Lead Scoring Case Study

Introduction

This assignment aims Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Business Objectives – 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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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. The typical lead conversion rate at X education is around 30%.

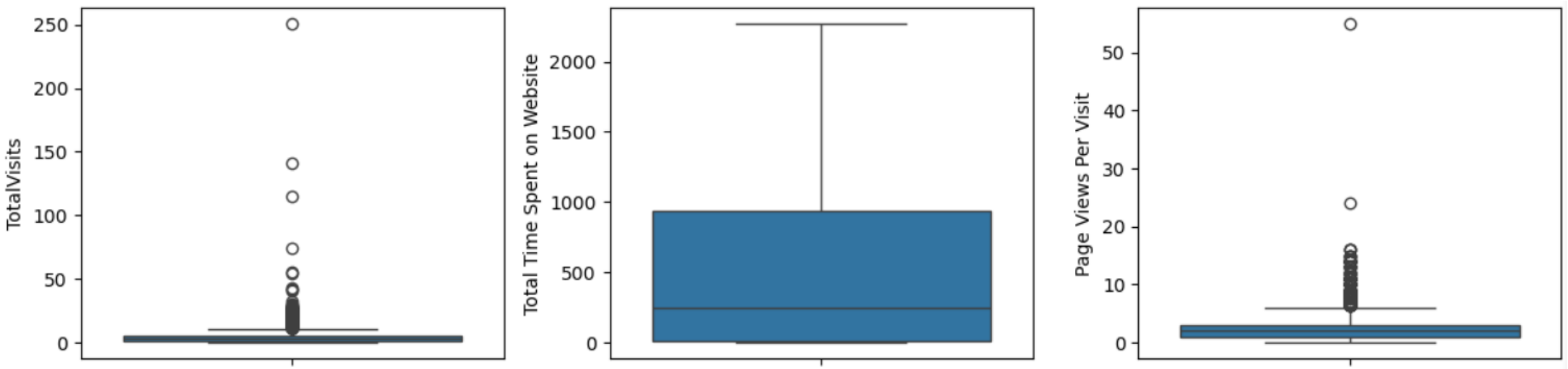
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If 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.

Analysis approach-

- Importing required libraries and Input data loading
- Inspect the Input data, Exploratory Data Analysis (EDA) and Visualization
- Data Preparation
- Train and Test data split
- Feature Scaling
- Looking at correlations
- Feature Selection using RFE and based on Manual Selection
- Model Building
- Model Evaluation and Prediction

Outliers - Boxplots

Outliers are values that are much beyond or far from the next nearest data points.

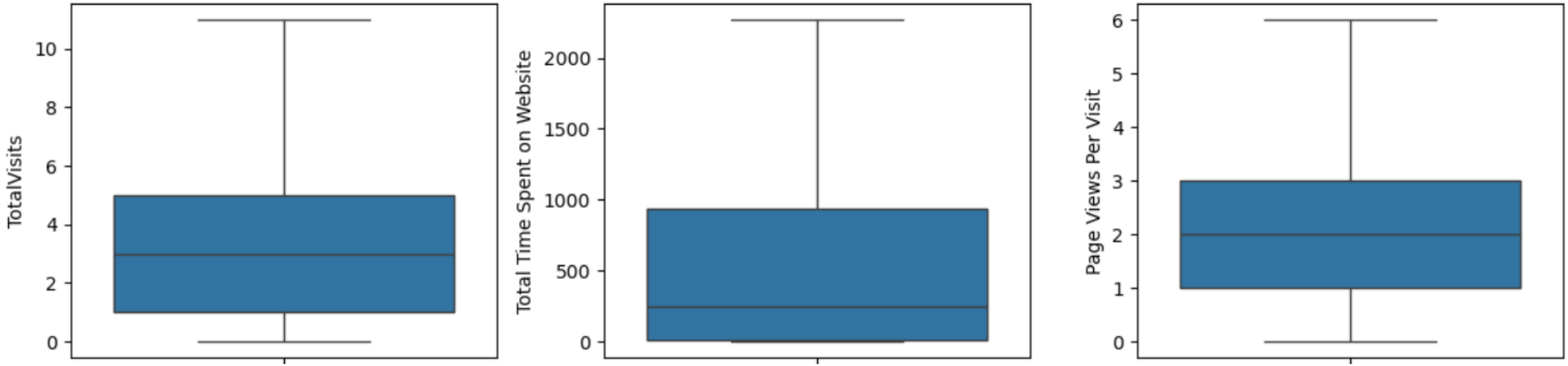


- It is observed the attributes **TotalVisits** and **Page Views Per Visit** have data points above the upper whisker respectively. Hence, assessing data further to understand the distribution.
- **TotalVisits** : ~3% records have TotalVisits value greater than Upper Whisker value.
- **Page Views Per Visit :** ~4% records have Page Viewes Per Visit value greater than Upper Whisker value.

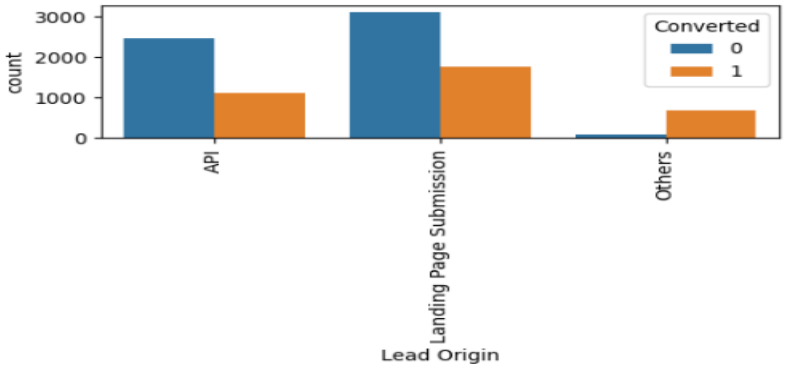
However, most of these visits are from Unemployed customers and logically these customers would be eager to upskill to get jobs. Hence NOT removing these records as outliers, instead capping the records with the upper limit values which will allow us to retain these records for further assessment

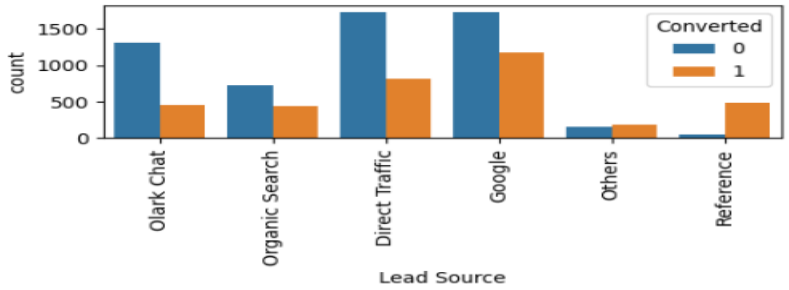
Data Distribution and Outlier Handling

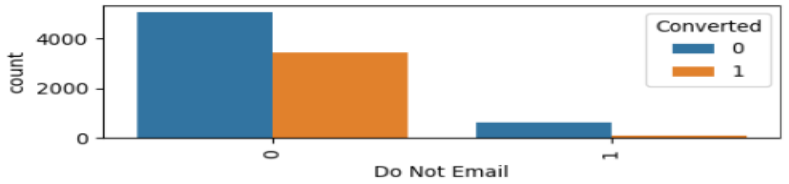
Outliers are treated based on the Capping (Boxplot method) approach for both the variables TotalVisits and Page Views Per Visit.

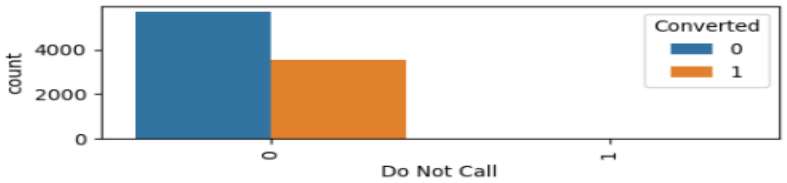


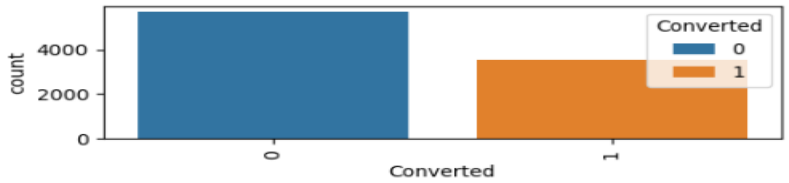
Univariate Analysis : analyze the single variable

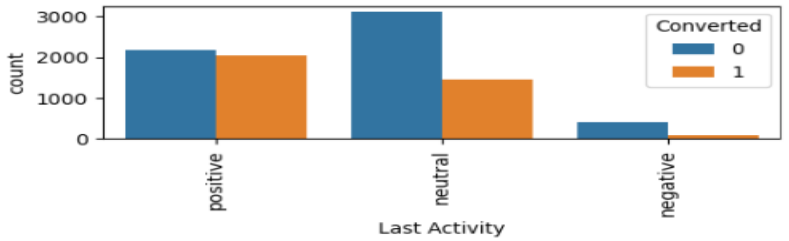


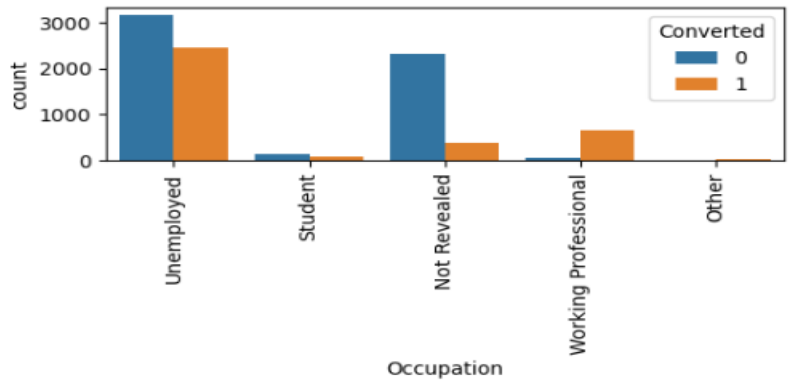


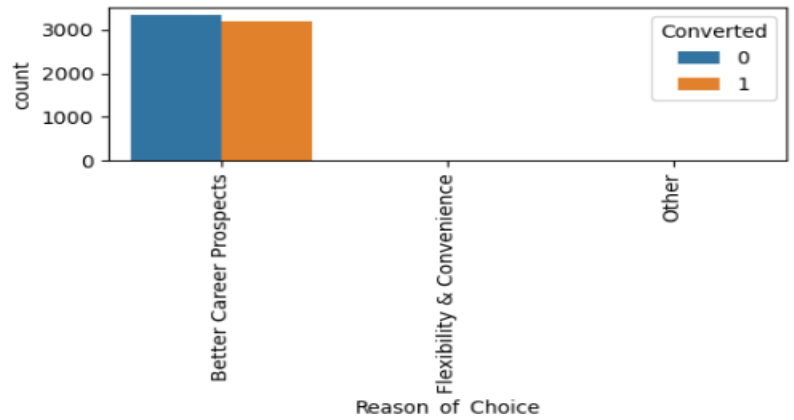


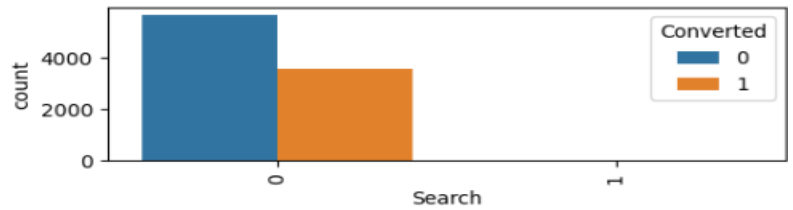








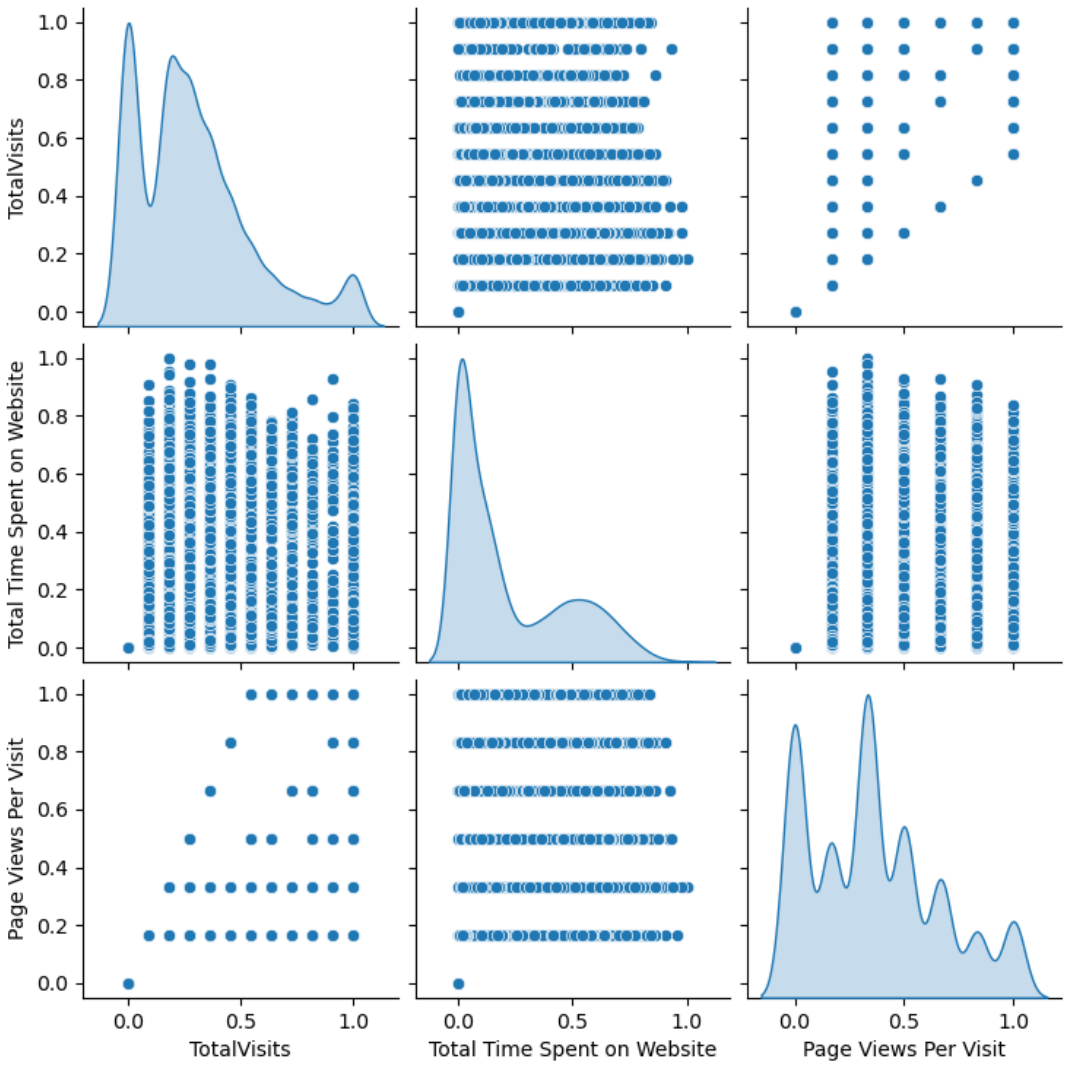


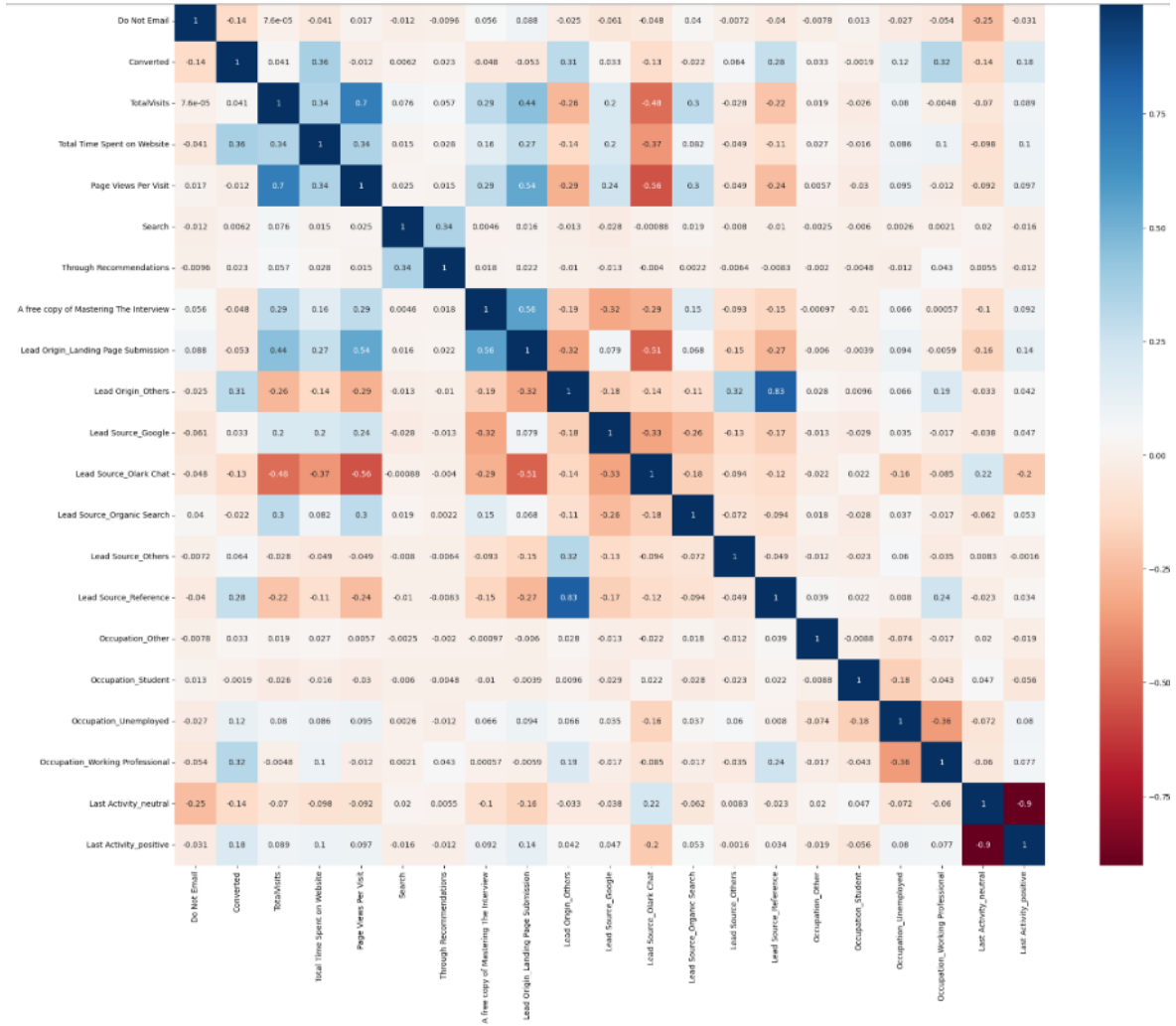


Multivariate Analysis : analyze the correlations

Based on the assessment and visualization -

- **Lead Origin_Other** and **Lead Source_Reference** are highly correlated (0.83)
- **TotalVisits** & **Page Views Per Visit** are highly correlated (0.70)





Model Building :

Based on the data analysis performed and mentioned in previous slides, the Multivariate Logistic Regression Models are built to solve the business problems stated in slide 2. The features selected for the Logistic Regression Model is based on following approaches:

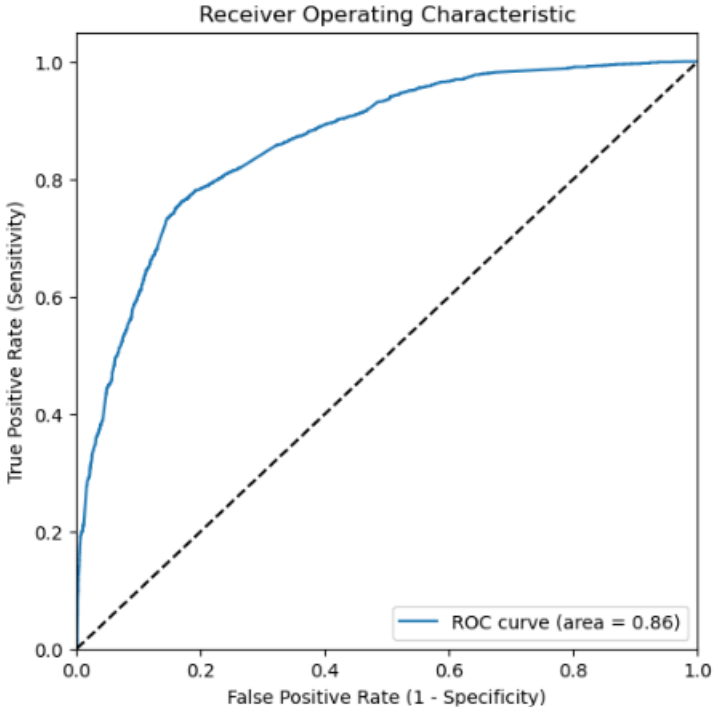
- automated feature selection(Recursive feature elimination) using SKLearn
- Manual feature selections based on the stats data

Two Models built with top 10 features and metrics evaluated the with final model (Model 2) highlighting the below outcome :

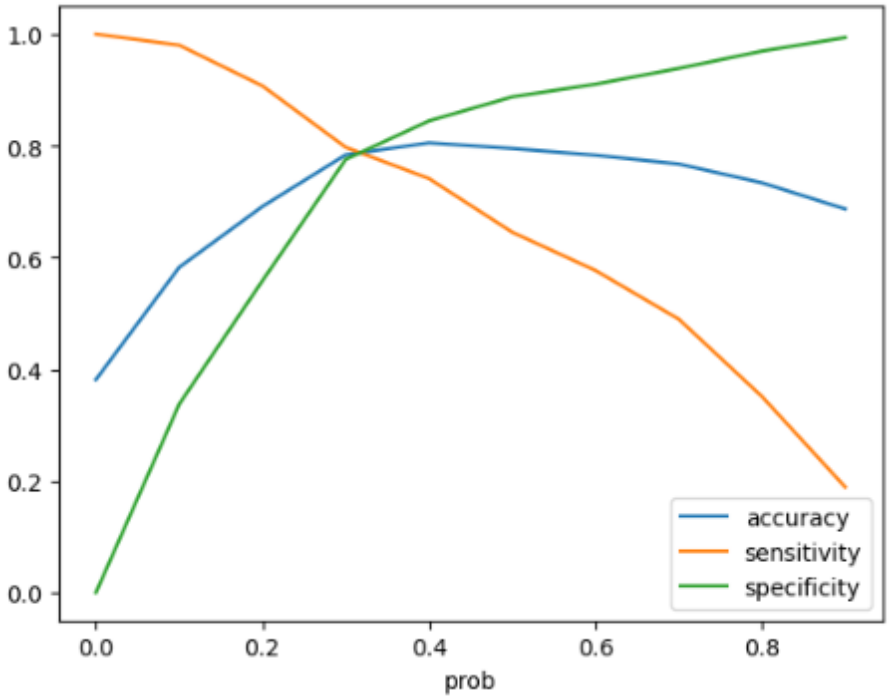
- Accuracy of the model 2 is **~80%** and the sensitivity metric is **~65%**. The cause of such a low sensitivity would be the arbitrarily chosen cut-off of 0.5. The threshold might not be the ideal cut-off point for classification and hence resulting in low sensitivity.
- The Optimal Probability Cutoff Point is assessed by calculating and plotting the Accuracy, Sensitivity and Specificity for various probability and the above model 2 is re-run with the cut-off of 0.32.
- The business objective is to increase in the lead conversions. Hence, achieving good sensitivity metric is important as the business wants to understand how well that an actual 'Lead' is predicted correctly and not miss out on these potential customers.

Graphs to interpret final model (Model 2)

Visualizing the ROC Curve generated based on the final model (Model 2) with cut-off value 0.50.



Plotting Accuracy, Sensitivity and Specificity for various probability to find the Optimal Probability Cutoff Point



- Good ROC curve is the one which touches the upper-left corner of the graph; so higher the area under the curve of an ROC curve, the better is your model. The above graph has decent area under the curve =0.86
- At about threshold of 0.32, the curves of accuracy, sensitivity and specificity intersect, and they all take a value of around 77-78%

Model Evaluation and Prediction

```
Model Evaluation Metrics on Train dataset with cutoff > 0.50
---------------------------------------
Confusion Matrix:
True Negative: 3552      False Positive: 450
False Negative: 875      True Positive: 1591

Overall model accuracy: 0.7951453308596166
Sensitivity / Recall:  0.6451743714517437
Specificity:  0.8875562218890555
False Positive Rate:  0.11244377811094453
Positive Predictive Value:  0.7795198432141107
Positive Predictive Value:  0.8023492206912131
```

```
Model Evaluation Metrics on Train dataset with cutoff > 0.32
---------------------------------------
Confusion Matrix:
True Negative: 3195      False Positive: 807
False Negative: 531      True Positive: 1935

Overall model accuracy: 0.7931354359925789
Sensitivity / Recall:  0.7846715328467153
Specificity:  0.7983508245877061
False Positive Rate:  0.20164917541229385
Positive Predictive Value:  0.7056892778993435
Positive Predictive Value:  0.857487922705314
```

```
Model Evaluation Metrics on Test dataset with cutoff > 0.32
---------------------------------------
Confusion Matrix:
True Negative: 1330      False Positive: 347
False Negative: 218      True Positive: 877

Overall model accuracy: 0.7961760461760462
Sensitivity / Recall:  0.8009132420091324
Specificity:  0.793082886106142
False Positive Rate:  0.20691711389385808
Positive Predictive Value:  0.7165032679738562
Positive Predictive Value:  0.8591731266149871
```

Result -
The metrics as show holds on the test dataset as well. Hence, the final model (Model 2) created for the Leads dataset is performing well with similar metrics captured for both the training and test datasets

Summary and Recommendations :

- Top 3 variables which contribute the most towards the probability of a lead getting converted:
  - What is your current occupation (Unemployed Professional)
  - Total Time Spent on Website
  - Last Activity being performed by the customer (Activity positive)
- Top 3 categorical variables which contribute the most towards the probability of a lead getting converted:
  - What is your current occupation
  - Last Activity being performed by the customer
  - Lead Origin
- Company during the peak season wishes the sales team make the lead conversion more aggressively and wants almost all of the potential leads to be converted. The sales team should leverage the LEAD SCOREs calculated to identify the hot leads easily. Higher the score higher is the chance of lead conversion to paying customer. Hence, the sales team should try to reach out these potential customers from all possible communication channels starting with phone calls providing valuable information and engaging the customers to expedite the customer purchases.
- Similarly, during the times where the company wants the sales team to continue the conversions with minimal phone calls and also focus on other tasks. The sales team can make use of the auto generated emails and DMs with course updates / details to retain leads engagement with minimal human intervention which will enable the sales team to also focus on the new work assigned. The sales team can get into calls with the leads in cases of the positive responses received by the potential customers.

Hence the X Education is suggested to consider these variables given the influence on the end results and plan accordingly to increase the conversions. Additionally, we can definitely explore more options in adding or deleting the features to get better results than the already existing Model 2 which is overall a good model based on the output generated.