LEAD SCORING ASSIGNMENT

ARUSHI SHREE

ABHISHEK PORWAL

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.

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.

Objective : Company requires objective is to 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. i.e. the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

Analysis Approach

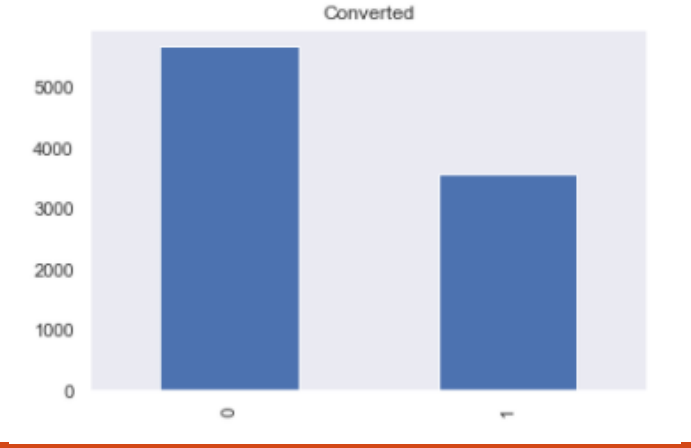
We'll perform **Logistic Regression** on the given data and generate the lead scores using the converted probabilities.

Steps to be performed:

- Reading and Understanding the data
- Data Cleaning
- Data Visualization
- Data Preparation
- Model Building
- Model Evaluation

Reading and Understanding the data

- Reading the dataset and understanding different variables and their data types.
- Get a brief idea how data is distributed in various columns, by checking the standard quantile values by describe() method.
- By checking the dataset we get to know that the data set consist of 9240 rows and 37 columns.
- The target variable i.e. “Converted” have around 38% of conversion rate.
- The data consist of many null values as well which had to be rectified in the further data preparation section.
-



Data Cleaning:

- In this part certain column values are checked for null or absurd values.

- Columns having null values greater than 3000 are dropped.

- Columns having skewed data (majority of same values) are dropped.

- Columns with null values less than 3000 are treated by mode/mean value replacement depending on if column is categorical or continuous.

- Few columns have value as “Select” which is absurd and depending upon count of values, these values are treated accordingly either replacing them my MODE of column value or by defining another value as “OTHERS”.

-Columns having very low numbers of null values. Rows having the null values are dropped in the data set.

-Resulting data set is having around 98% of data remaining for further analysis.

Dataset after Cleaning

Data set columns remaining after cleaning of data :

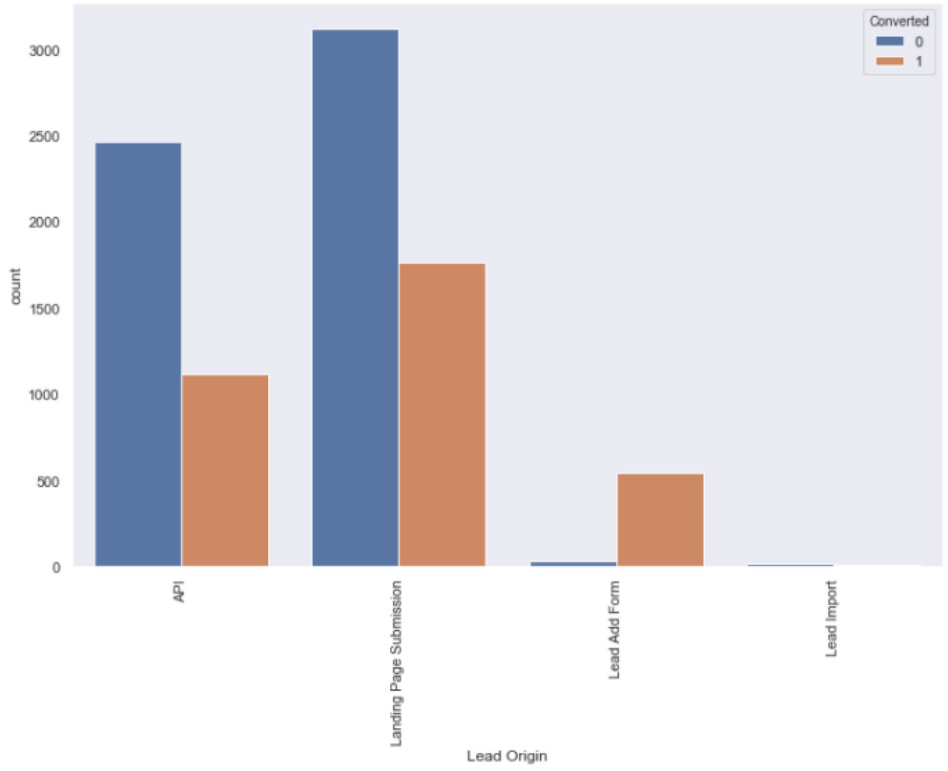
	Prospect ID	Lead Number	Lead Origin	Lead Source	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	A free copy of Mastering The Interview	Last Notable Activity
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	0	0.000000	0	0.000000	Page Visited on Website	Others	Unemployed	No	Modified
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	0	5.000000	674	2.500000	Email Opened	Others	Unemployed	No	Email Opened
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	1	2.000000	1532	2.000000	Email Opened	Business Administration	Student	Yes	Email Opened
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	0	1.000000	305	1.000000	Unreachable	Media and Advertising	Unemployed	No	Modified
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	1	2.000000	1428	1.000000	Converted to Lead	Others	Unemployed	No	Modified

Data Visualization:

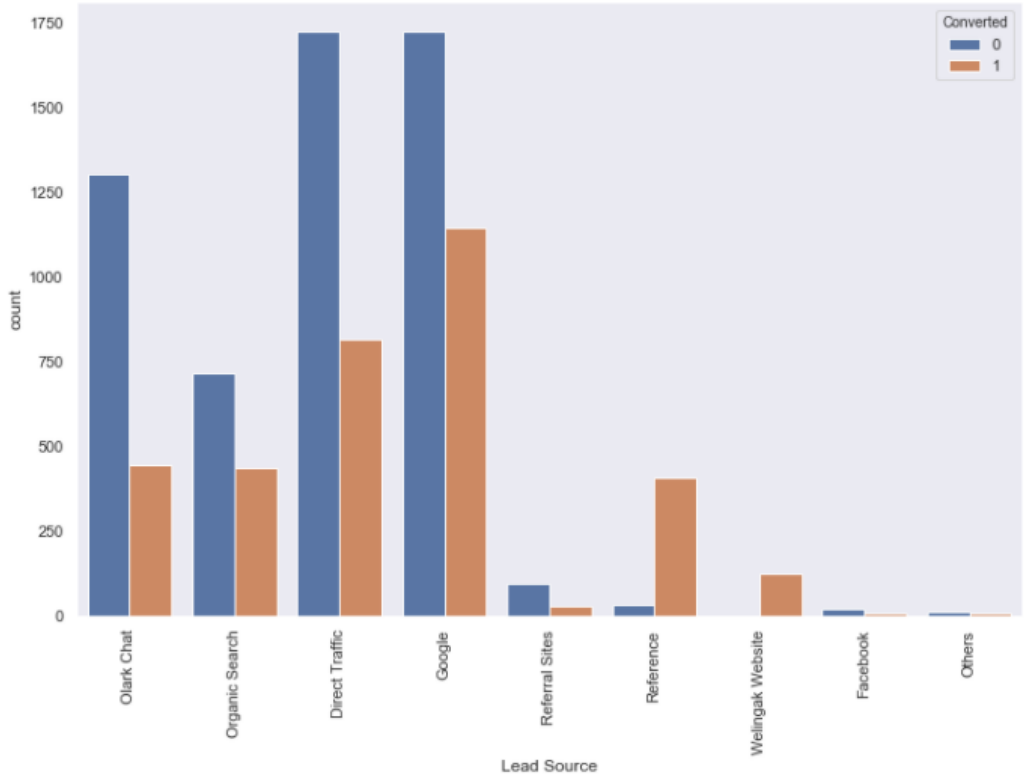
- A brief distribution of data can be visualized by plotting the count plots and boxplots with respect to our target variable.

- Outliers are also visualized for continuous variables and treated by capping them on a suitable quantile value.

Data Visualization:

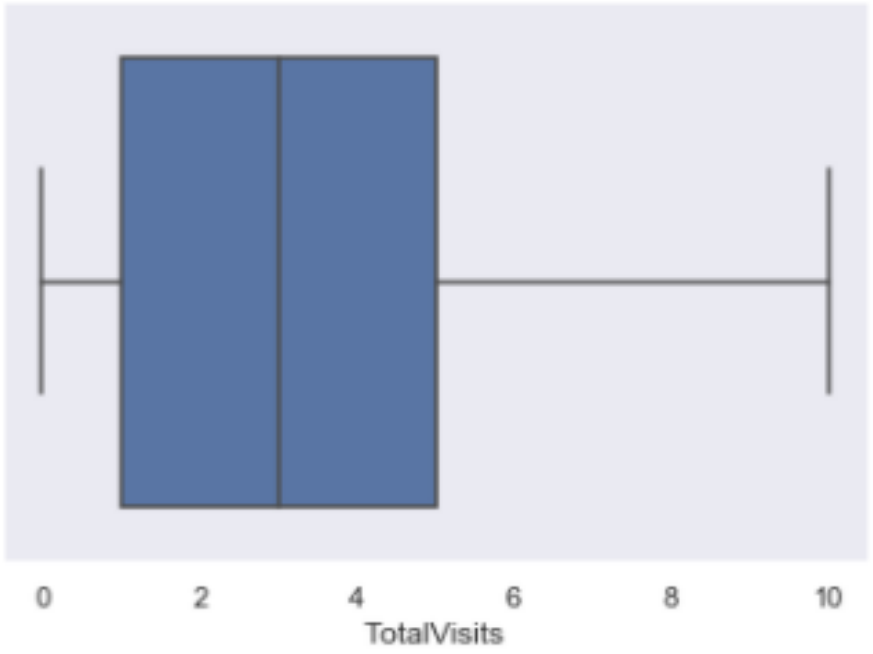


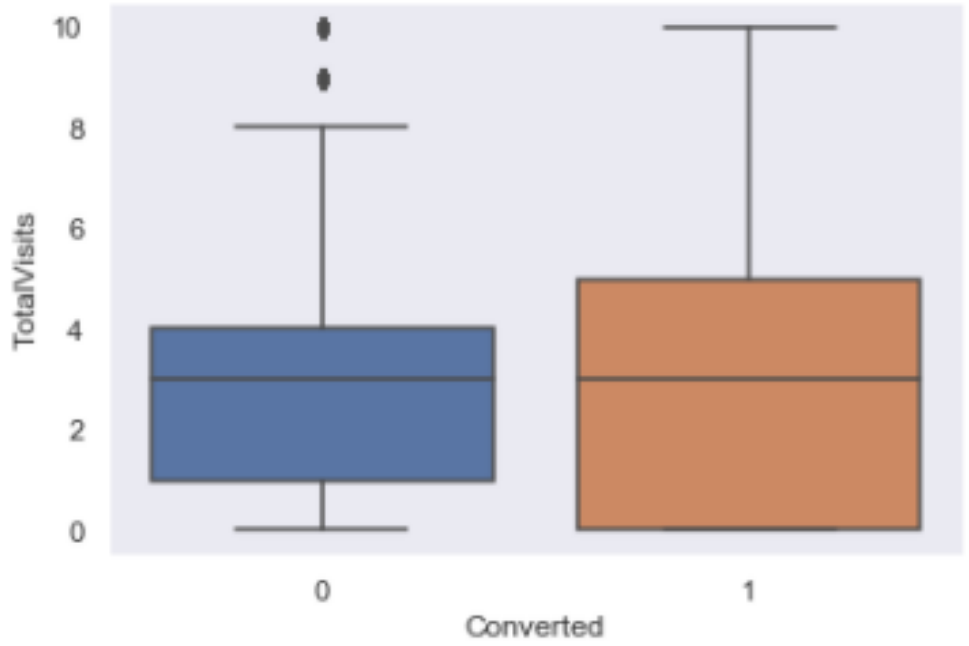
Majorly the leads are generated by API and Landing Page, team should focus on converting these leads.**
Also, Lead Add Form have most leads getting converted.**



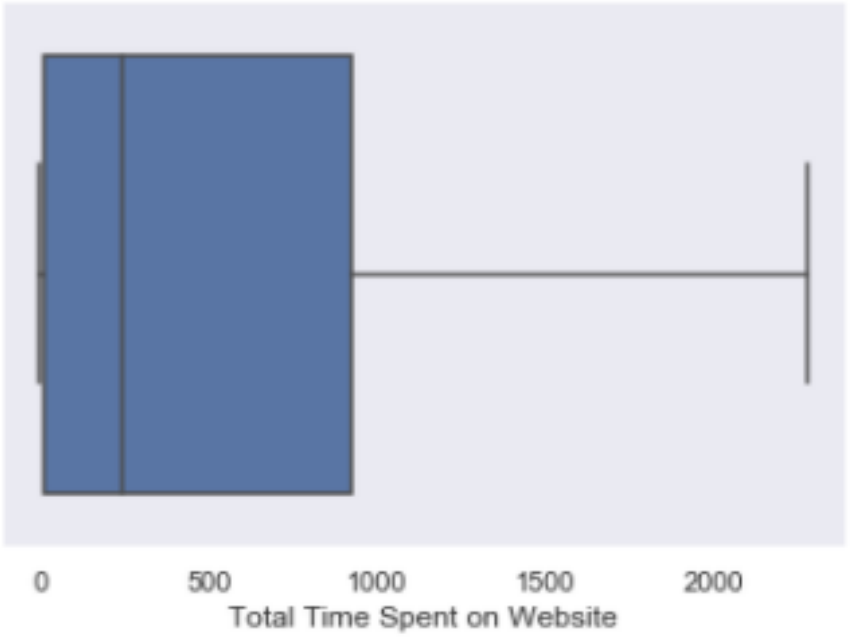
- Major leads are generated by Direct traffic and Google, team should focus on how to convert these.
- Leads are successfully getting converted generated via References and Welingak website.

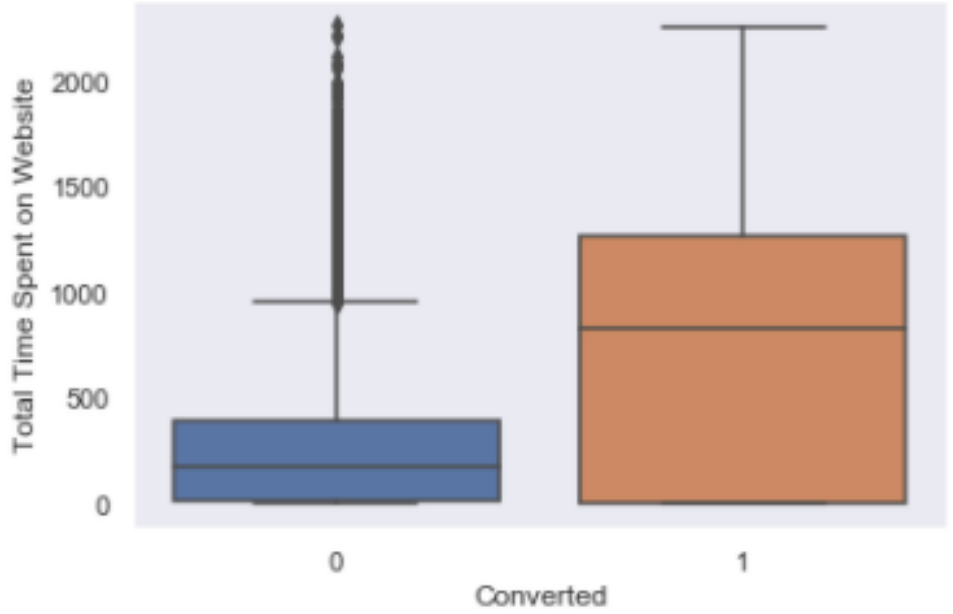
Data Visualization:





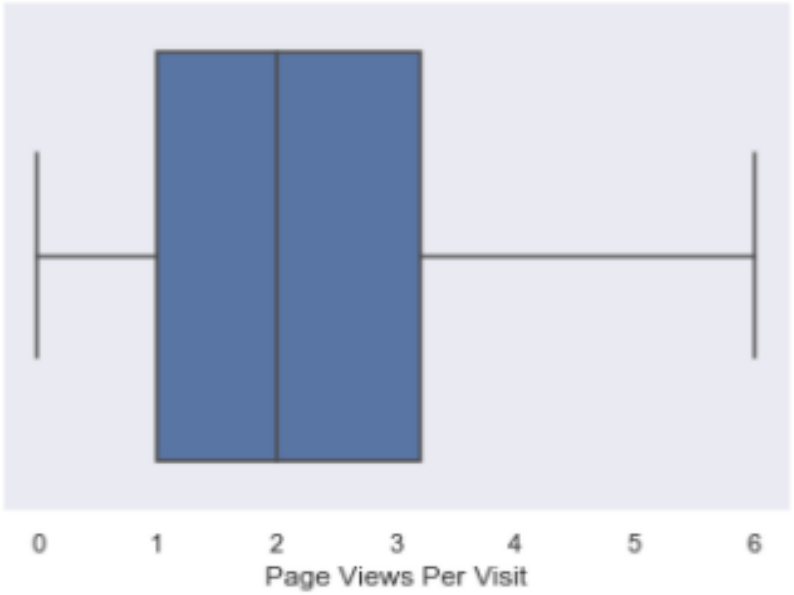
Data Visualization:

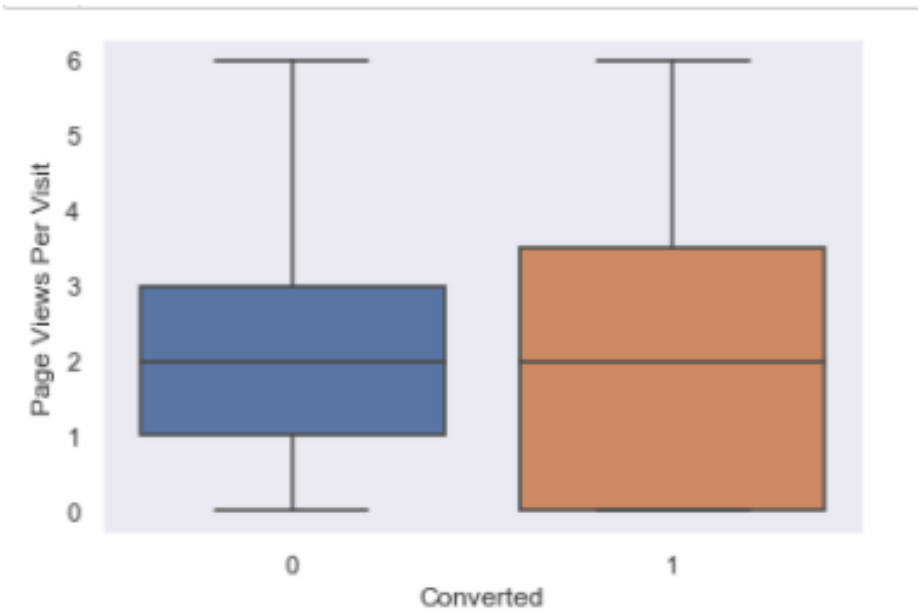


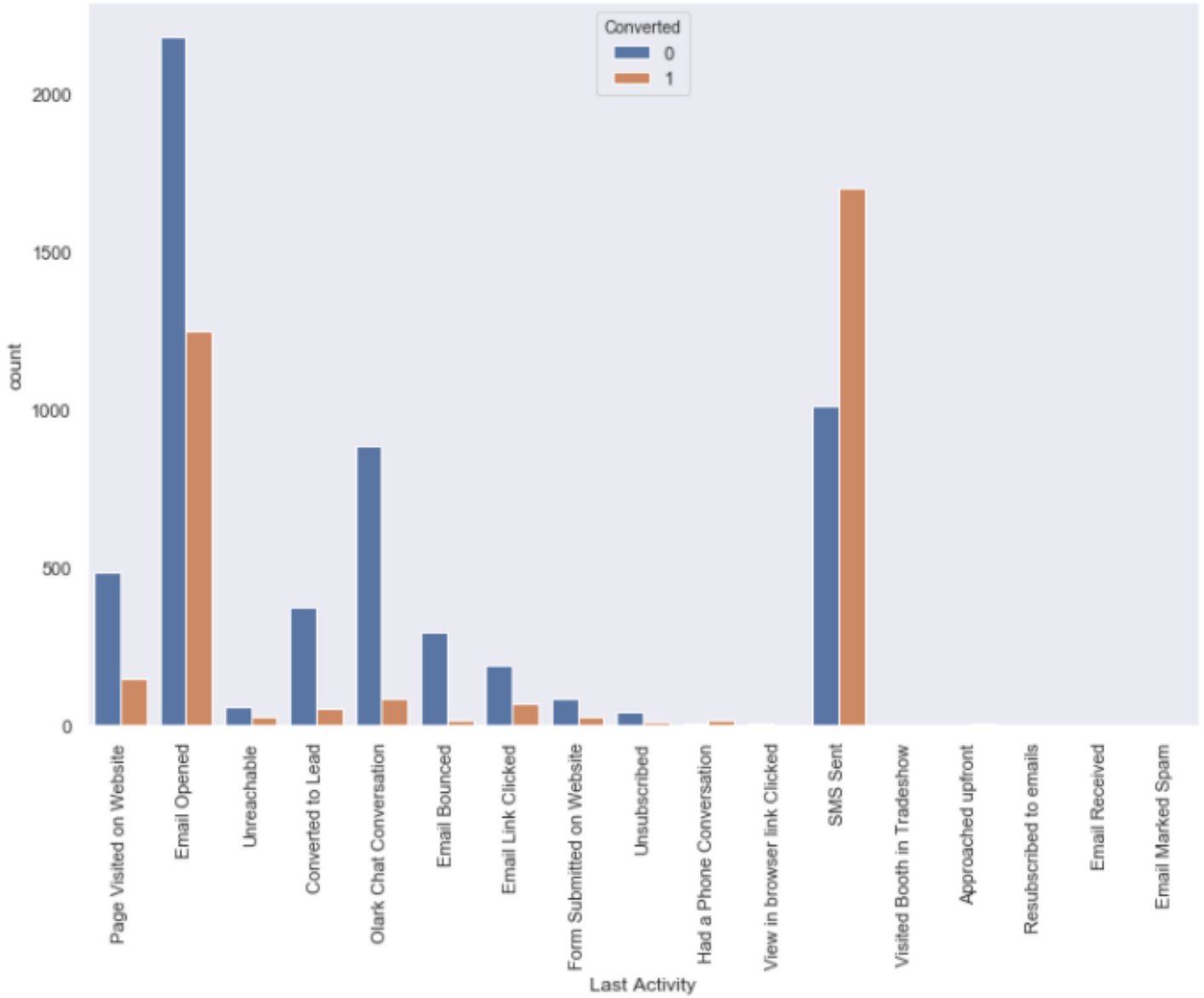


Its clear that, converted leads are those who spends a lot time browsing the and spending a lot of time over the platform.

Data Visualization:

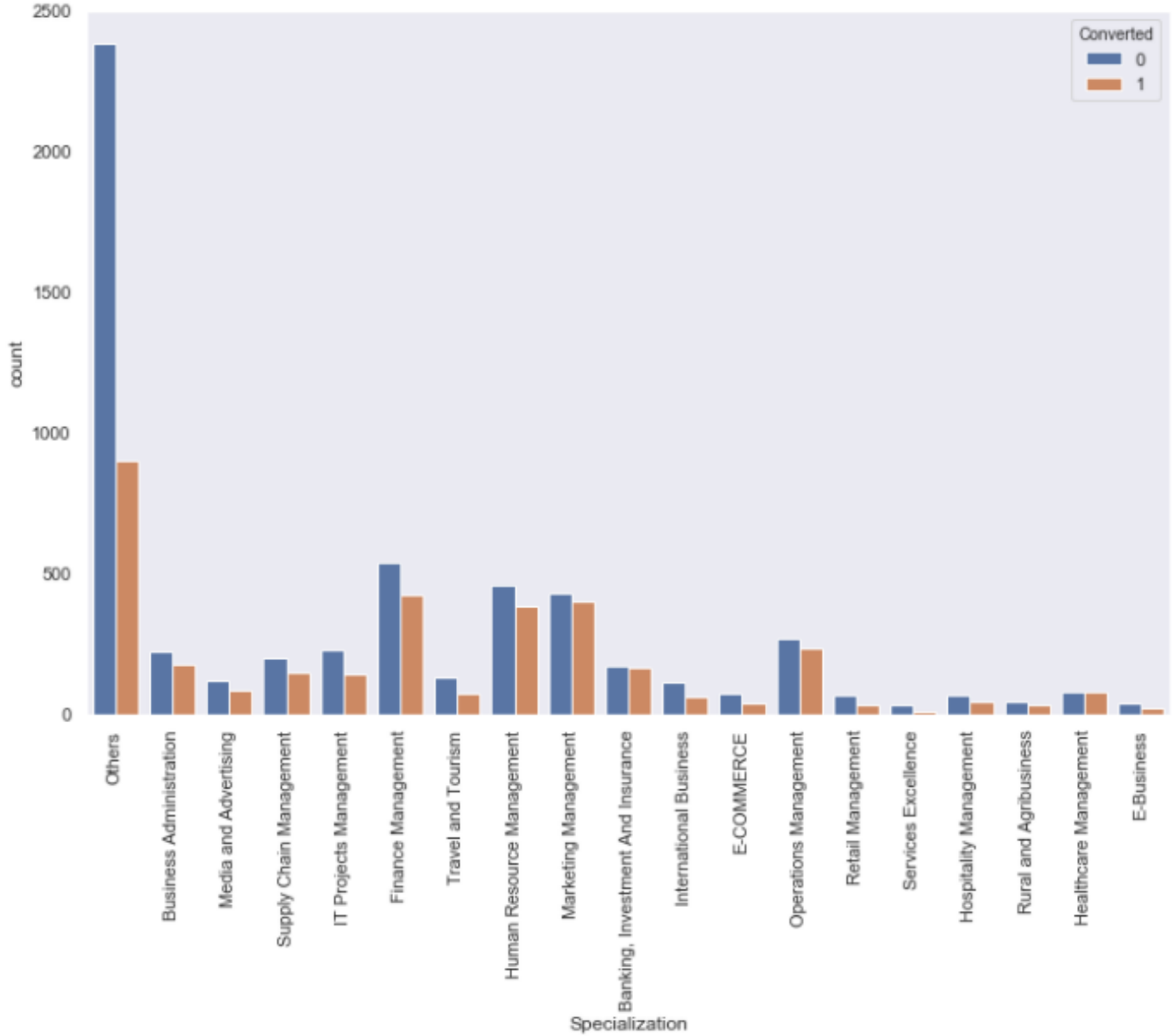






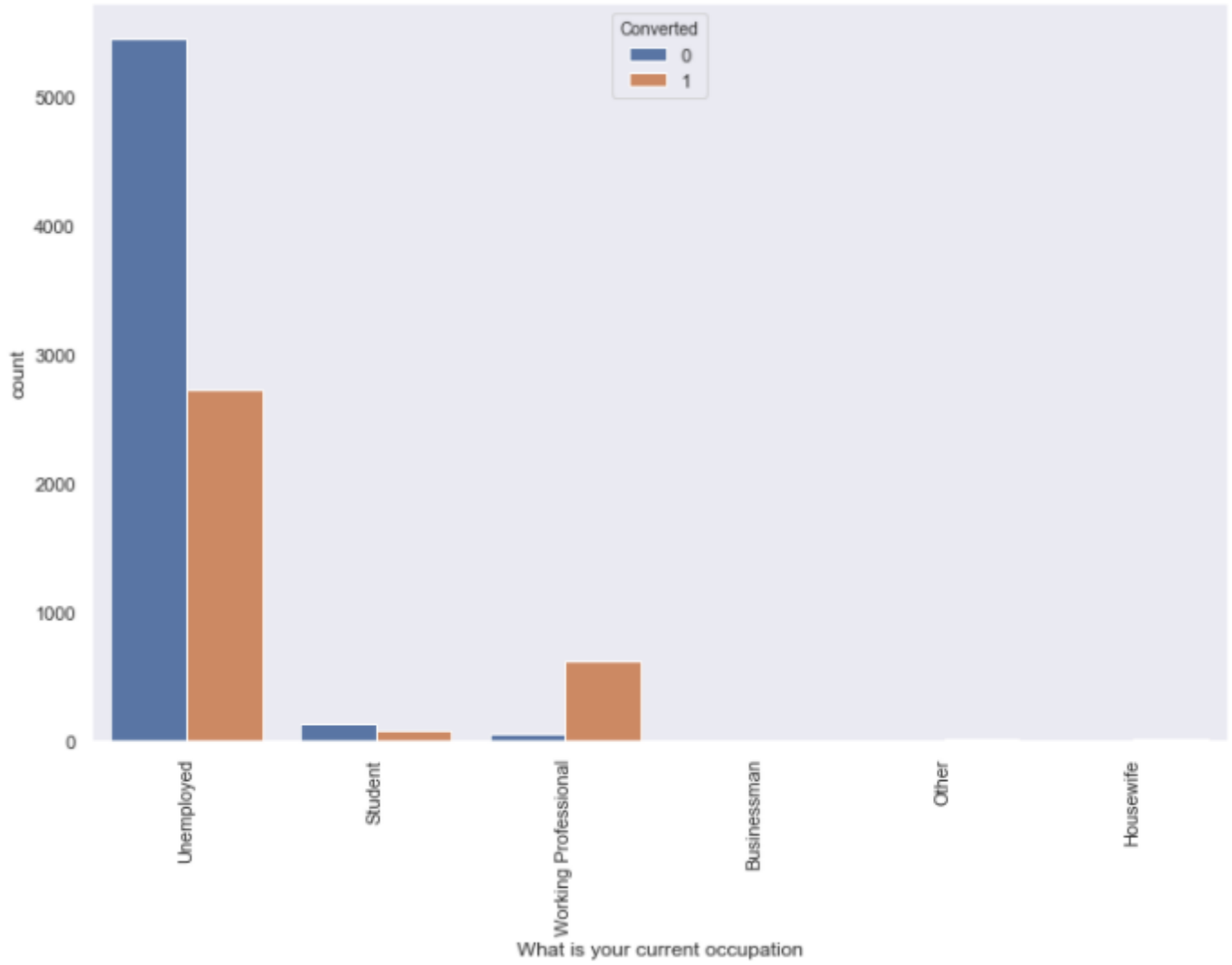
Data Visualization:

- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is almost 60%.



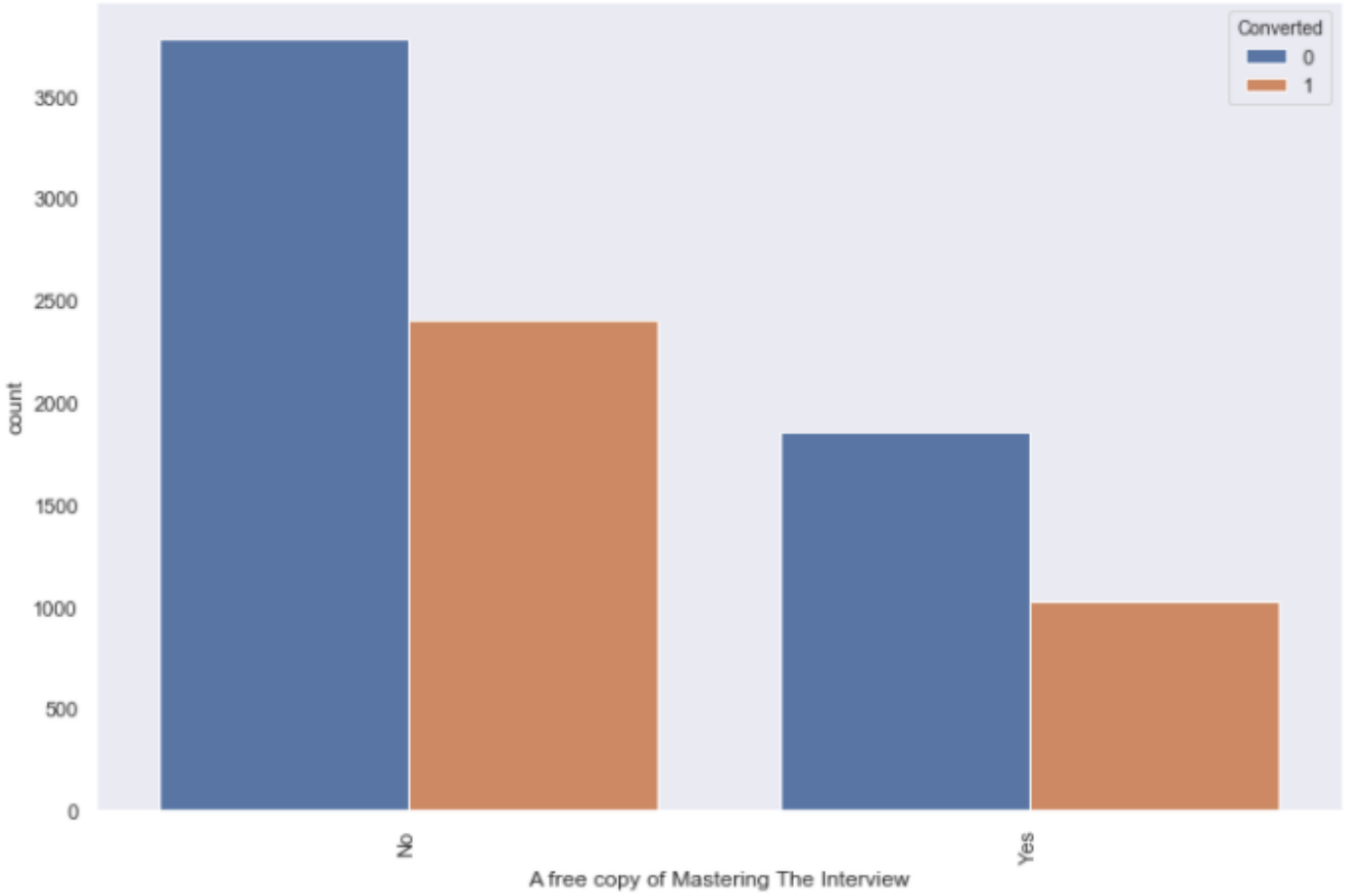
Data Visualization:

Focus should be more on the Specialization with high conversion rate.



Data Visualization:

- Unemployed leads are the most in numbers but have moderate conversion rate.
- Working Professionals going for the course have high chances of joining it.



Data Visualization:

- Those who didn’t subscribe for a free copy have greater lead conversions
- count of leads wanting a free copy is nearly half of those who opt for "NO".

Data Preparation

- Data is prepared by creating the dummy variables from the parent variables (categorical).

- Columns considered for dummy variable creation are ['Lead Origin', 'Lead Source', 'Last Activity’, 'What is your current occupation’, 'A free copy of Mastering The Interview', 'Last Notable Activity’].

- Dropping the parent categorical variables.

-Plot heatmap on correlation to get brief idea regarding redundancy.

- Making split of data set for Logistic modelling into train and test data(70:30 train-test split).

- Scaling the continuous variables by using Standard Scalar to analyze variables on a same scale magnitude.

Model Building

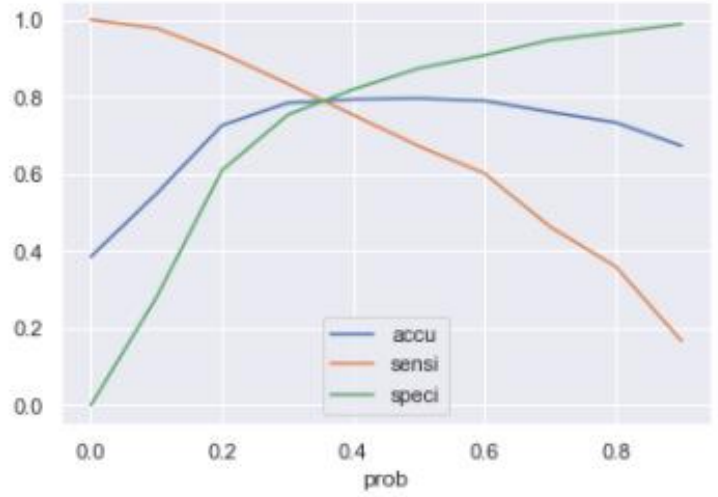
- Initializing the model building process by feature selection process which is done firstly by automated (R.F.E.) and taking the top 15 features.

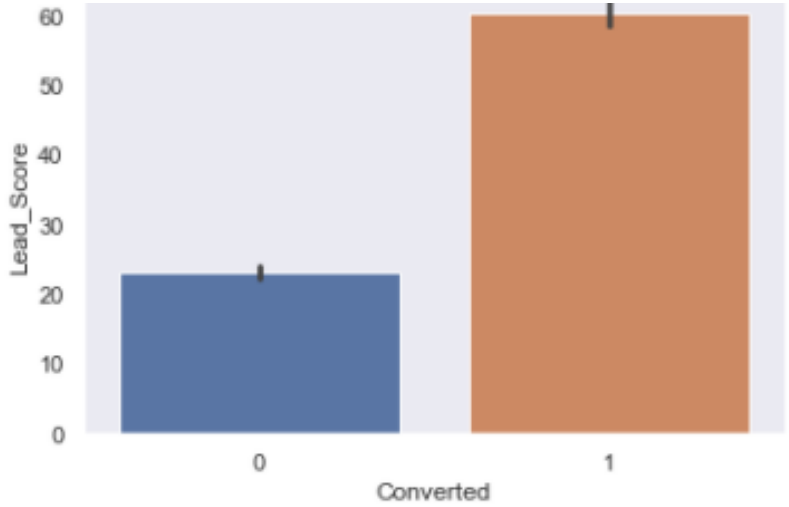
- The above top 15 features are then analysed manually by iterating for p-value and V.I.F. for each remaining feature.

- After manual and automated feature selection, we are left with top 10 features to build on our model upon.

- Below are the top 10 features selected:

```
Lead Source_Olark Chat
Last Activity_Olark Chat Conversation
Total Time Spent on Website
Last Notable Activity_SMS Sent
Lead Origin_Lead Add Form
Last Activity_Email Bounced
Lead Origin_Lead Import
Last Activity_Converted to Lead
Last Activity_Had a Phone Conversation
Last Notable Activity_Unreachable
```





Model Evaluation:

- Initial model evaluation is carried out by taking a probability cut-off point of 0.5, above which the model converts the leads and vice-versa.

- Confusion matrix is prepared and various parameters like Accuracy, Sensitivity, Specificity, Precision and Recall are checked.

- ROC curve is plotted to determine goodness of model which here is 0.86.

- Optimal threshold is found by plotting the line plot for accuracy, sensitivity and specificity whose intersection gave us the cut-off value of 0.3.

- By above threshold value, the successful conversion is given for all predicted probabilities >0.3.

- The Lead Score is then calculated as a percentage value of the predicted probabilities.

- From the Conversion distribution, we concluded that all the Leads having Lead Score >=60 can be considered as Hot Leads.

THANK YOU!