Lead Score Case Study

Team Members :

1. Vishwas Yedidya

2. Rijuta Wankar

Problem Statement

- 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.
- 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.

Business Problem:

- Find the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- Build a predictive model where a customer with high lead score have a high conversion chance and low score have low conversion chance
- The conversion rate should be 80 %

Methodology:

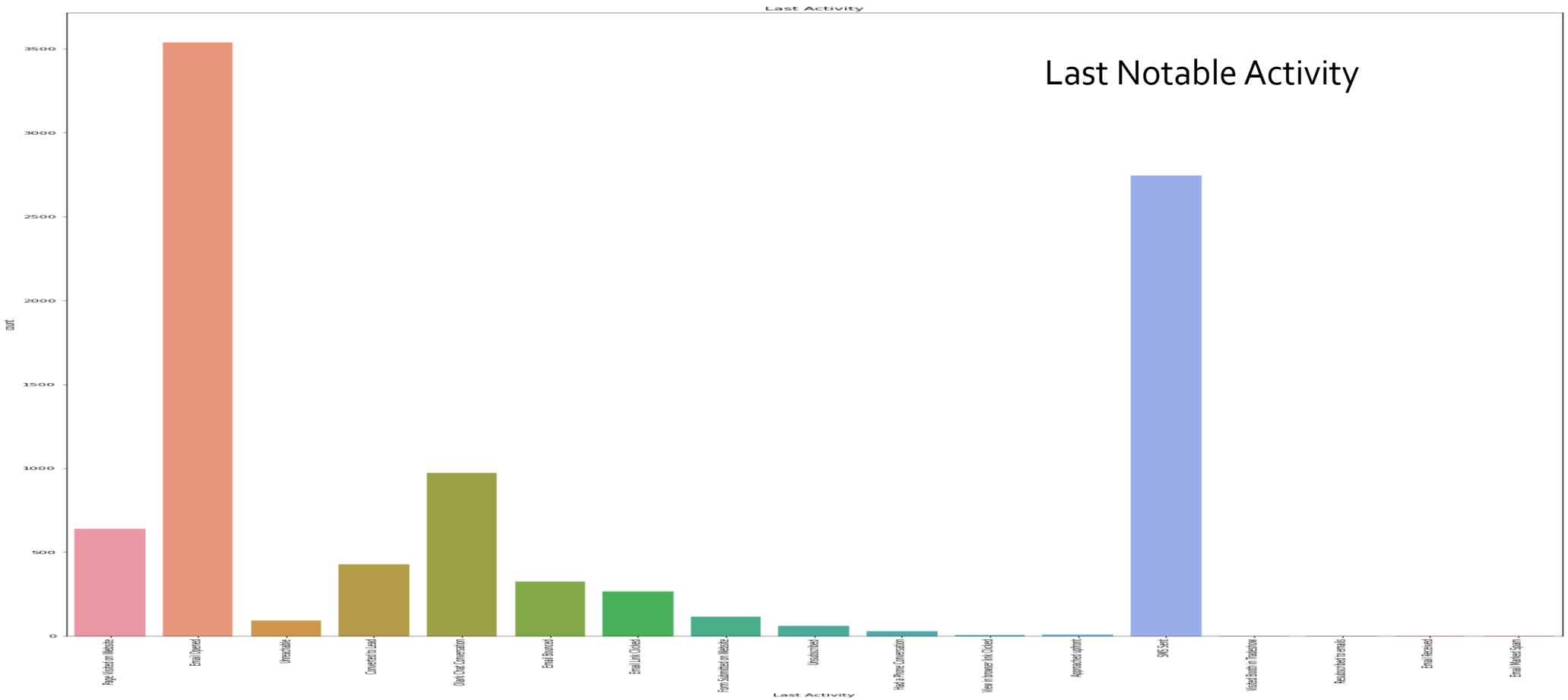
- **Data Cleaning**
  - Handling the missing data, null values, duplicate data. Handling the ‘Select’ level in the categorical column.
  - Dropping the columns that have high percent of missing data more than 40 percent of missing data and dropping them.
  - Check the number of unique categories in each categorical column.
  - For the columns with less percentage of missing, use some imputation technique.
  - Finally check for the outliers in the data.
- **Data Preparation(EDA)**
  - Create dummy variables for all categorical columns.
  - Perform train-test split.
  - Perform scaling. Check the percentage of rows retained in data cleaning process.
- **Modelling**
  - Use techniques like RFE to perform variable selection.
  - Build a Logistic Regression model with good sensitivity.
  - Check p-value and VIF.
  - Find the optimal probability cutoff.
  - Check the model performance over the test data.
  - Generate the score variable.

Data Preprocessing

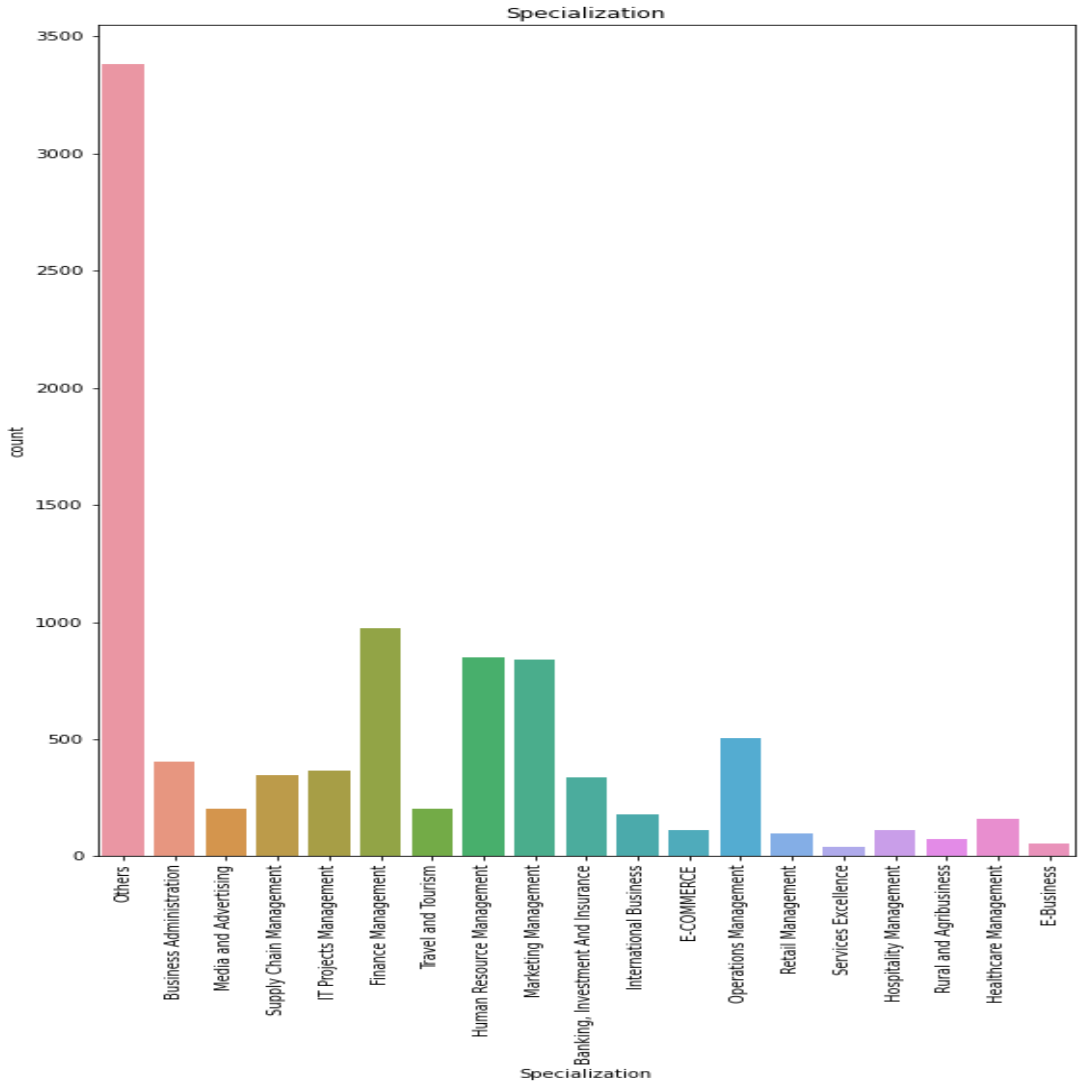
- The data consist of 37 rows and 9240 columns
- The data is had both numeric and categorical data which needs to be identifies and treated respectively
- Removing the columns ‘Prospect ID’ and ‘Lead Number’ since they are identifiers and not needed for the analysis
- We carried out the data imabalance check are decided to drop off some of the variables as there was major variance in the columns. W edropped off the variables like ‘Do Not Call’, ‘What amtters to you in the course’, ‘Search’, ‘Newspaper’, ‘X Education Forum’, ‘Newspaper Article’, ‘Digtal Advertisement’, ‘Magazine’.
- Dropped off the columns that had more than 30% of the data missing like the columns , ’ Asymmetrique Activity Index’, ‘Asymmetrique Profile Score’, ‘Asymmetrique Activity Score’,’ Asymmetrique Profile Index ‘, ‘Lead Quality’.

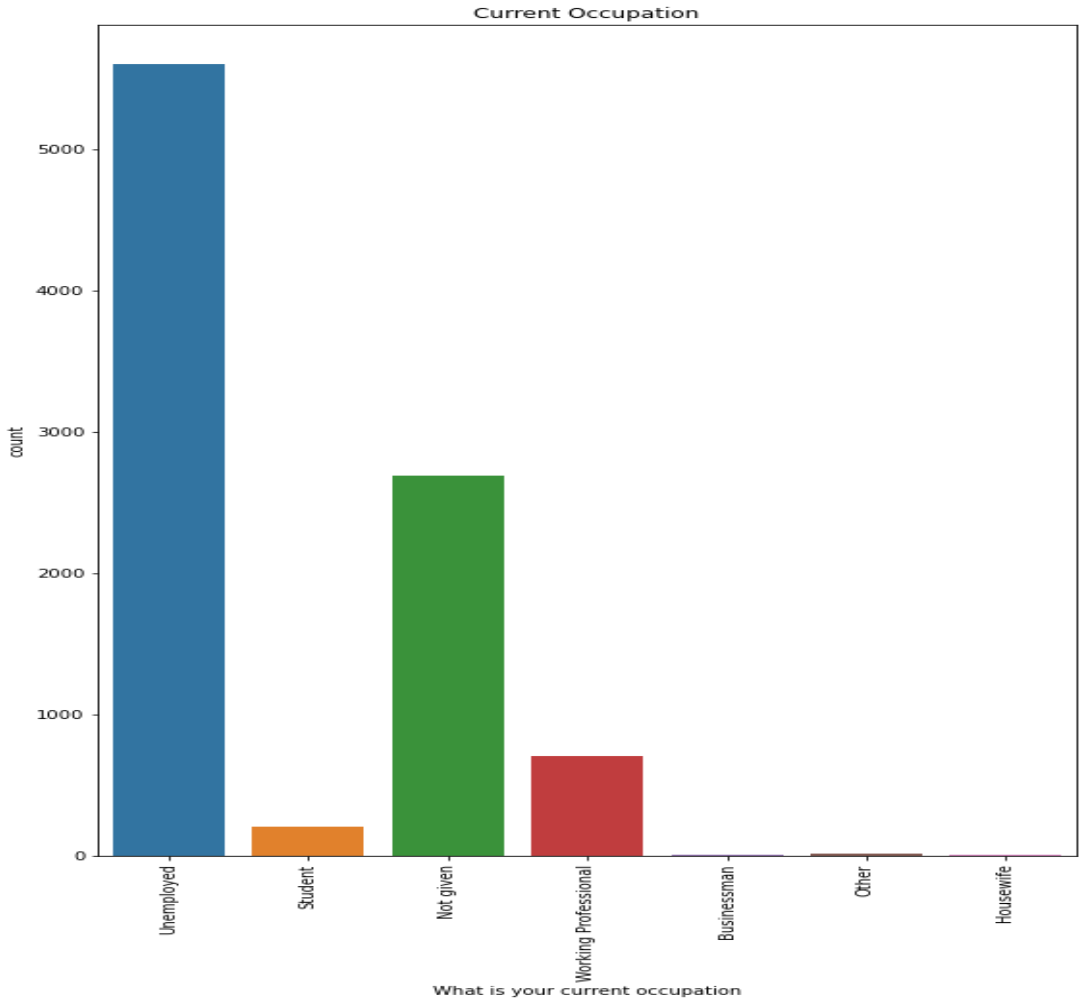
EDA

- Some of the columns that contribute to the model building

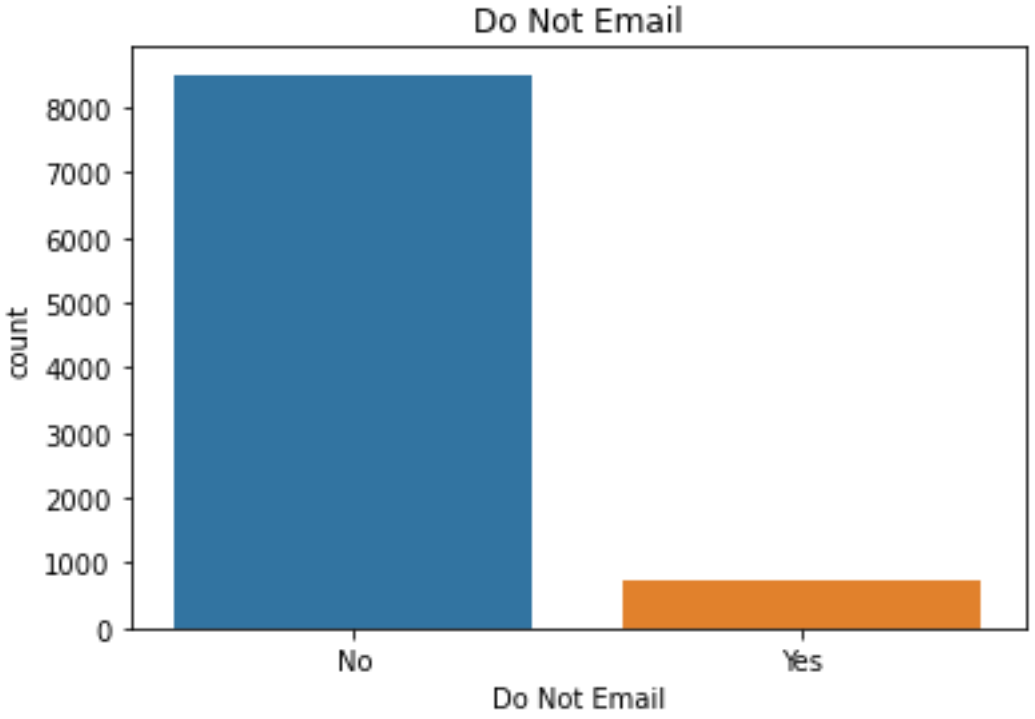


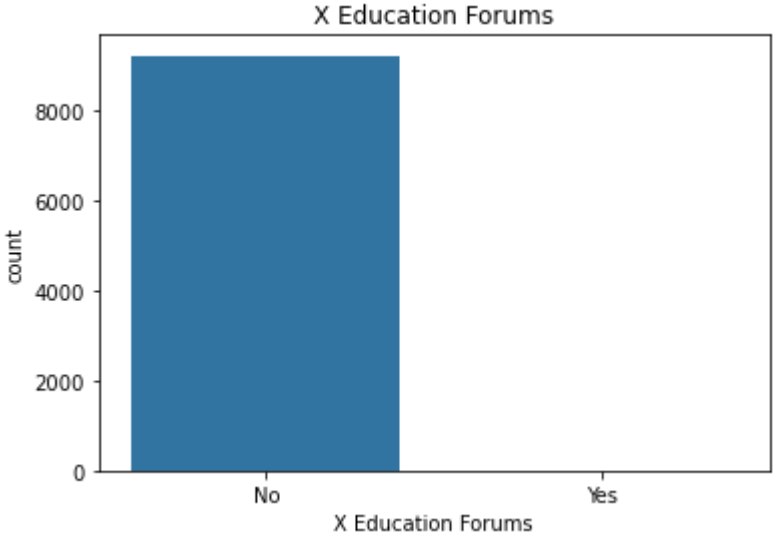
Specialization show the interest of the probable lead in the various fields and current occupation shows whether the leads are from working background or not. This can be contributing factor converting into a lead



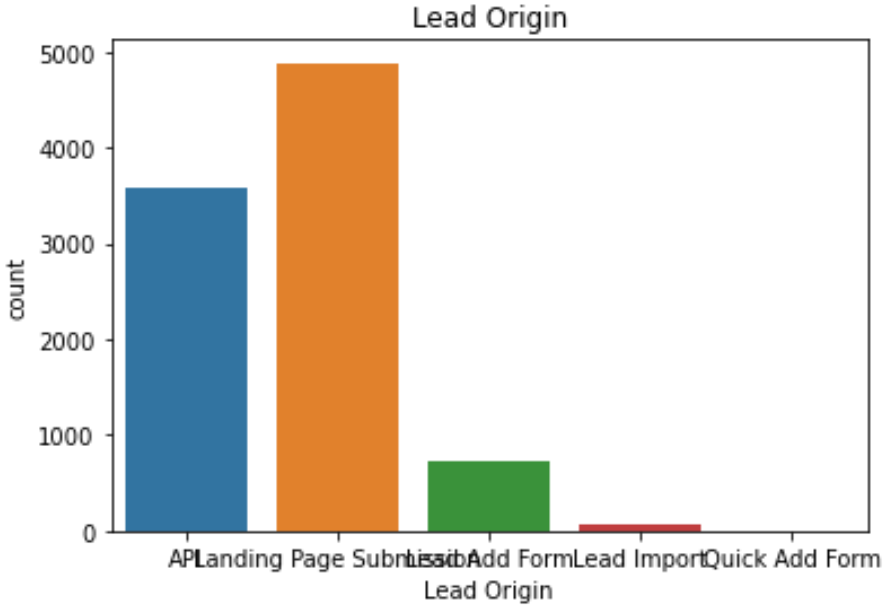


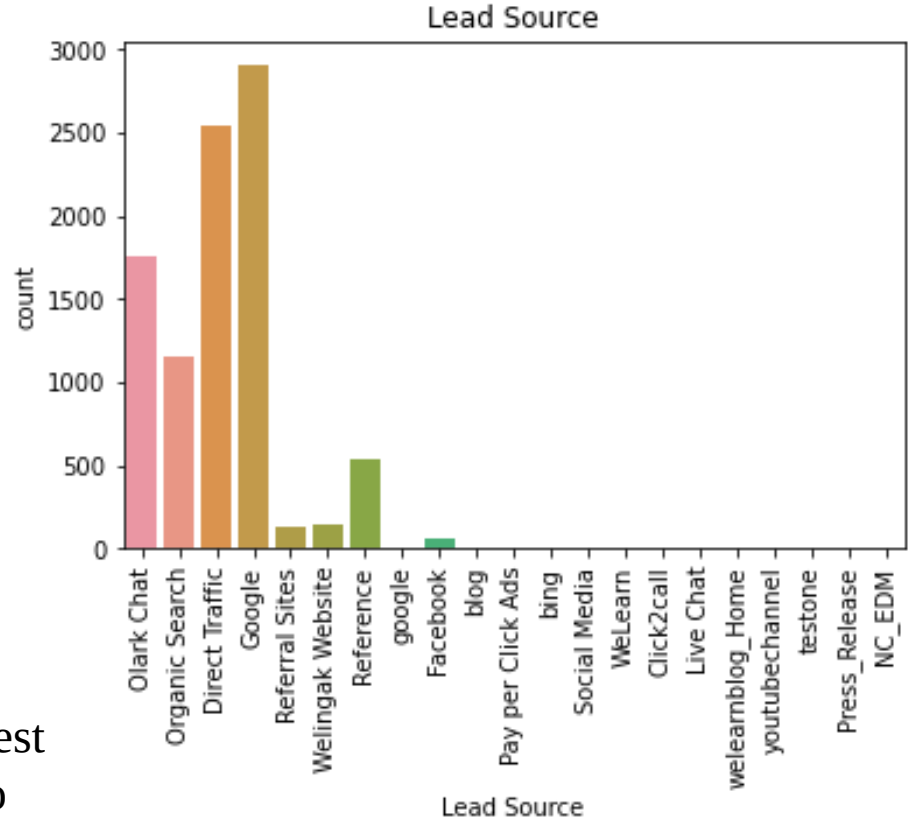
- The Leads coming from the sources are the varied according to the platforms provided to the leads
- The graphs show the data imbalance is some of the variables in the data





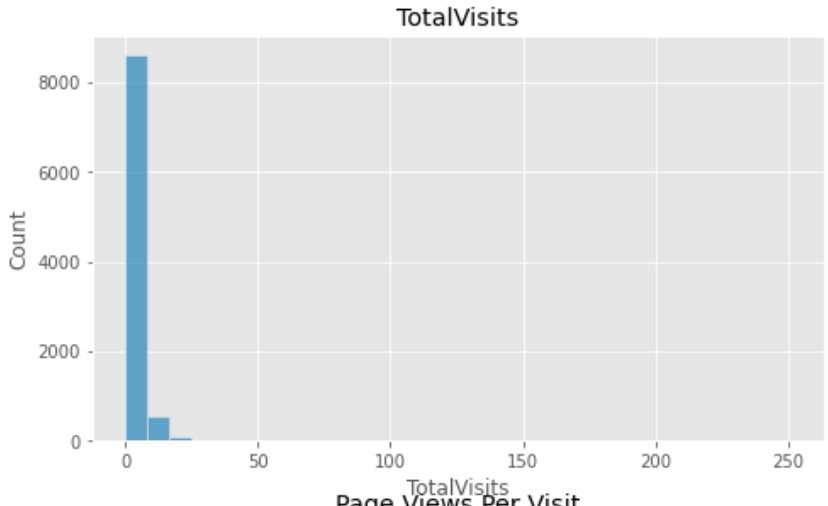
Categorical variables

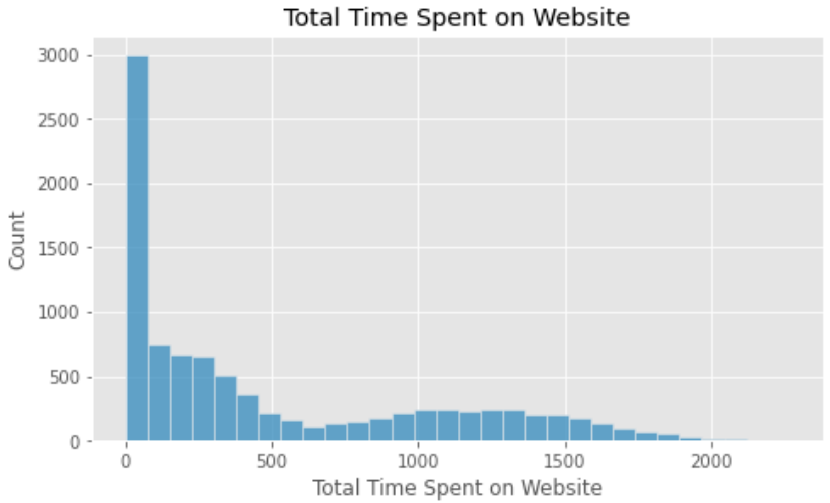


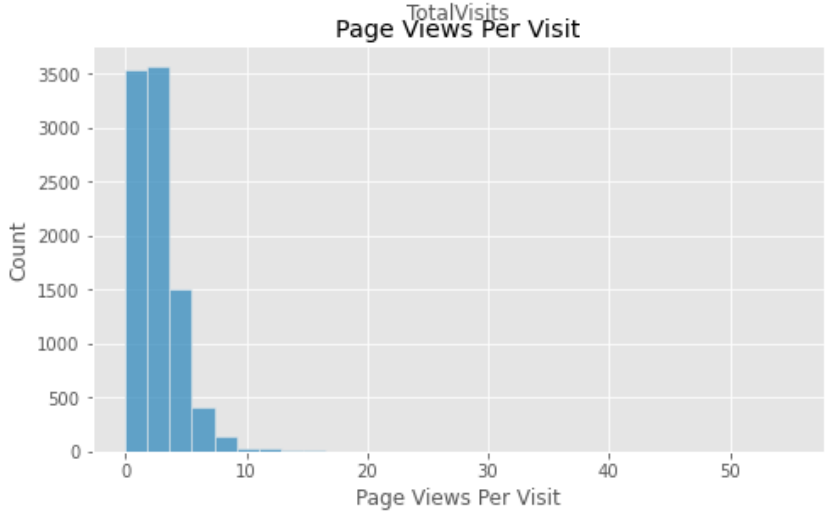


The categorical data shows that some sources are the best option for collecting the elad information as opposed to others which are hardly used by the leads

Numerical data Analysis

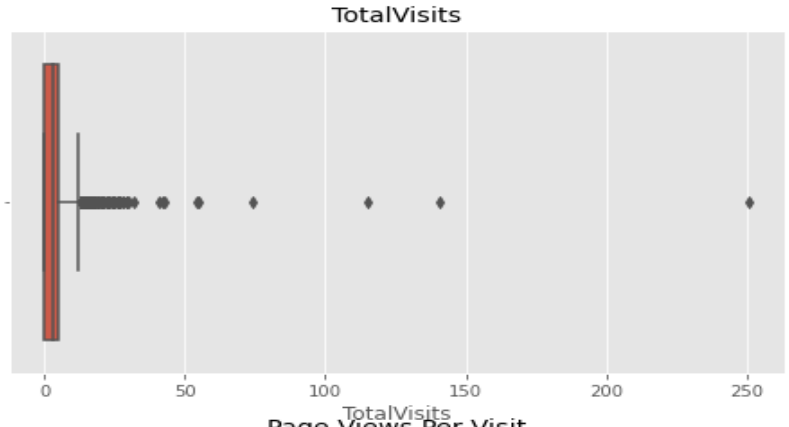


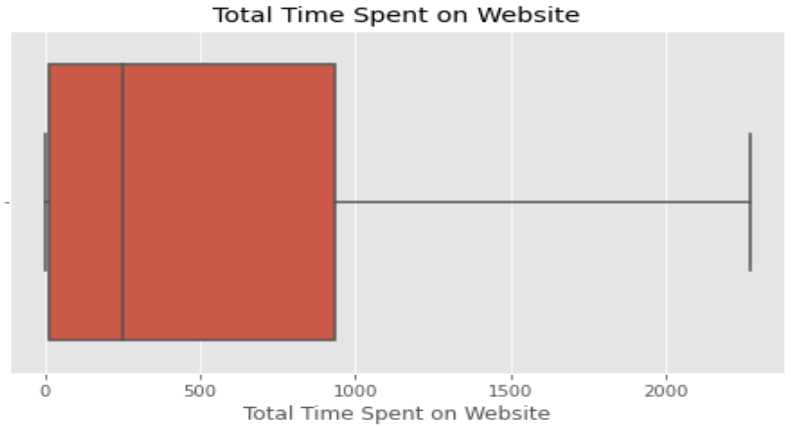


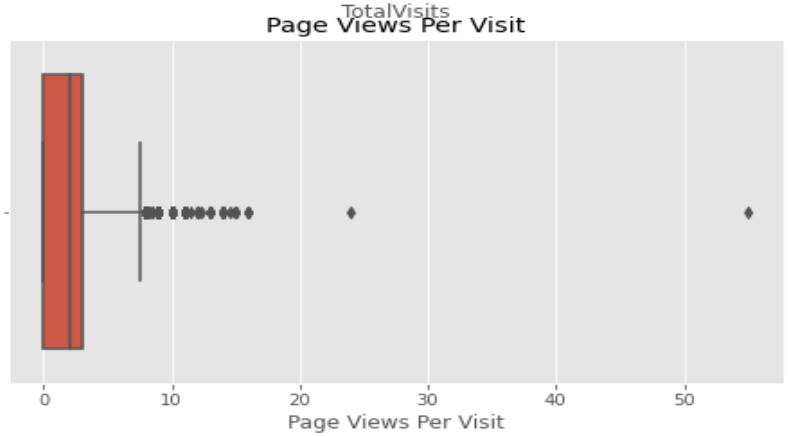


Some of the contributing factors in the finding the converting leads are the TotalVisists, Total time spent on websites and Page Views Per Visit

Outliers Handling







The outliers are handled by imputing the maximum values for the first two variables and using median value for the third varaibles in the graph mentioned here

Data Conversion

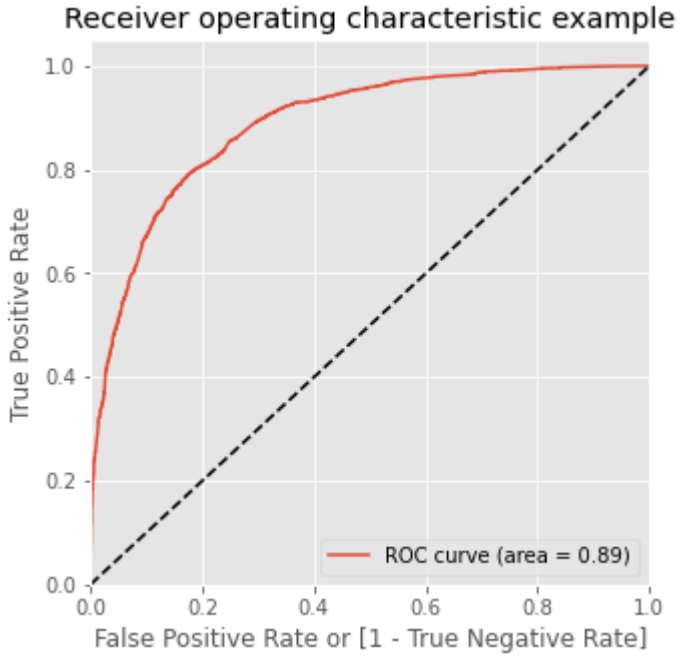
- After cleaning and normalizing the data we have 3 numerical variables
- The categorical variables having yes/no as the values are converted into 1/0 .
- Dummy variables are created for the variables whose falls into object category
- After data manipulation the total number of rows are 80 and the column are 9240
- These columns are further analysed using the p value and VIF factor for the consideration of the model building

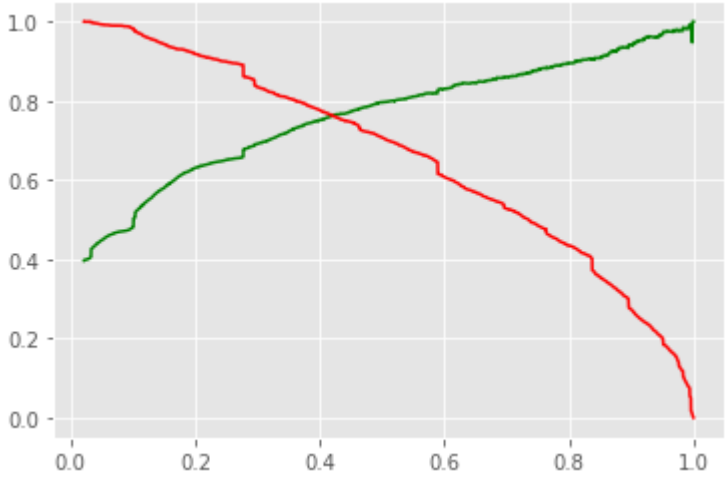
Building the Model

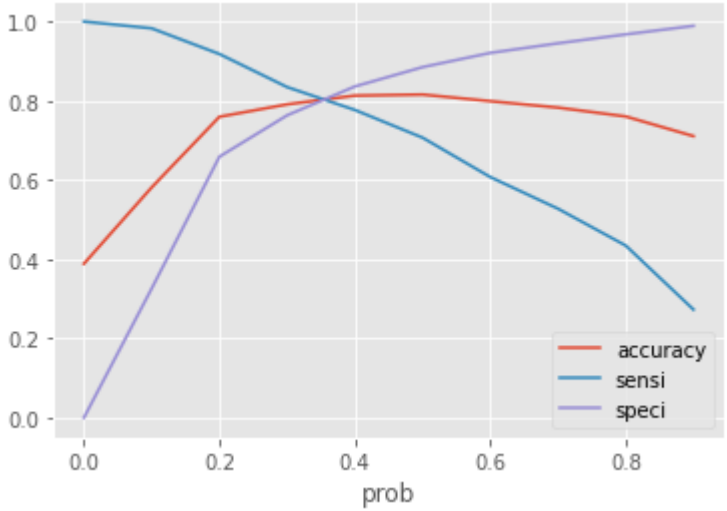
- The data has been split into train and test data using the 70:30 ratio
- We applied the Logistic regression model as the prediction model
- The parameters were selected using the RFE feature selection function. We have considered the rfe value to be 15
- The parameters were chosen using the variance inflation factor and the p-value method. The variables with vif less than 5 and the p-value less than 0.05 were chosen for the model building
- The model evaluation was performed on accuracy , precion and recall metrics
- The overall accuracy of the model turned out to be 81 % with precision of 74 % and the recall of 77%

ROC Optimzation

- The optimal cutoff value was found using the sklearns roc_curve function
- The optimal cut-off is chosen when the precision and the recall values are balance
- As we can see from the graph , the optimal cut-off is the point where the accuracy, sensitivity and the specificity meets.
- In our case this point came out to be around 0.35
- We then performed the similar analysis for precision and recall and selected the optimal cut-off to be 0.41







Conclusion

Following are the observations from the analysis we have done for the Lead_Scoring Case Study:

The variables that are more important with subject to target being converted are :

1. The total time spend on the Website.
2. Total number of visits made to the institute.
3. When the lead source was:
    1. Google
    2. Organic Search
    3. Direct traffic
    4. Welingak website
4. When the 'Lead origin' is 'Lead add format'.
5. When their 'current occupation' is as a 'working professional’.
6. When the last activity was
    a. SMS
    b. Olark Chat conversion