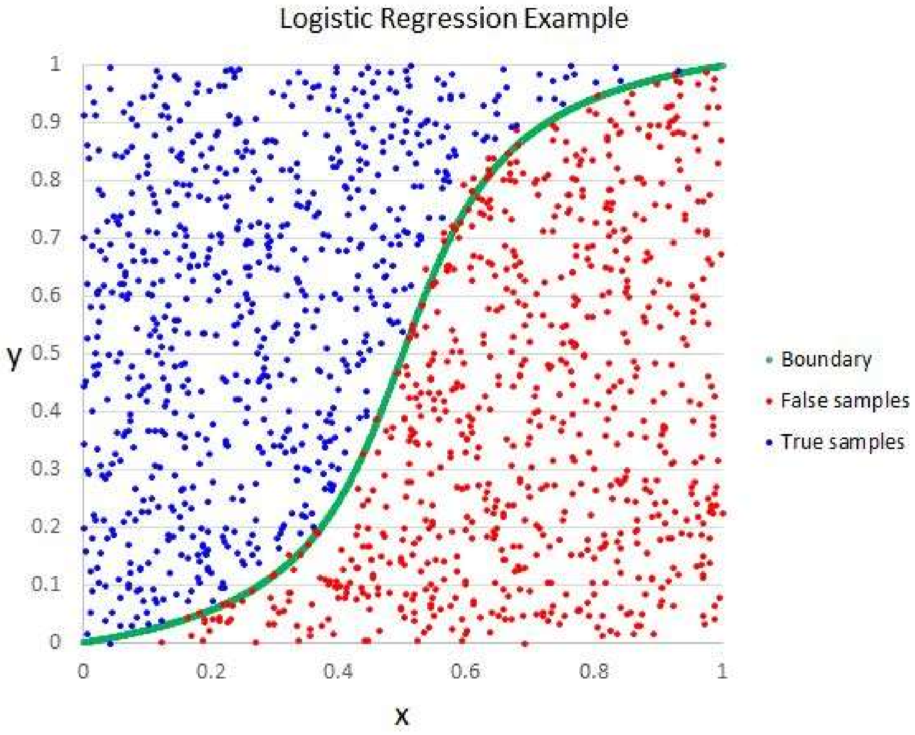
Logistic Regression –

Lead Scoring Assignment

- SNEHAL VIRWADEKAR (DSC 43)
- DEEPTHY T. BABU

Assignment- Technical & Business Analysis



I. BUSINESS PERSPECTIVE

Problem Statement

- X Education gets leads from various sources like Referrals , interested professional who visited the website etc.
- Although X Education gets a lot of leads, its lead conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- We need to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Data

- Leads.csv
  - Features – 37
  - Data Points : 9240
  - Target Column : ‘Converted’
    - 5679 – Label 0
    - 3561 – Label 1
  - 30 Features – Categorical
  - 7 Features - Numerical

Strategy

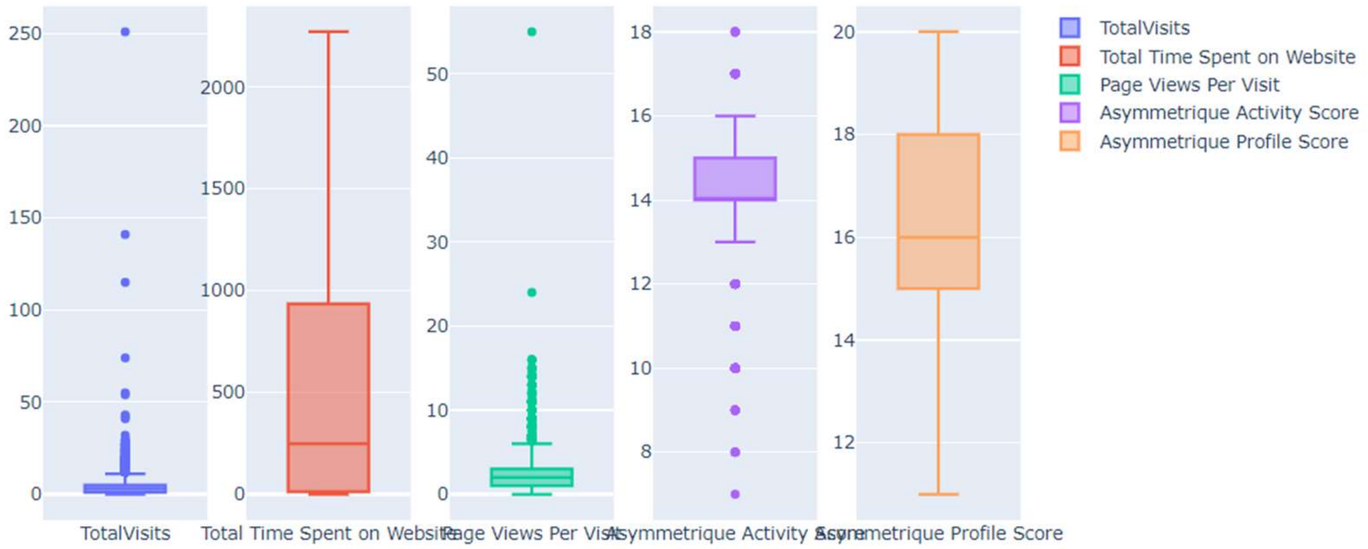
- Missing Data Handling –
  - Dropping Features with >45% Missing Data
  - GridSearchcv to find best Imputation strategy
- Outliers Trimming using IQR : Total Visits , Pageviews Per Visits
- EDA –using Pairplot, Boxplot
- Removing Constant and Quasi Constant Features : eg Magazine, Receive More updates about Course…etc
- Dealing with High Cardinal Features by Clubbing Rare Categories : eg Prospect ID, Leadnumber…etc
- Feature Scaling using Standard Scaler
- RFE For Feature Elimination.
- VIF – Removing Multicolinearity
- Trained model with reduced dataset
- Model Evaluation using F1 Score, Accuracy, Recall and Precision

Evaluation Metrices

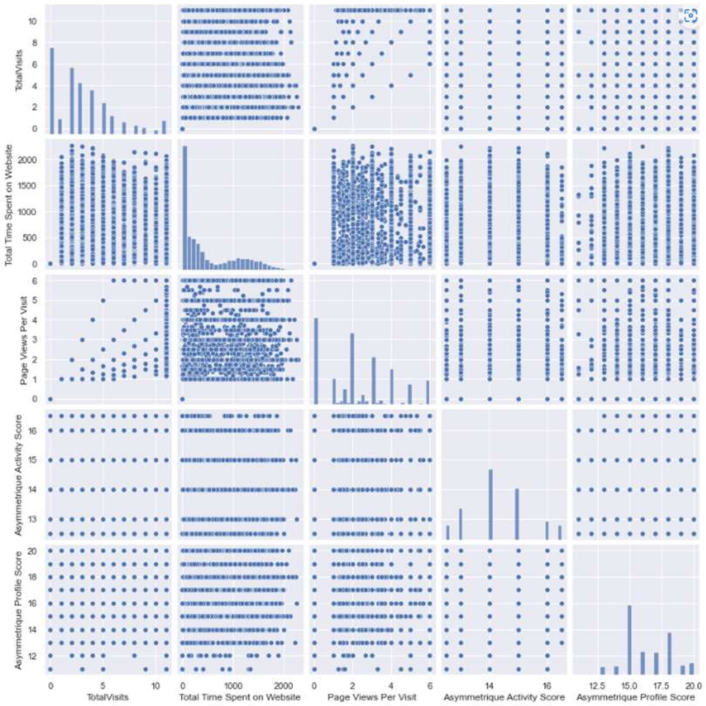
- Accuracy ~ 89.90%
- Precision ~ 89.55%
- Recall ~ 83%

Visualizations :

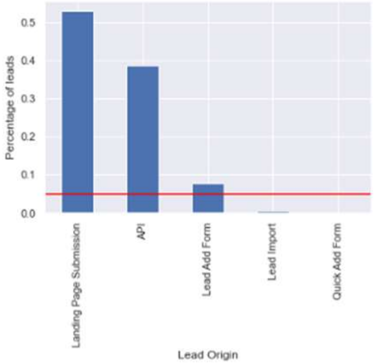
Outlier Analysis :

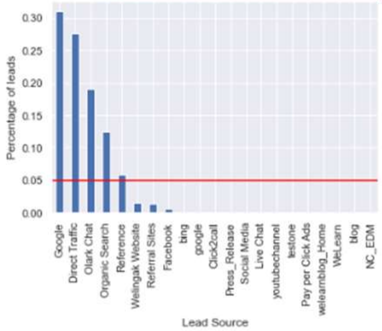


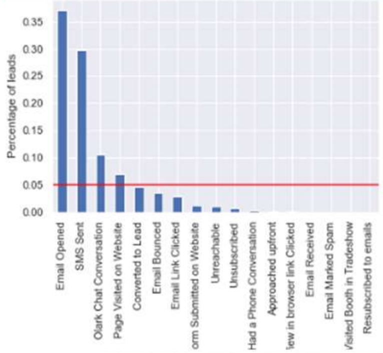
Pairplot :

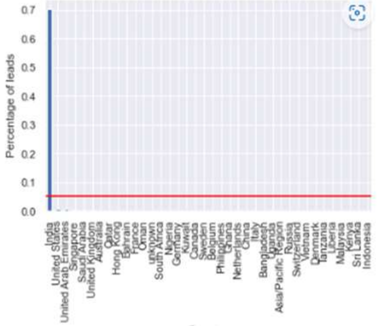

TotalVisits
Total Time Spent on Website
Page Views Per Visit
Asymmetrique Activity Score
Asymmetrique Profile Score
0
2
4
6
8
10
0
500
1000
1500
2000
0
1
2
3
4
5
6
13
14
15
16
12
14
16
18
20
0
5
10
0
1000
2000
0
2
4
6
14
16
12.5
15.0
17.5
20.0

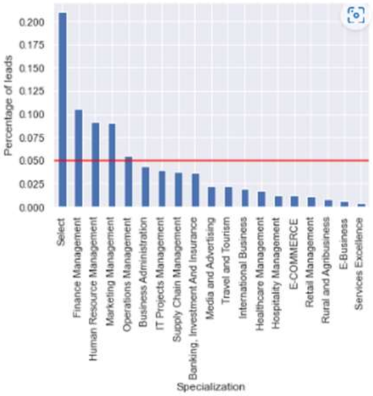
Features with Rare Categories :

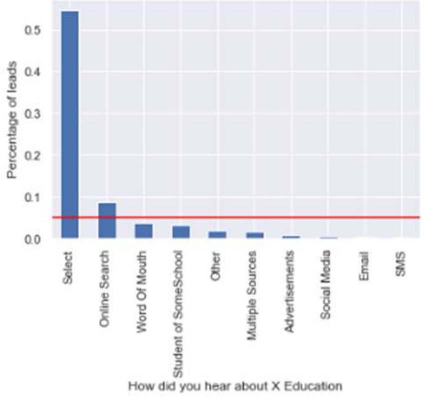


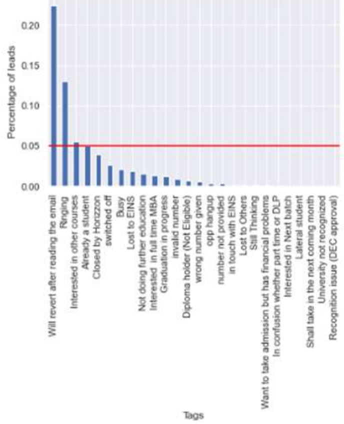


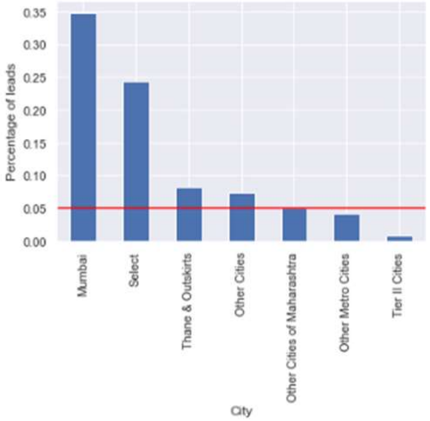






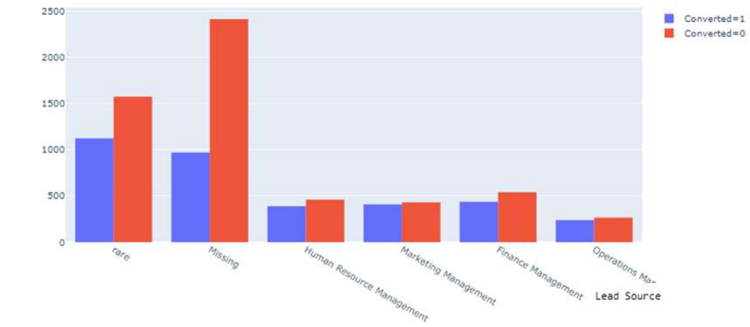




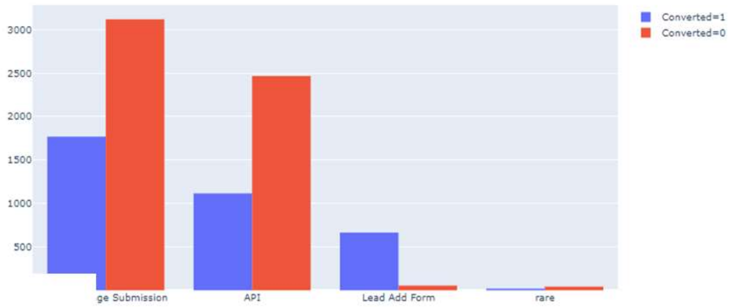


EDA : Categorical vs Target Variable(Converted)

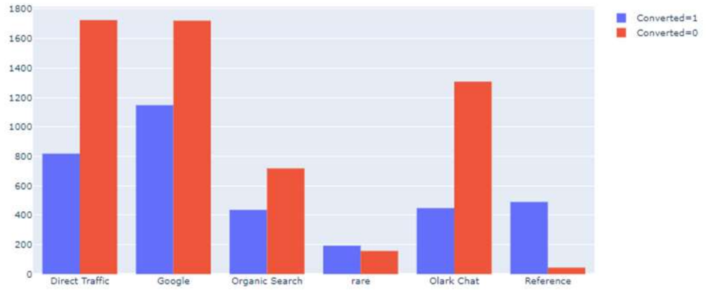
Specialization



Lead Origin



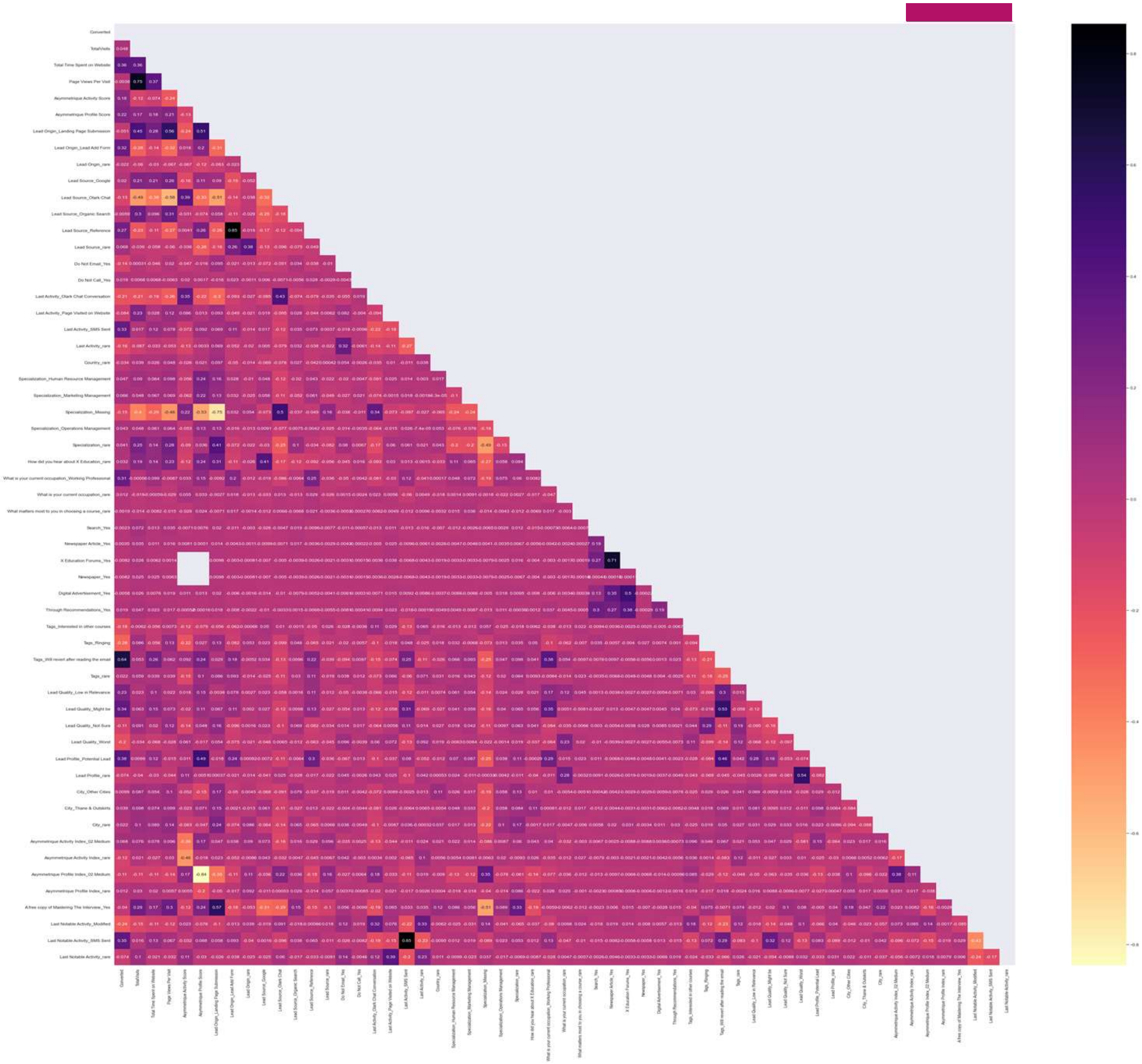
Lead Source



Observations From above plots

- Higher no of customers converted who were unemployed and Working
- Less customers converted when Last notable activity is Email opened and modified
- Less customers converted when Asymmetrique Profile Index = Medium
- Highest People Converted when Tags = Will Revert After Reading
- Highest People Converted when Lead Quality = Mightbe
- More People Converted when Lead Profile = Potential Lead
- More number Converted were from MUMbai CIty
- More number Converted were from GOOGLE and Direct Traffic

Correlations Heat map



Final Dataset after Preprocessing:

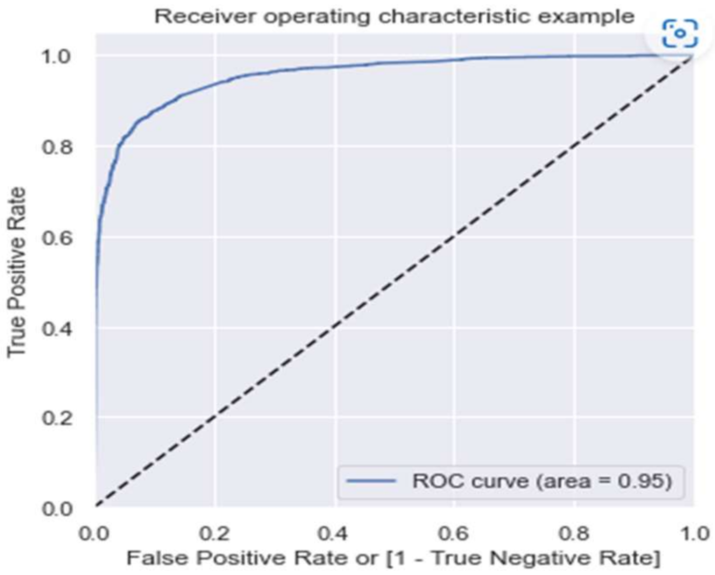
- Total Columns : 46
- Total Data Points (Rows): 9240

MODEL BUILDING

Model Building Strategy

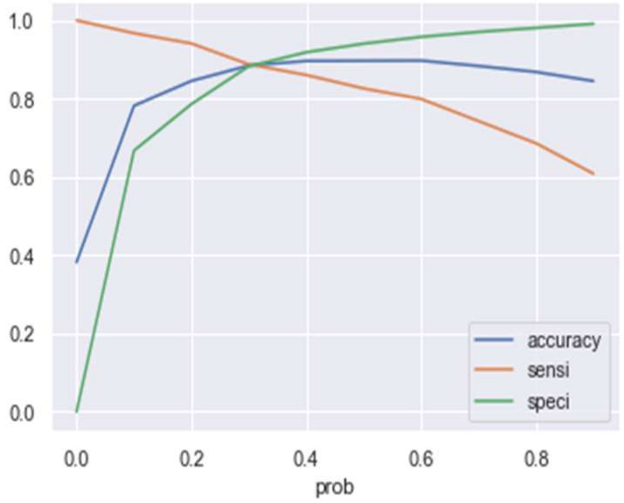
- Splitting into Train Test Split – 70:30 ratio of split
- RFE For Feature Selection : 15 variables as output
- Building model by selecting variable with p value < 0.05
- Applying Vif to remove multicolinearity
- Prediction on Test set :
- Evaluating using multiple Metrics
- Final Accuracy : 89.90%
- Recall : 83%
- Precision 89%

ROC Curve :


Receiver operating characteristic example
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.95)

Accuracy Sensitivity Specificity Curve

- Finding Optimal Cutoff point
- The sensitivity and specificity of a quantitative test are dependent on the cut-off value above or below which the test is positive.
- In general, the higher the sensitivity, the lower the specificity, and vice versa.



Important Features

- If Tags_Will revert after reading the email
- Lead Origin is from category Lead Add Form
- Tags – is Ringing
- Total Time Spent on Website is high
- Last Notable Activity is SMS Sent
- Lead Quality is not Worst
- Tags are Interested in other courses
- Lead Source is Olark Chat
- Lead Profile is Potential Lead
- Last Activity is Olark Chat Conversation
- Page Views Per Visit is more
- Do Not Email is Yes
- TotalVisits is more
- Lead Source is through Reference
- Asymmetrique Activity Indexis from rare

If X Education Focuses on above Features and Targets Customers based on these Behavior the Lead Conversion rate will increase and There will be many Hot Leads

Thank you