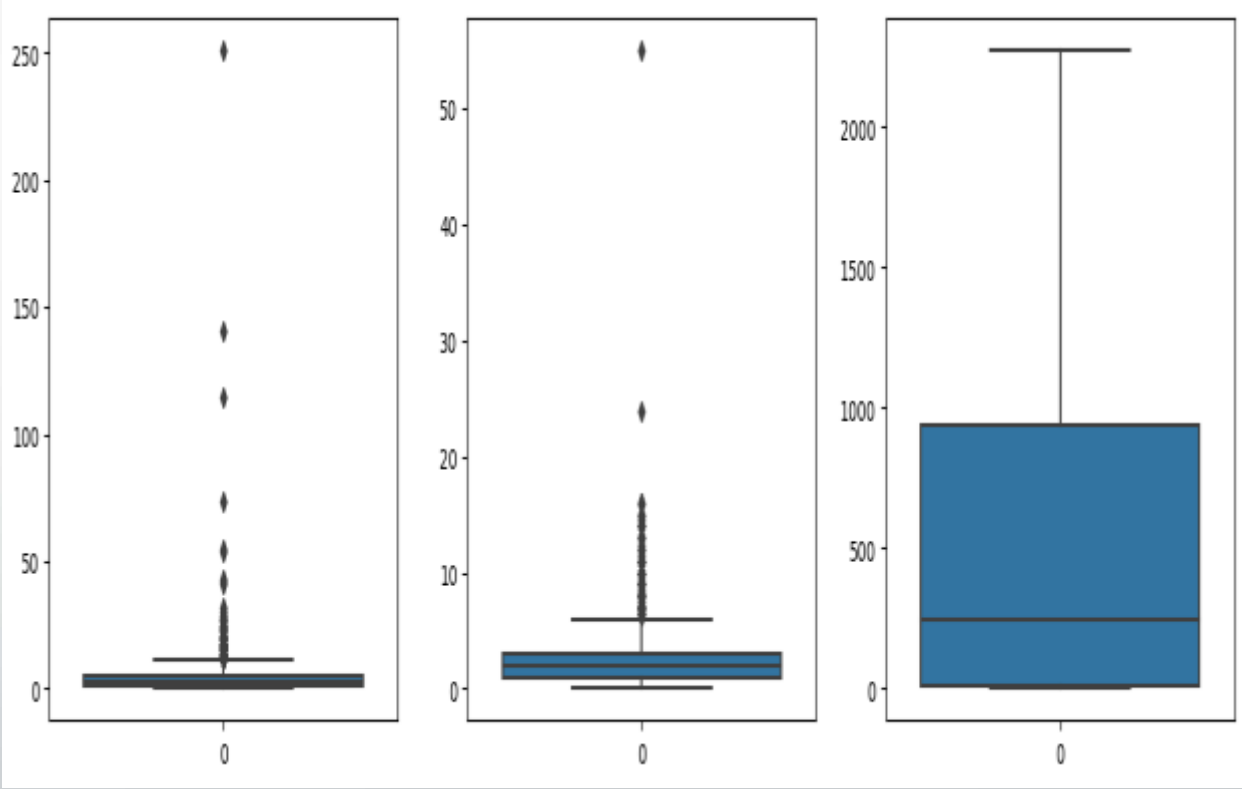
- Import the necessary libraries used for setting the data frame and for potting the graphs
- Import the library for Generalized Linear Model, Recursive Feature Elimination for removing the redundant features and stats model for adding the constant.
- Post this, import that data set that you want to analyse and perform the basic analysis in order to drop the columns that are not necessary
- After this, check regarding the shape, mean values by using the specific methods.

Dealing with the null values

- There are some values with “select” entry. We have deliberately replaced the select values with np.Nan to indicate it as a null value.
- The null values are then replaced with the mode and median of each of the columns
- The redundant columns which are not required are then dropped as they are having only one values
- There are some same words with capital as well as small letters they are replaced by a single word.
- At the end of this, we have all the columns filled with some or the other values and got rid of the null values present in the dataset.

Outlier Detection (Continuous variables)

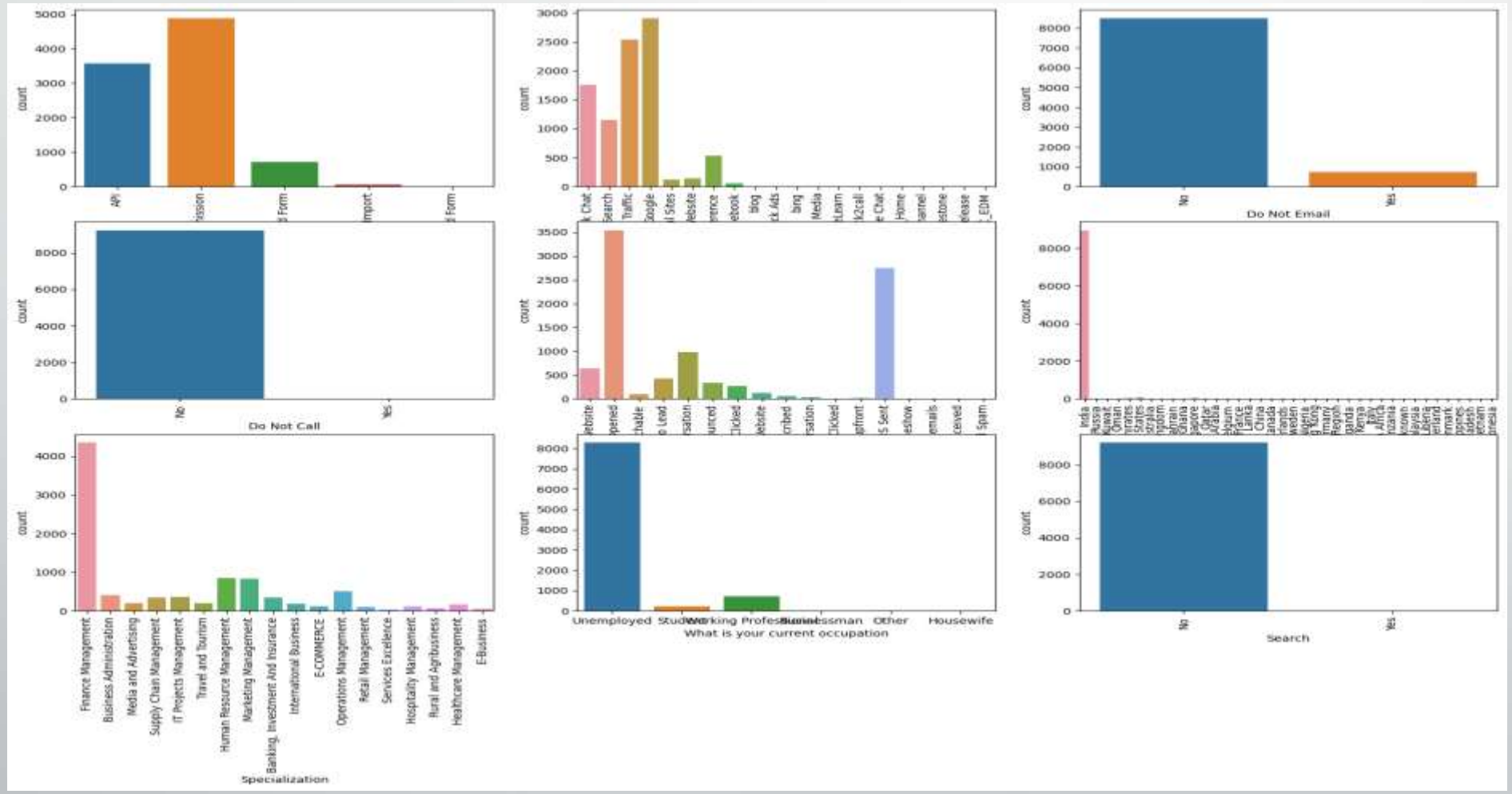
- There are some values that are at a very range which makes the dataset a bit vague.
- To handle these type of values we use a plotting technique known as Count plot and Box plot
- Box plot is used for checking the outliers in continuous values and Count plot for categorical variables.
- The median of the continuous variables is calculated and then the null values are replaced by this median



Outlier Detection (categorical values)

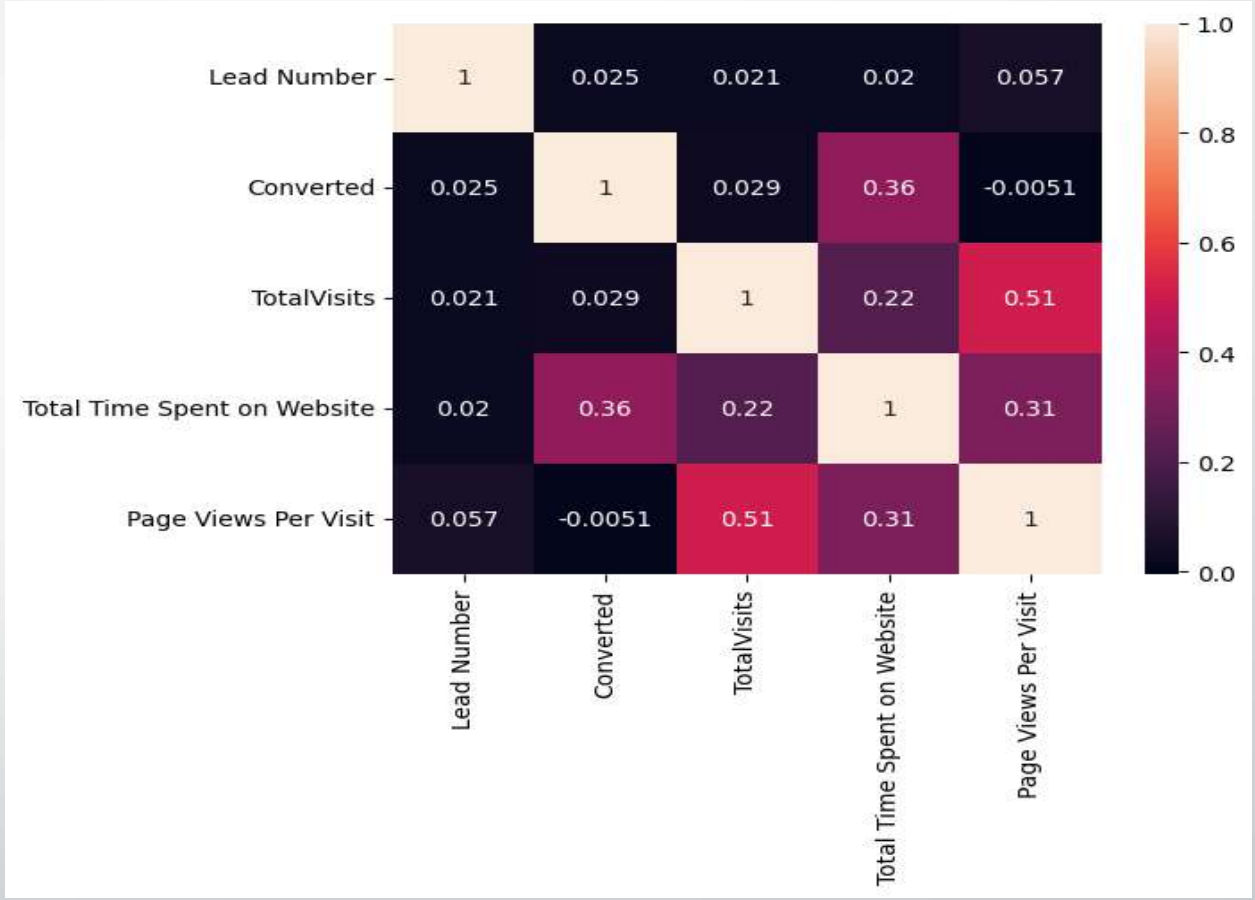
- Count plot is used for detecting and analysing the categorical variables which can be plotted even with the help of one variable present on X axis and other is the count on Y axis
- The columns that are present with only a single value and the columns with two values but the second value is as negligible as possible, these columns are dropped.
- This makes the data set more meaningful as we have only the important columns left with us.

Count Plot for the Categorical variables



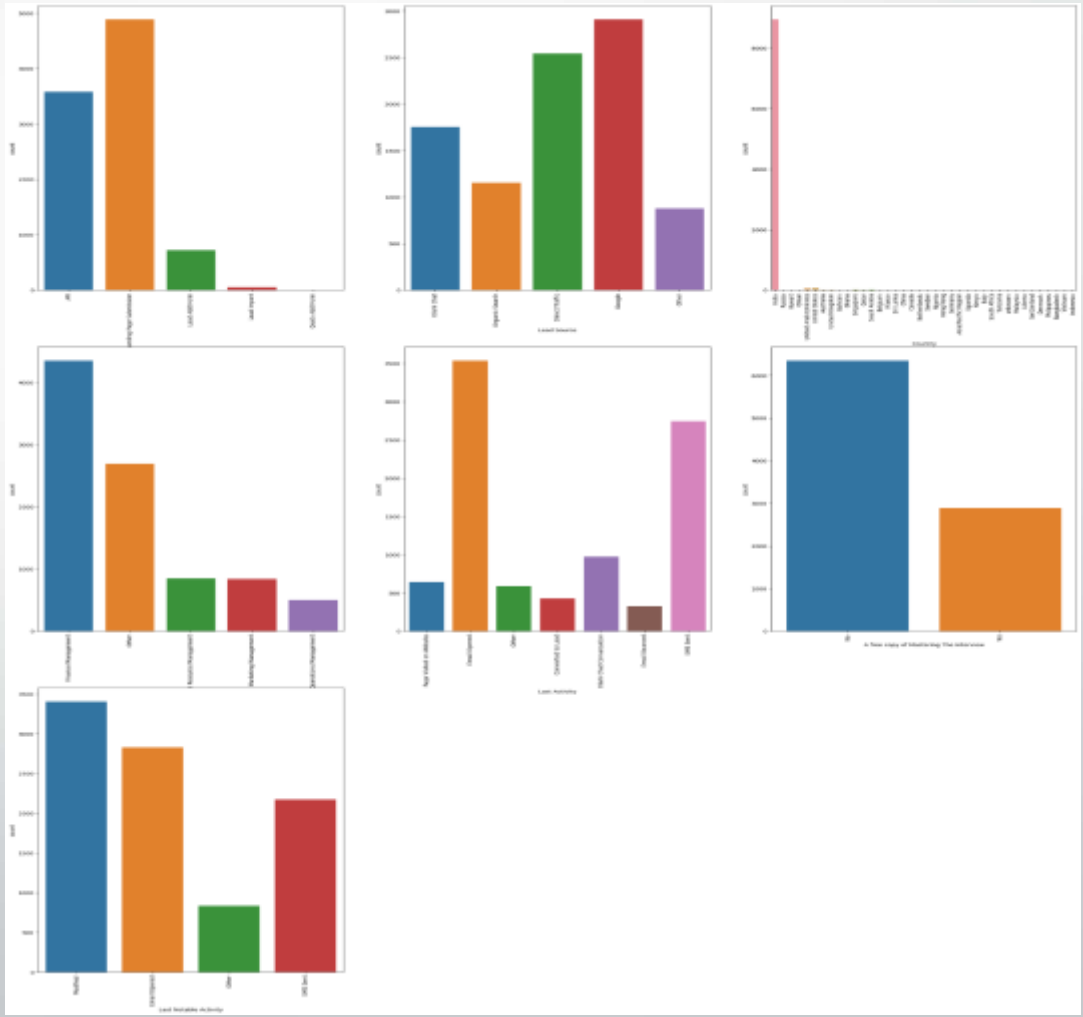
Plotting the correlations

- The columns that are highly corelated with each other are depicted by this plot.
- The columns that are highly corelated are dropped as these columns unbalances the data set.
- These are used for the continuous variables .
- After dropping all the columns we are left with the columns that contribute mostly towards the betterment of the model.
- We now have all the null values inspected and all the unwanted columns are dropped.



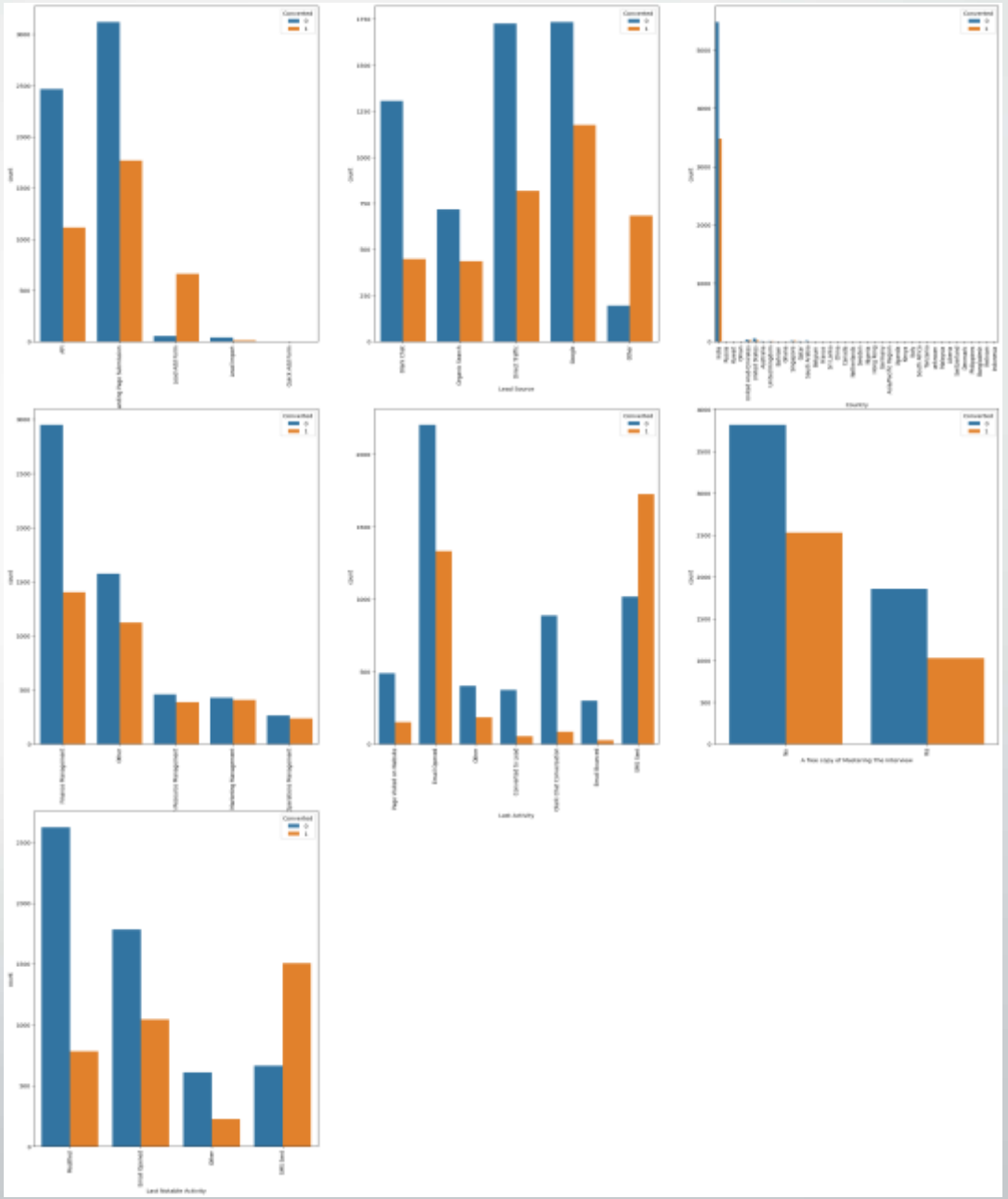
Univariate Analysis

- For the univariate analysis we use the count plot for depicting the analysis and visualizing the graphs.
- Only a single variable is used to visualize the outcomes and depending on the outcomes the insights are provided to the business.



Bivariate Analysis

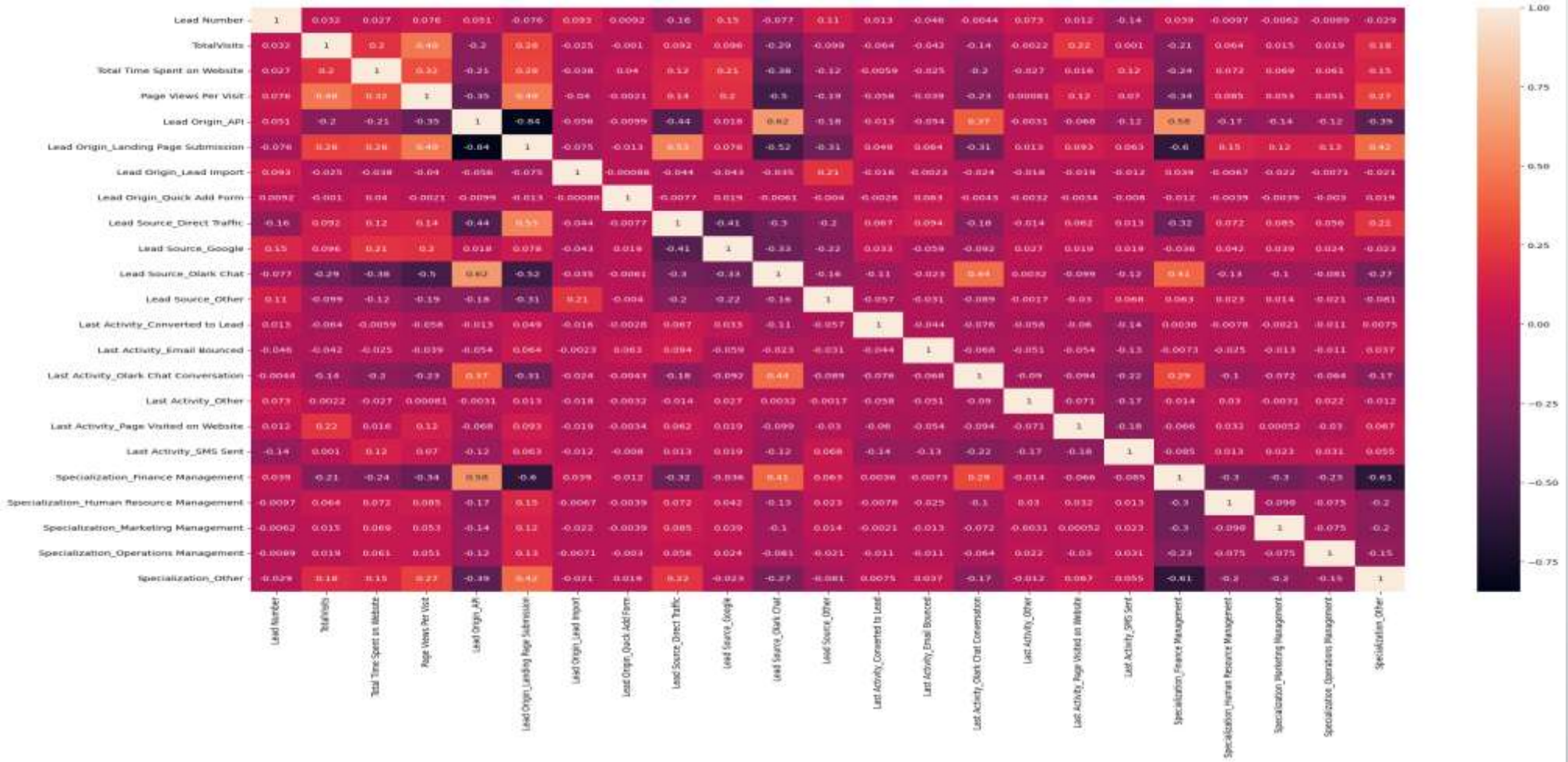
- Bivariate Analysis is same as the Univariate analysis but the only difference is that here two variables are used one is the target variable which is plotted on Y axis and other is the categorical variable plotted on X-axis.



Creating Dummy variables and splitting the dataset

- The leads data set contains a number of columns that have binary values 0 or 1.
- Also, these columns have various sections which can be depicted by creating dummy variables.
- These dummy variables represent each category of the columns.
- The data set is then separated into training set and test set with the help of train test split.

Correlation after the dummy variables



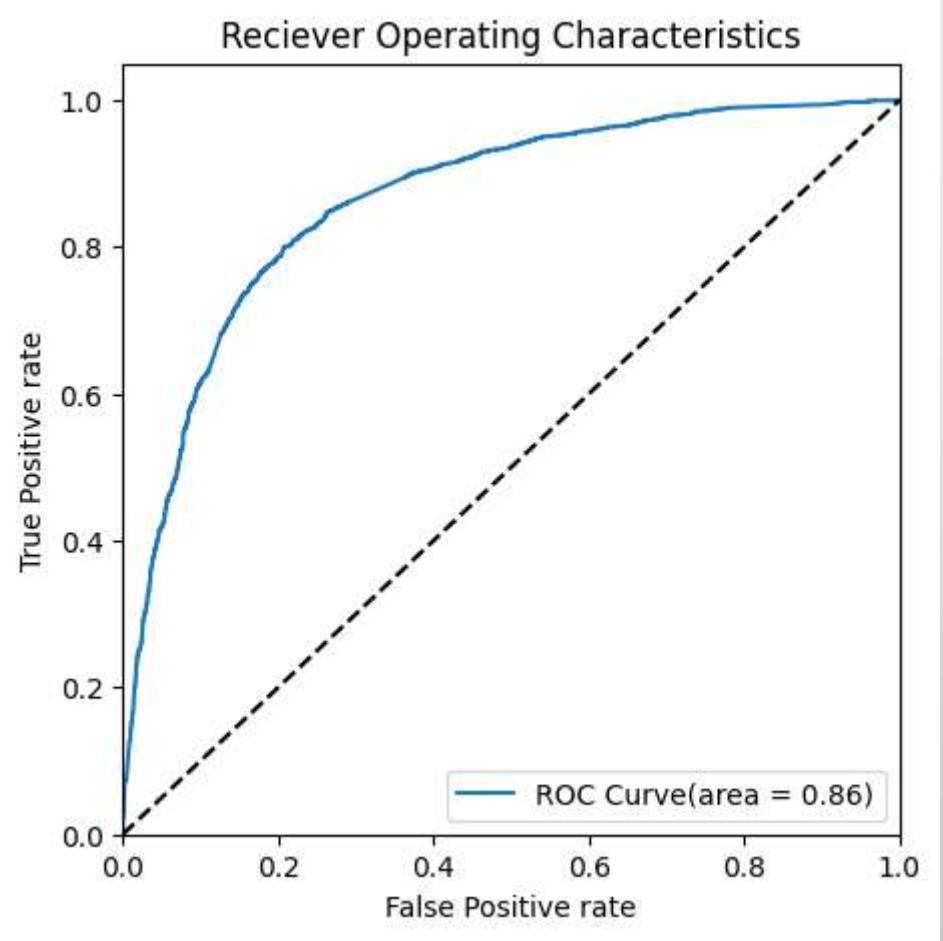
Building a logistic regression model

- The features that we have after performing the Exploratory Data Analysis are then passed through the Recursive Feature Elimination which provides a set of columns that are likely to contribute towards the model building.
- These features are then fitted to a Logistic Regression after adding constant to the dataset.
- Depending on the summary of the model and VIF calculations the columns with High VIF (Variance Inflation Factor) and High P-Values are discarded one by one each time after calculating both the parameters.
- This is how we obtain the features which best fit to the model depending on the VIF and P value. These two values should as small as possible.

Predicting the model on train set

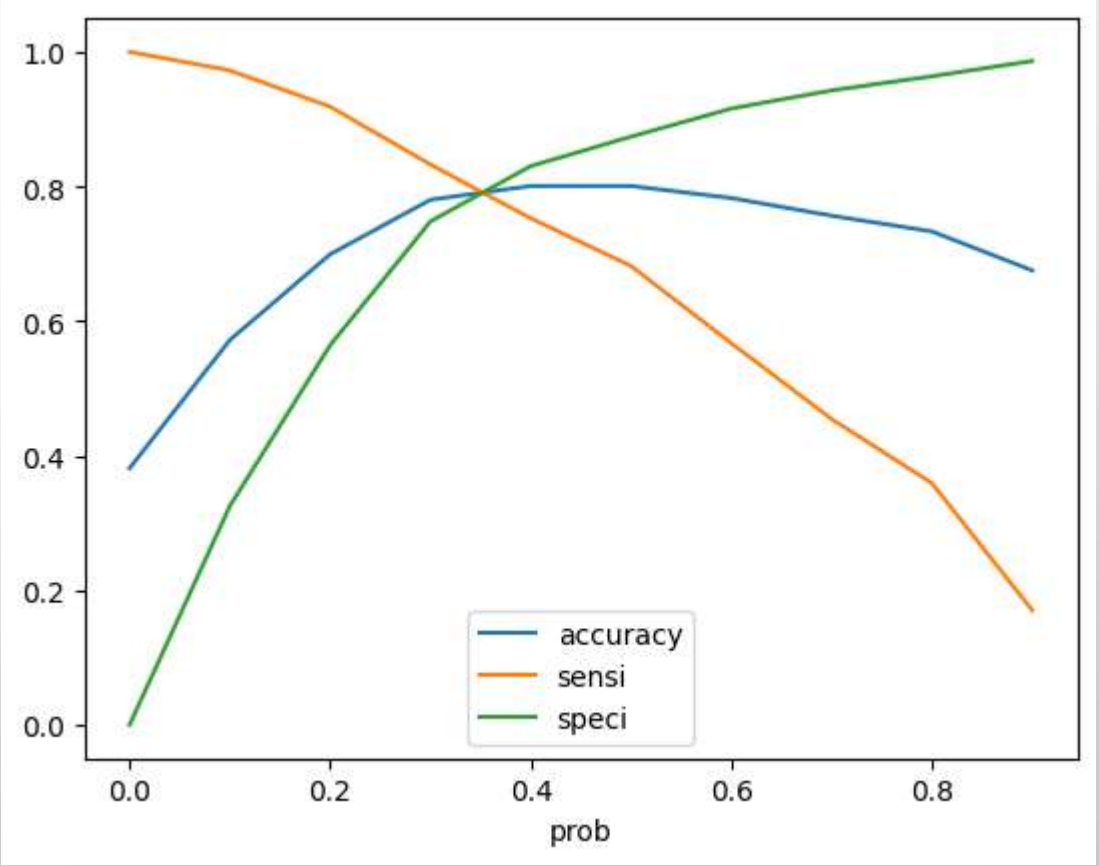
- The model that we have with us is then predicted over the train set and converted Probability is calculated.
- This converted probability is then predicted depending on the cut off value 0.5 by creating the data frame.
- Area of the curve tends more towards the upper left corner and hence the model is a best model.
- Also the model has a high area under the curve
- Accuracy: 80%
- Sensitivity: 68%
- Specificity: 87%
- True positive rate: 68%
- False Positive rate: 12%
- We have obtained a model where the FPR is less and TPR is more

Receiver Operating Characteristics Graph



Graph of Sensitivity Specificity and Accuracy

- This graph depicts the variation in Sensitivity, Specificity and Accuracy.
- As we can see from the graph, the Sensitivity and Specificity are inversely proportional to each other and as the probability increases the Accuracy increases to some extent and then remains stable.
- The cut off can be calculated from this graph which comes out to be 0.35



Revised metrices

The new metrices when calculated with the new cut off values are more accurate as the specificity has reduced which makes the model more efficient.

Accuracy: 79%
Specificity: 78%
Sensitivity: 8-%

The precision and recall values are also calculated
Precision: 68%
Recall: 80%

Predictions on Test Set

- The test set contains value that are out of bound, to convert those into a same range we scale the features using Standard Scaler which scales the data and transforms it
- The test data is then added with the constants to build the logistic regression model with the same instance created for the train set.
- A similar data frame is created where the probability of the Converted Predicted is calculated and depending on the new cut off value obtained, the Final Predicted values are calculated.
- The accuracy, sensitivity and Specificity for test data are calculated.
- Accuracy: 78%
- Sensitivity: 79%
- Specificity: 78%
- We can se that the model is doing good in test data as well
- Here the sensitivity tells that how actually converted and model predicted #them to be converted
- Our model is giving the 78% of accuracy where the actually converted members are predicted as converted

Assigning the lead score

- To assign the lead score, the two data frames are created one with predicted data on train set and other with the predicted data on test set.
- The lead score is assigned on the basis of the converted probability multiplied by 100

	Lead Number	Converted_Prob	Lead Score
0	5261	0.150310	15.03
1	2901	0.054709	5.47
2	6969	0.063946	6.39
3	1256	0.157724	15.77
4	1554	0.036980	3.70

Conclusion

- The most important features that helped in building the model were

- Total Visits
- Page Views per visit
- Total time spent on website
- Lead Origin Import
- Last Acitivity:Olark Chat conversation
- Lead source was: a. Google b. Direct traffic c. Organic search d. Olark Chat