LEAD SCORE ASSIGNMENT

SUBMITTED BY ADITYA VIKRAM

SUBMITTED BY KUNWAR URJASWIT

SUBMITTED BY SHYAM RAJAGOPAL

CONTENT

- Problem Statement
- Business Goal
- Strategy
- Conclusion
- Recommendation in Business Terms

PROBLEM STATEMENT

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process,

some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

BUSINESS GOAL

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

- The company needs a model wherein you a lead score is assigned 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%.

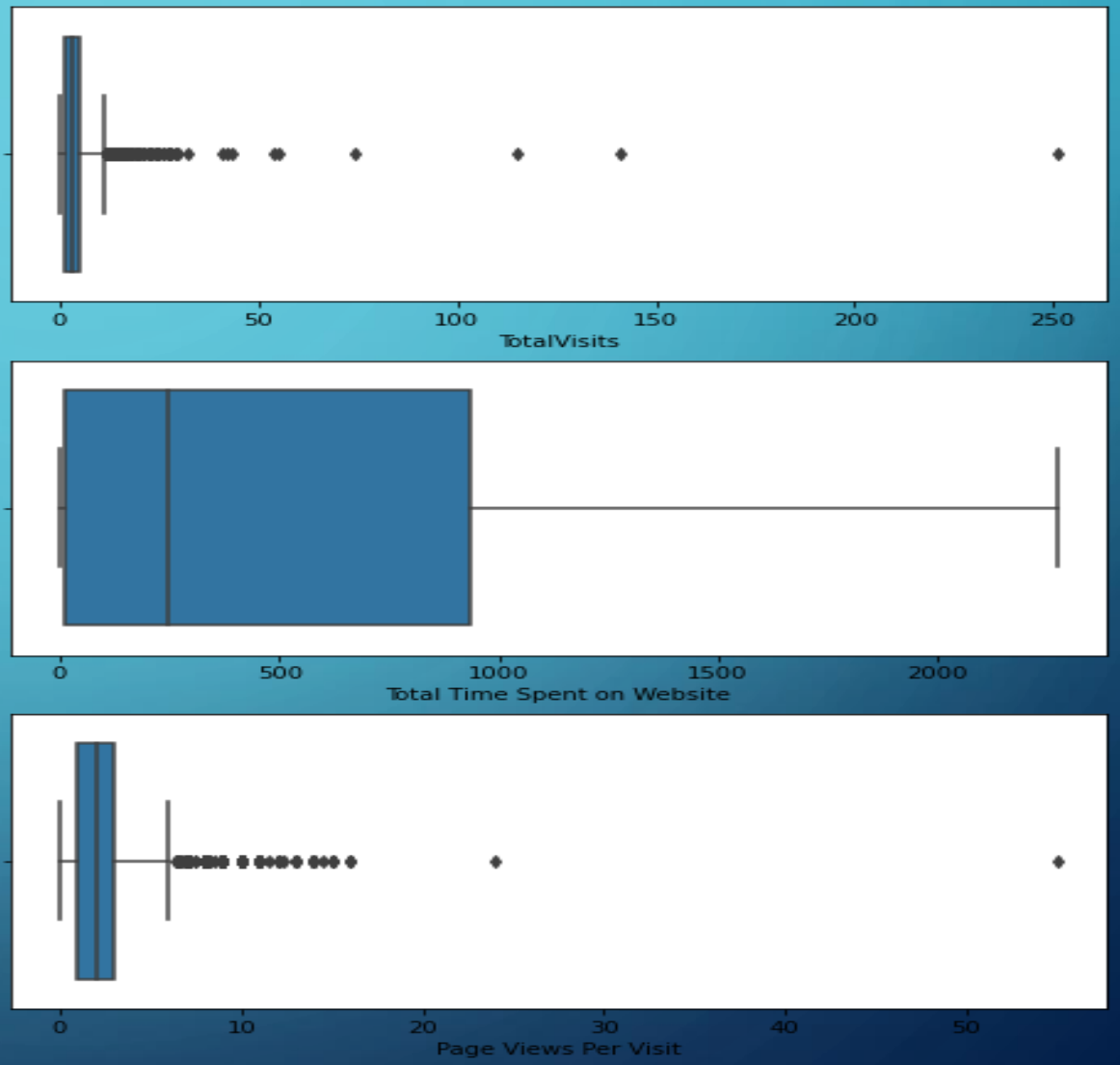
STRATEGY– CLEANING DATA & HANDLING NULL VALUES:

- The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information and data had unique identifiers
- Dropped one of the unique identifiers as it won’t help in decision making.
- Dropped columns having null values greater than 35%. Also, those columns were not important from business perspective.
- Few columns had high number of null values but not too high (<35%) and thus treated each column separately and dropped column which were heavily skewed.
- For rest of columns categorized few columns and combined together as the proportion was very less, this helped in reducing column.
- Further, there were few numerical column and categorical column which had a smaller number of null values thus imputed numerical column with median and categorical column with mode.

STRATEGY– EXPLORATORY DATA ANALYSIS:

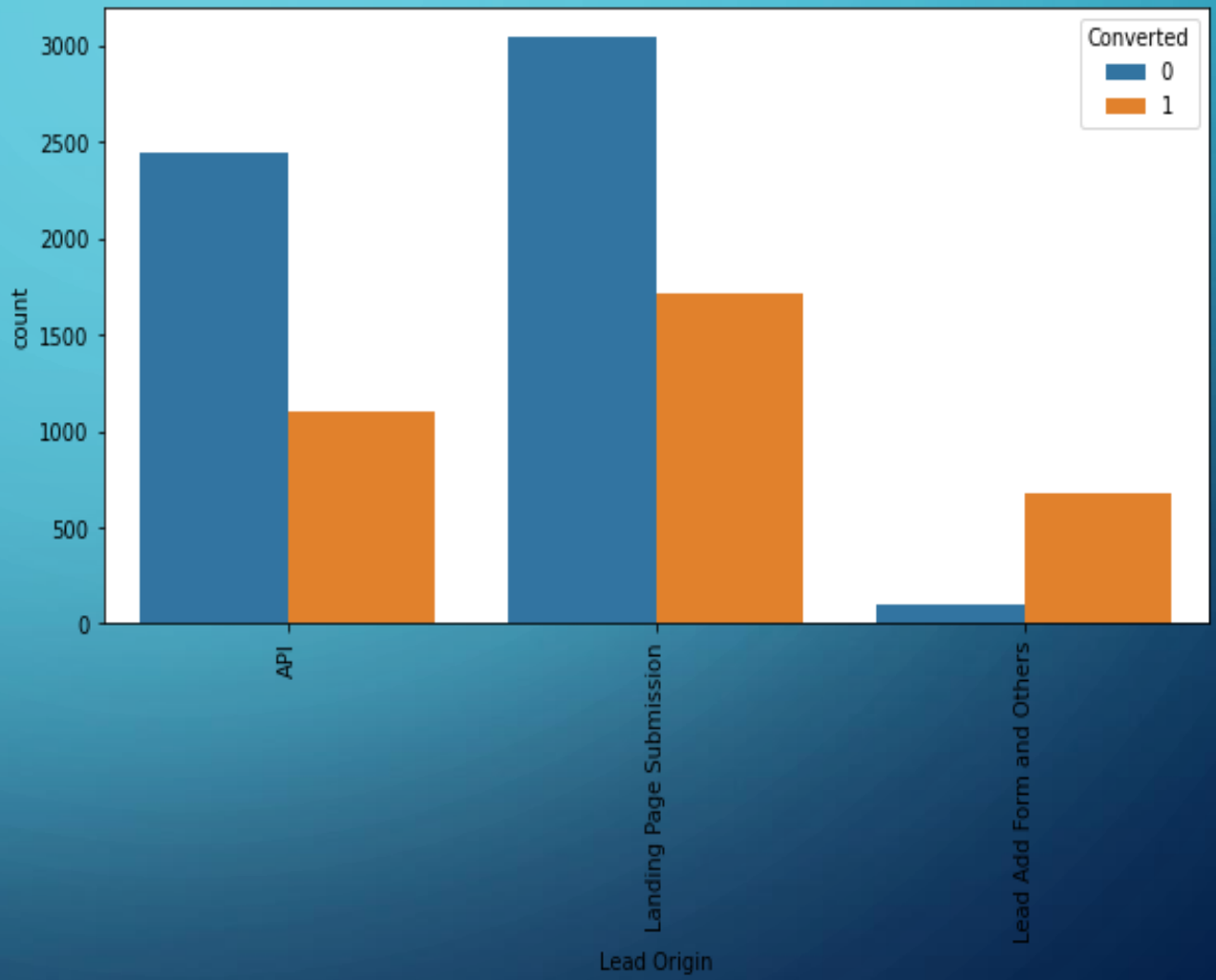
NUMERICAL COLUMNS

Plotted box-plot for numerical column which indicated outliers and thus treated outlier by capping at 99% percentile.



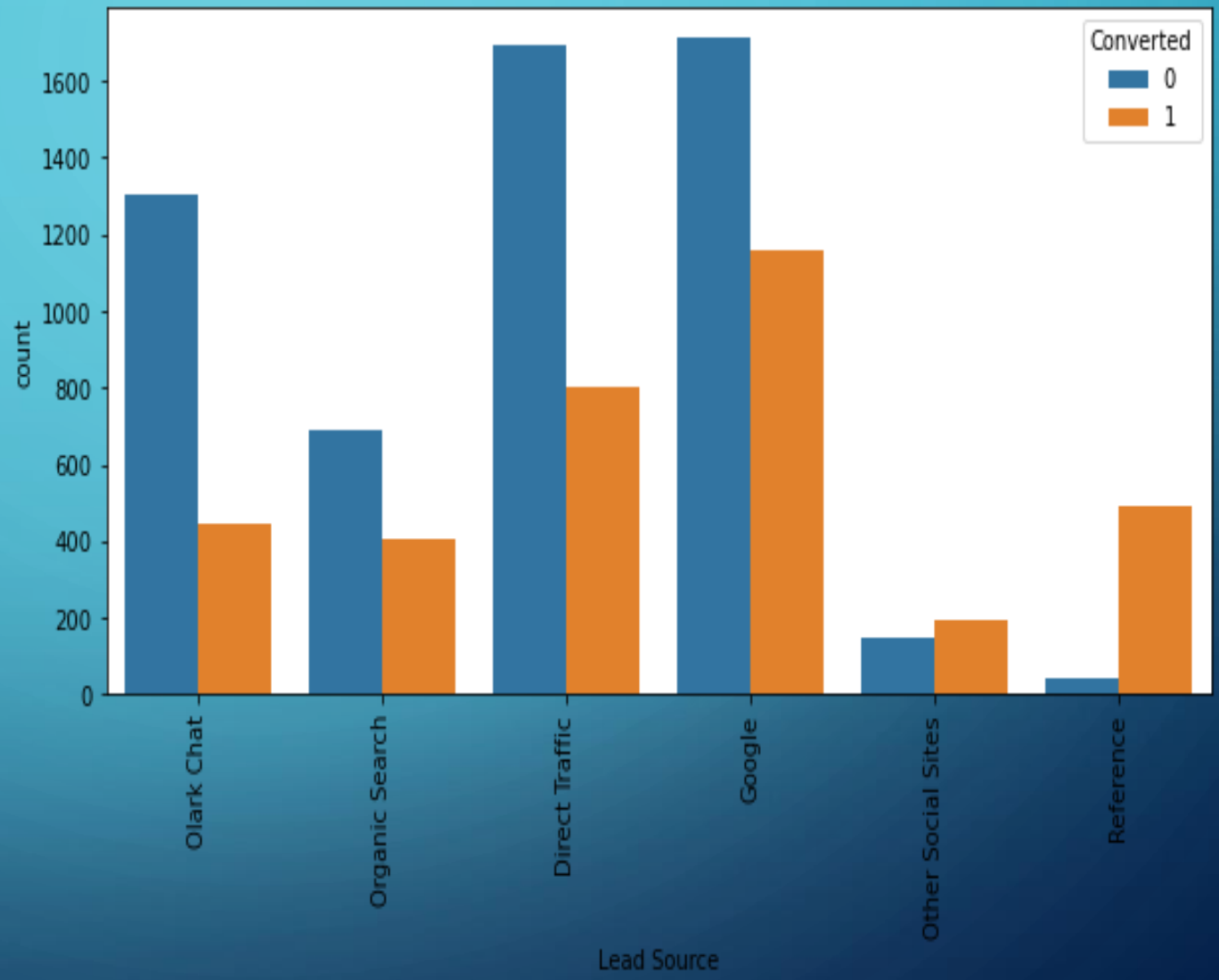
ANALYSIS

Plotted count-plot for categorical column with respect to target variable.



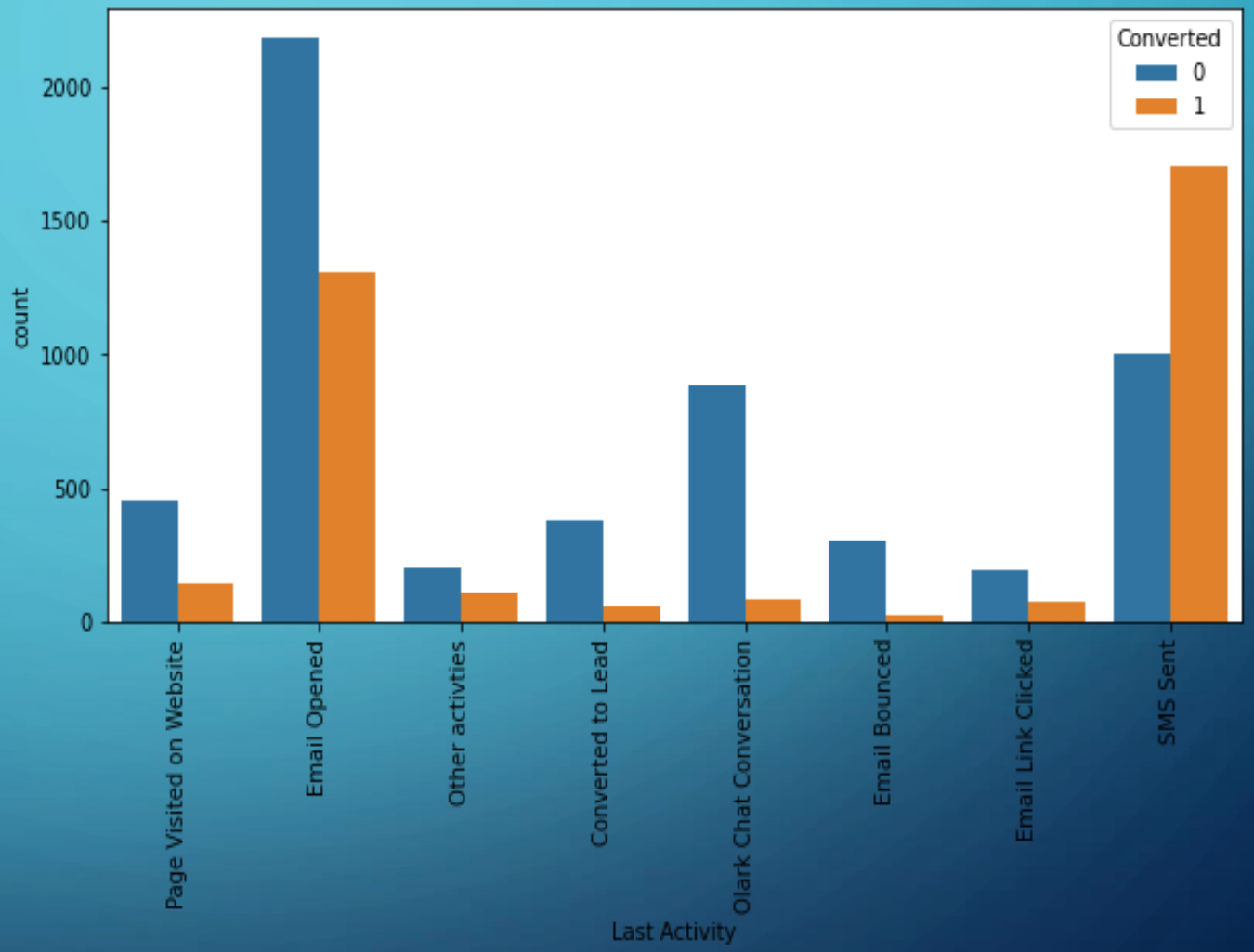
ANALYSIS

Plotted count-plot for categorical column with respect to target variable.



ANALYSIS

Plotted count-plot for categorical column with respect to target variable.

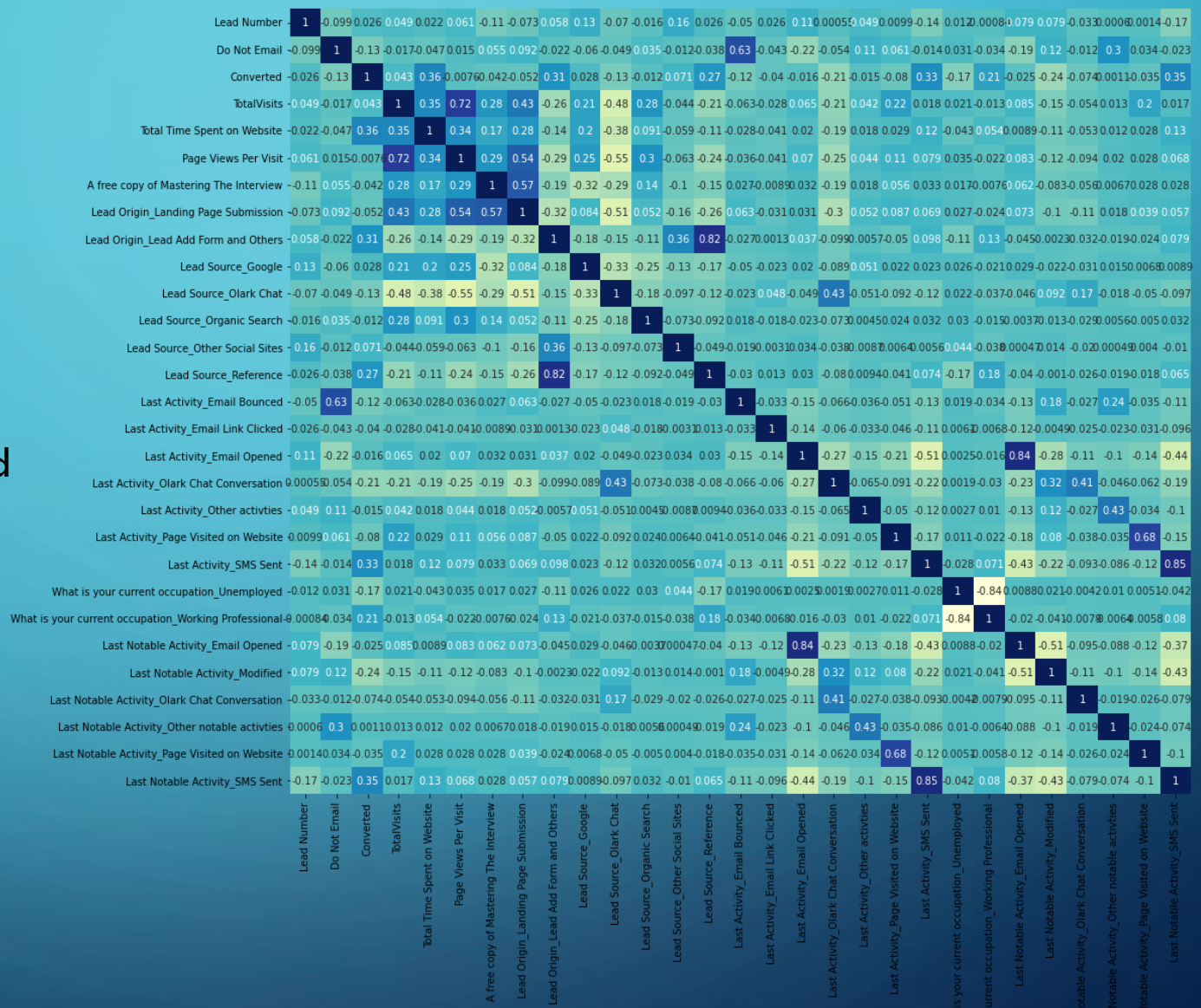


STRATEGY– DUMMY VARIABLE & TRAIN-TEST SPLIT:

- Converted binary valued column to 0 and 1.
- Created dummy variable for categorical columns except binary one.
- The split was done at 70% and 30% for train and test data respectively.
- Scaled for numerical columns.

CORRELATION

Correlation was plotted and dropped few columns which were highly corelated.



STRATEGY– MODEL BUILDING & EVALUATION:

- RFE was done to attain the top 15 relevant variables.
- Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
- 4th model was selected.
- Predicted the possibility of conversion on train dataset.
- A confusion matrix was made. Later on, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 78-80%.
- The optimum cut off value was found using ROC curve. The area under ROC curve was 0.88.
- After Plotting we found that optimum cutoff was 0.38 which gave
  - Accuracy 79.01%
  - Sensitivity 79.96%
  - Specificity 79.66%.

STRATEGY– PREDICTION ON TEST DATASET:

- Prediction on test dataset with new cut off of 0.45
- Created confusion matrix
- We get below matrices for test dataset with cut off 0.45:
- Accuracy 78.84%
- Precision 73.60%
- Recall 71.60%

CONCLUSION

VARIABLES IMPACTING THE CONVERSION RATE

- What is your current occupation_Unemployed
- Last Activity_Email Opened
- Last Activity_SMS Sent
- Total Time Spent on Website
- Last Notable Activity_Modified
- Lead Source_Olark Chat
- Last Notable Activity_Other notable activties
- Last Activity_Other activities
- Do Not Email
- Lead Source_Reference
- Last Notable Activity_Olark Chat Conversation
- Lead Source_Other Social Sites

RECOMMENDATION IN BUSINESS TERMS

RECOMMENDATIONS:

- A group can be made of "Hot leads" or " High priority" having conversion rate of 75% or more and sales team can focus on this section to as this will be converted easily.

- Second group can be made as "Intermediate priority having conversion rate between 40%-74% and sales team can provide them offers as with ample amount of work these leads have chance of getting converted.

- Third group can be made as "Low priority" having conversion rate of less than 40% and sales team can avoid these leads as they have less chances of getting converted and investing time and resource may result in loss of revenue.

THANK YOU !!