Lead Scoring Case Study

Problem Statement

- X Education Sells online courses to Industry Professional. The Company markets it’s on their own websites and search engines like Google.
- All these peoples fills the form on their website by providing the email address or phone numbers, they are classified as leads. Company also gets leads from their past referrals.
- Once these leads are acquired, employee from the sales team start making calls, writing emails etc. Through these process some of leads are converted or some of not. The typical lead conversion rates is only 30%.
- Their Lead conversation rate is very poor.

Business Objectives

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

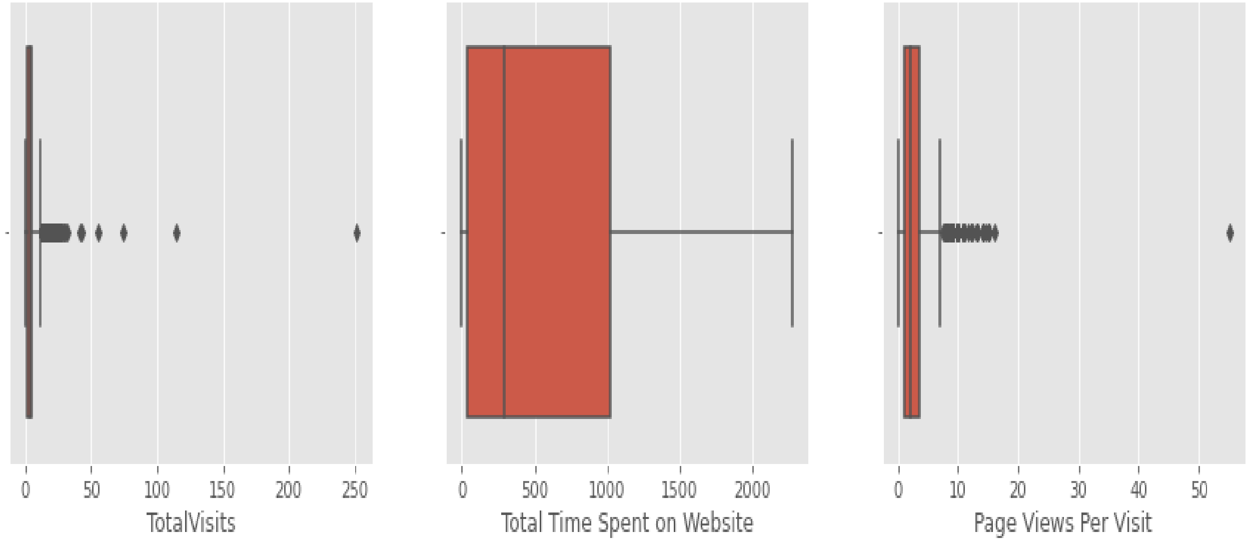
Approach

- Importing the data in Jupyter Note book.
- Data cleaning & preparations.
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data in Test & train.
- Building logistic Regression Model & calculate lead score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

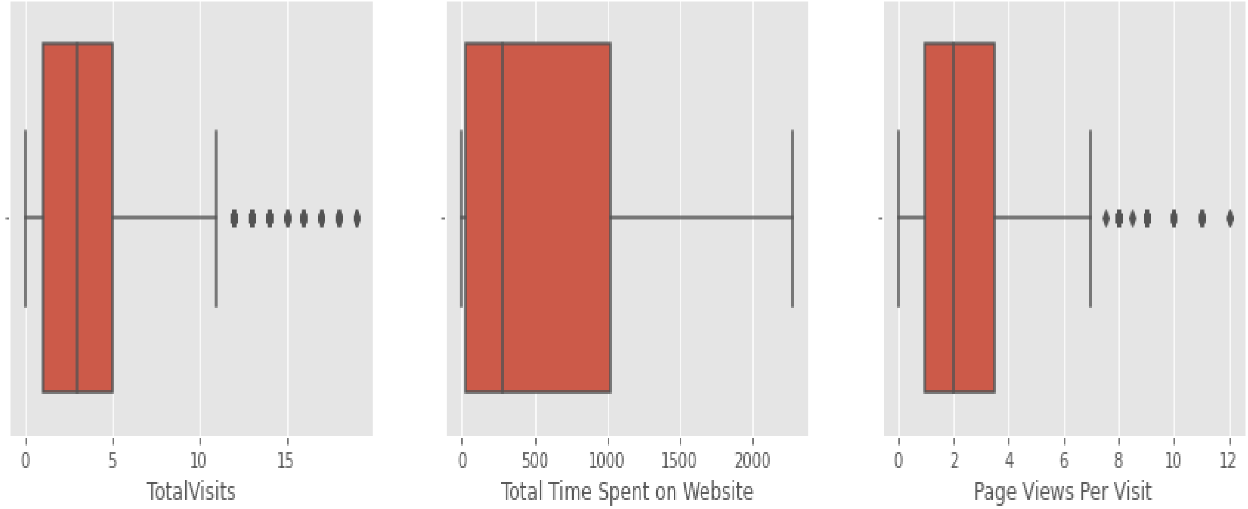
Exploratory Data Analysis

- Import data in Jupyter notebook.
- Data cleaning – Handling & removing null values.
- Removing unwanted columns from data set.
- Imputing null values.
- Outlier Treatment

Outlier Treatment



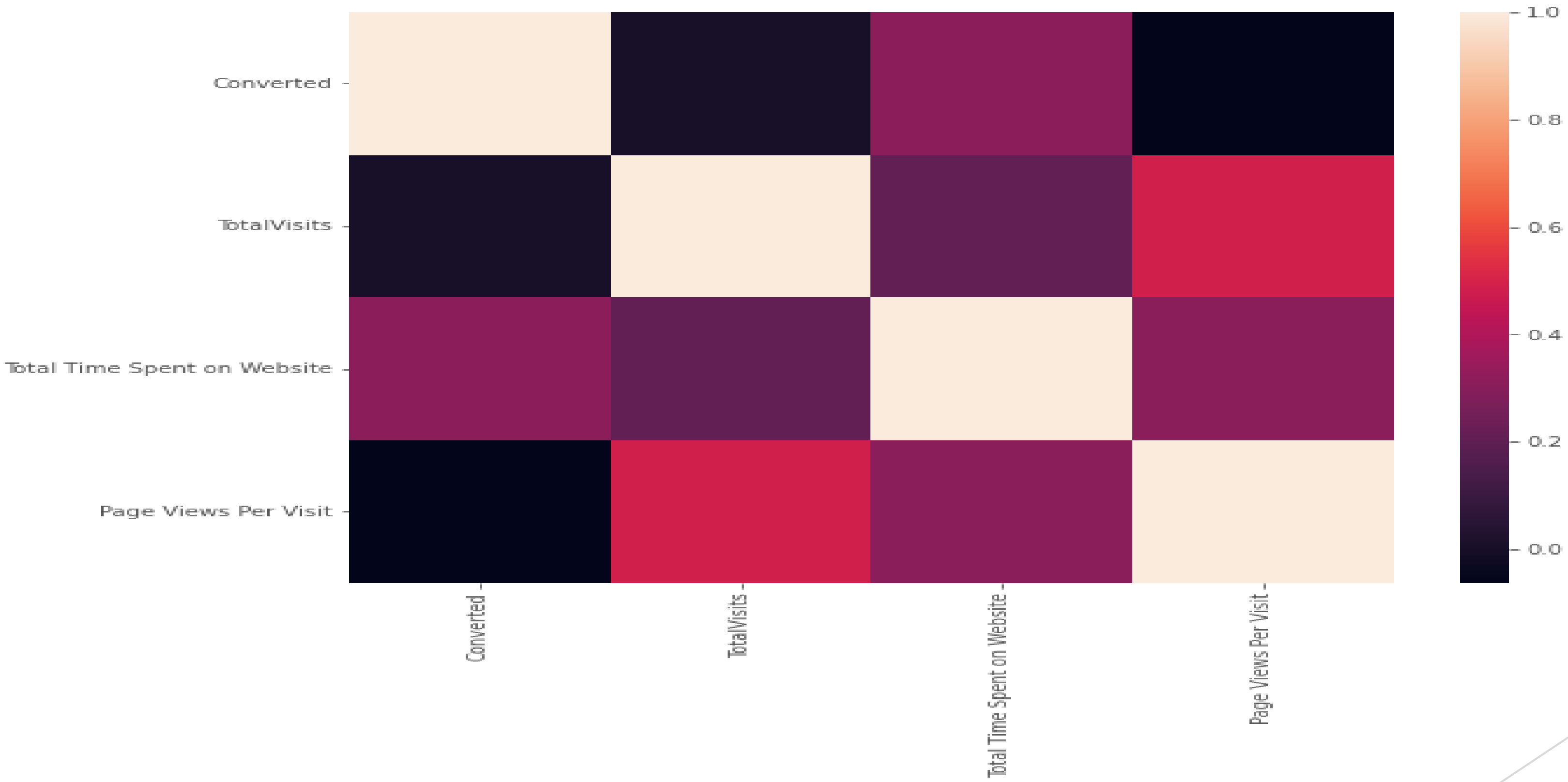
Total Visits & Page Views per Visit column contains outliers.



Boxplots after removing outliers from 10th & 90th percentile.

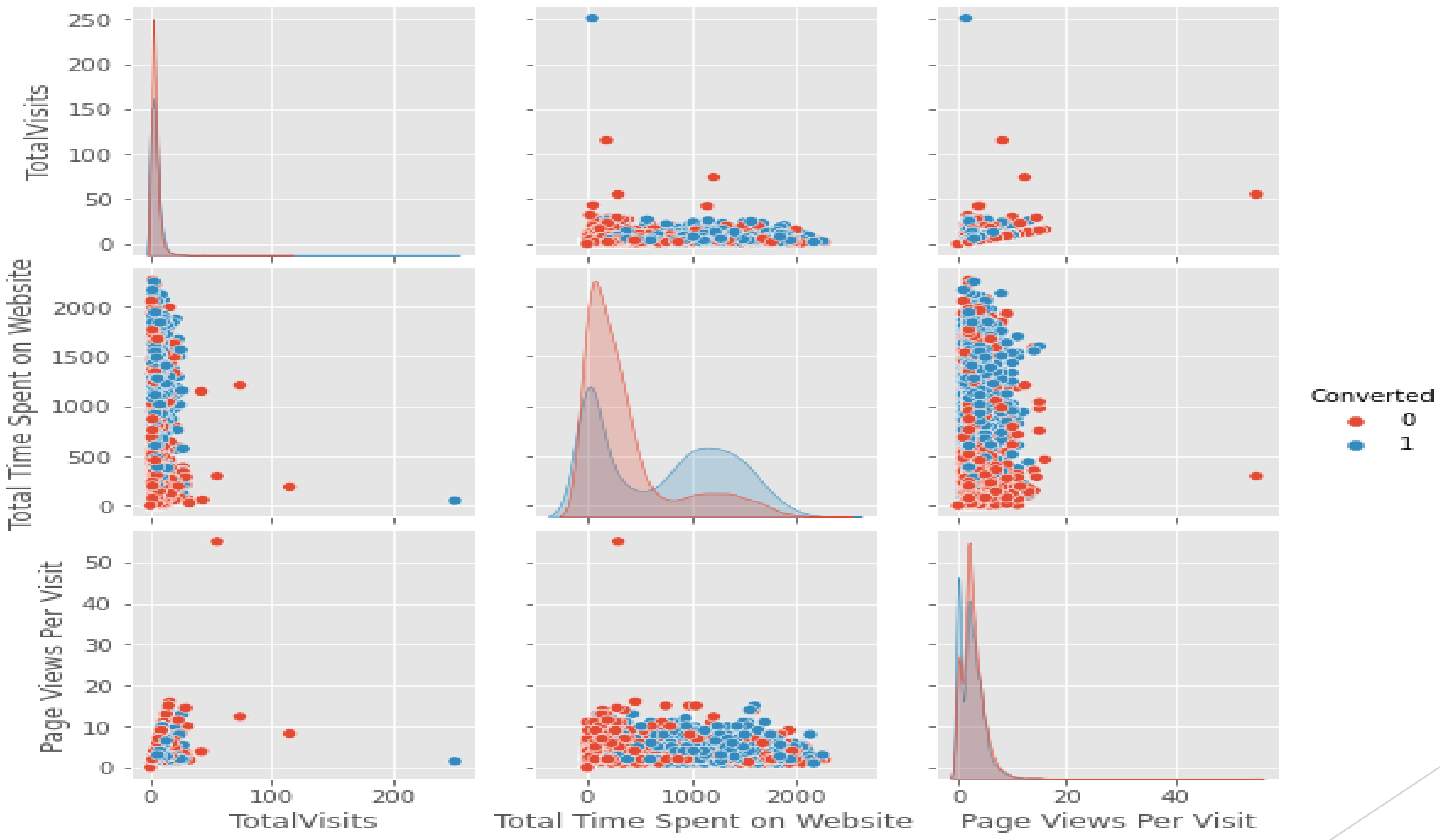
Heatmap

Heatmap of correlation between converted, Total Visits, Total Time Spent on Websites, Pages views per visit



Pairplot

Pairplot between converted, Total Visits, Total Time Spent on Websites, Pages views per visit
No linear correlation found.



Data Preparation

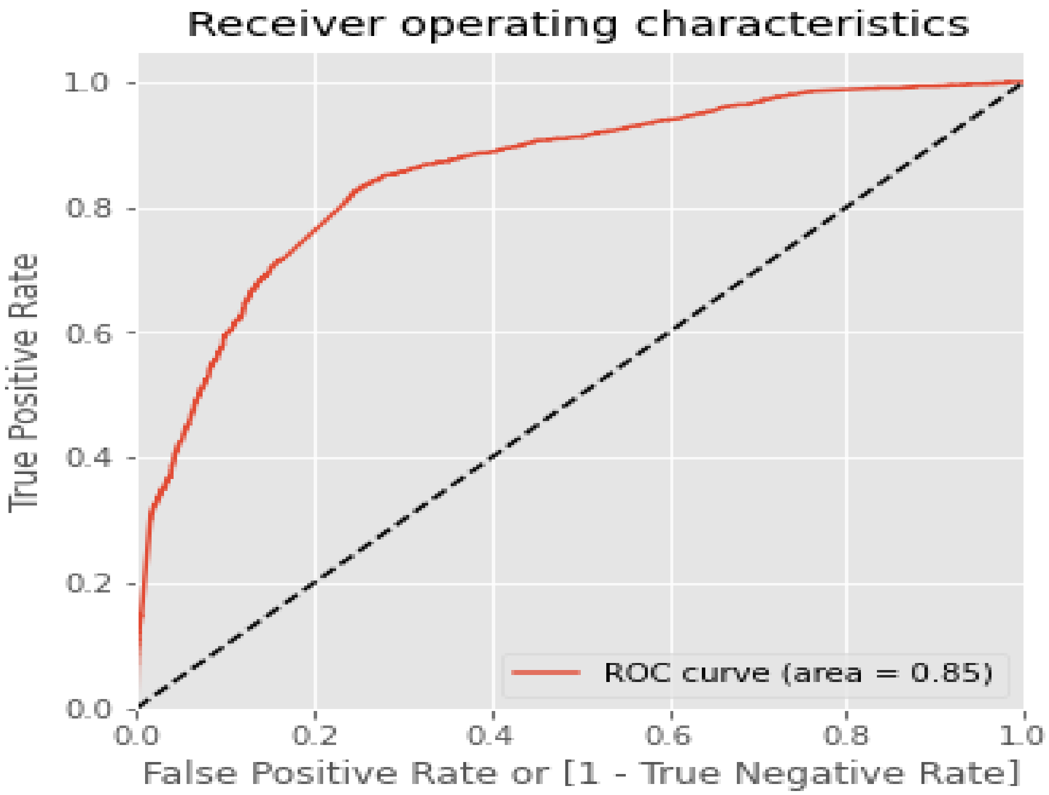
- Creating Dummy variable for categorical columns.
- Splitting the data into Test & Train Set.
- Scaling the data using Standard Scaler.

Model Building by Logistic Regression

- Feature Selection by using RFE
- Checking VIF
- VIF values grater than 5 & P-value greater 0.05 column dropped.
- Plotting ROC Curve.
- Finding out optimal cutoff values.

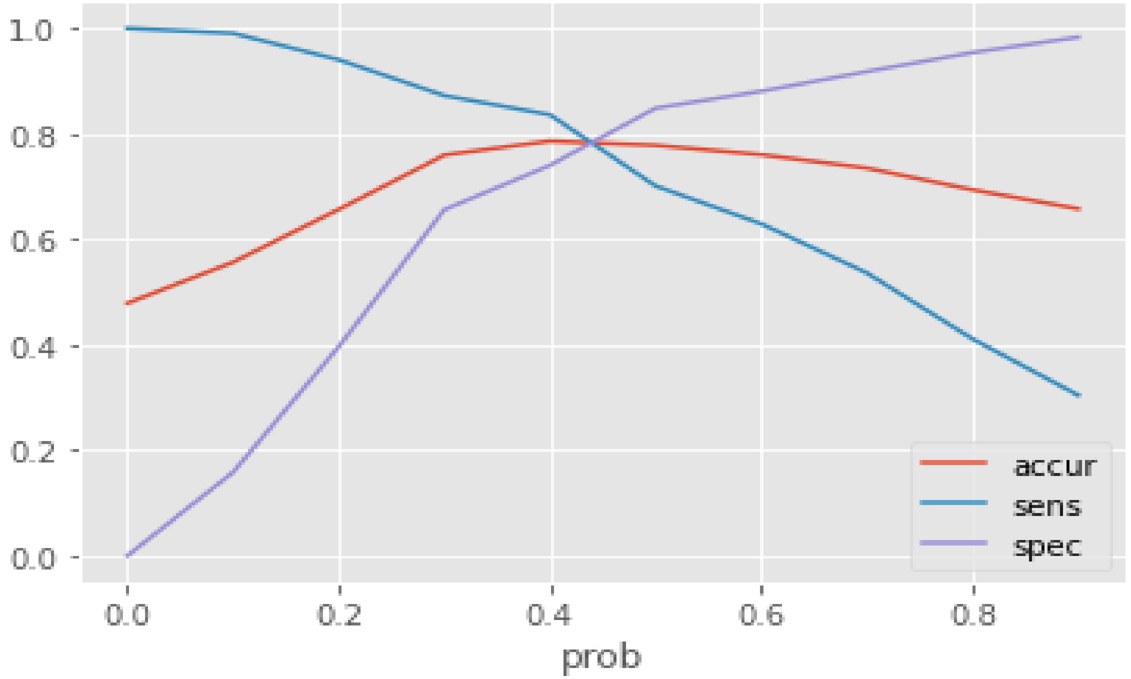
ROC Curve

- Area under ROC Curve is 0.85 which is quite good & represents good model.
- The closer the curve follows to left hand & top border of ROC space which suggest more accurate the test.
- The ROC Curve value should be closer to 1. We are getting good ROC value of 0.85 which represents good model.



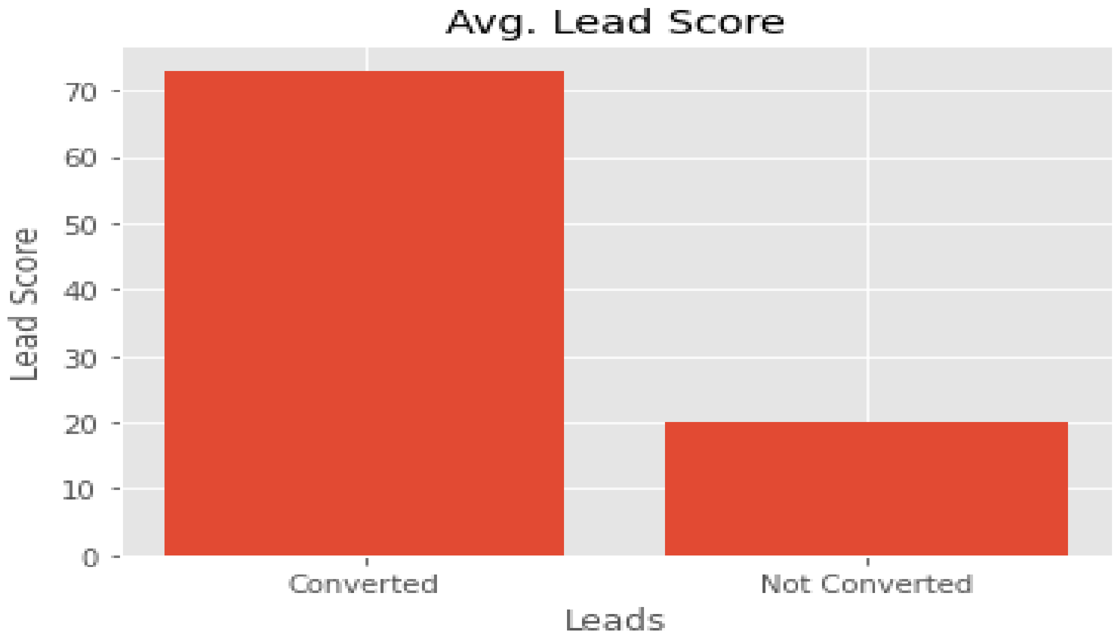
Optimal Cutoff value

- From the curve we can see that 0.42 is optimal point to cutoff probability.



Predicted Converted Leads

- The average lead scores of customers we where converted is 73.
- The average lead scores of customers we where not converted is 20.



Observation

- Accuracy , Specificity, Sensitivity values of Train data

1. Accuracy : 78%
2. Sensitivity: 83%
3. Specificity: 74%

- Accuracy , Specificity, Sensitivity values of Tes data

1. Accuracy : 77%
2. Sensitivity: 83%
3. Specificity: 74%

- The model seems to be performing well. which gives us confidence in recommending this model for making good calls

Thank you!