LEADS SCORING

CASE STUDY

BY:
Anush Goel
Rohit Sahu
Udit Birthare

PROBLEM STATEMENT

- X Education offers online education for professionals in the sector.
- Despite receiving a high volume of leads, X Education's conversion rate is low. For example, if they generate 100 leads each day, only roughly 30 of them get converted.
- To improve efficiency, the organization aims to find 'high potential', or hot leads.
- If they are successful in identifying this group of leads, the lead rate of conversion should increase since the sales staff will be focused more on connecting with prospective prospects instead of making calls to all people.

BUSINESS OBJECTIVE

- X education aims to identify the most likely leads.
- The goal is to develop a model that can identify hot leads.
- Deploying the model for future usage.

Solution Methodology

Data cleaning and Data manipulation

- Check and handle NA or missing values.
- Drop columns that contain a substantial number of values that are absent and are useless for the analysis.
- Whenever required, the values will be imputed.
- Check and supervise outliers in data.
- Check and eliminate duplicate data.

EDA

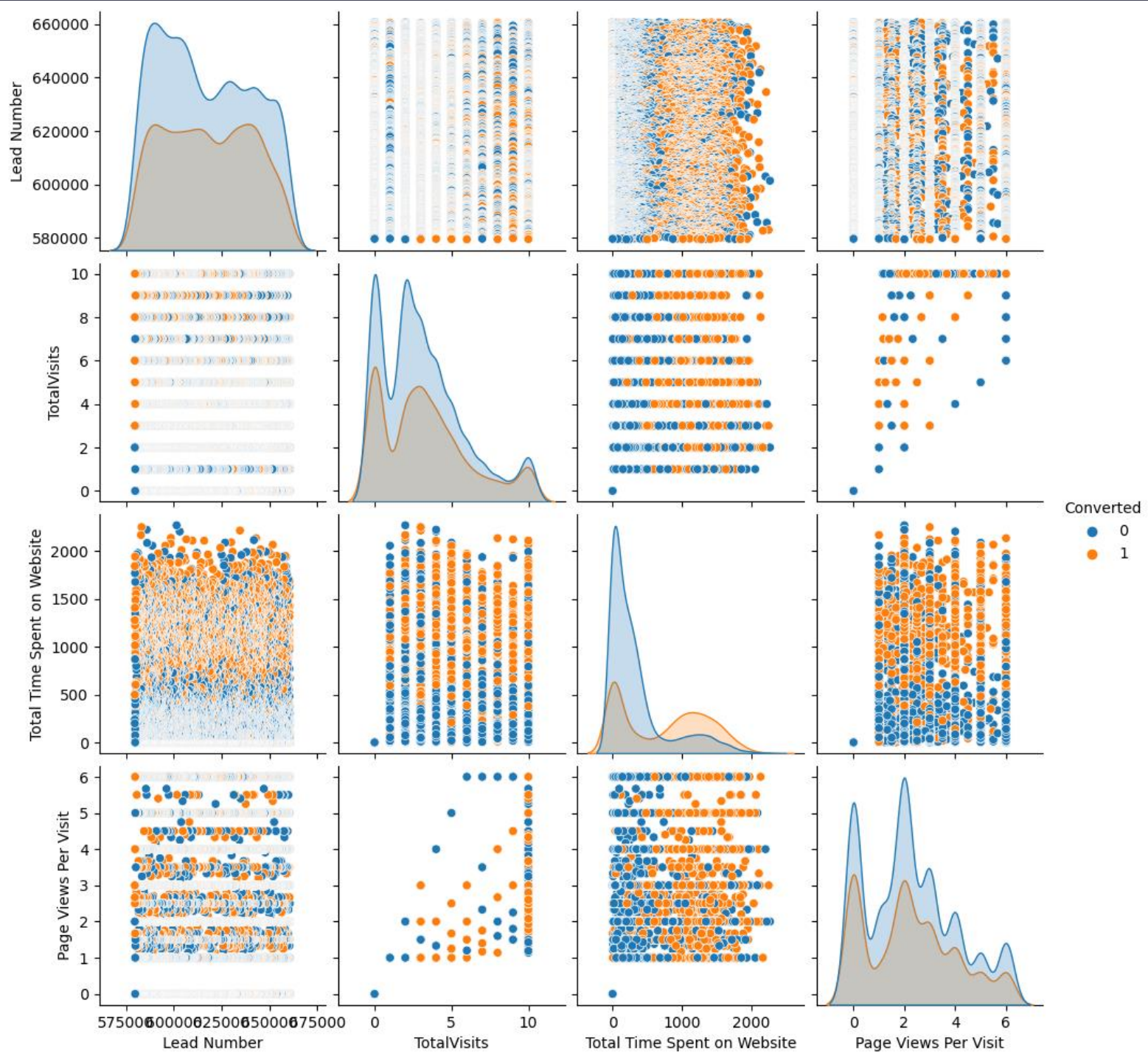
- Univariate data analysis: Plot of value count of variable .
- Bivariate data analysis: Analyze the relationship coefficient and patterns between parameters etc.
- Dummy Variables, Feature Scaling & Data Encoding.
- Classification technique: Logistic Regression used for the model making and prediction.
- Validation of the model: Confusion Matrix, Accuracy, Recall and Precisions
- Model presentation.
- Conclusions and recommendations.

DATA MANIPULATION

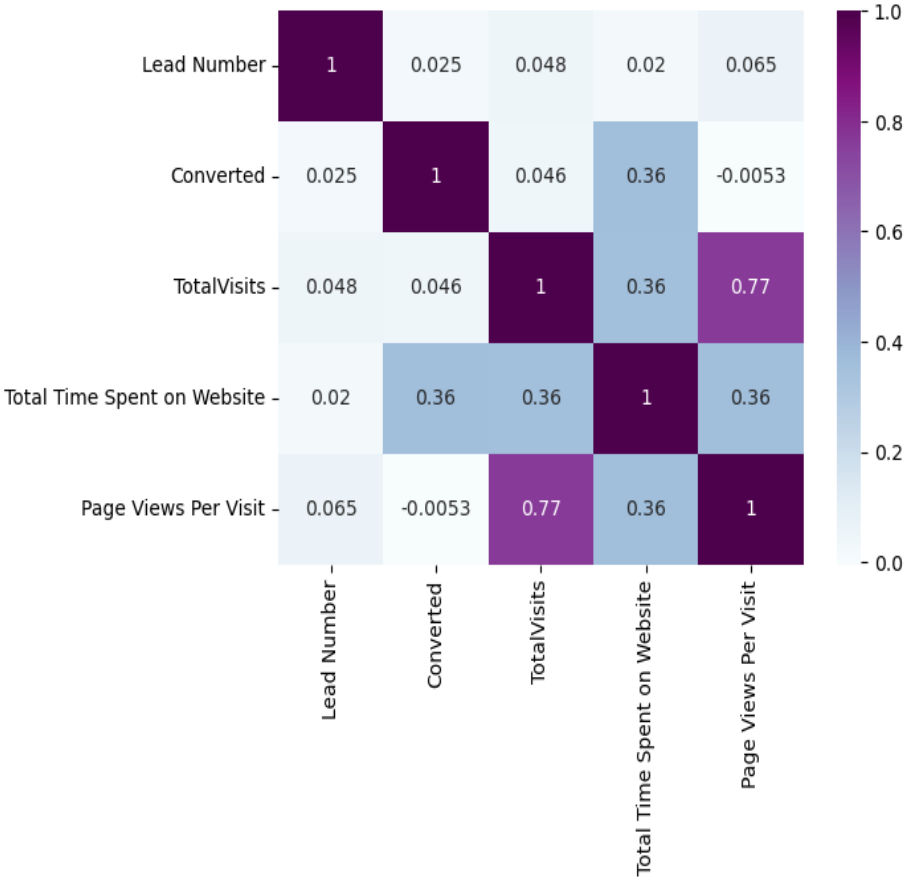
- Total count of rows : 37; Total count of columns : 9240.
- Single-value features include "Magazine," "Receive More Updates About Our Courses," and "Update My Supply“
- The words "chain content," "get updates on DM content," and "I agree to pay the amount through check" have been removed.
- Removed unnecessary fields "ProspectID" and "Lead Number" from the data analysis.
- The following features have been dropped due to high variance: "Do Not Call, "Search," "Newspaper, Article,“ "XEducation Forums," "Newspaper," & "DigitalAdvertisement.“
- Removed columns with over 35% missing values, including 'How did you learn about X Education' and 'Lead Profile'.

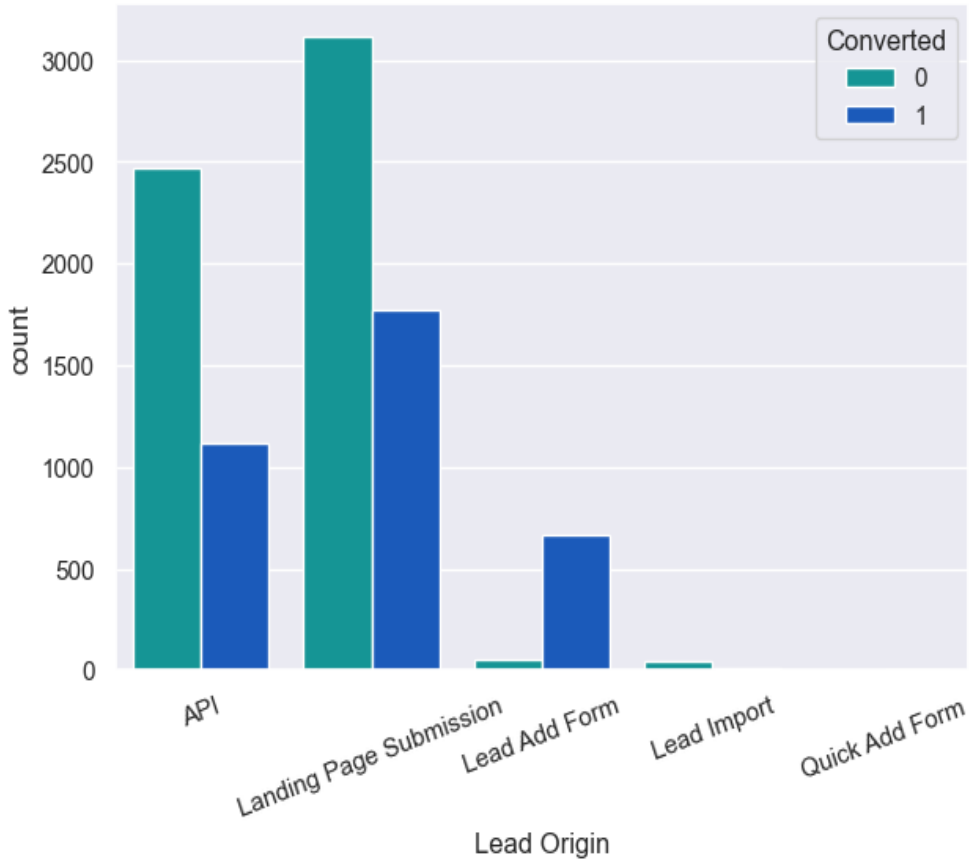
EDA

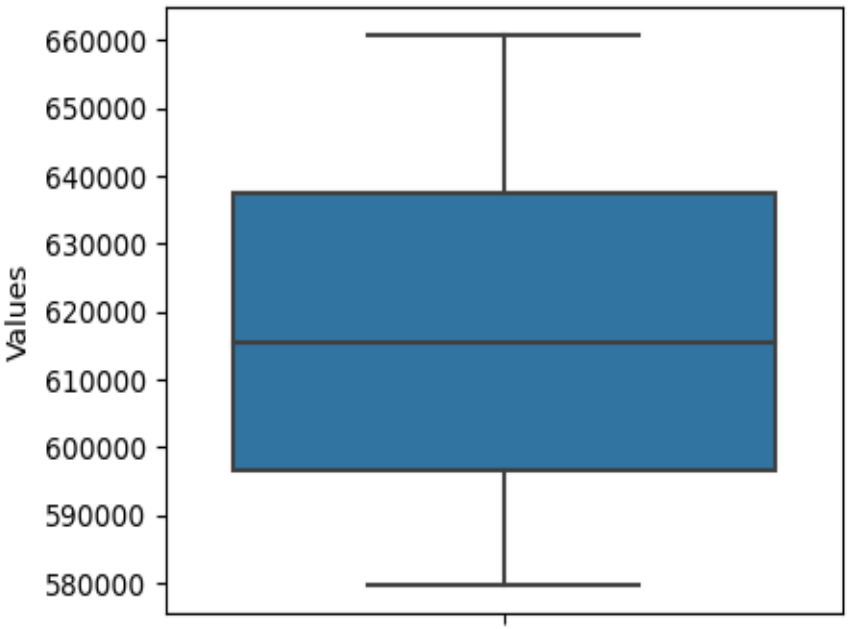
(EXPLORATORY DATA ANALYSIS)

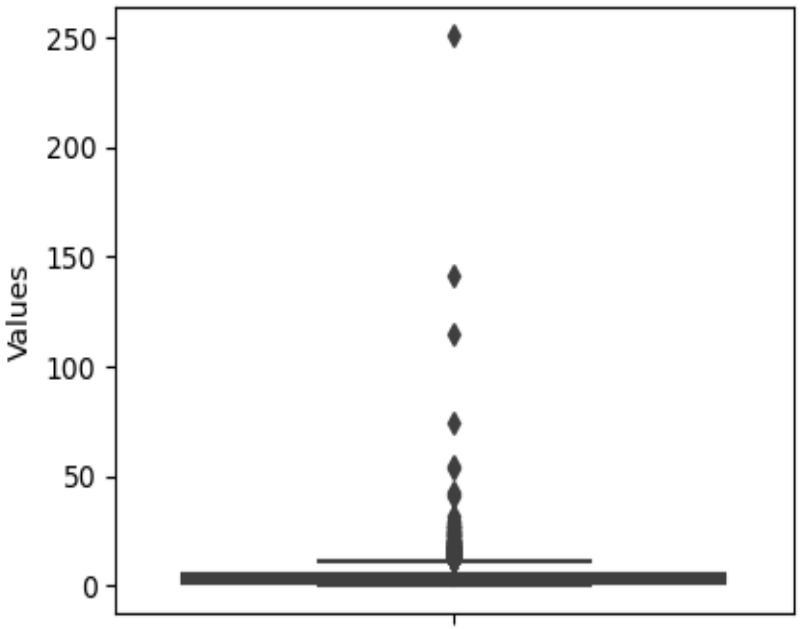


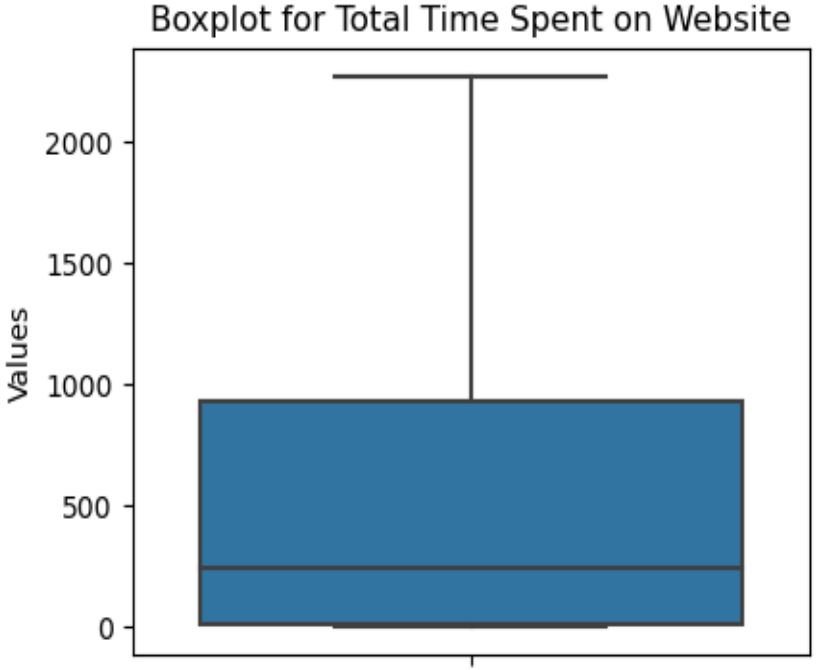
Heat Map & Bar Plot

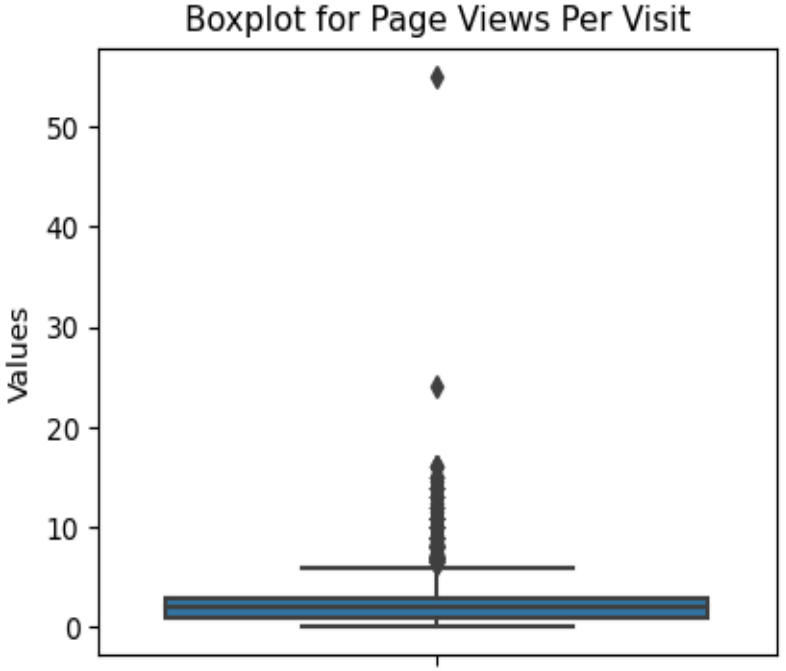





Boxplot for Lead Number
Values
660000
650000
640000
630000
620000
610000
600000
590000
580000


Boxplot for TotalVisits
Values
250
200
150
100
50
0

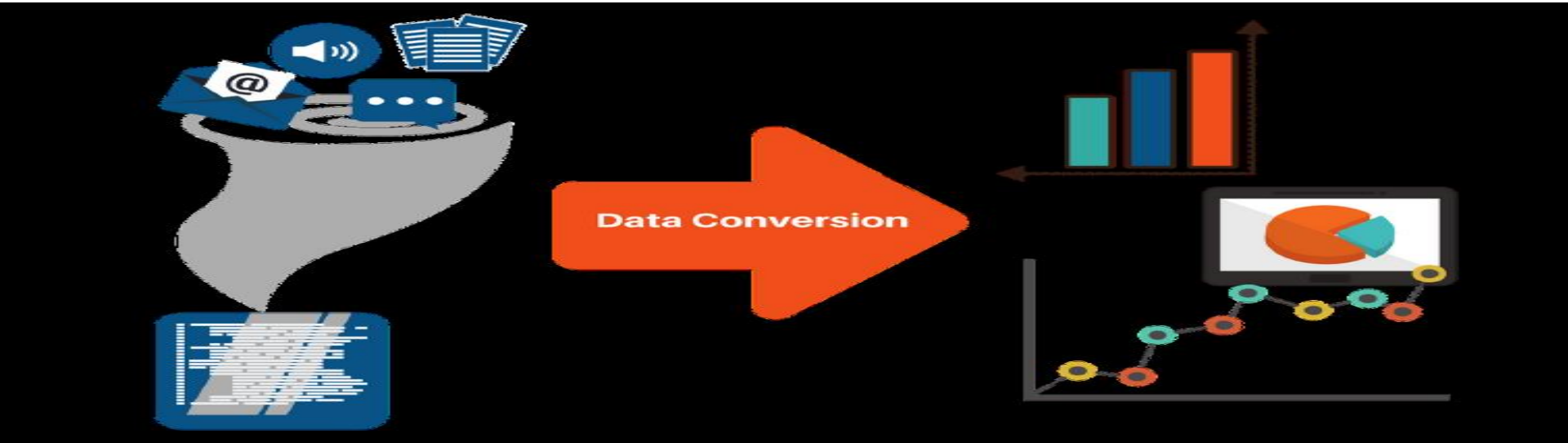

Boxplot for Total Time Spent on Website
Values
2000
1500
1000
500
0


Boxplot for Page Views Per Visit
Values
50
40
30
20
10
0

BOXPLOT

DATA CONVERSION

- Numerical Variables are Normalised
- Dummy Variables are created for Categorical variables
- Total Rows for Analysis: 9420
- Total Columns for Analysis: 57

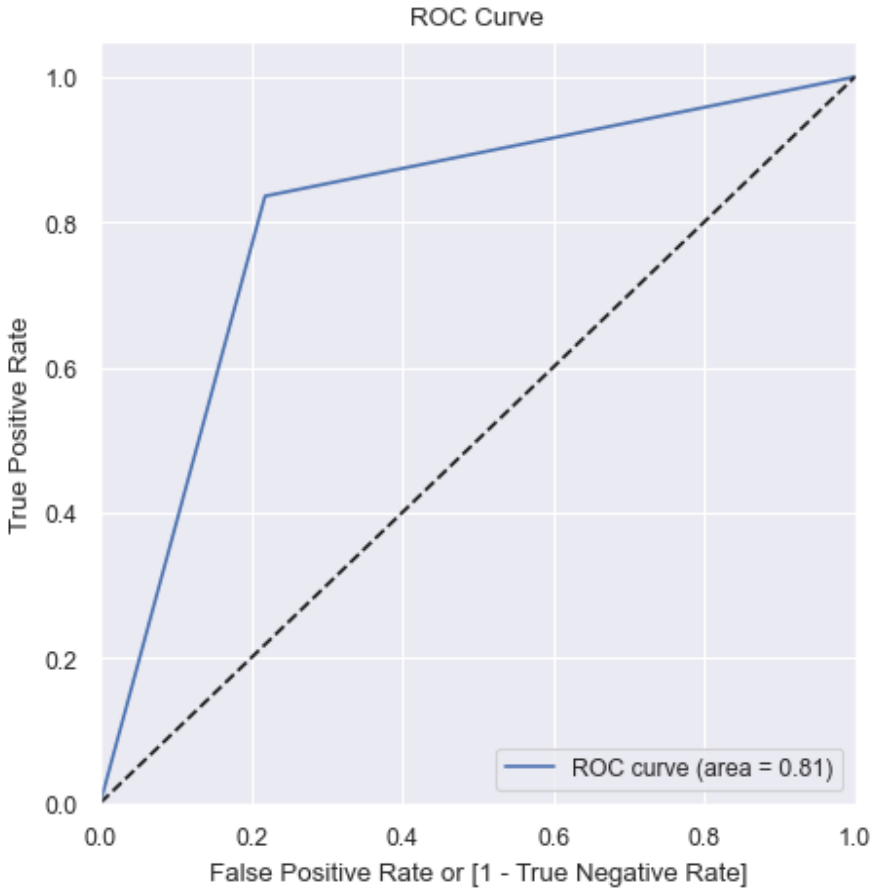


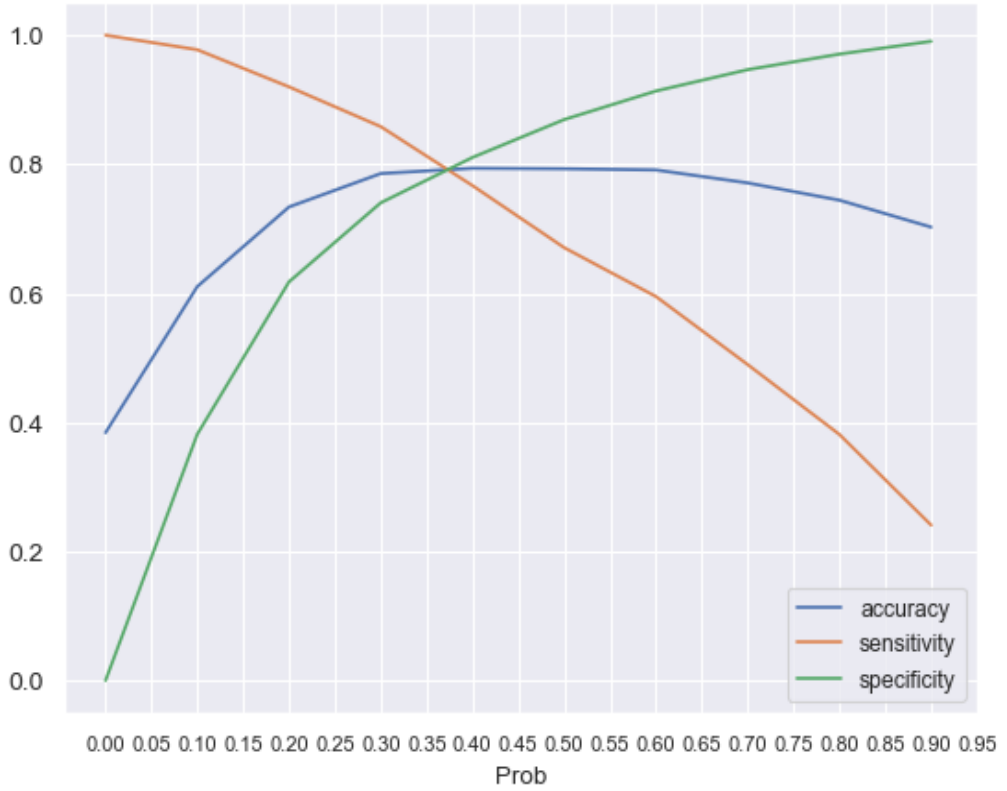
MODEL BUILDING

- Basic Splitting into X and Y
- Splitting the Data into Train-Test Split, in the ratio of 70:30
- Use RFE for Feature Selection
- Running RFE with 20 variables as output
- Building a model by eliminating variables with p-values larger than 0.05 and VIF values over five
- Predictions on test data set
- Overall accuracy 80 %

ROC Curve

Finding the Optimal Cut-off Point





- The optimal cut-off probability is
- Probability provides balanced sensitivity and specificity.
- The subsequent graph shows that the ideal cut off is 0.37.

CONCLUSION

It turned out that the criteria that contributed the most for possible customers were:

- The total time spend on the Website.
- Total number of visits.
- Last Activity: SMS Sent & Olark Chat
- Current Occupation: Working Professional

- With these factors in consideration, X Education has a great possibility of convincing customers to invest in the course they offer.