LEAD SCORE CASE STUDY

Group Members:

1. Yashus G
2. Soumya Shinde
3. Huynh Trong Huong Nhi

Problem Statement

- X Education sells online course to industry professionals. And the Company introduces its courses on several websites and search engine, such as Google, Olark Chat,… People who land on the websites and fill up information will be classified to be a lead. But these leads are very poor because there are about 30% of them are converted.
- To this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. The company focuses on communicating the promissing leads, instead of getting in touch with all leads.

Business objectives

- finding the most potential leads
- Building the model to identify the hot leads
- Apply the model in the company’s operation in the future

Solution methodology

- Data cleaning
  - Read the data
  - Check and handle missing values and unsuitable values
  - Drops columns with the highest percentage of missing values
  - Imputing missing values, if necessary
  - Check and handle outliers
- EDA
  - Univariate data analysis
  - Bivariate data analysis
- Data Preparation
  - Create dummies for all categories
  - Perform train-test-split
  - Perform scaling by standard method

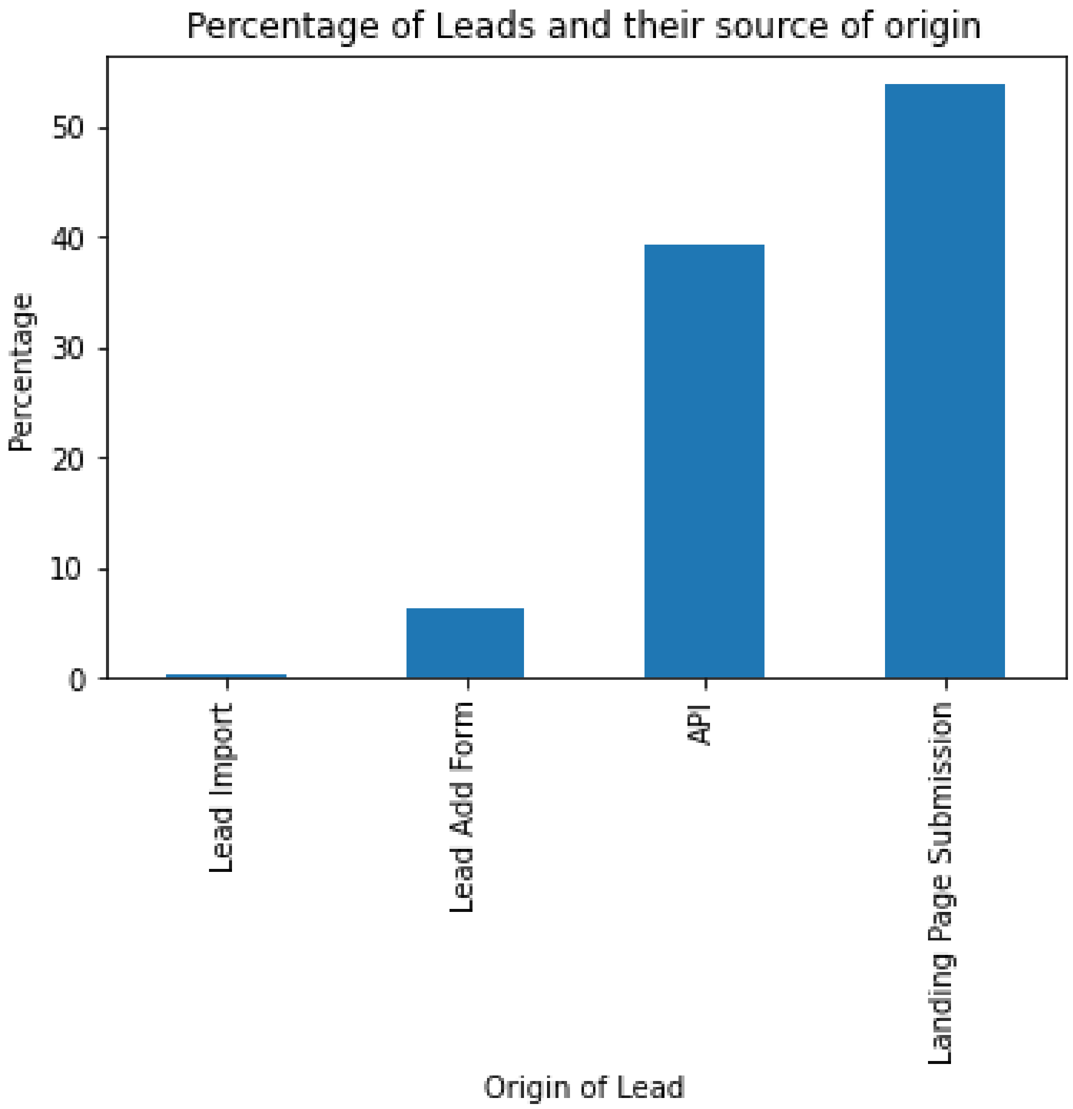
Solution methodology

- Modelling
  - Use RFE techniques to select variables
  - Build a Logistic Regression with Recall
  - Based on P-value and VIF to build the final model
  - Predict the independent variable
  - Find the optional probability cutoff
  - Based on test data, check the model performance (confusion matrix,score)
- Validation of the model
- Conclusion and recommendations

Understanding the model

- Leads data has 9240 columns and 37 rows
- Drop columns with missing values greater than 30%, except for columns with important information.
- Imputing columns with missing values less than 30%

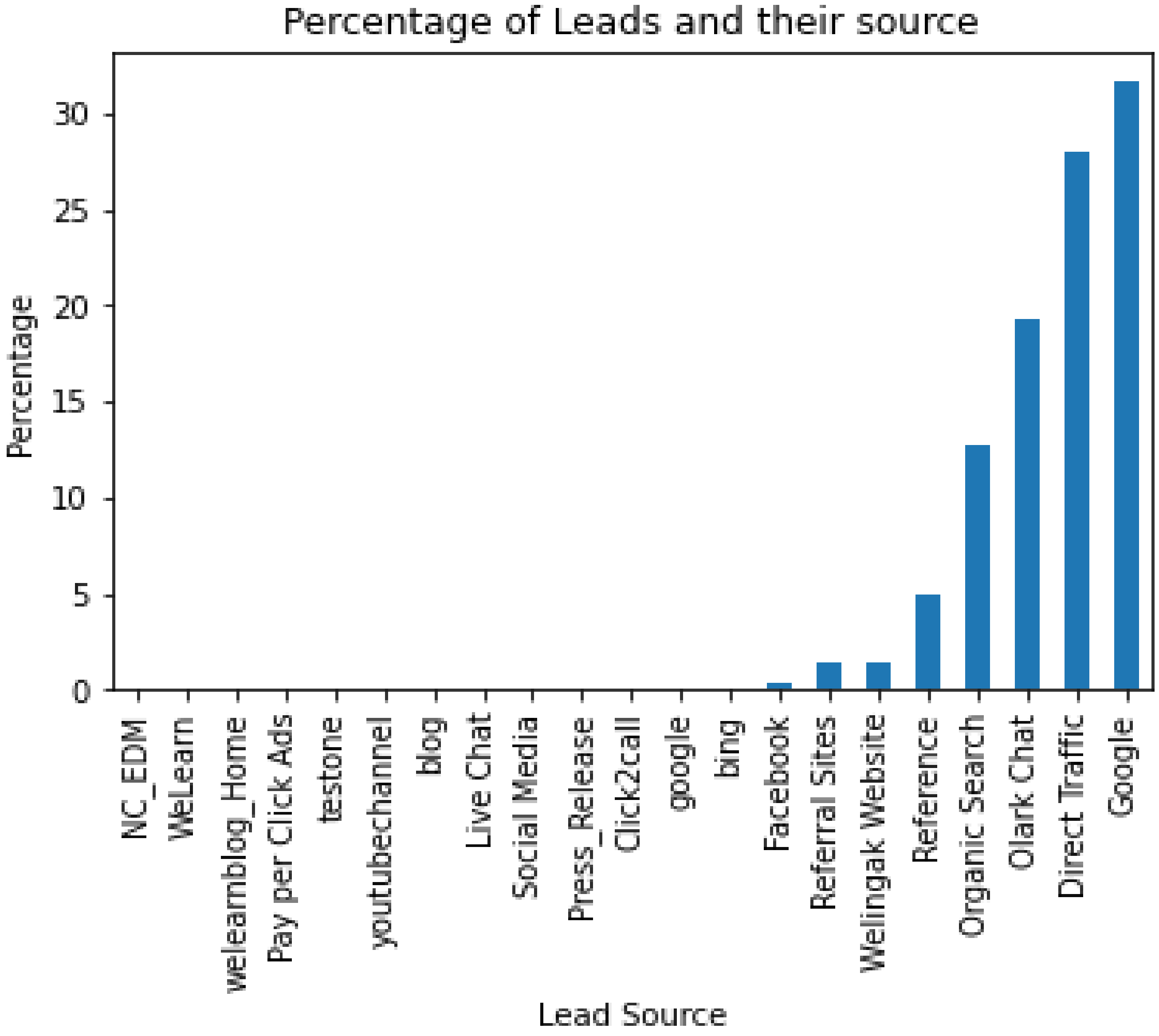
EDA



Conclusion:

From the above plot, it can be seen that a majority of the leads originate from Landing Page Submission(>50%), this is followed by API(~40%)

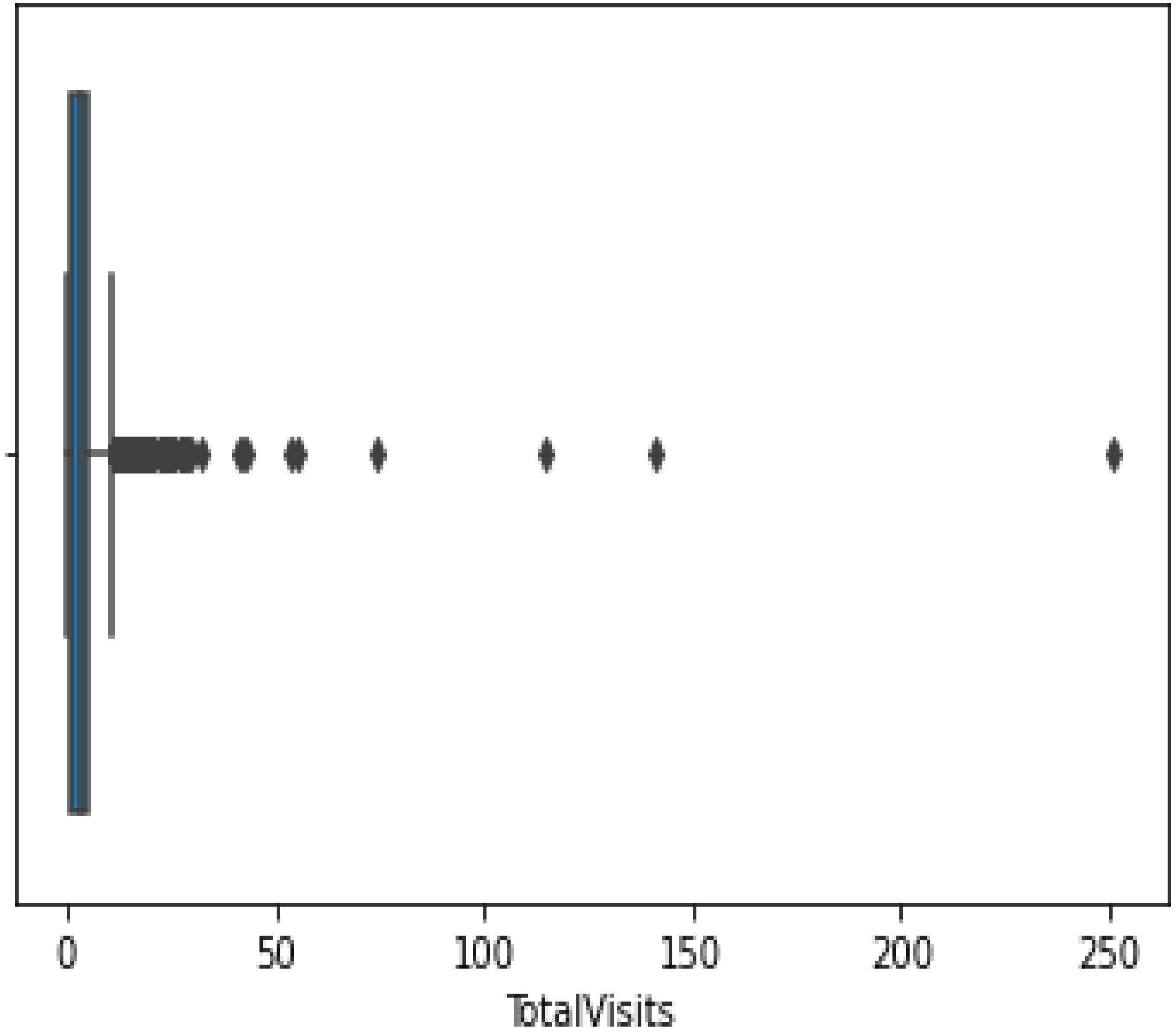
EDA



Conclusion:
From the above plot it can be seen that a majority of the leads are from Google and Direct Traffic. Thus these can be considered as important sources for marketing strategy decisions

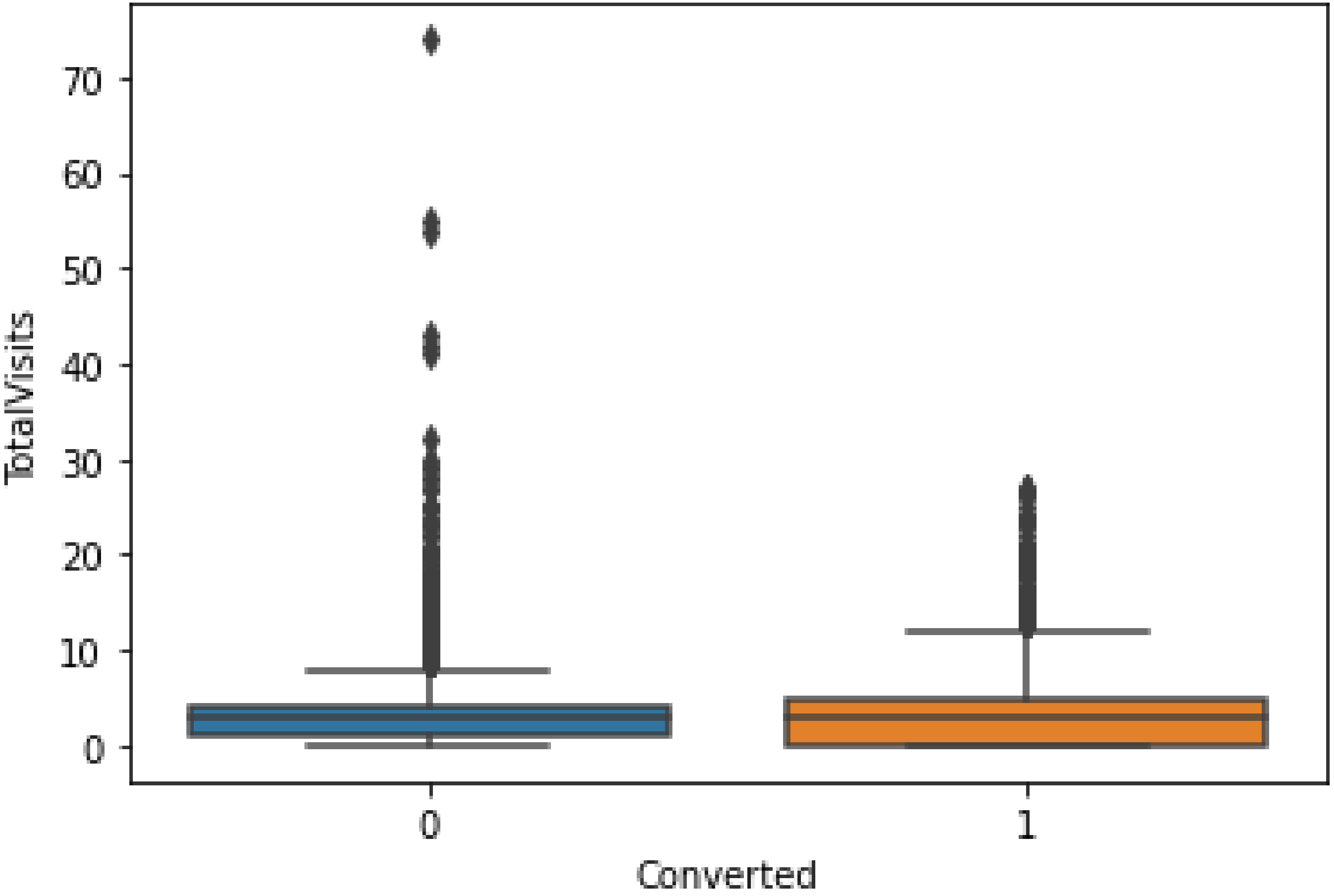
EDA

Conclusion:
From the above boxplot it can be seen that Values greater than 100 can be considerd as outliers and can be removed as these high number of visits can be considered as outliers(only 3 values out of more than 9000 data points)



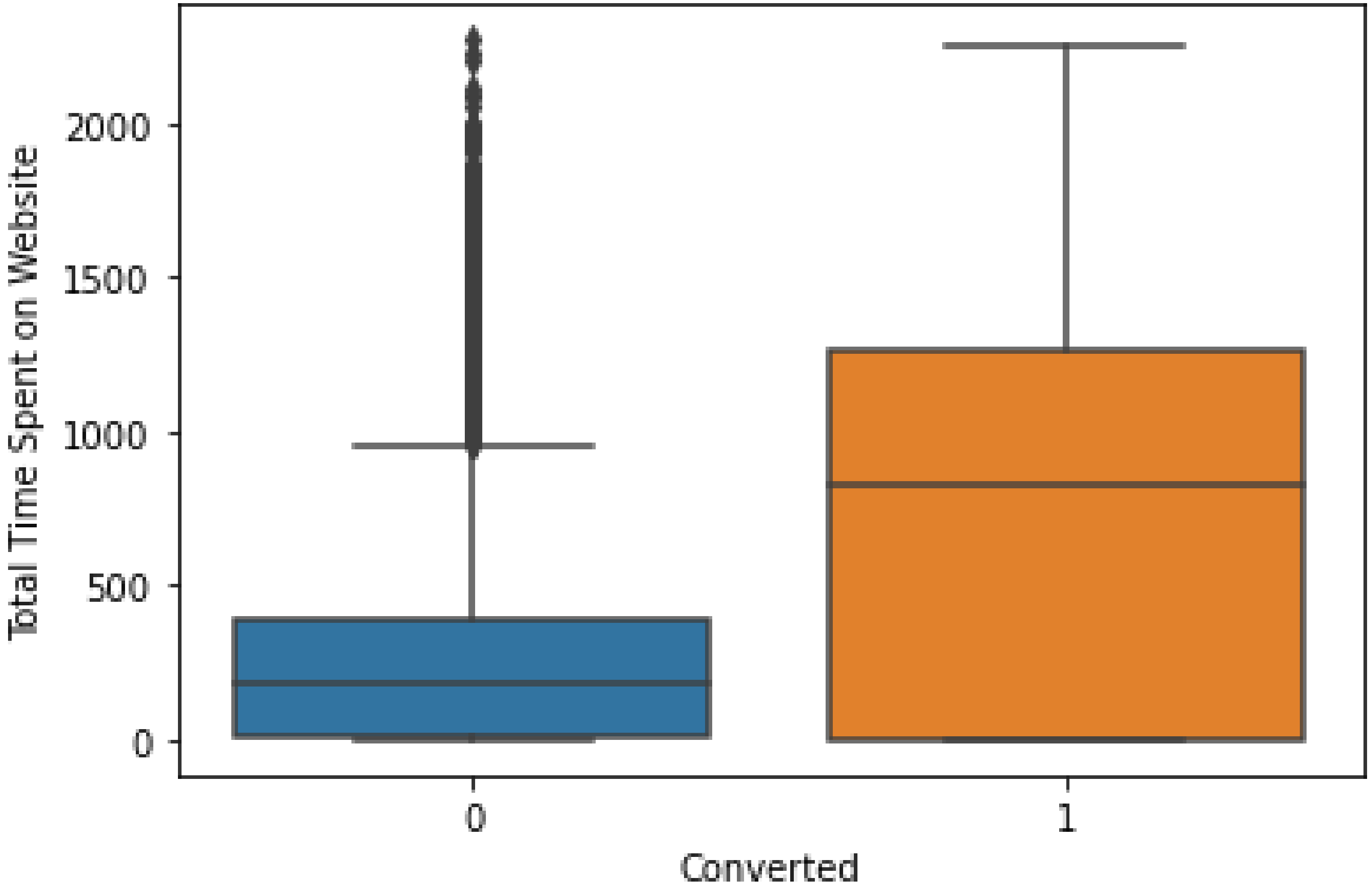
EDA

From the Above plot it can be seen that on an average, more number of visits implies a higher tendency for conversion.

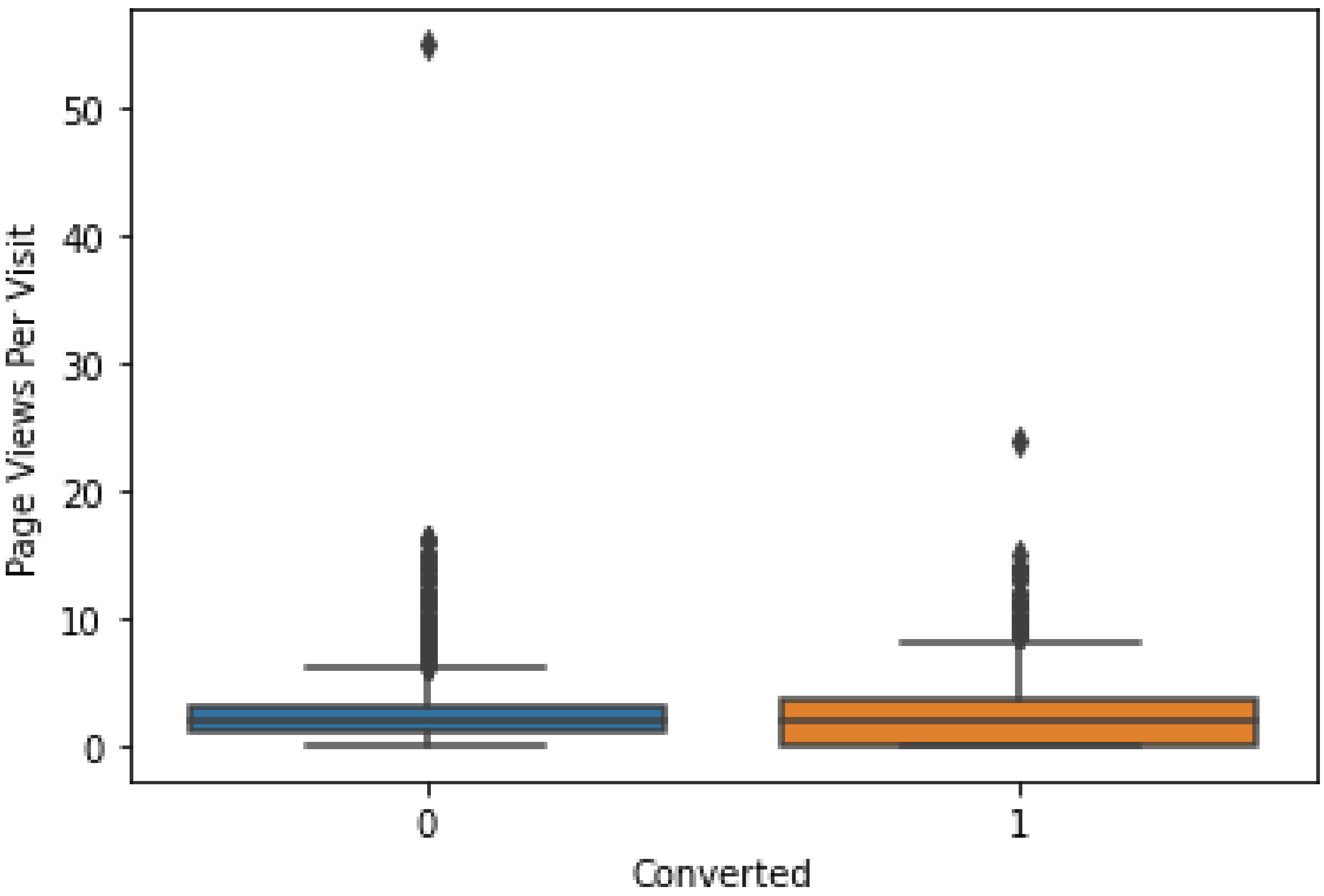


EDA

From the above box plot it can be clearly seen that a higher time spent on the website leads to a higher chance of successful conversion. Which is also inline with intuitive understanding.

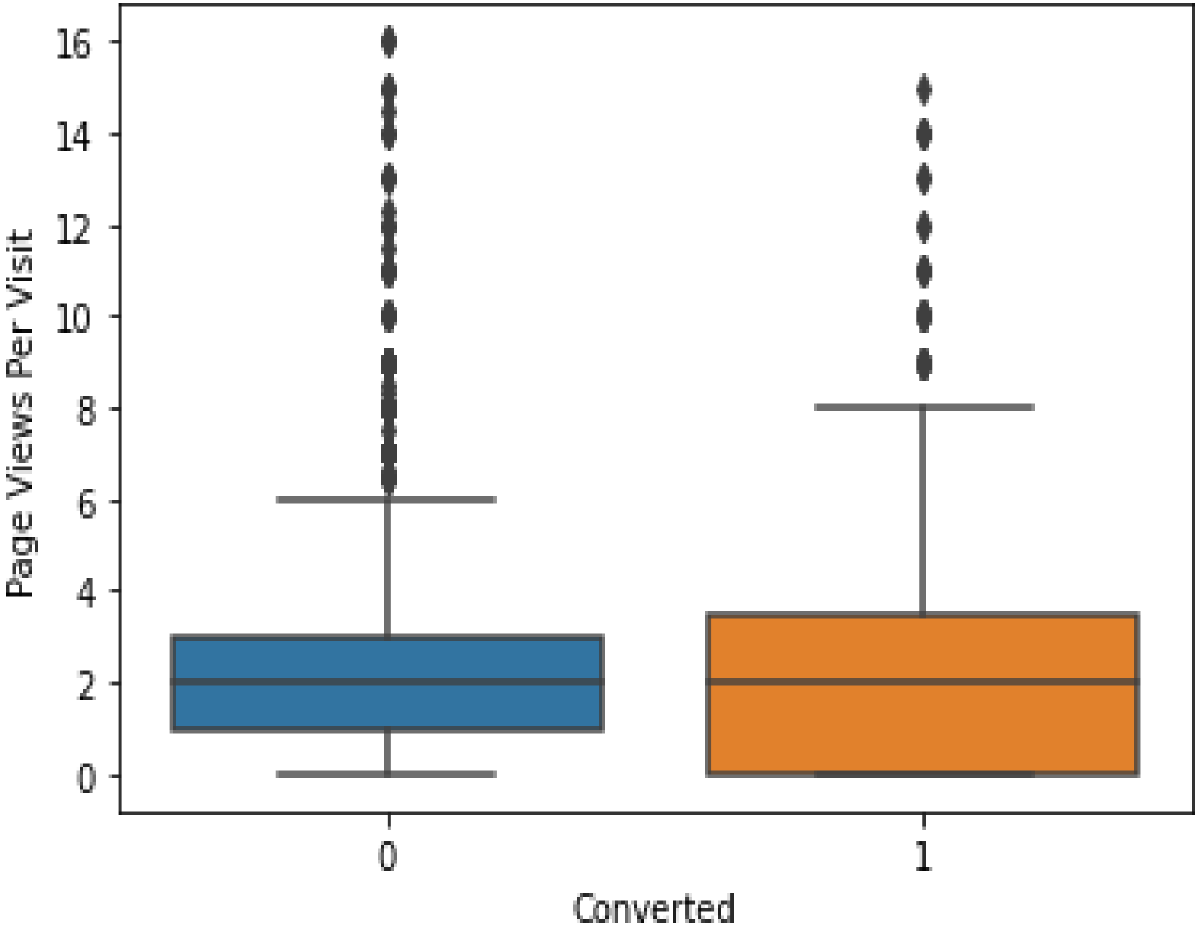


EDA



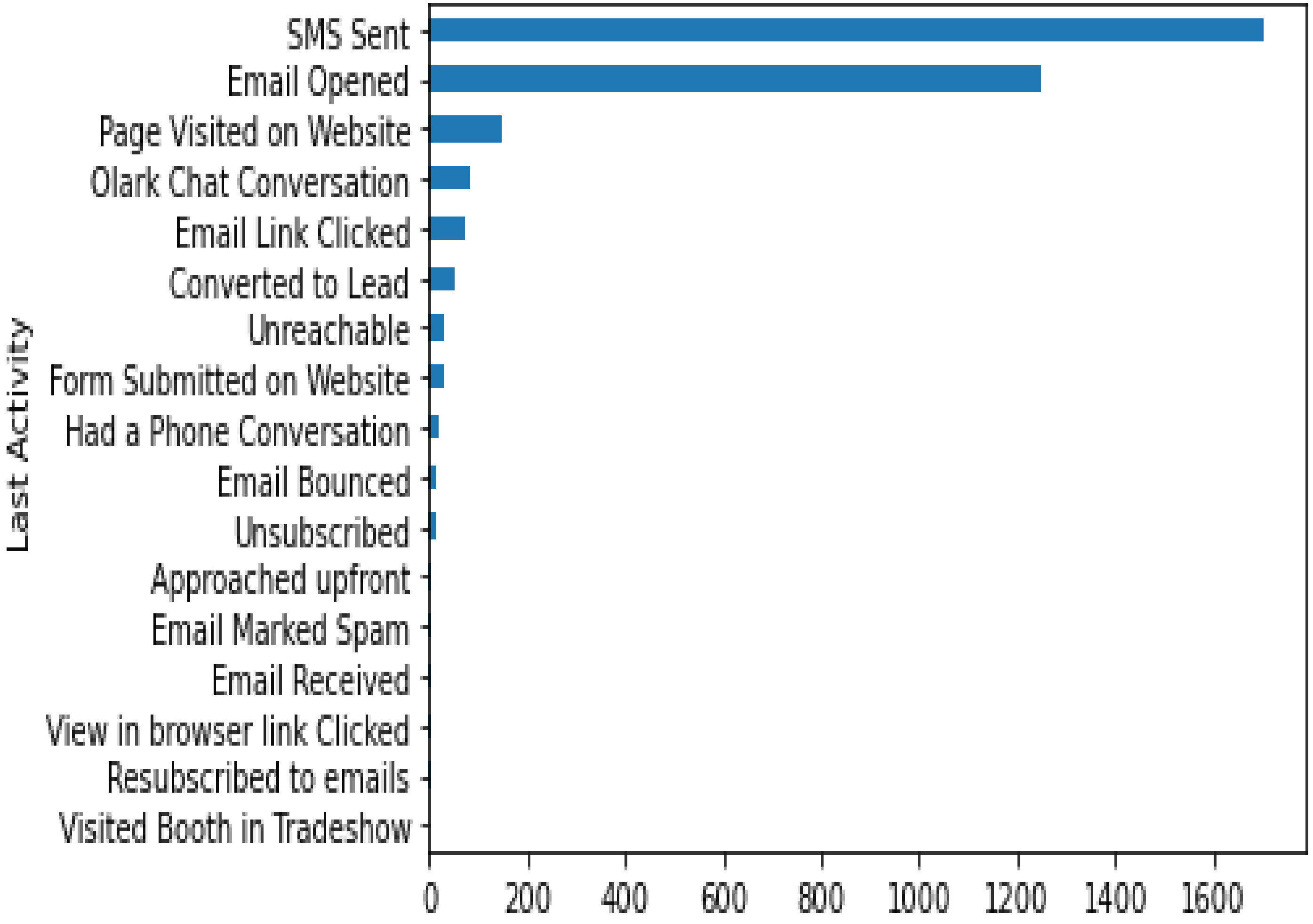
EDA

From the above plot, it can be seen that there is no major difference in the number of pages viewed per visit between successful and unsuccessful leads. Thus there should be a focus on improving quality of information provided per page rather than just increasing the number of pages. Focusing on Concise communication of information compared to Quantity of information.



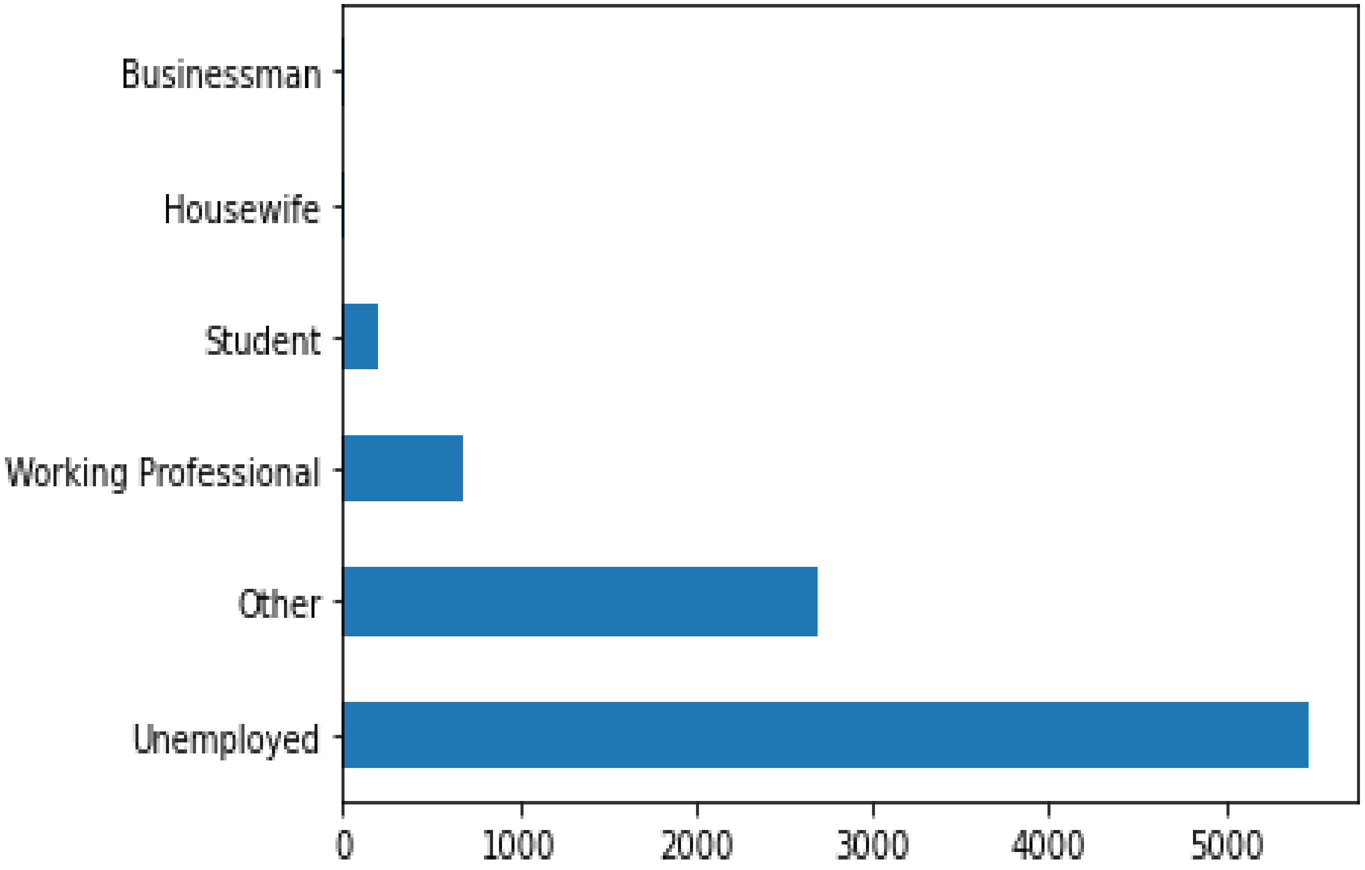
EDA

From the above plot, it can be seen that sending SMS and emails are better ways to achieve lead conversion. Which indicates that the sale of courses happens as a "Push marketing" rather than "Pull Marketing".



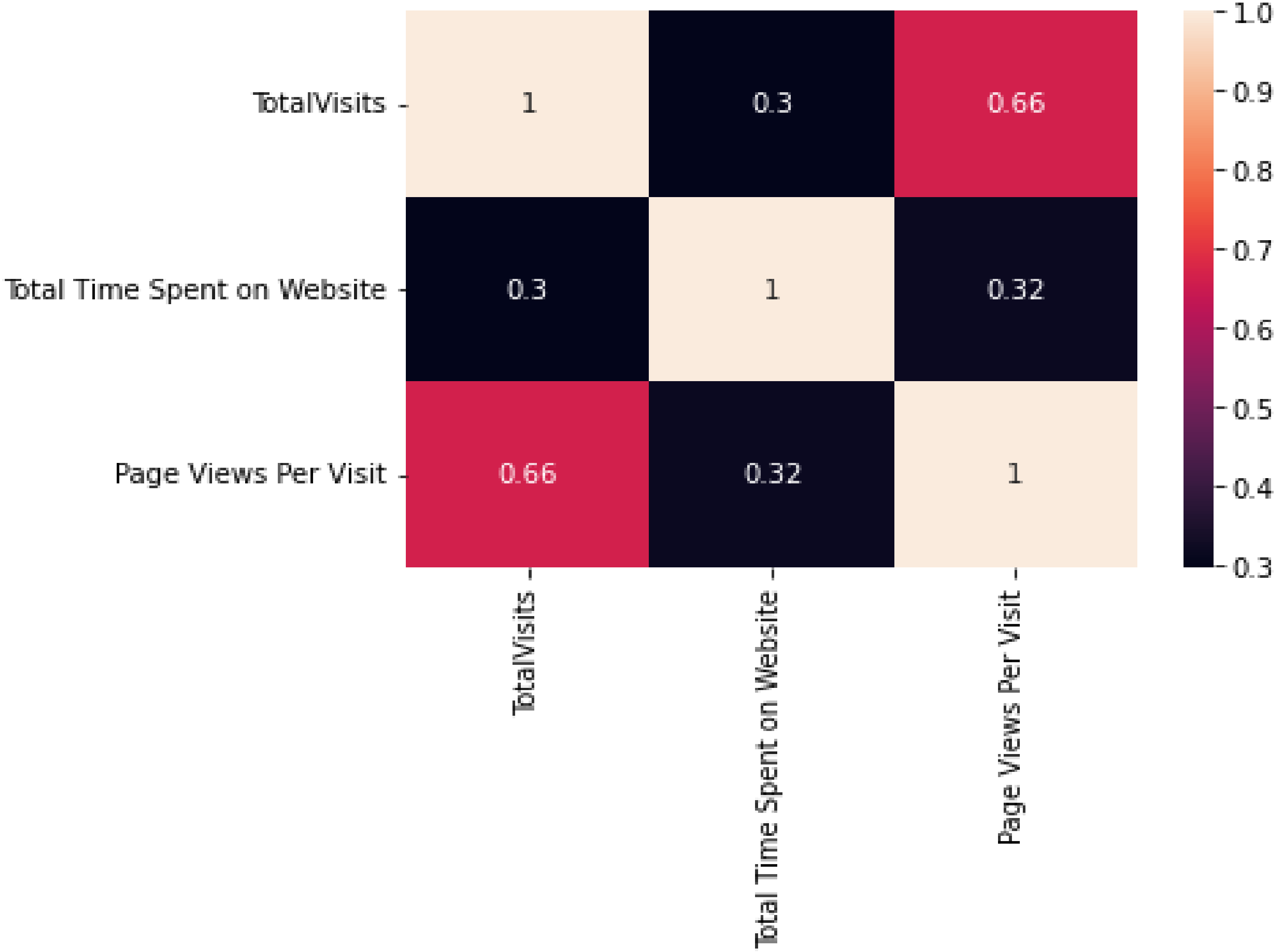
EDA

From the above plot it can be seen that People who are currently unemployed contibute to higher number of leads for the business. This information can be used to target potential learners and also in designing of courses to meet the requirements of the respective customer segment.



EDA

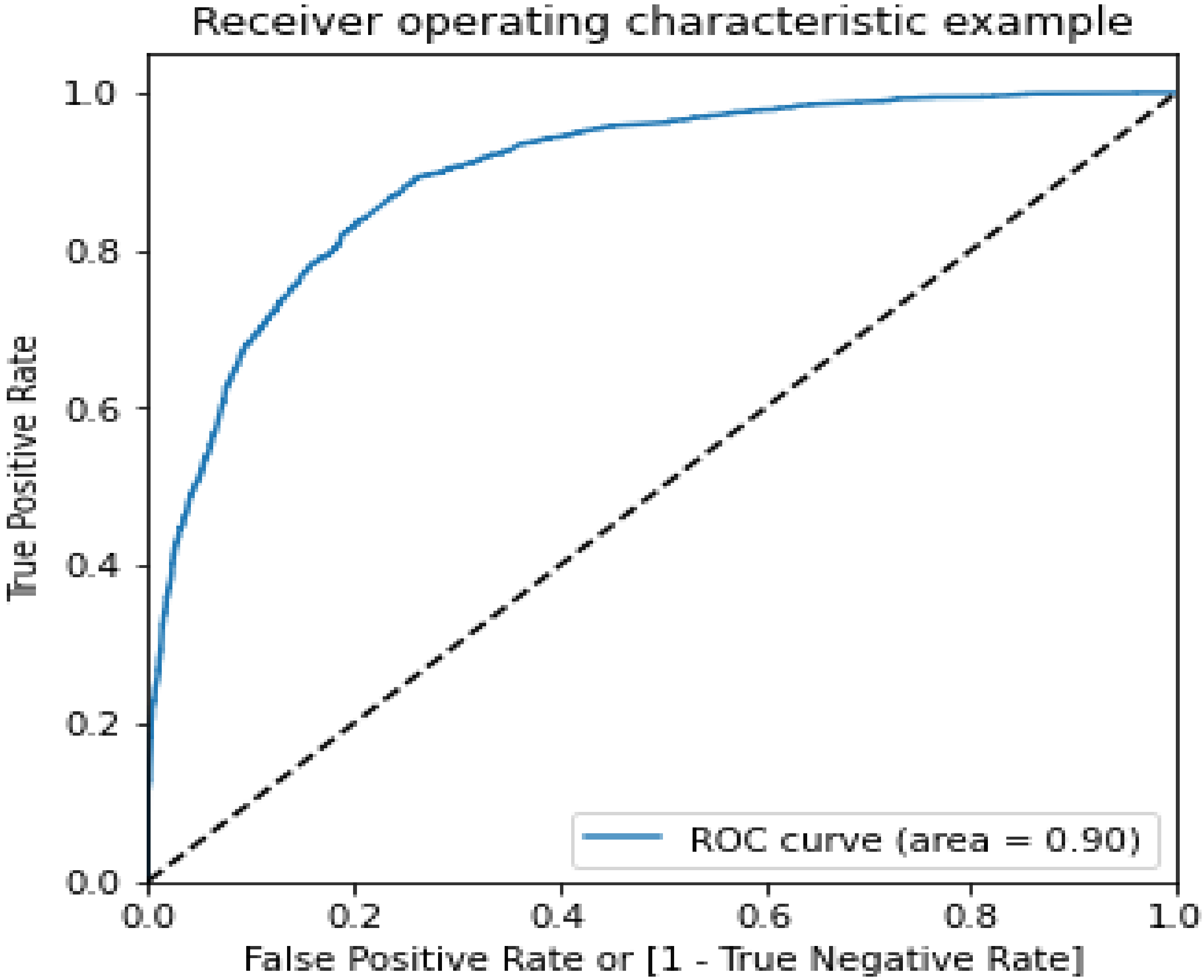
From the above heat map it can be seen that there is a noticable positive correlation, between "TotalVisits" and "Page Views per Visit"



- Data preparation
  - Creating dummy variable from all categorical variables
  - Target variable is converted
  - All remaining variables are independent variables
- Model building
  - Use RFE to remain 20 variables
  - Splitting the data into training and test sets by chosen 70:30 ratio
  - Building model by removing p-value greater than 5% and VIF greater 5
  - Prediciton on test data set
  - Overall accuracy is nearly 0.82

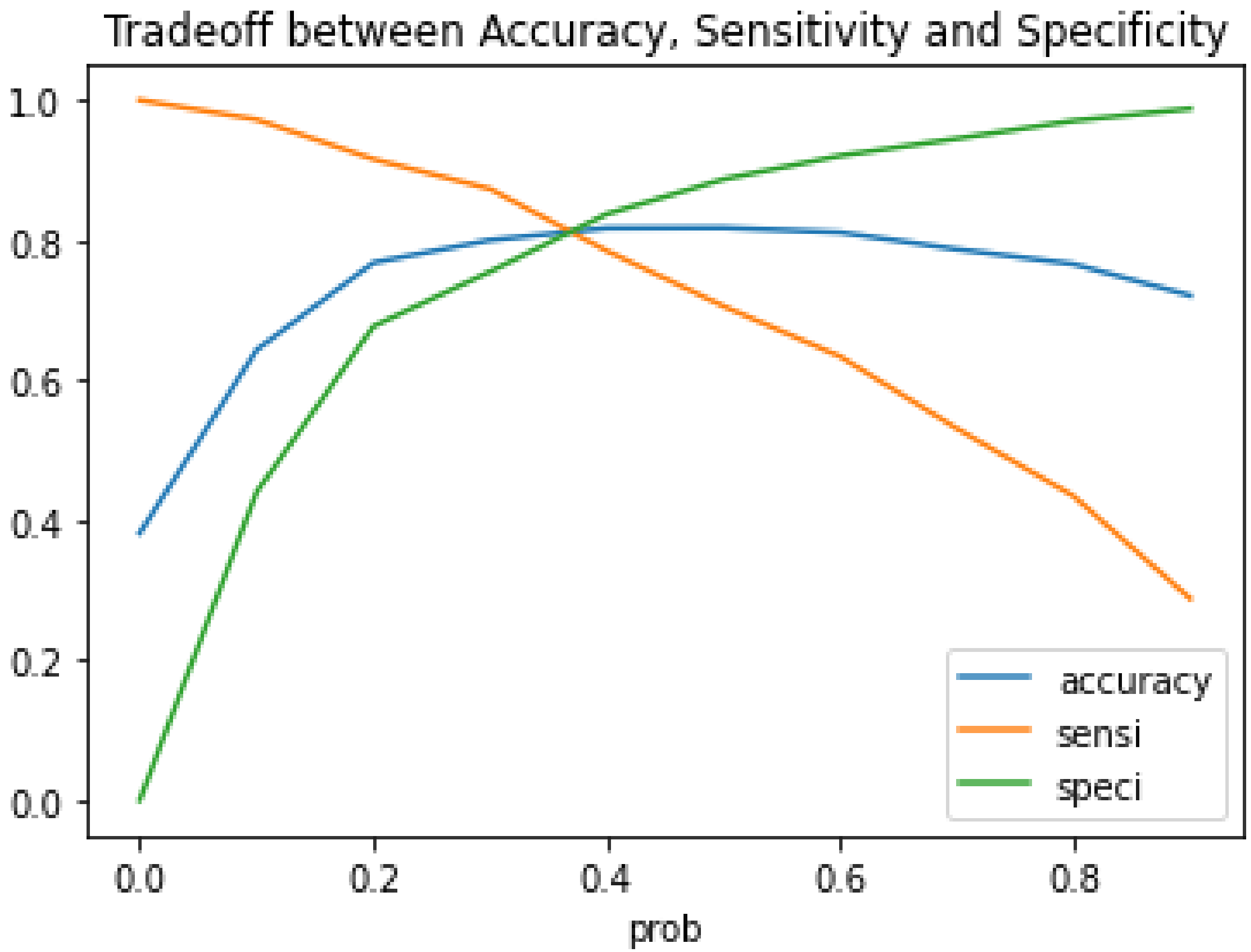
ROC CURVE

The Area Under Curve(AUC) is 0.9, which is a good value for AUC



EDA

From the plot it can be seen that a probability of 40% can be used as the cutoff for having a balanced sensitivity-specificity



Model evaluation

- Predict dependent variable
- Accuracy score is about 0.82
- Sensitivity of model is 0.78
- Specificity of model is 0.84
- Positive predictive value is 0.75
- Negative predictive value is 0.86
- False positive rate is 0.16
- The recommended threshold for the Lead Score for classifying as “Hot Lead” is 40.