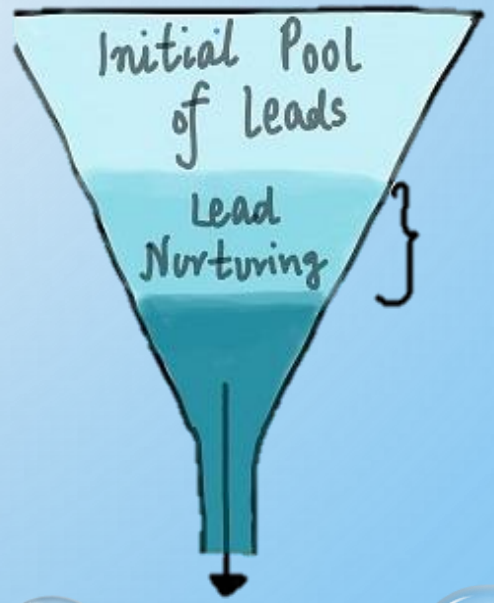
Lead_Scoring_Case_Study

Submitted by:
1. Manish Sahu
2. Sahel Siddique
3. Saichand Pratapagiri



PROBLEM STATEMENT:

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. They aim for a target lead conversion rate of 80%.
- 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 OBJECTIVE:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

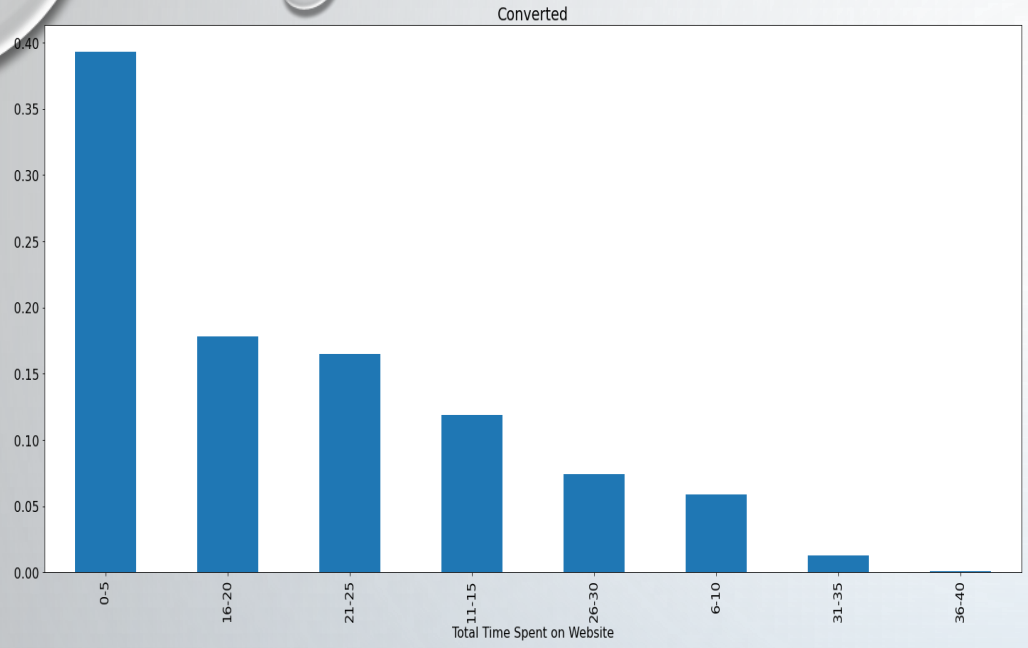
SOLUTION METHODOLOGY

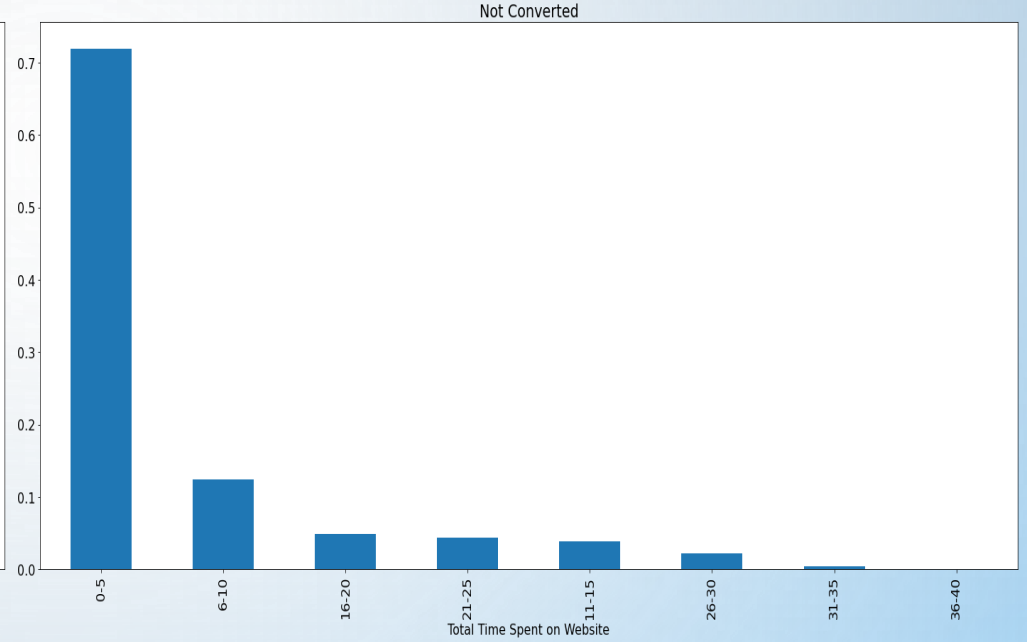
- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate and Multivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Finding the optimal cut-off and making predictions on text data.
- Conclusions and recommendations.

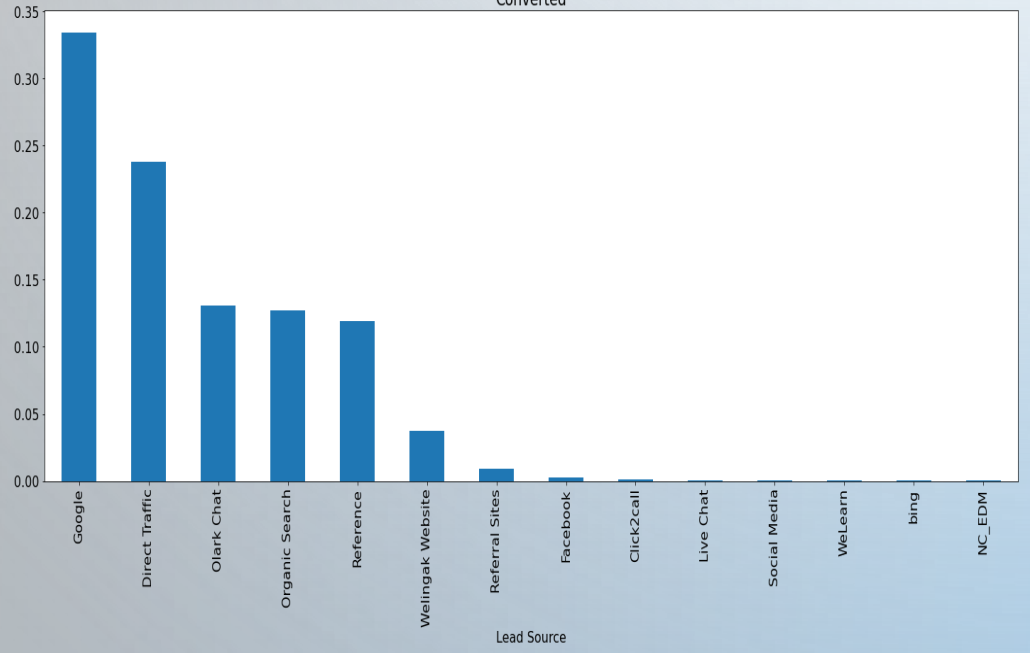
DATA MANIPULATION

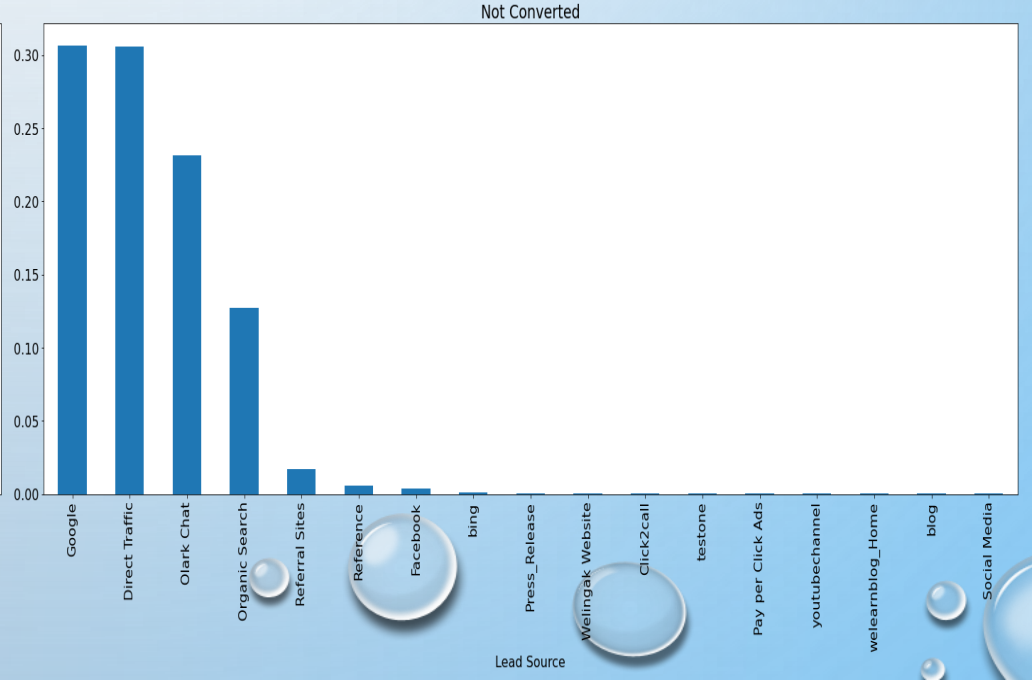
- Total Number of Rows =37, Total Number of Columns =9240.
- Dropping the columns having more than 33% as missing value such as 'Lead Quality','Asymmetrique Activity Index','Asymmetrique Profile Score','Asymmetrique Activity Score','Asymmetrique Profile Index', and 'Tags'.
- Columns such as 'I agree to pay the amount through cheque', 'Do Not call', 'Do Not Email', 'Newspaper', 'Newspaper Article', 'Get updates on DM Content', 'Receive More Updates About Our Courses', 'Digital Advertisement', 'A Free copy of Mastering The Interview', 'Update me on Supply Chain Content', 'Through Recommendations', 'Magazine', 'X Education Forums', 'Search' and 'How did you hear about X Education' have single value across all rows. These columns are of no use to perform analysis, etc. have been dropped.
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.

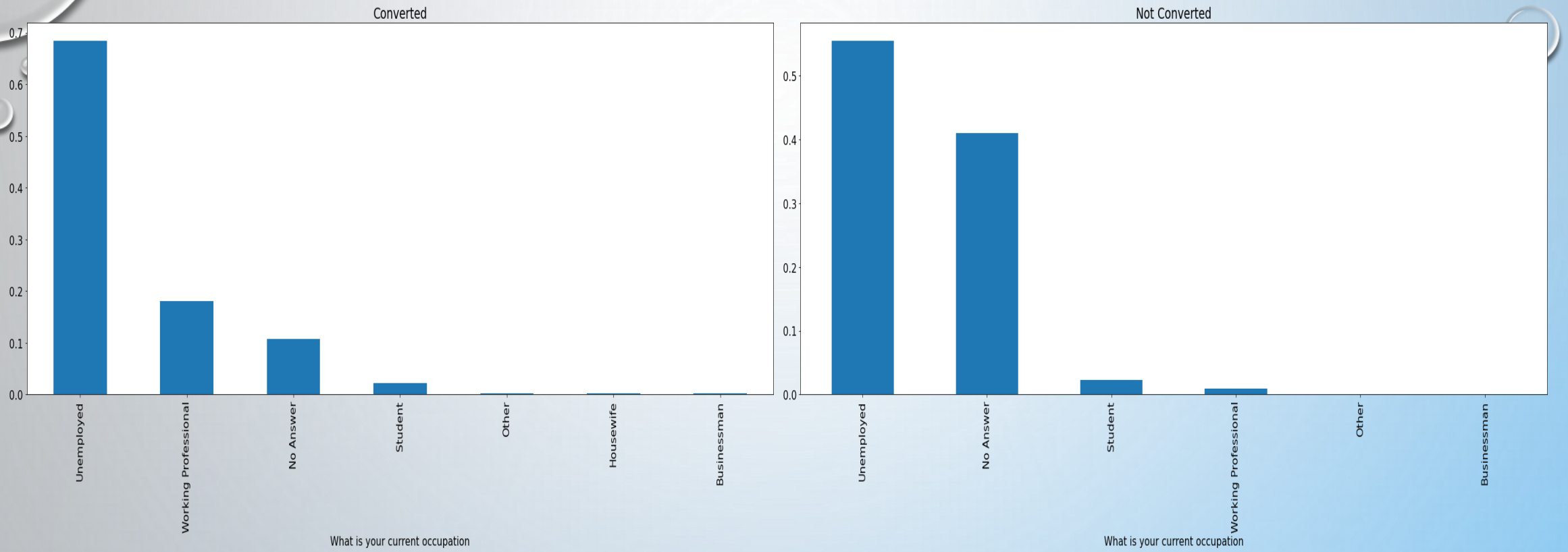
EDA - UNIVARIATE ANALYSIS







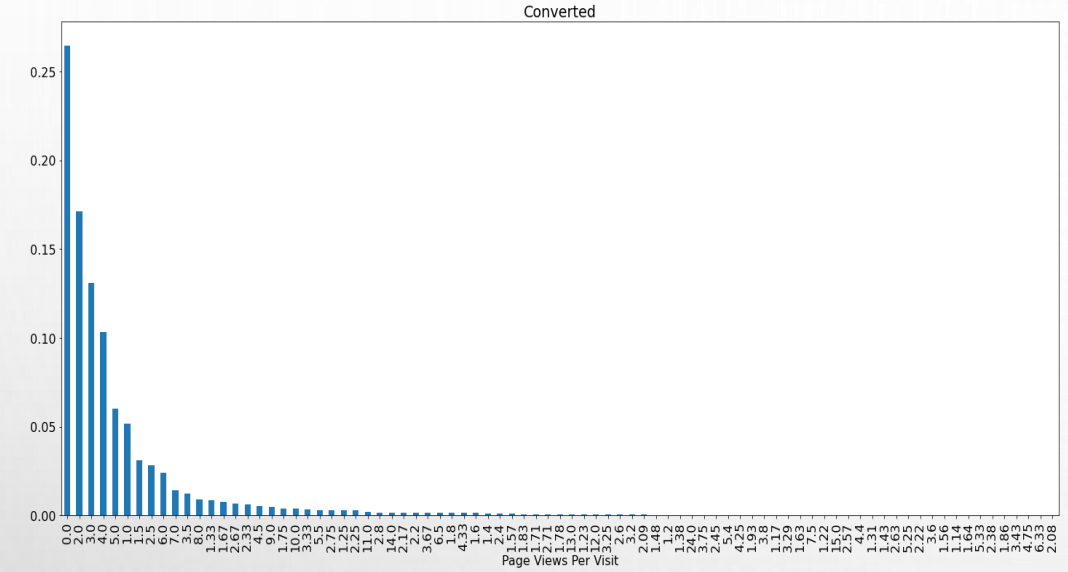


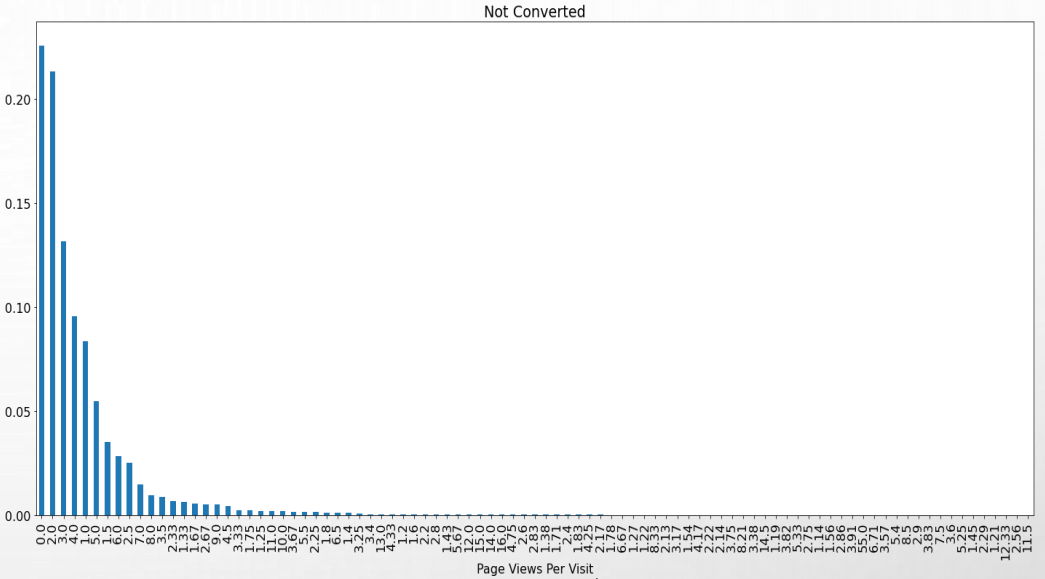


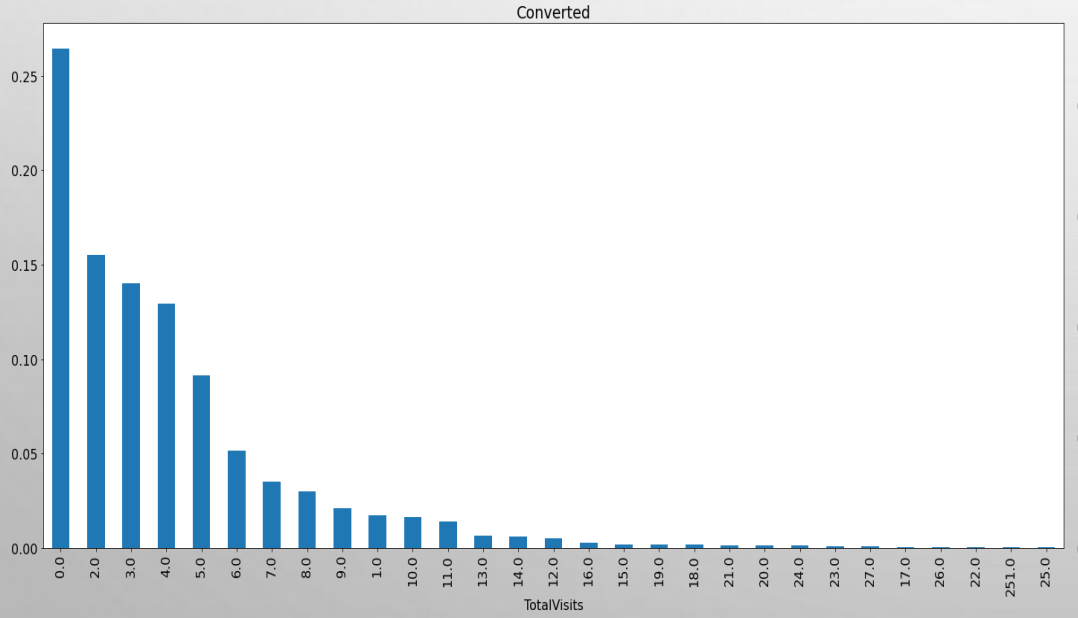
Observations :

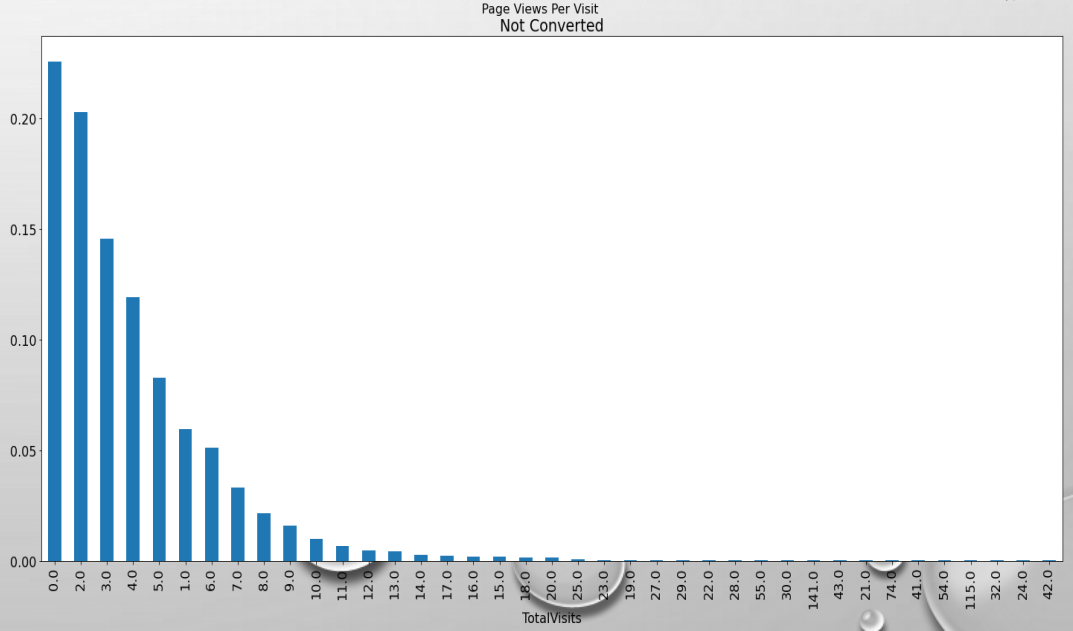
1) If total time spent on website greater than 15 minutes, there are more chances for lead to convert.
2) Working Professionals are more likely to convert.
3) Leads notified by SMS are more likely to convert.

EDA : NUMERICAL COLUMNS

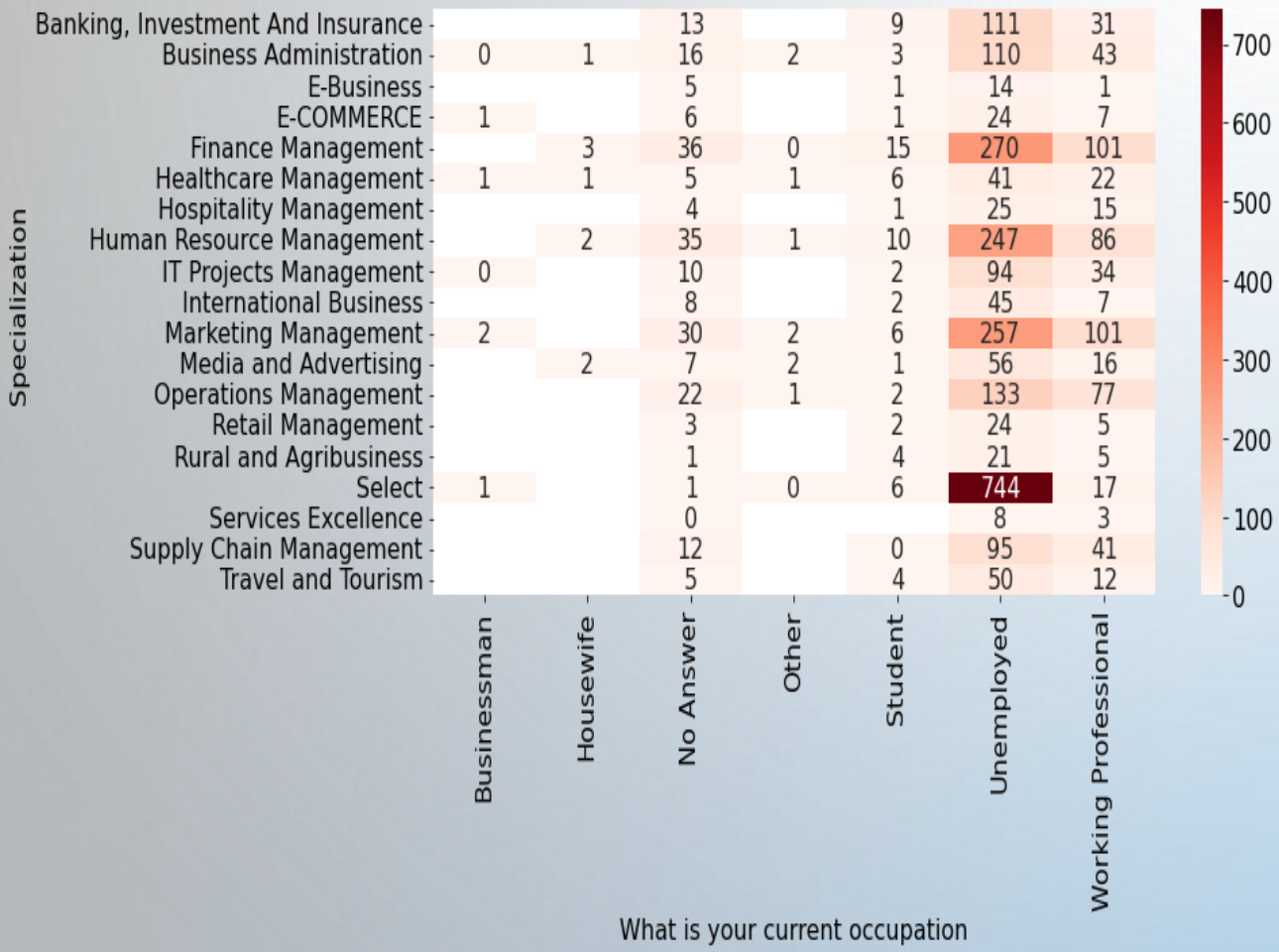


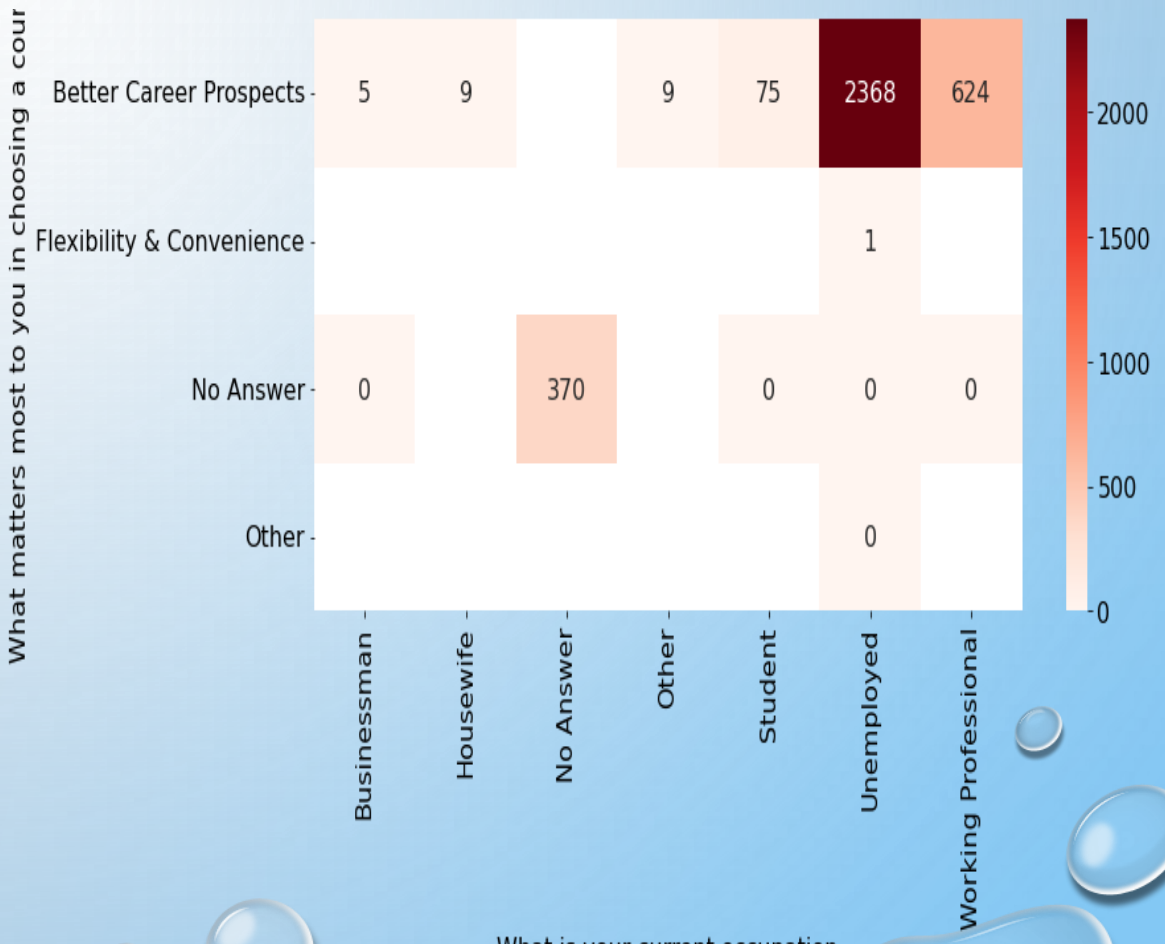






EDA- BIVARIATE ANALYSIS



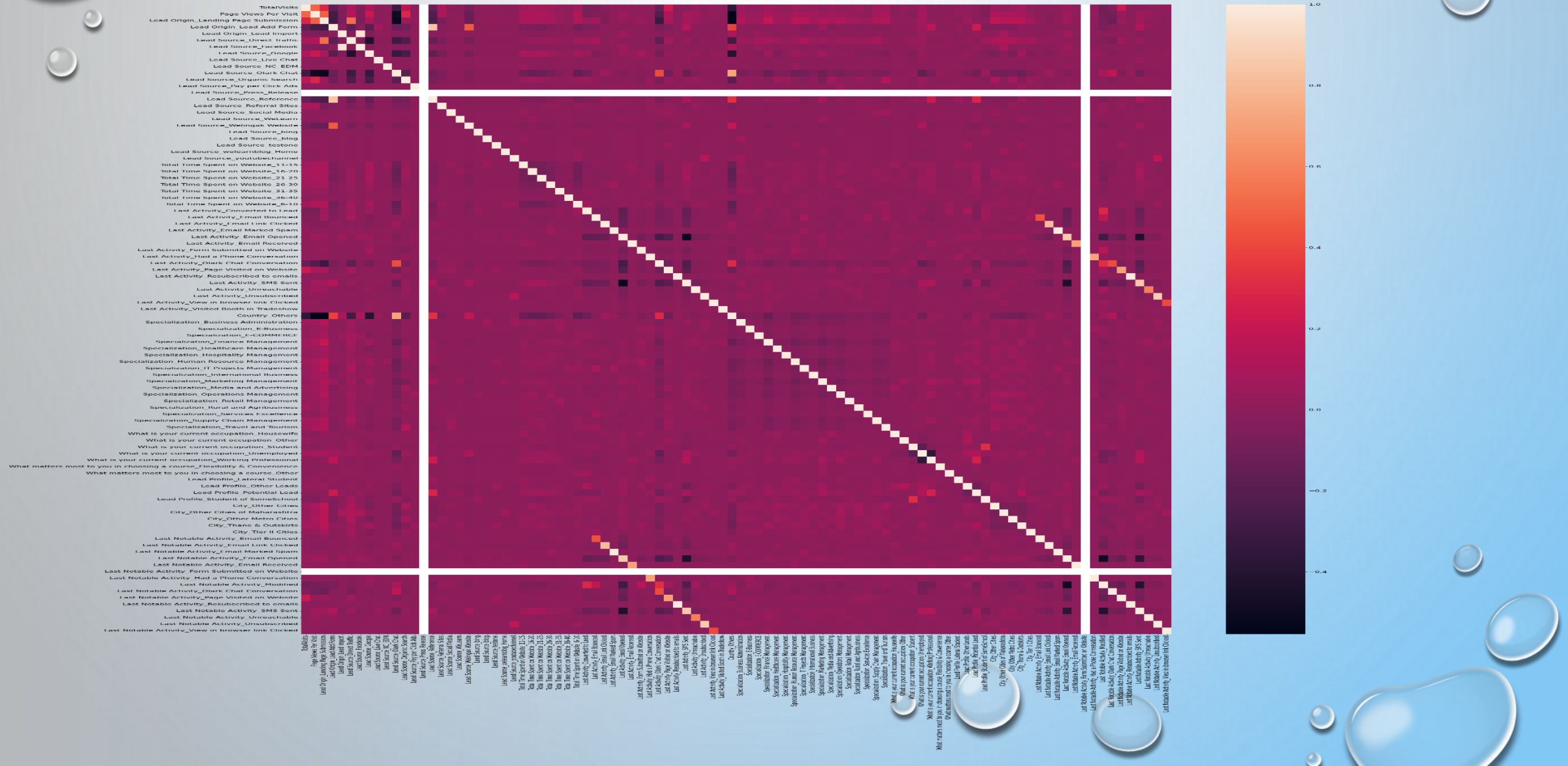


EDA – BIVARIATE ANALYSIS

Observations :

1) Unemployed people who aspire to take up specializations as Finance Management, Human Resources and Marketing Specializations are more likely to convert.
2) Unemployed people with Better Career Prospect are more likely to convert

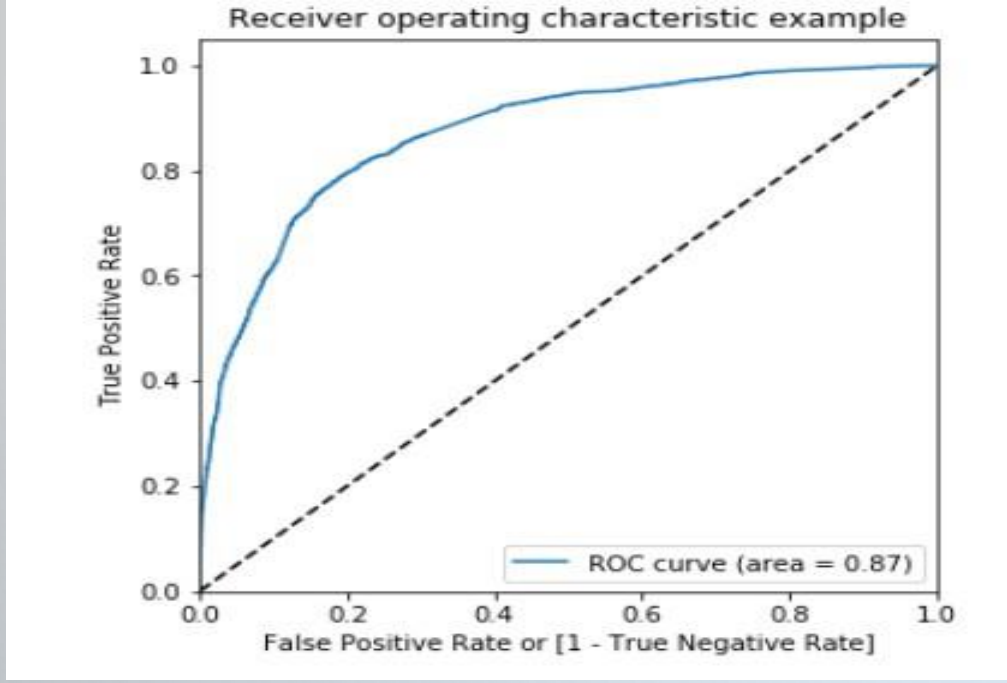
EDA –MULTIVARIATE ANALYSIS

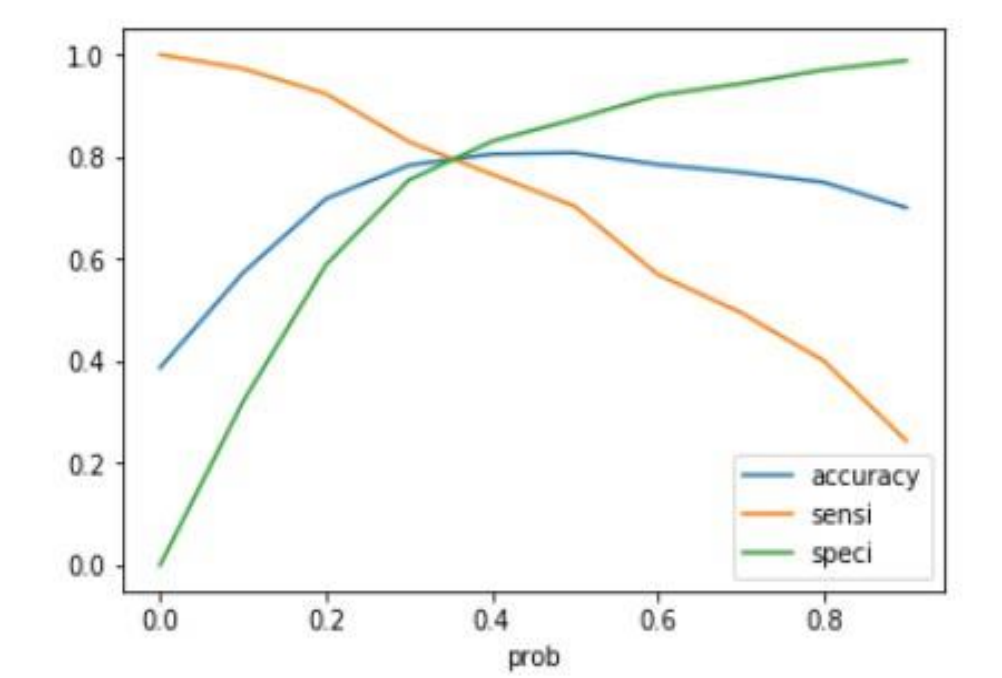


MODEL BUILDING

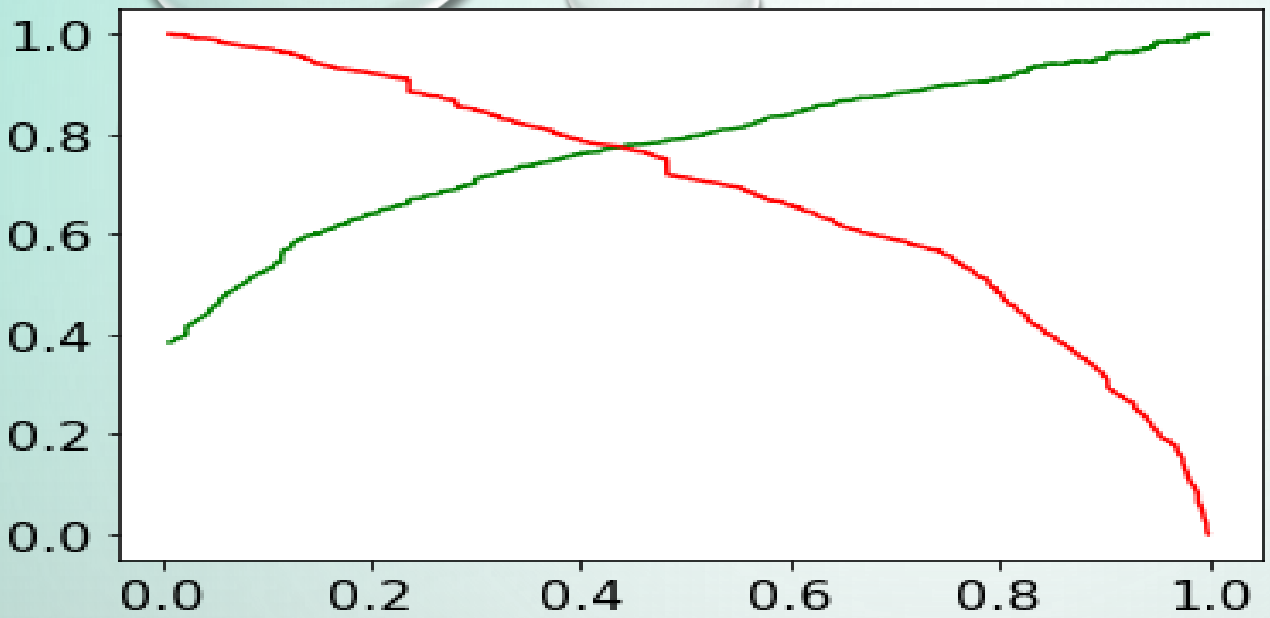
- SPLITTING THE DATA INTO TRAINING AND TESTING SETS
- THE FIRST BASIC STEP FOR LOGISTIC REGRESSION IS PERFORMING A TRAIN-TEST SPLIT, WE HAVE CHOSEN 70:30 RATIO.
- USE RFE FOR FEATURE SELECTION
- RUNNING RFE WITH 15 VARIABLES AS OUTPUT
- BUILDING MODEL BY REMOVING THE VARIABLE WHOSE P- VALUE IS GREATER THAN 0.05 AND VIF VALUE IS GREATER THAN 5
- PREDICTIONS ON TEST DATA SET
- OVERALL ACCURACY 81%

ROC CURVE





- Finding Optimal Cut off Point
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.



1) The trade-off between precision and recall occurs because improving one usually comes at the expense of the other. The choice of the threshold impacts the balance between precision and recall. By adjusting the threshold, we can make the model more conservative (increasing precision but potentially reducing recall) or more liberal (increasing recall but potentially reducing precision).
2) The threshold here selected is 0.41.

CONCLUSION

It was found that the variables that mattered the most in the potential buyers are :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  a. Direct traffic
  b. Organic search
  c. Welingak website
- When the last activity was:
  a. SMS
  b. Olark chat conversation
- When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.