Lead Score Case Study

- REVANTH R

Problem Statement:

- An education company named X Education sells online courses to industry professionals.
- The typical lead conversion rate at X education is around 30%. This is poor conversion of potential leads.

Objective:

- The company identified some potential leads which can become Hot Leads.
- Need to prepare and build a model to target potential leads.

The following steps were performed to build a model:

- Reading Data
- Exploratory Data Analysis
- Dummy Variables
- Split into Train – Test
- Scaling of Data
- Model Building
- Train Model Evaluation
- Test Model Evaluation

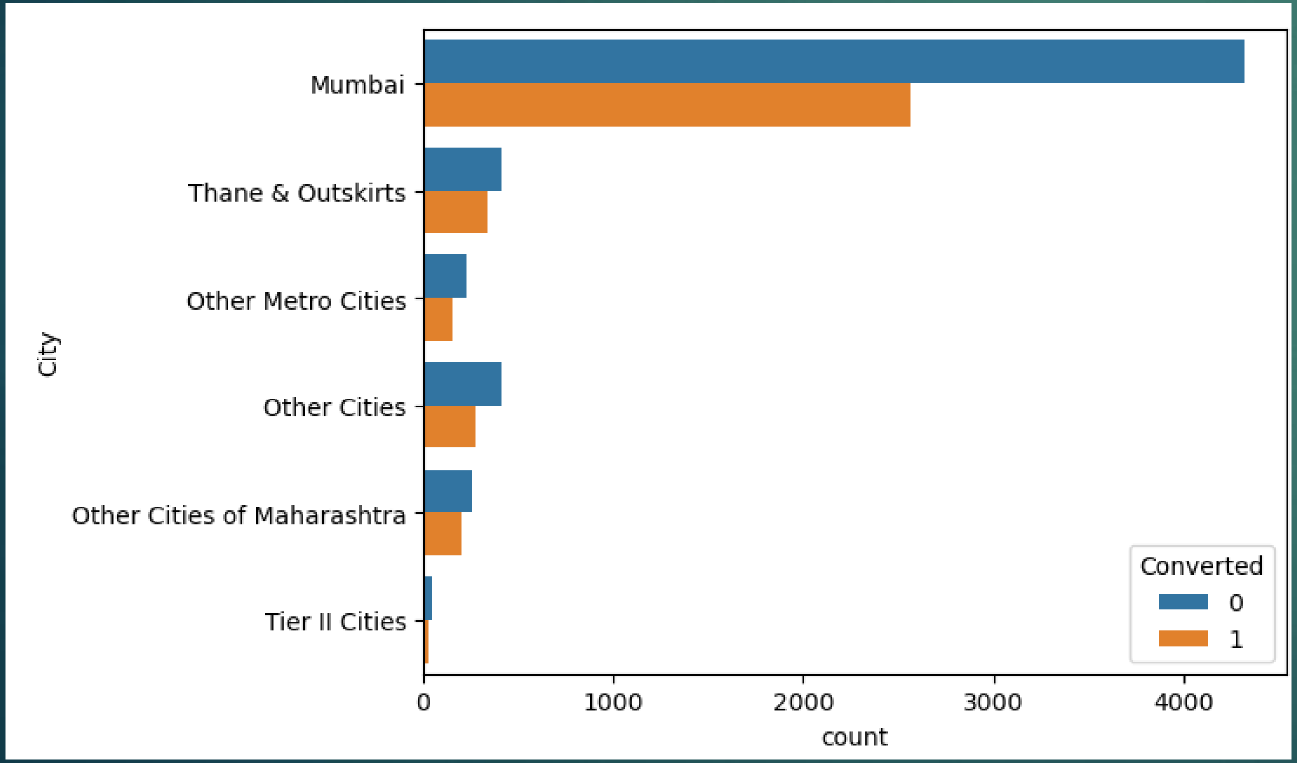
Following slides given brief description of each step.

Reading Data and Data Cleaning:

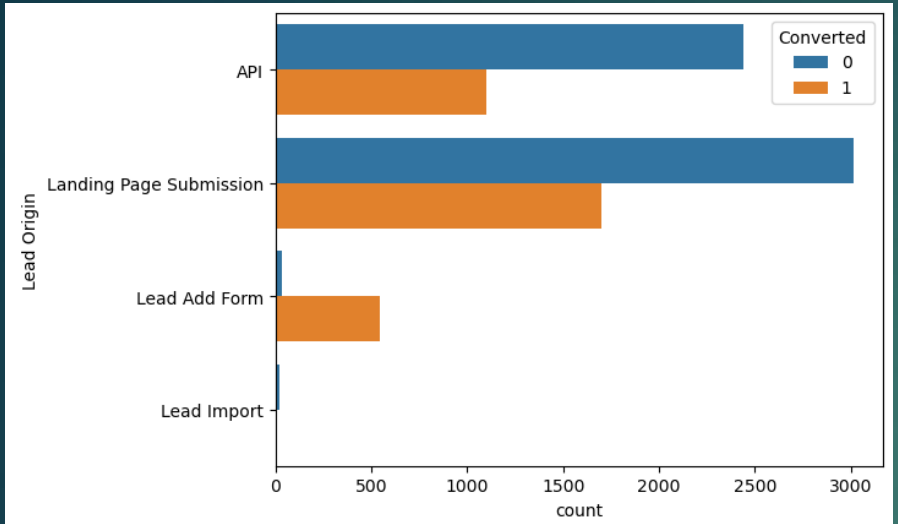
- Data is imported into python.
- Initially it contains 9240 rows and 37 columns of data.
- It contains missing values, some have select option in data.
- Select option is replaced by null values and more than 40% missing data columns were dropped.
- Categorical missing data is filled using most popular one.
- Some variables have very less data. It was combined with others. It is done as dummy creation will not create huge columns.
- Some columns have less than 2% data missing. Whole row of data was dropped as it will not make any effect on model building.

EDA:

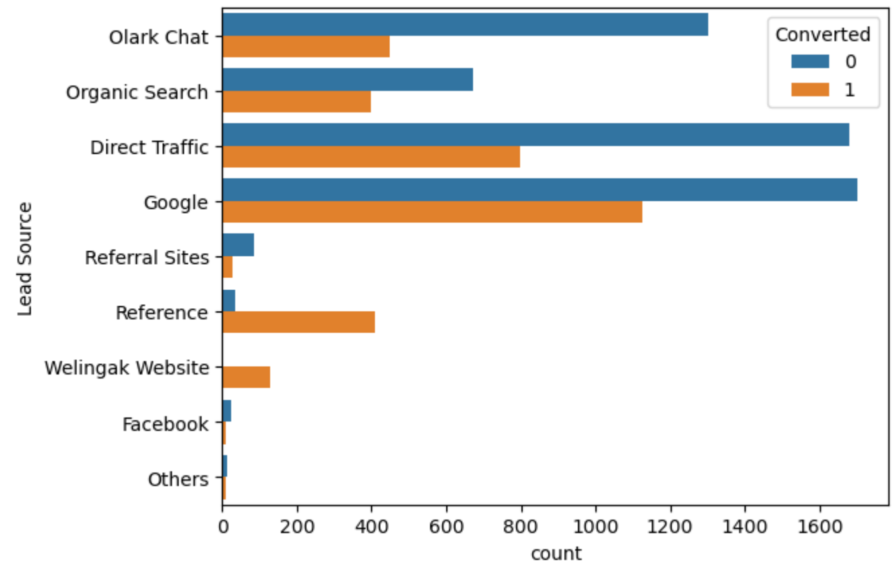
- Box Plot is used to identify the outliers in numerical columns.
- Count plot is used to compare the categorical variable with target variable.
- Some plots as shown below:



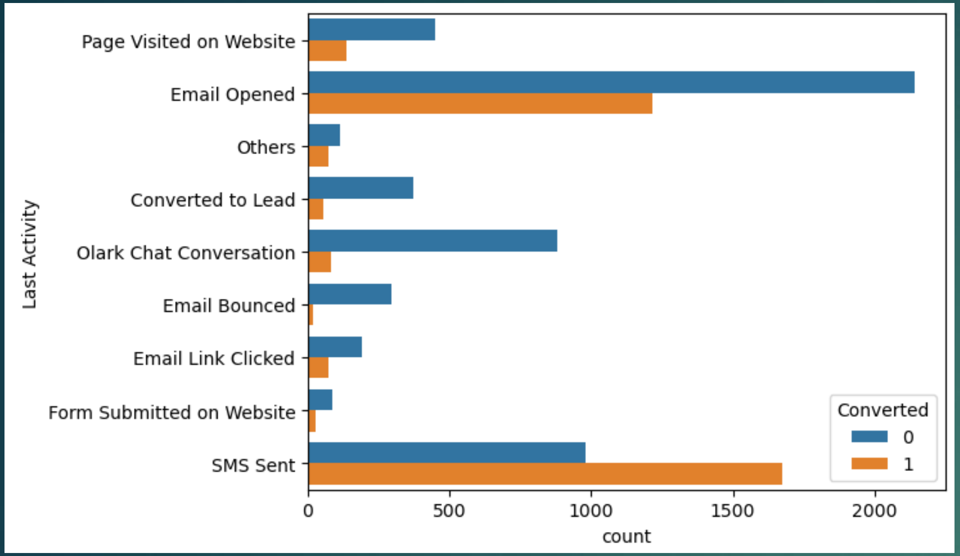
This shows Mumbai has got high number of leads converted to yes as compared to other cities.



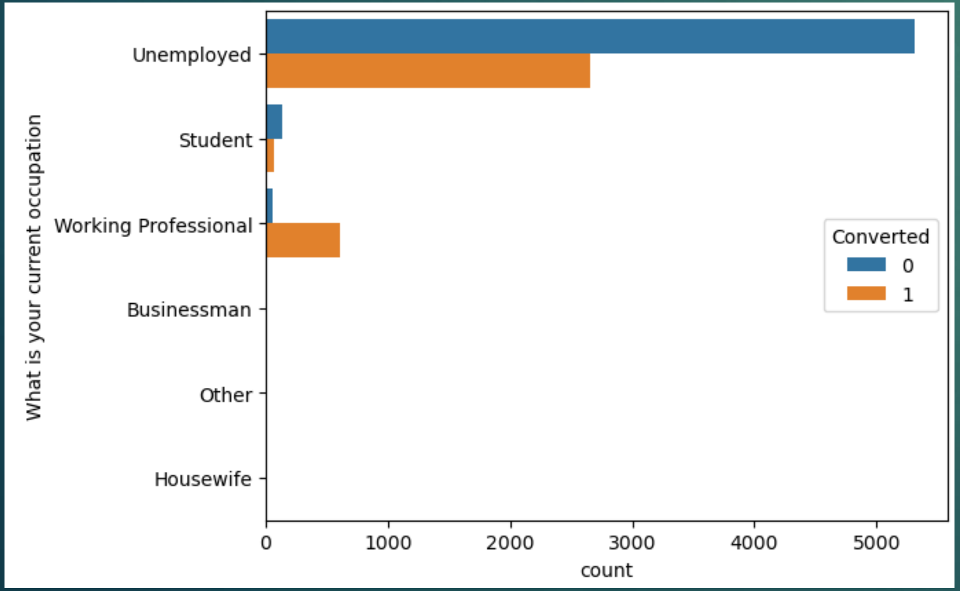
Among all, Landing Page Submission leads to more conversion to Yes as compared with remaining ones



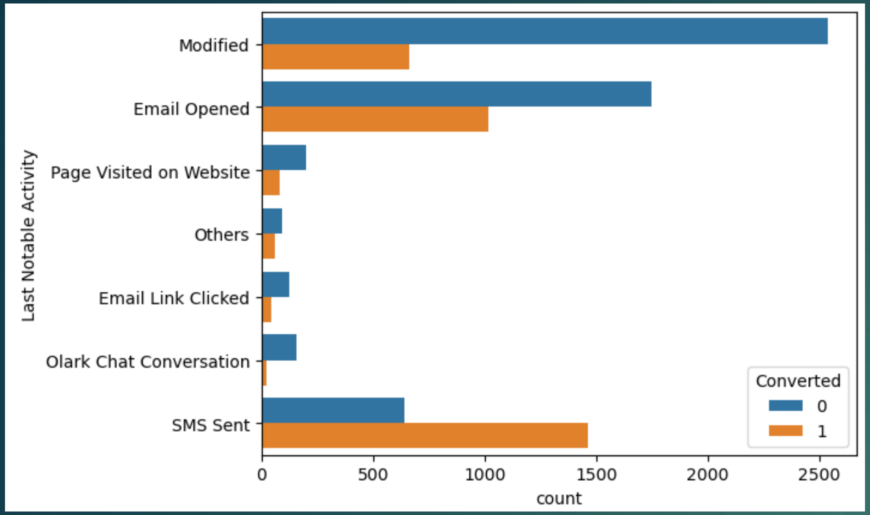
This shows google had contributed more conversion to yes as compared to others



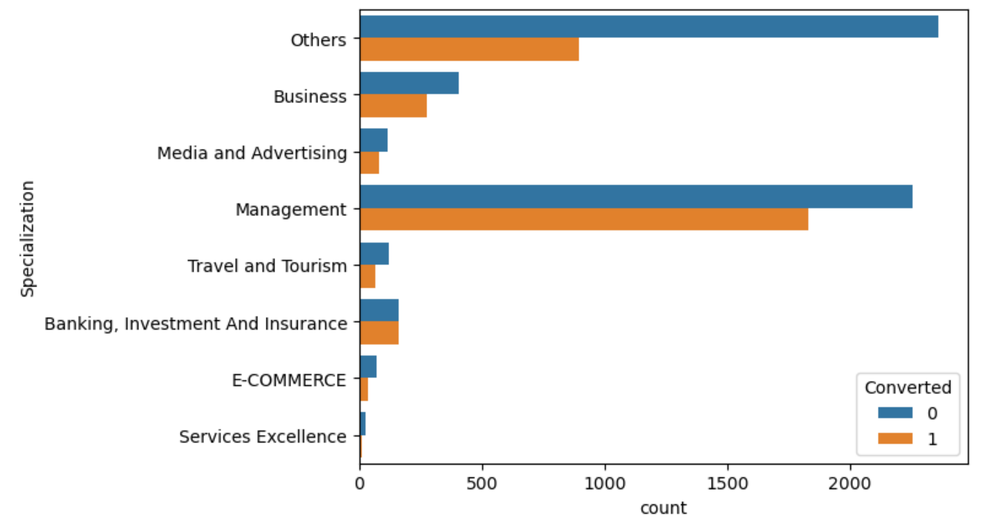
This shows Leads which had converted into Yes is by Email and SMS.



This shows working professional were more interested and more converted into Yes as compared to Others



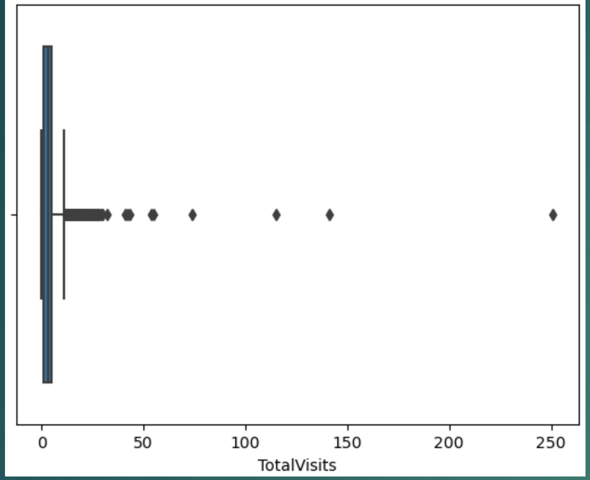
It shows majority are interested by SMS Sent option.



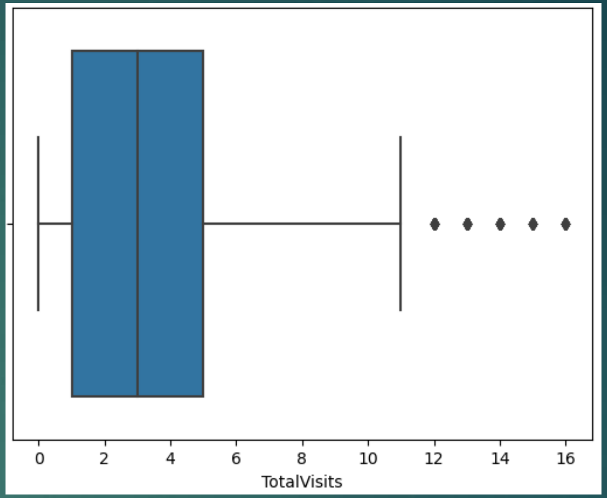
It shows management people are more interested as compared with the others.

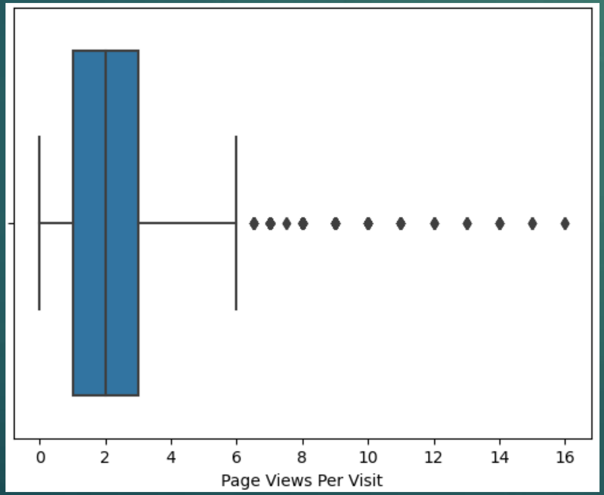
Outliers:

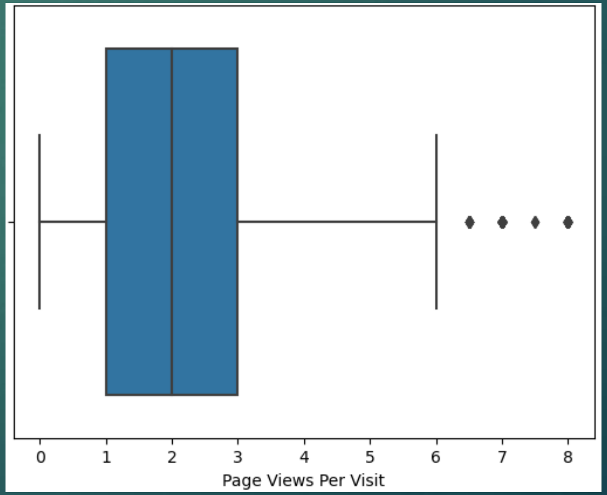
Before



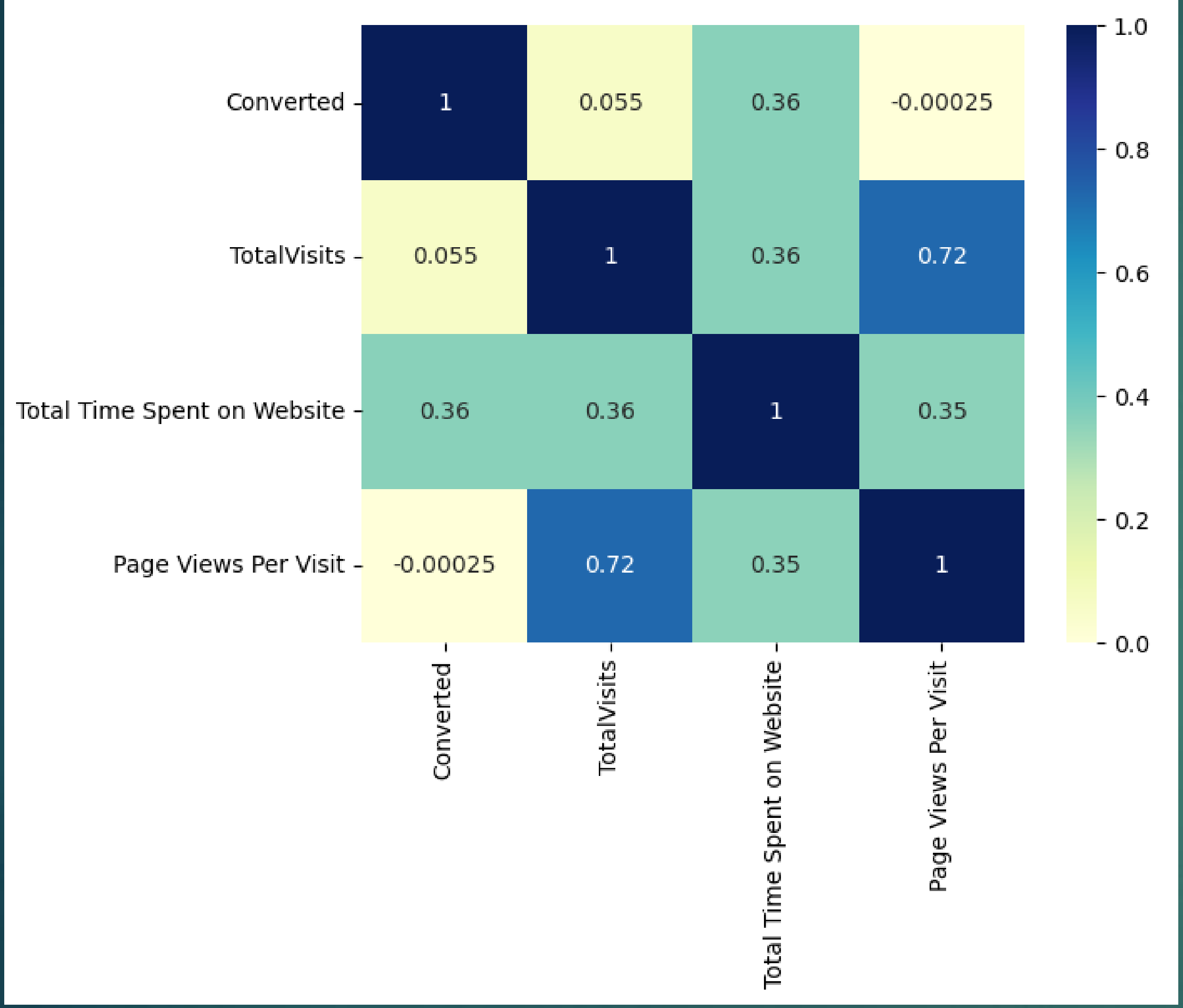
After





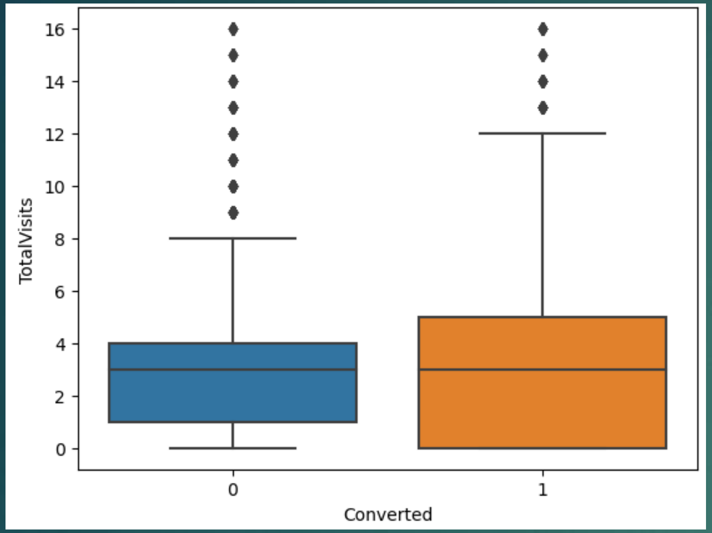


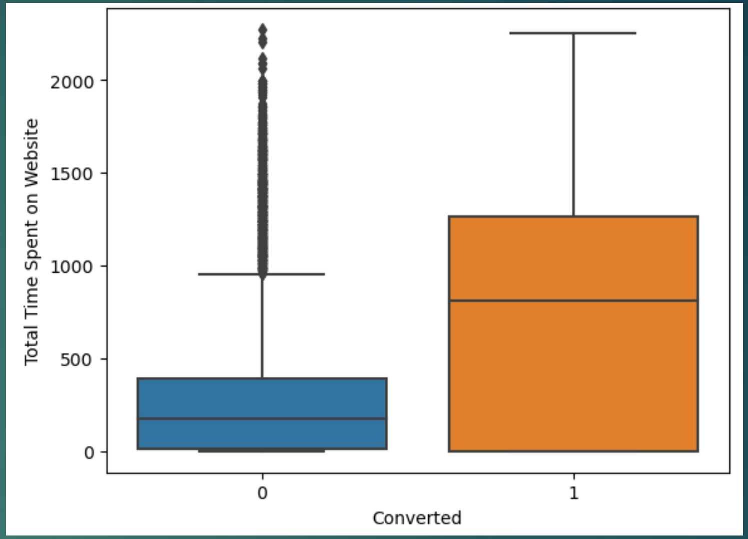
Heatmap Numerical Data Correlation



1. This shows TotalVisits Column and Page Views Per Visit Column were highly correlated.
2. Compared to all, Total Time Spent on Website is directly related with Converted column.

Boxplot – Numerical Vs Converted





Although median is same for both, but 3rd quartile is higher for Yes as compared with No. By looking at box plot, TotalVisits converted to Yes is more.

This clearly shows that 'Total Time Spent on Website' has shown more conversion to Yes as compared to No.

Dummy Variables and Scaling of Data:

- Dummy variables were created to make easy and have quantitative analysis of all categorical variables in the model.

Split into Train and Test Data:

- The Final data after cleaning, scaling, dummy variable creation, the whole data is divided into train and test data in the ratio of 70:30 respectively.

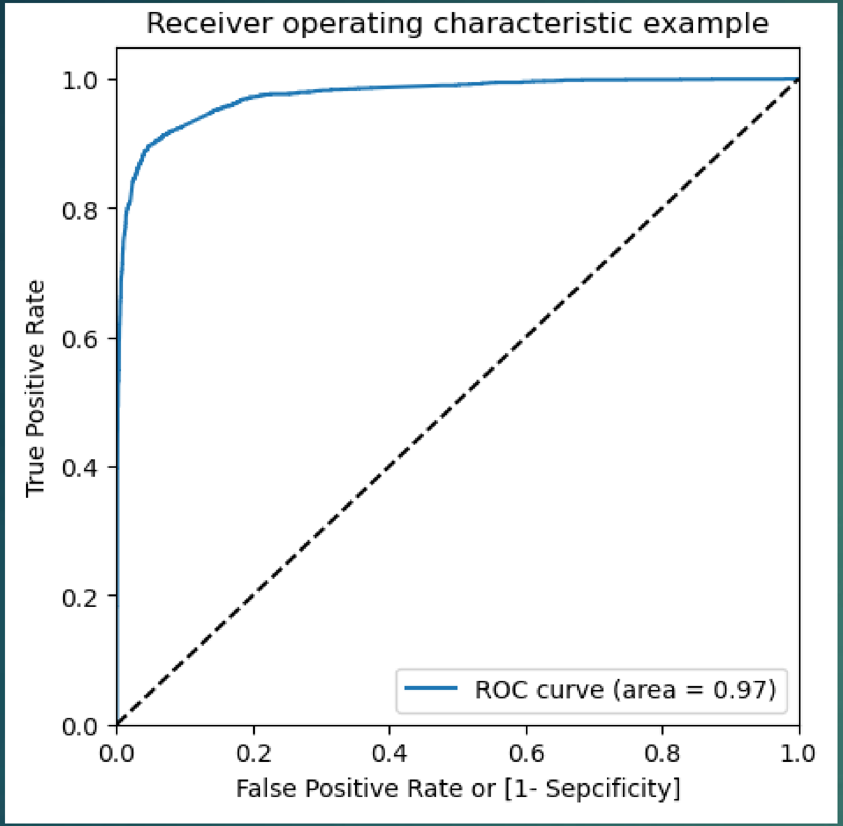
Scaling of Data:

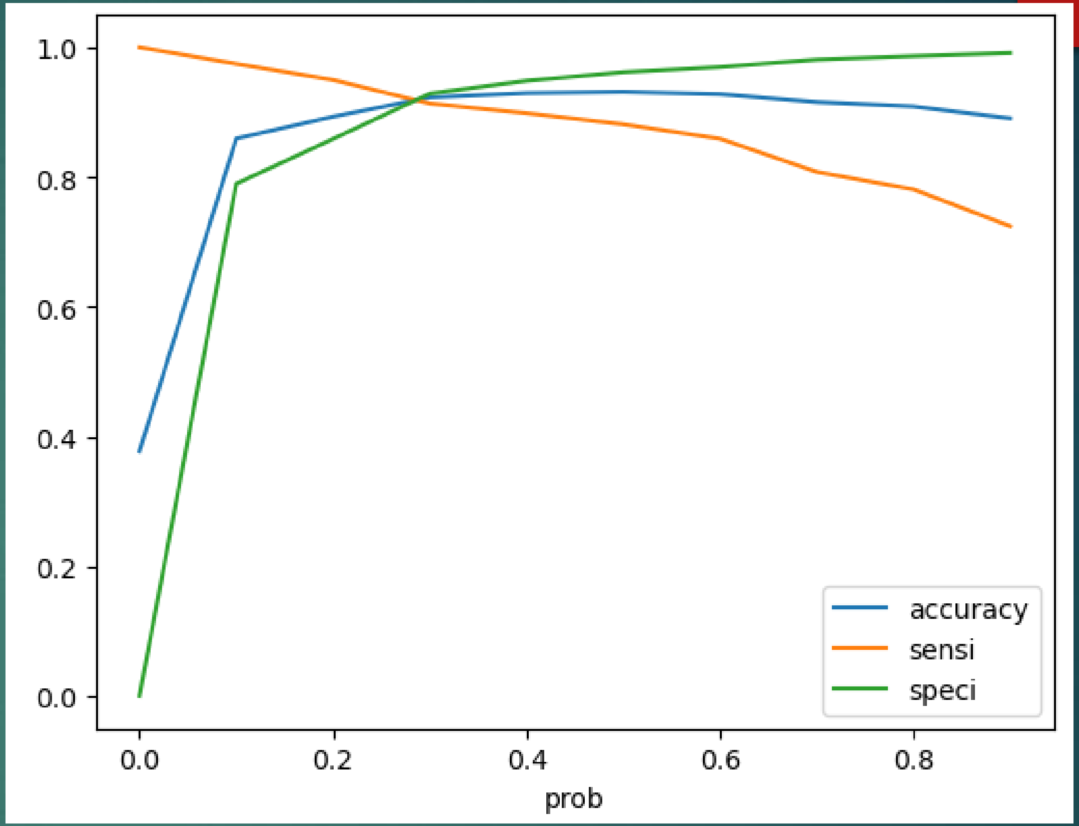
- Standard Scaling Fit transform is done on train data set.
- Standard Scaling transform is done on test data set.

Model Building:

- As Heatmap of whole data is very huge and will not get any observation.
- Hence recursive feature elimination (RFE) is done to reduce the data variables. It is reduced from 58 to 20.
- Logistic Binomial Regression is used to build the model as target variable converted is of only two (1/0)
- Based on p-value > 0.05 and VIF > 5, some variable were removed.

ROC Curve





Our Trained model is seems to be good as our ROC curve's value is 0.97 (It means it had covered almost 97% of graph area). I think this is very good.

From the curve above, 0.30 is the optimum point to take it as a cutoff probability

Train Model:

- Accuracy : 92.27%
- Specificity: 92.85%
- Sensitivity: 91.33%

Test Model:

- Accuracy : 91.23%
- Specificity: 91.63%
- Sensitivity: 90.59%

Model seems to be pretty fine.

Below are the variables which contributed the most are (in decreasing order)

- ➢ Tags_Closed by Horizzon
- ➢ Tags_Lost to EINS
- ➢ Lead Source_Welingak Website
- ➢ Tags_Will revert after reading the email
- ➢ Tags_Already a student

Below are the variables focused on more:

- ➢ Specialization – Management
- ➢ Specialization – Business
- ➢ Working Professional
- ➢ Lead through
- ✓ Google
- ✓ Chat
- ✓ SMS Chat