LEAD SCORING CASE STUDY

Submitted by-

KRISHNAPRAKASHA M N

PRAMIT GHOSH

PROBLEM STATEMENT:

- An education company named 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'.
- f 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:

- Xeducation 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.

STEPS :

- ❑ 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.

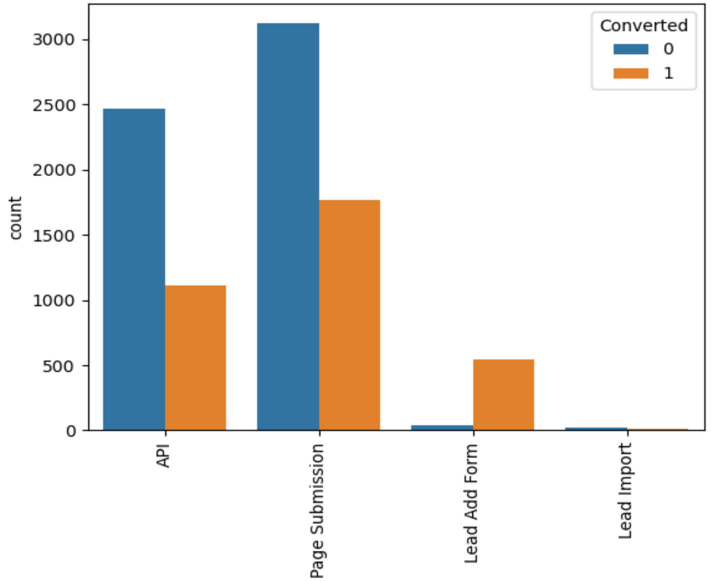
  2.Bivariate data analysis: correlation coefficients and pattern between the variables etc.

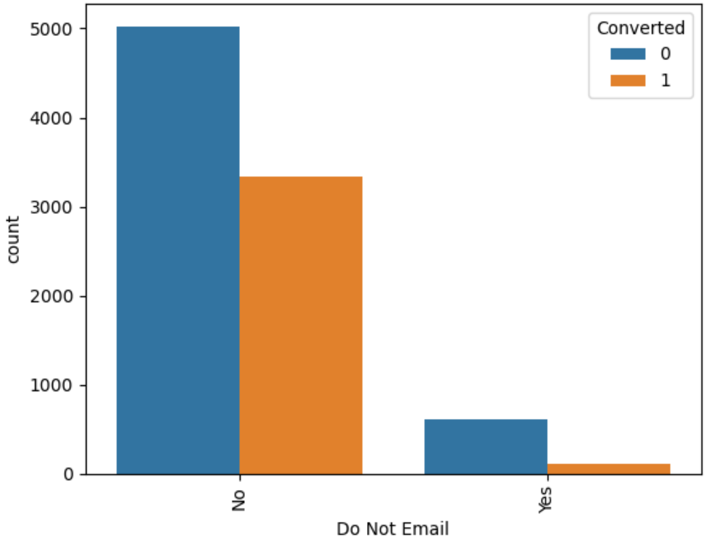
- ❑ Feature Scaling & Dummy Variables and encoding of the data.
- ❑ Classification technique: logistic regression used for the model making and prediction.
- ❑ Validation of the model.
- ❑ Model presentation.
- ❑ Conclusions and recommendations.

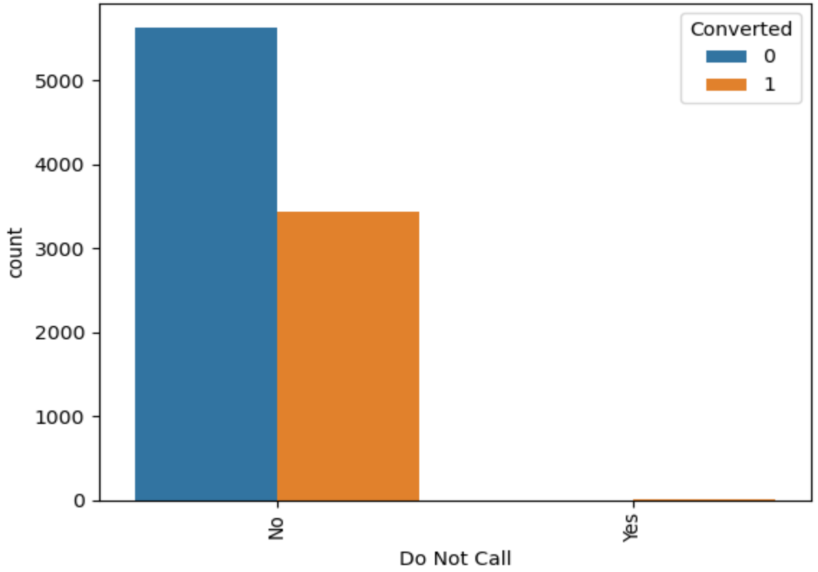
DATA CLEANING & MANIPULATION

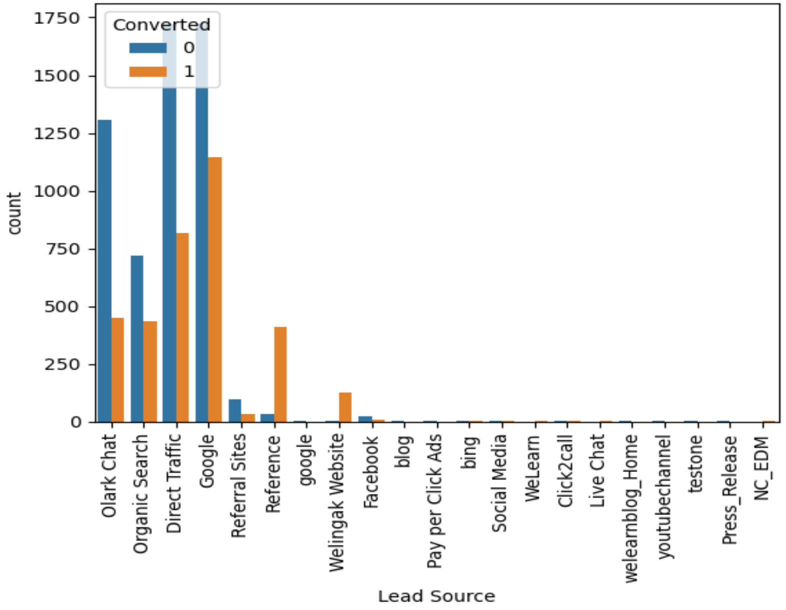
- Total Number of Rows =37, Total Number of Columns = 9240.
- Columns with single unique value have been dropped since they are not useful in the analysis.
- Removing the “Lead Number, Tags” which is not necessary for the analysis.
- Some columns have Select values which stored as null. Treated these values with np.nan.
- Dropping the columns having more than 40% as missing values.
- Columns with less than 40% missing values have been imputed with the mode value of the respective columns.

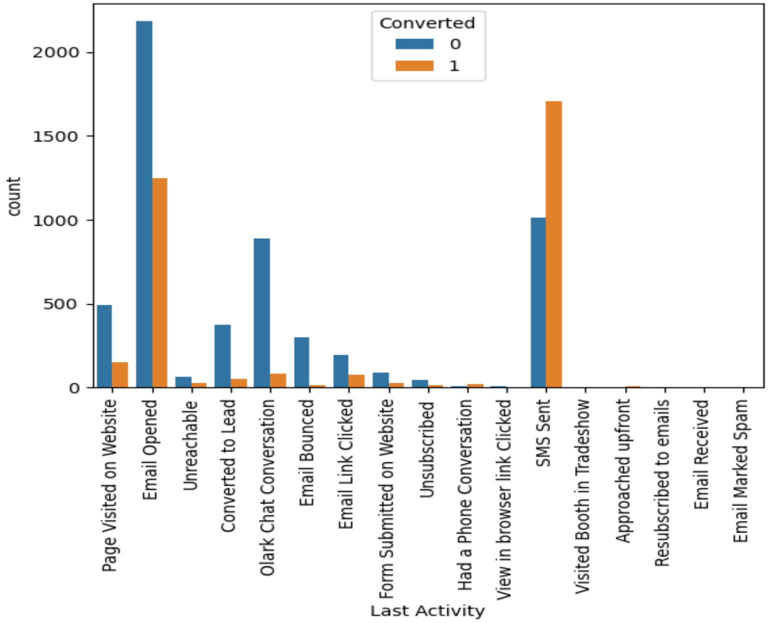
UNIVARIATE ANALYSIS

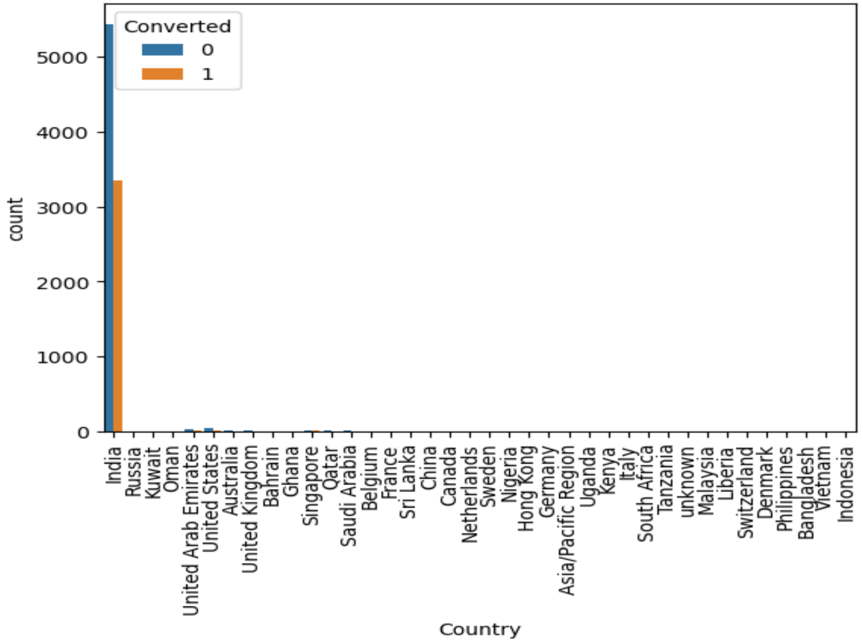




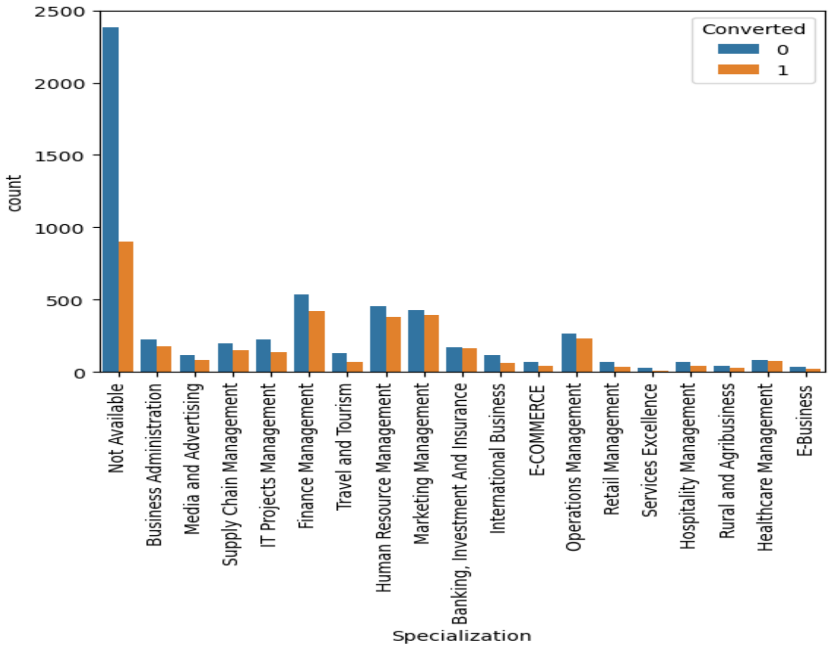


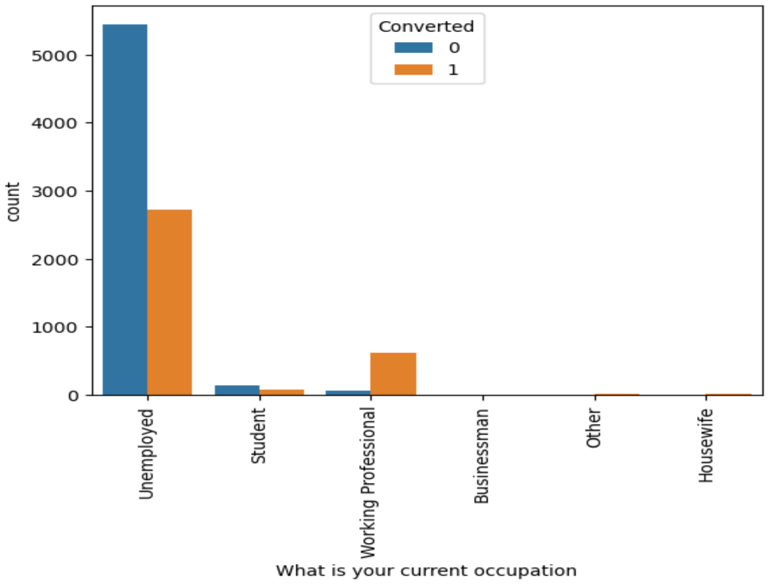


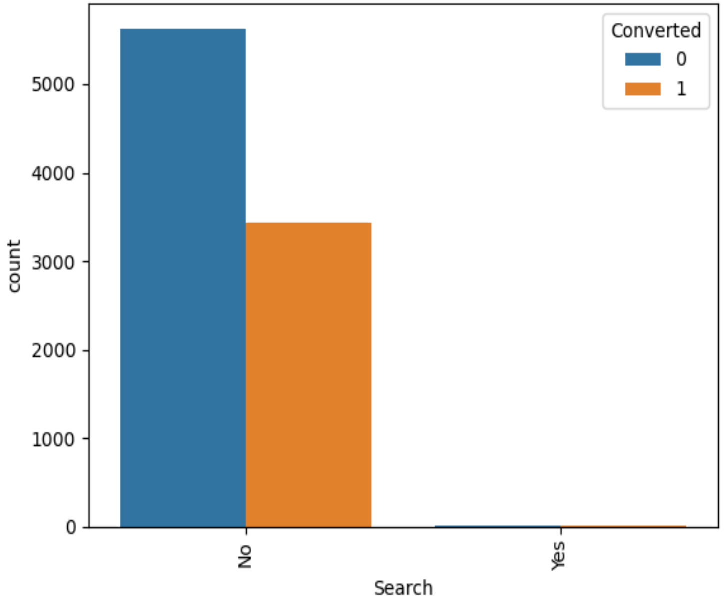


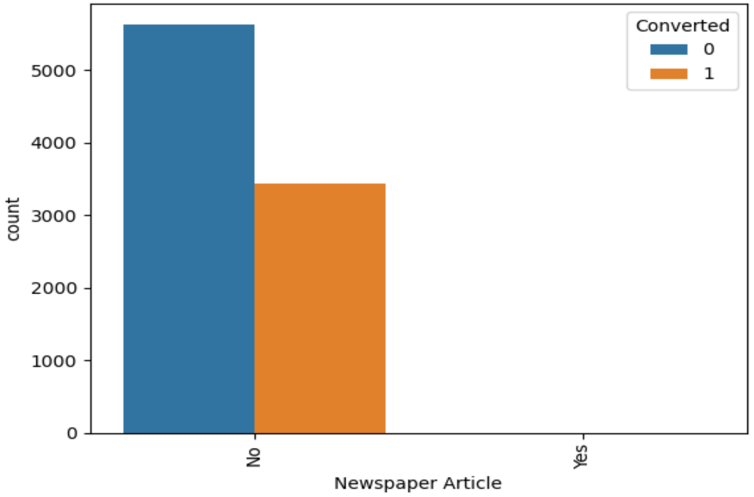


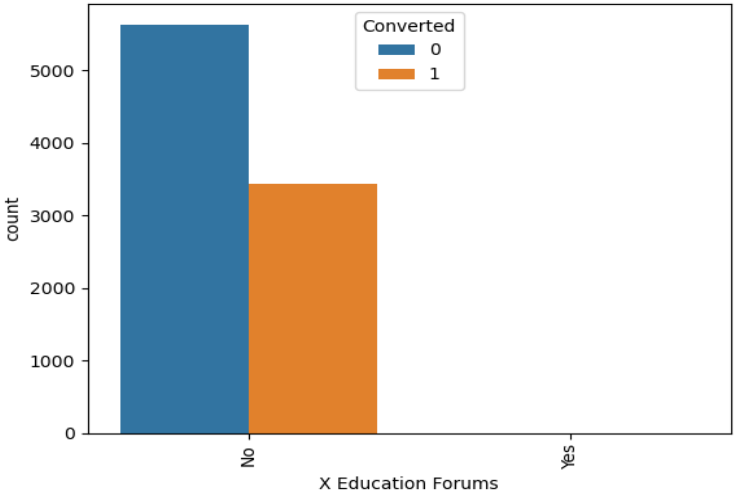
UNIVARIATE ANALYSIS

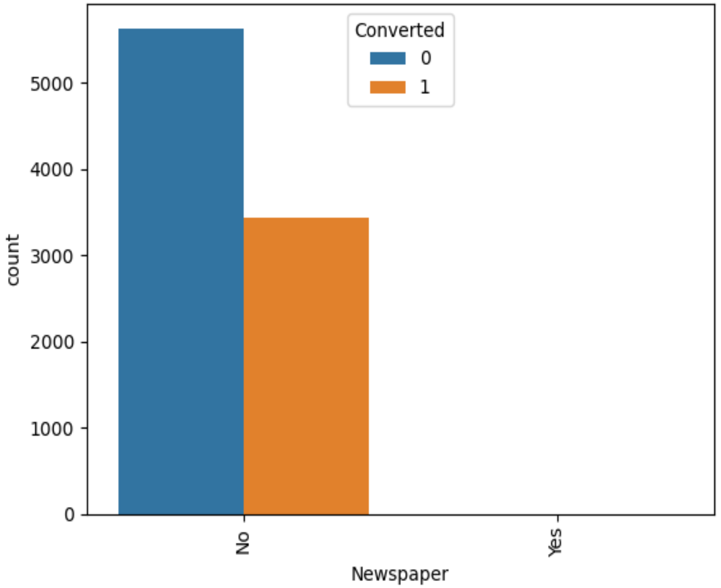




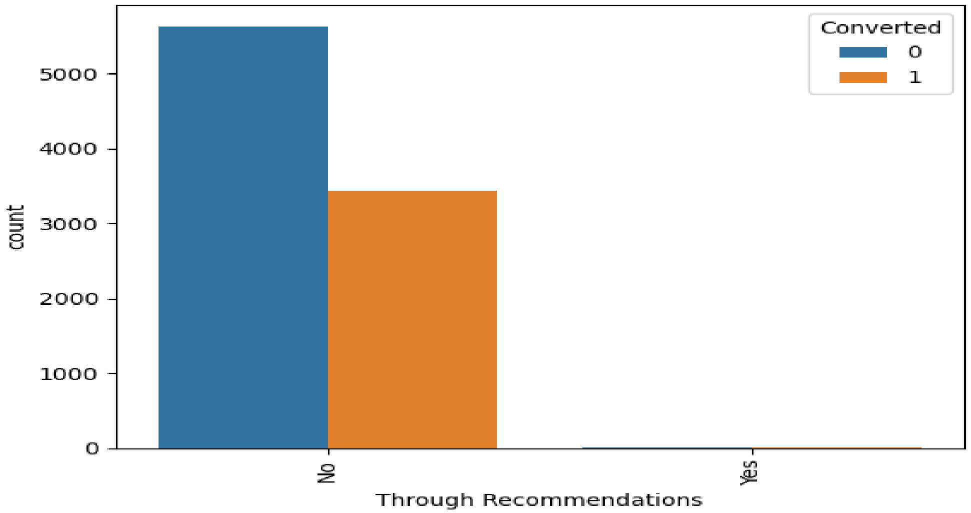


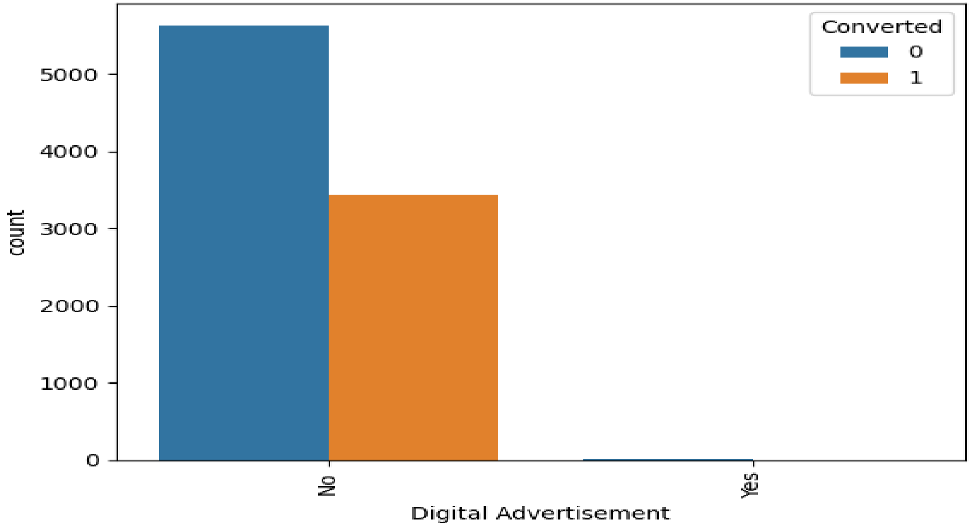


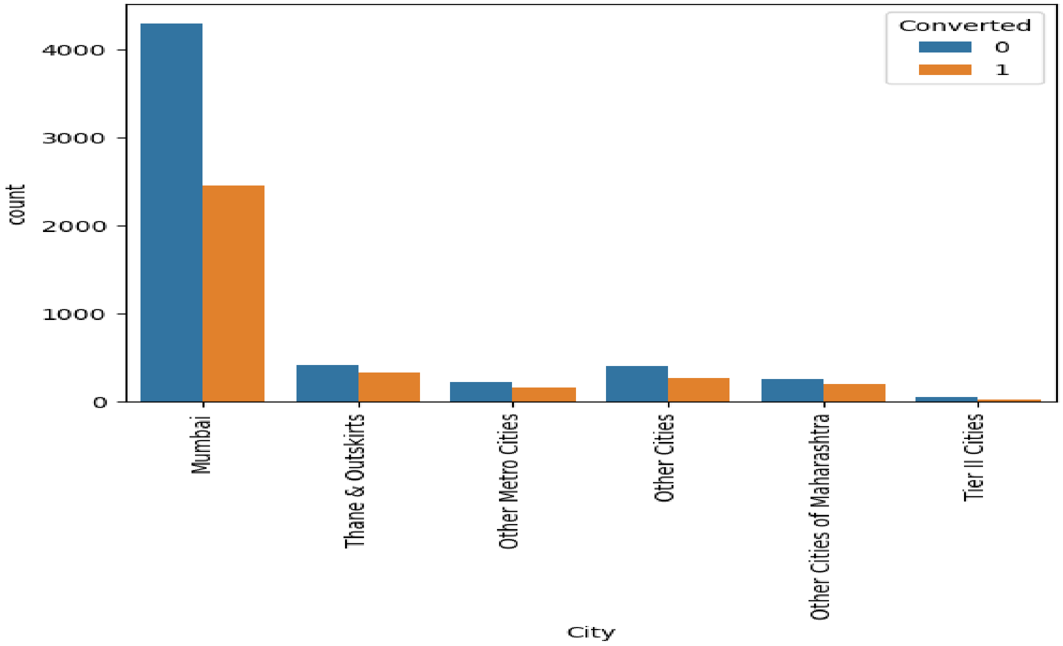


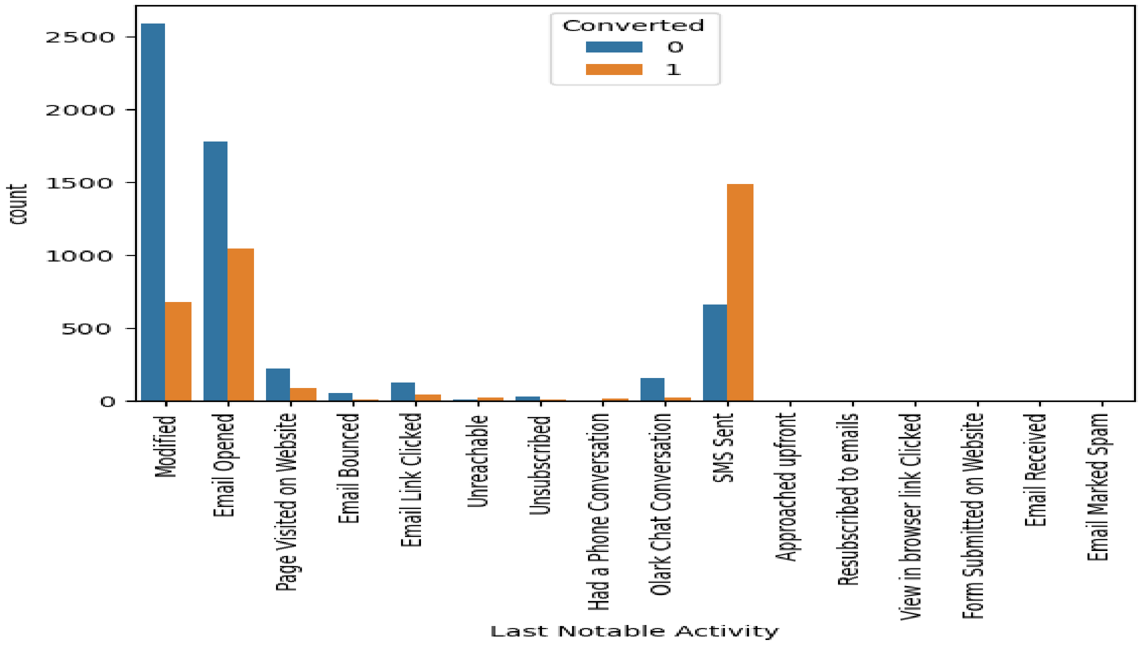


UNIVARIATE ANALYSIS









Observations from Univariate Analysis

- The API and Land Page Submission have less converted counts. Need to focus on improving the conversion. Lead Add form has more conversion but less in number.

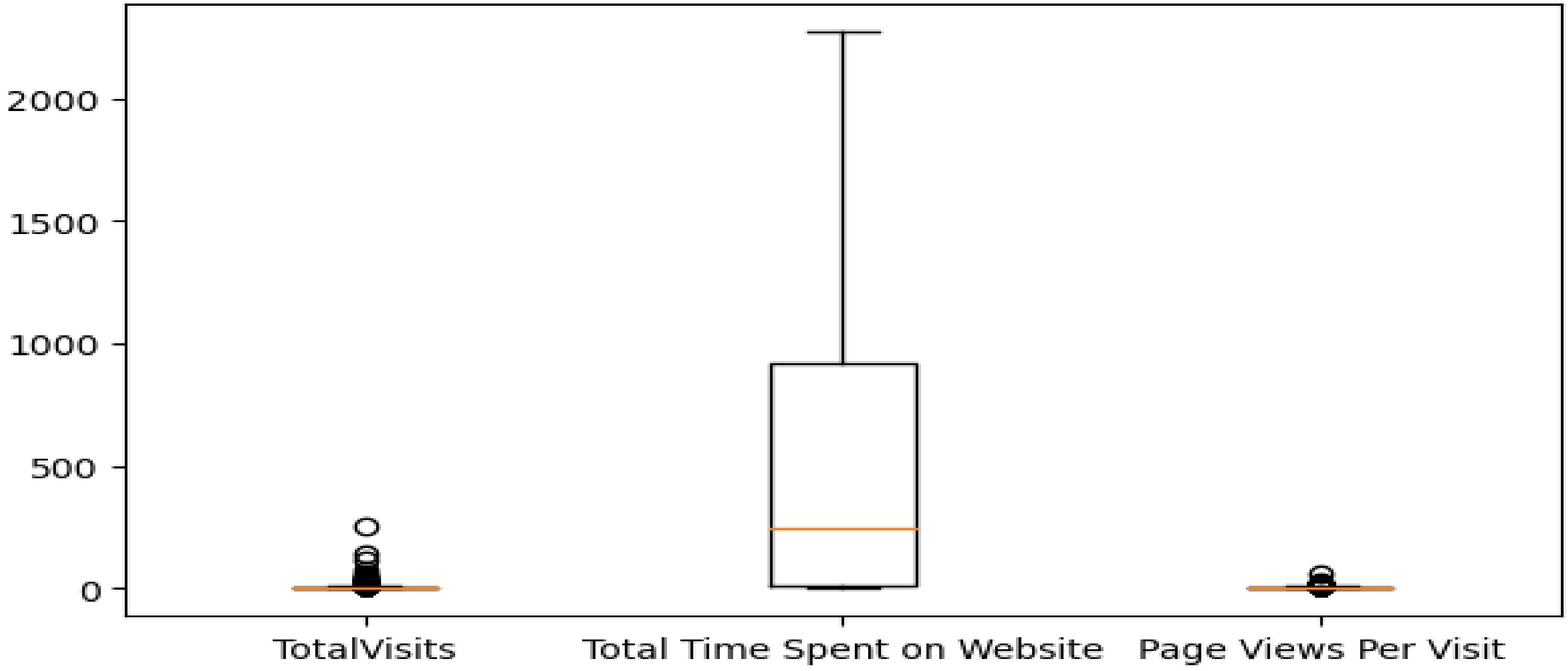
- Olark Chat, Organic Search, Direct Traffic, Google have less conversion rate. Need to focus on improving of mentioned datas. Referal Sites, Welingak Website have high conversion.

- Leads with SMS sent has more converision rates. Most of the leads have kept their email opened.
- No inferences to be made as this column is highly skewed by the data value India.
- Almost all of the specialization have both converted and not converted leads. Hence focus to be made more on converting the leads in specialization category.

- Professional have high conversion rate. Need to have high focus on convertion of Unemployed people.

- Need to have focus on people from all the cities to improve the converted rate.

OUTLIER HANDLING:



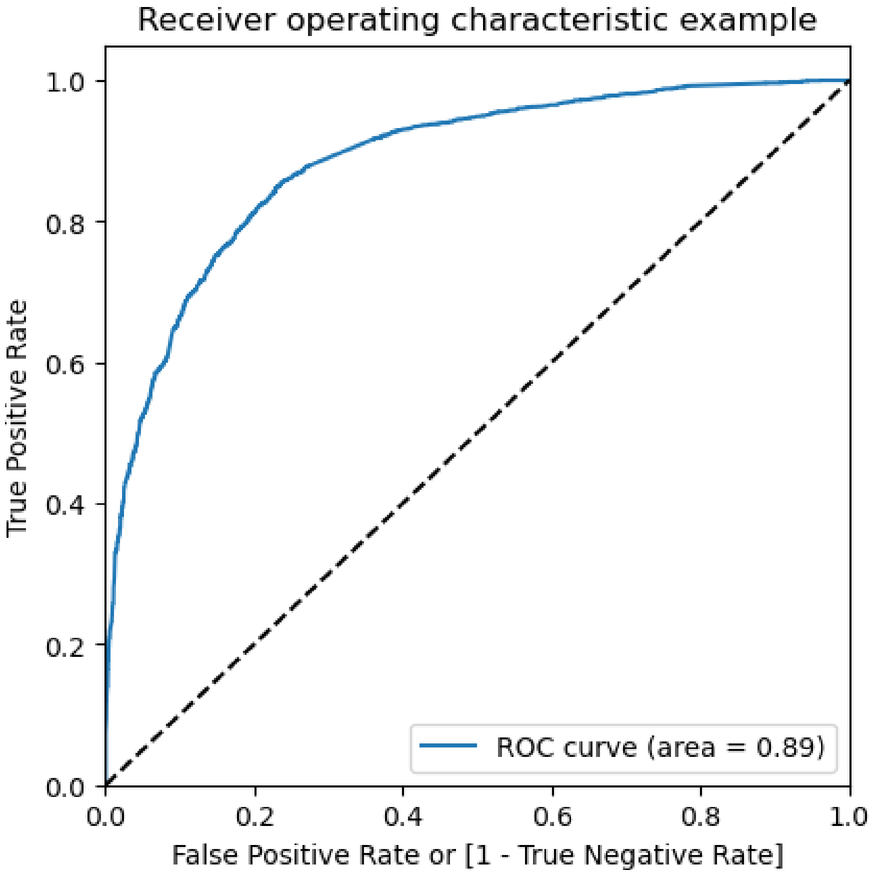
- Total Visits and Page Views Per Visit have outliers. Need to treat the outliers in these columns.

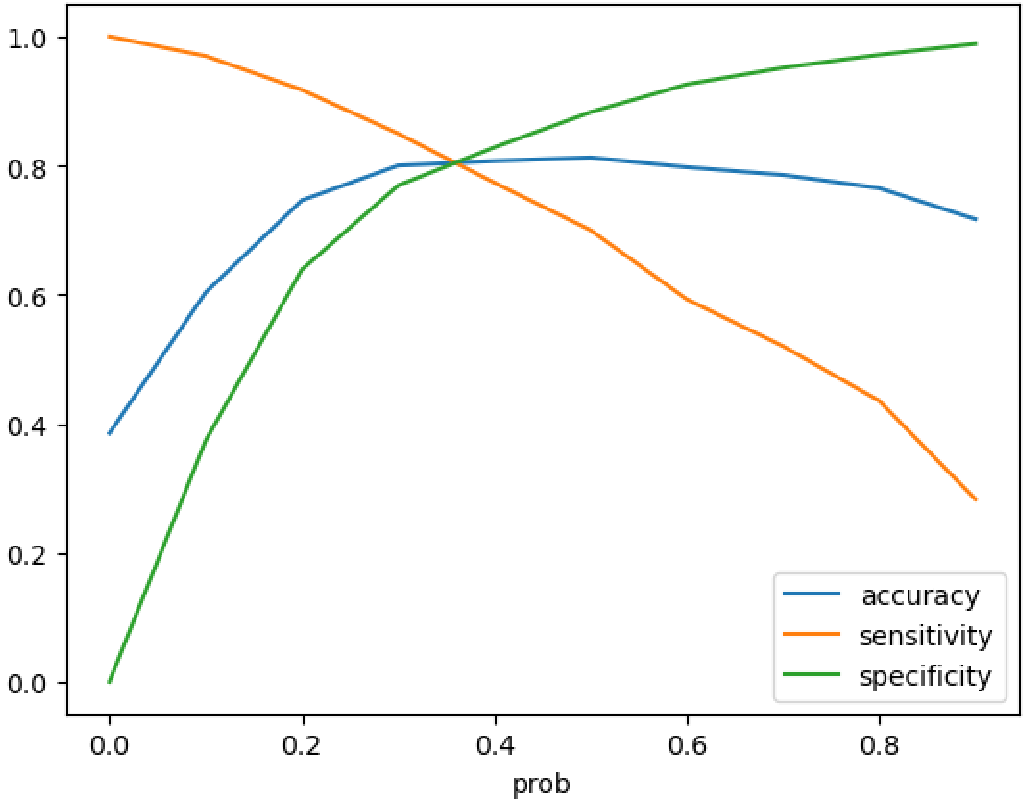
MODEL BUILIDING:

- Data Preparation carried out by converting binary variables to 0/1 and dummy variables creation done from categorical variables. Also done the scaling of numerical variables.
- Performed train-test split, we have chosen 70:30 ratio. Splitting the Data into Training and Testing Sets
- 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
- For Taining Data, Accuracy- 80.6%, Sensitivity – 80.7%, Specificity – 80.56%.
- For Testing Data, Accuracy- 80.38%, Sensitivity – 80.58%, Specificity – 80.27%.

Finding the Optimal Cut-off

- Optimal cut off probability is the probability where we get balanced sensitivity, specificity and accuracy.
- From the second graph it is visible that the optimal cut off is at 0.36.





CONCLUSION

- The model is able to give higher conversion rate of around 80%.
- We can make a decision on getting higher conversion rate based on the model that we have built.
- Welingak Website, Reference and Olark Chat are the Lead Sources which are most likely to get converted.
- Working Professionals are most likely to get converted from the occupation category.
- The company should contact to the leads whose last Notable activity shown as 'SMS Sent','Unreachable'.
- The company should contact to the leads whose last activity shown as 'Unsubscribed'.
- Leads who spent more time on the website are more likely to get converted.
- The company should not make calls to the leads Leads whose specialisation was Not Available.
- The company should not make calls to the leads Leads whose last activity was Olark Chat Conversation.
- The leads from Landing Page Submission lead origin are not likely to get converted.