LEAD SCORE CASE STUDY

By:-

RITIKA KUMARI

MAYUR PADORE

AVINASH SALVE

PROBLEM STATEMENT

- X Education sells online courses to industry professionals. The company marks its courses on several popular websites like google.

- X Education wants to select most promising leads that can be converted to paying customers.

- X Education gets a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion.

- To make this process more efficient, the company wishes to identify the most potential leads, known as ‘Hot Leads’.

- The company has had 30% conversion rate through the whole process of turning leads into cutomers by approaching those leads which are to be found having interest in taking the course.The implementation process of lead generating attributes are not efficient in helping conversion.

Business Objective

- X Education requires a model to be built for selecting most promising leads.
- For that they want to build a model which identifies the hot leads.
- Lead score to be given to each leads such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion.
- Deployment of the model for the future use.
- The Model to be built in lead conversion rate around 80% or more.

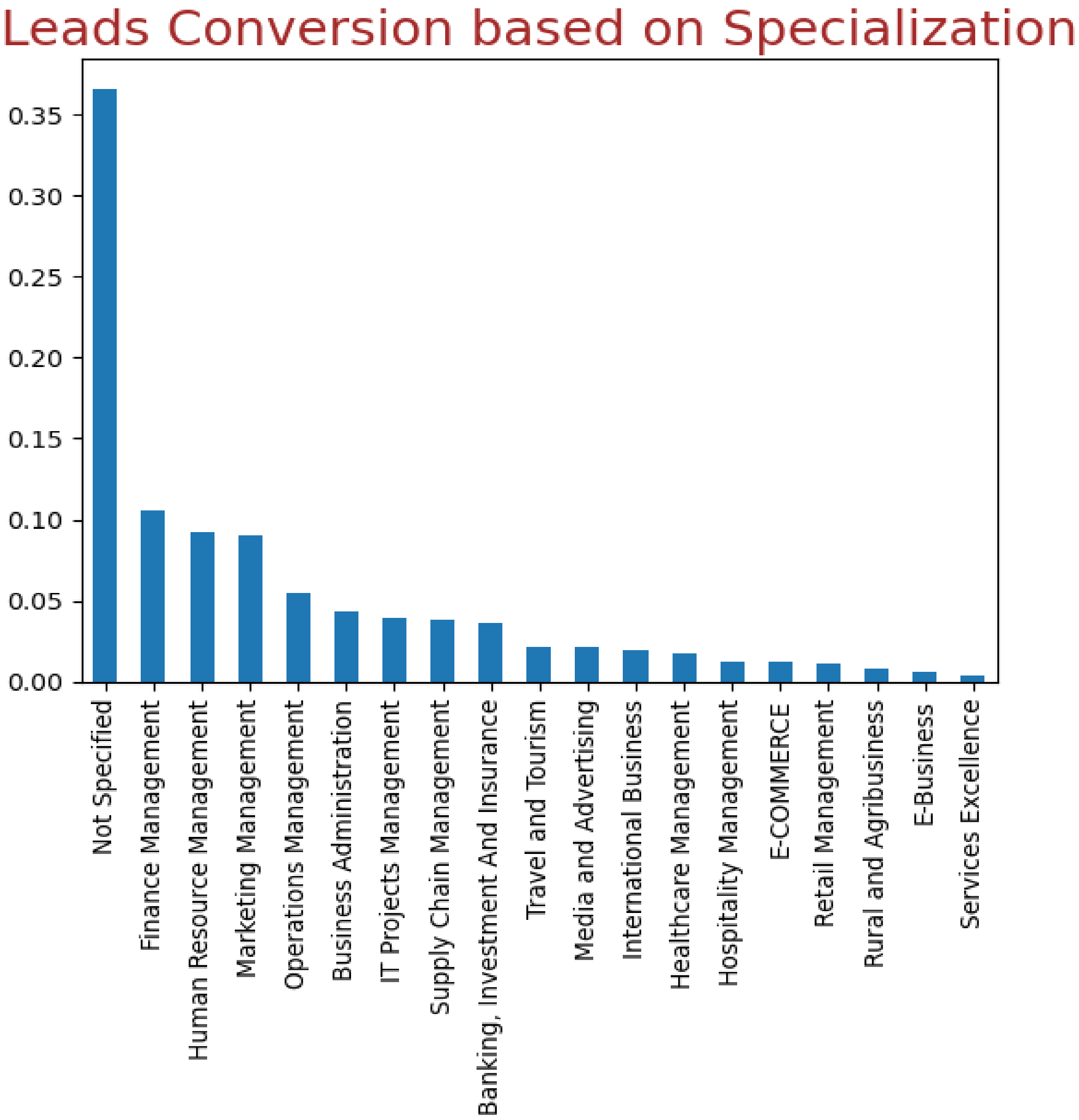
SOLUTION APPROACH

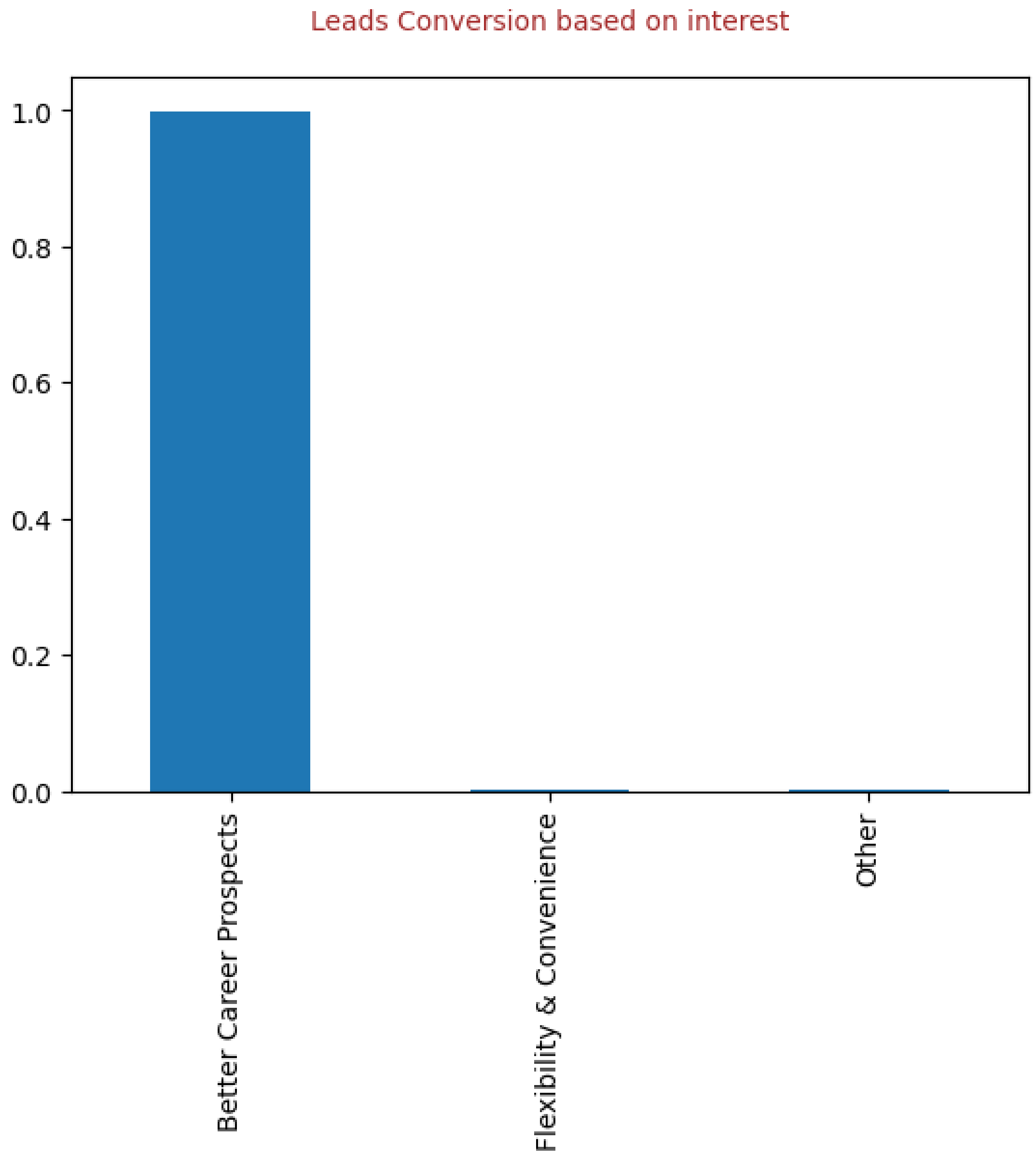
- Source the data for analysis
- Reading and understanding the data
- Data Cleaning & Treatment
- Categorical Variables Analysis
- Handle Binary columns
- Exploratory Data Analysis
- Check the outliers
- Data Preparation
- Model Preparation
- Model Building: Logistic Regression
- Feature Scaling
- Model Evaluation
- Predictions on test set

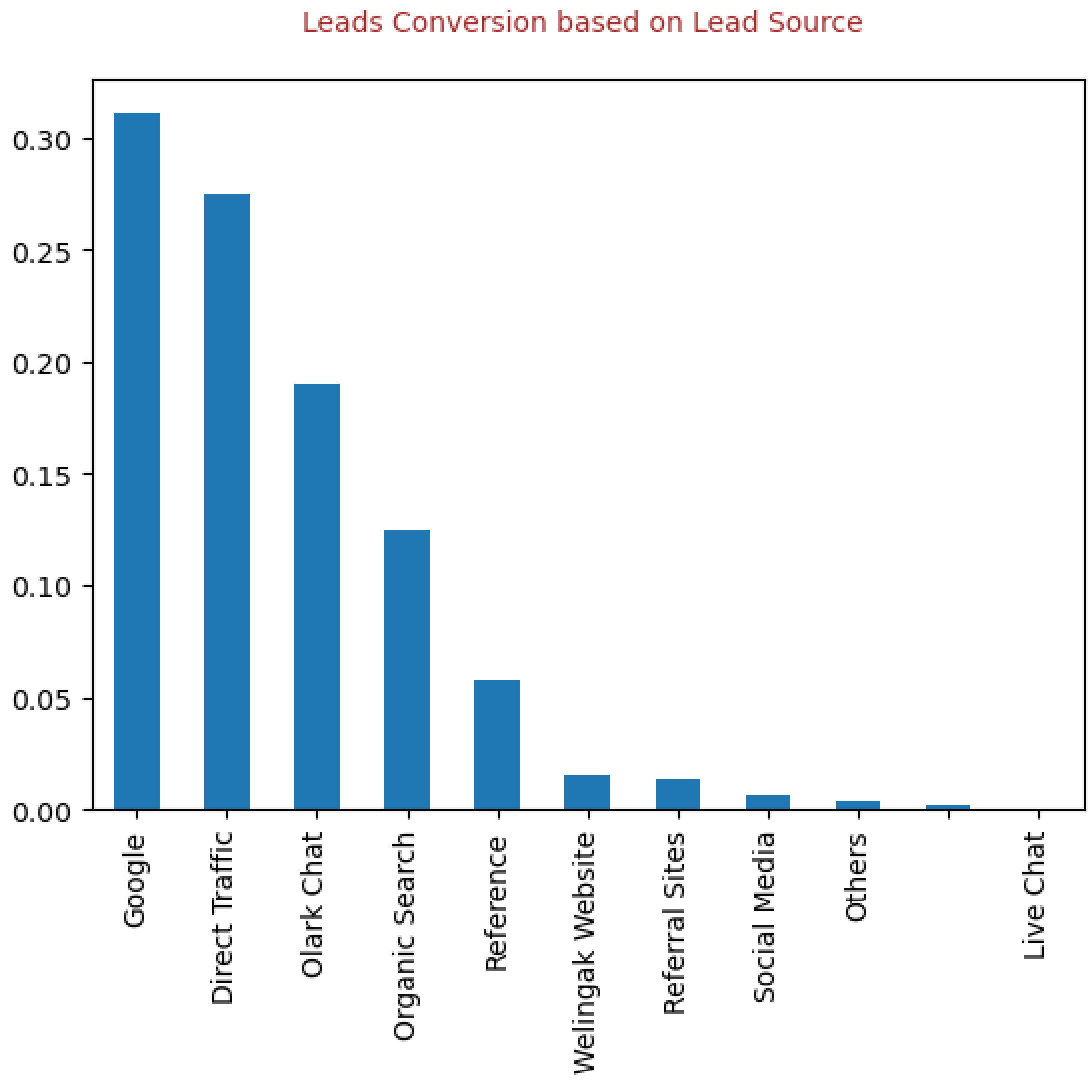
DATA CLEANING , SOURCING AND PREPERATION

- Read the data from CSV file
- Data cleaning handling null values & removing higher null values data
- Removing redundant columns in the data
- Imputing null values
- Outlier treatment
- Exploratory data analysis- approx. conversion rate is 38.4%
- Feature standardization

Categorical Variables Analysis:-

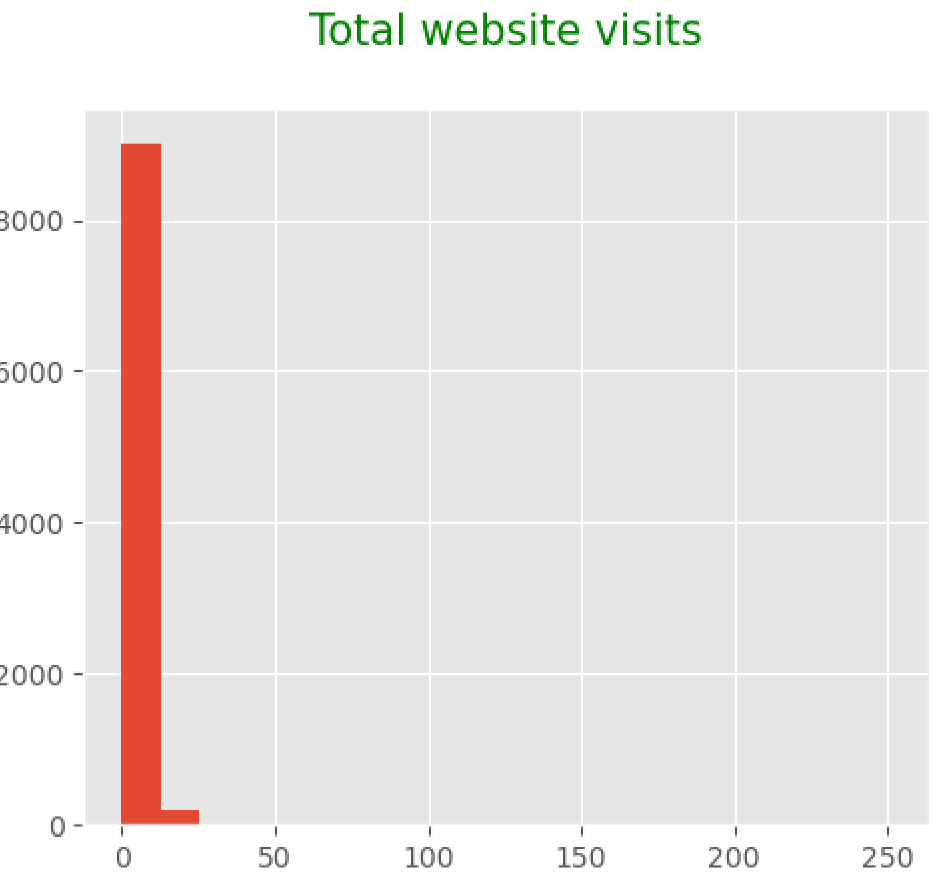


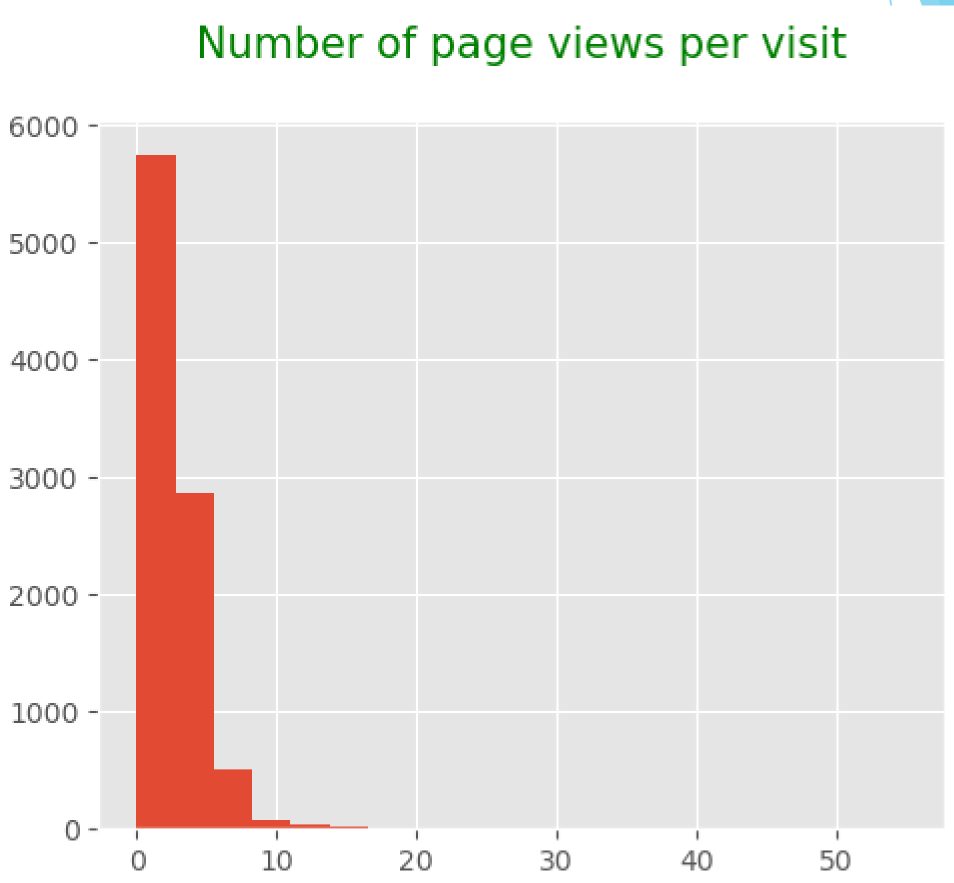

Leads Conversion based on interest
1.0
0.8
0.6
0.4
0.2
0.0
Better Career Prospects
Flexibility & Convenience
Other



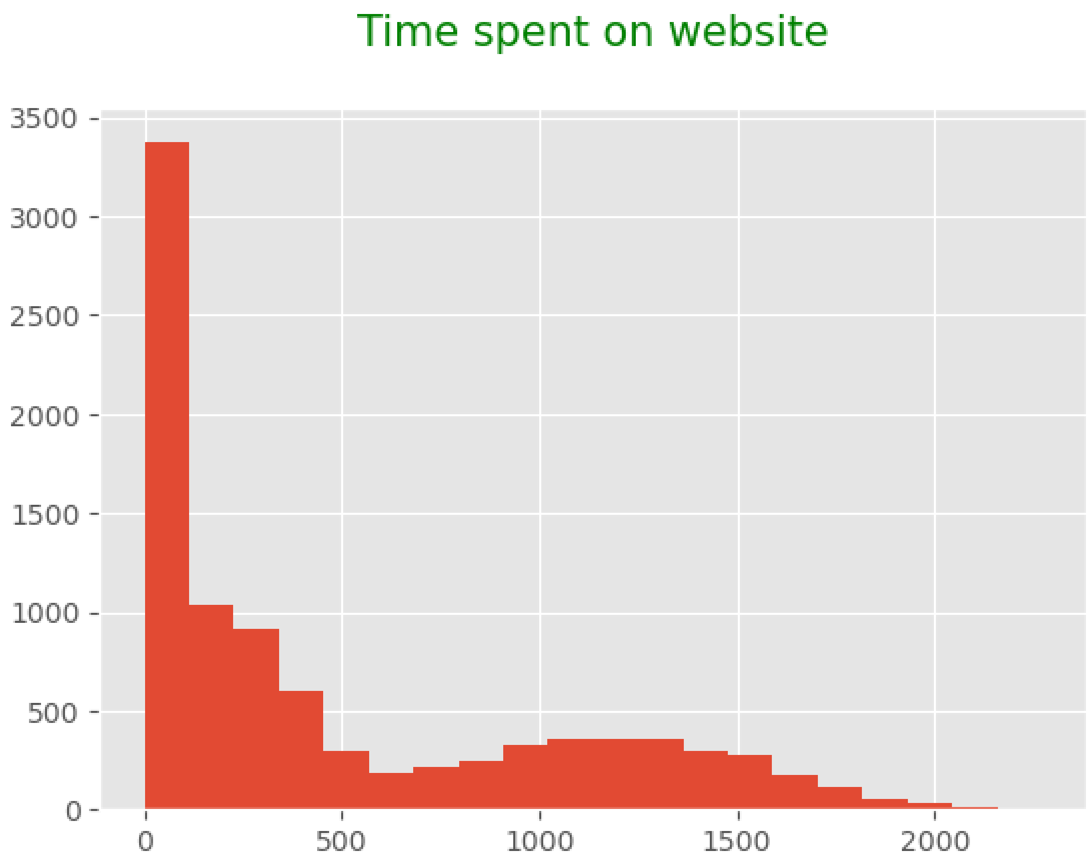
Observations : -Most of the leads generated are through Google and Direct traffic and the least through Live Chat -Welingak website ahs the most conversion rate -Lead conversion can be improved by maxising leads from Reference and welingak website -Focussing Olark chat, Organic search, Direct traffic, and google leads may increase the lead conversion.

Exploratory Data Analysis:-





Observations:-
High peaks and skewed data. There might be a possibility of outliers.

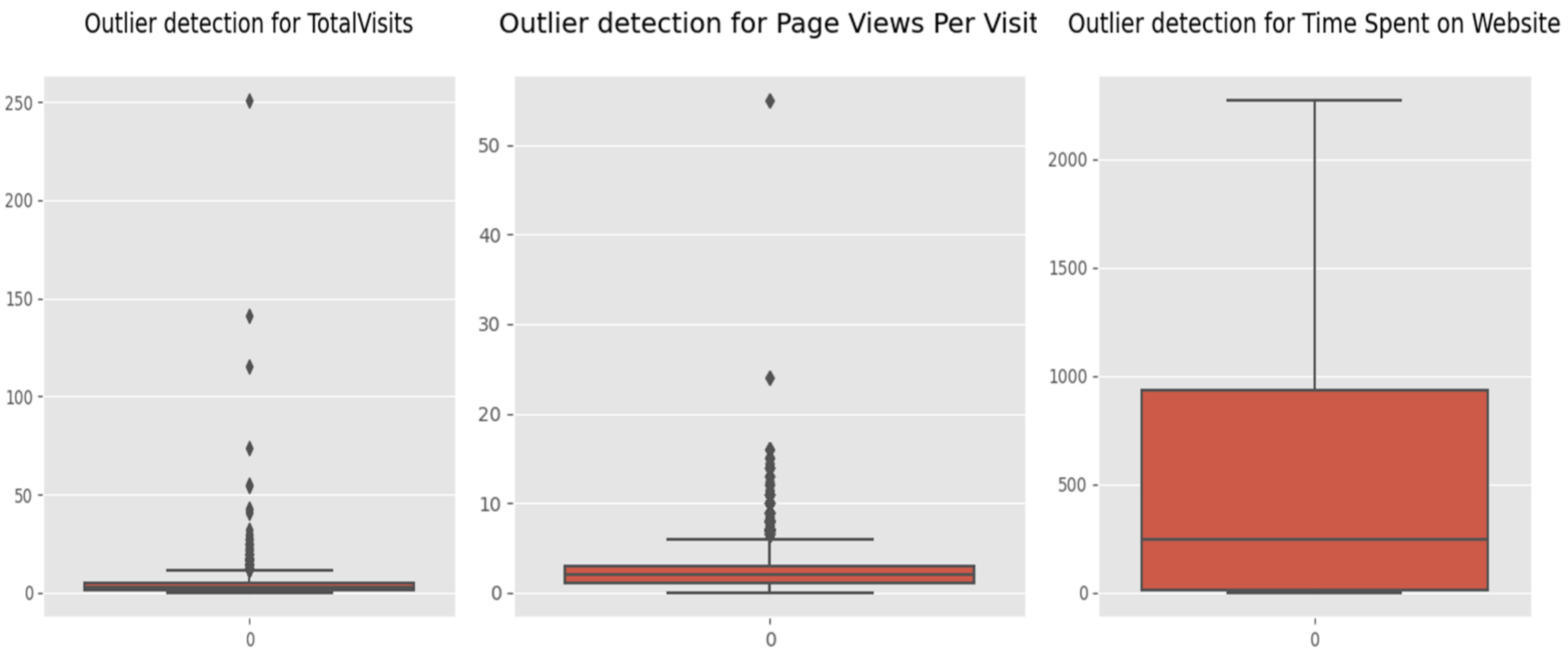


Time spent on website

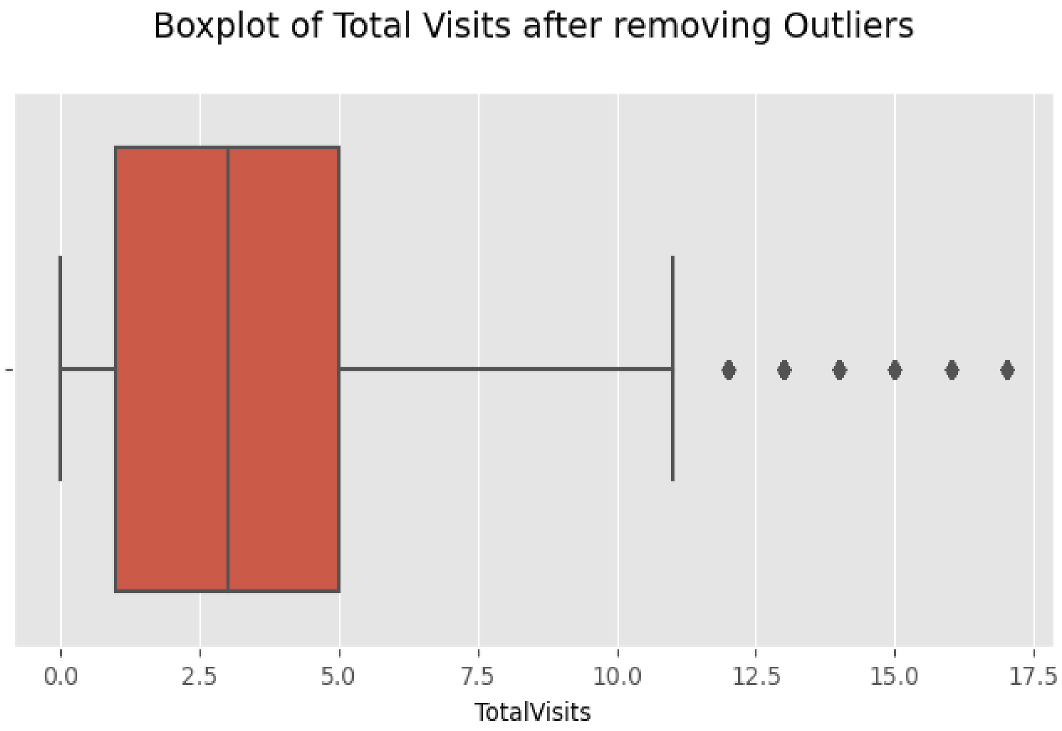
Observations:-

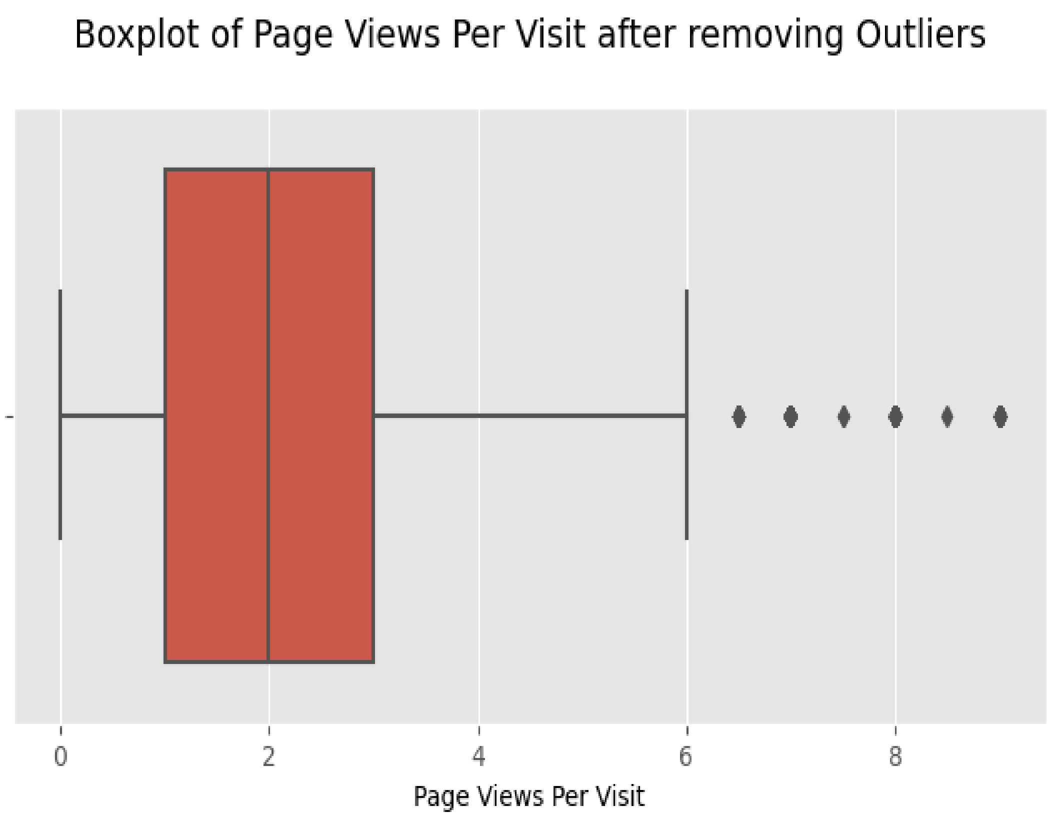
High peaks and skewed data. There might be a possibility of outliers.

OUTLIERS:-



TotalVisits, Page views per visit, time spent on website have Outliers .Looking at both the box plots and the statistics, there are upper bound outliers in both TotalVisits and Page Views Per Visit columns.





TotalVisits and page views per visit after removing Outliers.

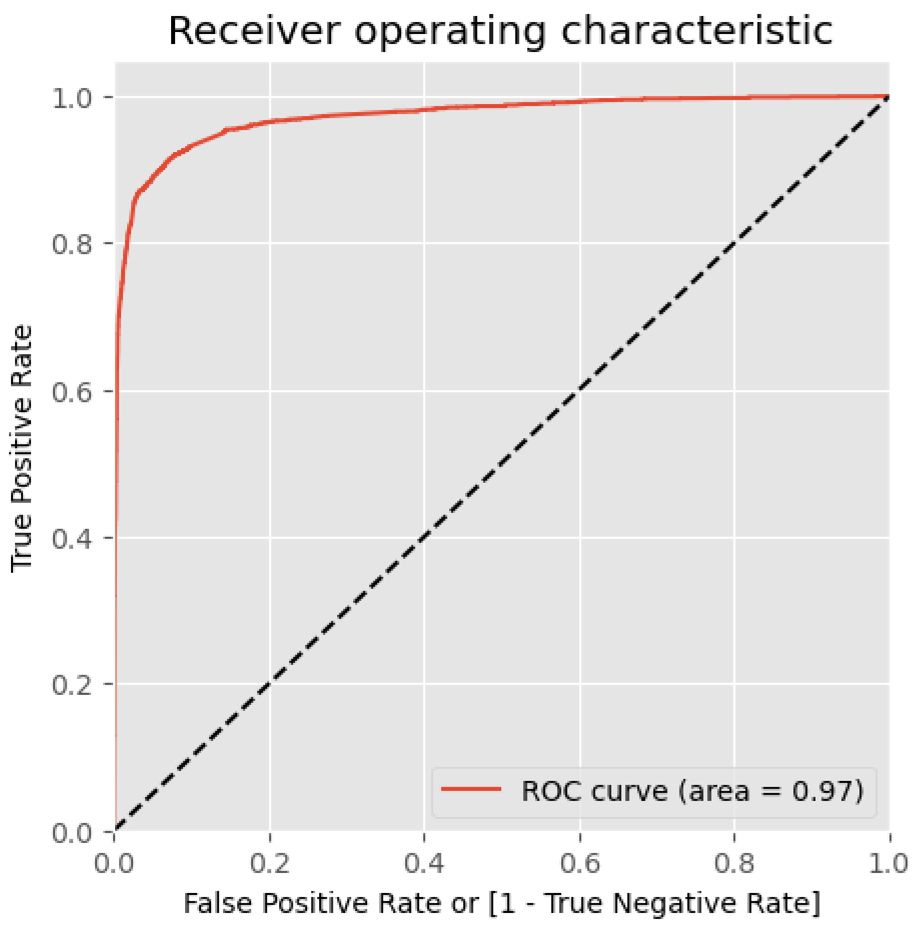
Data Preparation:-

- Converting Binary (Yes/No) to 0/1
- Creating dummy variables for categorical columns

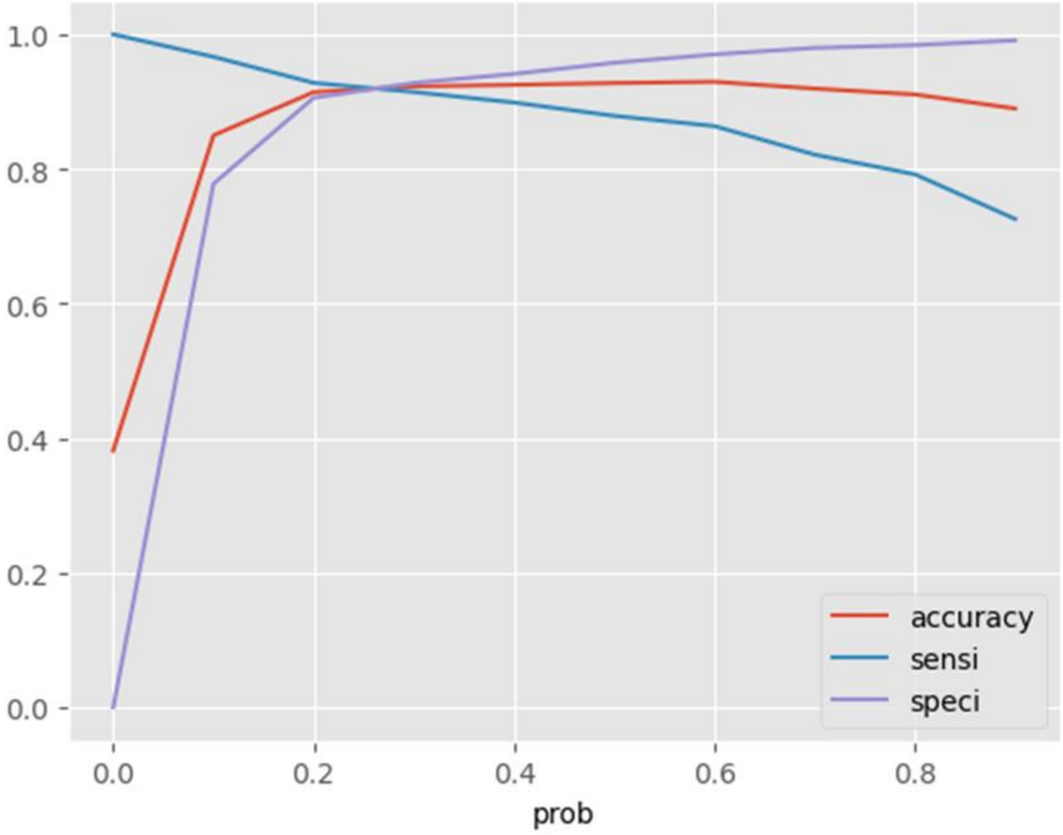
Model Building: Logistic Regression

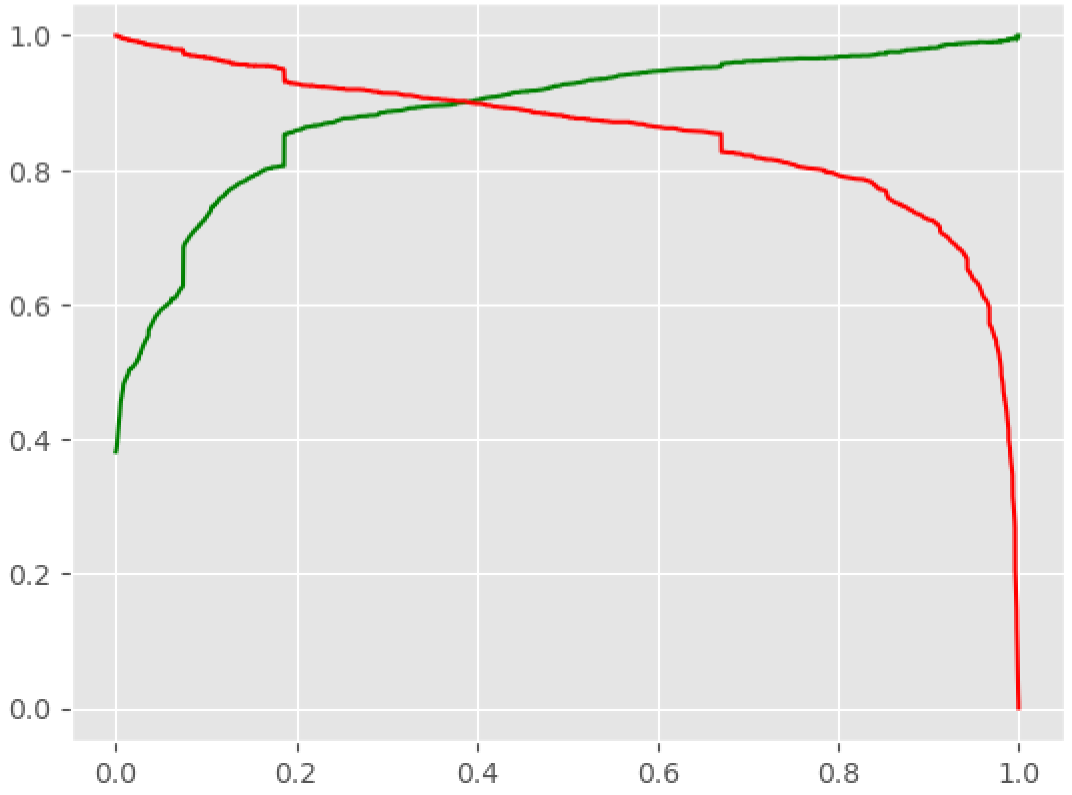
- Splitting the data into Training and Testing Set
- The first basic step for regression is performing a train-test split we have choosen 70:30 ratio,
- Use RFE for Feature Selection
- Running RFE with 15 variables as output.
- Building model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.
- Calculated accuracy, sensiitivity, specificity, precision ,recall & evaluate model

ROC CURVE


Receiver operating characteristic
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.97)

MODEL EVALUATION TRAIN





Accuracy : 92.26%
Sensitivity : 91.43%
Specificity : 92.78%
Precision:88.6%
Recall:91.4%

- Finding Optimal cutoff point
- Optimal cutoff probability is that probability where we get balanced sensitivity and specificity
- From the second graph it is visible that the Optimal cutoff is approximately at 0.35

Model Evaluation on the Test Dataset

Sensitivity and specificity on the Test Dataset:-

Accuracy : 92%
Sensitivity : 91%
Specificity : 92%

CONCLUSION

- While we have checked sensitivity, specificity, precision, recall metrics, we have considered the optimal cutoff based On sensitivity & specificity for calculating the final prediction.

- It was found out that the variables that mattered the most in the potential buyers are:-
- Total number of visits
- The total time spent on website
- When the last activity was-SMS ,Olark Chat Conversation
- When the lead origin is lead add form
- When the current occupation was- working professional, unemployed, student, other

- The threshold has been selected from accuracy, sensitivity, specificity measures and precision, recall curves.

- The model shows Sensitivity : 91.43%, Specificity : 92.78%

- The model finds correct promising leads and leads that have less chances of getting converted

- Hence overall model seems to be good