Lead Score Case Study

GROUP MEMBERS

1.MAYUR INGOLE

2.RABINDRA PRATAP HOTA

Problem Statement

- X Education sells online courses to industry professionals
- 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 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.

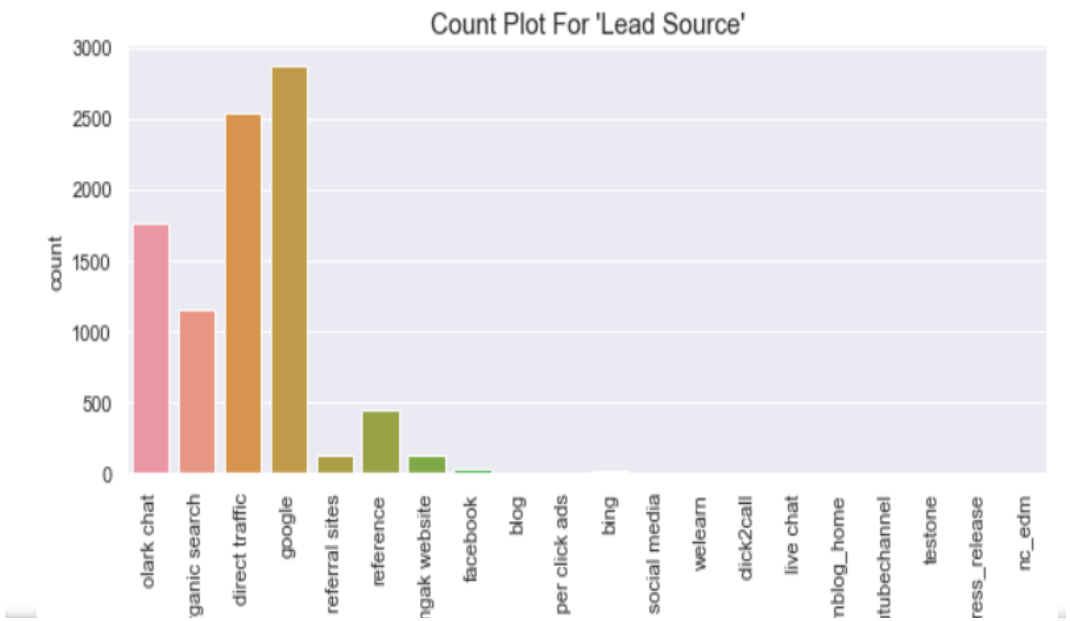
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

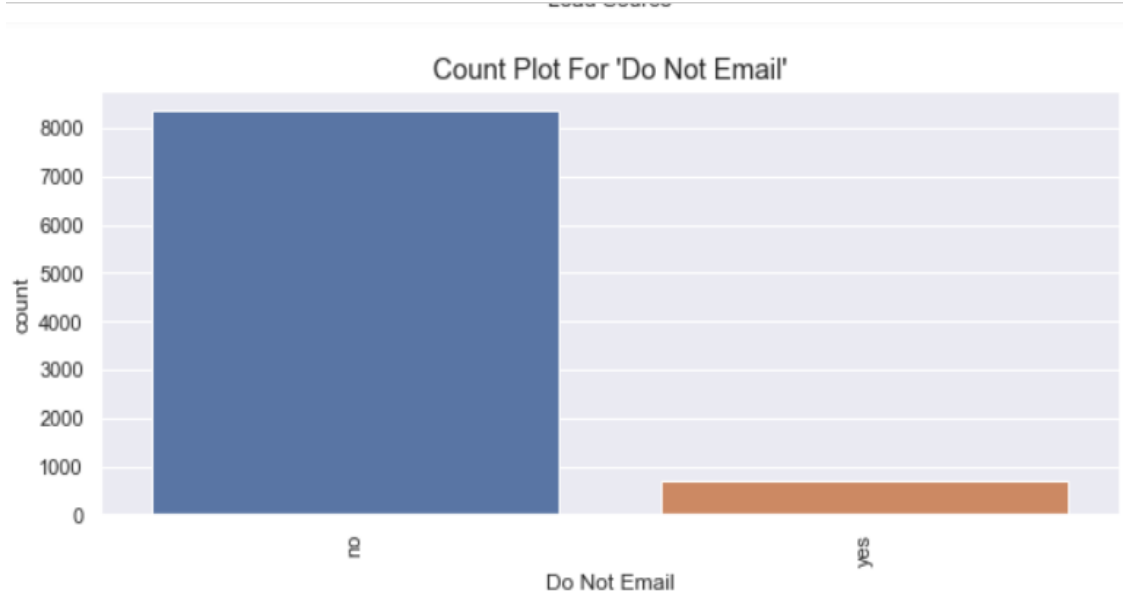
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

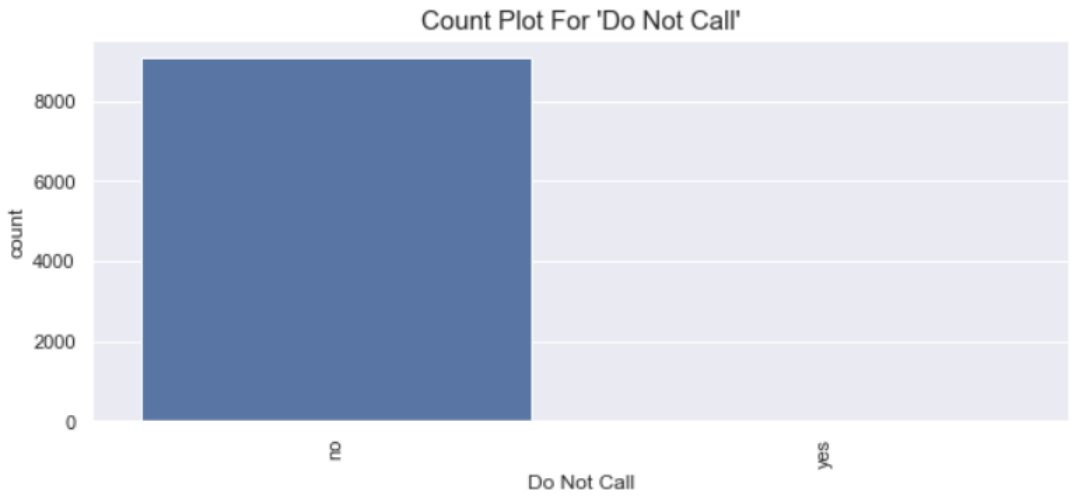
Data Manipulation

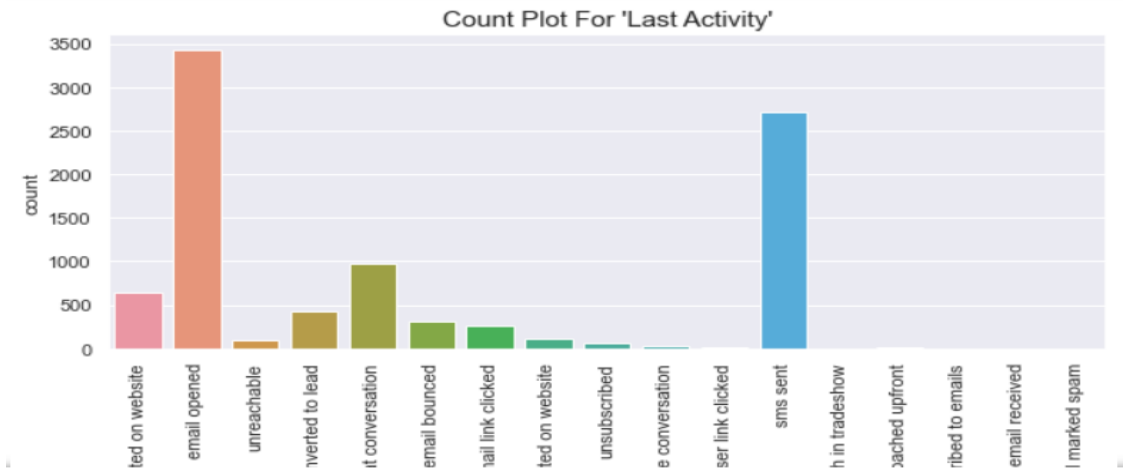
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”.
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

EDA - Categorical

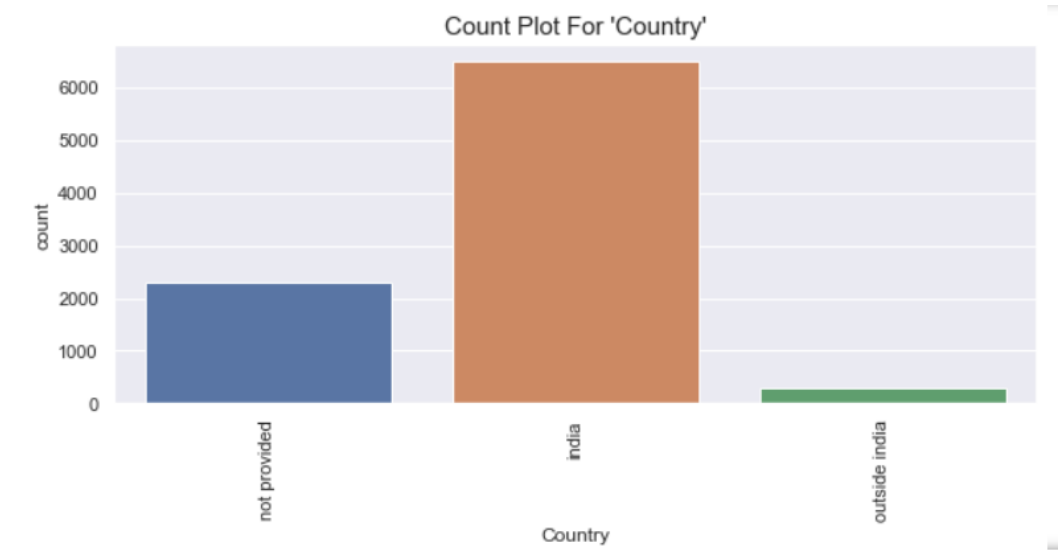


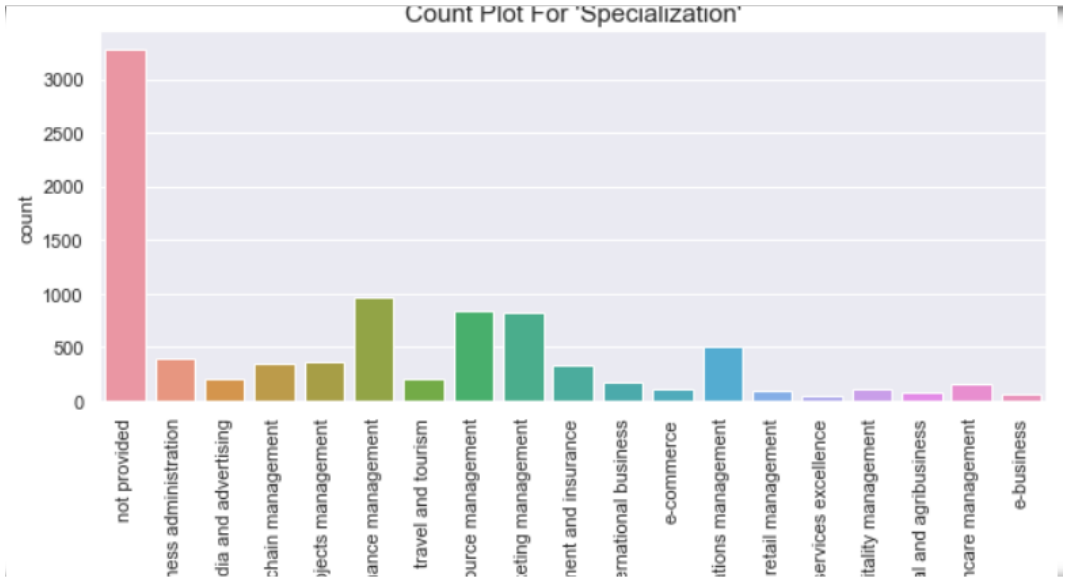


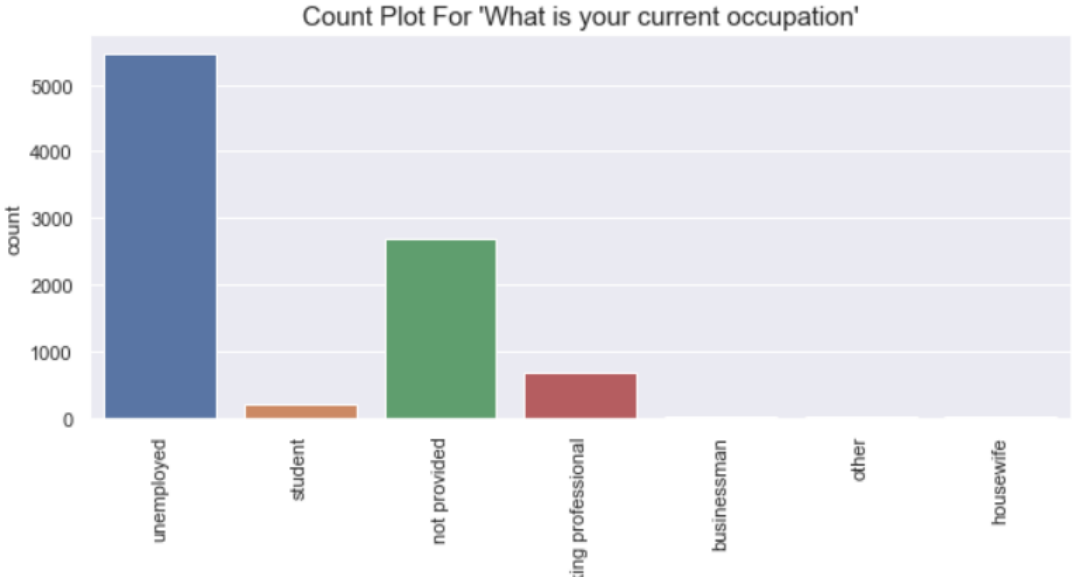


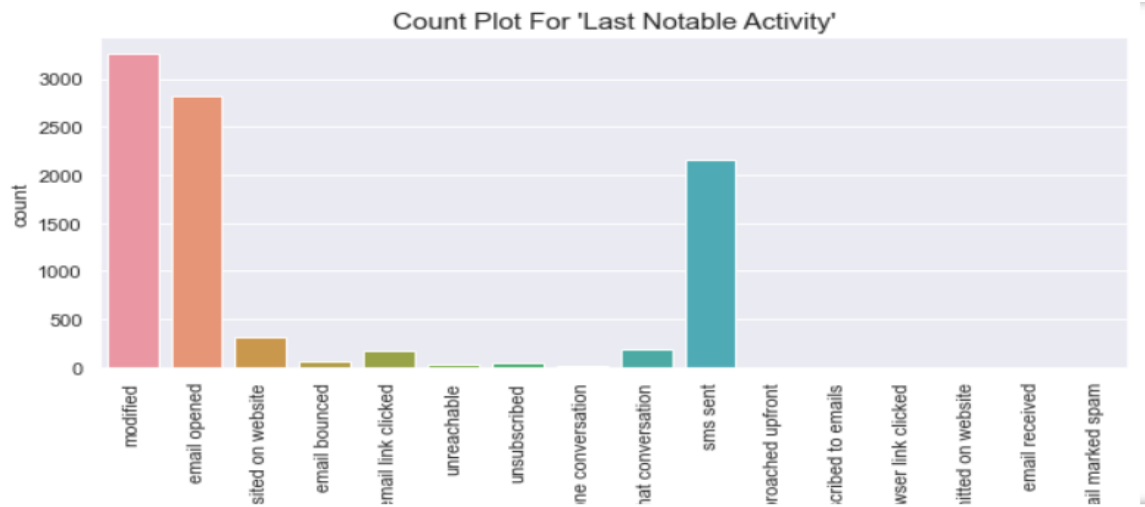


EDA - Categorical

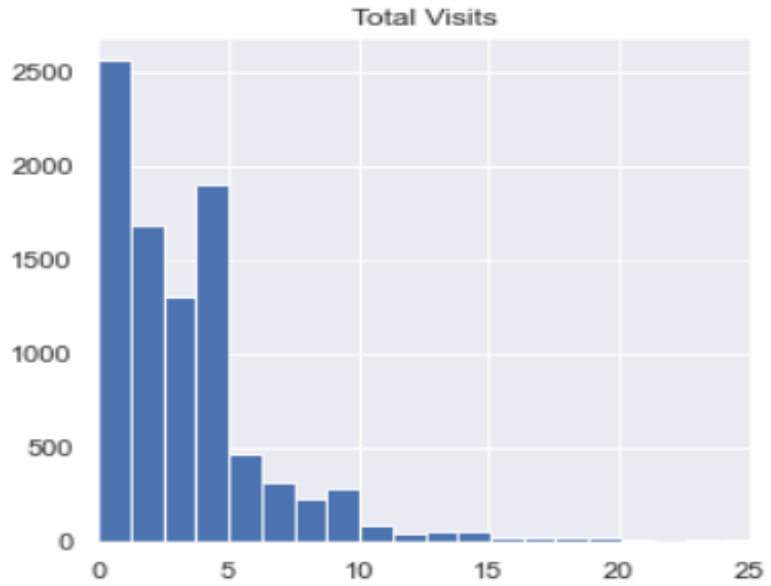


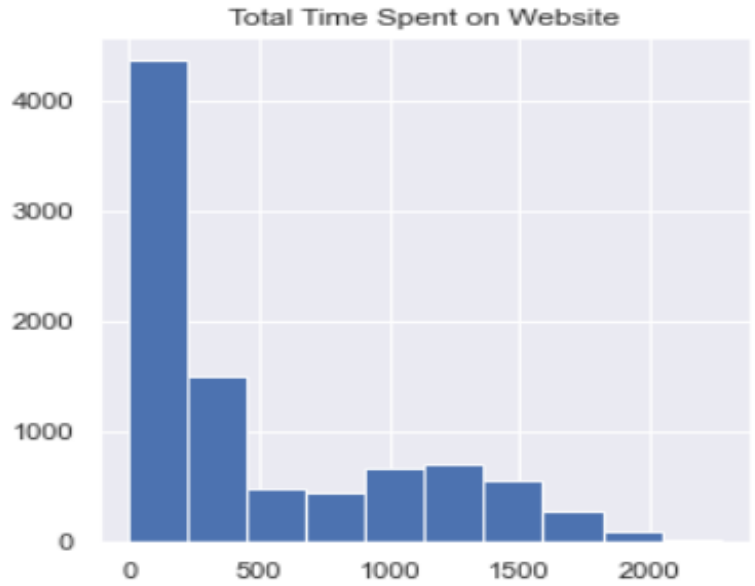


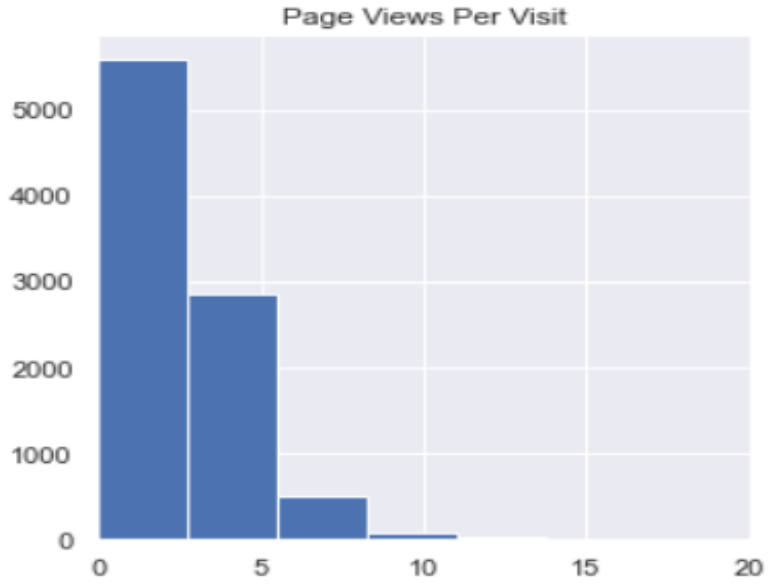




EDA- Numerical







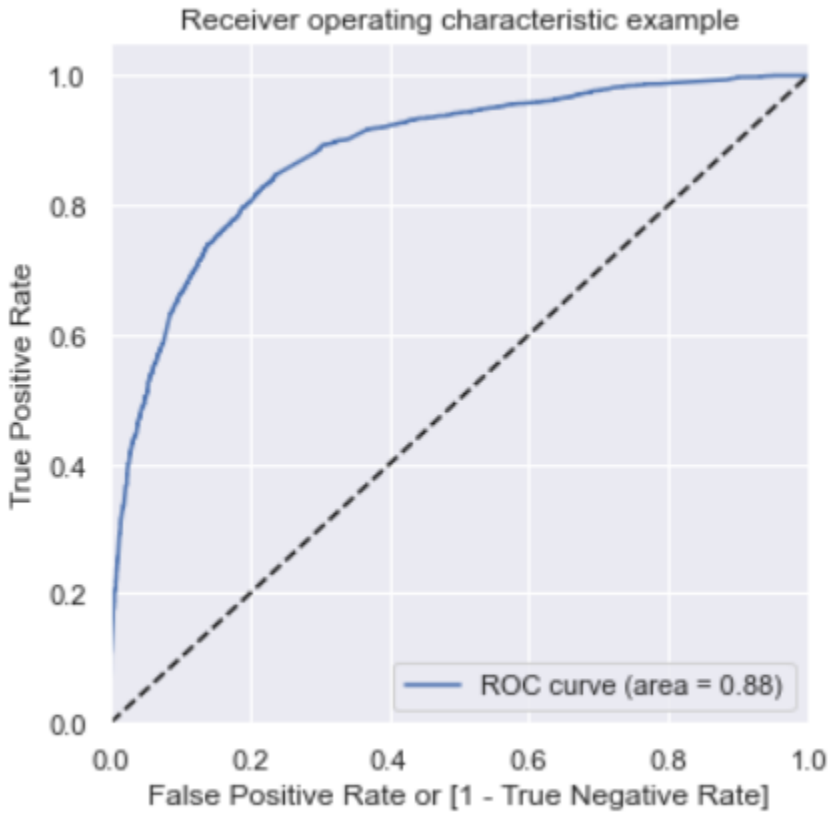
Data Conversion

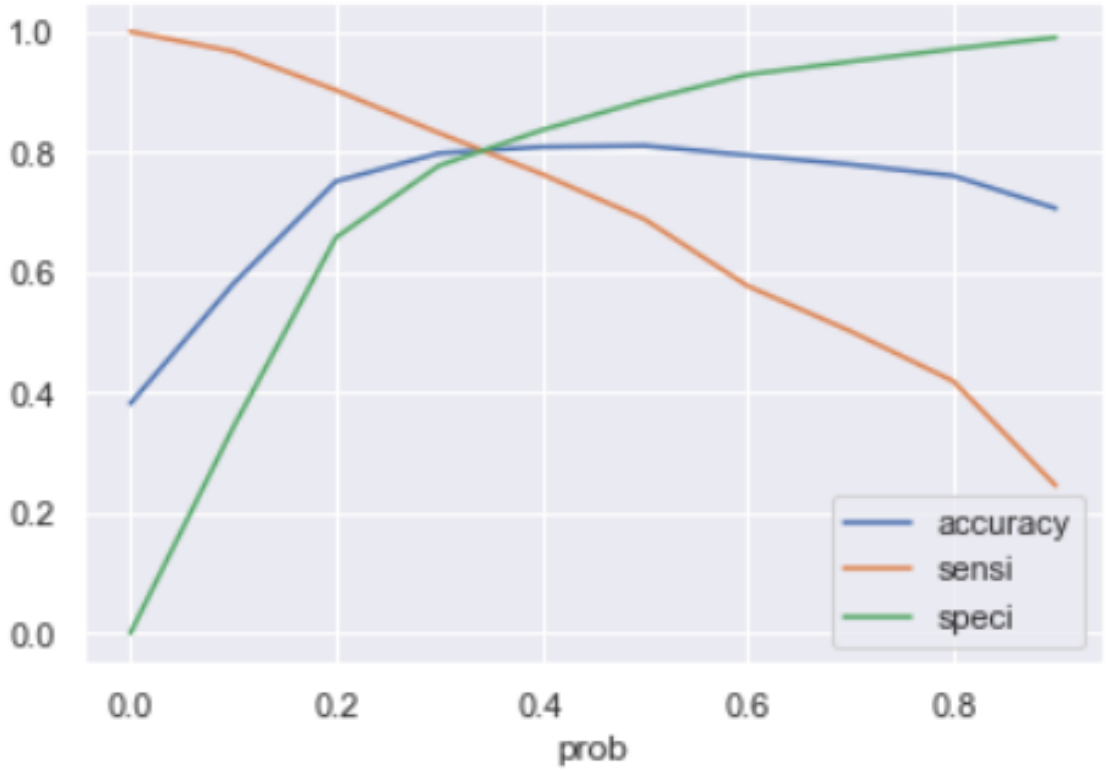
- Numerical Variables are Normalised – MinMaxScaler
- Dummy Variables are created for object type variables
- Total Rows for Modelling Analysis: 9063
- Total Columns for Modelling Analysis: 83

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for 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 vifvalue is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC CURVE





- 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.

Conclusion

The variables that mattered the most in the potential buyers are:

- The total time spend on the Website.
- When the lead origin is Lead add format
- When their current occupation is as a working professional
- Total number of visits.
- When the lead source was:
    a. Google
    b. Direct traffic
    c. olark chat
    d. Welingak website
- When the last activity was:
    a. Email Bounced
    b. Olark chat conversation
- When the last notable activity was:
    a. Email linked clicked
    b. Email Opened
    c. Olark chat conversation
    d. Activity modified
    e. Page visited on website