LEAD SCORING CASE STUDY

- PRANAV REDDY, MOHAMMED SAIF, MEGHANA REDDY

PROBLEM STATEMENT

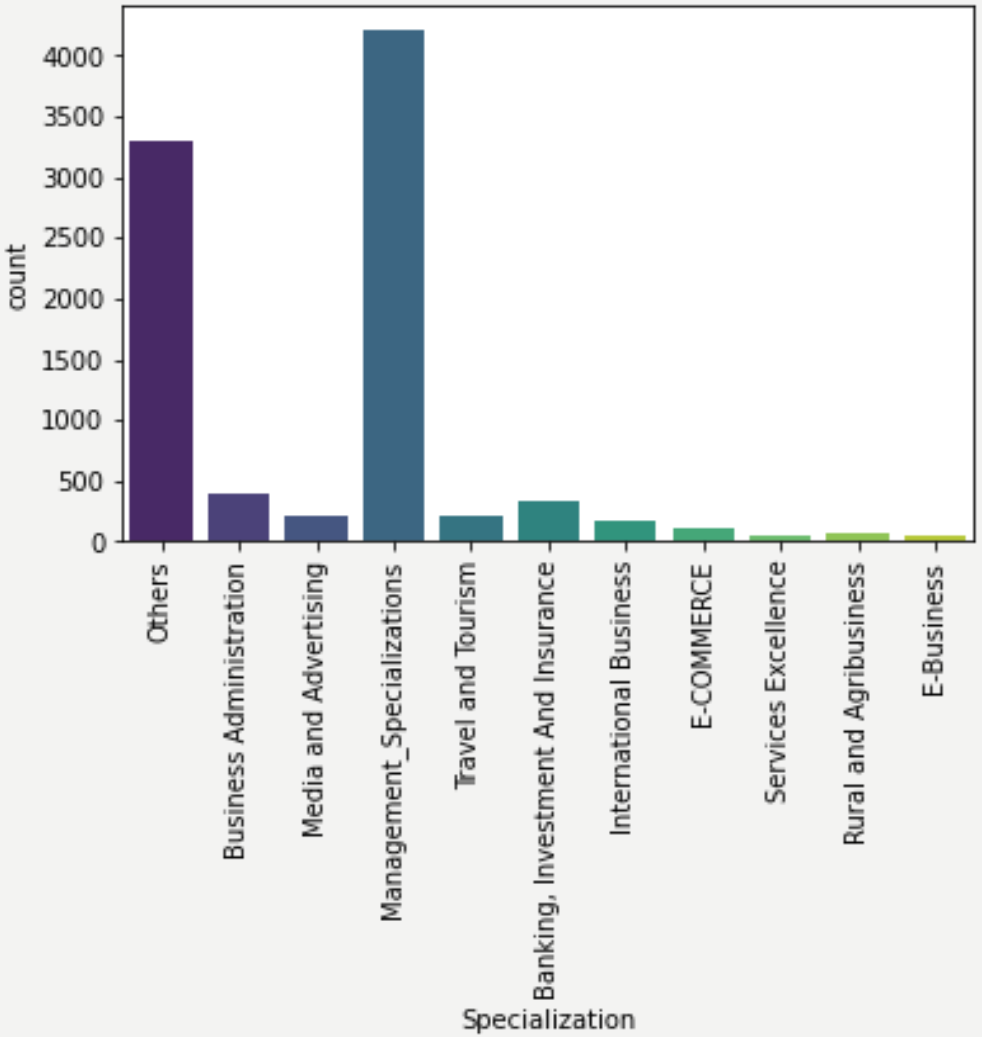
The problem at hand is to identify and prioritize the leads that are most likely to convert into paying customers. This requires developing a model that assigns a lead score to each lead, where higher scores correspond to higher chances of conversion. X Education's current lead conversion rate is about 30%, and the CEO expects a target conversion rate of around 80% by focusing on high-potential leads. The data provided contains various attributes related to leads, such as source, time spent on the website, visits, and last activity, with a target variable indicating whether the lead was converted.

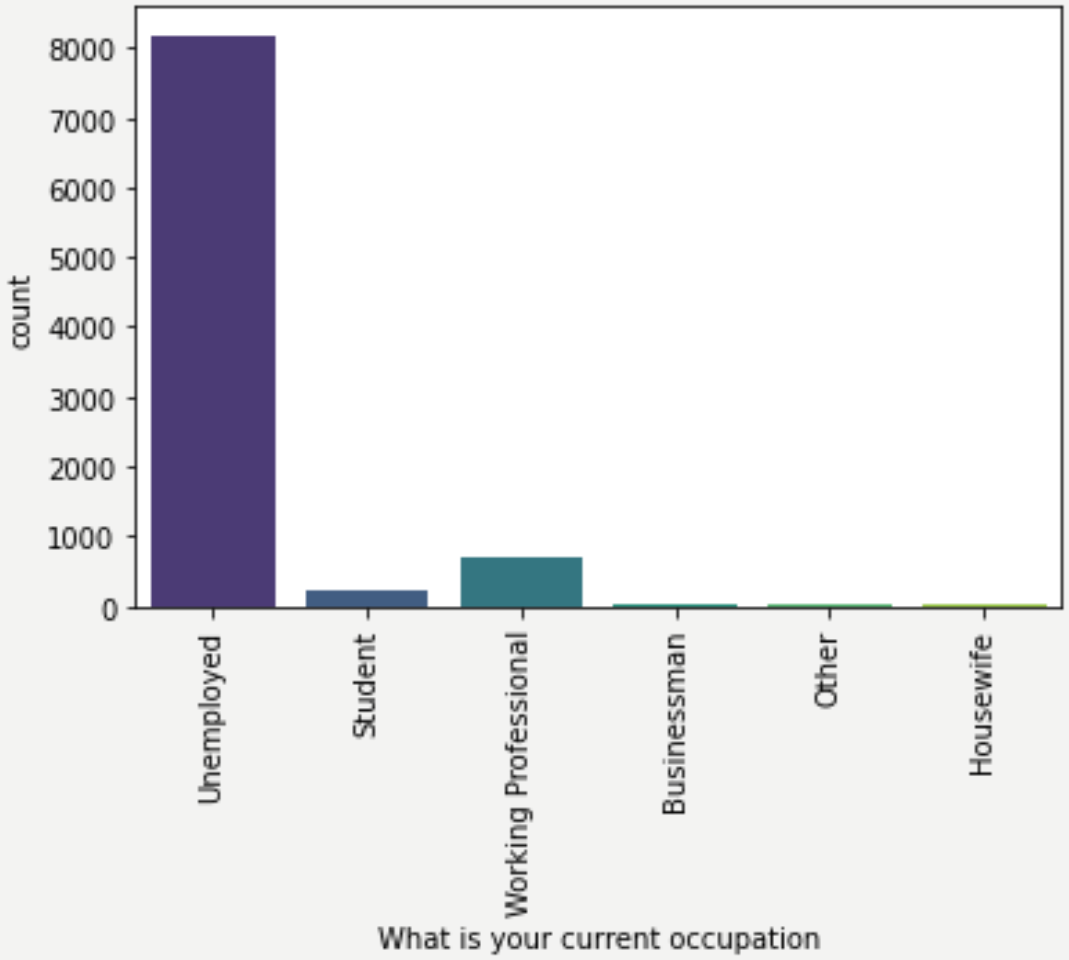
ANALYSIS APPROACH

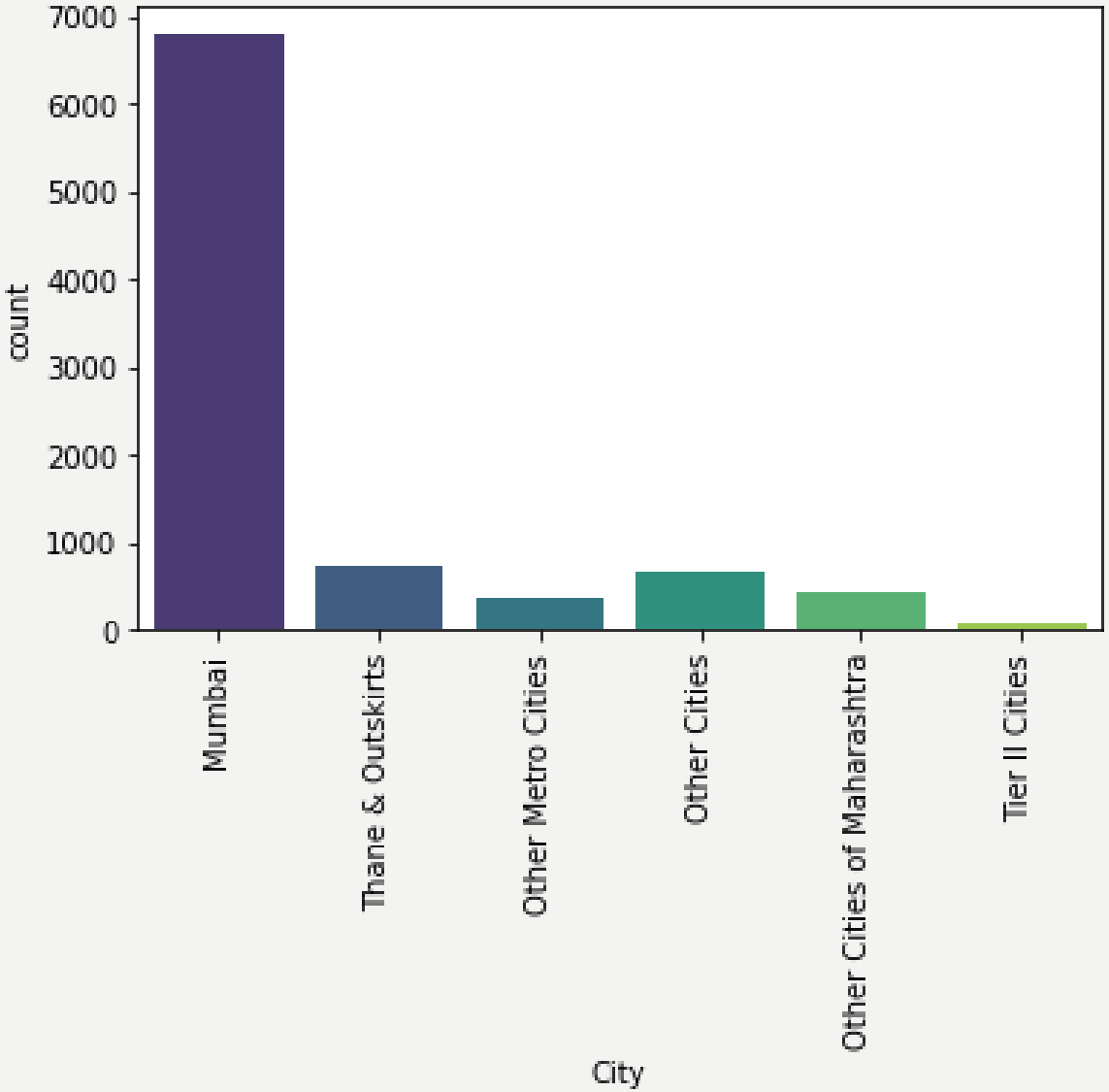
- Importing Data
- Data Inspection
- Data cleaning
- Exploratory Data Analysis
- Creating Dummy Variables for the Categorical Variables
- Model Building Using Logistic Regression
- Prediction On Test Dataset
- Assigning Lead Score with respect to Lead_Num_ID
- Finding out the Hot Leads which should be contacted
- Conclutions

KEY INSIGHTS FROM EDA

UNIVARIATE ANALYSIS

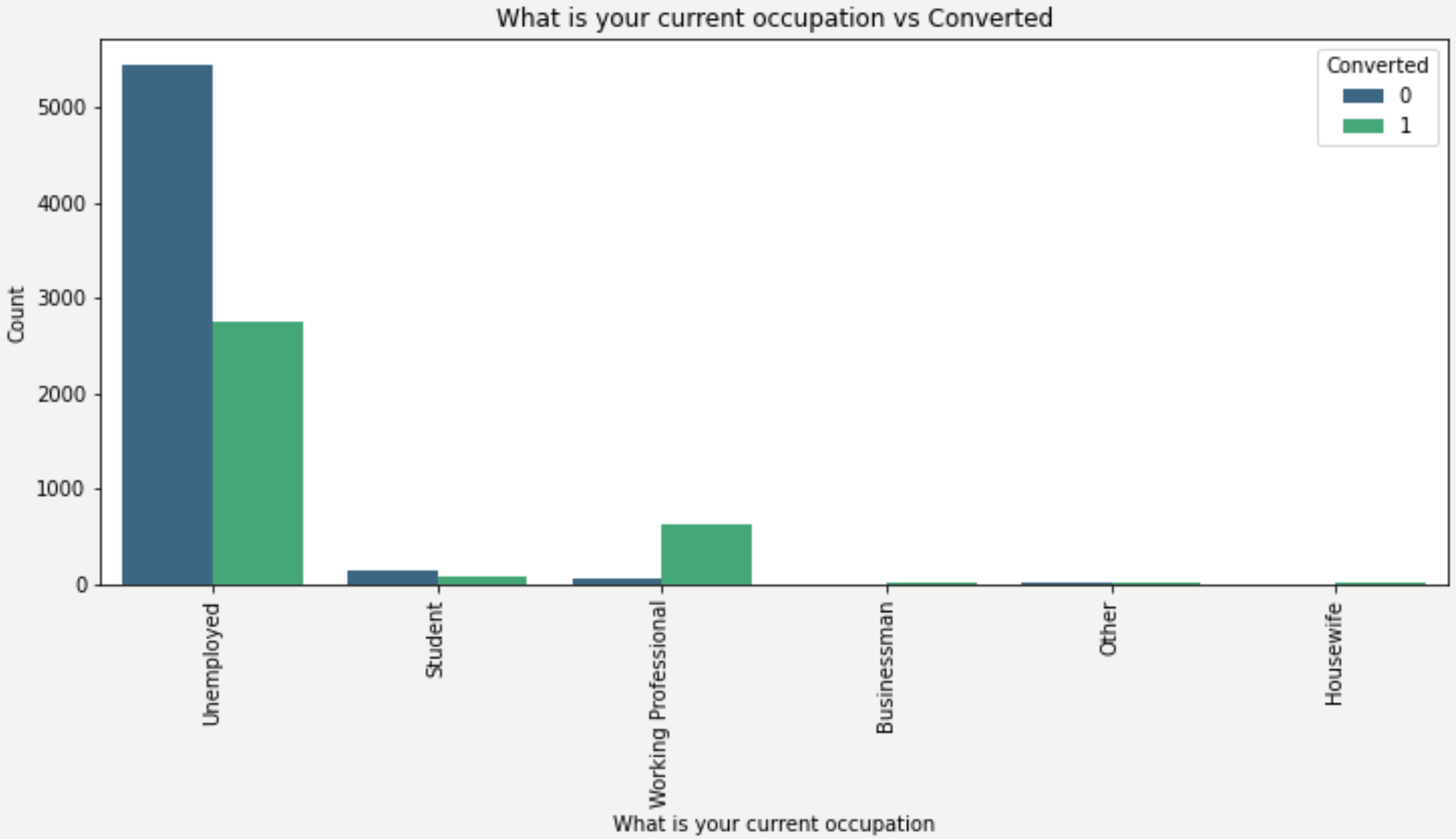






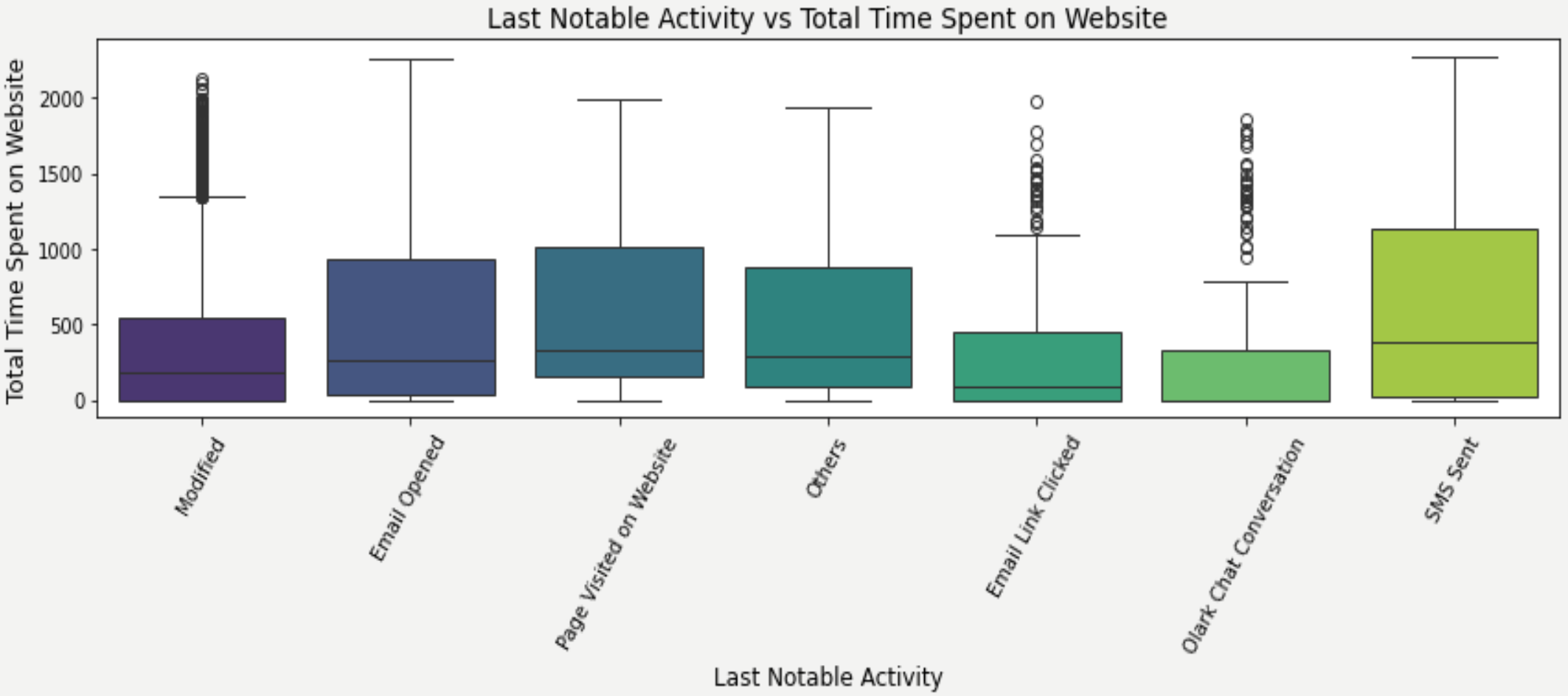
- Google has the highest number of leads as the source, which is logical since most people primarily use Google as their search engine.
- The majority of leads were generated through "Landing Page Submission" and "API" origins.
- Most customers are categorized as unemployed.
- A large number of customers have their specialization listed as "Others" or "Management Specialization."
- Majority of the customers are from Mumbai city.

BIVARIATE ANALYSIS



- Leads originating from the Lead Add Form have a higher conversion rate compared to those from API and Landing Page Submission.
- The conversion rate is highest for leads from the Reference Source, while those from Google, Direct Traffic, and Olark Chat have relatively lower conversion rates.
- SMS Sent has a higher conversion rate compared to Email Opened.
- Customers with "Management" and "Other" specializations show a comparatively higher conversion rate.
- Working professionals have a higher conversion rate than unemployed individuals, likely because they are more aware of current market demands and aim to upskill themselves accordingly.

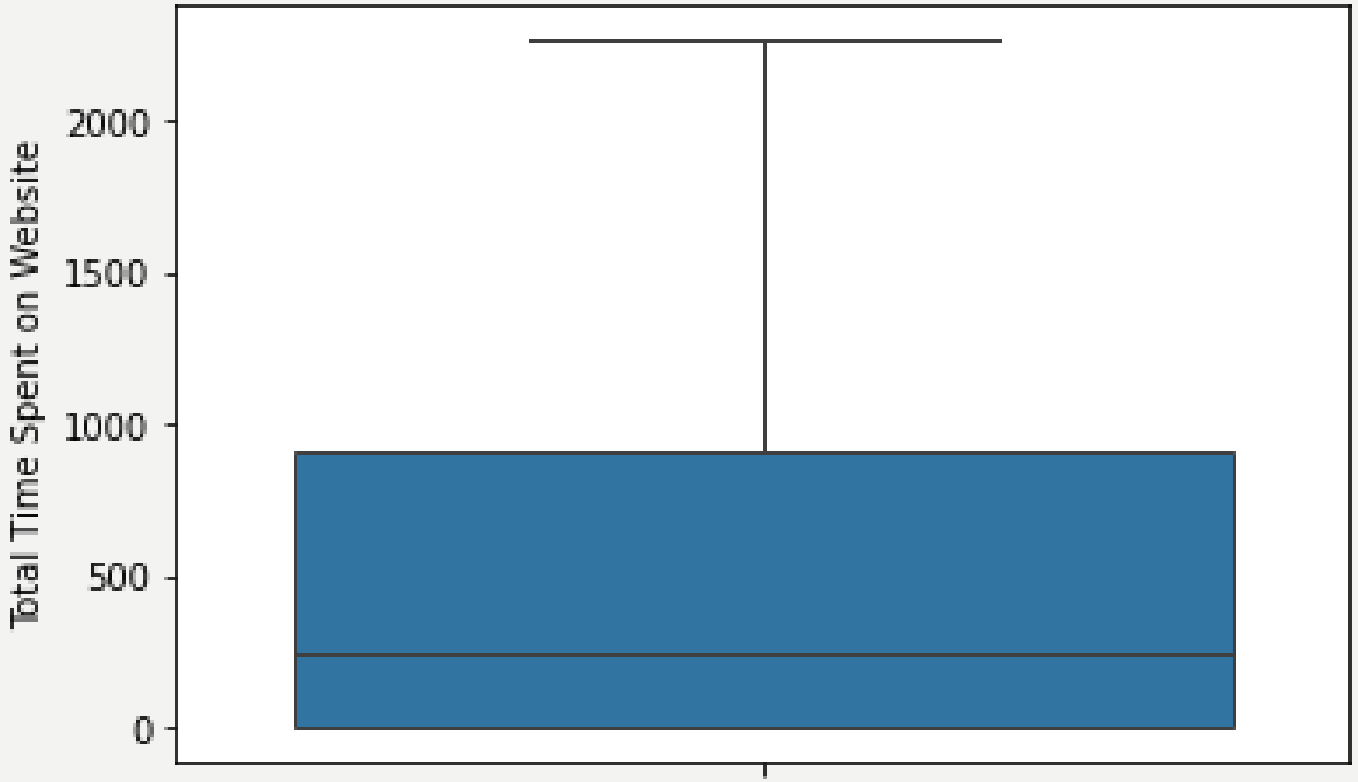
BIVARIATE ANALYSIS (CATEGORICAL VS CONTINUOUS)



•We can see that above boxplots highlight significant differences in numerical variables across various categorical variables that helps us in identifying which groups have higher engagement or activity levels.
•We can also see outliers which indicate unique cases or extreme behaviors within certain categories, which needs to be treated before modelling.

- 'TOTAL VISITS' AND 'PAGE VIEWS PER VISIT' HAVE THE MOST CORRELATION WITH EACH OTHER.

OUTLIER DETECTION AND TREATEMENT



We can see there are no outliers on Total Time Spent on Website.

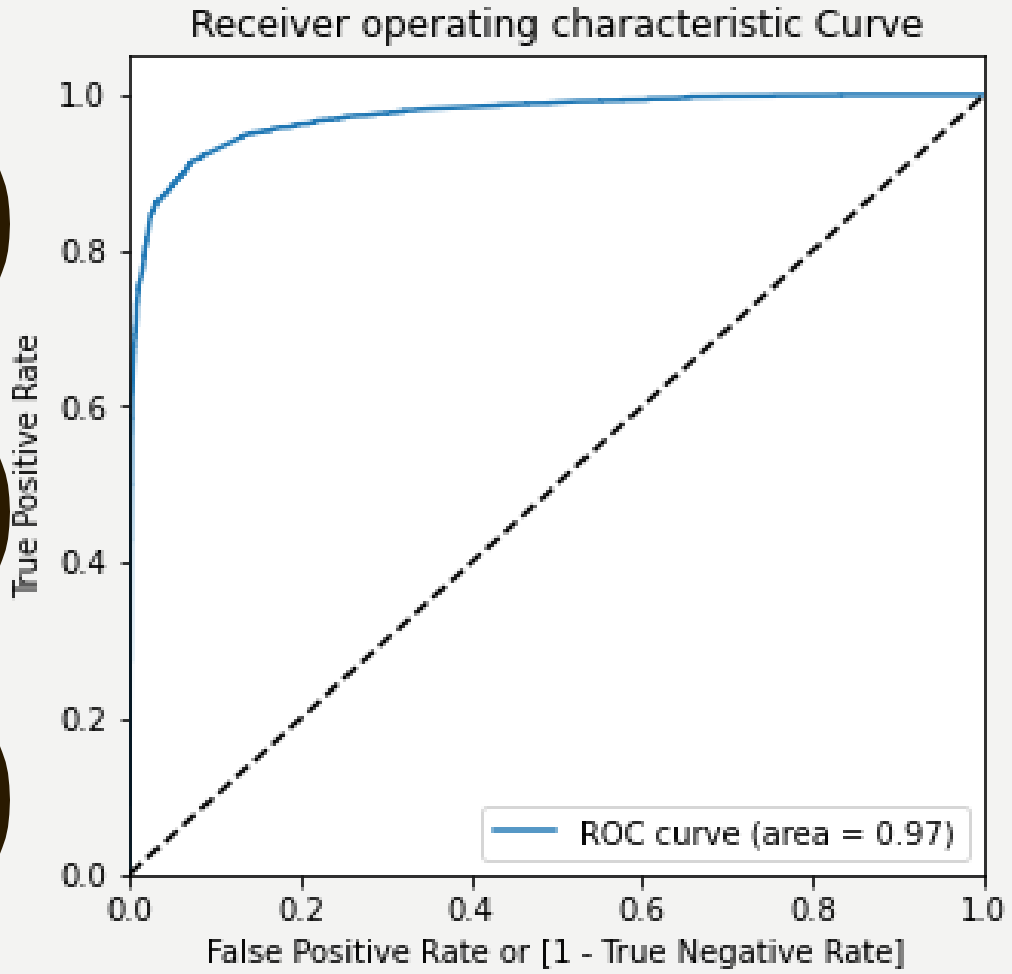
CONFUSION MATRIX

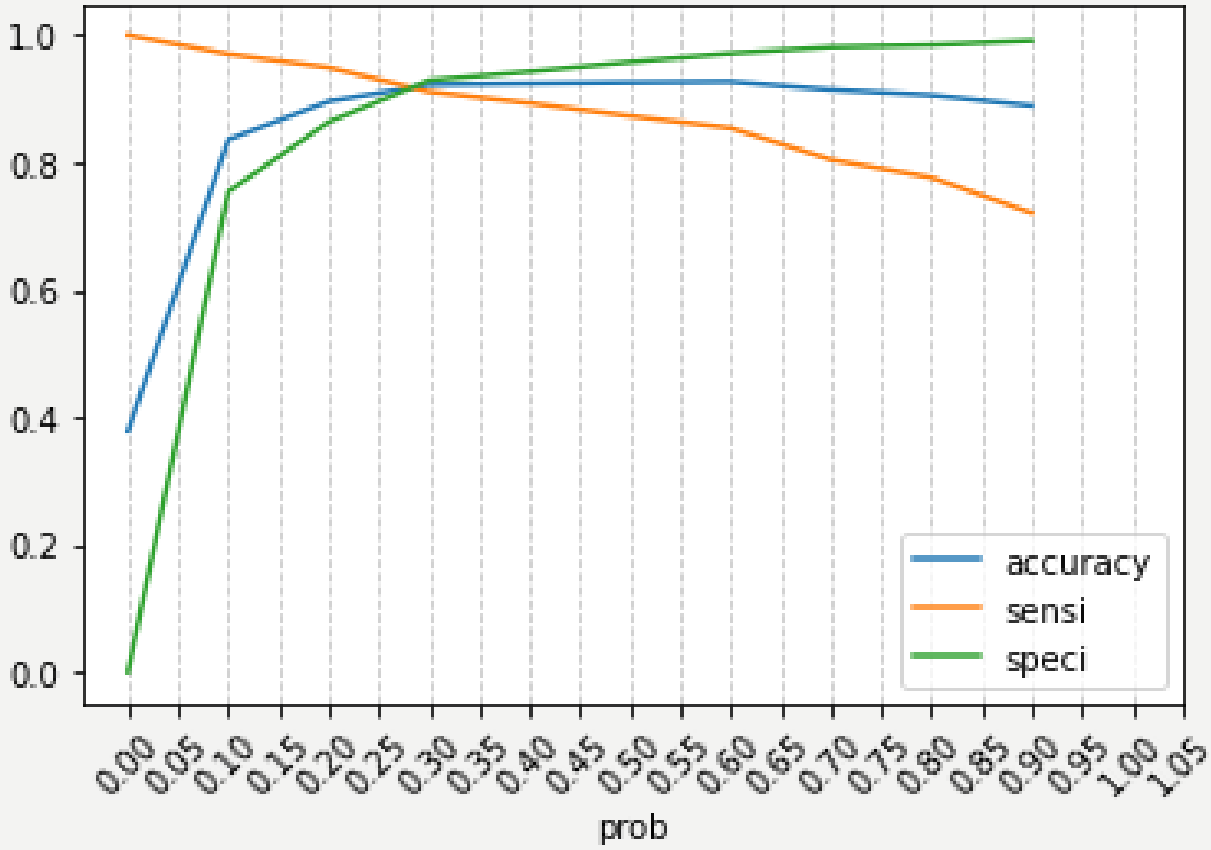
COMPARISION OF THE EVALUATION METRIC RESULTS OBTAINED ON TRAIN AND TEST DATASET:

- Train Dataset
  - Accuracy --> 91.98
  - Specificity --> 92.16
  - Precision --> 87.68
  - Recall --> 91.67

- Test Dataset
  - Accuracy --> 92.65
  - Specificity --> 92.52
  - Precision --> 88.52
  - Recall --> 92.87

ROC CURVE & OPITIMAL CUTOFF





- **ROC-AUC Score:** 0.97
- **Optimal Cutoff:** 0.5

BUSINESS IMPLICATIONS & RECOMMENDATIONS

- Actionable Insights: Focus on high-scoring leads to maximize conversions.
- Allocate sales resources efficiently.
- Impact: Improved conversion rates and reduced time spent on low-priority leads.
- Use lead scoring in CRM systems.
- Regularly update the model with new data for better accuracy.
- Train sales teams to prioritize based on lead scores.

Conclusions

- According to final model, the variables that are important for verifying the Hot Leads are:
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email
- Lead Source_Welingak Website
- Lead Origin_Lead Add Form
- Last Activity_SMS Sent
- Lead Source_Olark Chat
- Total Time Spent on Website

THANK YOU