Lead Score Case Study

- Muruganand Kumaran & Karthik Veluchamy Sarguru

Problem Statement

- X Education sells online courses to industry professionals.
- Company gets lots of enquiries and leads for courses but only 30 percentage is conversion rate.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If hot leads are identified correctly then the conversion rate might go up.
- In this way company wants to communicate only with hot leads instead of calling and enquiring everyone.

Solution Approach

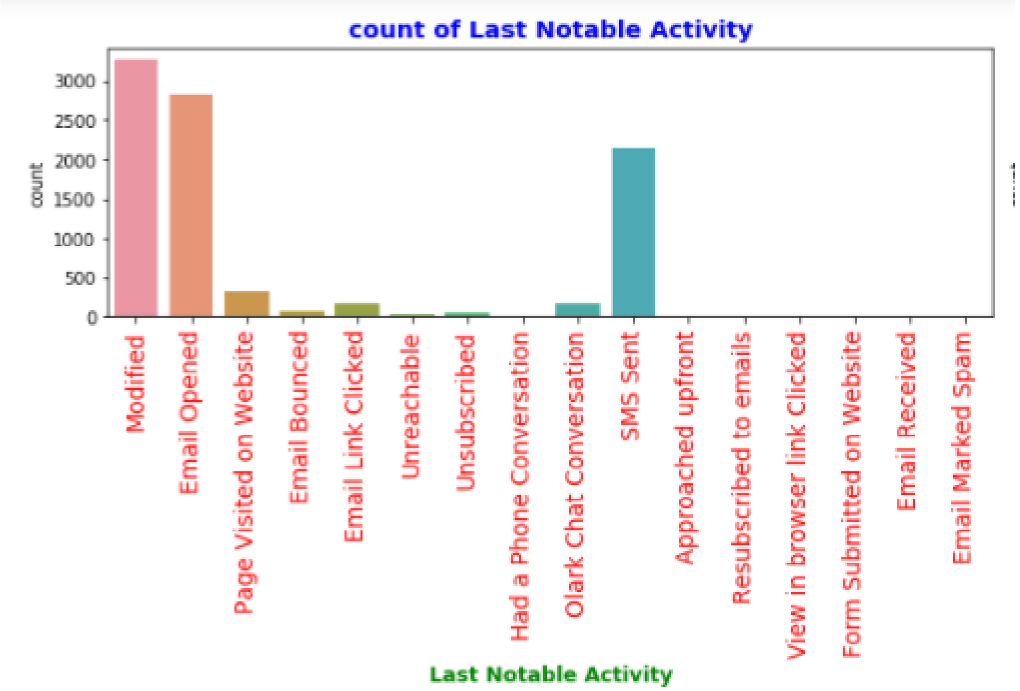
- Data Cleaning and Modification
  - Remove columns with just one value.
  - Check the null percentage in column and drop those which are above 35%.
  - Handle null values in categorical data by assigning new category “no values”.
  - Check and handled outliers by considering data only up to 95 percentile, as there was a large variation from 95 to higher percentile(outliers).
- EDA
  - Univaraiate analysis was done by plotting counts of features to check the distribution.
  - Bivariate analysis was done by plotting all the features across Converted to see how the distribution of conversion was across features.
- Model Building
  - 70:30 split was done for Train and Test set
  - Numerical features were scaled using Standard Scalar.
  - Logistic Regression was used and RFE was done to select top 15 features and then manual method was used to drop features with high VIF and p value.
  - Based on Accuracy, Sensitivity, specificity arrived at optimal cut off value for the model to decide on Conversion.
  - Above parameters were derived from Confusion matrix.
  - Final model performance was evaluated by checking the accuracy, sensitivity , specificity for both Test and Train data set to be around 80%.

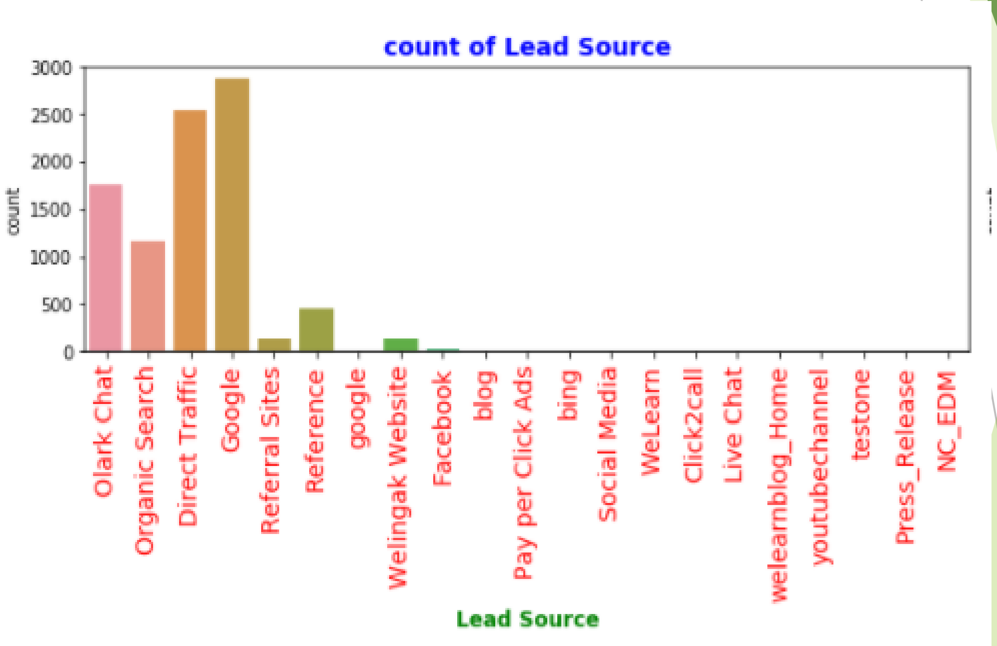
Data Cleaning and Manipulation

- Columns with values as “select” as assigned as null.
- Single value columns like “Magazine”, “Receive More Updates About our Course”, Update me on supply chain content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” are dropped.
- From value counts dropped below columns as they do not have enough variance to provide any meaningful affect on model
  - Do Not Call,What matters most to you in choosing a course, Search, Newspaper Article, X Education Forums, Newspaper, Do Not Email, Digital Advertisement, Through Recommendations
  - Dropped columns with null value greater than 35 percent.

EDA

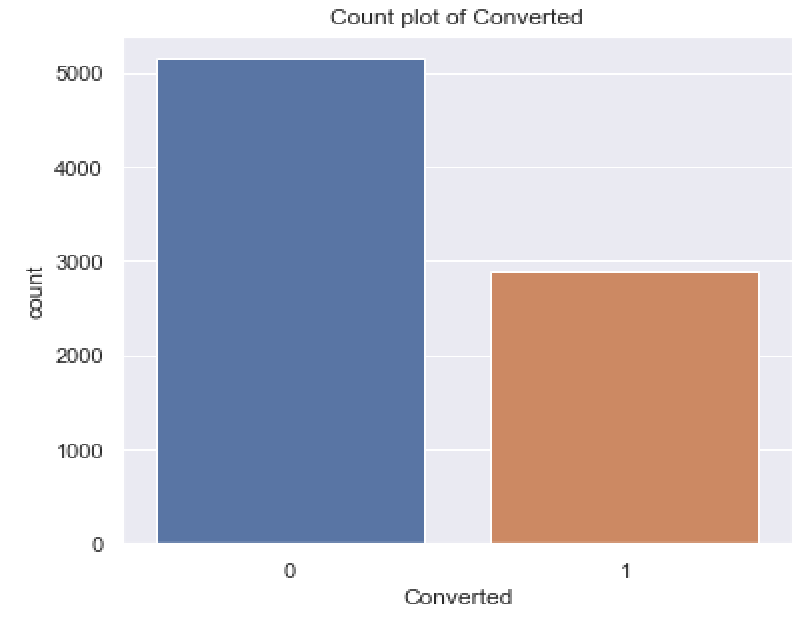
- Plot of Count of Last Notable Activity and Lead Source provides information on how the data is spread, these features were converted to dummies and used in modelling, as they have enough variance.

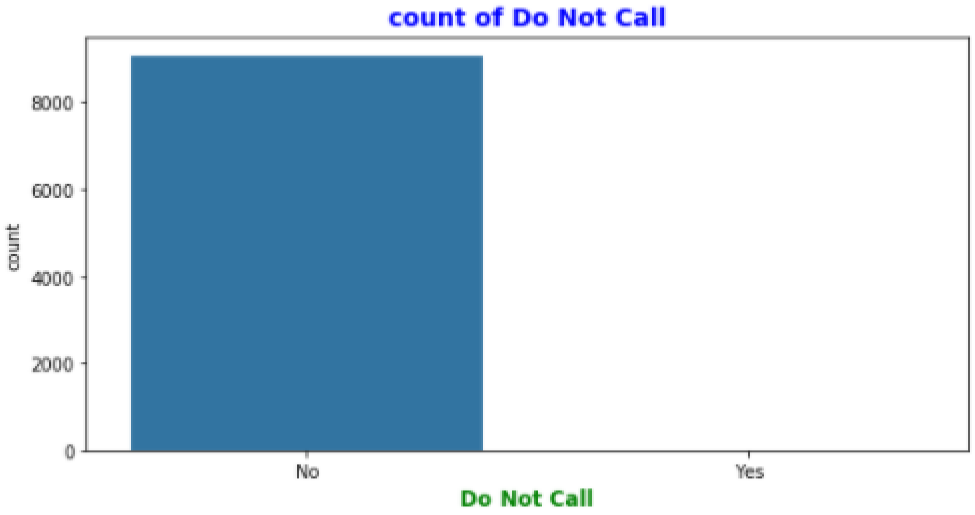


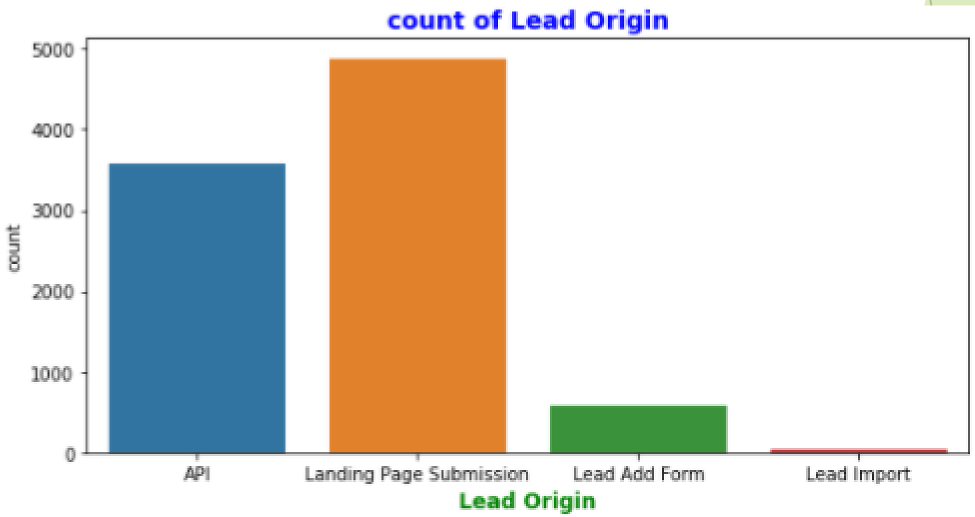


EDA Continued

- Converted ratio is very less as seen in plot
- Similar to “Do Not Call” column many columns with less variance were dropped.
- Only columns with variance similar to “Lead Origin” were retained for modelling and dummies were created for them

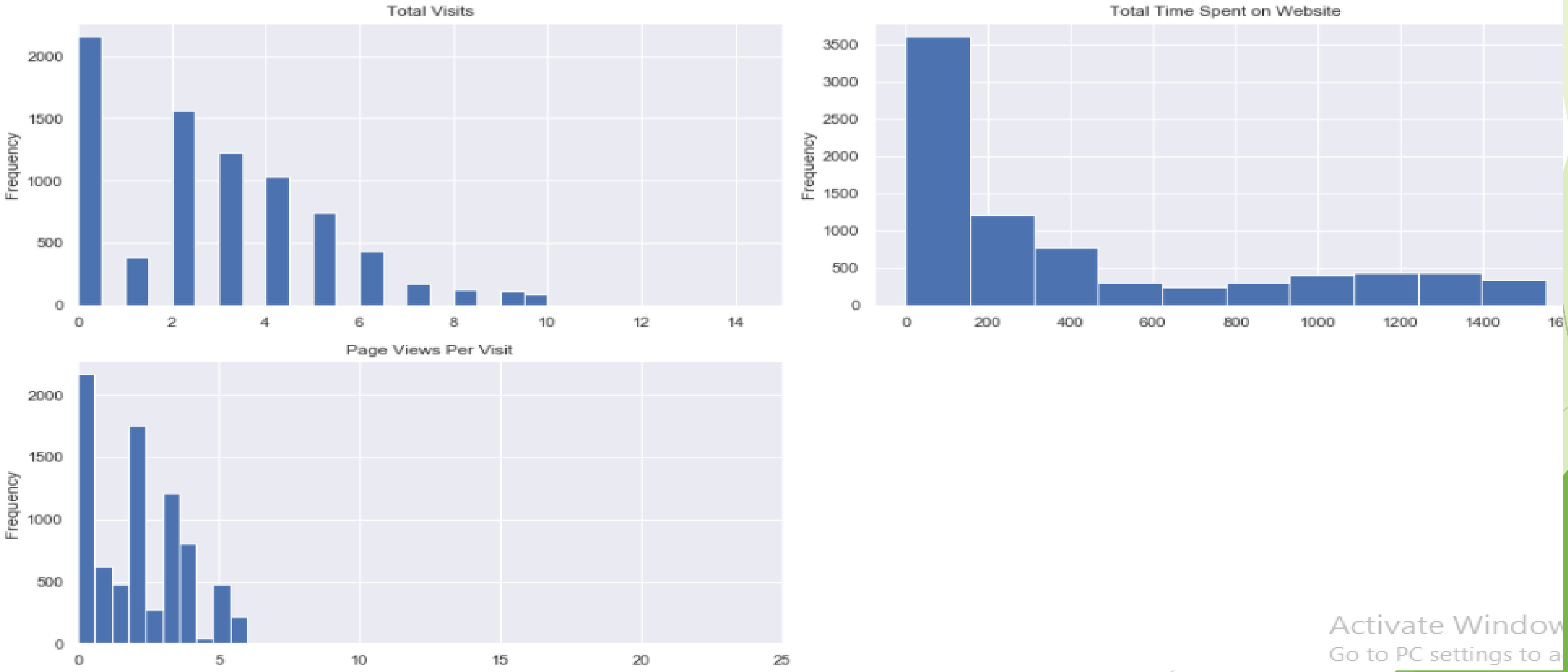






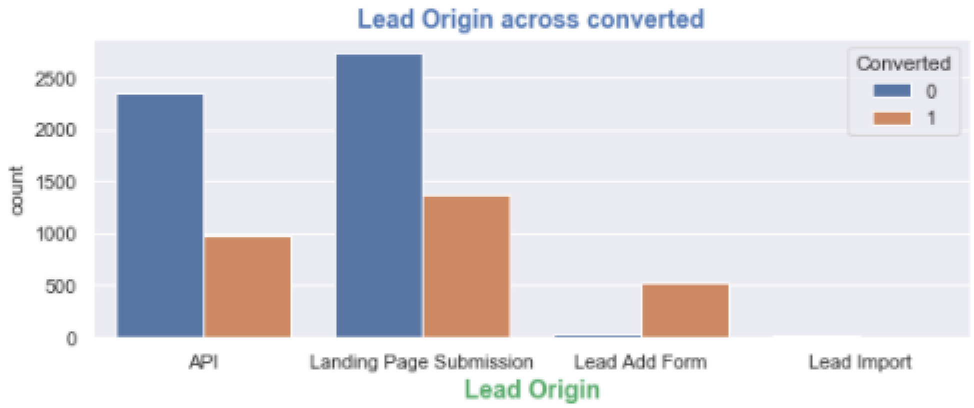
EDA of Numerical Columns

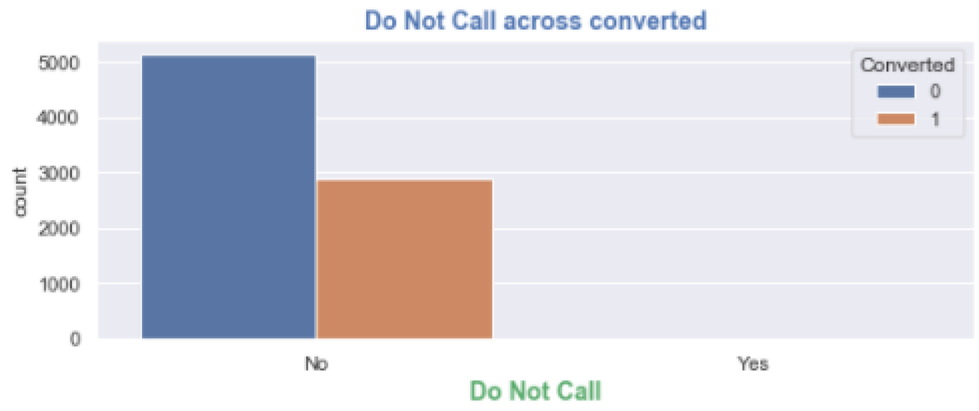
- Numerical columns show the data is spread and its not accumulated at a point and hence offer better prospect in deciding the outcome of Model.
- Total Visits, Total Time Spent on website, Page views per visit are columns used.

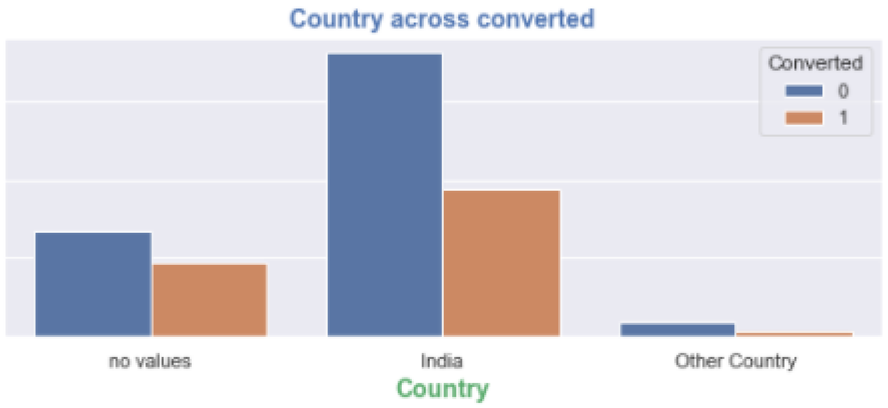


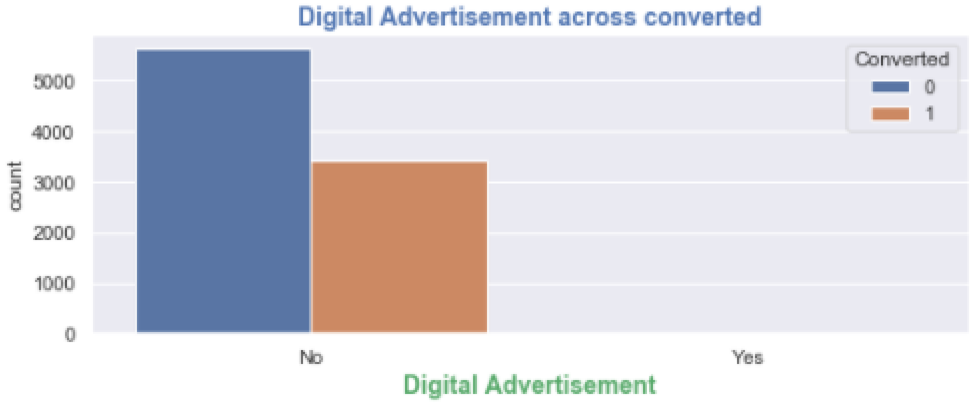
Relationship between Categorical and Converted

- Do Not Call and Digital Advertisement shows previously proved point of less variance
- Lead Origin and Country plotted across converted show good amount of variance which would be usefull.

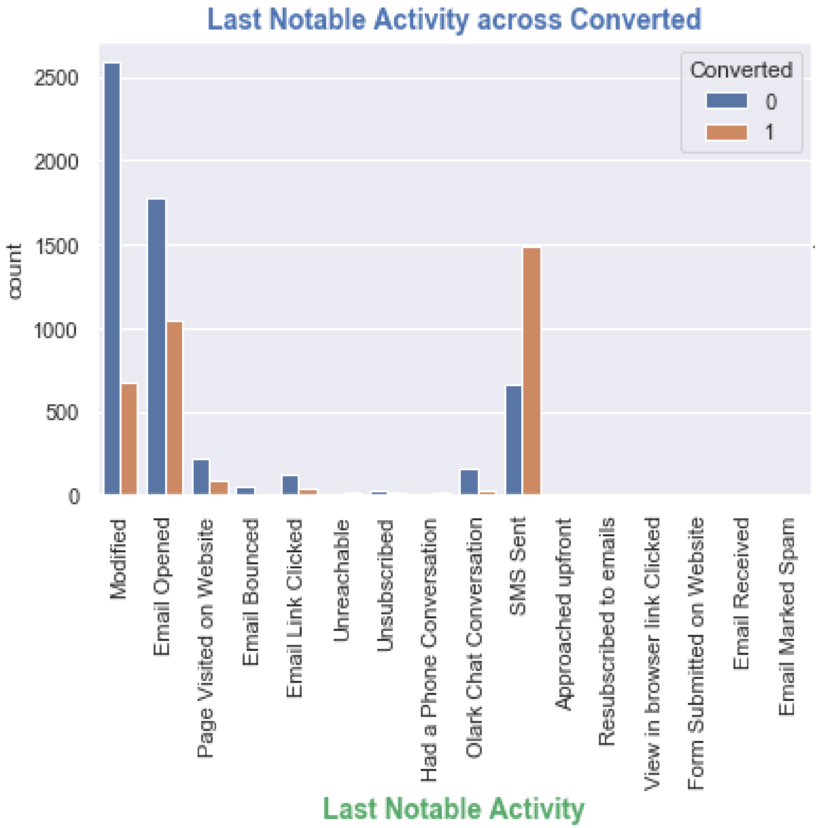

Lead Origin across converted

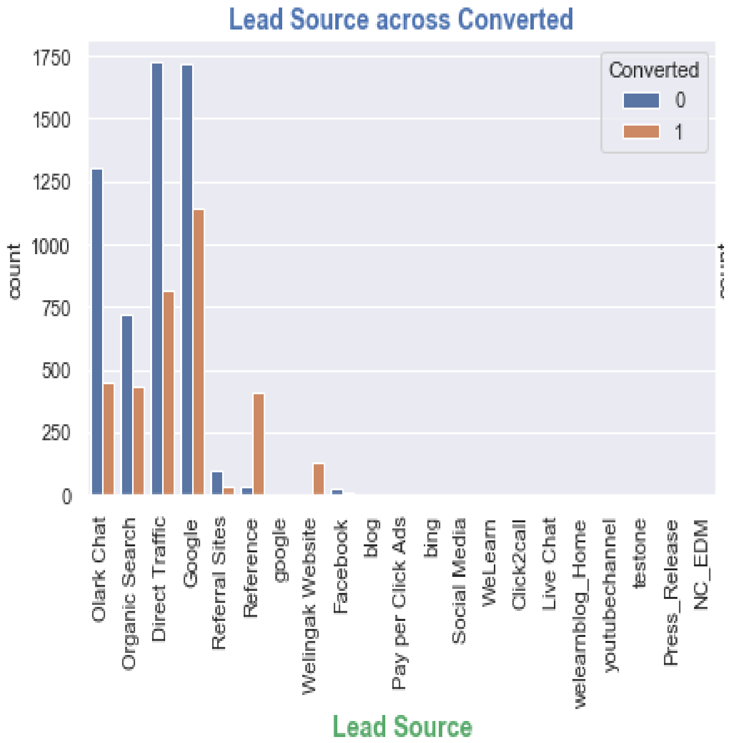

Do Not Call across converted

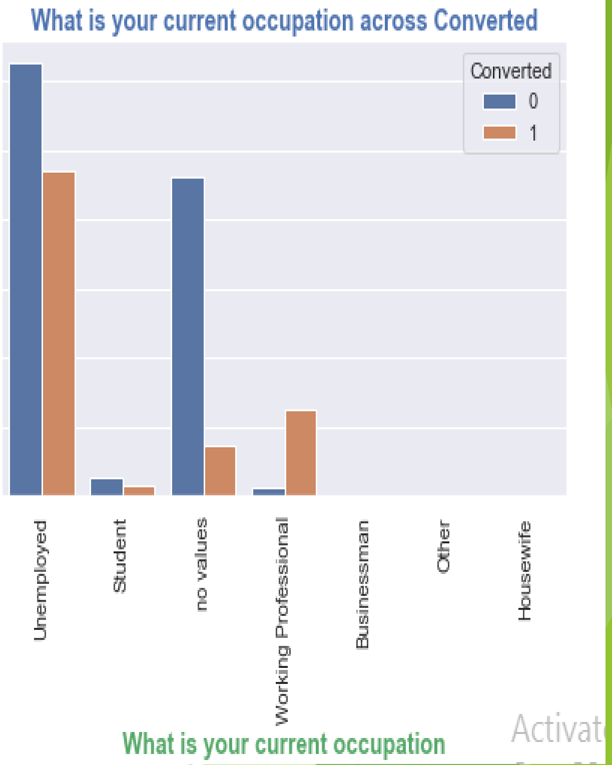

Country across converted


Digital Advertisement across converted

- Last Notable Activity- we can see high conversion rate for SMS Sent and Email opened
- Lead source – Google, Organic chat has higher conversion.
- Unemployed and Working professional have higher conversion.

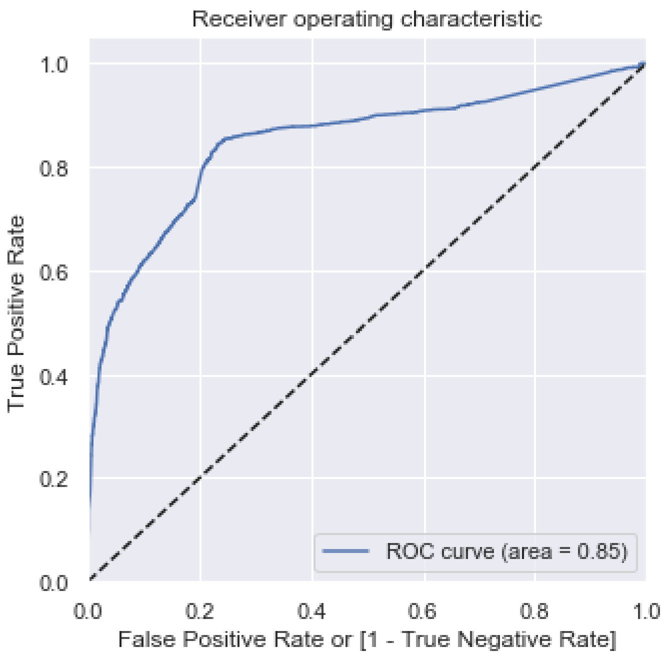


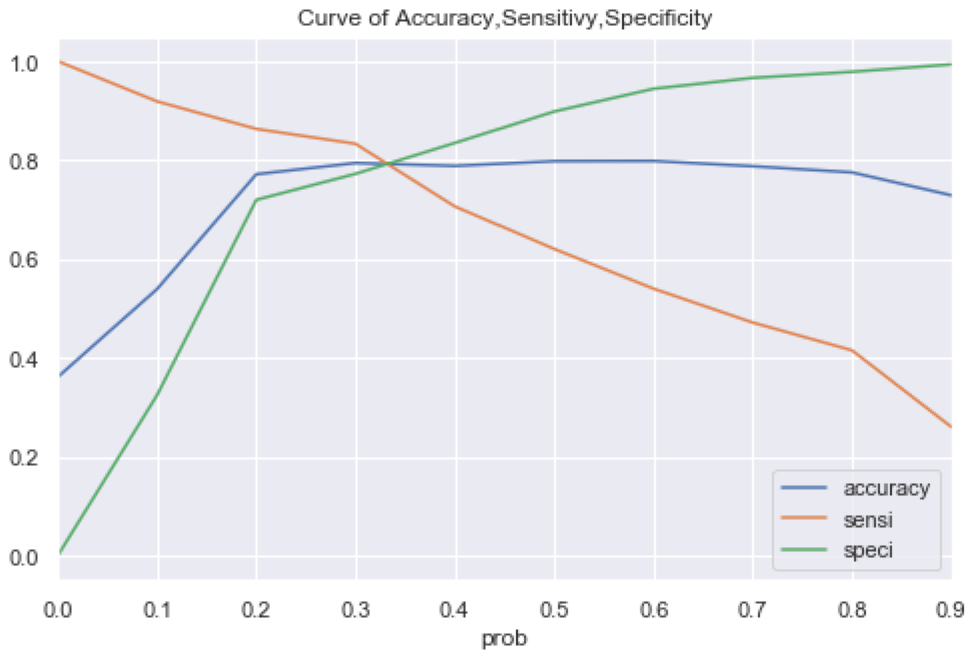




Model Building

- After splitting of test and Train data set.
- RFE used to get top 15 and manual process followed to get less than 10 features , features dropped on VIF (correlation) and p values.
- Cut off for probability was derived based on curve below where Sensitivity, specificity, Accuracy are balanced.
- 0.33 was choose as Optimal cut-off.
- Accuracy of 80% was achieved.
- ROC curve proves model is good as its way above median line.





Suggestions to improve Conversion

- The Variables that should be looked upon to improve conversion rate
- Total Time Spent on Website
- Lead Source
  - Welingak Website
- Lead Origin
  - Add Form
- Last Notable Activity
  - SMS Sent
  - Had A phone Conversation
- Current Occupation is Working Professional
- Current City is Mumbai
- People from Other Country
- Keeping all above features X Education can improve there profits and conversion rates.
- Interns could be asked to look on these features to improve conversion rates and when targets achieved use only these features to make a call.