LEAD SCORING CASE STUDY

LOGISTIC REGRESSION

SUBMITTED BY

DASARI INDIRA DEVI

PROBLEM STATEMENT

- X-Education sells the online courses to industry professionnals
- X- Education gets lot of leads,its leads conversation is very poor
- 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 conversation rate should go up as the sales team will now be focussing more on communicating with the potential leads rather than making calls to everyone.

Business Objective

- ✓X-Education wants to know the most promising leads.
- ✓To build a model so that we can identify the Hot Leads
- ✓Deployment of the model for the future use

PROBLEM SOLVING METHOLOGY

Data Sourcing ,Cleaning, Preparation

- Source the data for analysis
- Clean and prepare the data(proper imputation methods for missing values)
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization

Feature Scaling

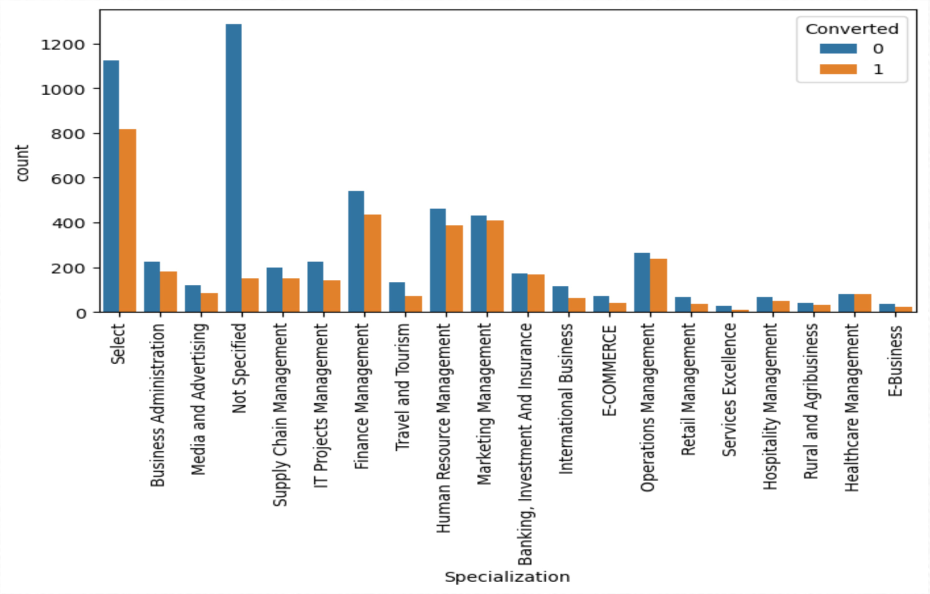
- Feature Scaling of Numerical data
- Splitting data into train and test dataset

Model Building

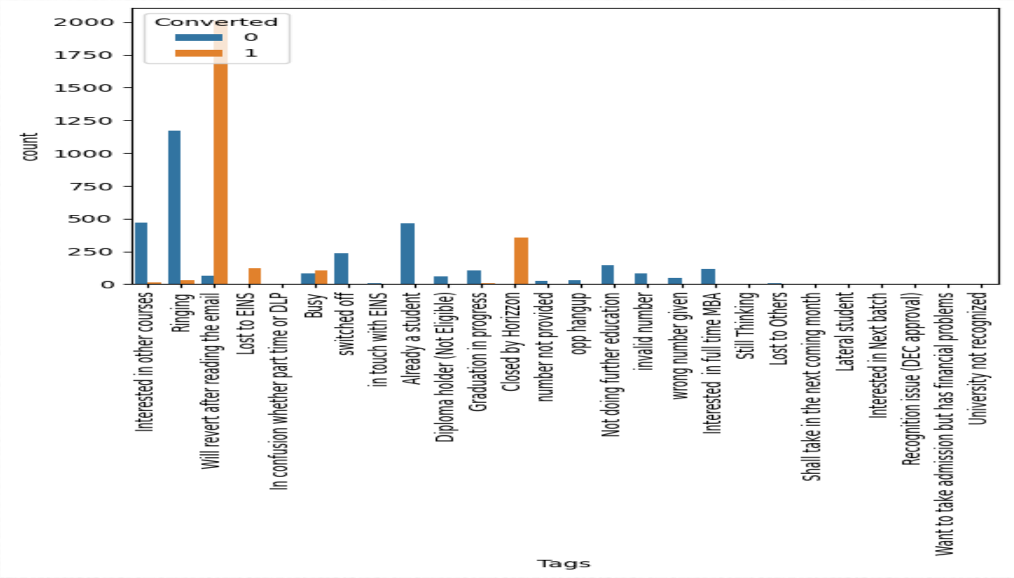
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy,sensitivity,specificity,precision and recall

Exploratory Data Analysis

Univariate (Specialization ,Converted)

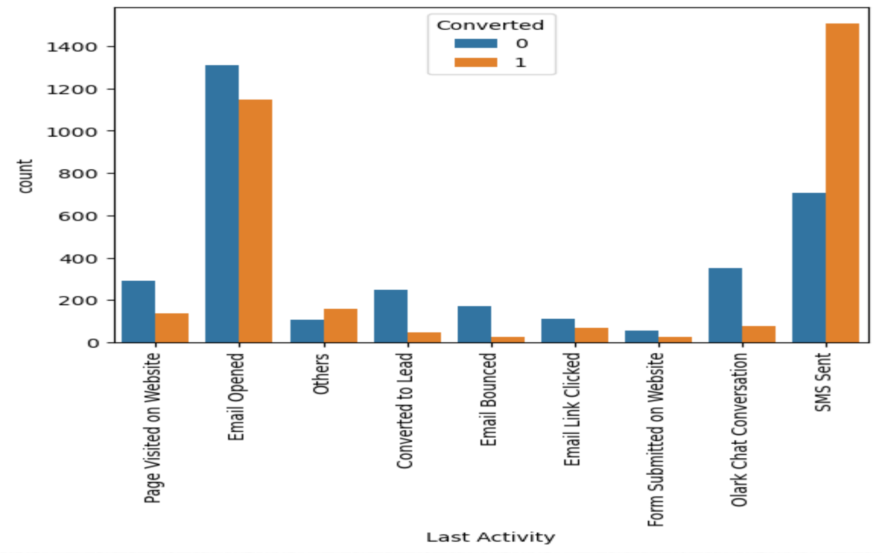


Univariate(Tags,converted)

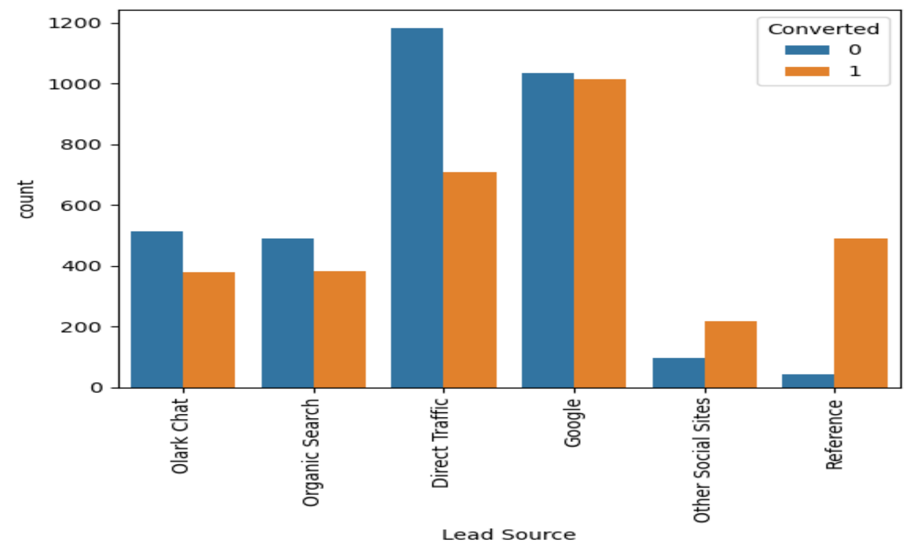


Potential Leads are Email sent,s ms sent, google, Direct Tarrif

Mostly Email and sms are converted

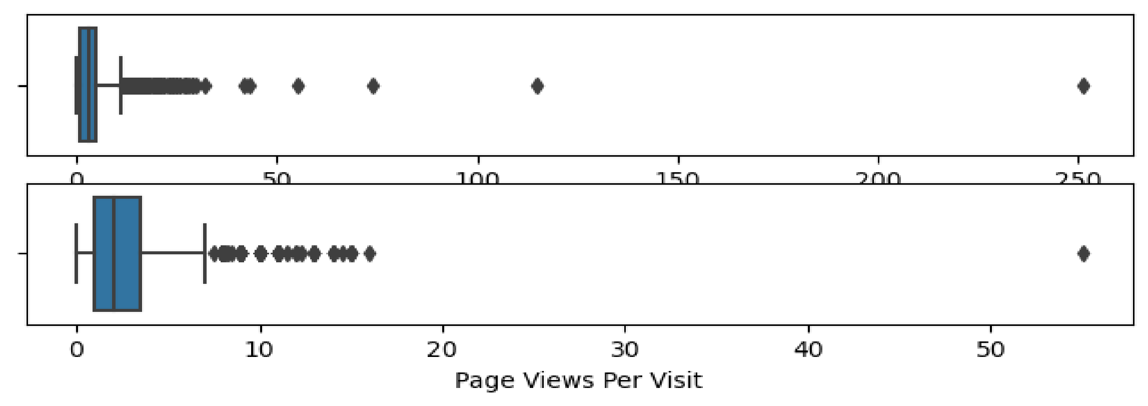


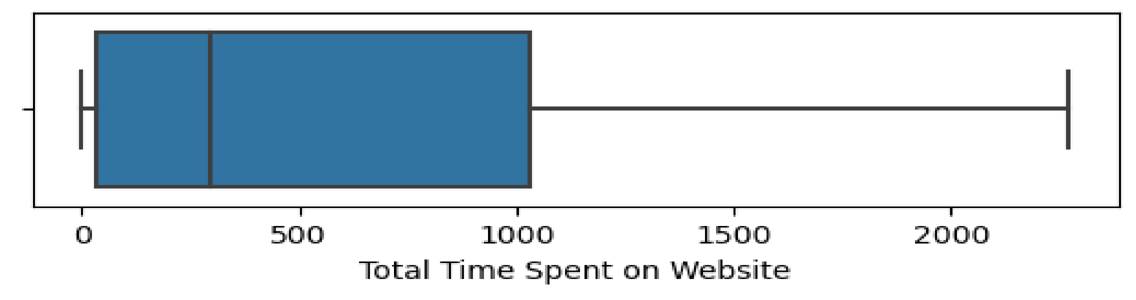
Mostly Google,Direct Tariff are converted

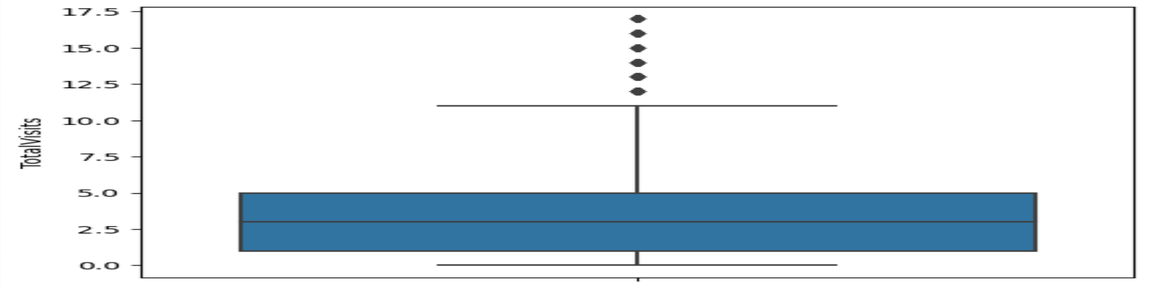


Outliers
(Box Plot)

- Outliers are identified using Box Plot for Page views per visit,Total Time spent in Website,Total Visits
- We have identified and detected using IQR range and mostly occurred after 99%
- Capping concept is used here for outliers and analysis proceeded with converted



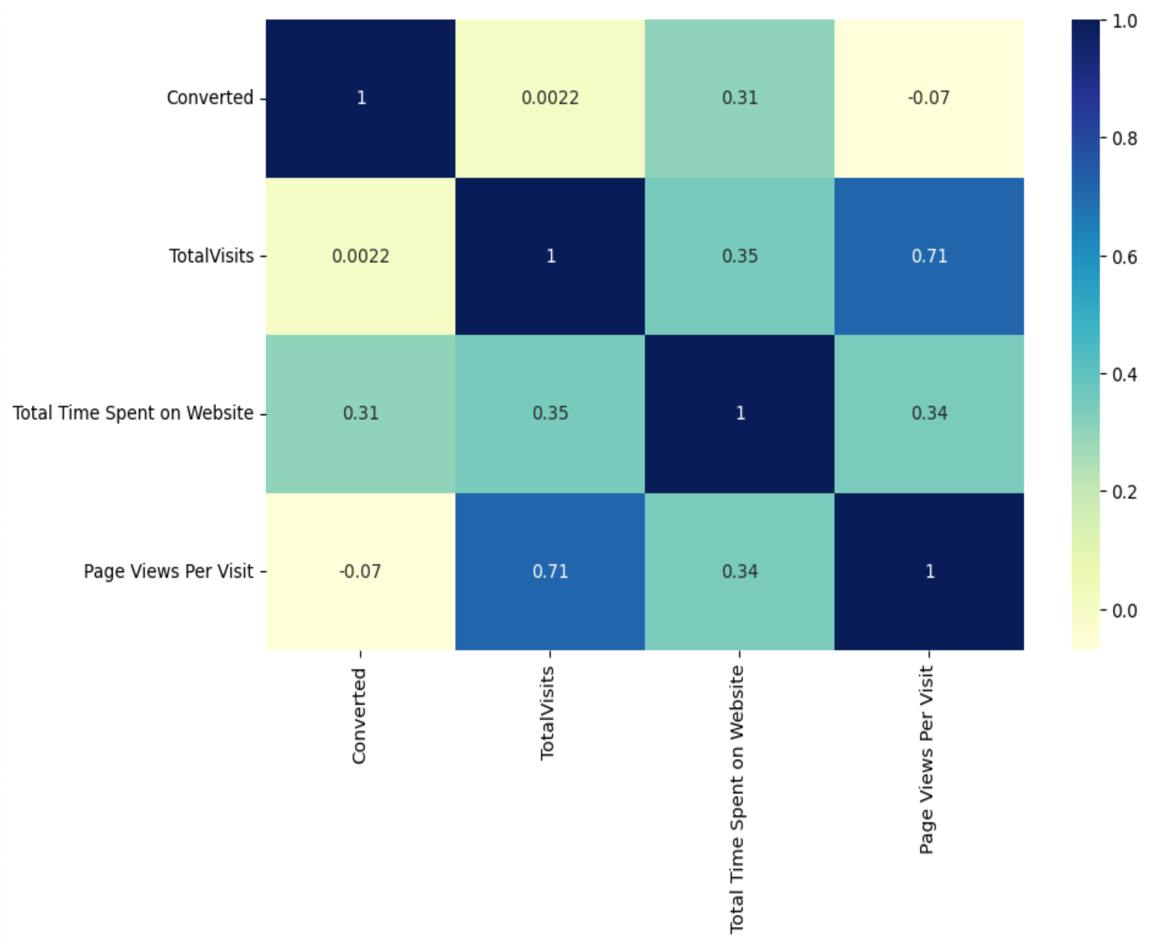




Heat Map

Total visits,Time Spent On Website,Pages viewed per visit

- From the heat map we can clearly predict that total visits and page views per visit
- Now we can drop one variable so that model will be accurate
- From the analysis mostly total time spent on website ,total visit will increase probability of buying a course



Data Conversion

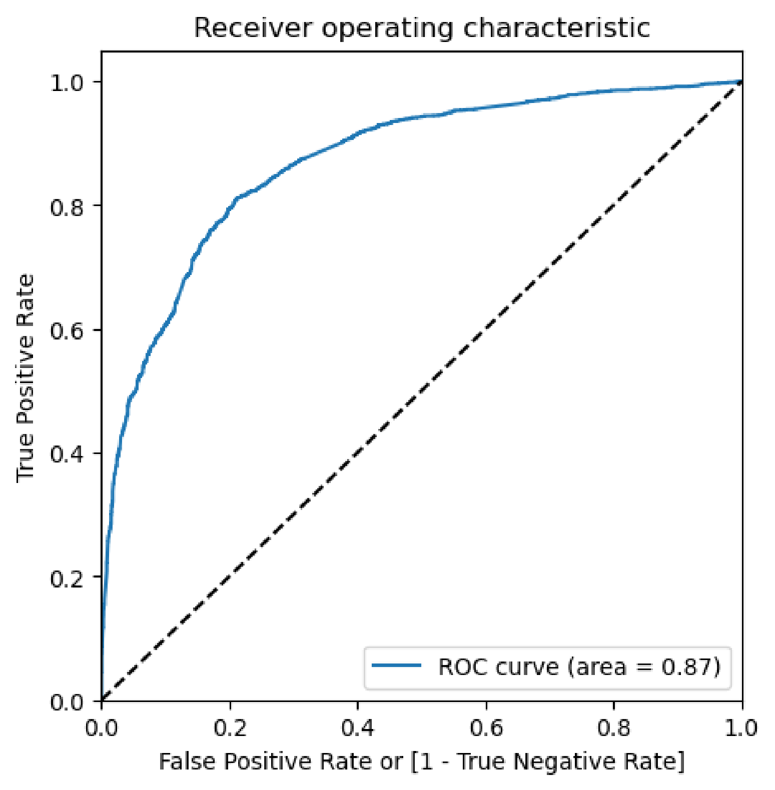
- Numerical values are normalised
- Dummy variables are created for object type variables

Model Building

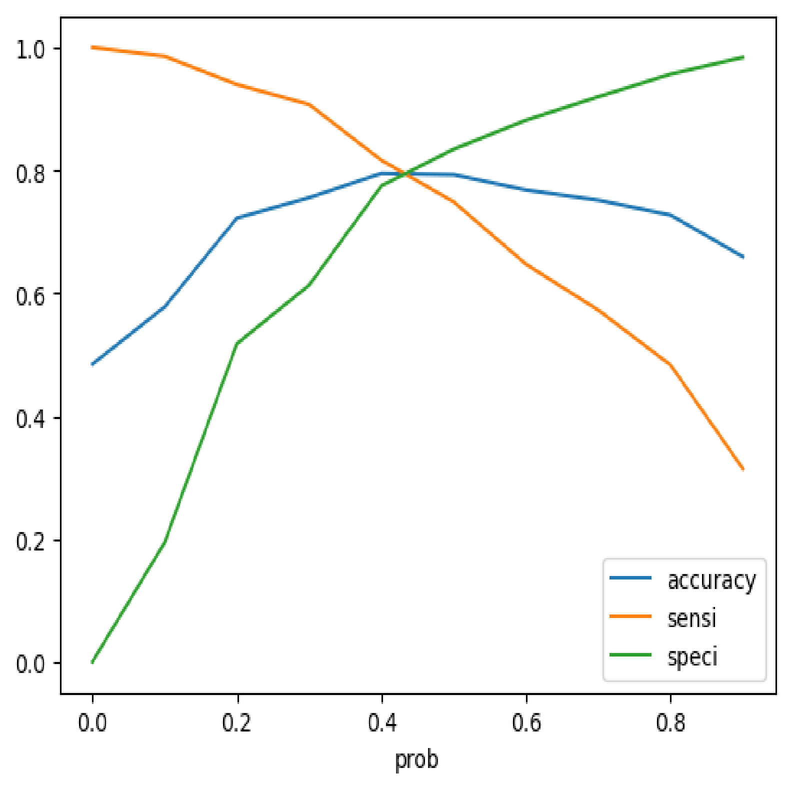
- Splitting the data into Training and Testing Data
- Train-Test split is performed as 70:30 ratio
- Using RFE for feature selection
- Running RFE with 15 variables as output
- Building model by removing the P-value greater than 0.05 and VIF greater than 5
- Predictions on Dataset

Roc Curve

ROC (True Positive rate and False Positive Rate



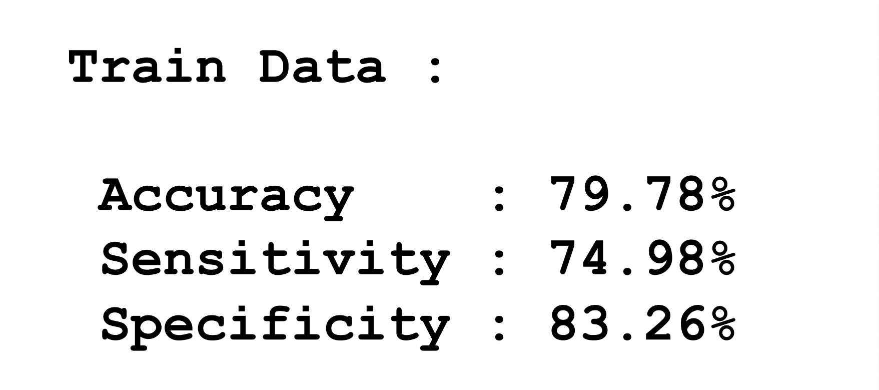
Accuracy,Sensitivity,Specificity



Finding the optimal cutoff point

Optimal cut off probability is that probability where we get balanced sensitivity and specificity

From the second ROC curve it is visible that optimal cut off is approximately at 0.3



```
Train Data :

 Accuracy     : 79.78%
 Sensitivity : 74.98%
 Specificity : 83.26%
```

```
Test Data:

Accuracy : 73.78%
Sensitivity : 89.98%
Specificity : 66.26%
```

Conclusion

Potential Buyers:

- Total time spent on wesite
- Total number of visits
- Last activity as sms sent
- Current Occupation

  Working Professionals

  Student

  Unemployed

Keeping the above factors we can increase the potential buyers as paying customers