LEAD SCORING CASE STUDY

GROUP MEMBERS

- KHANTIL SHAH
- VYSAKH VENUGOPAL

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals
- Some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- The company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- 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 higher lead score have a higher conversion chance and the customers with 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%.

Lead Conversion Process - Demonstrated as a funnel

GOALS OF THE CASE STUDY

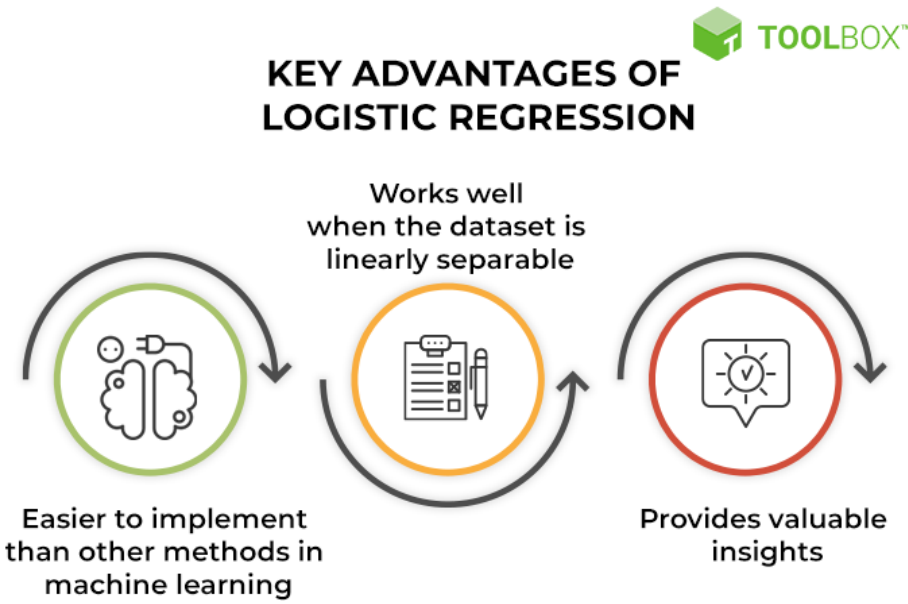
- There are quite a few goals for this case study.

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

- Business Objective
  - The company wants to know most promising leads
  - Identify hot leads and build a model around it
  - Model to be deployed for future use

APPROACH/METHODOLOGY

- Data understanding
- Data cleaning and manipulation
  - Check null and missing values and handle it
  - Identify duplicate data and deal with it
  - Identify large amount of missing values and not useful for analysis and drop columns
  - Imputing values if necessary
  - Handling outliers
- Exploratory data Analysis
  - Univariate data analysis – value counts etc.
  - Bivariate data analysis – correlation coefficients and patterns between variables
- Dummy Variables and feature scaling (encoding data)
- Classification method – Logistic regression used for model preparation and prediction
- Validating the model
- Model
- Concluding remarks along with observations



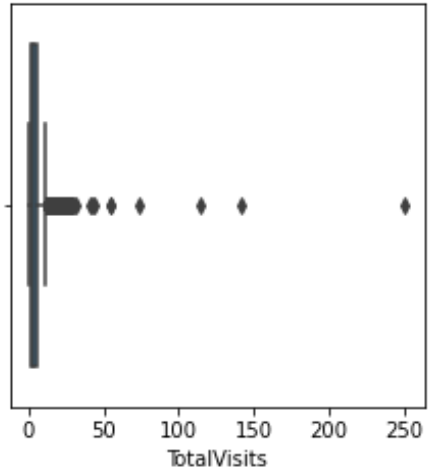
DATA UNDERSTANDING, CLEANING AND MANIPULATION

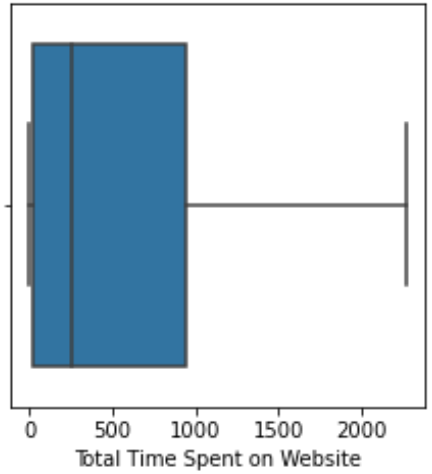
- Total No. of Columns = 9240, Total No. of Rows = 37
- Data types = Float 64, Int64 3, Object 30
- Missing/null values (removing all the columns with more than 40% missing values)

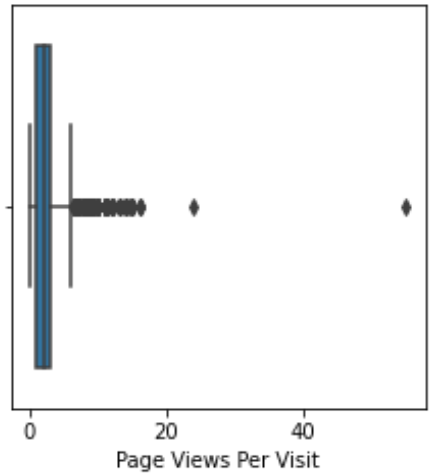
Lead Quality	51.59
Asymmetrique Activity Index	45.65
Asymmetrique Profile Score	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Index	45.65
Tags	36.29
Lead Profile	29.32
What matters most to you in choosing a course	29.32
What is your current occupation	29.11
Country	26.63
How did you hear about X Education	23.89
Specialization	15.56
City	15.37
Page Views Per Visit	1.48
TotalVisits	1.48
Last Activity	1.11
Lead Source	0.39

- Dropped columns with high missing values – Lead Quality, Asymmetric Activity Index, Asymmetric Profile Score, Asymmetric Activity Score, Asymmetric Profile Index, How did your hear about X Education, Lead profile, City, Specialization, Tags, What matters most to you in choosing a course, Country,
- Creating a new column - the case due to customer not specifying their occupation, we could create a new category 'Unspecified' for the missing values.
- Total Visits and Last Activity columns missing values will be imputed with median value

HANDLING OUTLIERS

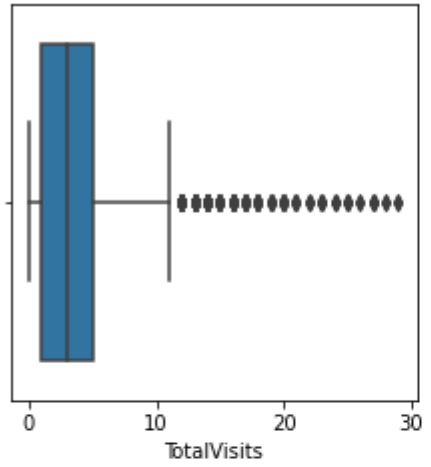


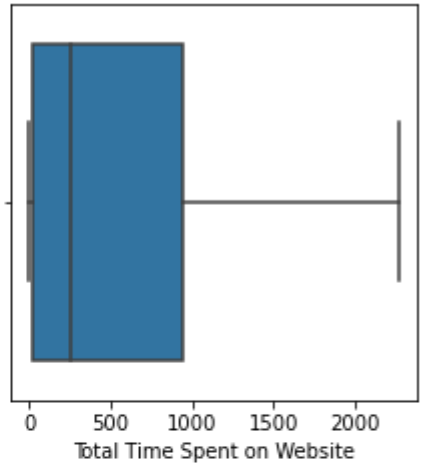


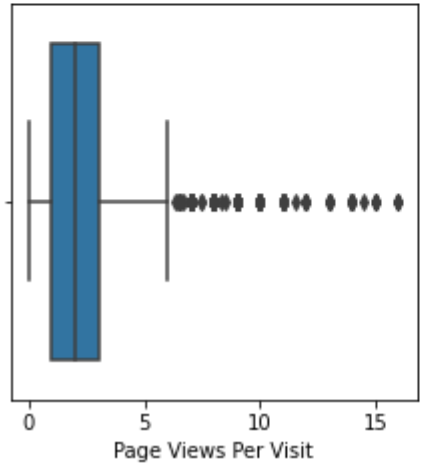


We can see that, columns Total Visits and Page Views Per Visit have some outliers at the higher side.

We will select the threshold as 30 and 20 respectively for the columns Total Visits and Page Views Per Visit for removing the outlier data points, as most of the values seems to be relevant.

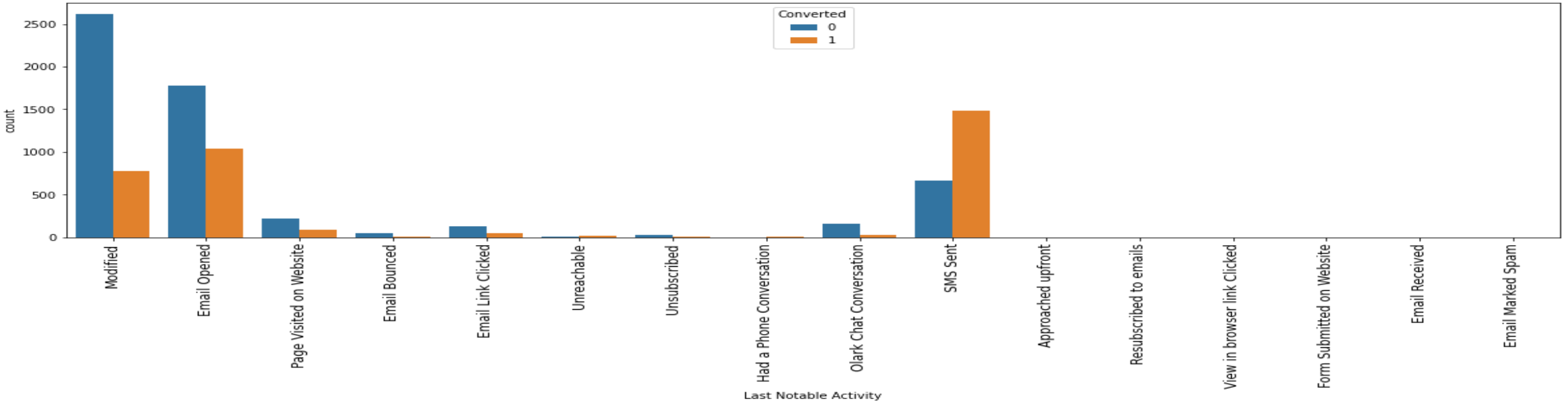






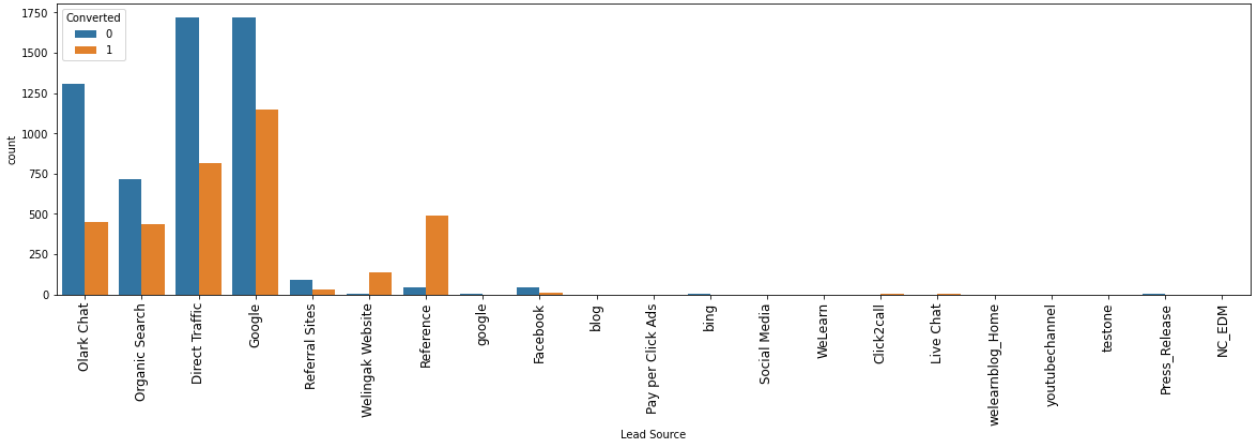
All the outliers have been handled here.

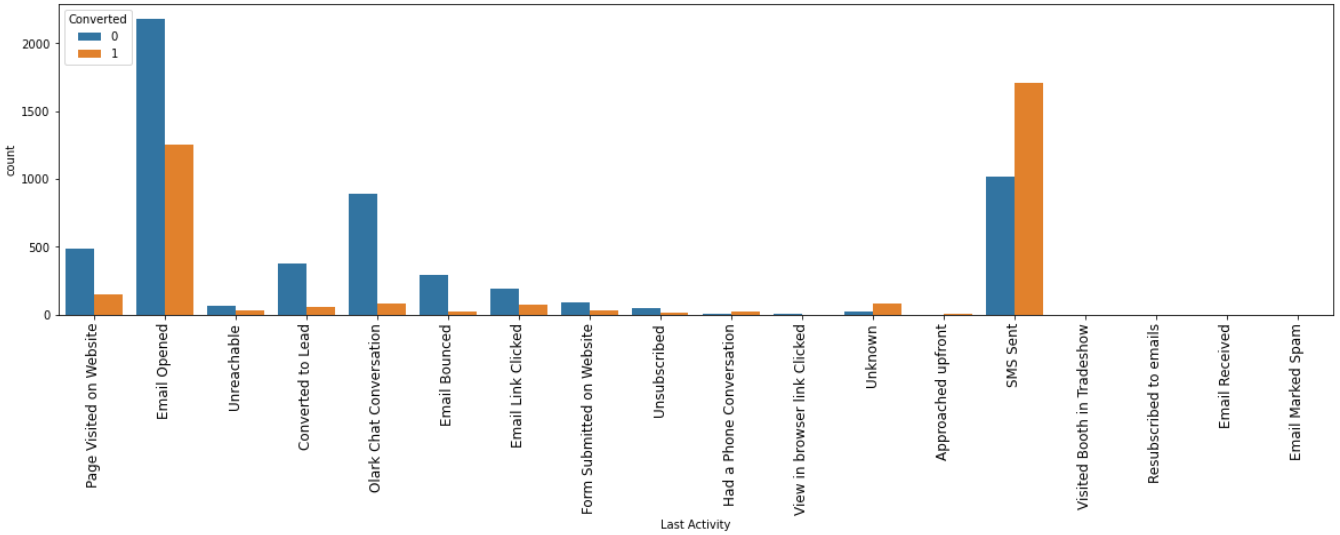
EXPLORATORY DATA ANALYSIS



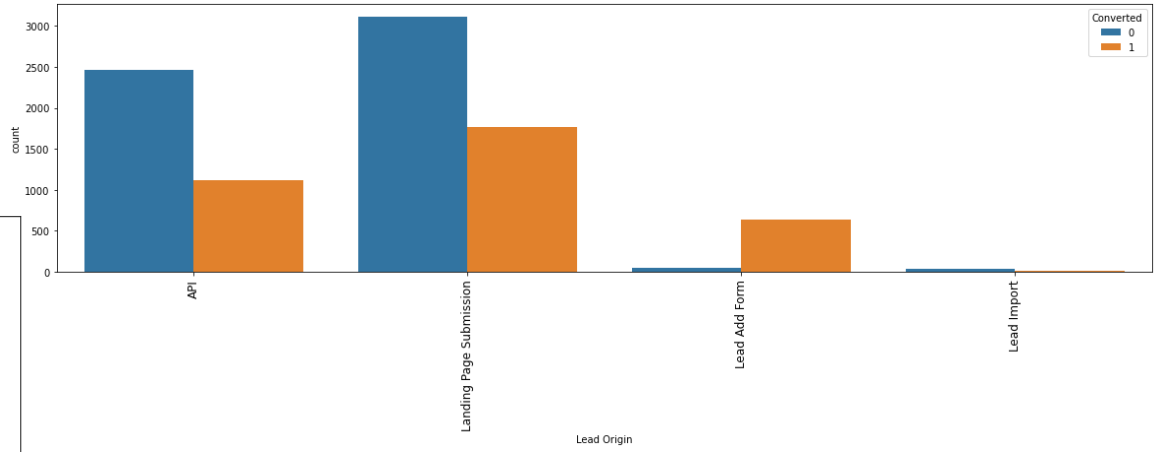
- Lead Origin Lead Add Form seems to provide more converted leads.
- Google seems to be the most popular Lead Source. However, Reference has more leads converted than not.
- Most of the leads have Last Activity as Email Opened. However, leads with Last Activity as SMS Sent have converted more often than not.
- Unemployed leads have been the most and have near same amount of converted and not converted leads. However Working Professionals have more leads converted than not.
- Similar to the Last Activity, SMS Sent has more leads converted than not.

EXPLORATORY DATA ANALYSIS





Categorical Variable Relation



EXPLORATORY DATA ANALYSIS

- We can see that there are certain columns which are highly imbalanced. Hence, we will drop these columns along with the identifier column Lead Number.
- We will keep the column Prospect ID for now and drop later after the train-test split.
- Dropped columns - 'Lead Number', 'Do Not Email', 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'
- Total No. of Columns = 9192, Total No. of Rows = 11

DUMMY VARIABLES

- Numerical Variables are normalised
- Dummy variables are created for object type variables
- Total No. of Rows = 66, Total No. of Columns = 9192 (for further analysis)

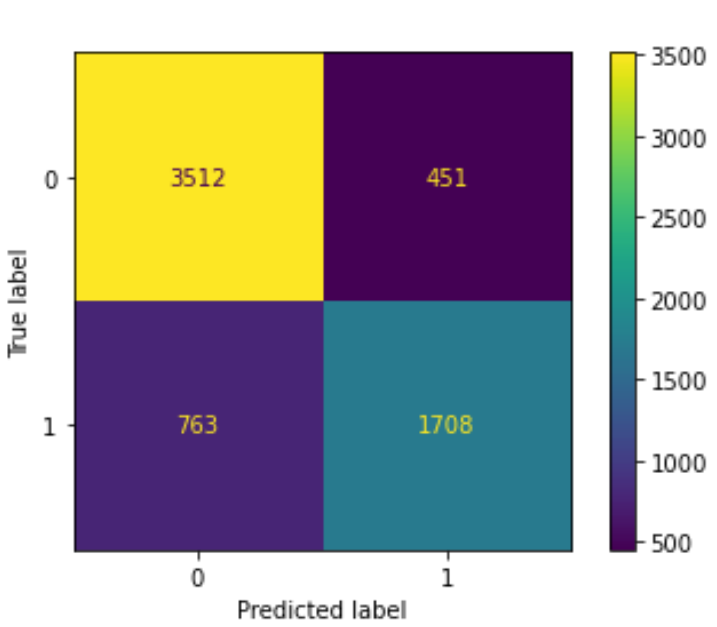
MODEL BUILDING

- Splitting the data into Train and Test sets
- We have chosen 70:30 ratio for Train and Test split
- RFE used for feature selection
- Feature Scaling - We have used the Standard Scaler technique for scaling the numerical features. This will centralize all the feature values around 0. We will use Standard Scaler utility from sklearn.preprocessing API.

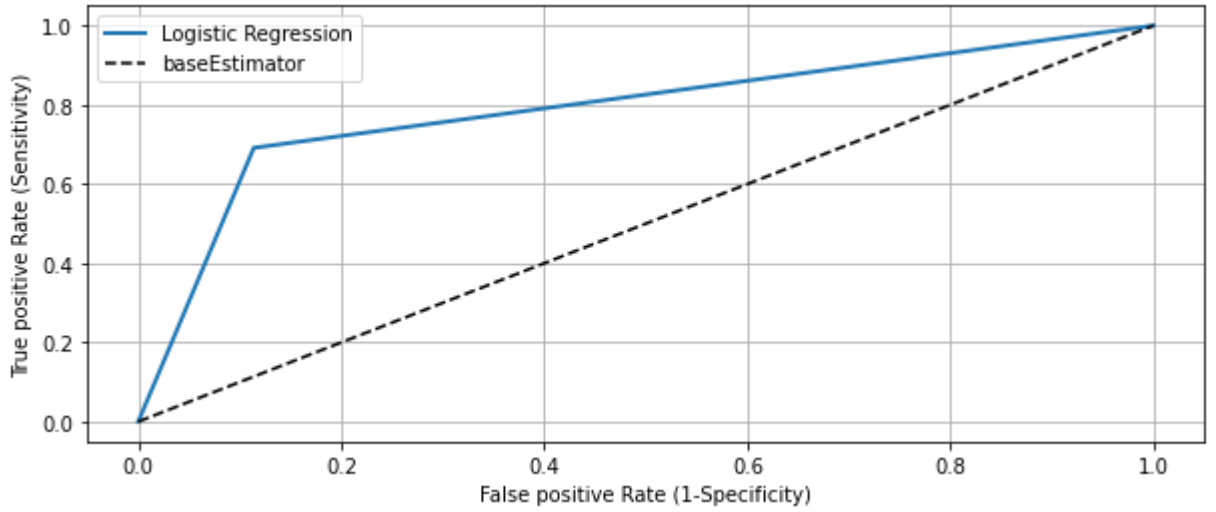
Build model using statsmodel for manual feature selection

- We will use the following two metrics for feature selection.
  - p-value < 0.5
  - VIF < 5
- For feature selection, we use the following algorithm:
  - If p-value > 0.5 and VIF > 5, drop the variable
  - Else if p-value > 0.5, drop the variable
  - Else if VIF > 5, drop the variable
  - Else Keep the model

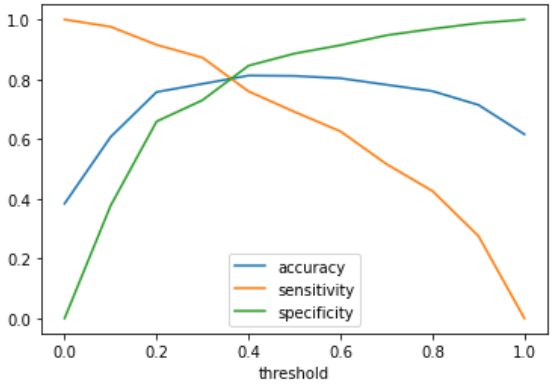
MODEL BUILDING



Confusion Matrix



ROC Curve



plot accuracy, sensitivity and specificity and the optimum threshold value.

Optimal cut off is at 0.33

CONCLUDING REMARKS AND OBSERVATION

- **For Test Set:**
  - **Accuracy:** 81.58%
  - **Sensitivity:** 80.36%
  - **Specificity:** 82.34%
  - **Precision:** 0.74
  - **Recall:** 0.80
  - **F1-score:** 0.77

The classification report of the model

	precision	recall	f1-score	support
0	0.87	0.82	0.85	1699
1	0.74	0.80	0.77	1059
accuracy			0.82	2758
macro avg	0.80	0.81	0.81	2758
weighted avg	0.82	0.82	0.82	2758

- **Conversion Rate for the leads is more than 80%**
- Variables that mattered the most are Total Times Spend, Total Number of visits, Lead Source etc.
- Variables that moderately mattered are Last Activity, Lead Origin, and Current Occupation
- The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model