LEAD SCORE CASE STUDY

CASE STUDY PARTNER:

VIREN GANDHI

VRUSHALI RAJKAR

YOGESH NALAGE

BATCH: DS C62

BATCH ID: 5674

STUDENT NAME: VIREN ASHOK GANDHI

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.
- 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.
- 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 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.

BUSINESS OBJECTIVES:

- X education wants to know most promising leads.
- The company needs a model wherein you a lead score is assigned 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%.
- Deployment of model for the future uses

Methodologies followed to analysis:

- Data cleaning and data manipulation.

  1.Check and handle duplicate data.

  2.Check and handle NA values and missing values.

  3.Drop columns, if it contains large amount of missing values and not useful for the analysis.

  4.Imputation of the values, if necessary.

  5.Check and handle outliers in data.

- EDA

  1.Univariate data analysis: value count, distribution of variable etc.

  2.Bivariate data analysis: correlation coefficients and pattern between the variables etc.

  3. Using different plot like count/bar/pie/hitmap

Methodologies followed to analysis:

- Feature Scaling & Dummy Variables and encoding of the data.
- Splitting the data into Test and Train dataset
- Classification technique: logistic regression used for the model making and prediction.
- Evaluation the model by using different metrics like Specificity and sensitivity or precision and Recall
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

Methodologies followed to analysis:

Data Sourcing, clearing and preparation:

1. Read the data from sources
2. Convert the data in clean format so it is suitable for analysis
3. Remove duplicate data if any
4. Outlier treatment
5. EDA
6. Feature standardization

Feature scaling and splitting of data:

1. Feature scaling of numeric data
2. Splitting data in to train and test at ratio of 70:30

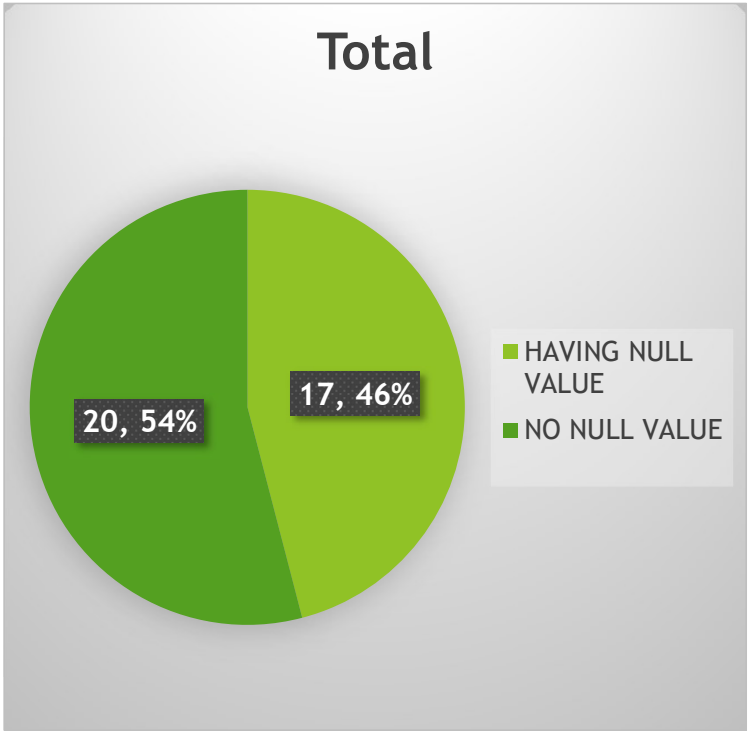
Model Building:

1. Feature selection using REF
2. Determine the optimal model using logistic regression
3. Calculation of various metrics like accuracy, sensitivity, specificity, precision and recall
4. Evaluation of model

Results

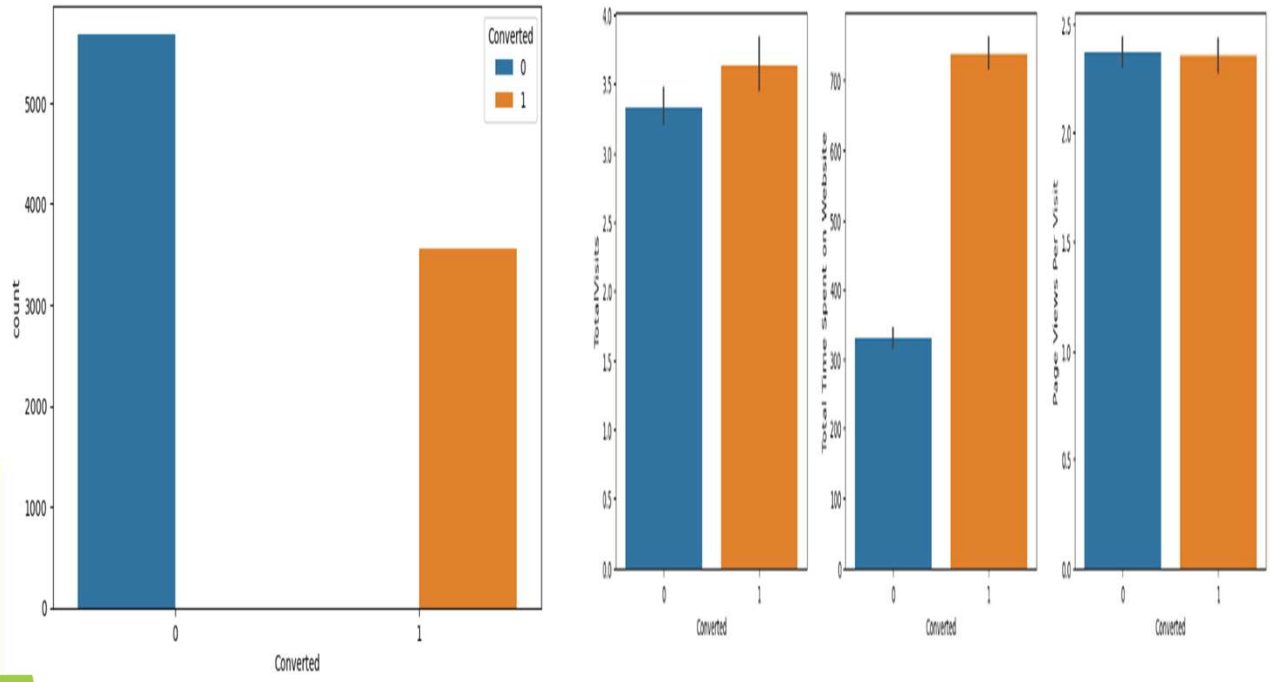
1. Determine the lead score and check if target final rate more than 70% conversion rate.
2. Evaluate final prediction on the test set using threshold limit

Read and understand the data:



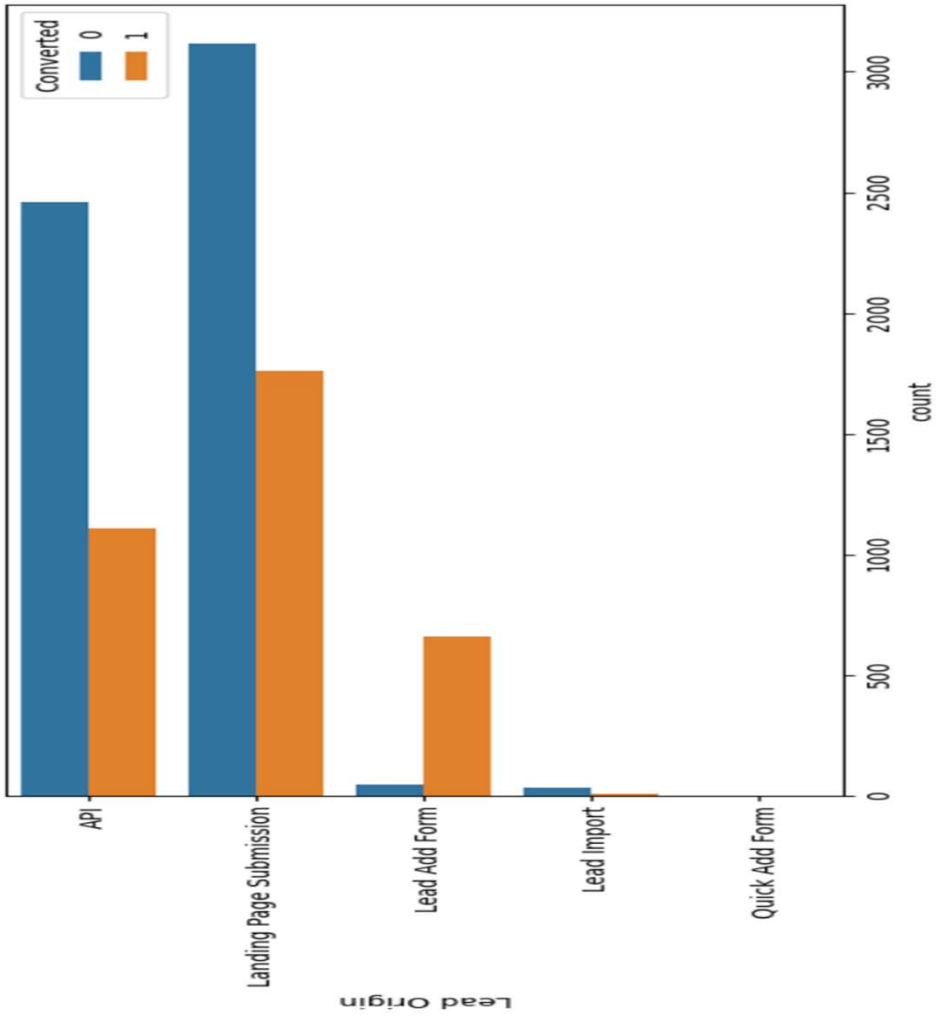
1. There is total 9240 rows and 37 columns.
2. Out of 37 there are 7 numerical and rest 30 are categorical variables.
3. Out of 37 there are 17 heading which having null rows and rest 20 are without null row.
4. There are outliers present mostly in 'totalVisits','Total Time Spent on Website' and 'Page Views Per Visit’.

Exploratory Data Analysis



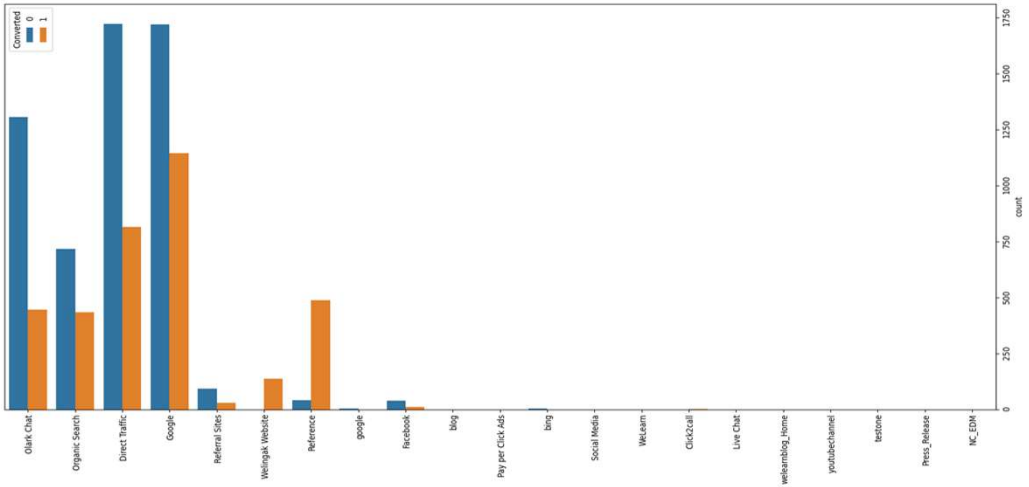
1. We have around 39% conversion rate in total.
2. We have mojor lead conversion from Total Visits, Total Time Spent on Website, Page Views Per Visit

Understanding lead conversion and lead origin

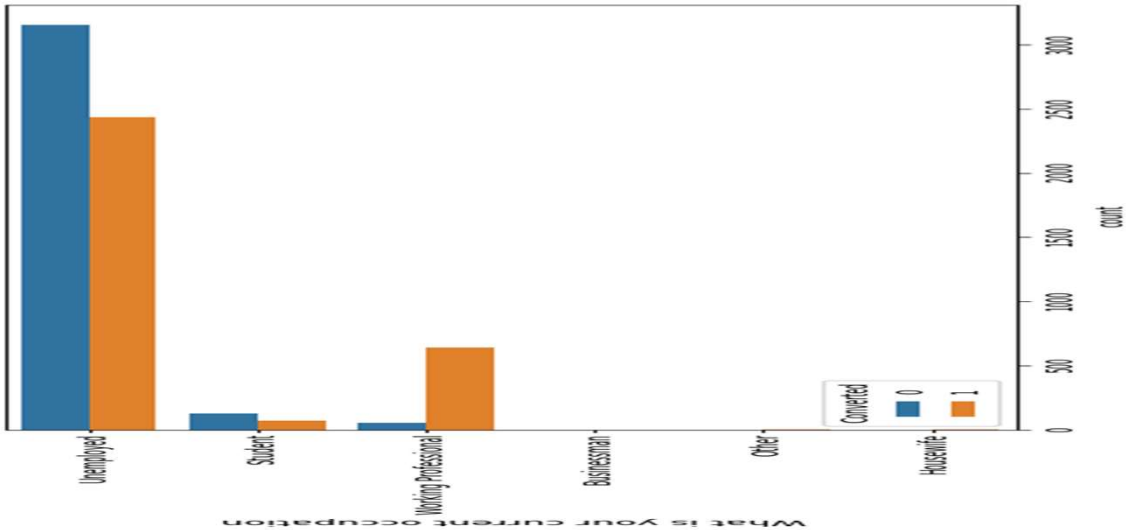


In lead origin maximum conversion happened from landing page submission

Understanding lead conversion and lead Sources

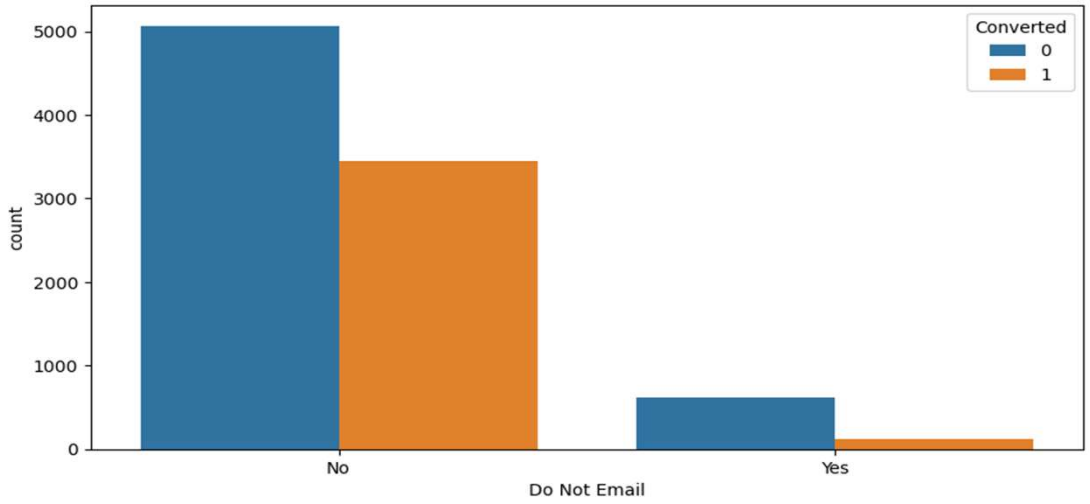


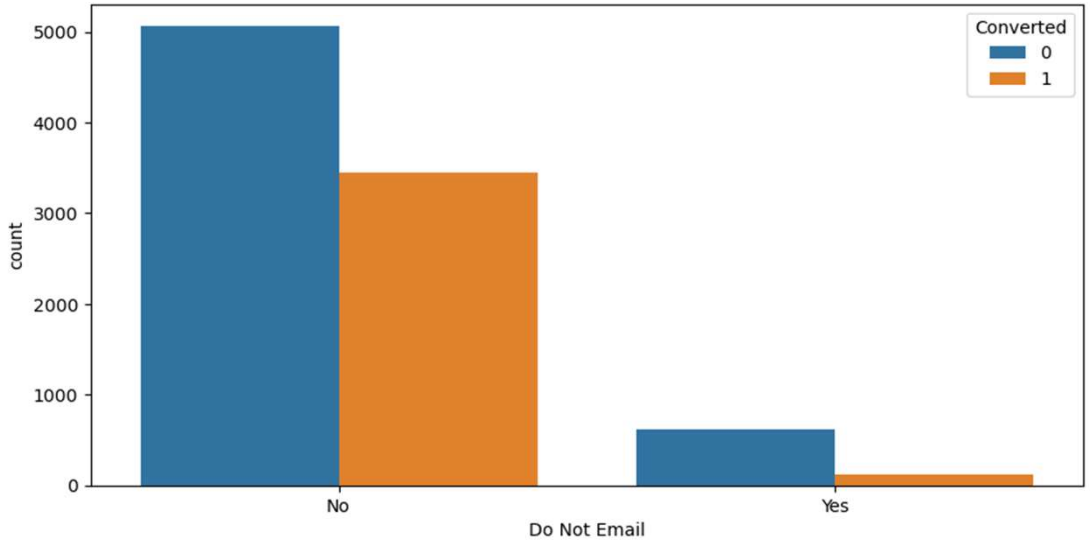
From the given graph it is clearly visible that major lead conversion in the lead sources is from “Google”



From the given graph it is clearly visible that major lead conversion in the lead sources is from “Unemployed group”

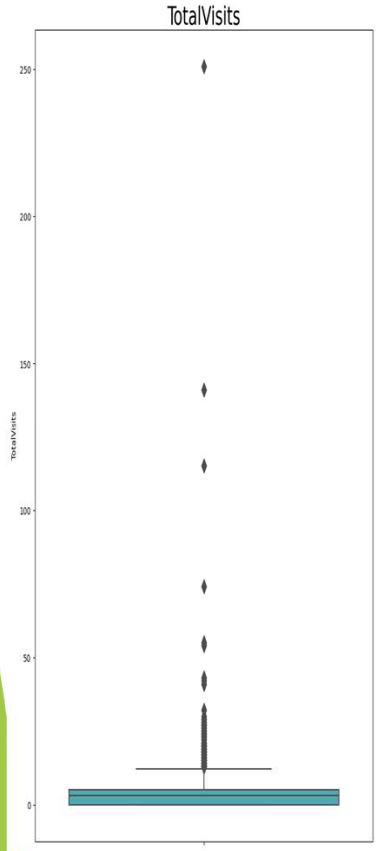
Understanding lead conversion and lead Sources

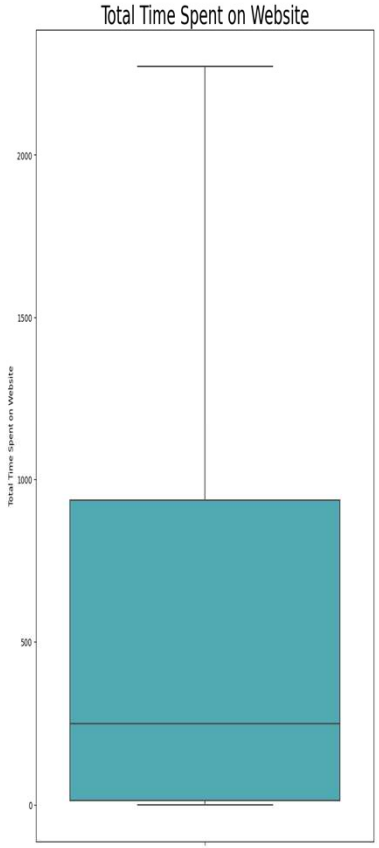


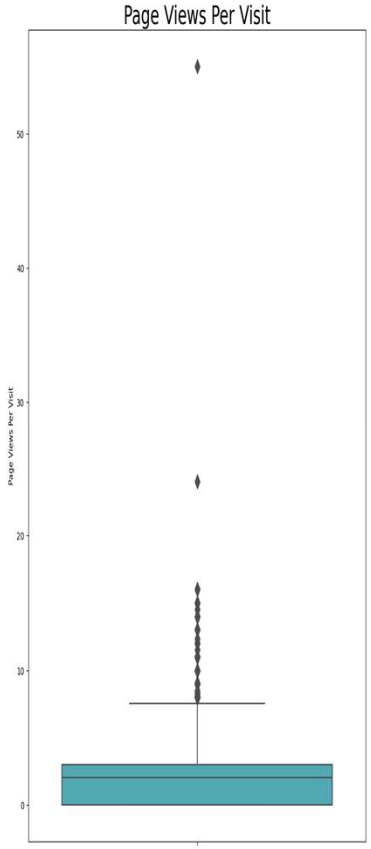


From the given graphs it is clearly visible that major lead conversion has happened from Email sent and calls made

Data Cleaning





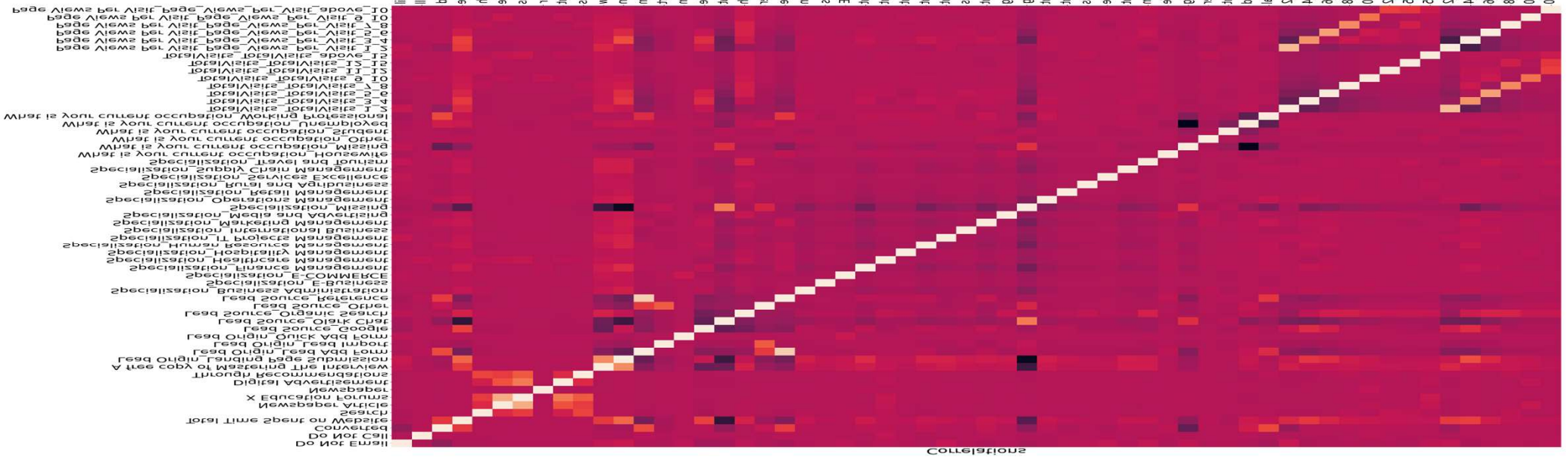


There are some interesting facts regarding the data cleaning and preparation:

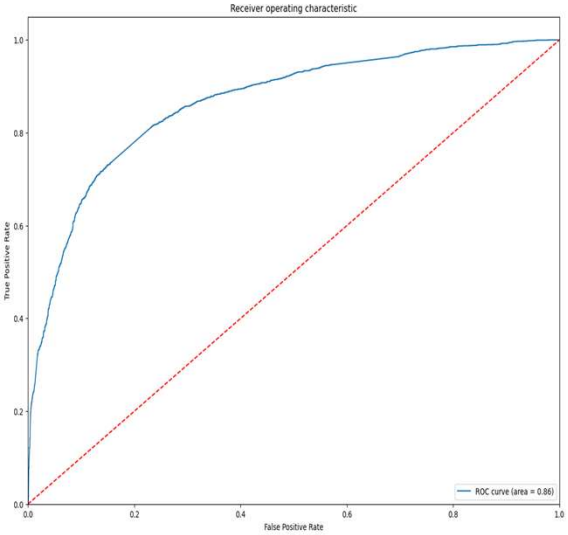
1.Cleaning the dataset by removing the redundant variables/features.
2.Remove columns having more than 40% null values
3.Imputing missing values as per column data available
4.Converting yes/no category column in to binary 0 and 1.
5.Deal with outliers.
6.Create dummy variable.
7.Remove all redundant and repeated colums.

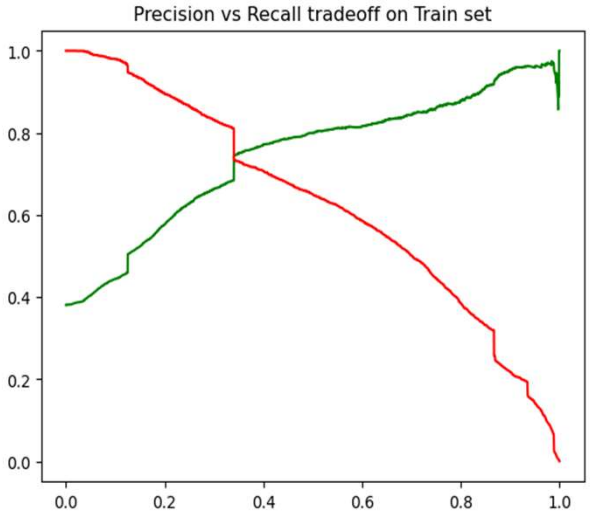
Data Preparation

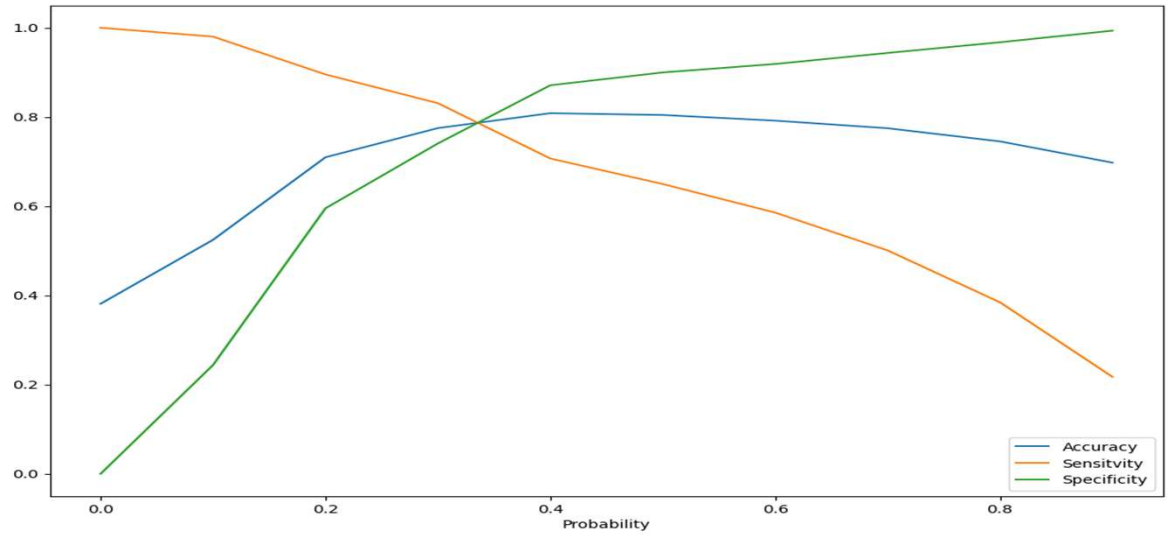
Split the dataset into train and test dataset and scaled in dataset
Checking the correlation among the variables
Checking highly correlated features



MODEL EVALUATION-ROC.SENSITIVITY AND SPECIFICITY





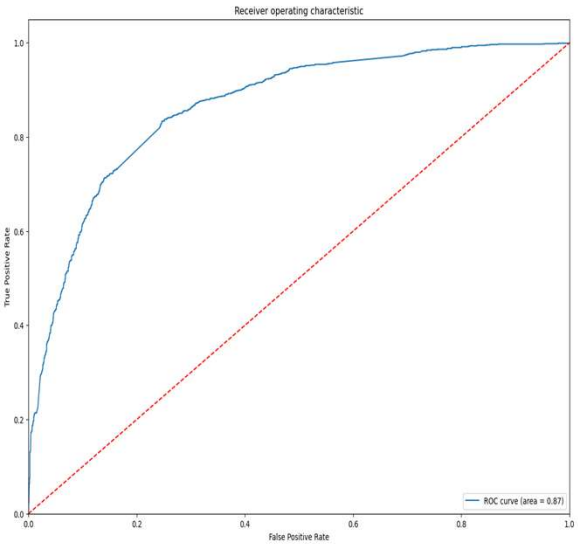


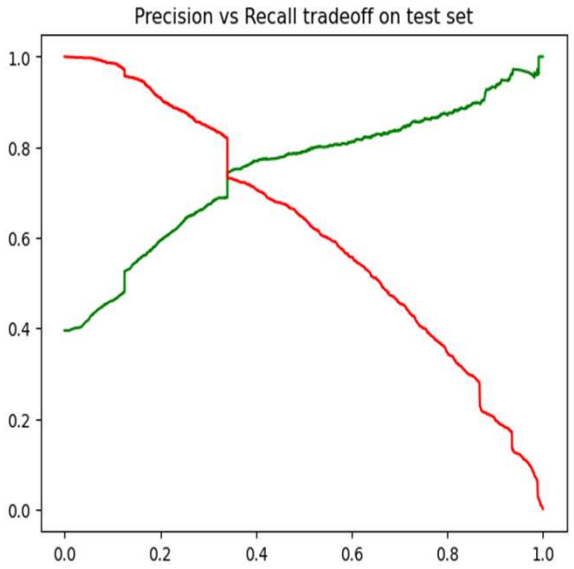
Confusion Matrix

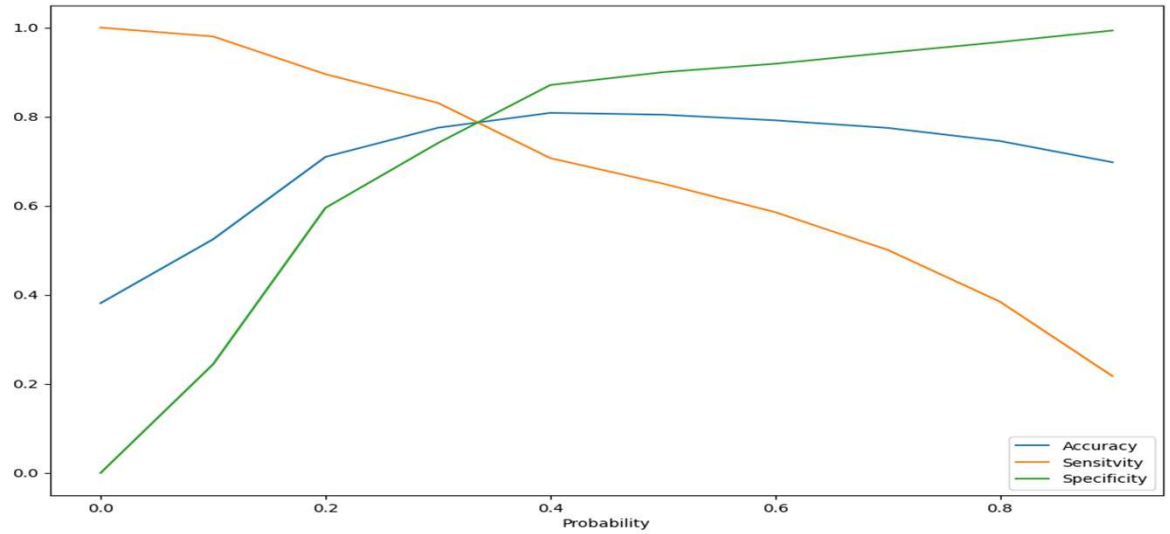
2964	1038
417	2049

Accuracy:- 81%
Sensitivity:- 83%
Specificity:-74.06%
Precision:-66%
Recall:-83%
Train Set Accuracy: 77%

MODEL EVALUATION TEST DATA-ROC,SENSITIVITY AND SPECIFICITY







Confusion Matrix

1228	449
170	925

Accuracy:- 81%
Sensitivity:- 84.47%
Specificity:-73.23%
Precision:-67.32%
Recall:-84.47%
Train Set Accuracy: 77.67%
FI SCORE: 74.92%

Metrics Comparison between Train data set and Test data set

Train Data Set Metrics:

Sensitivity:- 83.09%
Specificity:-74.06%
Precision:-66.38%
Recall:-83.09%
Accuracy: 77.5%

Test Data Set Metrics:

Sensitivity:- 84.47%
Specificity:-72.23%
Precision:-67.32%
Recall:-84.47%
Accuracy: 77.67%

CONCULSION:

After the EDA and Model evaluation we have come to the following conclusion:

- The Sensitivity and Specificity, Accuracy, Precision and Recall score we got from test set are almost accurate.
- We have high recall score than precision score which is a sign of good model.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
  - Lead Origin_Lead Add Form
  - Total Time Spent on Website
  - What is your current occupation_Working Professional