LEAD SCORING CASE STUDY

By:
MEGHANA S. M.

SANNAN DABIR

BHARTI RANI

PROBLEM STATMENT:

- An education company, X Education sells online courses to industry professionals. The company markets its courses on various websites and search engines such as Google

- Once people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals

- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. The typical lead conversion rate at X education is around 30%

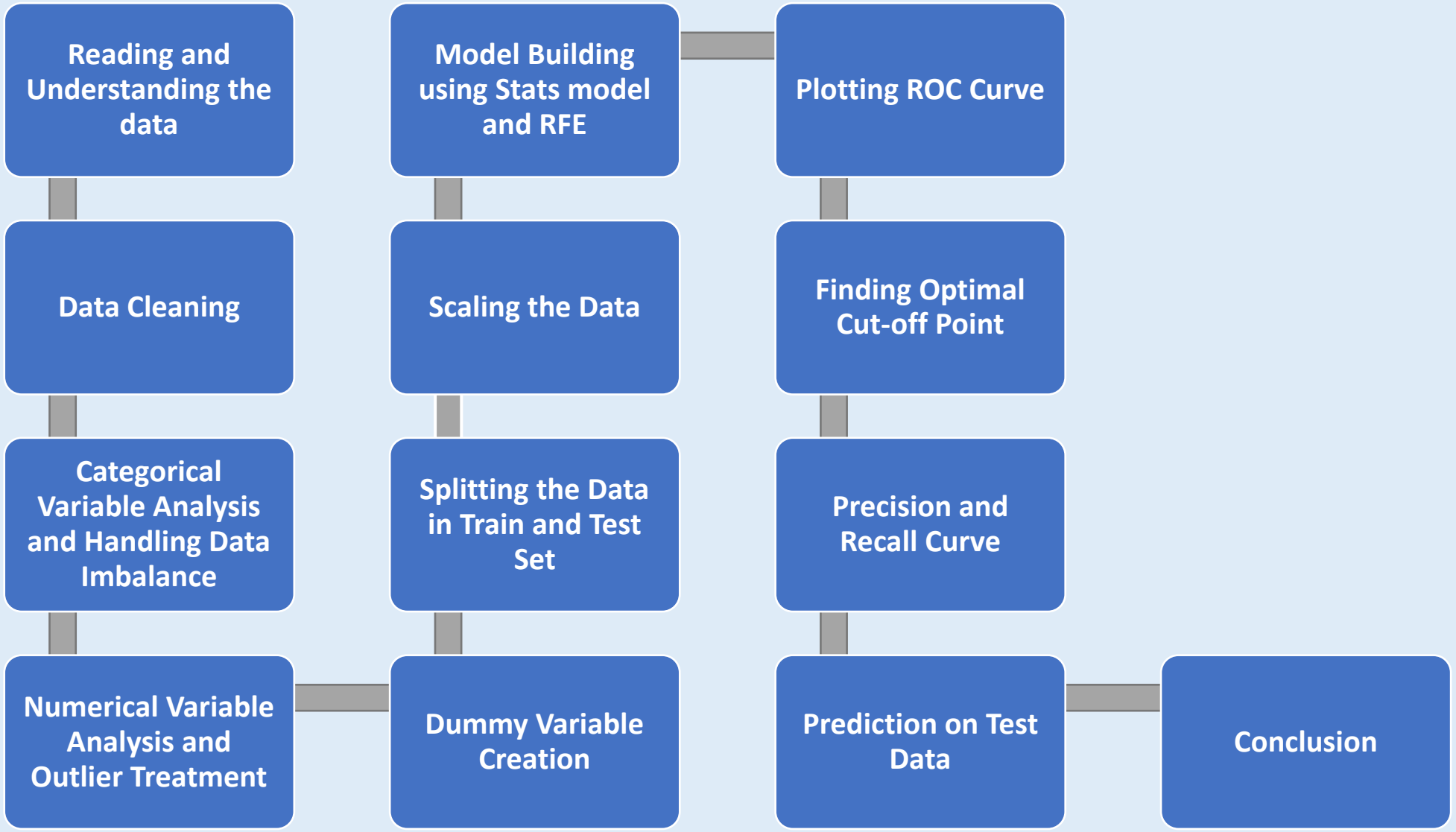
BUSINESS GOALS:

- Company wishes to identify the most potential leads, also known as “Hot Leads”

- The company needs a model wherein a lead score is assigned to each of the leads such that the customer with a higher lead score have a higher conversion chance and the customer with a lower lead score has a lower conversion chance

- Deployment of the model for the future use

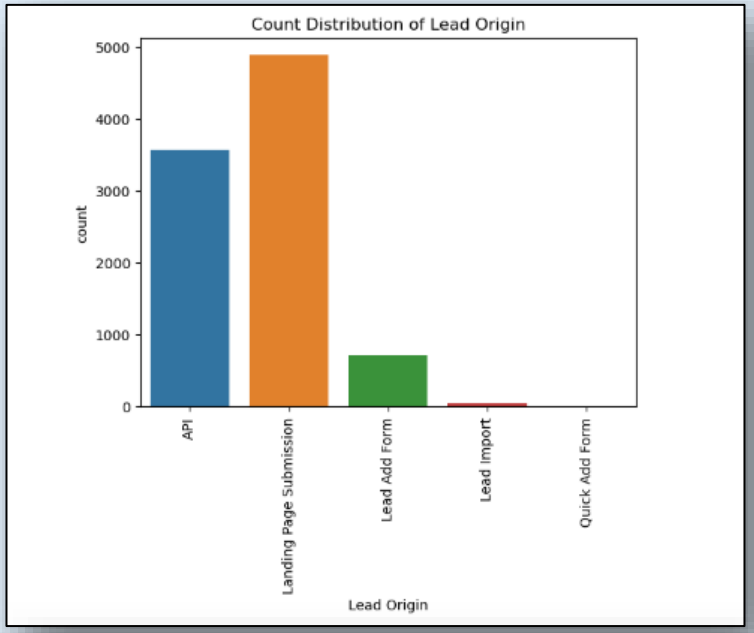
OVERALL APPROACH:

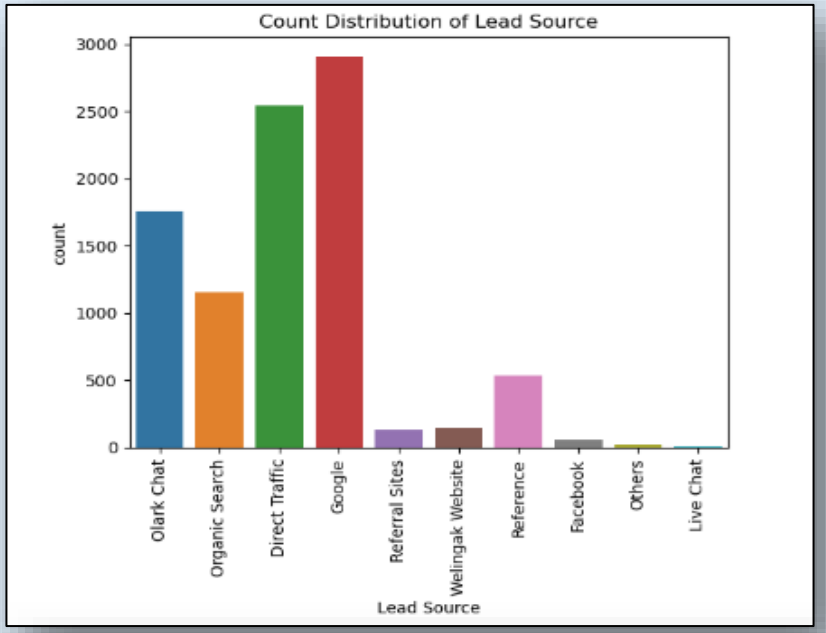


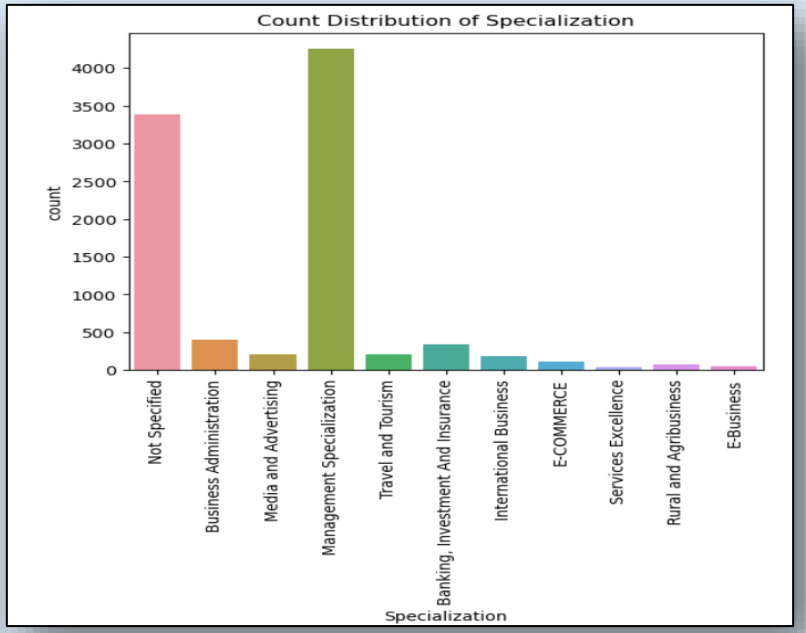
PROBLEM-SOLVING METHODOLOGY:

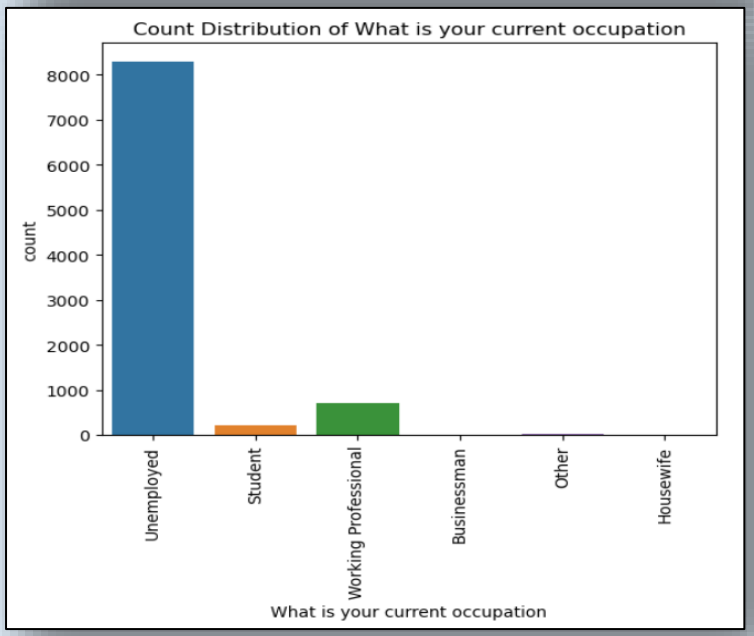
- Importing Data and Necessary Libraries
- Data cleaning and data manipulation
  - Check and handle NA values and missing values. In this dataset, “Select” values were present which were basically null values and were considered as null values.
  - Dropping columns containing a large number of missing values i.e. more than 40% of overall data.
  - Imputing the null values with the mode of the variable in case of categorical variables and median value in case of numerical variables.
  - Removing the variables having very high data imbalance.
- EDA and Data Visualization
  - Categorical Variable Analysis.
  - Numerical Variable Analysis.
  - Checking and handling outliers in data.
- Creation of Dummy Variables and encoding of the data.
- Scaling the Numerical Variables.
- Classification technique: logistic regression is used for the model making and prediction.
- Validation of the model using different parameters such as accuracy, sensitivity, specificity, recall, precision, etc.

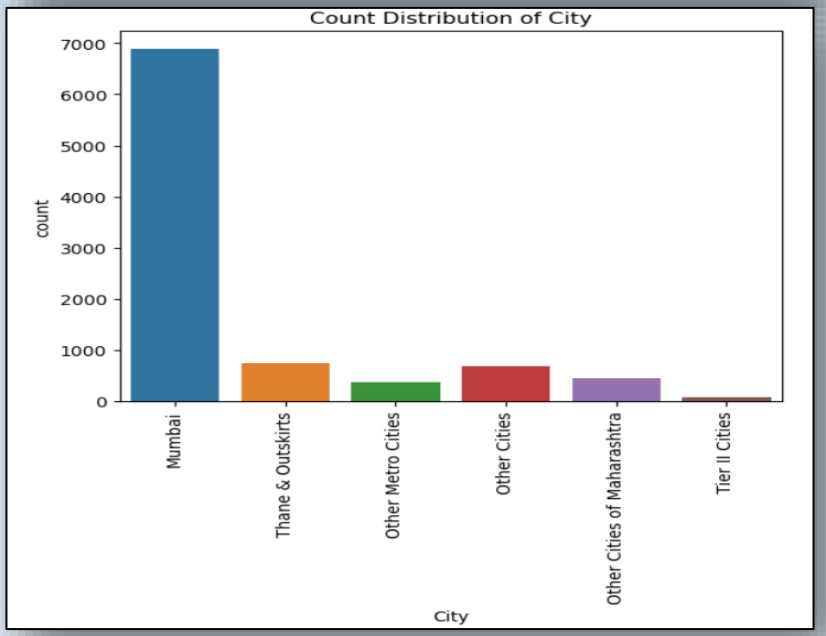
Analysis of Categorical Variables:



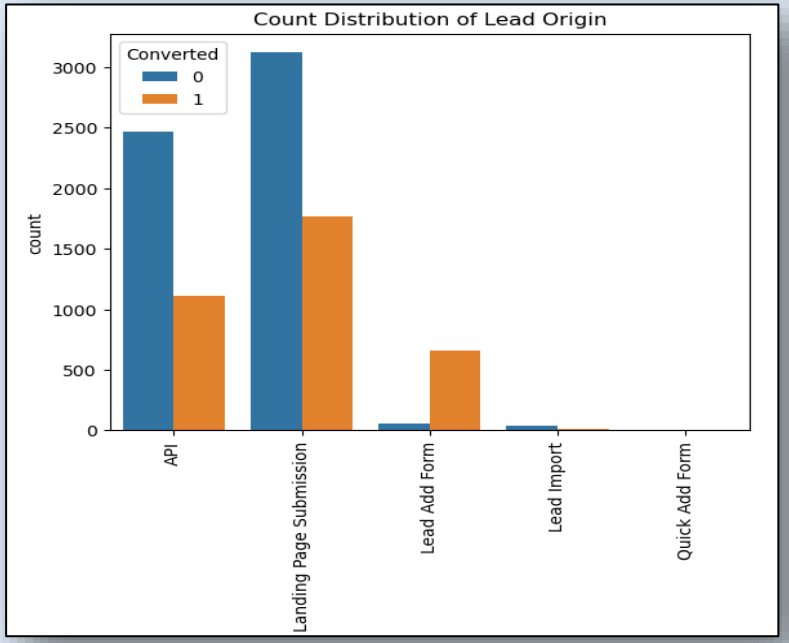


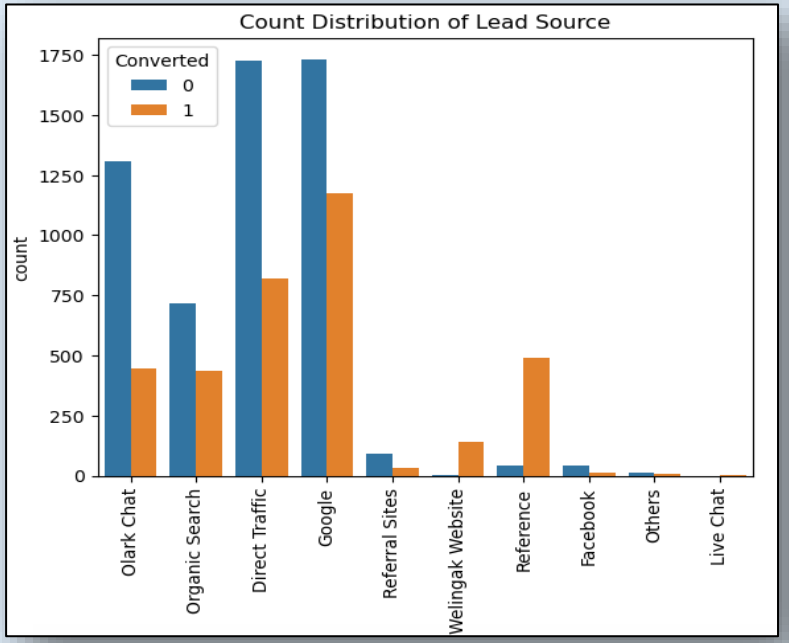


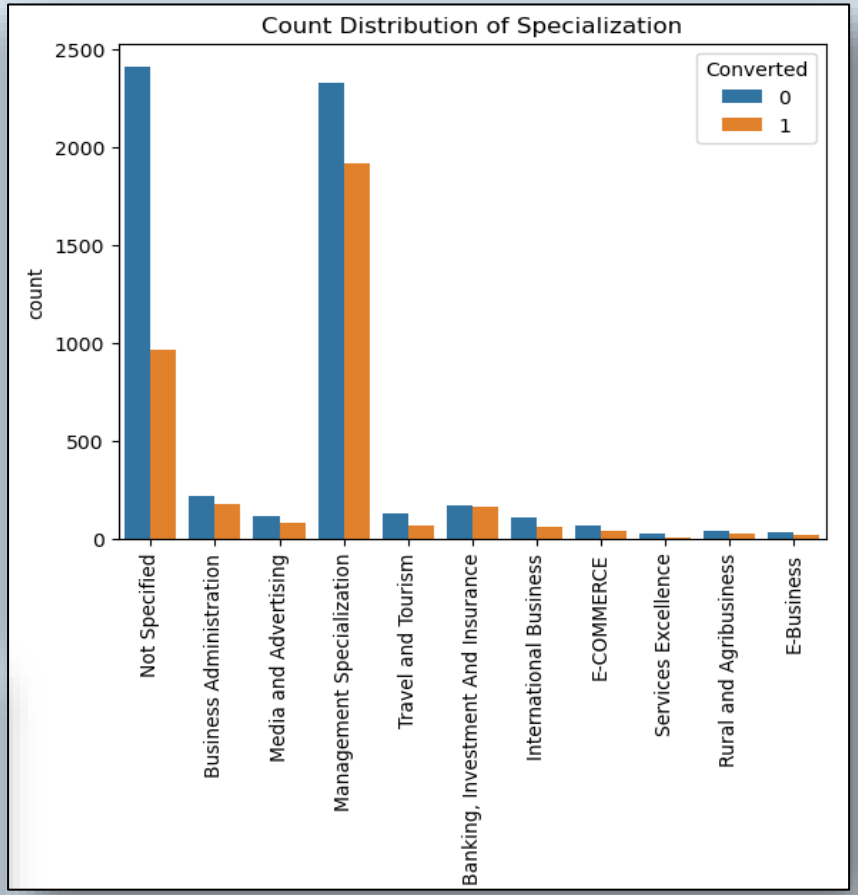


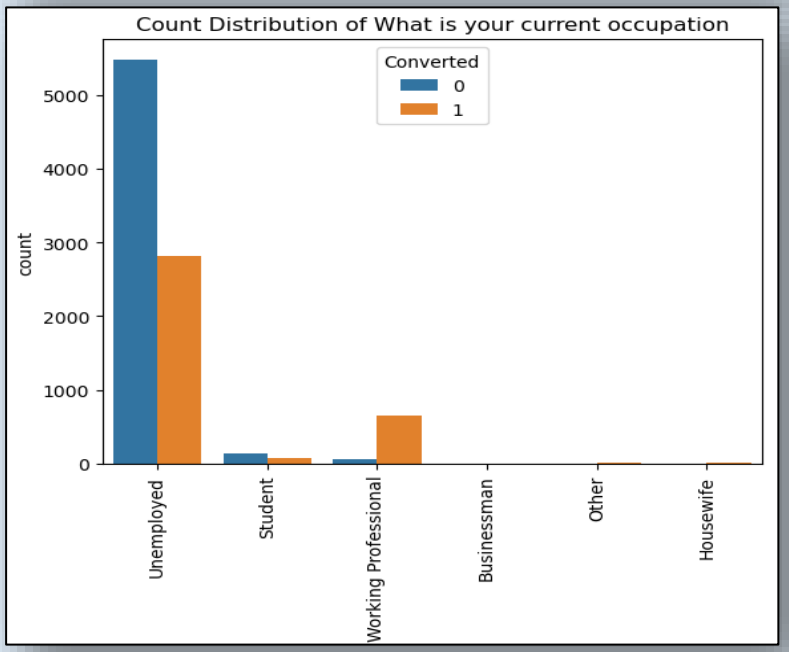


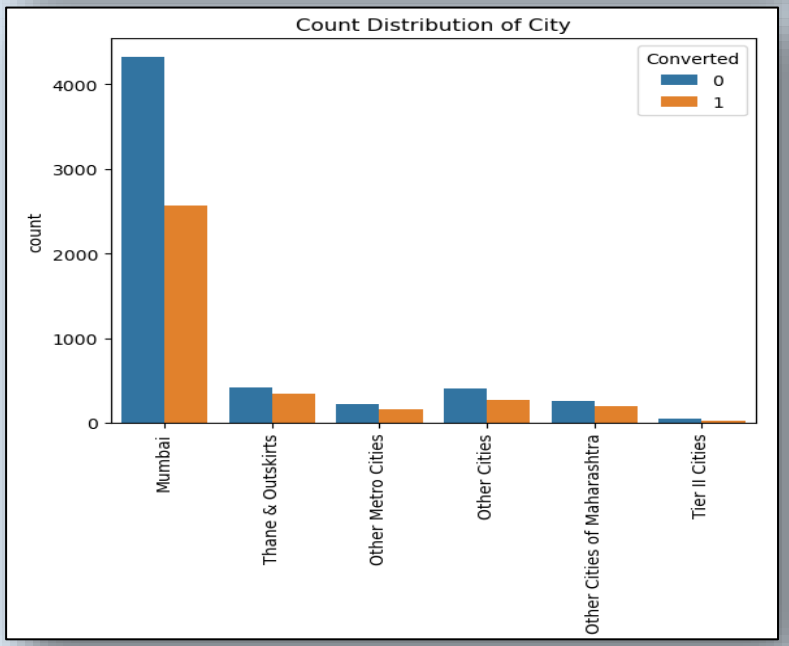
Analysis of Variables w.r.t. lead conversion:











Insights from the Analysis of Categorical Variable:

1. Lead Origin:
- Most of the leads have been generated from API and Landing PageSubmissionsn.
- Lead Add from has a very High Conversion rate followed by Landing Page Submission and API.

2. Lead Source:
- Maximum number of leads are generated by Google, Direct Traffic, and Olark Chat.
- References and Welingak Website have the highest conversion rate.

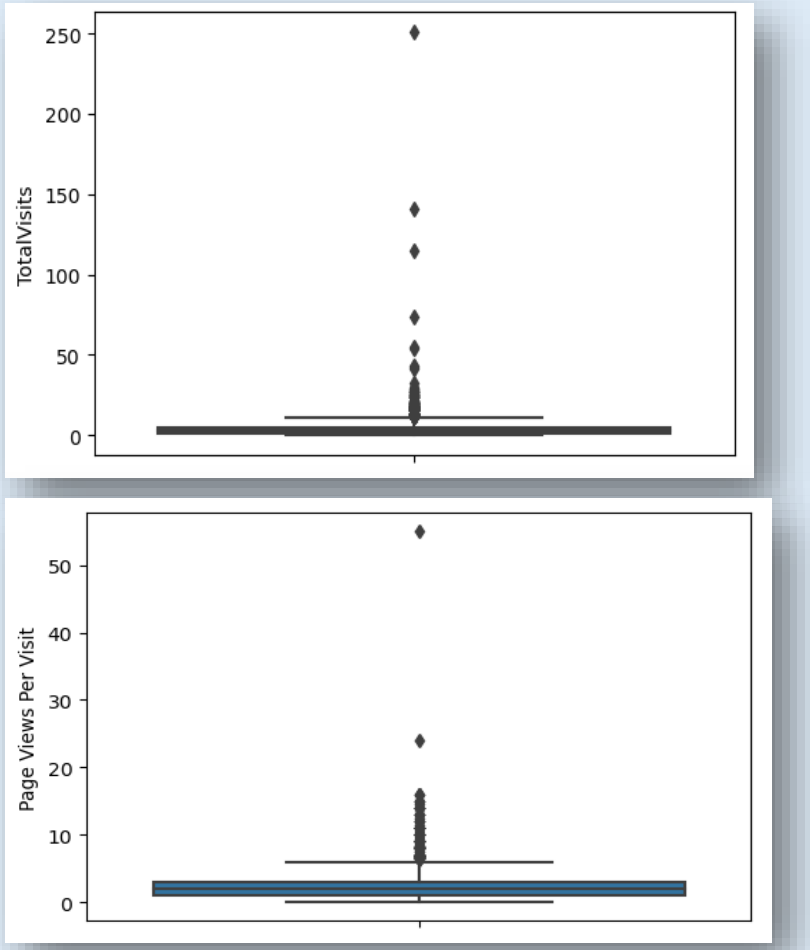
3. Current Occupation:
- Most of the leads are generated from the Unemployed Population.
- Working professionals have the highest Lead Conversion Rate.

4. City:
- Maximum leads are generated from Mumbai and it also has the highest lead conversion rate.
- Leads generated from other cities are very less as compared to Mumbai.

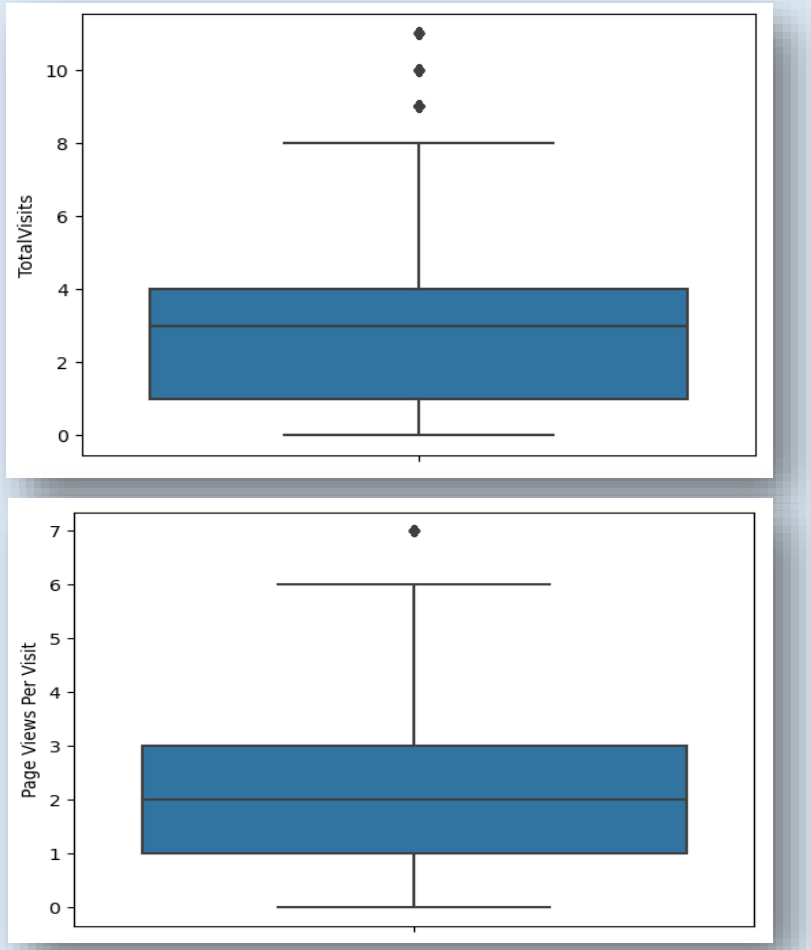
5. Specialization:
- The highest number of leads are generated from Management Specialization and it also has the highest conversion rate.

Analysis of Numerical Variables:

Before Outlier Treatment:

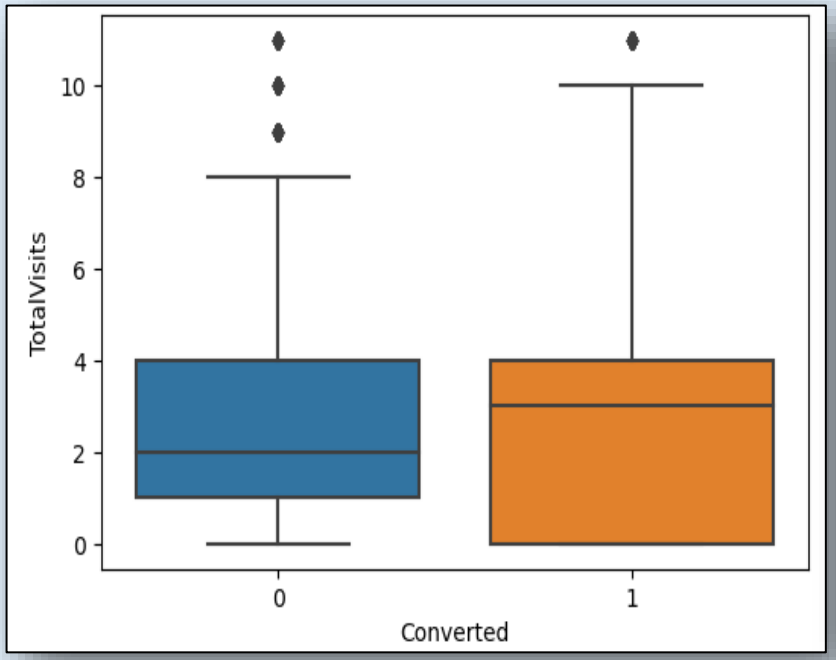


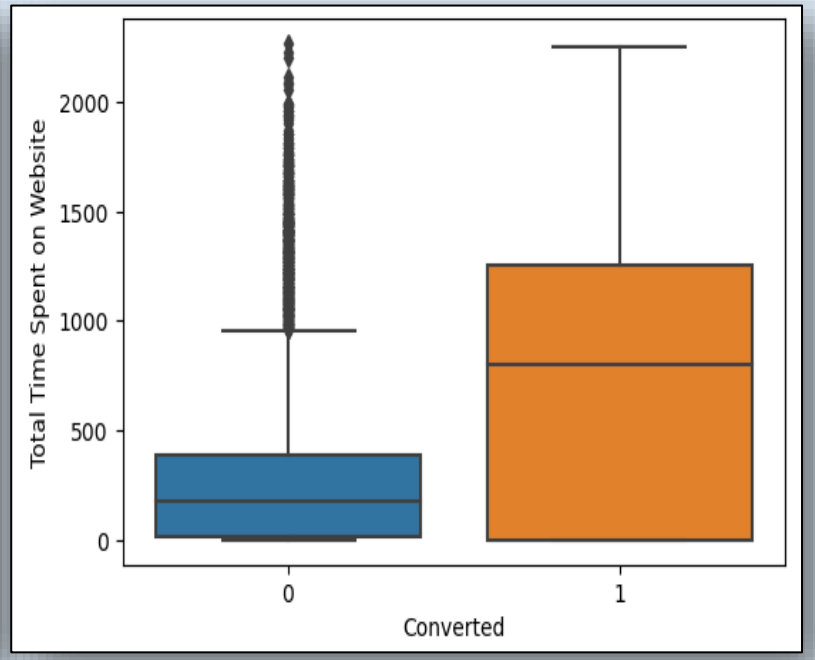
After Outlier Treatment:

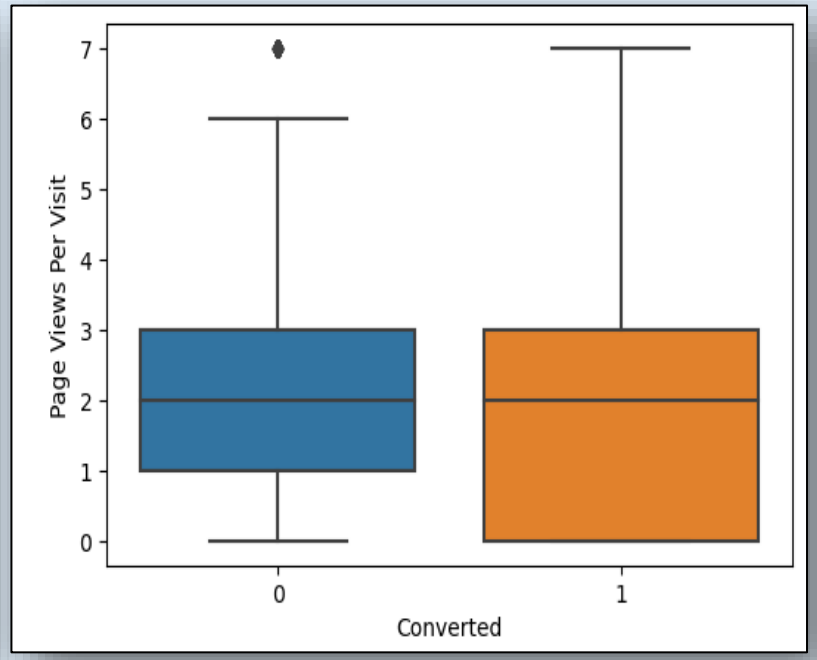


Top 3% Outliers have been removed from both the Page Views Per Visit and Total Visit Variable.

Analysis of Numerical Variables w.r.t. leads converted:





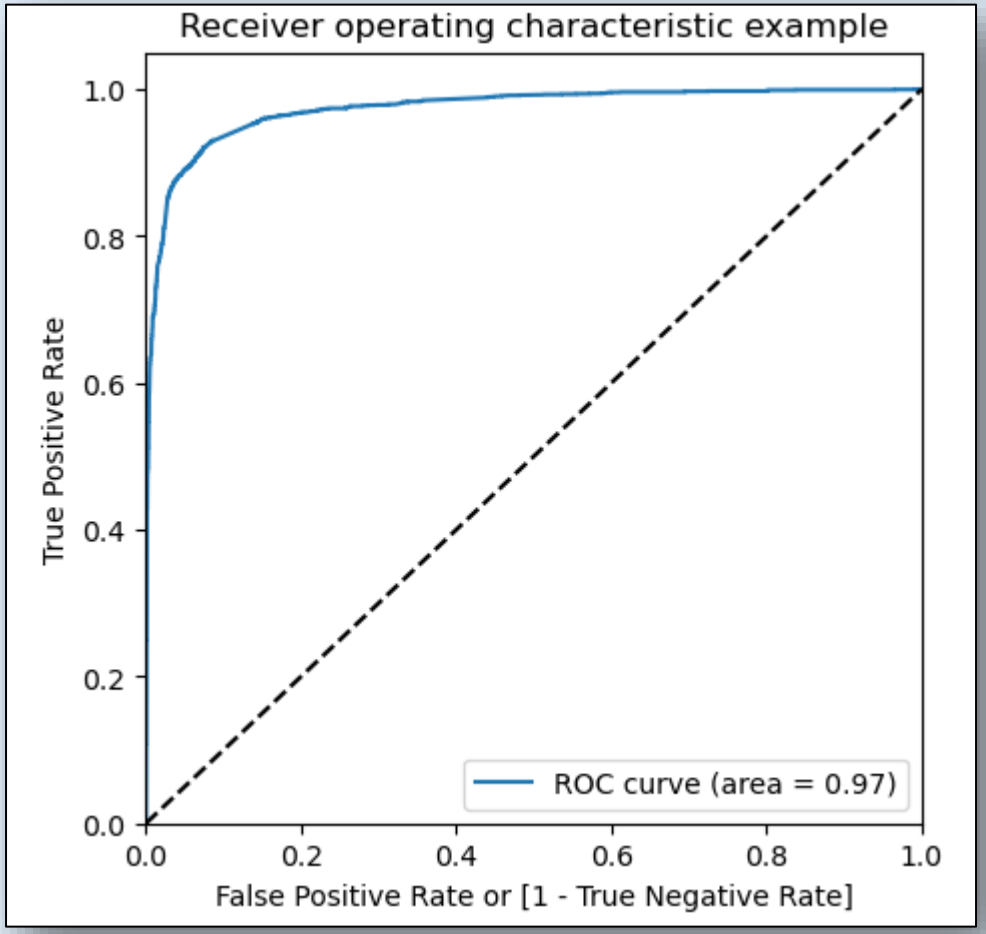


- Leads spending more time on the Website are more likely to convert. The overall conversion rate can be increased by making the website more interesting and interactive to the visitors.
- Not much can be concluded from the above charts for TotalVisits and Page Views Per Vist Variables.

MODEL BUILDING:

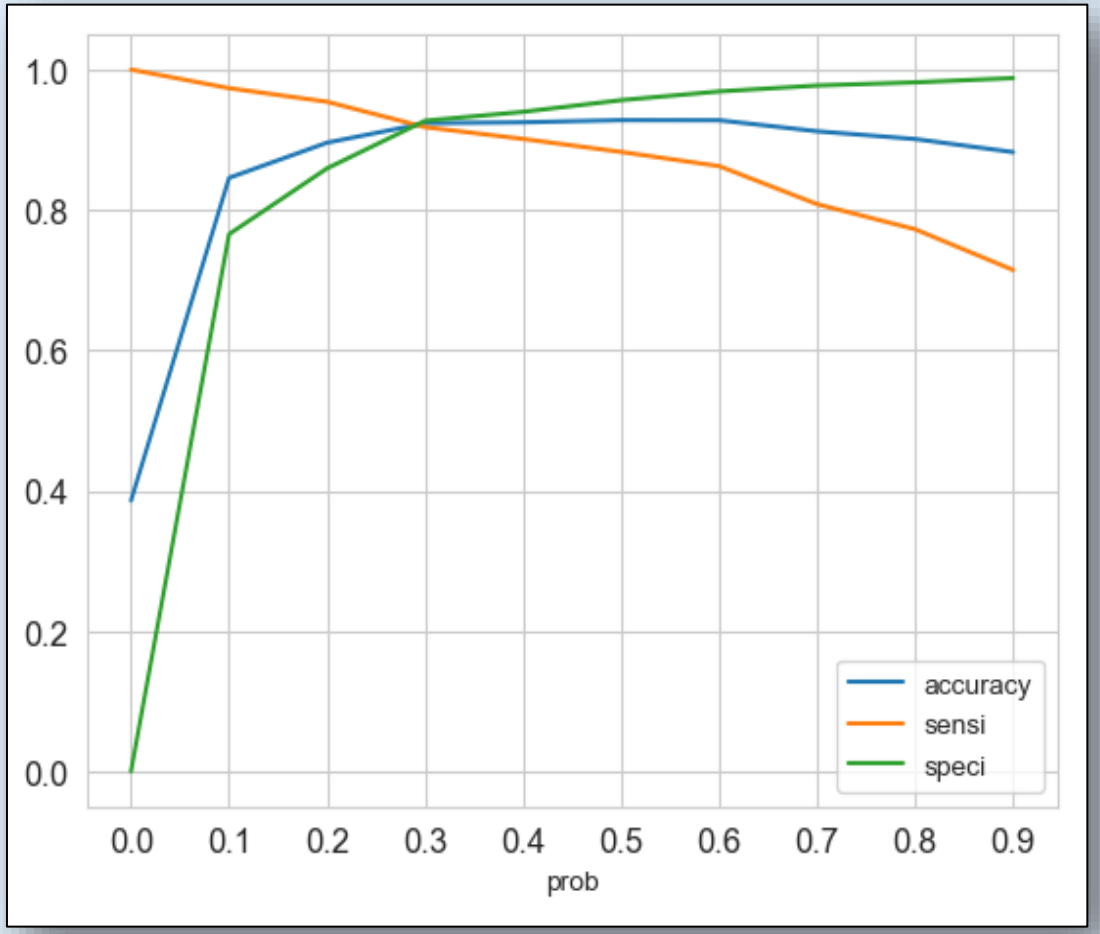
- Dummy Variables have been created for all the categorical columns of the data set.
- The dataset is split into a train set and a test set in the ratio of 70% and 30% respectively.
- MinMax Scaling is used on the numerical columns of the dataset excluding the dummy variables.
- After iterating through three models, the third model was the best fit with all the p-values and VIF values within the acceptable range.
- Observation of the Train Dataset:
    1. Accuracy: 92%
    2. Sensitivity: 91.8%
    3. Specificity: 92.71%
    4. Precision: 88.77%
    5. Recall: 91.8%
    6. False Positive Rate: 7.29%
    7. Positive Predictive Value: 88.77%
    8. Negative Predictive Value: 94.74%
    9. ROC Curve Value: 0.97
- Observation on Test Data:
    1. Accuracy: 92.39%
    2. Sensitivity: 91.8%
    3. Specificity: 92.71%
    4. Precision: 88.77%
    5. Recall: 91.8%

ROC Curve



The ROC Curve is 0.97 which is a good value indicating a good predicting model.

Optimal Cut-off Point



From the above plot, we can see that the optimum point is coming as 0.3 to take it as a cut-off probability.

Conclusion and Recommendation:

- The model is predicting the conversion rate very well and can be recommended to the CEO to make calls basis on it.
- To increase the overall lead conversion, the following points can be helpful:
  - More leads from API and Landing Page Submission needs to convert and more leads can be generated from Lead Add Form, Lead Import, and Quick Add Form.
  - More leads need to be converted from Sources such as Google, Organic Search, Olark Chat, and Direct Traffic whereas more leads need to be generated from Referral Sites, References, and Welingak Website.
  - Converting more leads from the Unemployed population and increasing the number of leads from Students and Working Professionals.
  - Generating more leads from cities other than Mumbai and converting more leads from Mumbai.
  - More leads need to generate from other specializations and most of the leads generated are from Management Specialization.

THANK YOU