




LEAD SCORING CASE STUDY

By

- Varun Reddy
 - Vaibhav Sharma
 - Ramsai Vajrapu
- 



Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%. Now, this means 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

- Lead X wants us to build a model to give every lead a lead score between 0 -100. So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.



Problem Approach

- Importing and Reading the Data
- Exploratory Data Analysis (EDA)
 - Data Cleaning
 - Categorical Columns Analysis/Handling
 - Numerical Columns Analysis/Outlier Treatment.
- Data Preprocessing for Modeling
 - Dummy Variable Creation
 - Train – Test Split
 - Data Scaling
- Model Building
 - RFE & Stats Models
 - Predictions and Probabilities
 - Testing on Test Data
- Final Observations

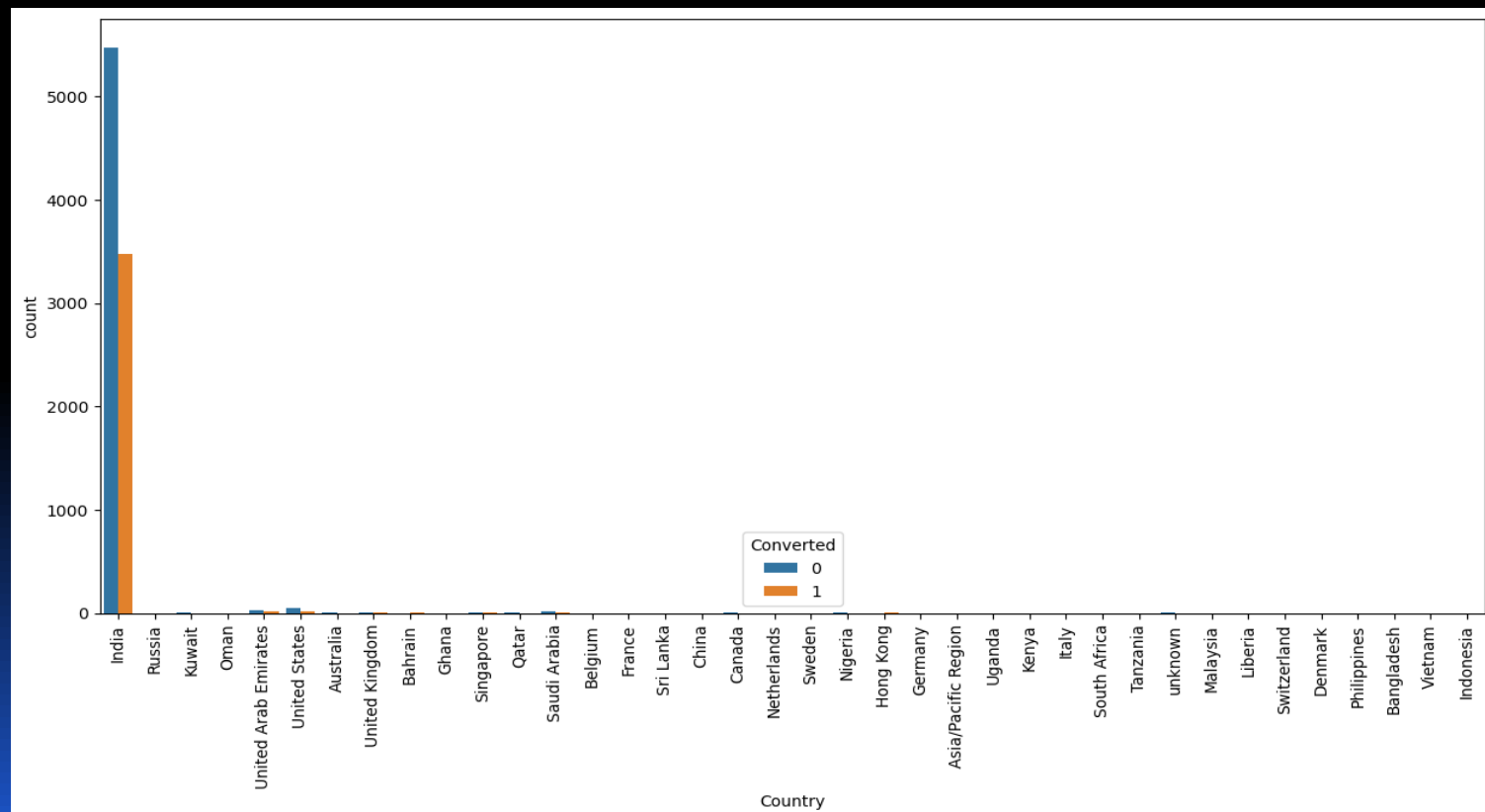


EDA – Data Cleaning

- Dropped 'Prospect ID' and 'Lead Number' as they have unique values across all entries and are used to identify a specific person.
- Updating all the 'Select' values options into Nan.
- Dropped all the columns with more than 45% missing values.

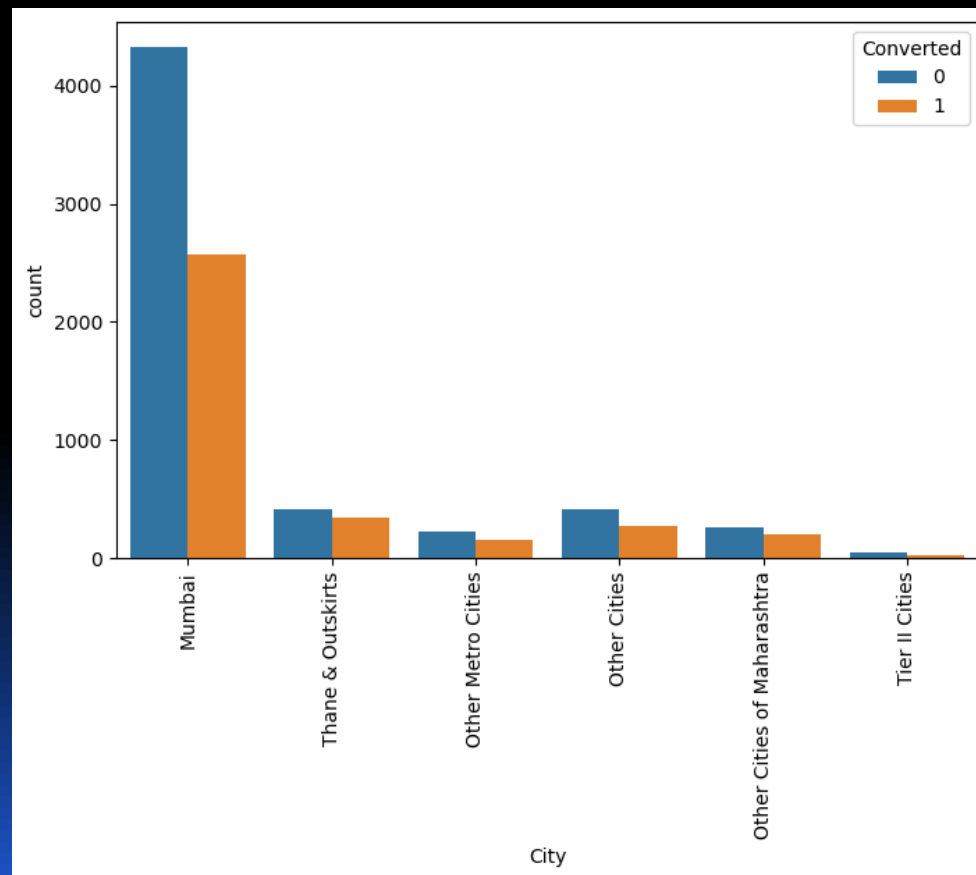
EDA – Categorical Columns

- Country
 - Imputed Nan values with India.
 - Dropped country column as India is the value for around 97% of the entries.



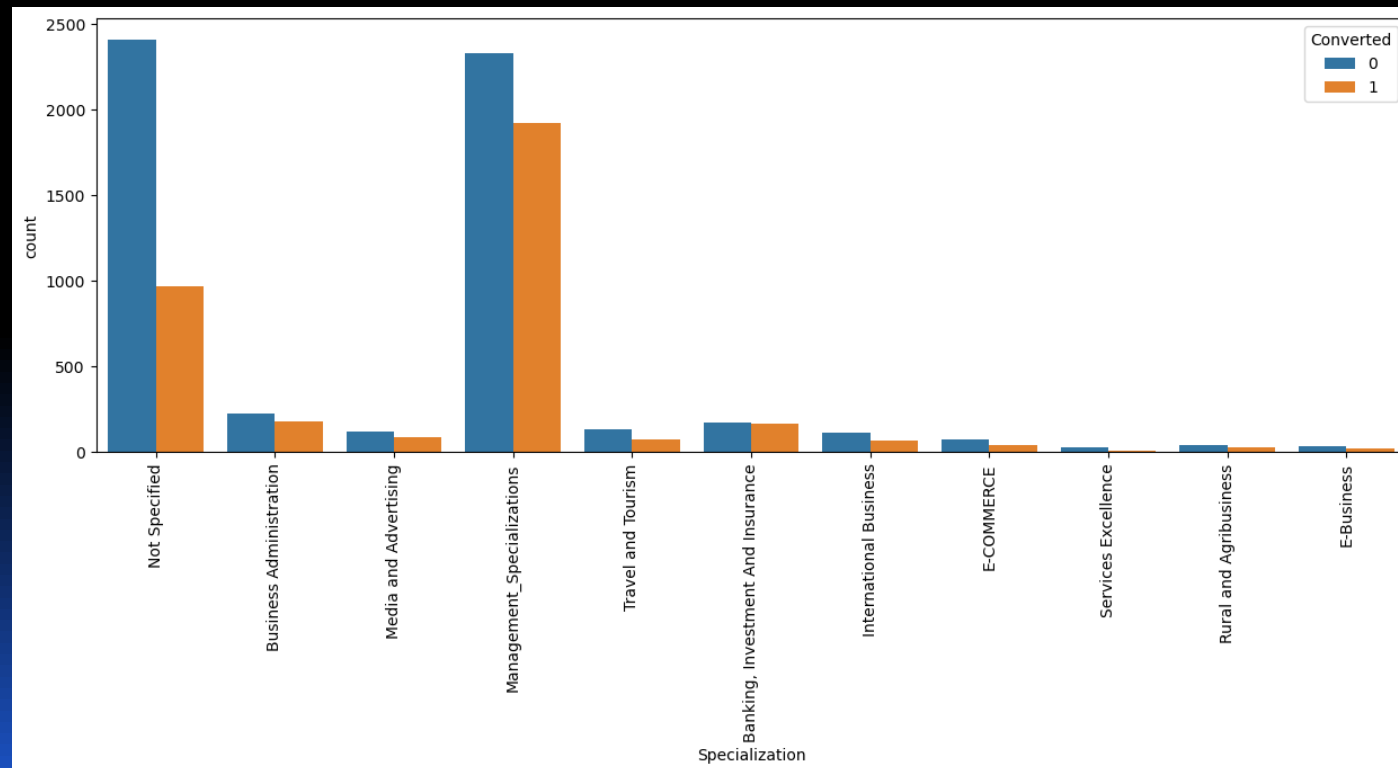
EDA – Categorical Columns

- City
 - ▣ Imputed Nan values with Mumbai.



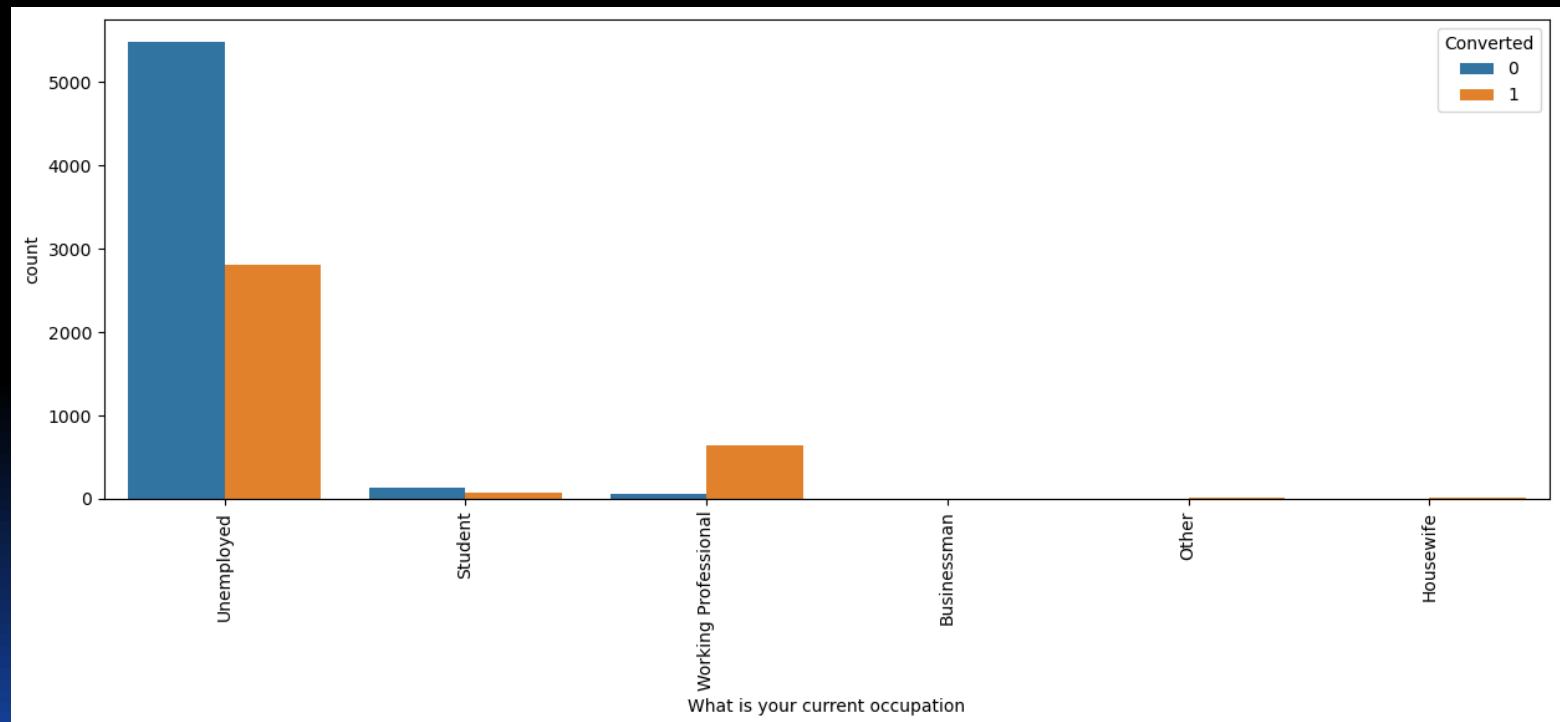
EDA – Categorical Columns

- Specialization
 - Imputed Nan values 'Not Specified'.
 - Merged all the management type categories into one category as 'Management_Specializations'.



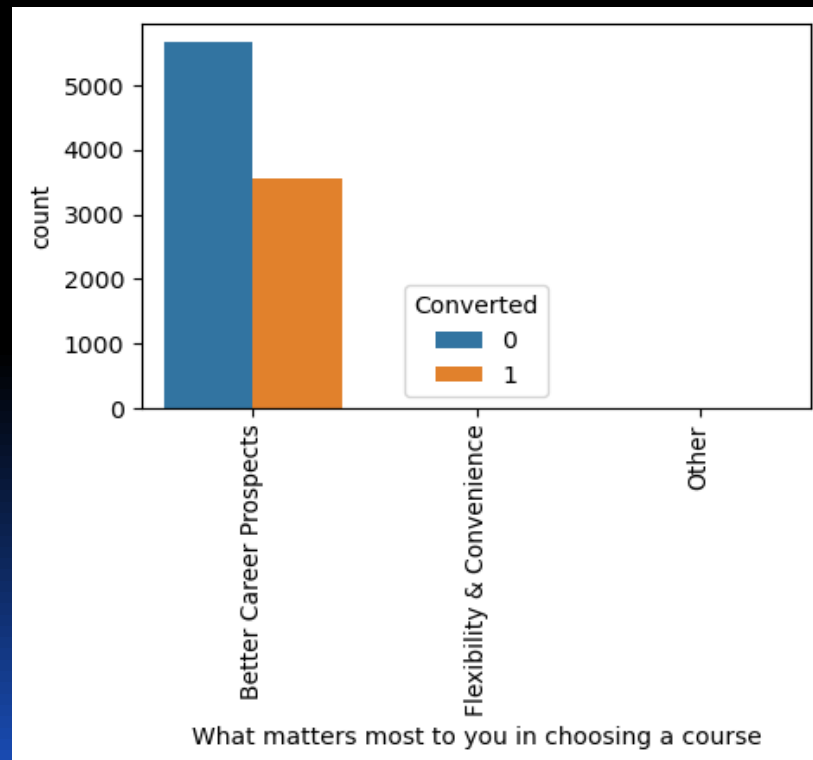
EDA – Categorical Columns

- What is your current occupation
 - Imputed Nan values as 'Unemployed'.



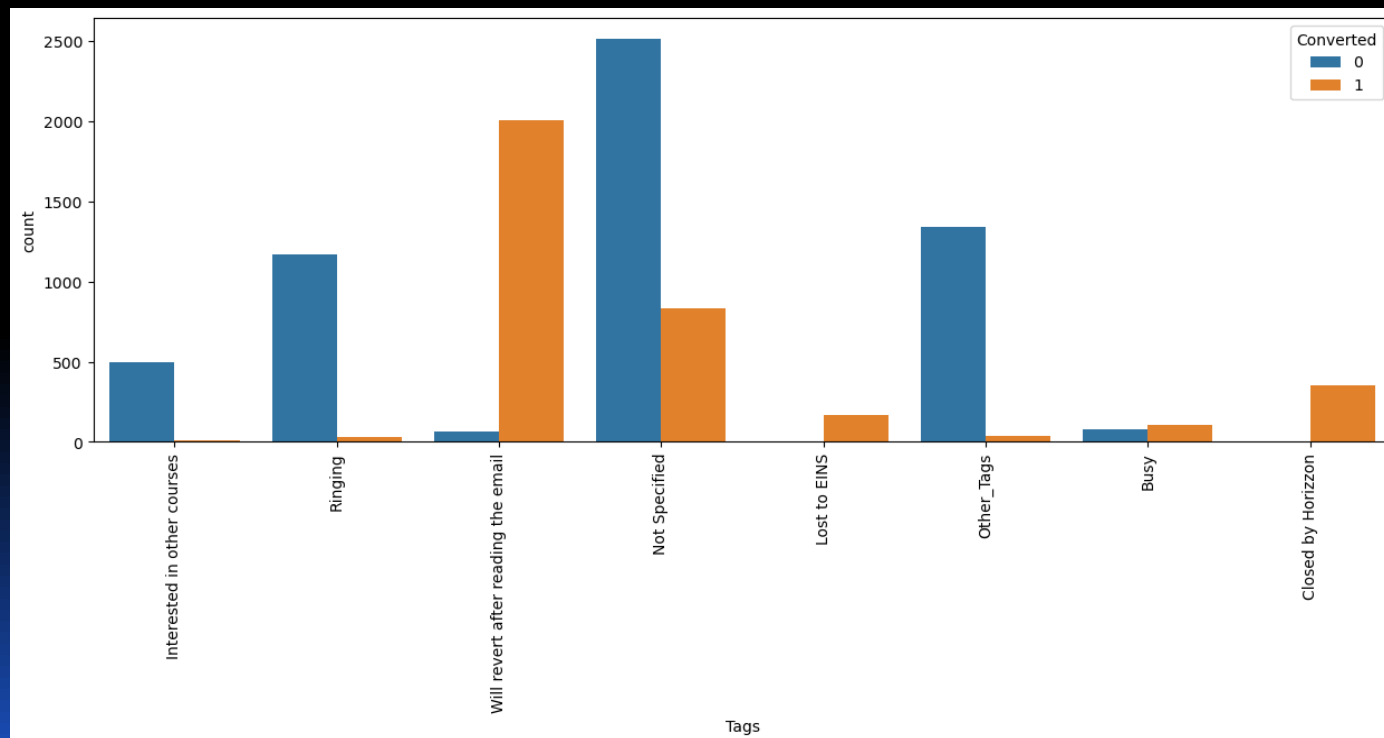
EDA – Categorical Columns

- What matter most to you in choosing a course
 - Imputed Nan values as 'Better Career Prospects'.
 - Dropping the column as only one value is present in almost 99.9% of the entries.



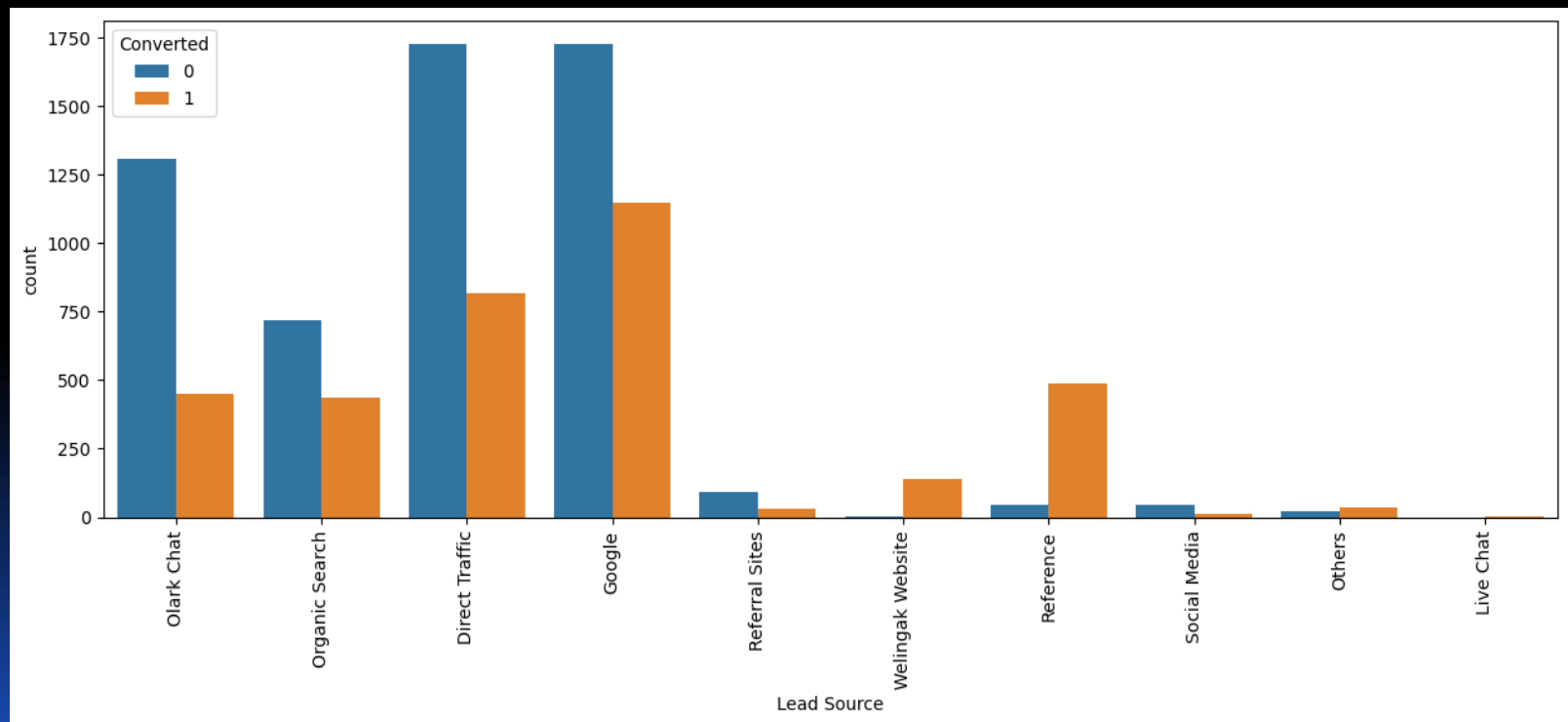
EDA – Categorical Columns

- Tags
 - Imputed Nan values as 'Not Specified'.
 - Merging all low count Tags into 'Other _Tags'.



EDA – Categorical Columns

- Lead Source
 - Imputed Nan values as 'Others'.
 - Naming adjusted and merged all the low count categories into 'Others'.

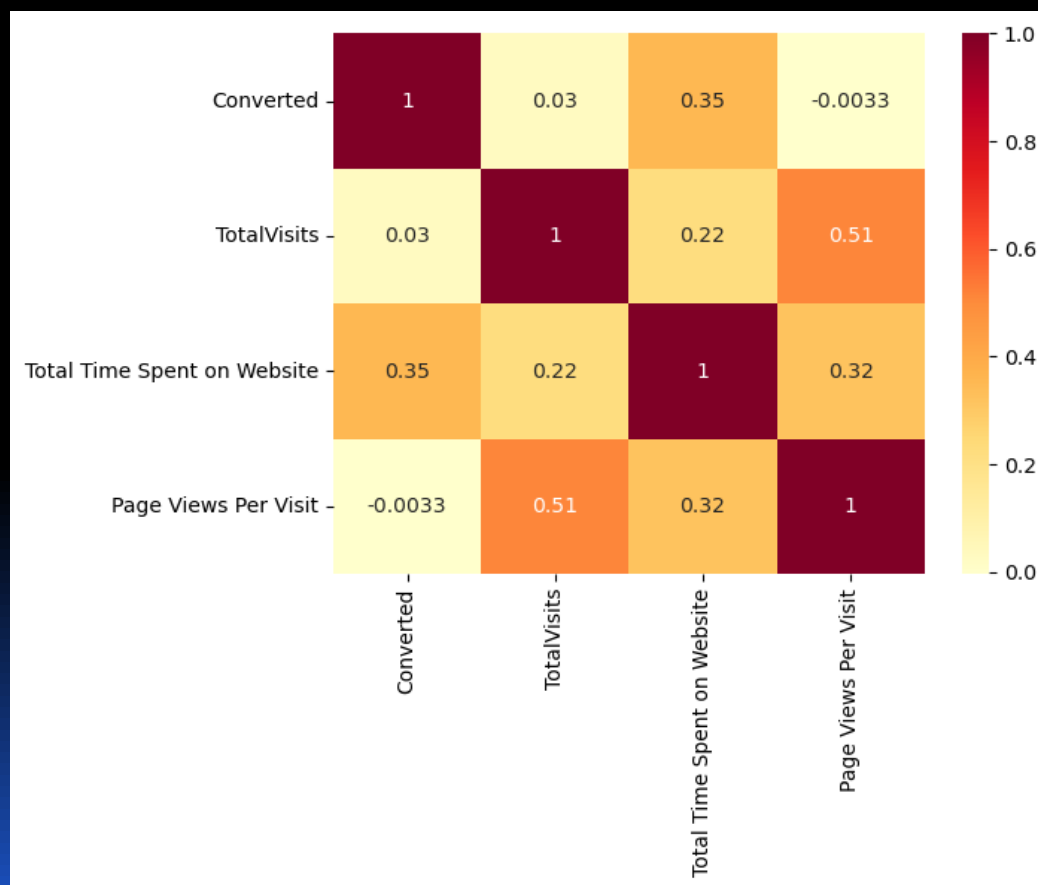


EDA – Categorical Columns

- Last Activity
 - Imputed Nan values as 'Others'.
 - Merged all the low count categories into 'Others'.
- Nan values are less than 2% in remaining null values columns, dropped all the rows with Nan values, as it won't effect the analysis much.
- Dropped all the below columns as they were having only one value across all the data entries ('Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', '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').

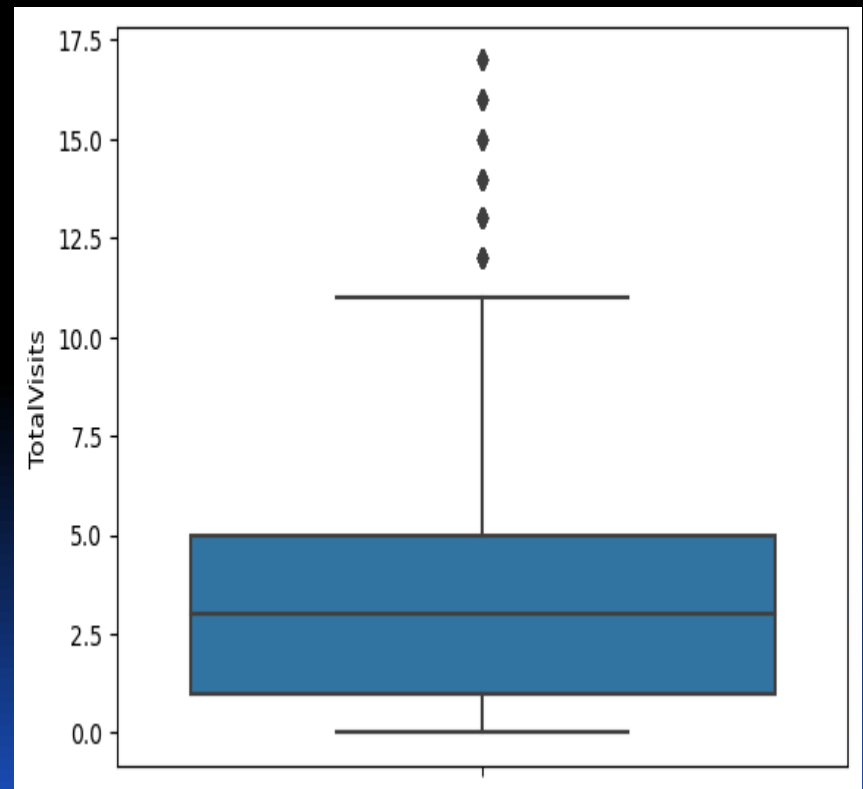
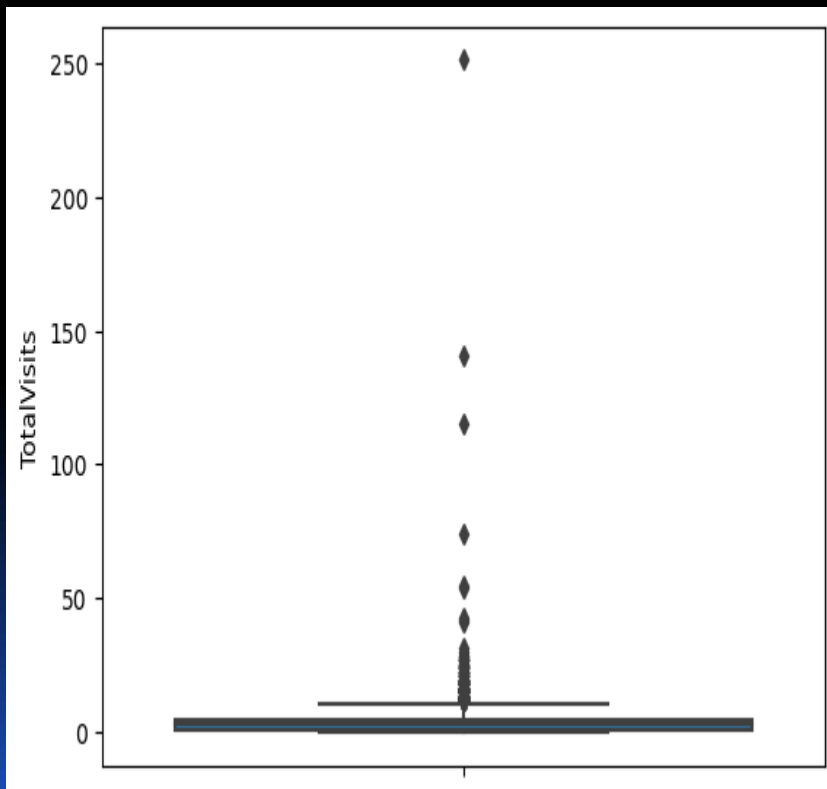
EDA – Numerical Columns

- Correlations mapped into a Heat Map for numerical columns.



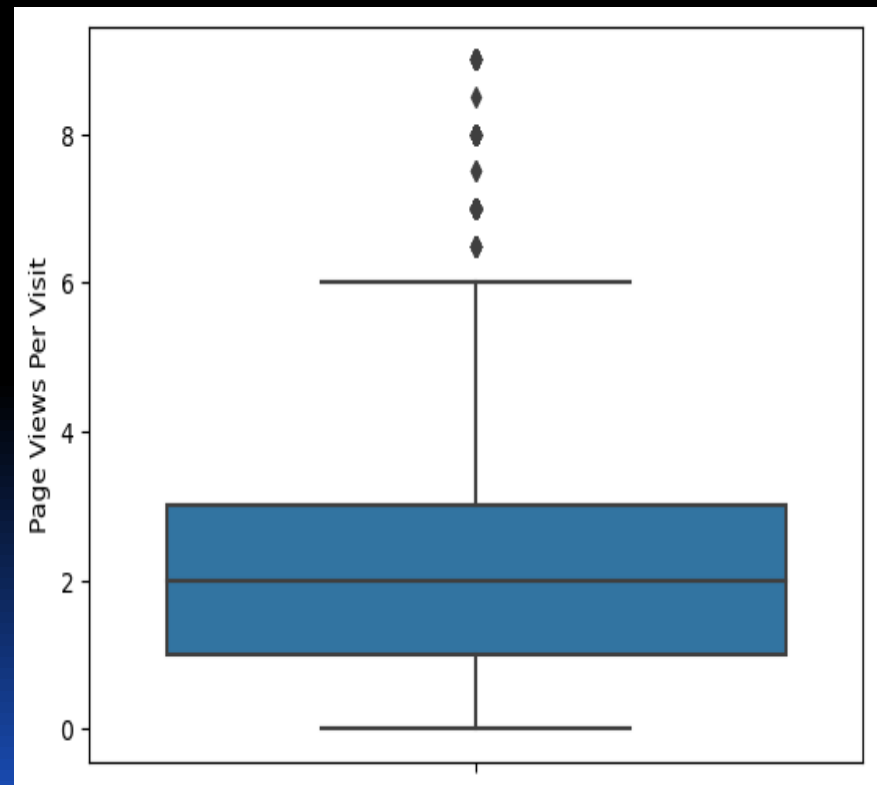
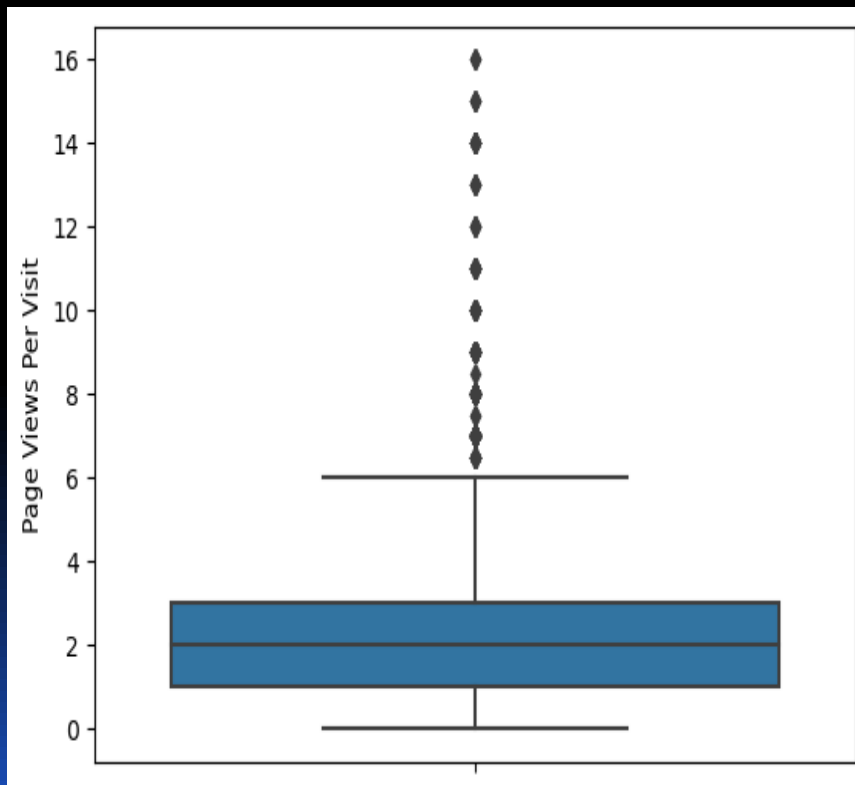
EDA – Numerical Columns

- Total Visits
 - High valued outlier are removed (one percentile from top and bottom)



EDA – Numerical Columns

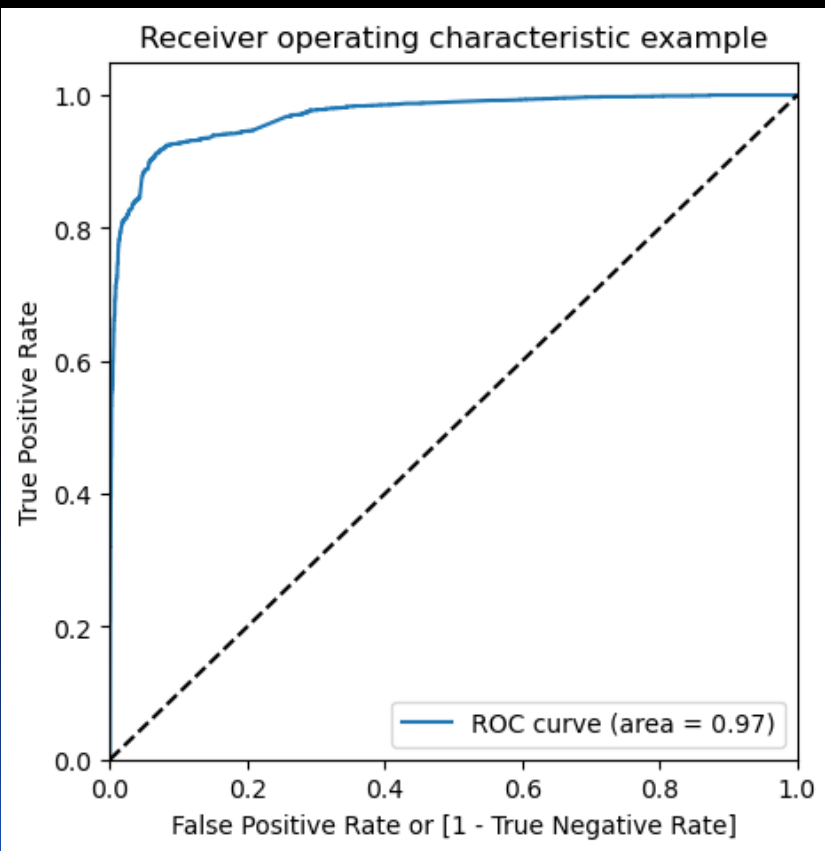
- Page Views Per Visit
 - Outlier treatment – values are removed (one percentile from top and bottom)



Model Evaluation

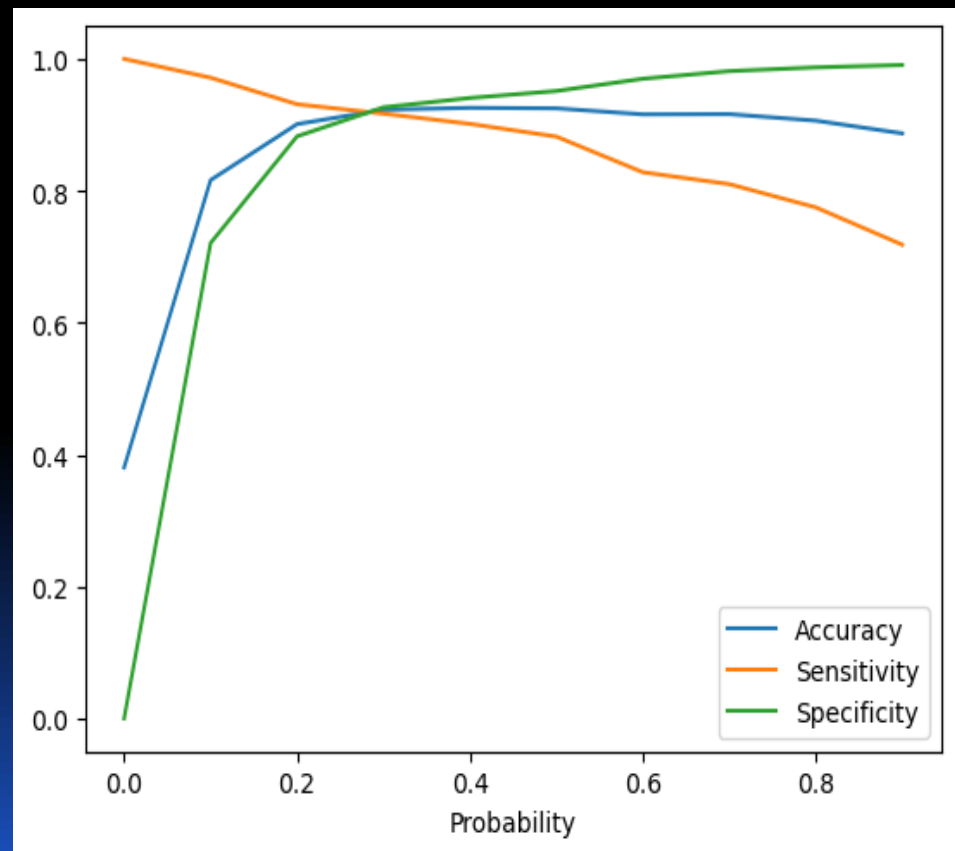
- ROC Curve

- Curve is close to 1 and area is 0.97



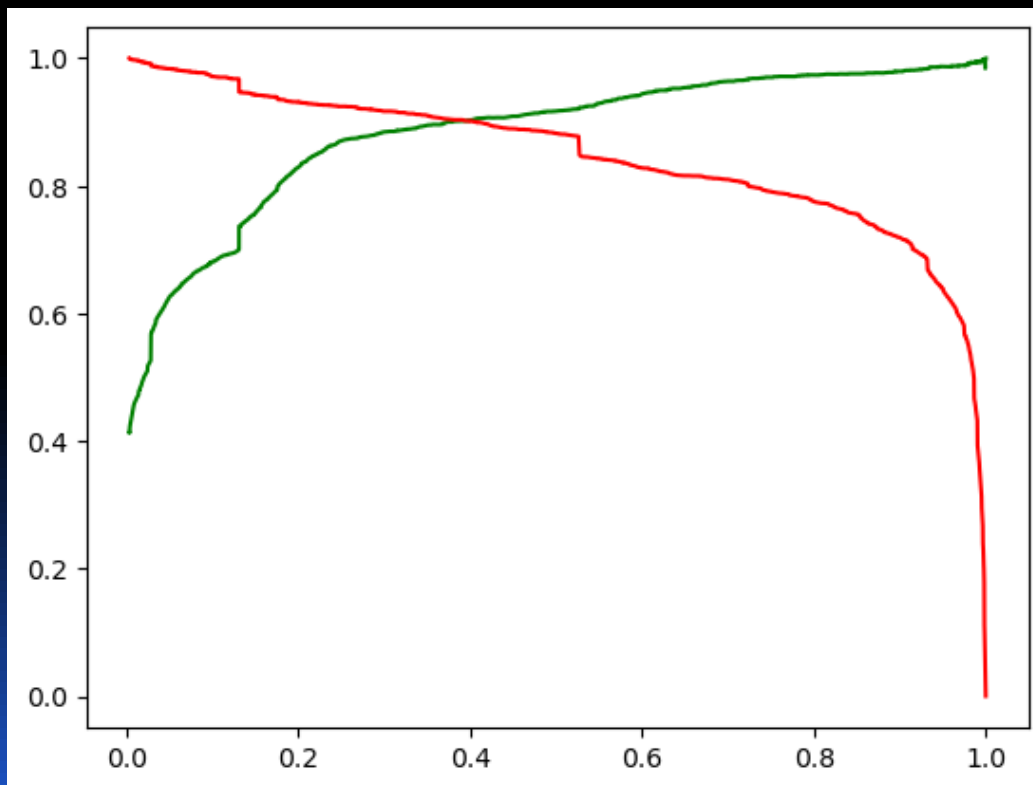
- Optimal Cutoff Point

- Point of intersection is close to 0.3, hence the optimal cutoff point is 0.3



Model Evaluation

- Precision – Recall Curve
 - Precision for cutoff 0.3 is 88.47%
 - Recall for cutoff 0.3 is 91.69%





Observations

- Train Data :
 - ▣ Accuracy : 92.29%
 - ▣ Sensitivity : 91.69%
 - ▣ Specificity : 92.65%
 - ▣ Precision : 88.47%
 - ▣ Recall : 91.69%

- Test Data :
 - ▣ Accuracy : 92.70%
 - ▣ Sensitivity : 91.68%
 - ▣ Specificity : 93.31%
 - ▣ Precision : 89.21%
 - ▣ Recall : 91.68%



Final Features List

- Lead Origin_Lead Add Form
- Tags_Will revert after reading the email
- Last Activity_SMS Sent
- Last Notable Activity_Modified
- Lead Source_Direct Traffic
- Lead Source_Welingak Website
- Tags_Other_Tags
- Total Time Spent on Website
- Tags_Closed by Horizzon
- Tags_Ringing
- Tags_Interested in other courses
- Tags_Lost to EINS
- Last Notable Activity_Olark Chat Conversation