Lead Score Case Study

Group Members:

- Ankit Prakash Sharma
- Keerthana Prasanna Kumar
- Sowmya Vinayak

Problem Statement and Business Objective

- X is an online course provider
- X advertises its course in various platforms and get leads via various platforms
- However, X has difficulty in converting all the leads. For eg: only 30 out of 100 leads are converted
- To make the conversion rate more efficient, X needs to know the 'Hot leads' who has more possibility to get converted
- This will enable better sales for X and sales and marketing team will involve in converting these 'Hot leads' in a better way than all the leads.
- Hence, X wants to build a model to identify most potential leads and deploy the model for better sales.

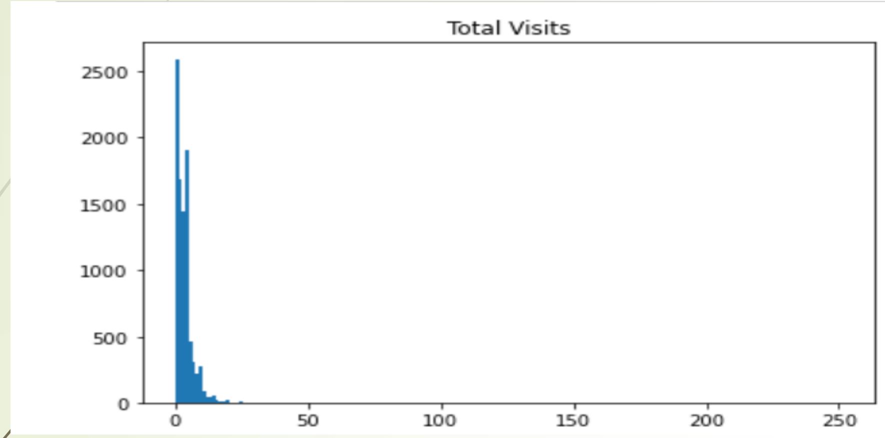
Methods used for solution

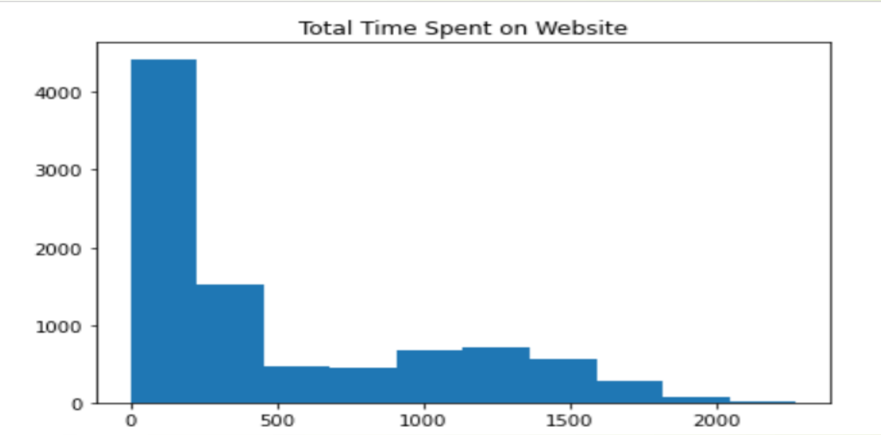
- Data cleaning and manipulation
  - Imputing or removing missing data
  - Removing the columns which has lesser impact
  - Dropping columns with high imbalance
  - Combining columns having low percentage into single column
  - Outlier handling
  - Checking duplicates
- EDA (Univariate and bivariate analysis)
- Model building using logistic regression
- Model evaluation
- Conclusions

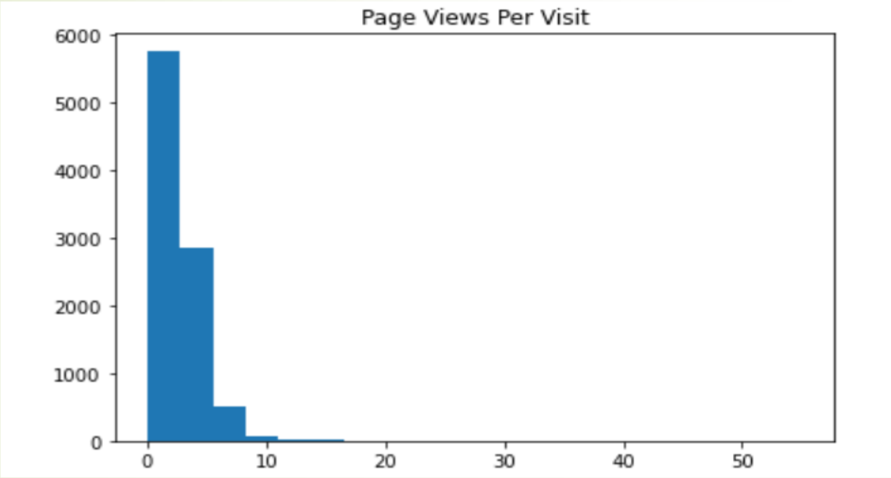
Data Cleaning and Manipulation

- Replacing all the data which has ‘select’ with ‘nan’
- Dropping ‘Specialisation’, ‘How did you hear about X education’ and ‘City’ columns as the missing values are more than 35%
- Checking unique values in categorical values.
- Dropping columns with high imbalance
- Combining columns having low percentage into single column
- Replacing null values mode for categorical variables and mean or median for numerical variables

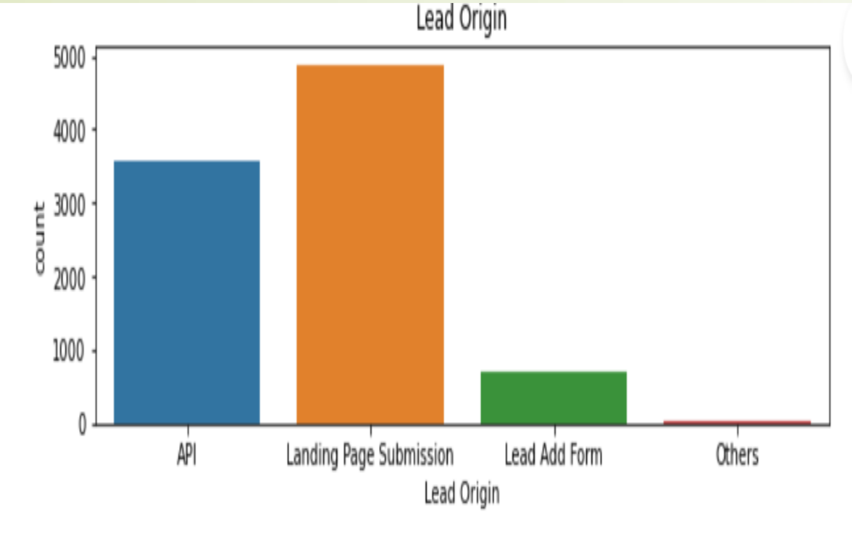
EDA

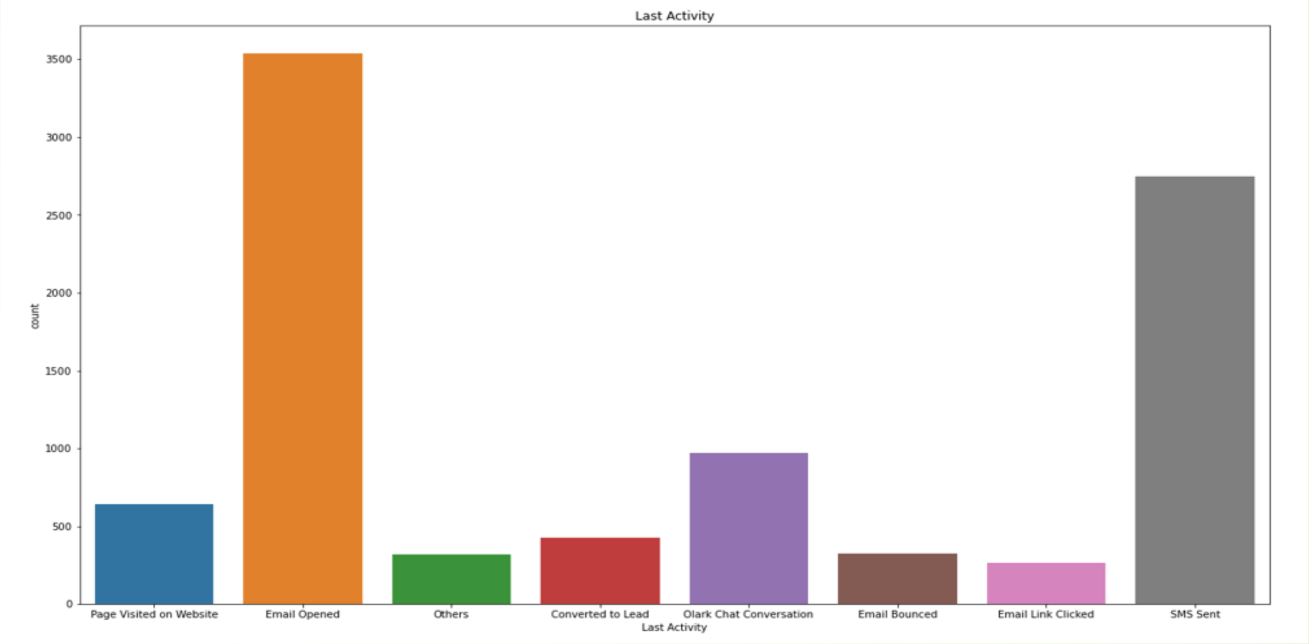


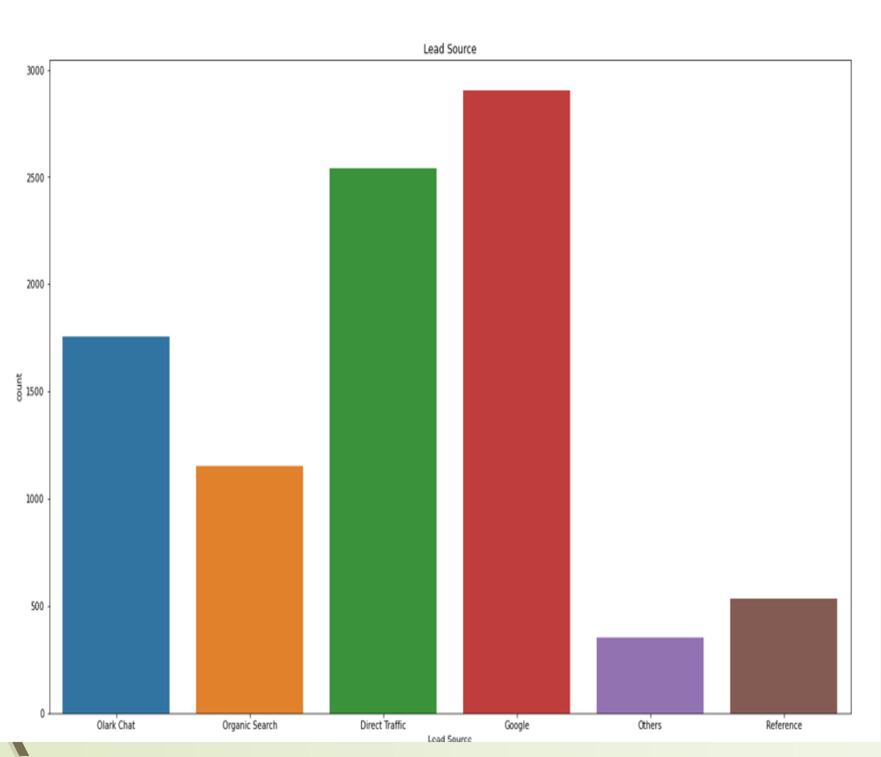




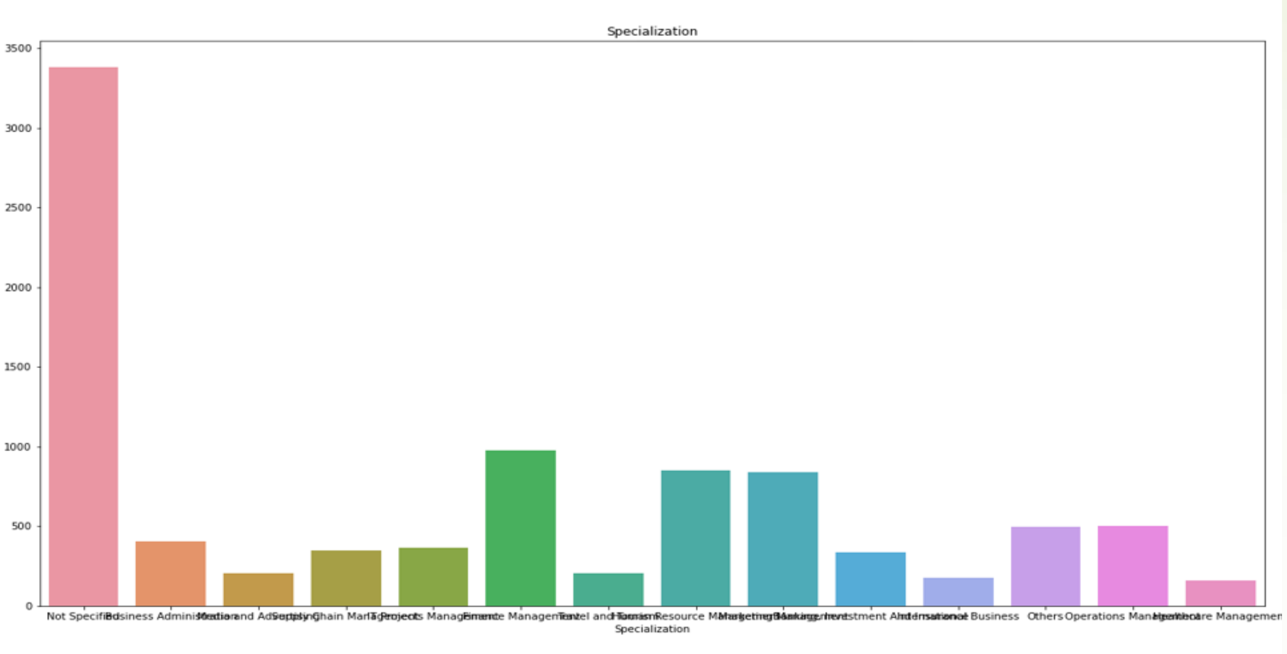
- Total visits increases but decreases later
- Max probability for Pageviews per visit is between 3-5
- Probability of time spent on website is around 0- 300 seconds

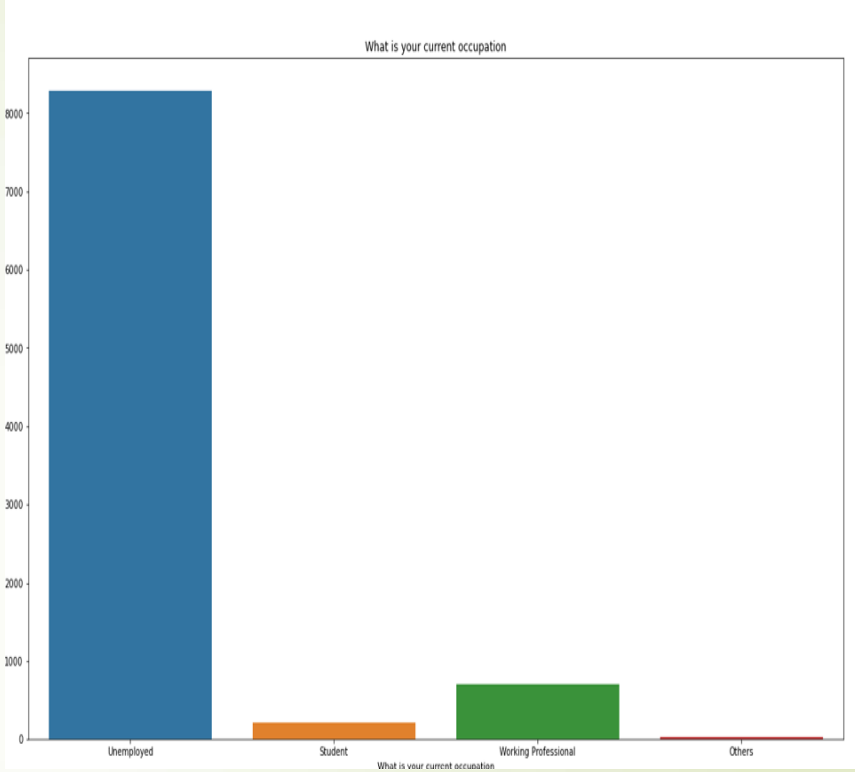


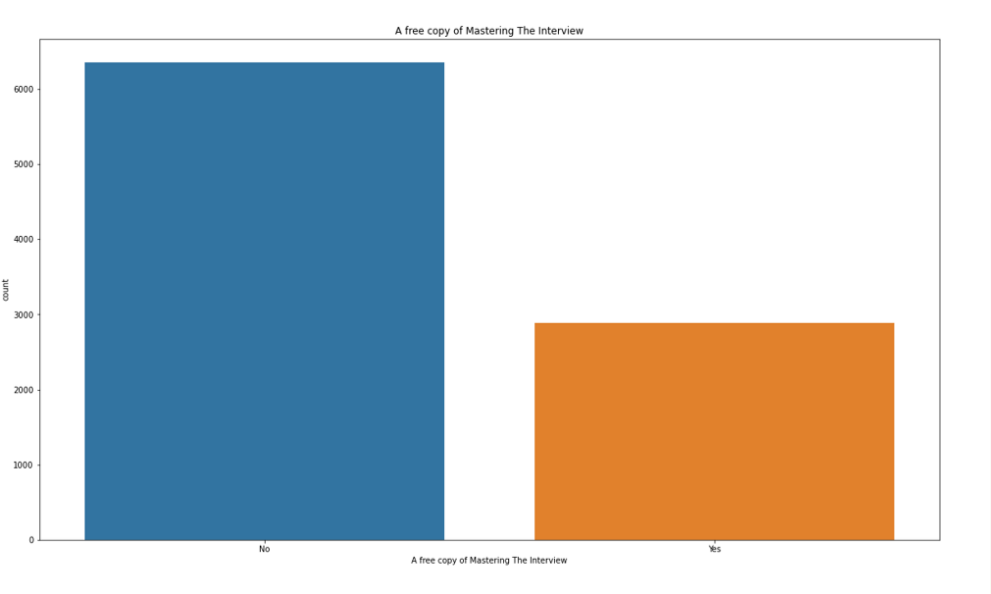




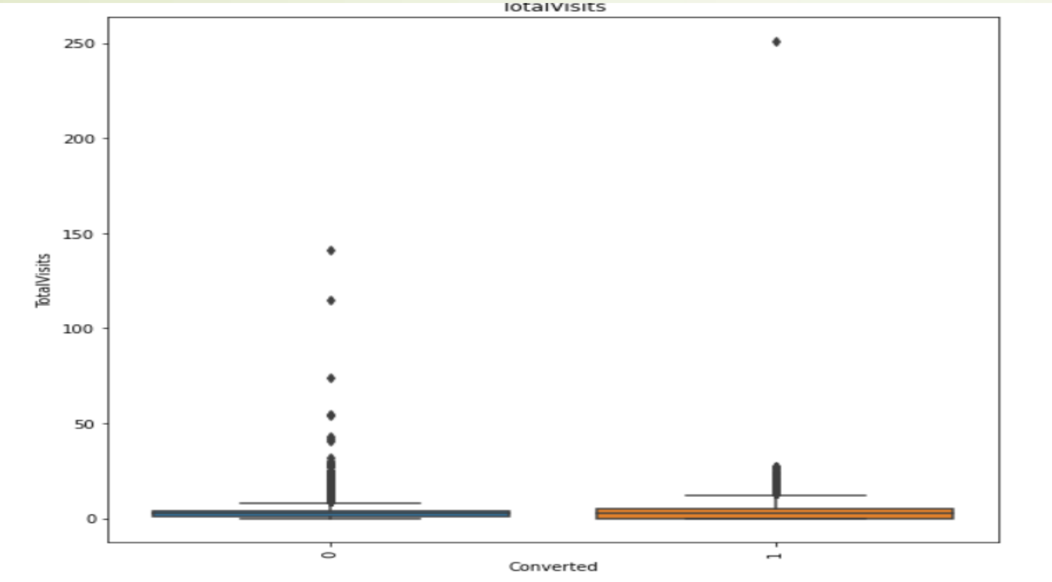
- Landing page submission has better conversion
- Google is the most important source for lead conversion
- Probability of conversion is better when communicated via email or sms

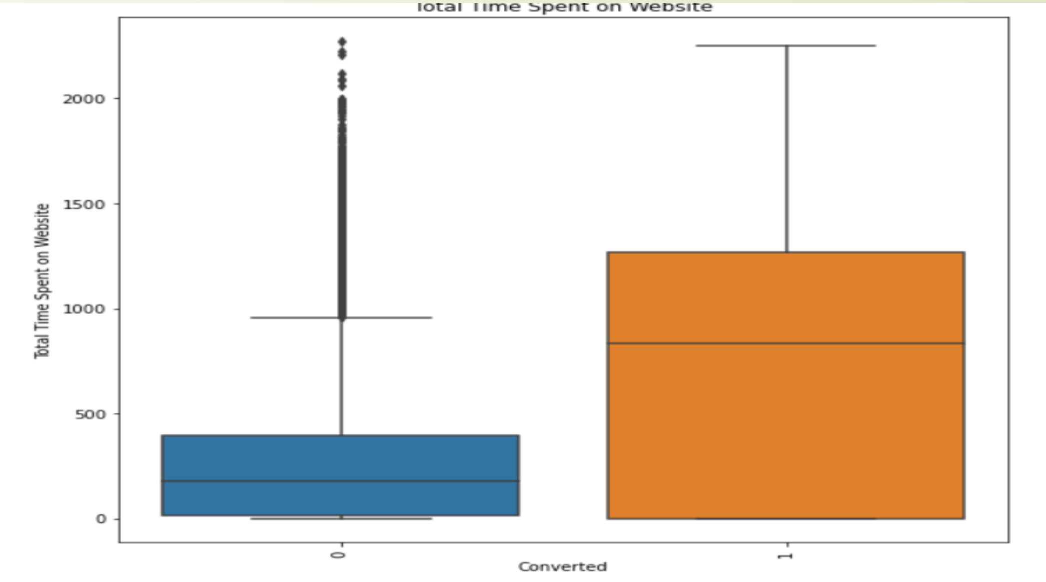


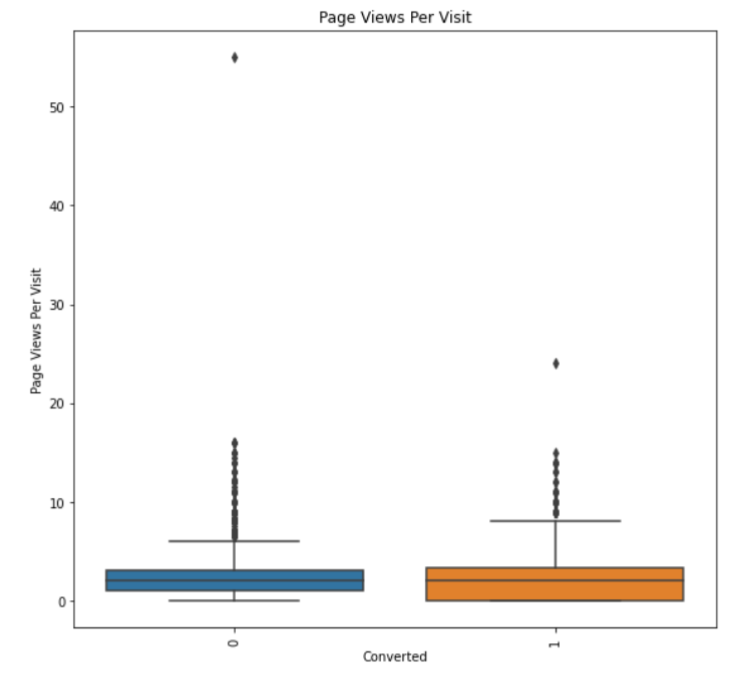




- More on target should be on unemployed and working professional for better conversion





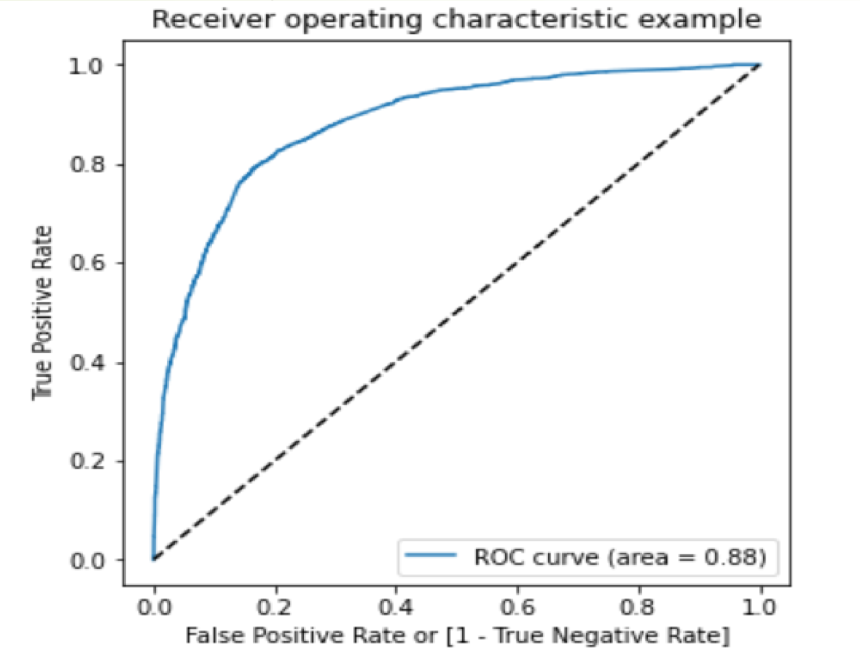


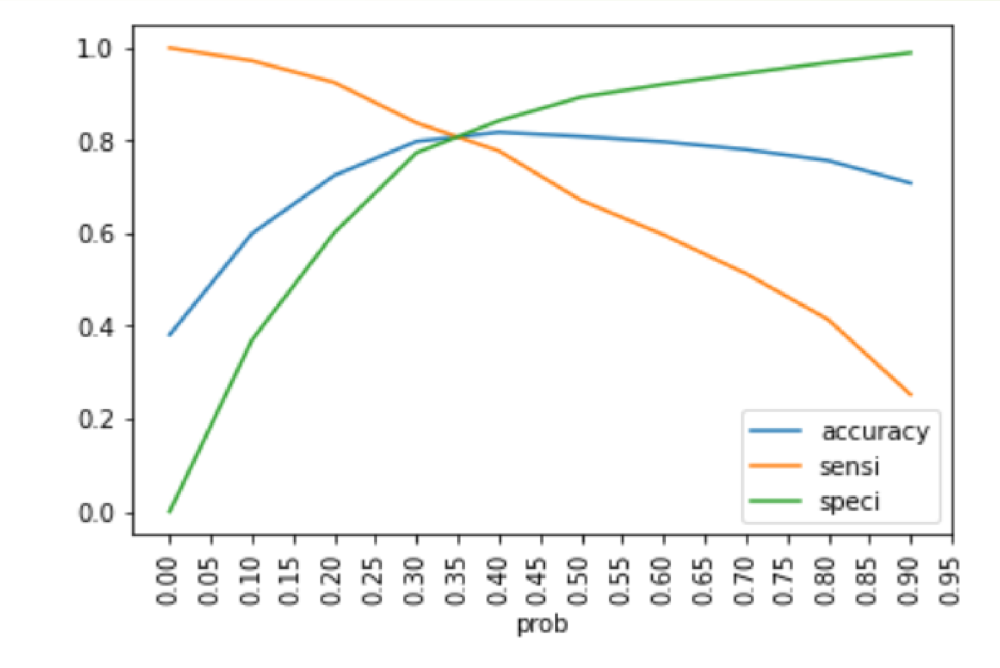
- The mean is found to be higher in case of Converted people rather than non-converted people.
- The average total visits for both converted and non converted people is found to be the same.
- The average page views for both converted and non converted is found to be the

Model Building

- Data is split into train and test set
- The regression is performed on train – test split which is considered at 70-30 in our model
- Before building the model, conversion rate is 38.5%
- RFE is built with 14 variables for final output
- Model is built whose p-variable is greater than 0.05 and vif is greater than 5
- Final accuracy is 81.19%

ROC Curve





- From the second point, it is observed that the optimal point is at 0.35