Lead Score Model Case study

By
Sasanapuri Sravya

Problem Statement

- An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein you need to 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of Case Study

- 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.

Step 1:Importing the Necessary Libraries:

- Imported all the necessary libraries and warnings.

Step 2: Inspecting the Data:

- Observed the columns and duplicates in the dataset.

Step 3:Exploratory Data Analysis:

Data Cleaning:

—Identifying Missing Values:

- There are few columns with level called 'Select' which means that the customer had not selected the option for that particular column which is why it shows 'Select'. These values are as good as missing values and hence we will convert 'Select' values to Nan.

—checking percentage of null values in each column:

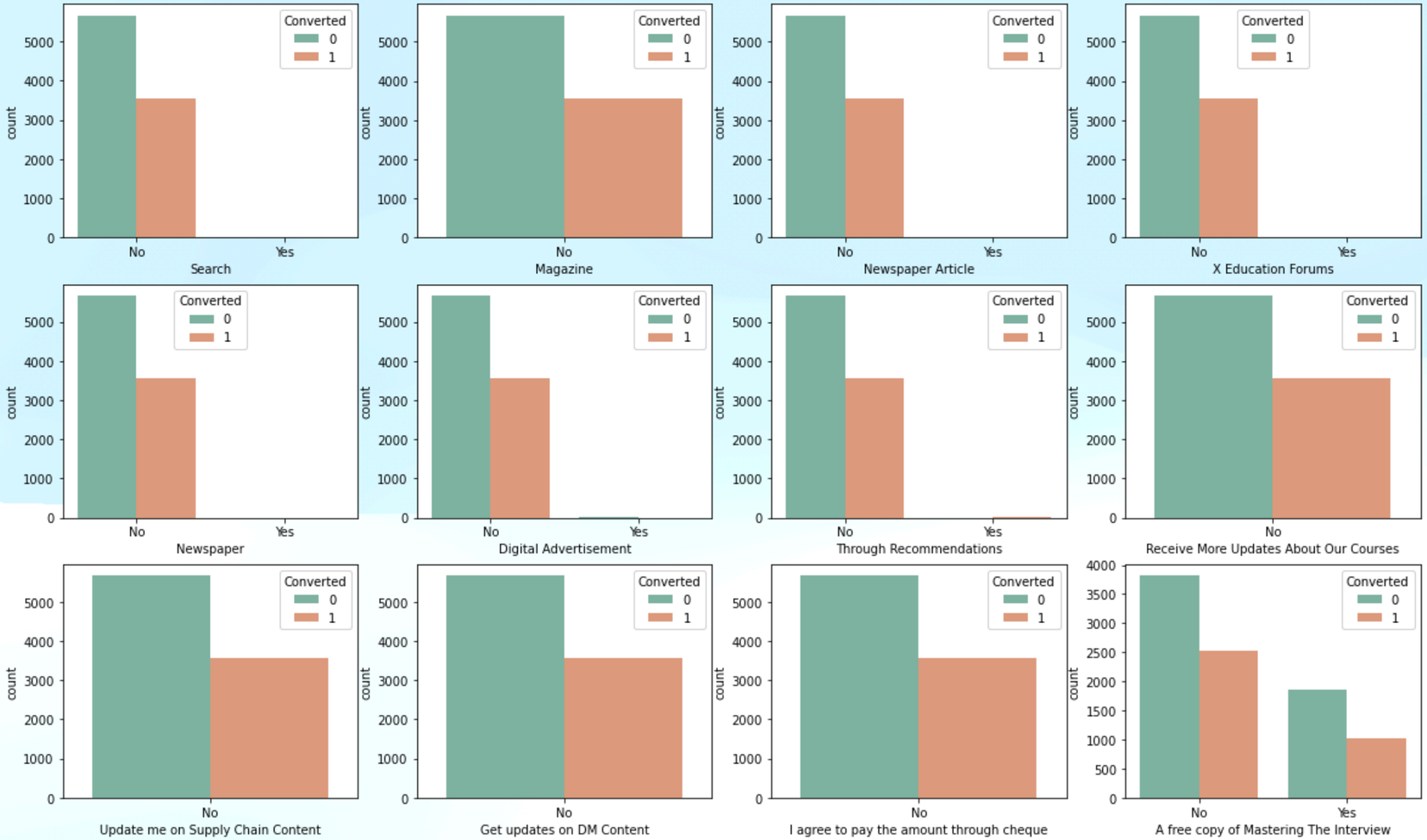
- There are many columns with high percentage of null values, dropped them as they are not useful.

—Dropping Columns with Missing Values >=35%:

- Dropped all the columns with more than 45% missing values.
- Checked the percentage of null values in each column after dropping columns with more than 45% missing values.

—Categorical Attributes Analysis:

- Visualzing variables for imbalancing

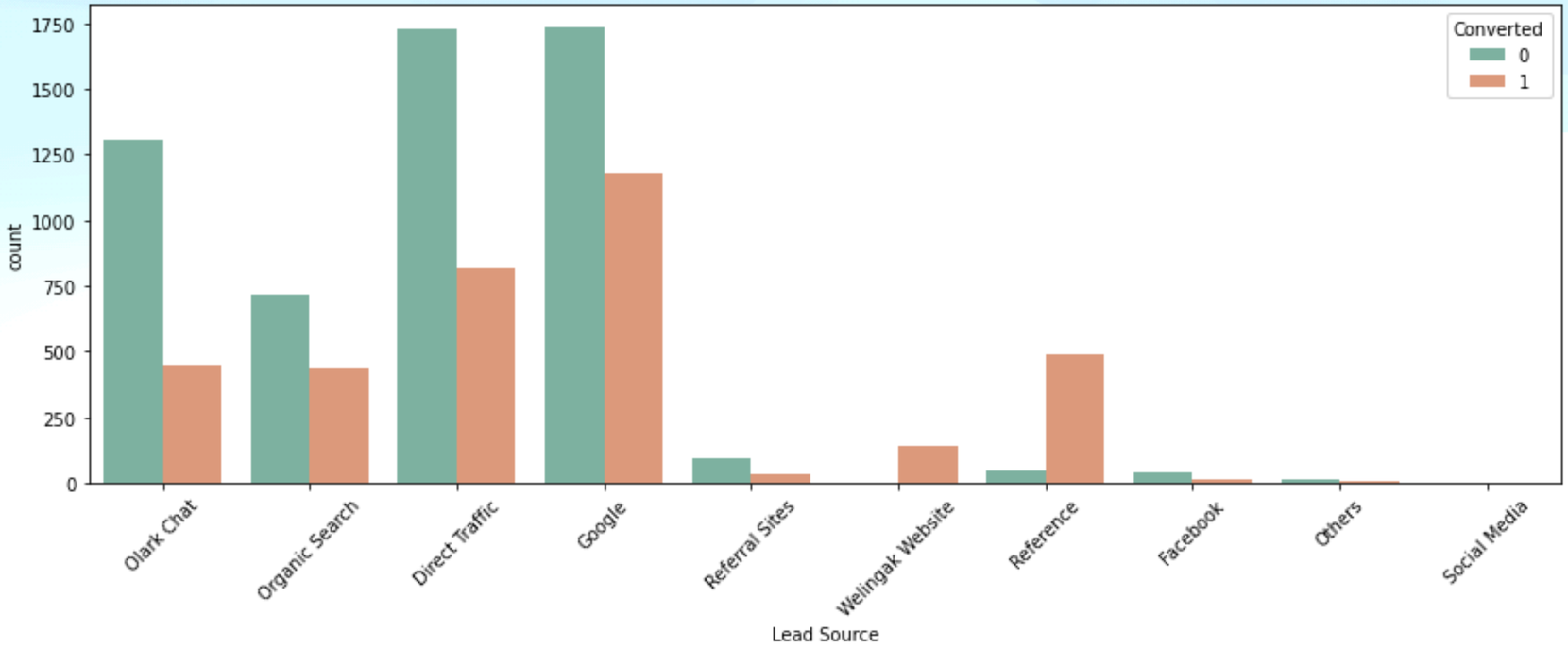


Inference

- For all these columns except 'A free copy of Mastering The Interview' data is highly imbalanced, thus dropped them.
- "A free copy of Mastering The Interview" is a redundant variable so included this also in list of dropping columns.

—Lead Source Column

- Google is having highest number of occurences, hence imputed the missing values with label ‘Google'.
- Replaced Nan Value with Google.
- Lead Source' is having same label name 'Google' but in different format i.e 'google', So converted google to Google.
- Combined low frequency values to Others.
- Visualized count of Lead Source Variable based on Converted value.



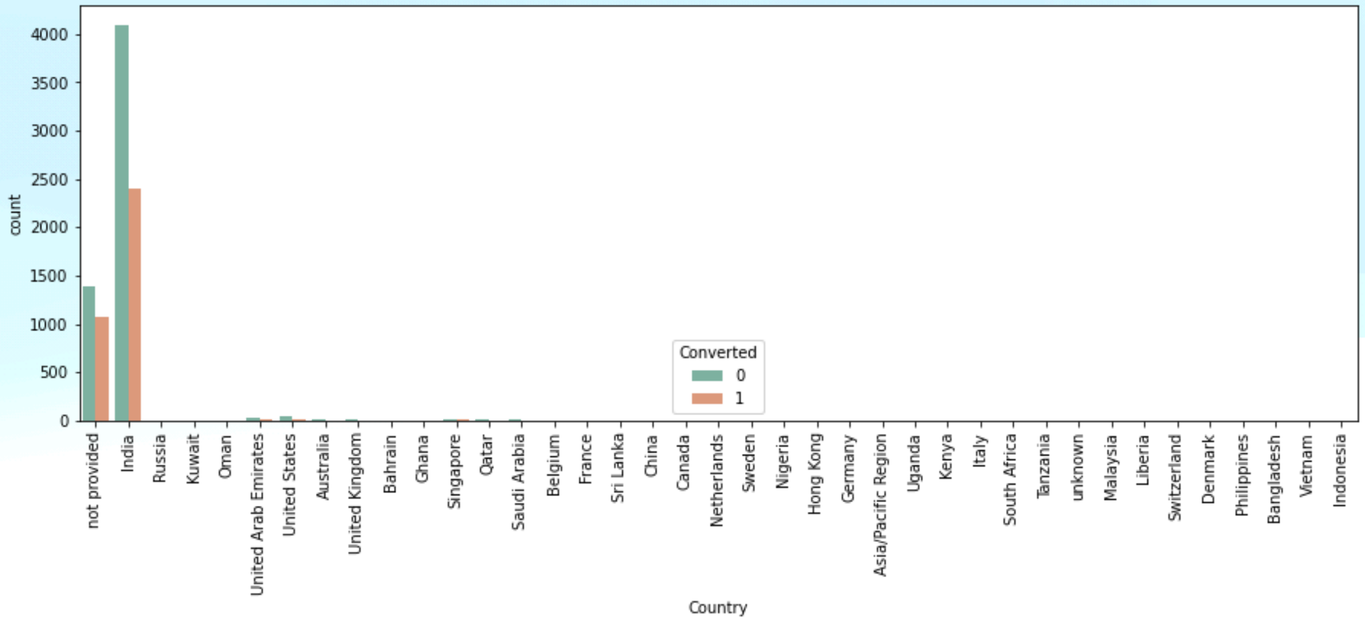
Inference

- Maximum Leads are generated by Google and Direct Traffic.
- Conversion rate of Reference leads and Welinkgak Website leads is very high.

—Country

- Since, missing values are very high , imputed all missing values with value 'not provided

- Visualized Country variable after imputation



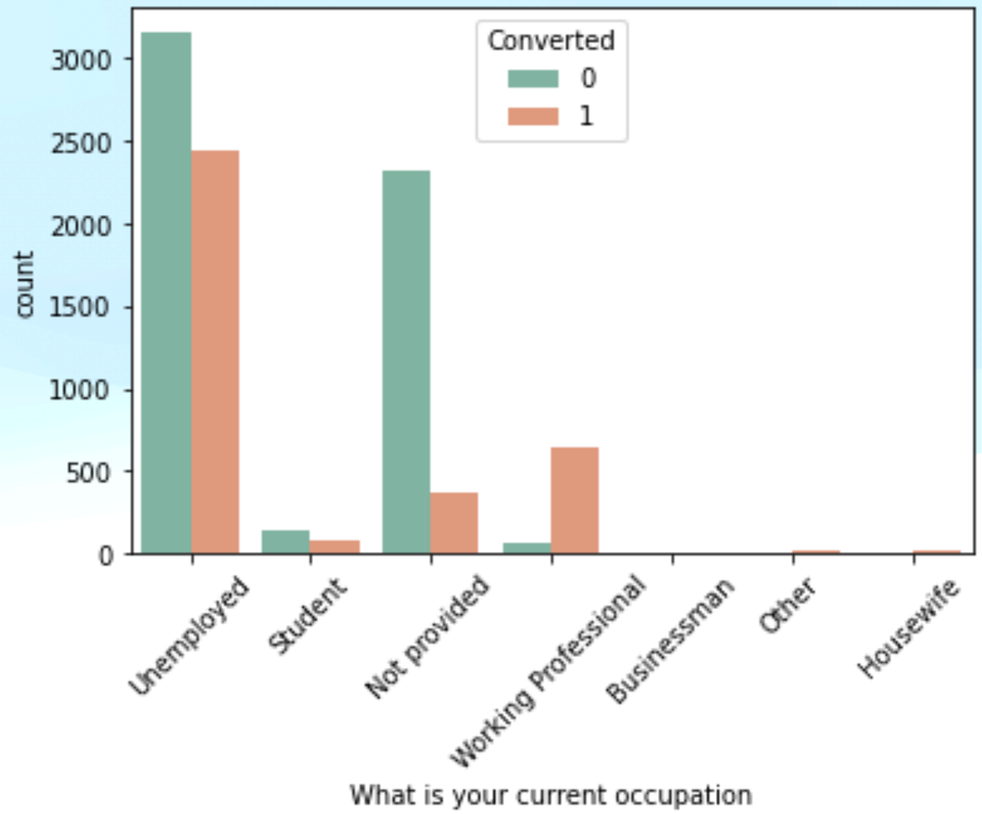
Inference

- As we can see that most of the data consists of value 'India', no inference can be drawn from this parameter.Hence,dropped this column.

- Created a list of columns to be droppped.

—What is your current occupation

- Checked value counts of 'What is your current occupation' column
- Since no information has been provided regarding occupation, hence replaced missing values with new category 'Not provided'.
- Visualized count of Variable based on Converted value



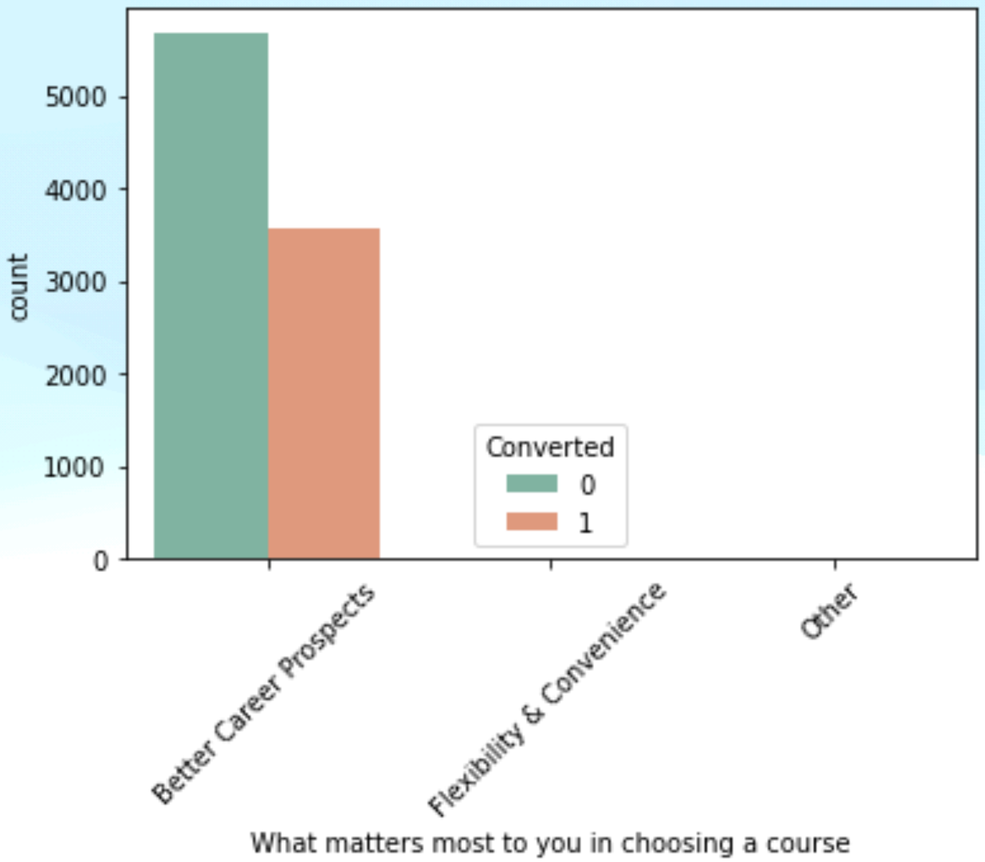
—Inference

- Maximum leads generated are unemployed and their conversion rate is more than 50%.
- Conversion rate of working professionals is very high.

—What matters most to you in choosing a course

- Checked value counts of 'What matters most to you in choosing a course’.

- Clearly seen that missing values in the this column can be imputed by 'Better Career Prospects’



•Visualized count of Variable based on Converted value

—Inference

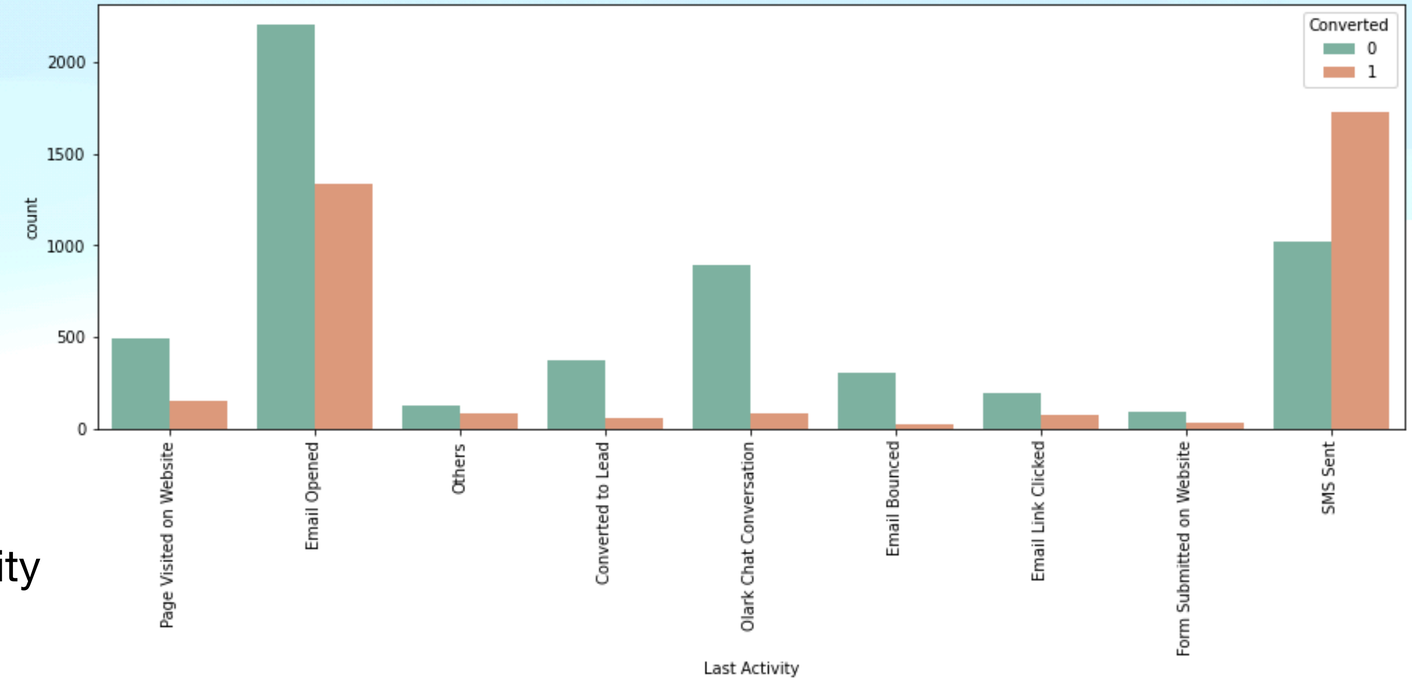
This column spread of variance is very low , hence it is dropped.

—Last Activity

- Checked value counts of Last Activity.
- Replaced Nan Values with mode value "Email Opened”.
- Combined low frequency values.
- Visualized count of Last Activity Variable.
- Dropped this column since it is a sales team generated data.

Inference

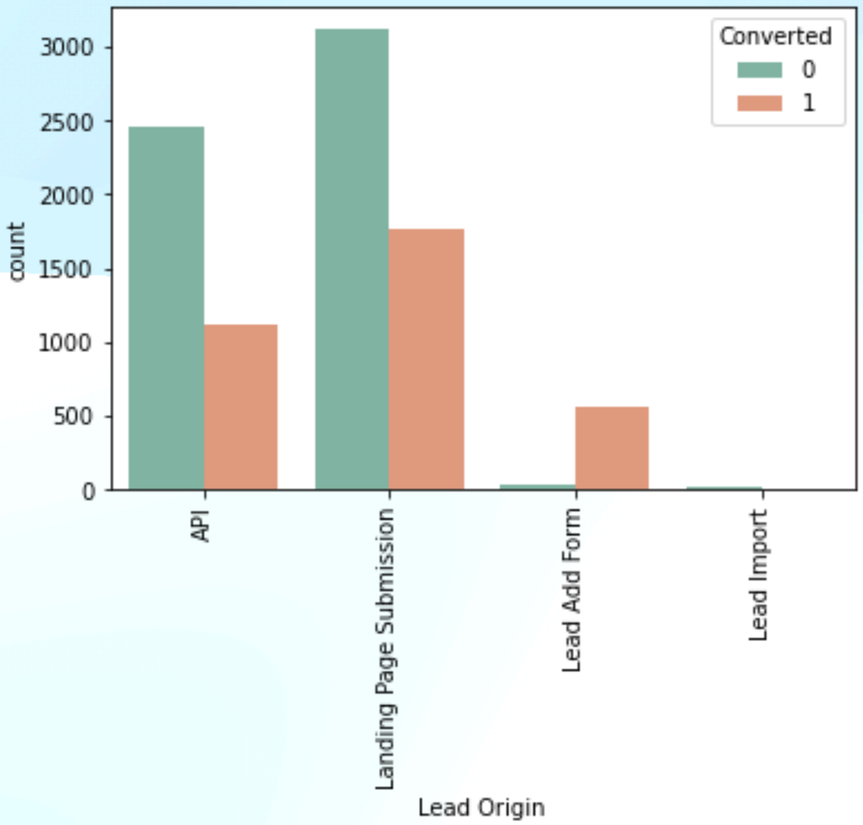
- Maximum leads are generated having last activity as Email opened but conversion rate is not too good.
- SMS sent as last acitivity has high conversion rate.

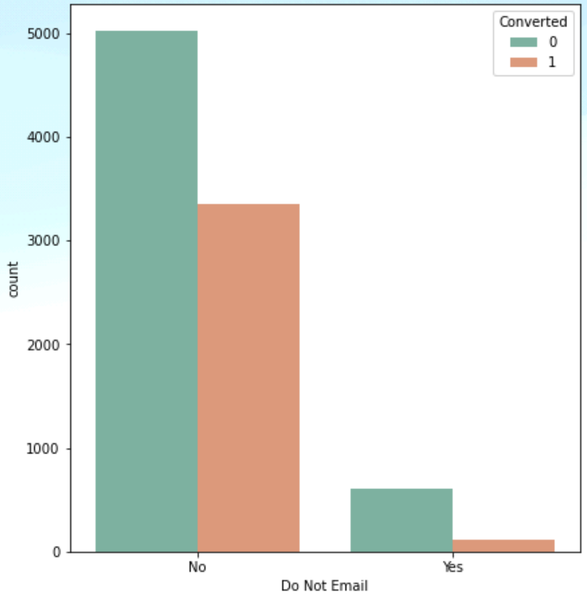


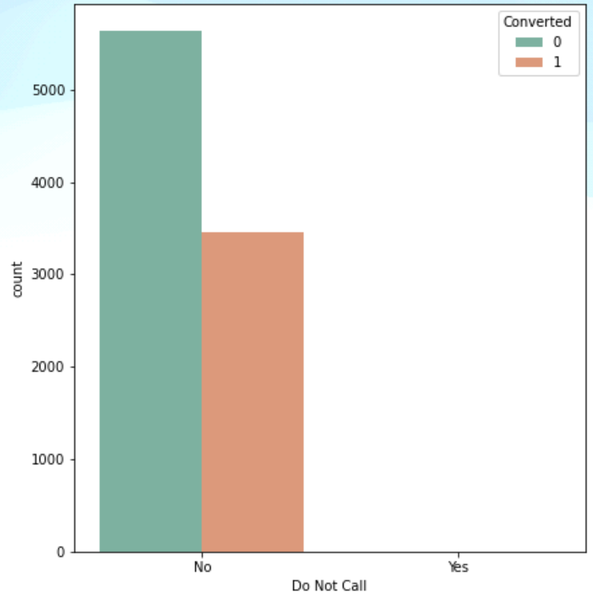
- Checked the Null Values in All Columns after imputation
- Remaining missing values percentage is less than 2%,hence dropped those rows without affecting the data.
- ReChecking percentage of Null Values in All Columns and found zero null values in remaining columns.

—Lead Origin,Do Not Email &Do Not Call

Do Not Call Column is dropped since it is highly skewed.





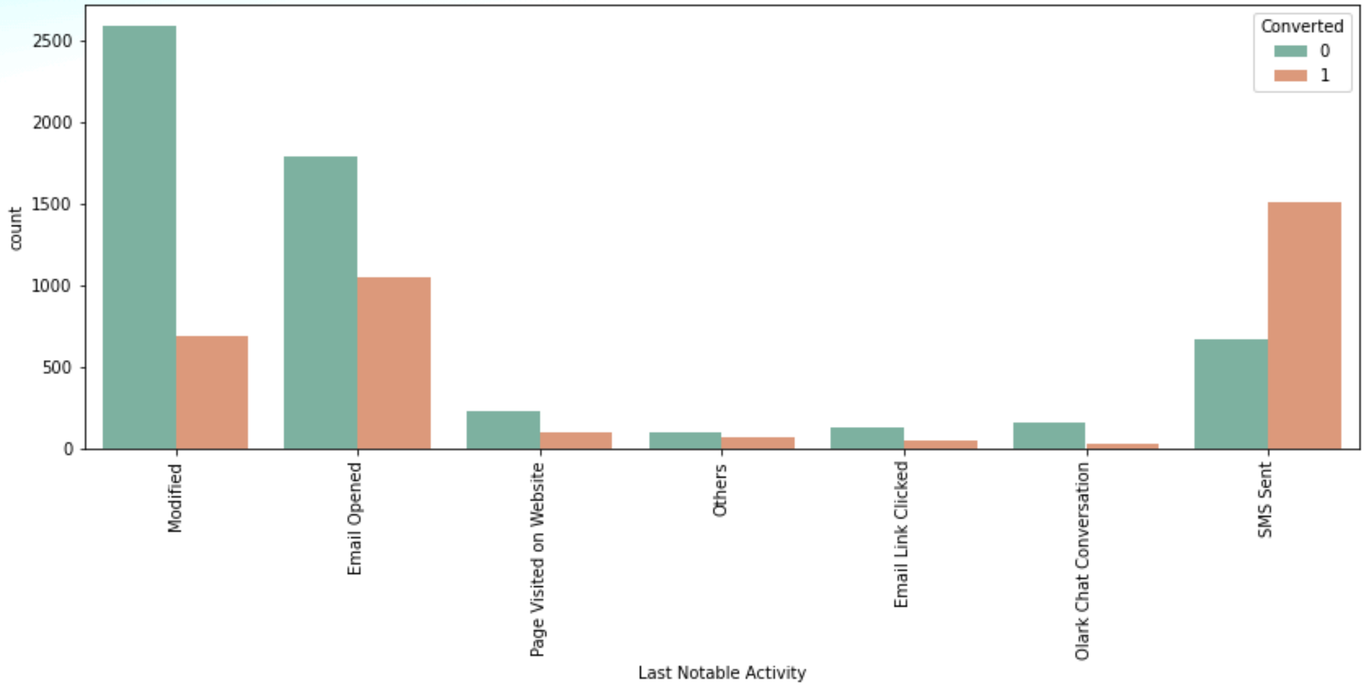


—Last Notable Activity

checked value counts of last Notable Activity,c lubbing lower frequency values and visualized count of Variable based on Converted value.

Inference

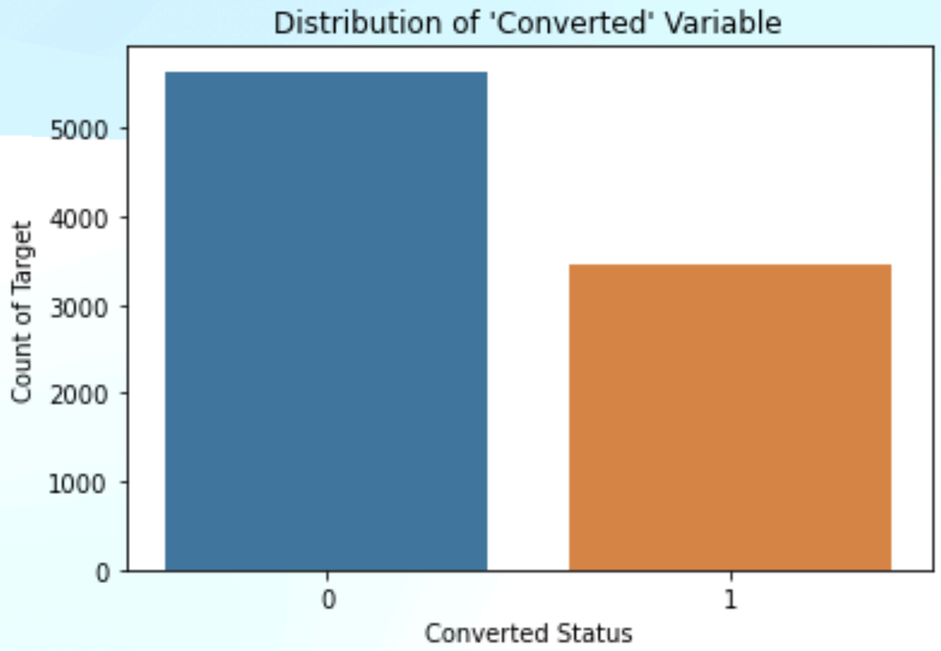
- Maximum leads are generated having last activity as Email opened but conversion rate is not too good.
- SMS sent as last acitivity has high conversion rate.
- Hence dropped the column as this is a sales team generated data.



—Numerical Attributes Analysis:

—Converted

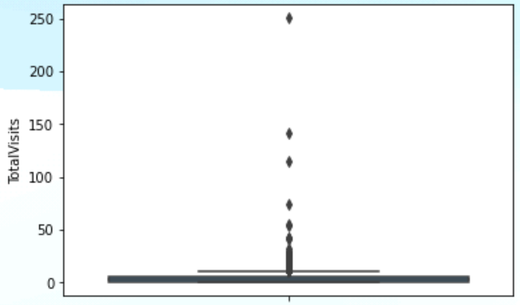
- Converted is the target variable, Indicates whether a lead has been successfully converted (1) or not (0).
- Visualized Distribution of 'Converted' Variable and find out the conversion rate as 38%.

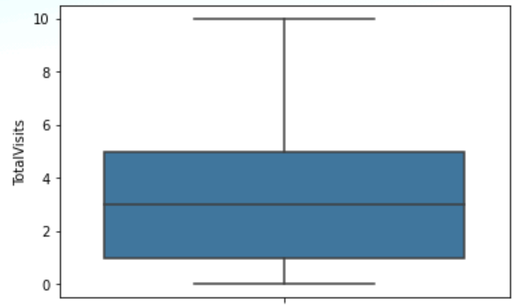


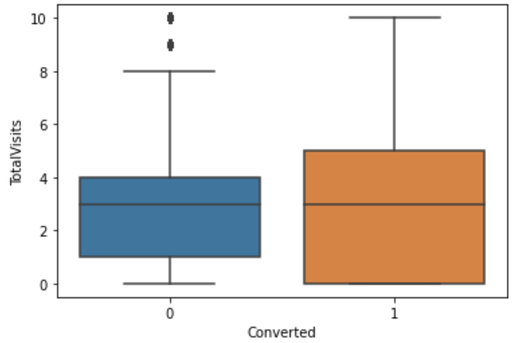
-Checking correlations of numeric values using heatmap

-TotalVisits

- **Visualized spread of variable Total Visits.**
- **Outlier Treatment: cappied the outliers to 95% value for analysis.**
- **Visualized variable after outlier treatment.**
- **Visualized TotalVisits w.r.t Target Variable 'Converted'**





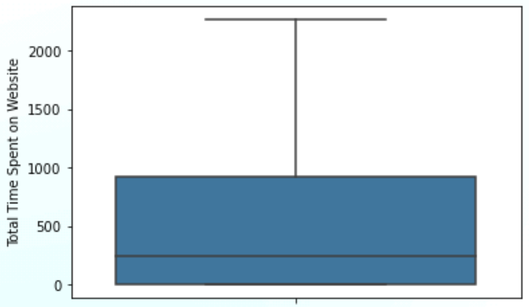


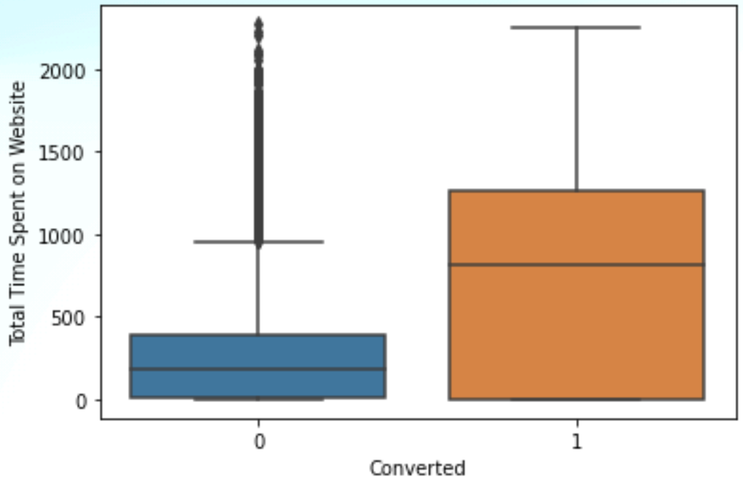
Inference

As the median for both converted and non-converted leads are same , nothing coclusive can be said on the basis of variable TotalVisits.

—Total time spent on website

- Checked percentiles for "Total Time Spent on Website”
- Visualized spread of variable 'Total Time Spent on Website’
- Visualized 'Total Time Spent on Website' w.r.t Target Variable 'converted'



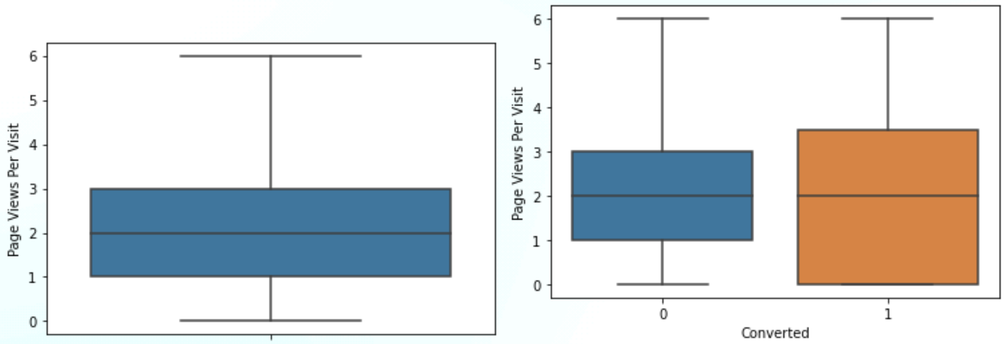


Inference

As can be seen, leads spending more time on website are more likely to convert , thus website should be made more enagaging to increase conversion rate.

—Page Views Per Visit

- Visualized spread of variable 'Page Views Per Visit’.
- Outlier Treatment: capping the outliers to 95% value for analysis.
- Visualized variable after outlier treatment.
- visualizing 'Page Views Per Visit' w.r.t Target variable 'Converted'



Inference

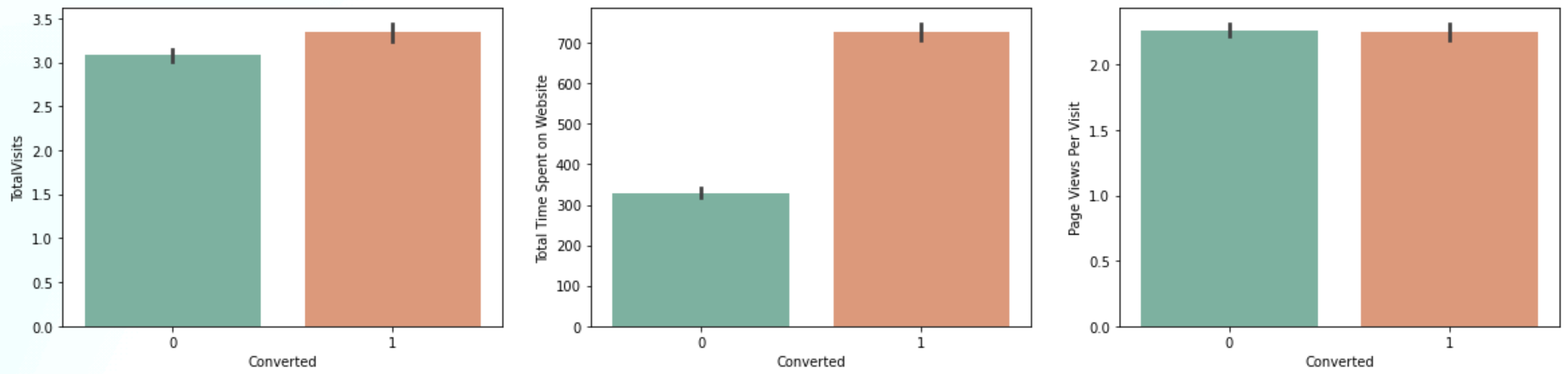
- Median for converted and not converted leads is almost same.
- Nothing conclusive can be said on the basis of Page Views Per Visit.

—Now checked the conversions for all numeric values

Inference

The conversion rate is high for Total Visits, Total Time Spent on Website and Page Views Per Visit

Now, all data labels are in good shape , we will proceed to our next step which is Data Preparation



Step 4: Data Preparation

—Converting some binary variables (Yes/No) to 0/1

—Dummy Variable Creation:

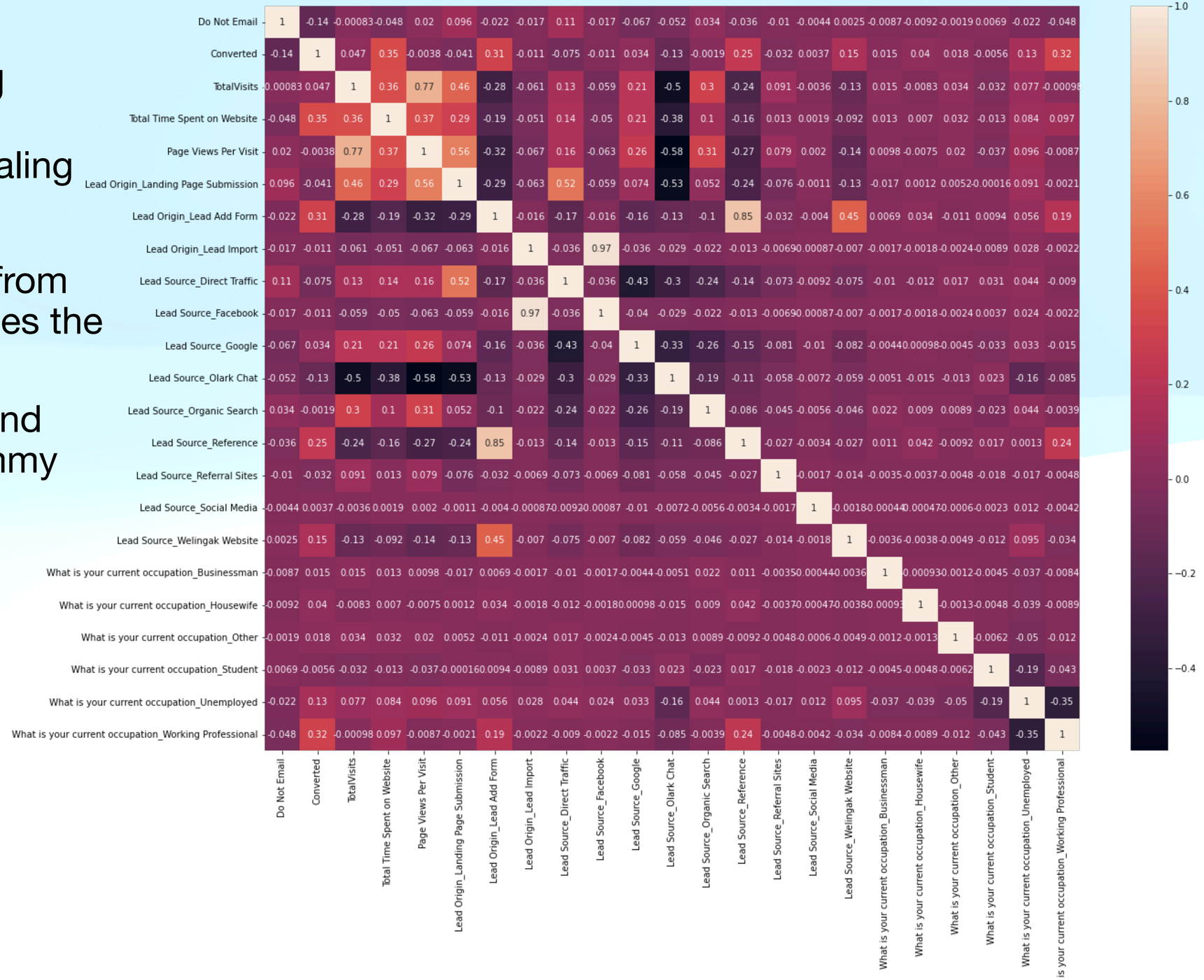
- Got a list of categorical columns foy creating dummy
- Created dummies and dropped the first column and adding the results to the master dataframe
- Dropped the original columns after dummy variable creation

Step 5: Test-Train Split

- Imported library for splitting dataset
- Putting feature variable to X
- Putting response variable to y
- Splitted the data into train and test

Step 6: Feature Scaling

- importing library for feature scaling
- Scaling of features
- Checking the conversion rate from 'converted' column as it denotes the target variable
- Visualised Correlation Matrix and dropped highly correlated dummy variables.



Step 7: Model Building using Stats Model & RFE

- Imported necessary library
- Printed list of RFE supported columns
- Built Model 1 with p-value of variable “What is your current occupation_Housewife” is high, so we dropped it.
- Built Model 2 with high p-value of variable "Lead Source_Welingak Website" , so we dropped it.
- Built Model 3 with high p-value of variable 'What is your current occupation_Businessman' so we dropped it.
- Built Model 4 with high p-value of variable 'What is your current occupation_Other' so we dropped it.
- Built Model 5 and the Model 5 seems to be stable with significant p-values, we shall go ahead with this model for further analysis.

—Predicting a Train model

- Getting the Predicted values on the train set
- Calculated converted Probability values.

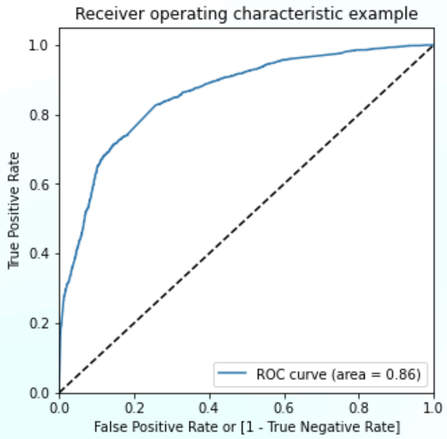
—Calculated Metrics -Accuracy, Sensitivity, Specificity, False Positive Rate, Postitive Predictive Value and Negative Predictive Value

- Printed Confusion matrix and found overall accuracy as 80%.
- sensitivity of our logistic regression model is 64%.
- specificityof our logistic regression model is 89%.
- False Postive Rate - predicting conversion when customer does not have convert is 10%.
- positive predictive value is 79%.
- Negative predictive value is 80%.

—PLOTTING ROC CURVE

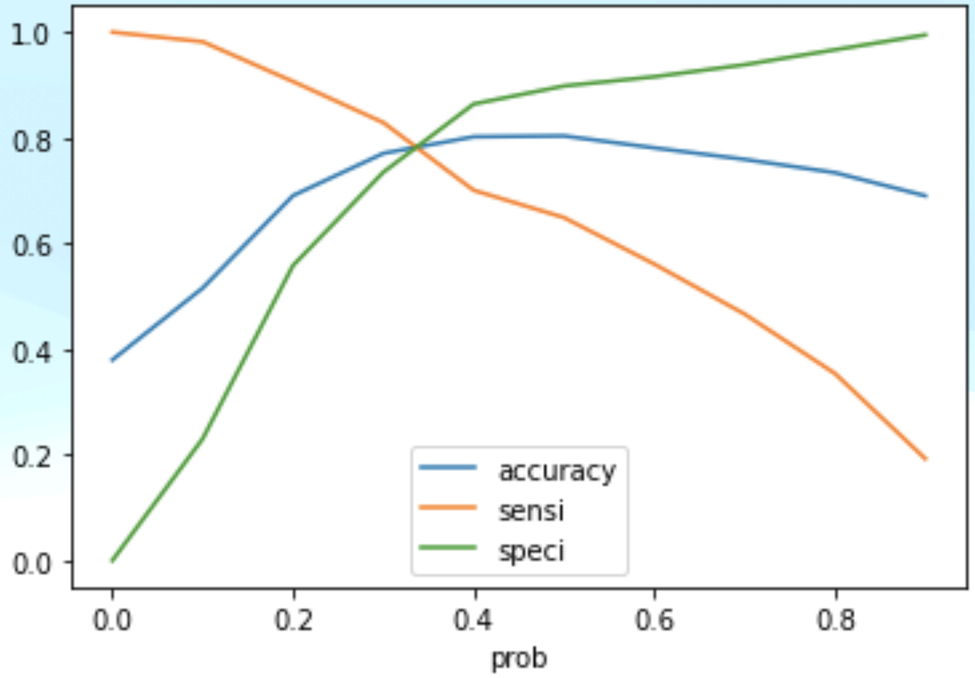
An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- The ROC Curve should be a value close to 1. We are getting a good value of 0.86 indicating a good predictive model.



—Finding Optimal Cutoff Point

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.
- Created columns with different probability cutoffs.
- Calculated accuracy sensitivity and specificity for various probability cutoffs.
- Plotted accuracy sensitivity and specificity for various probabilities.
- Calculated Lead Score,From the curve above, 0.3 is the optimum point to take it as a cutoff probability.



- Checked if 80% cases are correctly predicted based on the converted column.
- got the total of final predicted conversion / non conversion counts from the actual converted rates
- checked the precentage of final_predicted conversion is 82%
- Hence, we can see that the final prediction of conversions have a target of 83% conversion as per the X Educations CEO's requirement . Hence, we can say that this is a good model.

Overall Metrics - Accuracy, Confusion Metrics, Sensitivity, Specificity, False Postive Rate, Positive Predictive Value, Negative Predicitive Value on final prediction on train set.

- overall accuracy is 77%.

Inference:
So as we can see above the model seems to be performing well. The ROC curve has a value of 0.86, which is very good. We have the following values for the Train Data:

- Accuracy : 77.05%
- Sensitivity :82.89%
- Specificity : 73.49%

Some of the other Stats are derived below, indicating the False Positive Rate, Positive Predictive Value,Negative Predictive Values, Precision & Recall.

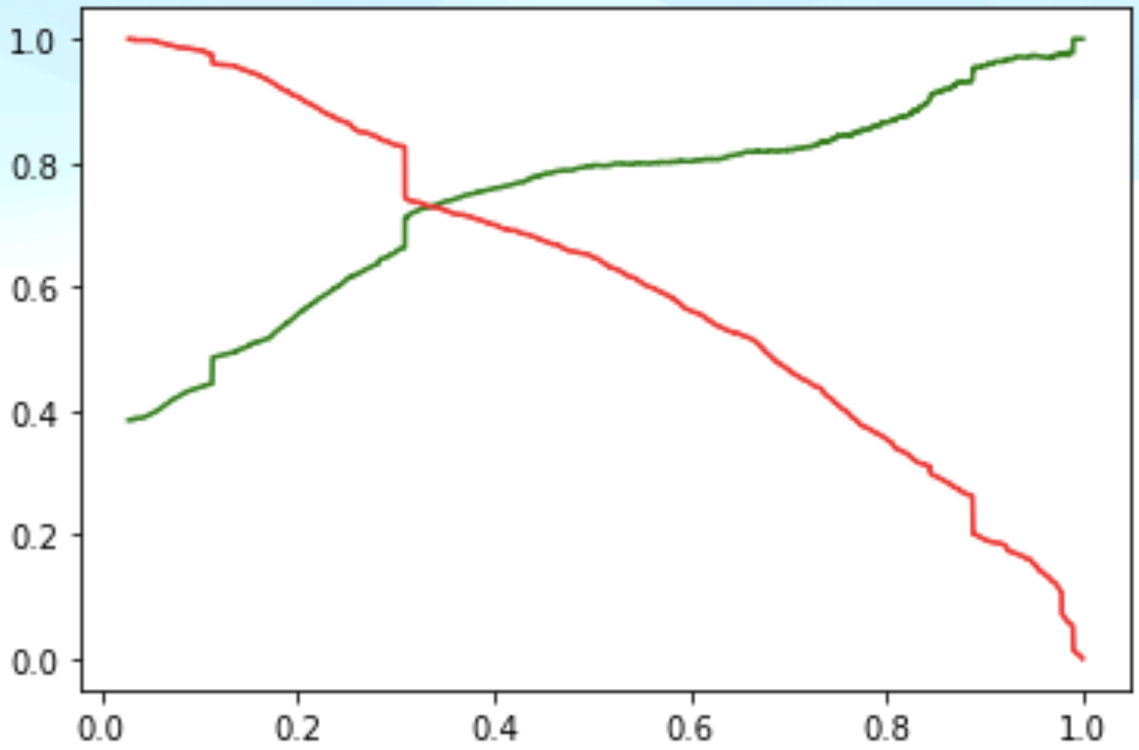
- False Postive Rate - predicting conversion when customer does not have convert is 26%.
- Positive predictive value is65%.
- Negative predictive value is 87%.

—Precision and Recall

- Precision is 65.6%
- Recall is 82%
- precision_score is 65.6%
- recall_score is 82.8%

—Precision and Recall Trade-off

- Imported precision recall curve from sklearn library
- Created precision recall curve



—Predictions on the test set

- scaling test set.
- Converted y_pred to a dataframe which is an array.
- Converted y_test to dataframe.
- Appended y_test_df and y_pred_1.
- Assigned Lead_Score.
- Checked if 80% cases are correctly predicted based on the converted column.
- got the total of final predicted conversion or non conversion counts from the actual converted rates.
- Checked the precentage of final_predicted conversions on test data is 83.01%.
- Hence we can see that the final prediction of conversions have a target rate of 83% (same as predictions made on training data set.)

—Overall Metrics - Accuracy, Confusion Metrics, Sensitivity, Specificity, False Postive Rate, Positive Predictive Value, Negative Predicitive Value on final prediction on train set.

- Checked the overall accuracy as 77.5%.
- the sensitivity of our logistic regression model as 83.01%.
- calculate specificity as 74%.
- precision_score is 66.43%.
- recall_score is 83.01%.

Inference:
After running the model on the Test Data these are the figures we obtain:

- Accuracy : 77.52%
- Sensitivity :83.01%
- Specificity : 74.13%

Conclusion:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the
  optimal
  cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to
  the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is
  around 80%
- Hence overall this model seems to be good.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
-Lead Origin_Lead Add Form
-What is your current occupation_Working Professional
-Total Time Spent on Website

THANK YOU