Lead Scoring Assignment

Approach of the analysis

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- 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%.

Problem Statement

- Now, although 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.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you 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%.

Approach

Step1: Read, Understand and Prepare the data

- Data is imported and understood using the data dictionary.
- There are few “select” values in the dataset which are as good as null values. They are converted into actual null values.
- Null value percentage is calculated for each column. The columns that have more than 50 percent of null values are straight away dropped.
- Additionally, columns that have the data that was populated by sales team are also dropped because that data comes after the sales team approaches the lead and we don’t want such data in our model.
- These are the columns that are dropped from above two steps:

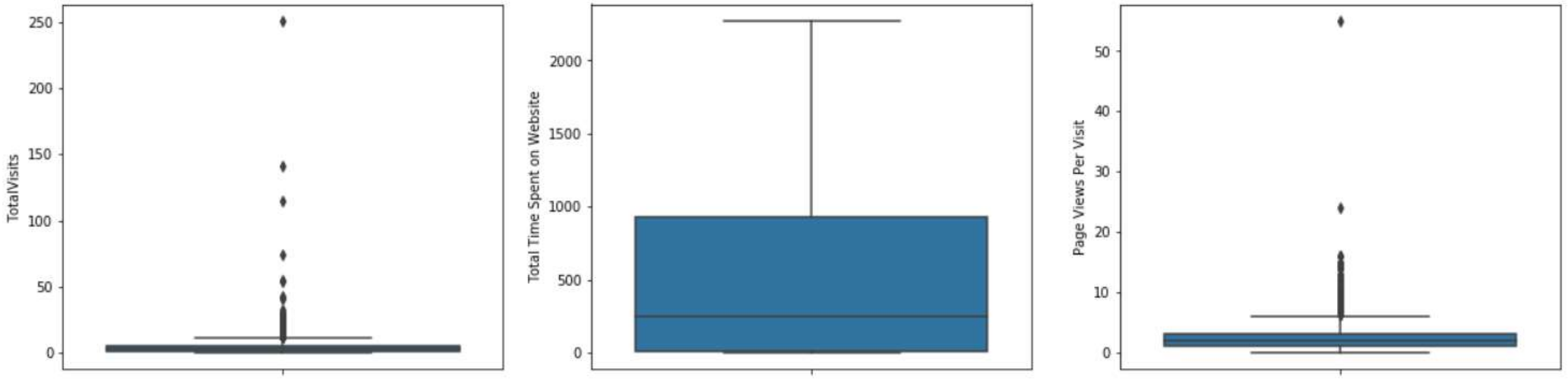
'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score', 'Lead Profile', 'Tags', 'How did you hear about X Education', 'Lead Quality'

- Columns that are highly skewed are dropped. There are a lot of columns that have high skewness. They are all dropped as they add no value to the prediction. Below are the highly skewed columns

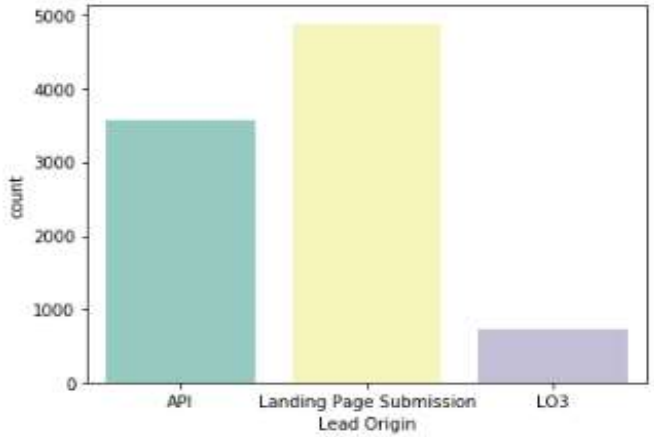
'Do Not Call', 'Country', 'What matters most to you in choosing a course', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'

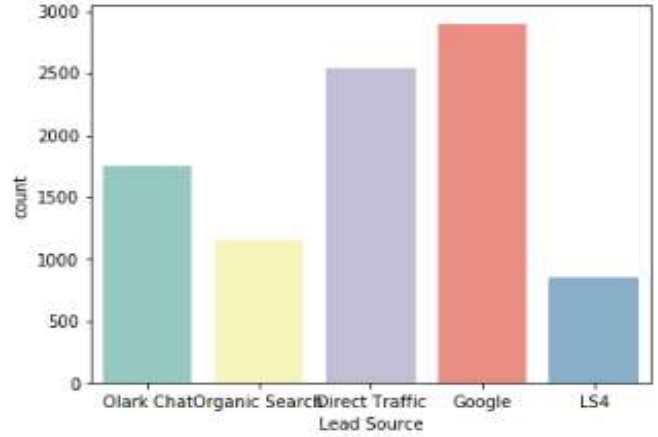
- From the resulting dataset, all the rows that have at least 5 null values have been dropped.
- For each column of the resulting dataset, missing values have been replaced with their corresponding mode.
- Few categorical columns have very high number of categories with majority of the categories having very less percent of data. They all are combined as one single category and this category will be dropped after creating dummies.

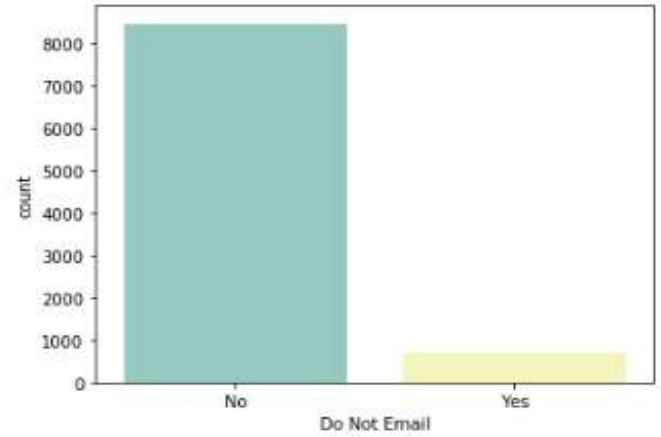
- Numerical columns are visualized using box plot and identified the outliers.

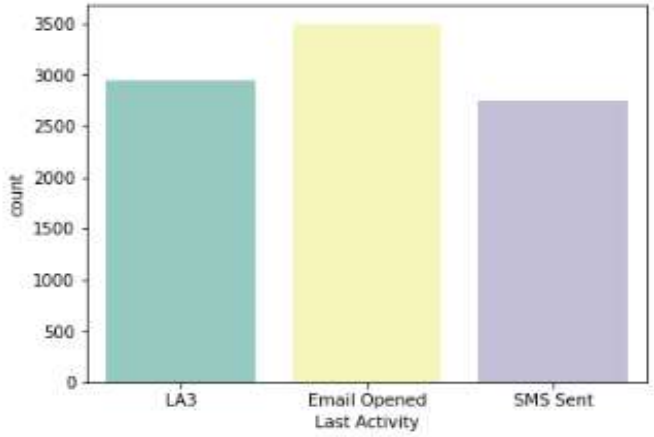


- Categorical columns are visualized using count plot to understand the distribution of different categories. This visualization is shown below.

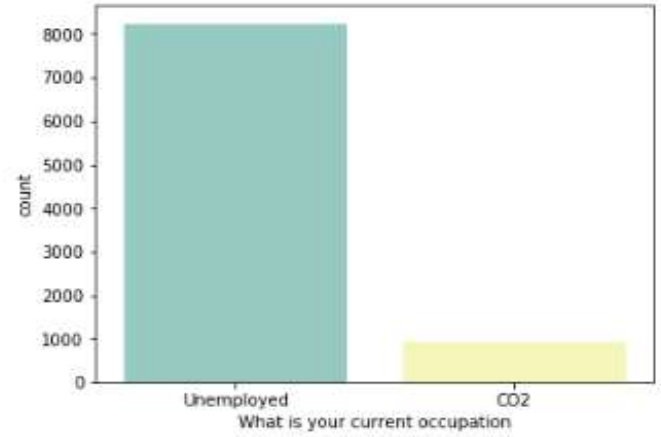

count
API
Landing Page Submission
LO3
Lead Origin

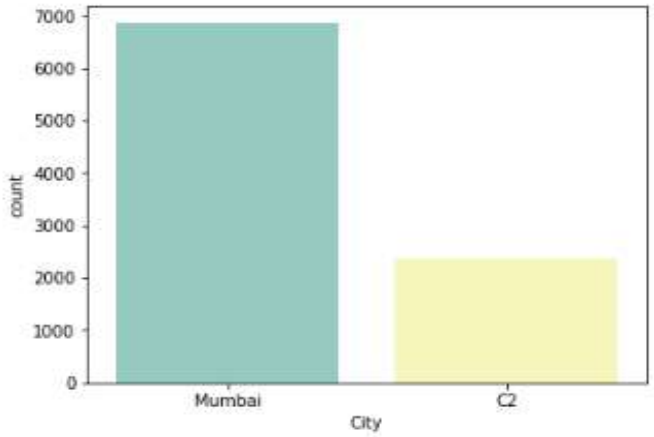

count
Olark Chat
Organic Search
Direct Traffic
Google
LS4
Lead Source

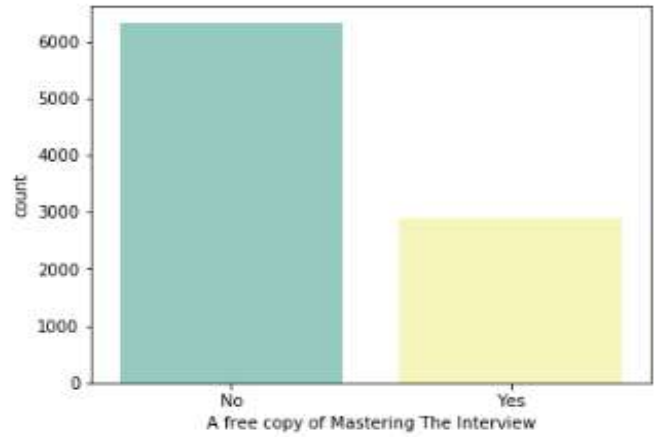

count
No
Yes
Do Not Email

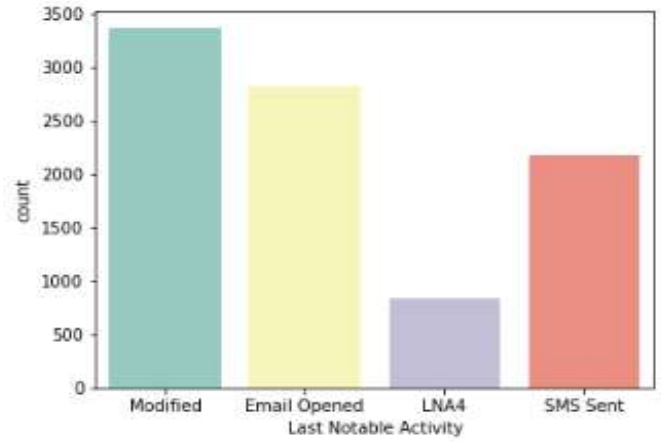

count
LA3
Email Opened
SMS Sent
Last Activity


count
Finance Management
S2
Specialization

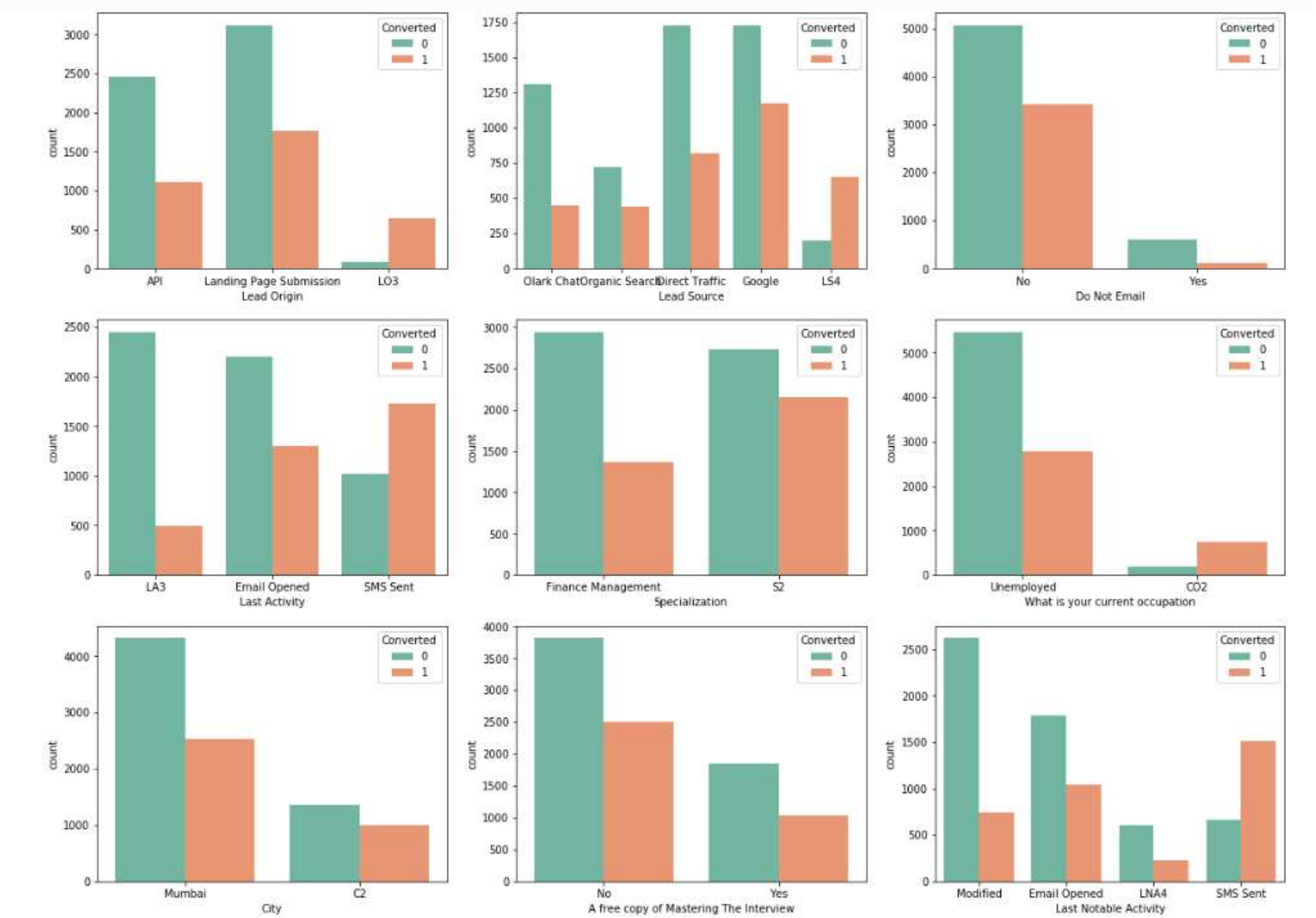

count
Unemployed
CO2
What is your current occupation


count
Mumbai
C2
City

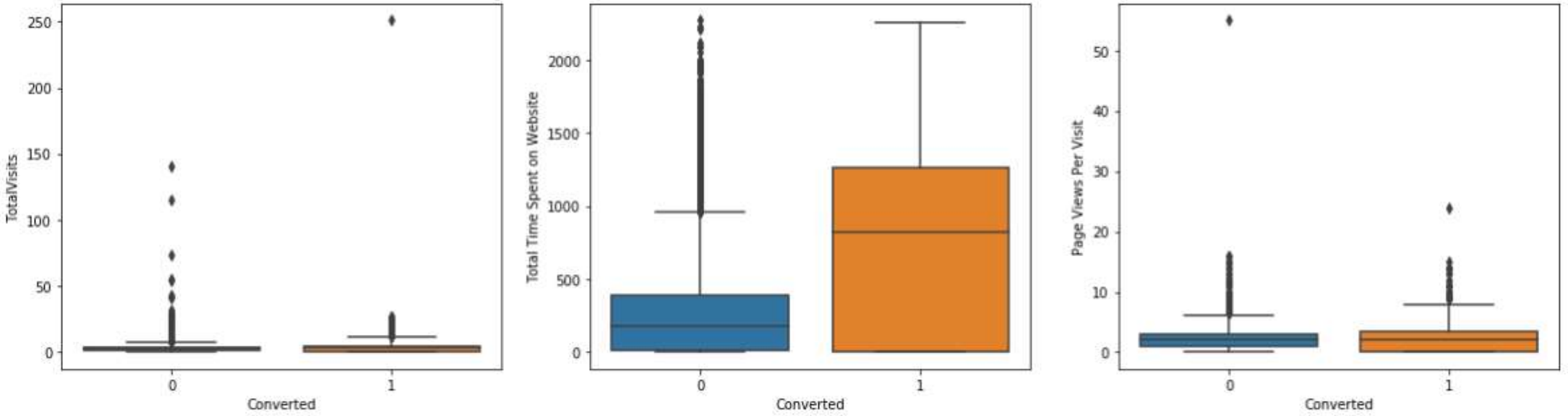

count
No
Yes
A free copy of Mastering The Interview


count
Modified
Email Opened
LNA4
SMS Sent
Last Notable Activity

- All the categorical columns are visualized with the target variable “converted” to understand the number of leads each category has got.

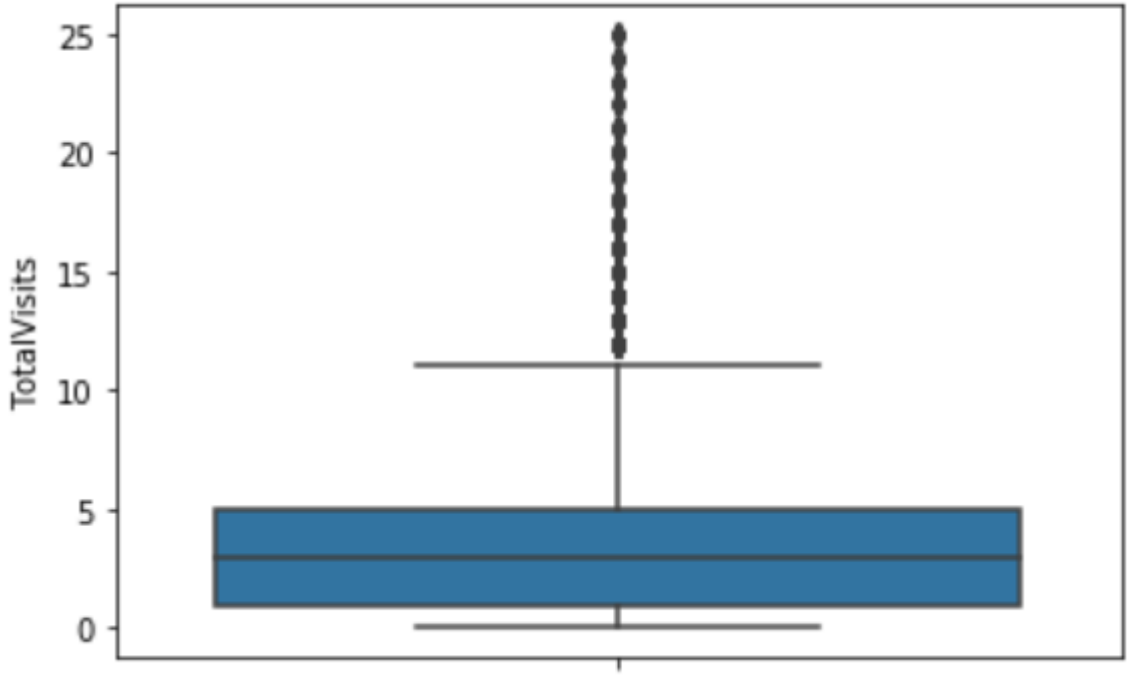


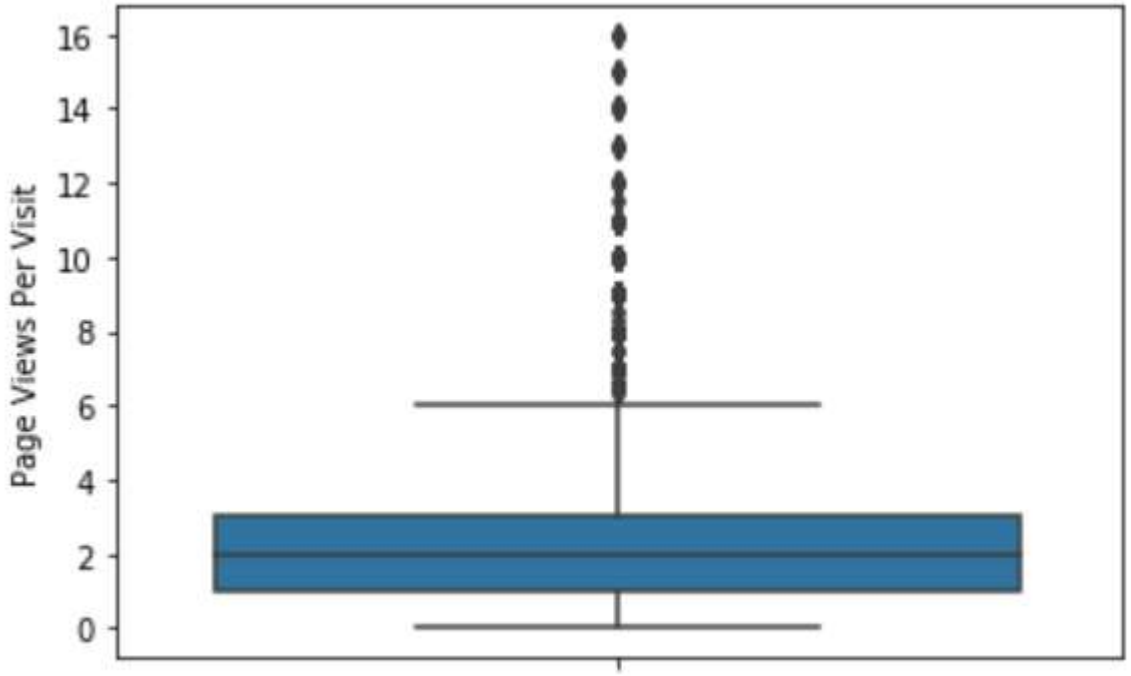
- The numerical columns are visualized with the target variable “converted” to understand how each numerical column behaves for a converted lead and for a non converted lead.



- The values in the binary columns are converted to 0 and 1
- Dummy variables are created for all the remaining categorical variables.

- For ‘Total Visits’ and ‘Page views per visit’, majority of the outliers are continuous and only few of them are actually outliers. So all the extreme outliers are removed based on custom threshold values. Below is the box plot that shows those two columns after removing the outliers.





Step2: Model Building

- The dataset is split into train set and test set and feature scaling is performed for numerical columns using standard scaler.
- The number of columns in train set is initially reduced to 15 using the automated approach of RFE.
- Using the columns that RFE gives, an initial model is built. This model shows that p-values of all the columns are zero but the VIF values are pretty high. So the columns with high VIF are dropped one by one and a model is built after each column is dropped and new VIF values are verified.

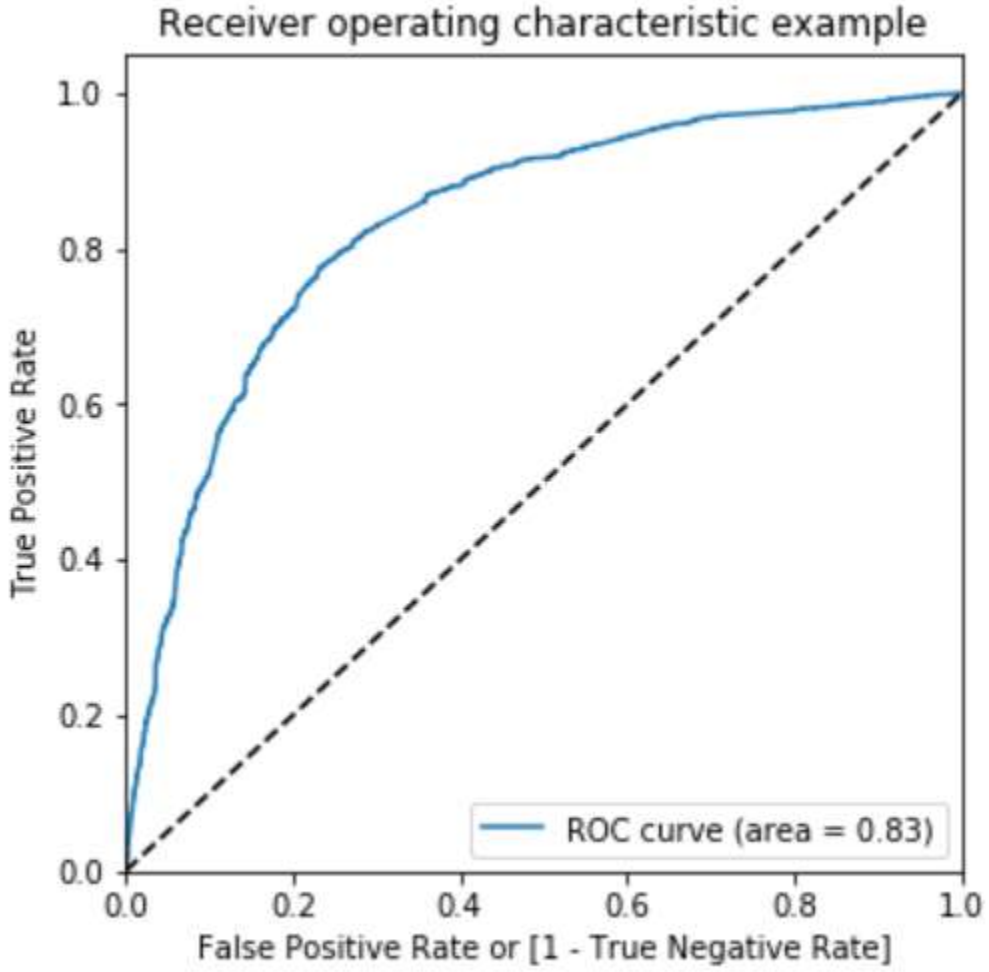
'Last Notable Activity_Email Opened', 'What is your current occupation_Unemployed', 'Last Notable Activity_SMS Sent', 'Lead Origin_Landing Page Submission'

- The above columns are dropped based on VIF values, and after this, the p-values of 'Lead Source_Olark Chat' and 'TotalVisits' are increased. So 'Lead Source_Olark Chat' is dropped and after that 'TotalVisits' is also dropped. A new model is built on remaining columns.
- At this point, the VIF values and p-values looks good and this model is finalized.

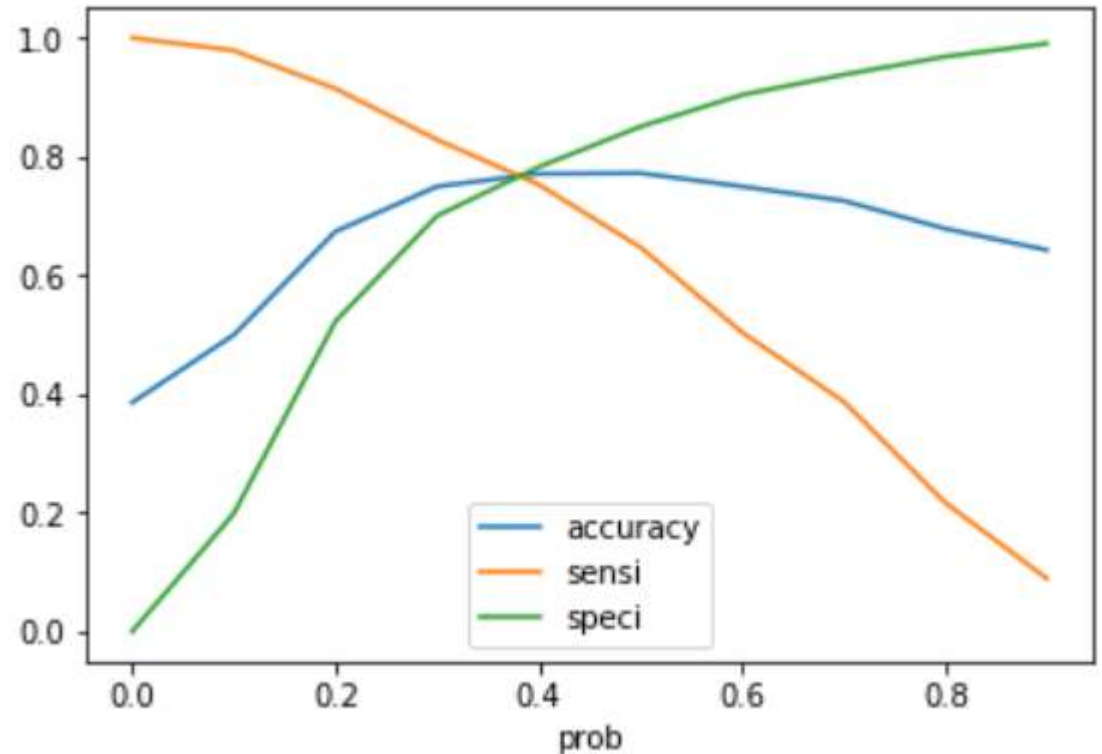
Step3: Model Evaluation

- Prediction of probabilities of lead conversion (target variable) are calculated based on the train set
- Probabilities of prediction are converted into a lead value (0 or 1) based on a randomly chosen cut off value (0.5)
- For these predictions, confusion matrix is evaluated and sensitivity(or recall), precision, f1 score are calculated.
- At this point, sensitivity of the model is less because the probability cut off value is randomly chosen. The sensitivity will be increased once the ROC curve is plotted and an optimal cut off value is calculated.

- ROC curve is plotted and optimal cut off value is found to be 0.3. Below graphs show the ROC curve accuracy, sensitivity, specificity for various probabilities.



	prob	accuracy	sensi	speci
0.0	0.0	0.385765	1.000000	0.000000
0.1	0.1	0.499611	0.978603	0.198783
0.2	0.2	0.673727	0.914413	0.522566
0.3	0.3	0.749727	0.828421	0.700304
0.4	0.4	0.771064	0.752927	0.782454
0.5	0.5	0.772154	0.647558	0.850406
0.6	0.6	0.749727	0.504239	0.903905
0.7	0.7	0.725588	0.387566	0.937880
0.8	0.8	0.678555	0.217198	0.968306
0.9	0.9	0.642735	0.089221	0.990365



- The optimal cut off point is 0.4 but as we require more sensitivity for our model (because we can afford to call some customers who are probably not going to convert but we can not afford to loose those customers that are definitely going to convert), 0.3 is chosen.
- In addition to this, a lead score is calculated for each lead based on its probability value. It is basically a value that tells what is the percentage of likeliness that a lead is going to convert.
- Confusion matrix is evaluated for the final model and at this point, the sensitivity of thee train set came out to be 82.84, precision is 63.45, f1 score is 71.86

Step4: Making Predictions on Test Set

- Numerical columns are scaled before making predictions.
- Probability values are predicted on test set using the same columns as the final model and a lead score is calculated for test set as well
- Confusion matrix is evaluated for test set and the sensitivity of the test set came out to be 83.55, precision is 62.74 and f1 score is 71.67
- Now a ‘final_scores’ data frame is made with lead scores of train and test sets combined including the row ids of each score. This row id indicates which lead the score refers to in the original data frame.
- Now the ‘final_scores’ data frame is sorted based on the row ids so that it can be the scores can be added as a new column in the entire data frame.

Business Insights of the model

- ‘Last Activity_SMS Sent’, ‘Total Time Spent on Website’, ‘Last Activity_Email Opened’ are the most important variables that sales team must focus on in order to convert a lead.
- ‘Lead Source_Direct Traffic’, ‘Do Not Email’, ‘Last Notable Activity_Modified’ are the three top negative variables for a lead conversion.
- Each lead with a lead score of more than 30 is considered a hot lead. The sales team must concentrate on those that have lead score of greater than 30.