Lead scoring case study

Upgrad 2023

Problem Statement

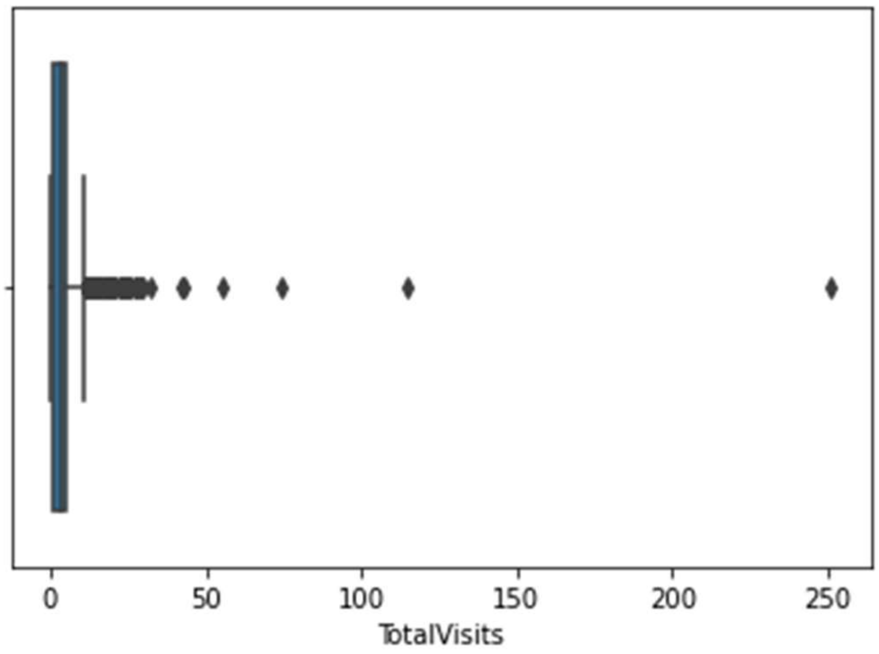
- An Education company named X Education sells online courses to industry professionals. Also, company has been marketing its courses in various websites & search google.
- People who fill up a form by providing their email address & phone number becomes a lead
- Based on multiple emails & phone calls, around 30% of leads gets converted into paying customers
- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, we must nurture the potential leads to get a higher lead conversion.
- The company's target is to get a ballpark of target lead conversion rate of 80%.
- Based on the given datasets, we must build a model wherein we 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.
- Also, our target variable is 'Converted' column which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

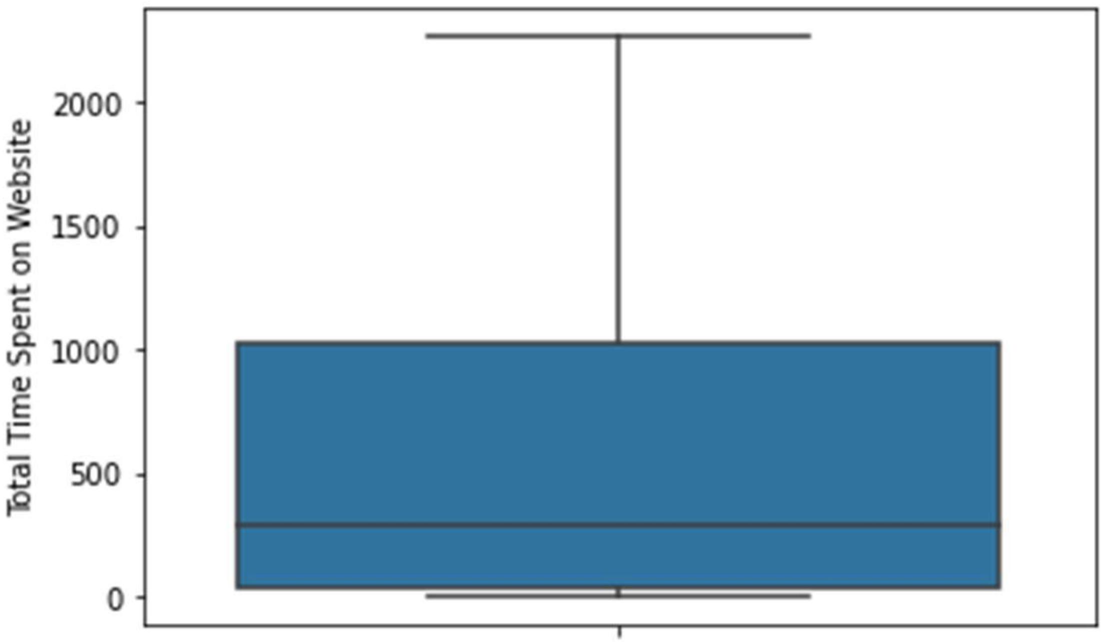
Analysis Approach - I

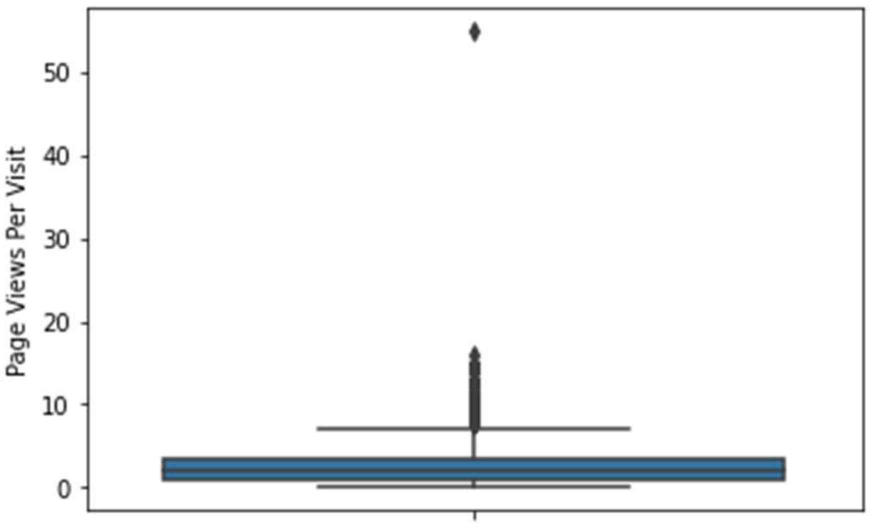
- Our first and main approach was to run all necessary packages as per need, post we uploaded the datasets in Jupyter Notebook.
- Post that understanding the datasets & statistical values and its descriptions and data types
- We found that few columns have null values with the help of info command.
- After running null command for the entire datasets, we could see some columns have 40-50 % of rows with null values. Imputing that we will not be good approach, so we decided to drop those columns from our datasets.
- We dropped Tags & Assymetric columns from our dataframe
- Next, we moved on with the other columns where null values were not high, so we decided to remove the rows from our datasets.
- Also, variables which has high count Select values : those columns has been dropped as well. It will act as biased for our datasets.
- After cleaning the datasets & removing unwanted columns like Prospect id, lead number which will not be used for our analysis.
- We moved on with our Exploratory data analysis by plotting few graphs for continuous & categorical variables & categorical variables with our target variable.
- Here, we investigated the quality of data where we find correlations, trends and outliers in my datasets.

Analysis Approach - I

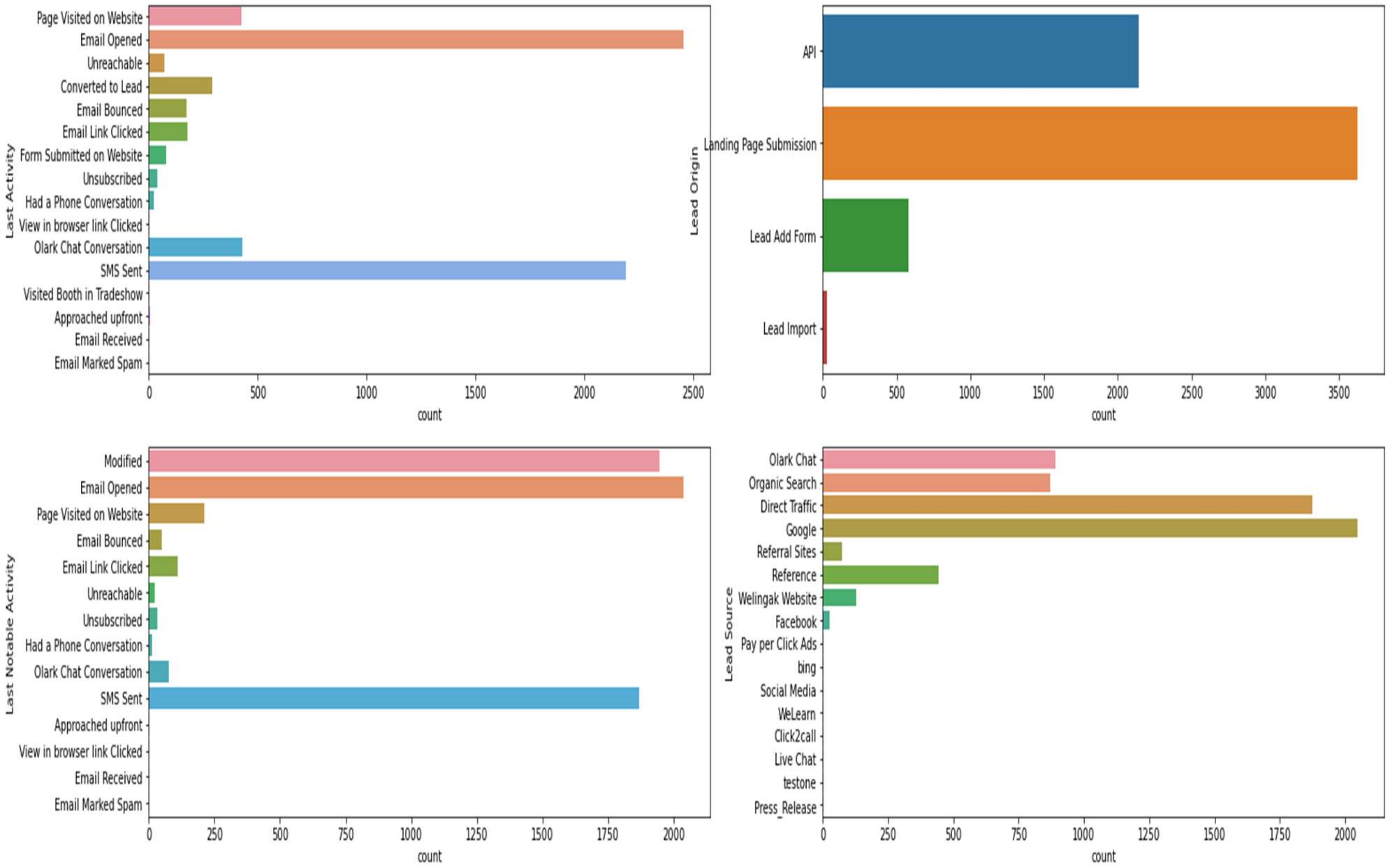
- Next, we started with pre-processing the datasets. Since it’s a classification problem we decided to use machine learning techniques for our datasets by building logistic regression model.
- Firstly, our data preprocessing steps we had to create dummies for categorical variables. By making them into continuous variables which will helps us to include in our data model.
- Next, we splitted the datasets with 70% of train data and 30% of test data.
- We will build our model based on train datasets & evaluate on testing datasets to check its accuracy
- After train-test data split, we had to scale our continuous variables where all variables can be in same measurement terms. We used min-max scaler approach.
- Post that, we trained our model using Automatic RFE Approach since we had a huge number of variables post dummies creation. Based on RFE approach, we went ahead with manual elimination of variables using p value method or VIF method.
- Variables which has high p-value were eliminated & Variables which has high VIF > 5 were eliminated from the model one by one.
- Since it’s a logistic regression model, we had to derive at optimal probability value for our model.
- We used to two approach to get the optimal cut off for our model i.e ROC curve with plotting graph for accuracy, specificity and sensitivity of their probability
- Next, we used Precision-recall trade off view to arrive at the optimal point
- At the end, we test the model on our test datasets to check its accuracy score, confusion matrix, sensitivity & specificity scores.



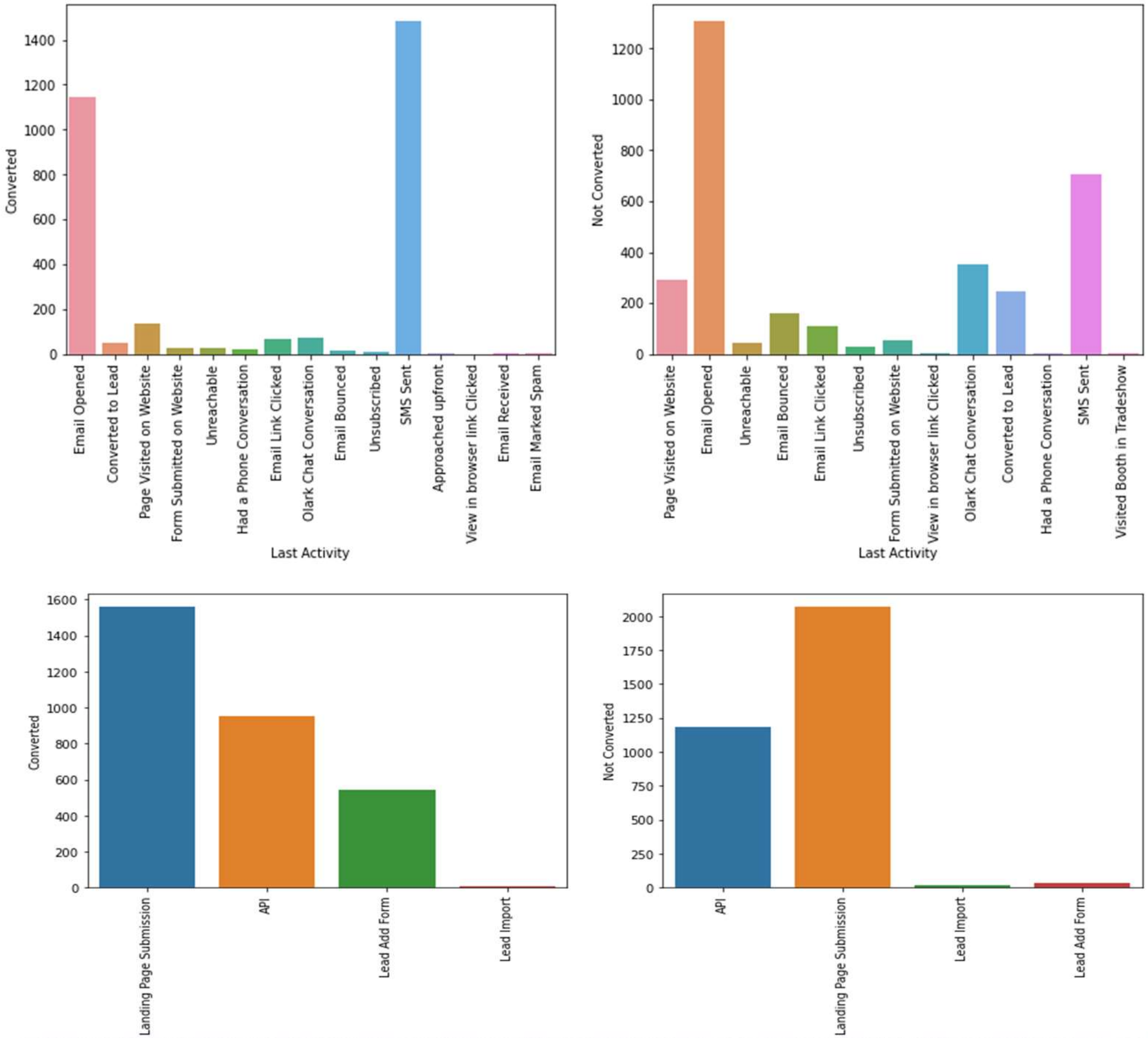




- Plotted boxplots for continuous variables. Except for Total time spent on websites, we found some outliers in Total visits & Page views per visit. That needs to be treated accordingly, since all 3 variables are relevant for the model.



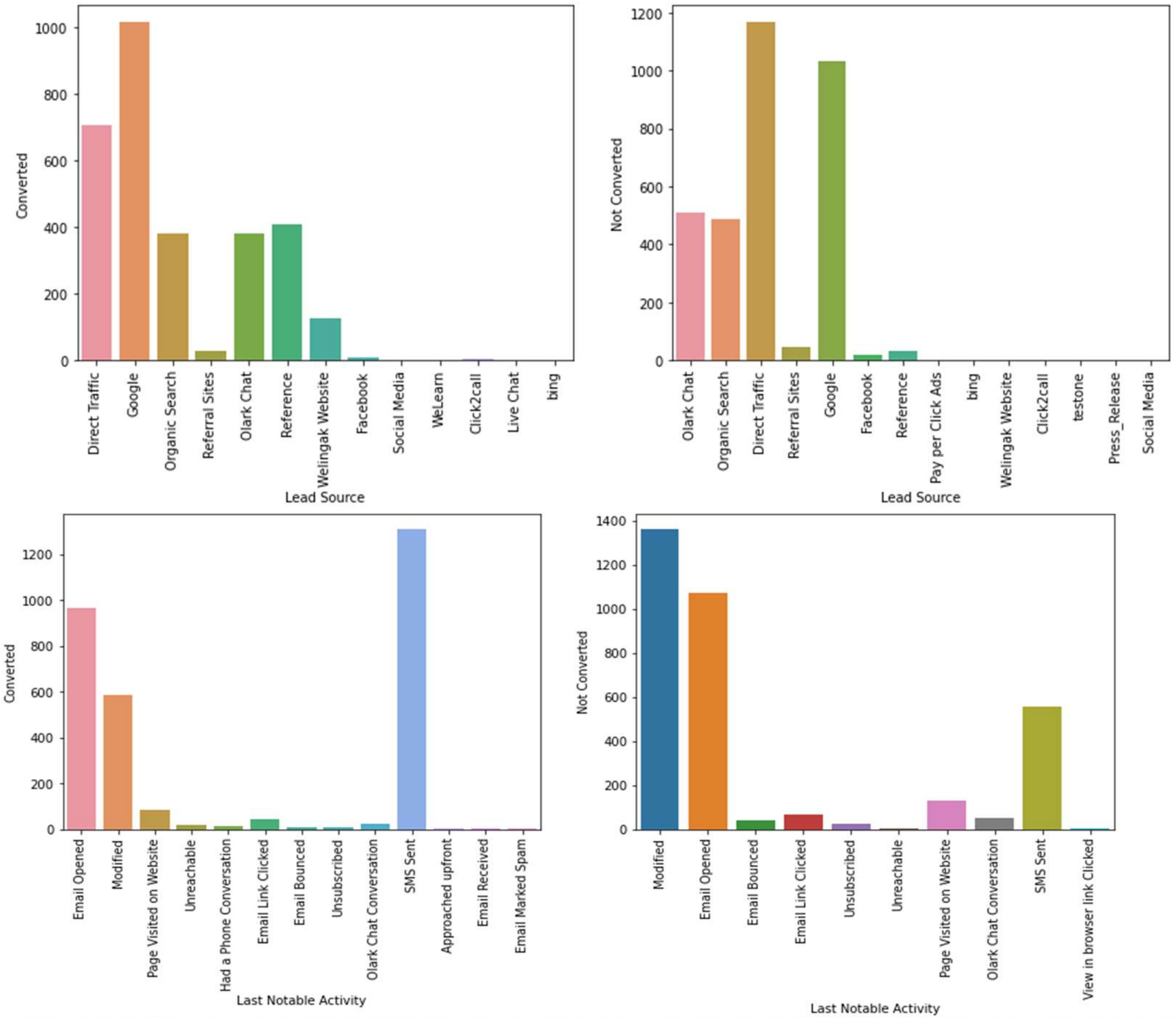
- Majority of lead origins are of Landing page submission
- Majority of Lead source are from Google
- Most of the lead's last activity was Email Opened and SMS sent



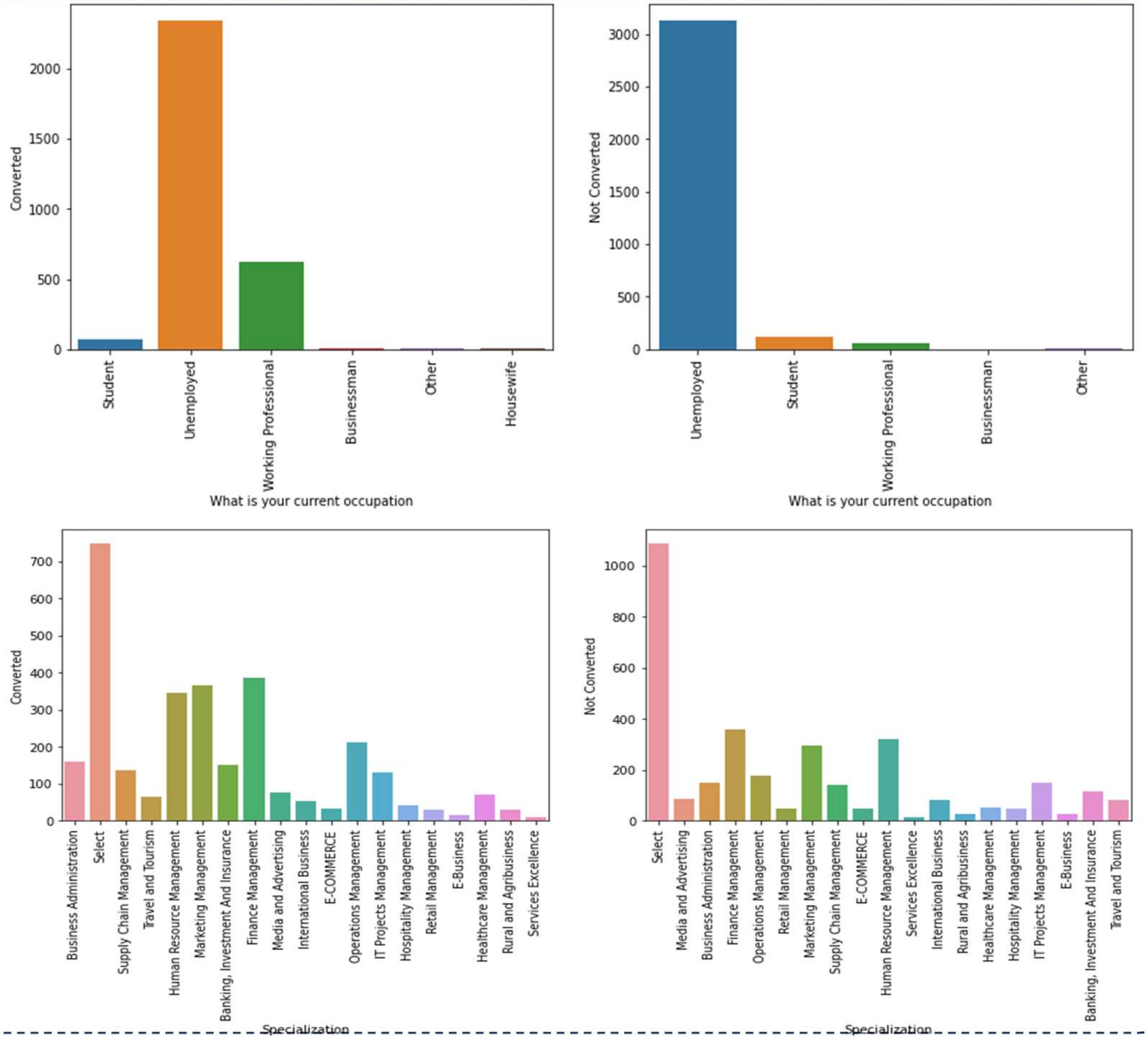
- For Non converted ones : Email opened is the last activity done by user has the highest count.

- For converted ones : SMS sent is the last activity done by user has the highest count.

- For both converted & non converted : Landing page submission has the highest count for Lead origin

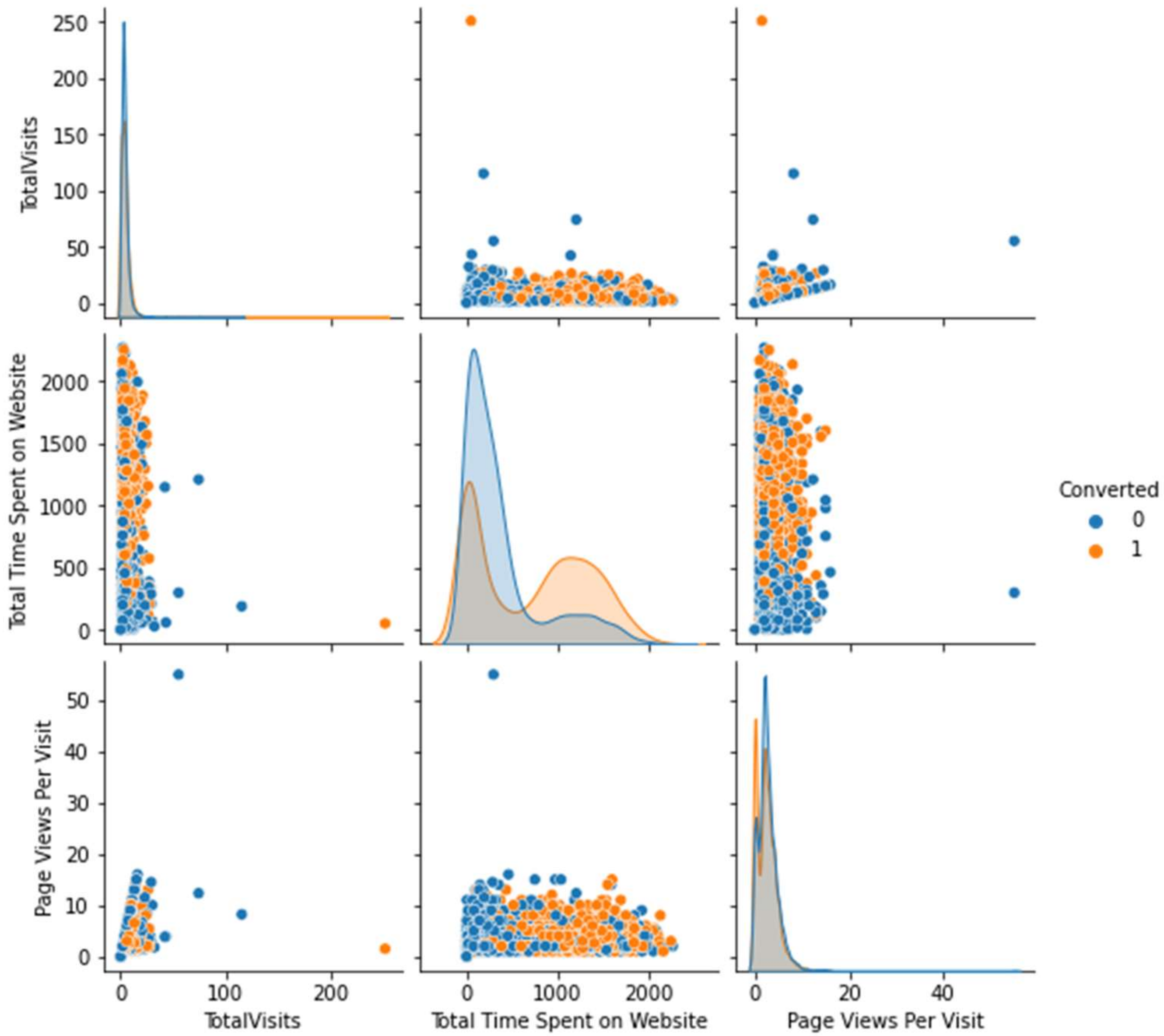


- Google is the highest ones for the leads getting converted
- For Non converted ones: lead source is Direct traffic
- Modified has the highest count for non converted in last notable activity
- SMS sent stand as highest count for converted in last notable activity



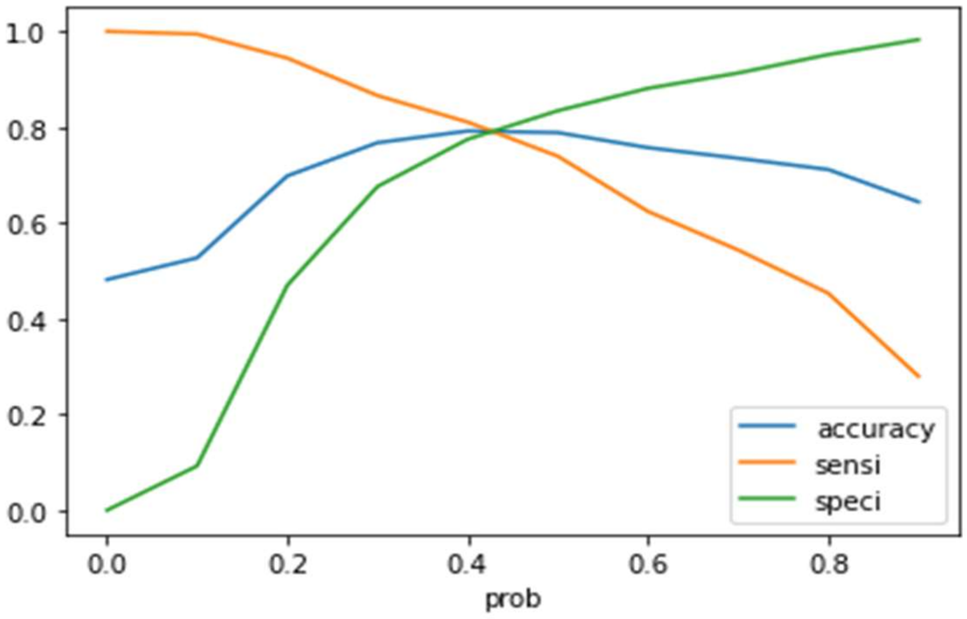
- Current occupation for both converted & non converted belongs to Unemployed

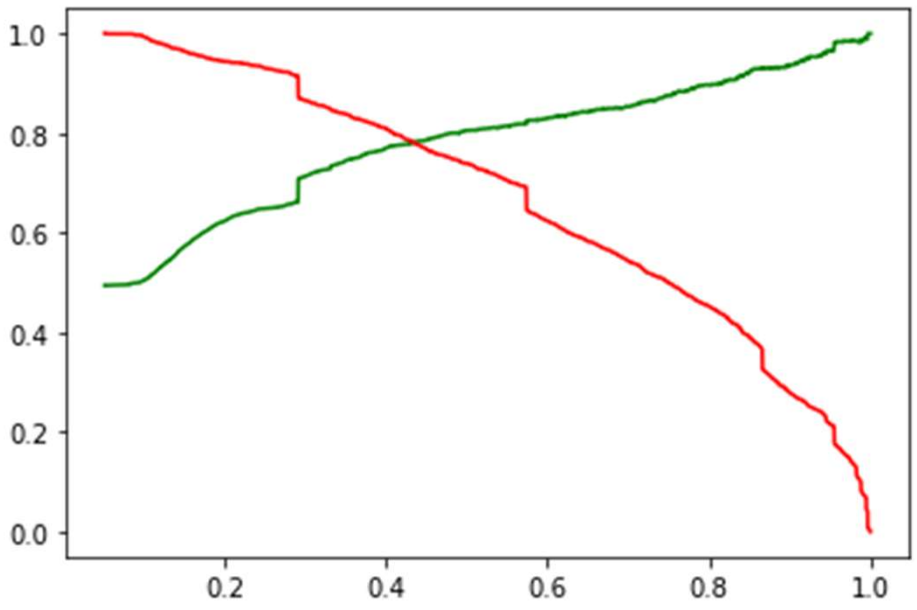
- Specialization for both Converted and Non converted ones belong to Finance & Marketing management



- There doesn’t seem to be a unique relationship between continuous variables for both converted & non converted ones.

- But Total visits & Page views per visit seems to show an upward trend indicating a positive relation.





By plotting the probabilities for Accuracy, sensitivity & specificity : we receive an optimal cut off 0.42 whereas trade off between precision and recall gave an optimal point of 0.44. Let's look at the below table:

Prob of Accuracy, Sensitivity and Specificity	Train : 0.42	Test : 0.42	Precision/Recall trade off	Train : 0.44	Test: 0.44
Accuracy score	0.79	0.78	**Accuracy score**	0.79	0.79
Sensitivity	0.79	0.78	**Precision**	0.78	0.78
Specificity	0.79	0.79	**Recall**	0.78	0.77

Business Results and Recommendations

- Its very important for the company to target those leads where they spend more time on the website and total number of visits to website is also more. There is a high chance of leads getting converted into prospect customers.
- Students & Unemployed leads should not be targeted as they won't lead the company to higher conversion rate.
- Olark chat & Welingkak website as a lead source should be targeted more.
- The lead origin identifier as “Lead add form” should be targeted more & it will give more business to the company.