Introduction

This document covers the lead scoring case study

- Document explains the case study conducted in following manner :
  - Problem Statement
  - Approach to the case study
  - Model Building
  - Results of Analysis

Analysis by :

Abhishek Kare
Ashisha
Mallika Gollapalli

Problem Statement

- Education Companies such as X education find it hard to acquire more registrations.
- X Education has online courses to offer for industry professionals. Many professionals land on their website and browse their courses.
- X education has the strategy of reaching out to leads through search engines, several websites. Lead information is captured for further filtering and processing .
- Despite all the process in place to acquire leads , conversion rate is not as good as expected by the company.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

Assumptions

- Data is loaded by keeping the file in the local path .
- Based on the data dictionary , few of the columns are identified for Analysis. There could be more columns, but based on business relevance few of the columns are analysed and coded in notebook.
- Only selected graphs are taken into ppt. Notebook covers more analysis and graphs .
- Treatment for Missing and null values in Lead data is done for columns where missing data percentage is more than 40%.
- Few of the categorical column has data as ‘Select’ these are treated as missing values as the actual value is not available.
- Based on the Target variable ‘Converted’ which is a categorical variable the model to be built will be Logistic regression model .

Analysis Approach

Approach for Logistic Regression is as mentioned in the below Steps

Steps	Approach
Step 1	**Data Understanding** :- • Load the data and perform the basic sanity of the data checking for size, type of data types, data Quality check
Step 2	**Data Cleaning and Manipulation** :- • Missing value checks • Data type checks • Handling Outliers
Step 3	**Data Analysis / EDA**:- • Univariate Analysis • Bivariate Analysis • Multivariate Analysis / Correlations.
Step 4	**Data Preparation:-** • Creating Dummy variables for the categorical columns • Dropping the unwanted columns for the model building.
Step 5	**Model Building** • Model building using RFE • Iterate the model build till optimal p value and VIF values are achieved
Step 6	**Calculation of Metrics** • Plotting of ROC curve, finding optimal cut-off • Calculation of accuracy, precision, Recall, sensitivity and specificity and probability and Lead score.
Step 7	**Predictions on Test set** • Recalculate all metrics on Test set and compare with Train data metrics.

Data Understanding and Data Cleaning

- As a first step , Data dictionary has been referred to understand the different columns and meaning of those columns.

- Next step, The data is loaded in the Jupyter Notebook.

- Following are the observations of the Data

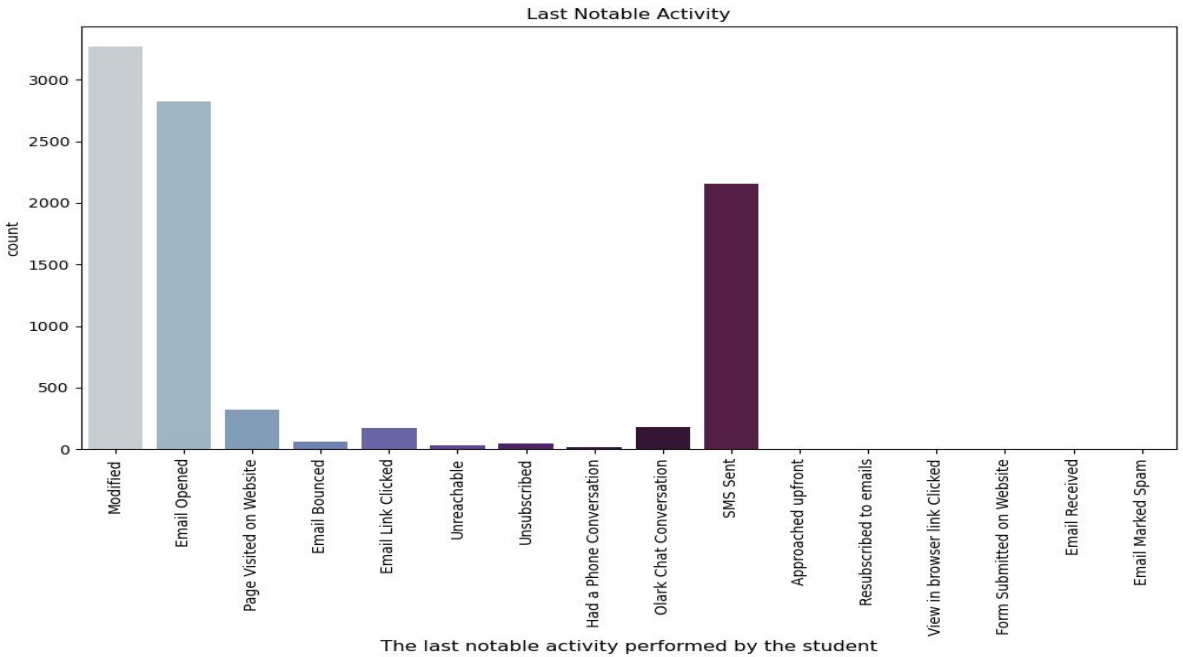
  - Lead scoring data set has 37 columns and 9240 rows.
  - Majority of the Columns are Object and very few numerical columns.
  - Lead Number and Prospect ID are unique numbers and they will be dropped for model building
  - Country column has majority same value and City data is not very relevant. Both will be dropped for model building.
  - Columns below are dropped as they have same data across all rows
    Update me on Supply Chain Content
    Get updates on DM Content
    I agree to pay the amount through cheque
    Receive More Updates About Our Courses'

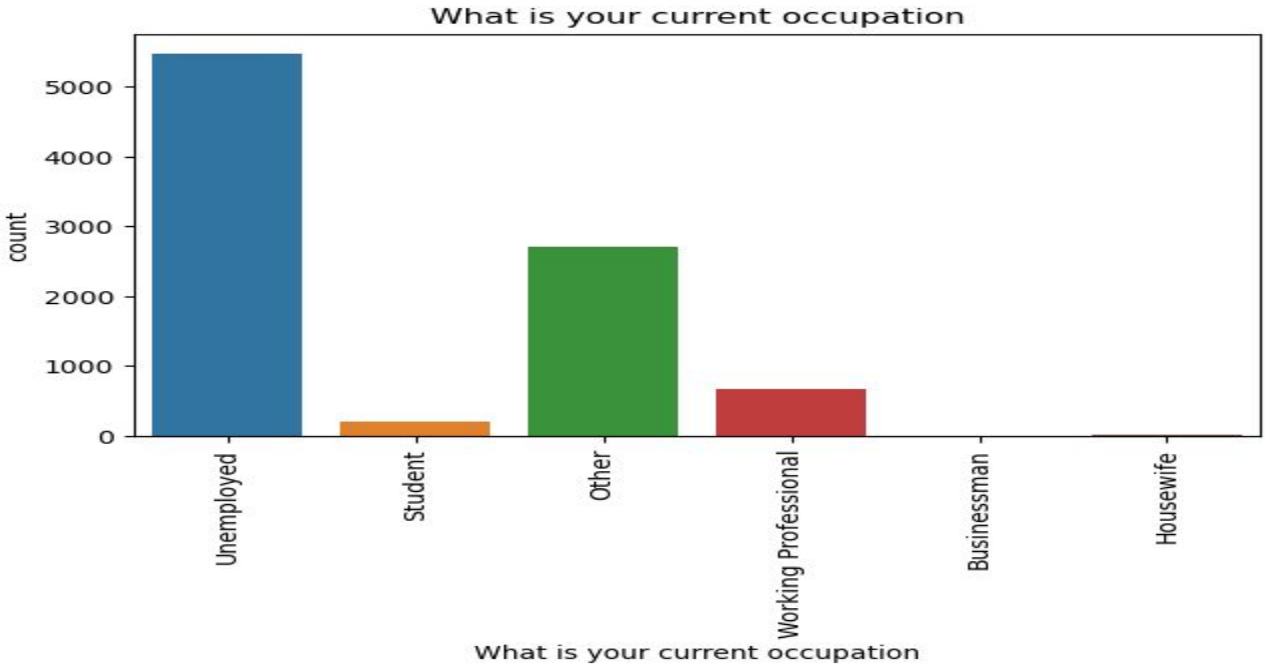
Missing value Treatment

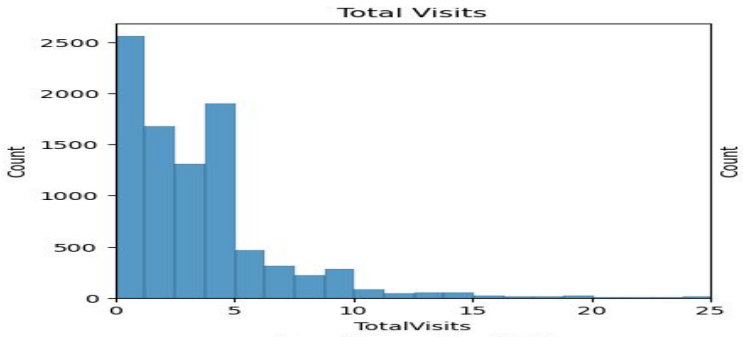
Missing Value treatment - Threshold of 40% is considered to handle the missing values.

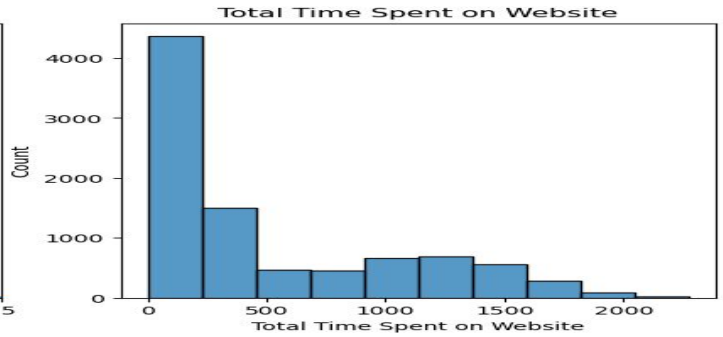
- Following columns have more than 40% so they will be dropped.
  - Lead Quality, Lead Profile , How did you hear about X Education .
- Specialization – has 15% missing values . ‘it has been imputed with ‘Others’.
- What is your current occupation- This has 30% missing values and has been imputed with ‘Others’
- What matters most to you in choosing a course – This has 29% missing values and has been imputed with mode ‘Better career prospects’ as 70% of the data is having this value.
- Tags – Having 37% null values , it is less than the threshold for missing value treatment but all the values are skewed and hence dropping the same.

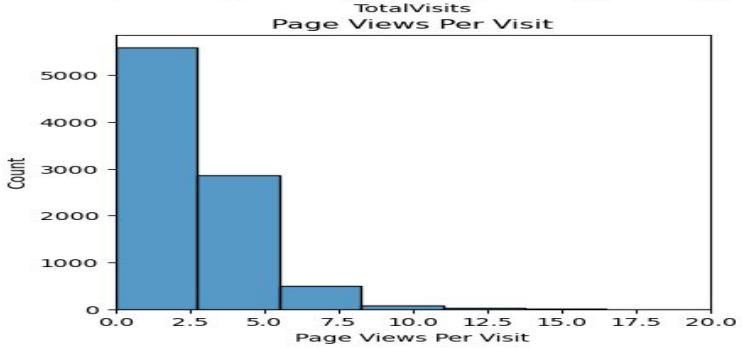
EDA





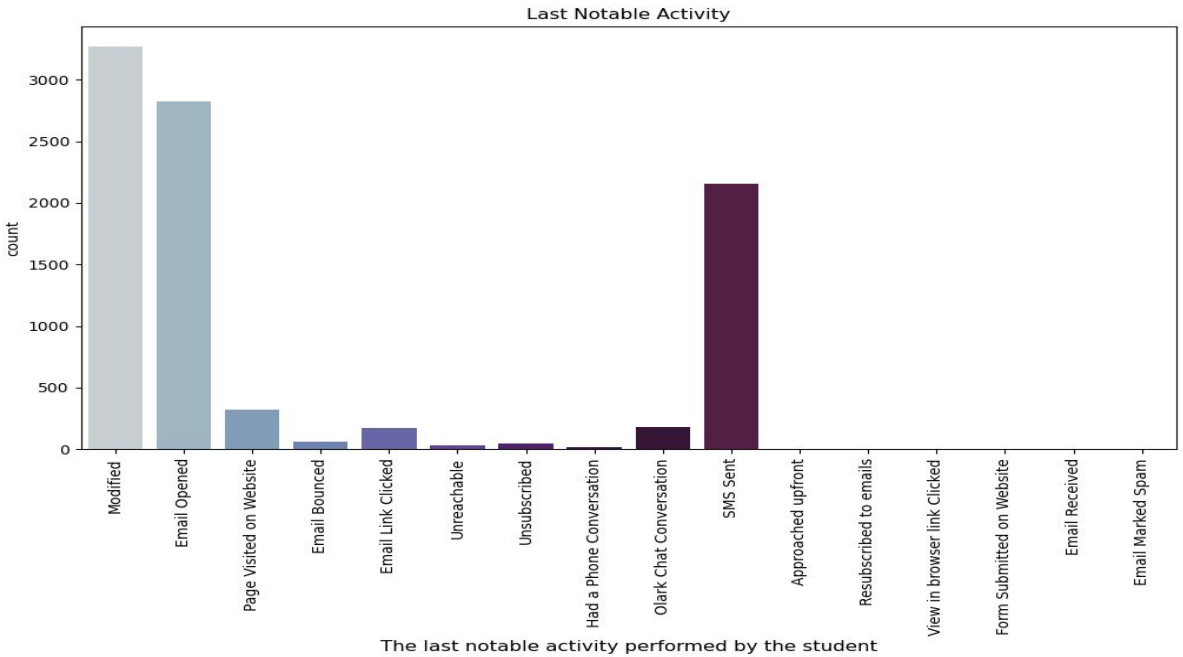


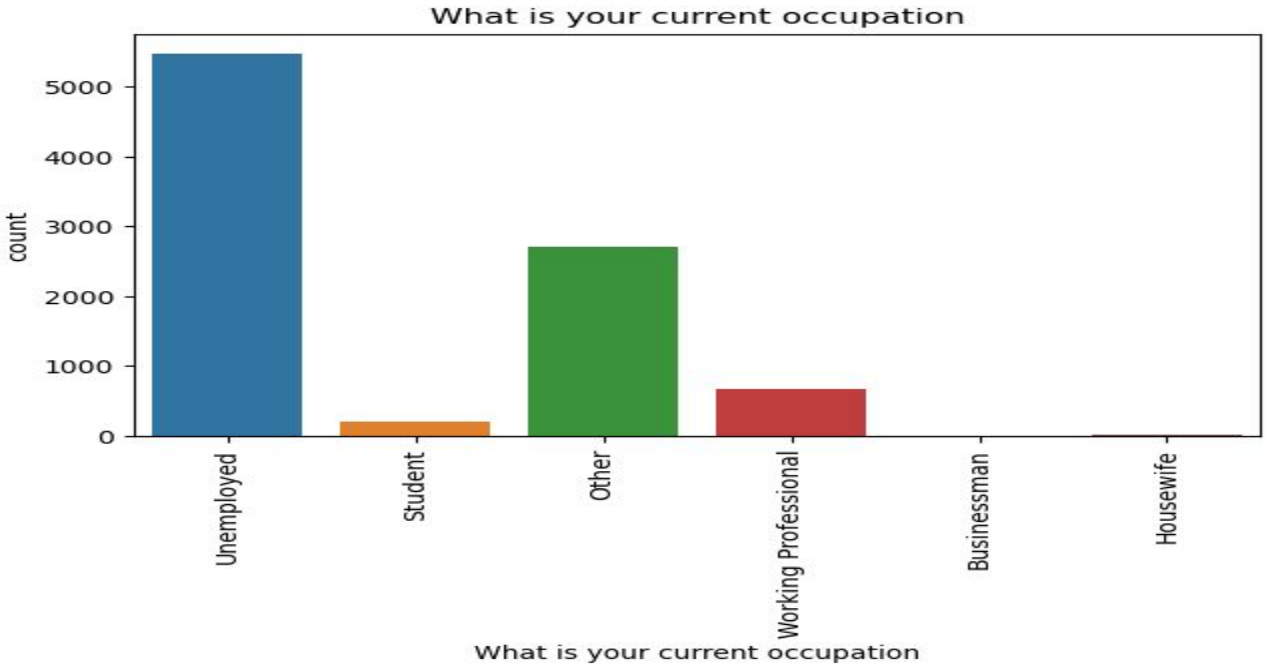


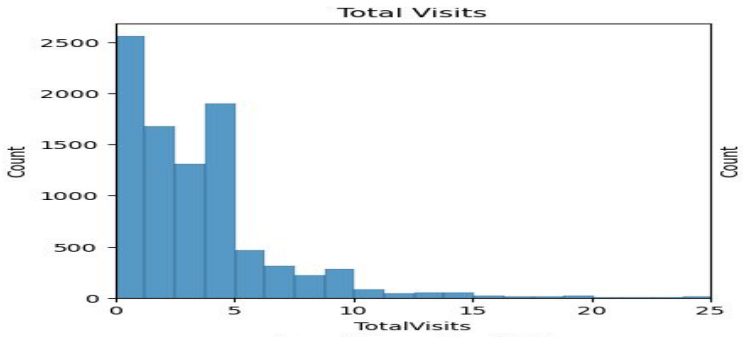


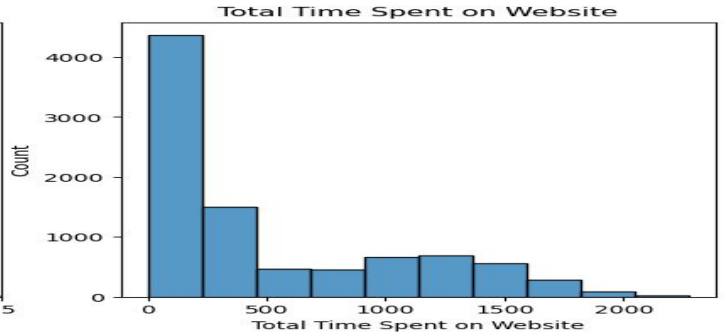
- Based on the Analysis majority of the leads occupation is unemployed.
- Most of the leads last notable activity is modified or Email opened or Email sent.
- Not all leads have visited the website and have not spent enough time on website.

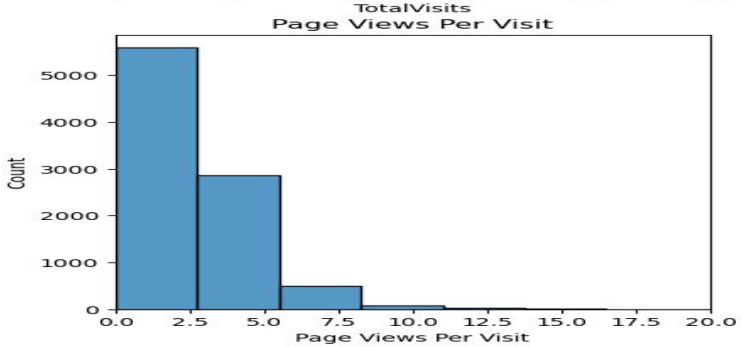
EDA





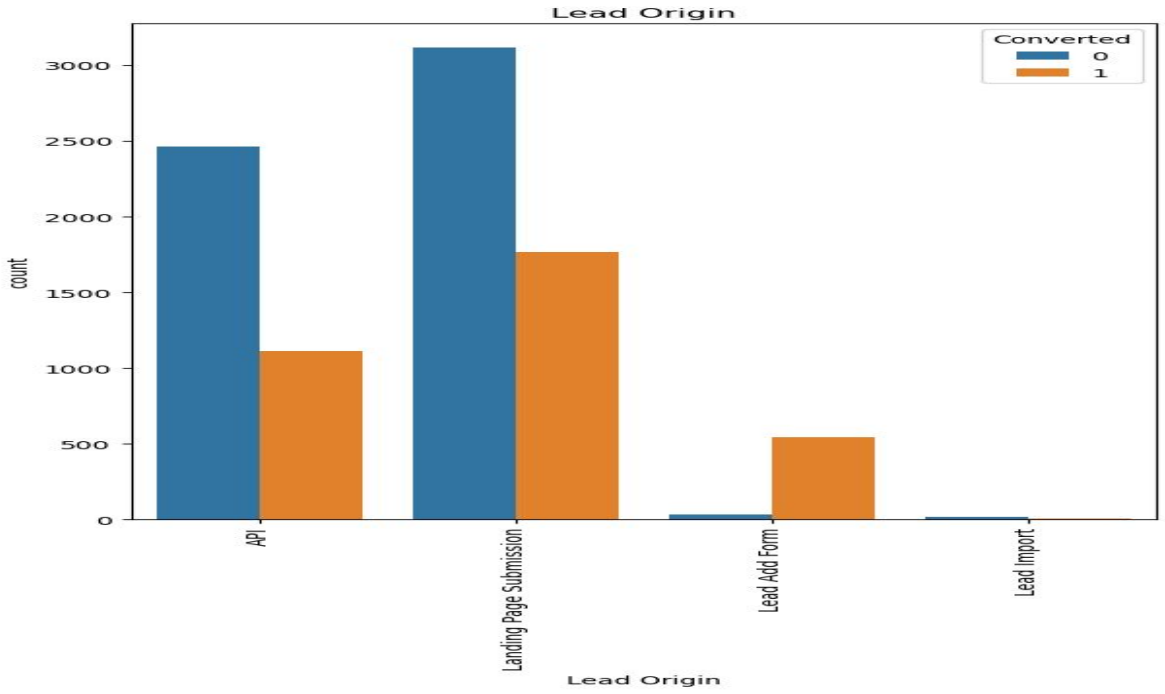


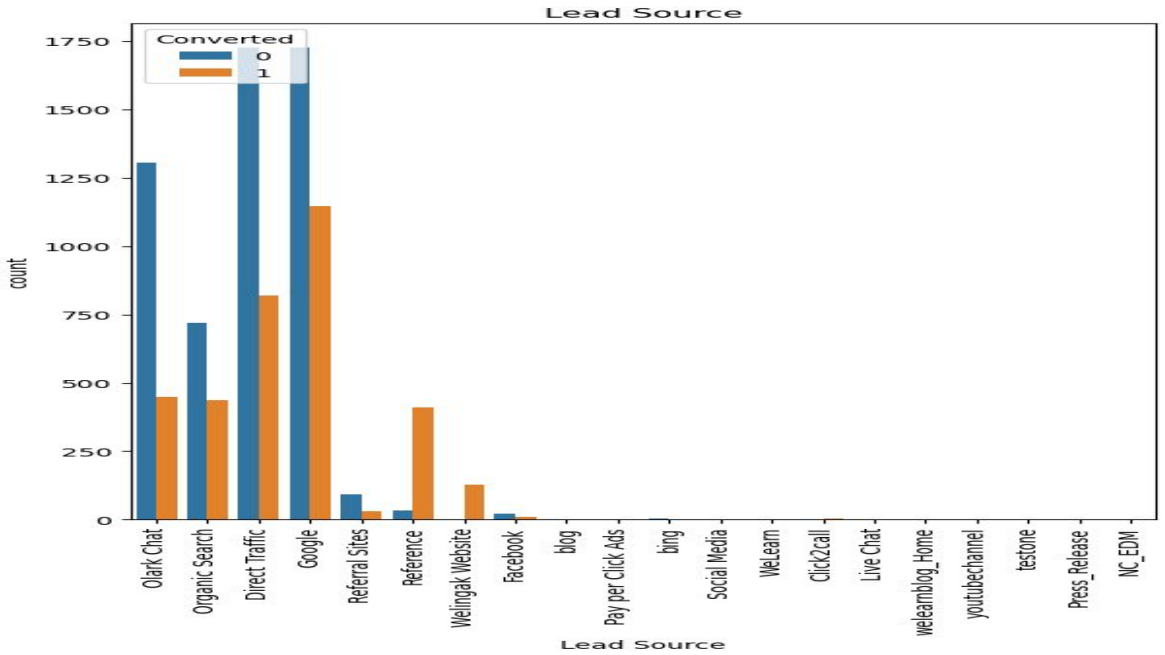


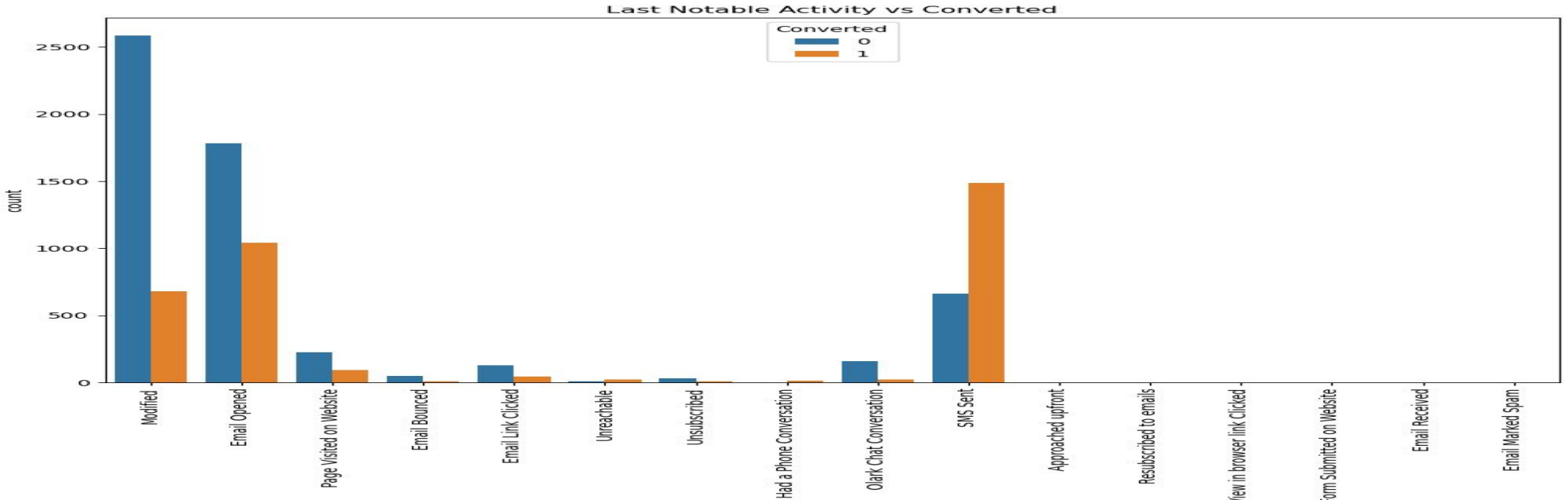


- Based on the Analysis majority of the leads occupation is unemployed.
- Most of the leads last notable activity is modified or Email opened or Email sent.
- Not all leads have visited the website and have not spent enough time on website.

EDA







- From the Analysis it is observed that converted leads are from google and direct traffic sources. Most Converted leads are originating from landing page submission and they have spent time in visiting the Email and SMS

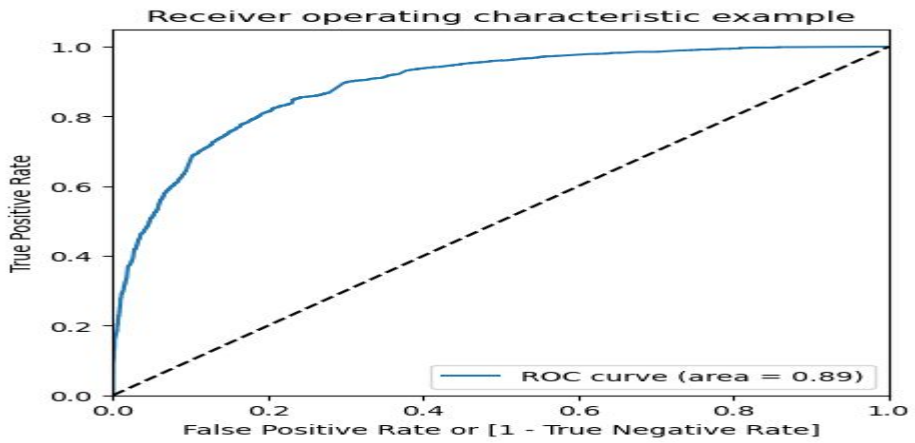
Data Preparation and Model Building

- Data Preparation for model building
  - First step in model building is to ensure that the dataset is proper.
  - For categorical data with 2 values ‘Yes’ and No’ are converted to 1 and 0 respectively . Following variables are converted to binary data. 'Do Not Email', 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'A free copy of Mastering The Interview’
  - MinMaxScaler function was instantiated to scale all the numeric variables.
  - For all multiple categorical values dummy variables will be created.
  - Total columns after the data preparation is 92 .

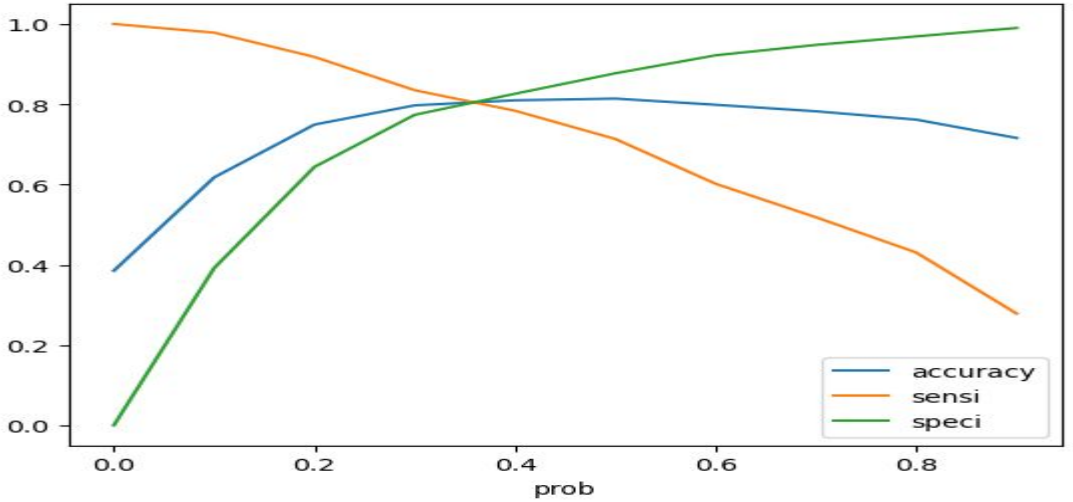
- Model Building
  - First step in building the model is to create the Training data and Test data .
  - Ratio chosen to create this data is 70:30 for training and test respectively
  - Using RFE as feature selection , 20 features are selected .
  - Iteratively Model was rebuilt till the optimal values of p value and VIF were achieved. P value with in 0.05 and VIF values less than 5.
  - Final model contains 17 features

ROC Curve & Optimal Cutoff and Metrics

ROC Curve - Test is accurate as the curve is closer to the left border and the top border.



Finding Optimal Cutoff



From the graph Optimal cutoff can be considered as 0.35.
With the cutoff following metrics are calculated :

- accuracy - 80%
- sensitivity - 81%
- Specificity - 79%
- Precision - 71%
- Recall - 81%

Sample Lead Scores are calculated as follows : Below Table depicts the converted value , probability of conversion , predicted conversion and final prediction on conversion for the prospect Leads.

Prospect ID	Converted	Conversion_Prob	predicted	final_predicted	Lead Score
3009	0	0.098844	0	0	10
1012	0	0.358302	0	1	36
9226	0	0.193187	0	0	19
4750	1	0.712869	1	1	71
7987	1	0.674673	1	1	67

- Predicted value is calculated based on the cut-odd 0.5
- Final predicted value is based on optimal cut-ff 0.35
- Lead score below is calculated as probability *100 .

Test Predictions

- Test Data i.e the 30% of the data is considered to make predictions.
- Numeric test data is scaled and prediction is performed .
- Metrics are calculated as below

  - accuracy - 80%
  - sensitivity - 79%
  - specificity - 80%
  - precision - 70%
  - Recall - 79%

Sample Lead Scores calculated on Test data are shown below :
Tables depicts the converted value , probability of conversion , final prediction on conversion and lead score for the prospect Leads.

Final predicted value is based on optimal cutoff 0.35
Lead score below is calculated as probability *100 .

Prospect ID	Converted	prob_rate	final_predicted	Lead_Score
3271	0	0.129947	0	13
1490	1	0.966715	1	97
7936	0	0.113668	0	11
4216	1	0.8418	1	84
3830	0	0.066738	0	7

To Conclude the metrics of both train data and test data are matching. There is no deviation. The model seems to fit.

Conclusion

- Observation on the converted leads is as follows :
  - Most of the converted leads have visited website and have spent time on website
  - Lead source is from Google , direct traffic , References, welingak website.
  - Working professional, students and unemployed are interested in the courses.
  - Last Notable activity is email opened , SMS sent and modified.
- Recommendations:
  - Company should invest on website to ensure that leads visit and spend more time on website. There is more chances of leads getting converted when more details are available on website.
  - Sourcing from references and other search engines can continue and remain in momentum.
  - Working professionals can be targeted as leads and provide them flexible courses .
  - Using the logistic regression model that is built lead score can be generated based on which sales team can work on the Hot leads to ensure the conversion.
  - With the above changes X education should be able to attract more conversion of leads.