LEAD SCORE CASE STUDY

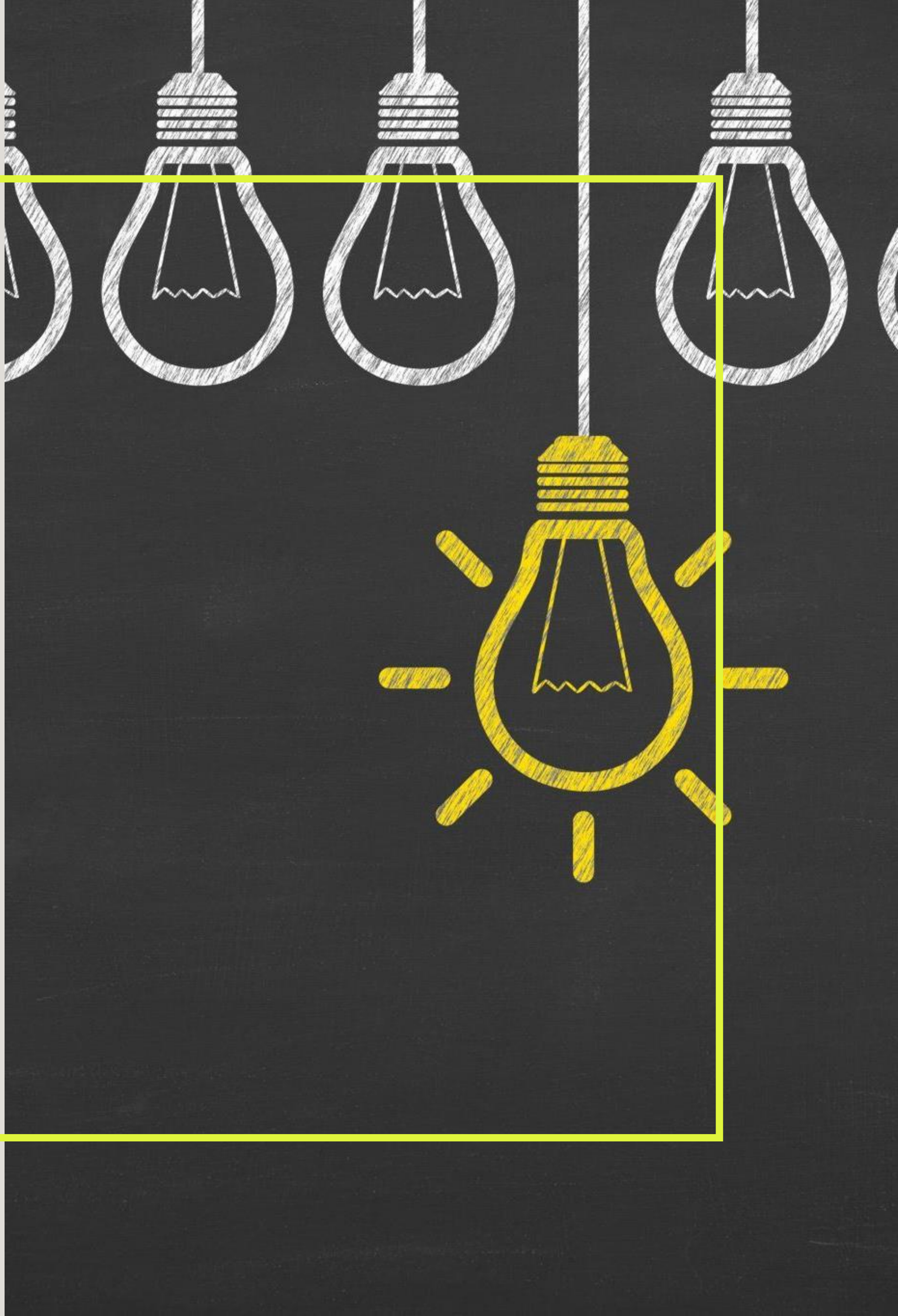
Harithakiran Pakki

Harsh Kant Tiwari

Vrushali Harshe

PROBLEM STATEMENT

- X Education specializes in offering online courses tailored to industry professionals. Despite the company's significant lead generation efforts, its current lead conversion rate remains suboptimal.
- In an effort to enhance the efficiency of this process, X Education aims to distinguish the leads with the greatest potential, commonly referred to as 'Hot Leads.'
- By effectively identifying this subset of leads, X Education anticipates a notable increase in the lead conversion rate. This improvement will be achieved by redirecting the sales team's focus towards cultivating relationships with these high-potential leads, thereby reducing the need for broad outreach efforts and ensuring more fruitful interactions.



BUSINESS OBJECTIVE

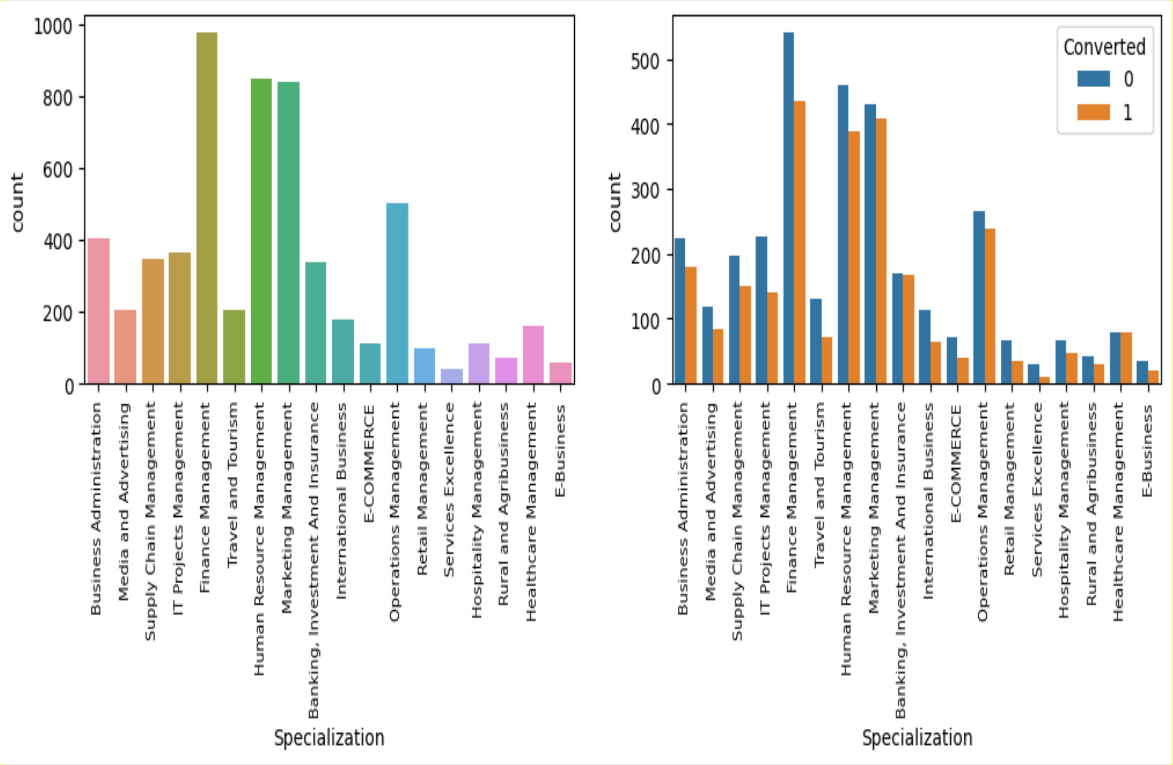
- X Education seeks to discern and prioritize the most promising leads in their lead acquisition process. To achieve this objective, the company is actively engaged in the development of a sophisticated model designed to identify and categorize 'Hot Leads' accurately.
- Furthermore, X Education plans to deploy this model for future utilization, thereby enhancing their lead conversion efforts by directing resources towards engaging with high-potential leads, thereby optimizing sales team performance and efficiency.

METHODOLOGY

- Data cleaning and data manipulation.
- Handle duplicate data.
- Handle NA values and missing values.
- Drop columns, if it contains large number of missing values.
- Imputation of the values.
- Handle outliers in data.
- EDA
- Univariate data analysis
- Bivariate data analysis

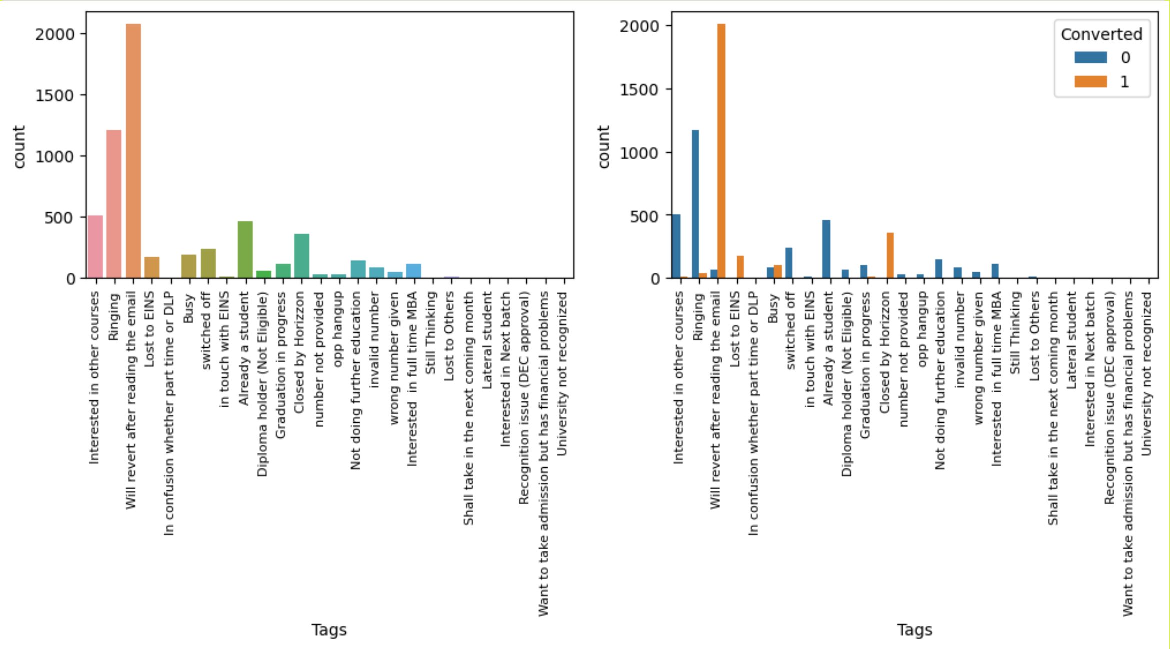
METHODOLOGY

- The data preprocessing stage involves Feature Scaling and the creation of Dummy Variables for appropriate data encoding.
- Subsequently, the chosen classification technique, Logistic Regression, is employed for model construction and prediction.
- A rigorous validation process is conducted to assess the model's accuracy and effectiveness.
- Following validation, the model's findings and insights are presented comprehensively.
- In conclusion, the study offers valuable insights, and based on these findings, recommendations are provided to guide informed decision-making and further actions.



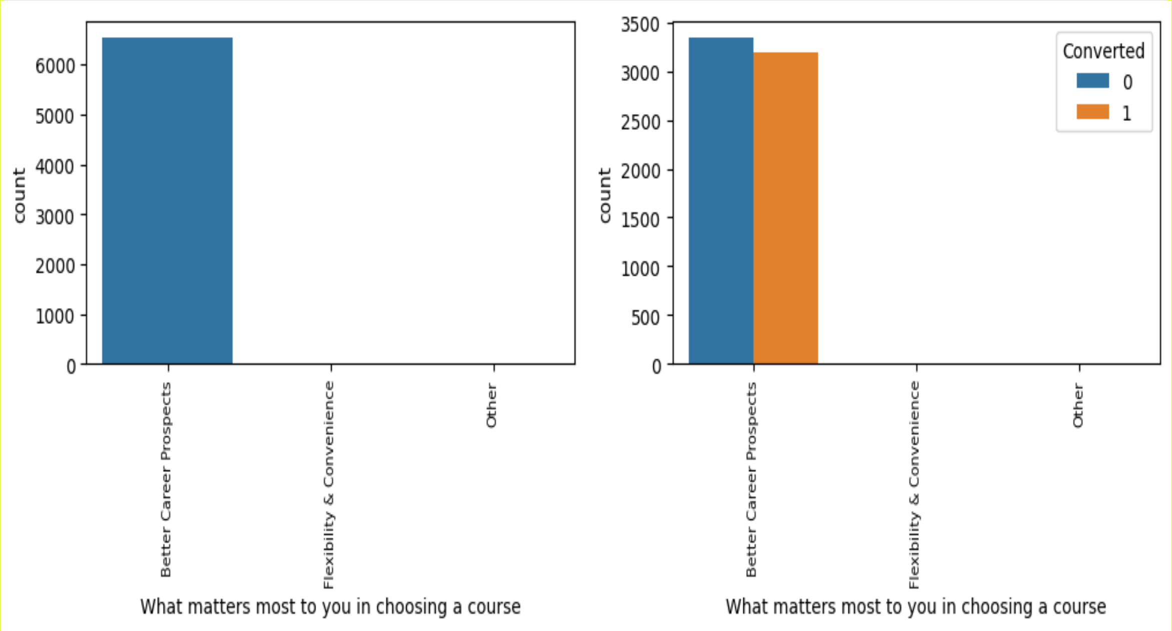
CATEGORICAL PLOT

These two categorical plots show the number of courses offered by the administration and the second plot shows the conversion rate for each courses.

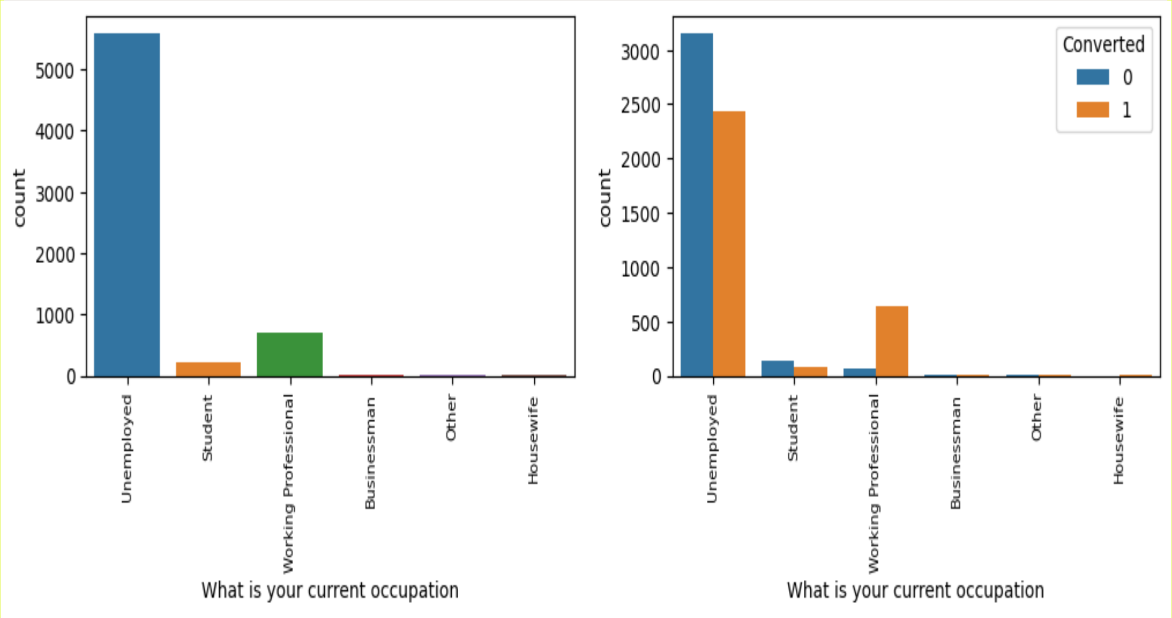


CATEGORICAL PLOT

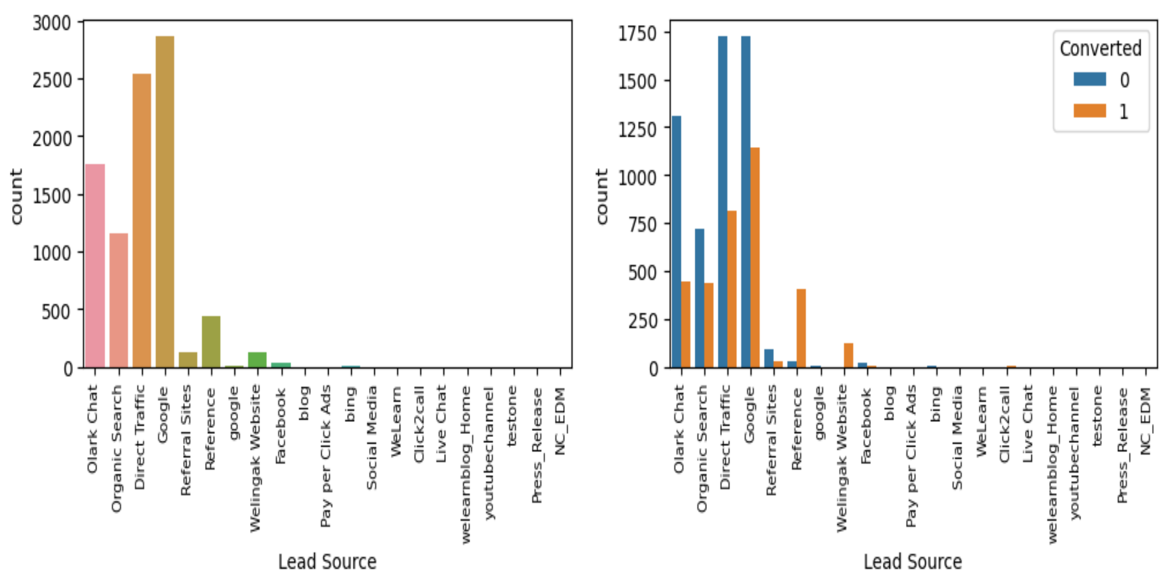
These two categorical plot shows the behaviors for customers and conversion rate as per the behavior with respect to how they respond to them.



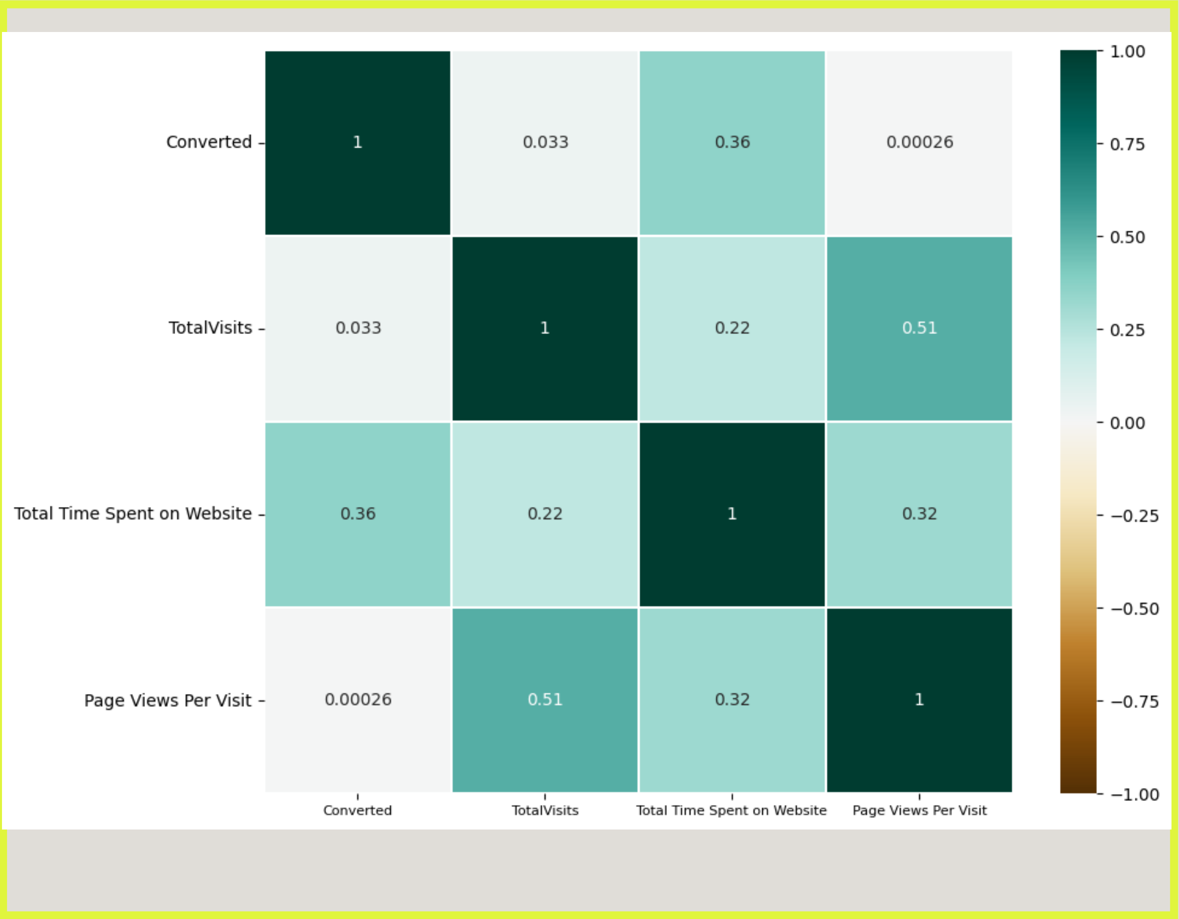
This plot shows the reason for people choosing the course and their respective conversion.



This categorical plot shows the occupation for the leads and as per this we see that people who are unemployed are at the highest and similarly the second plot shows their conversion rate.

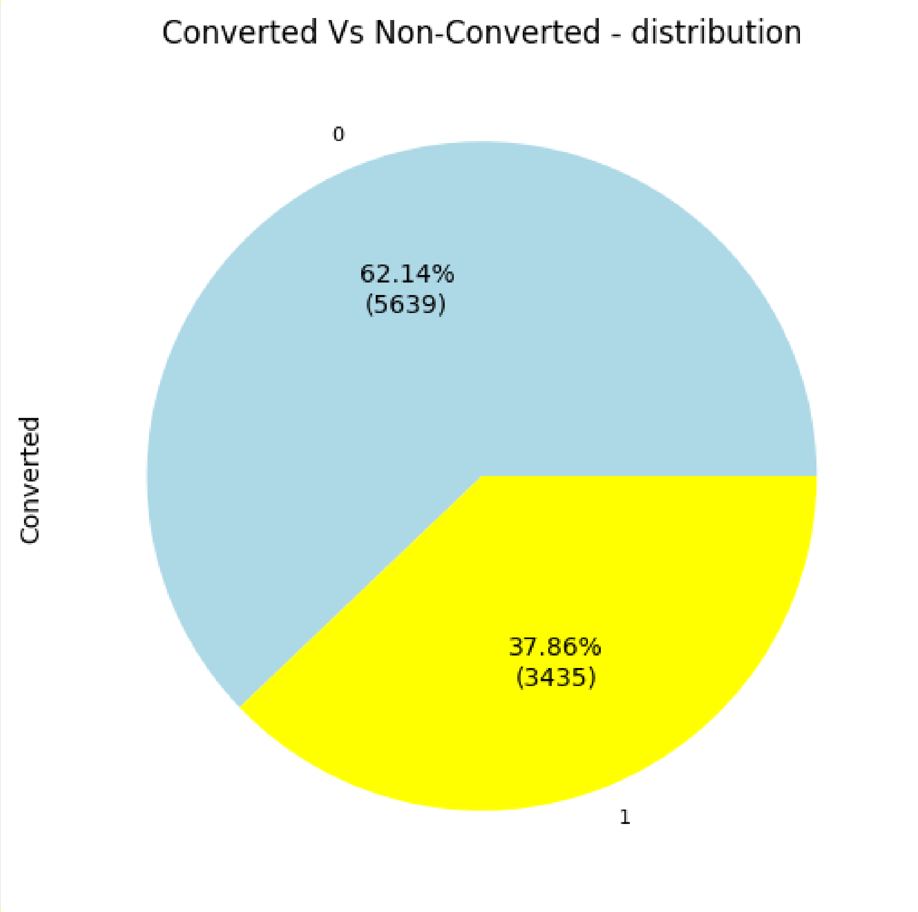


This categorical plot shows the lead sources and if we observe, we notice that google is one of their highest lead source and the second plot shows the conversion rate of leads based on the source.

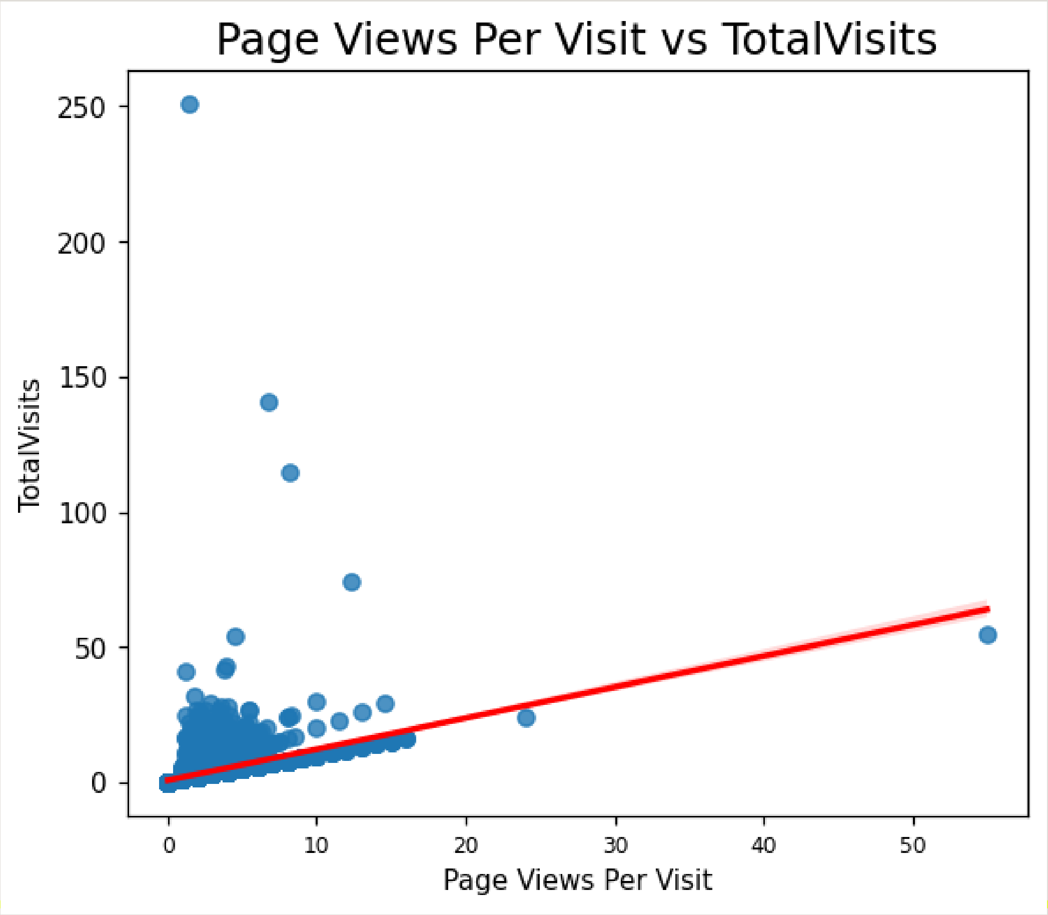


HEAT MAP

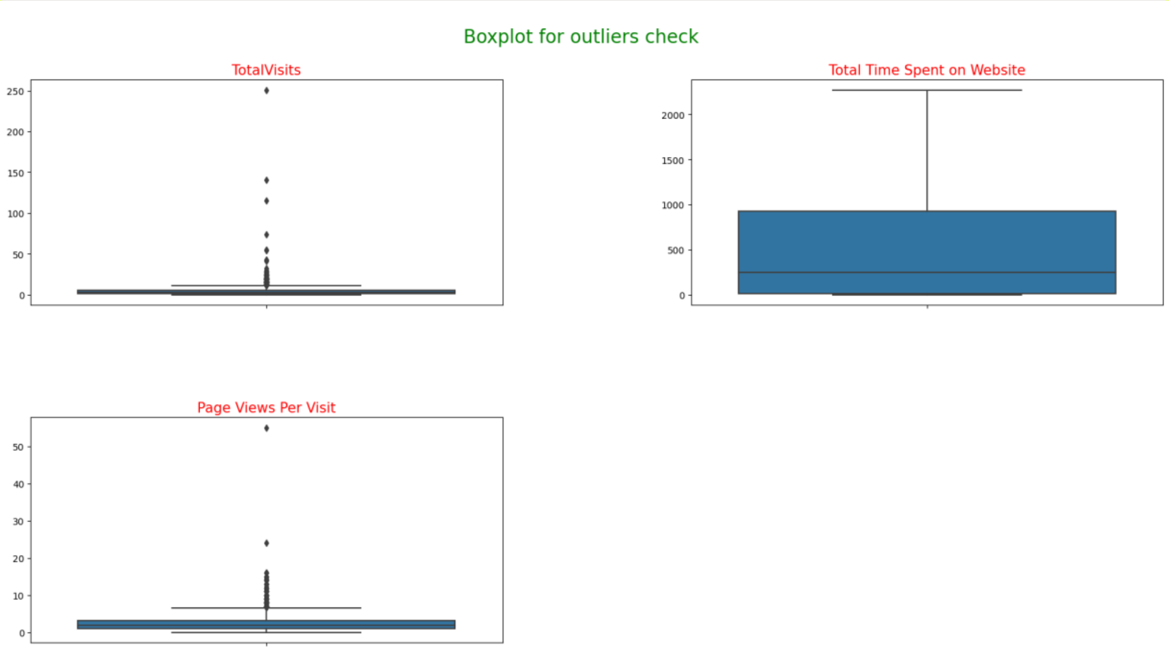
- The metric 'Page Views Per Visit' demonstrates a correlation with the 'TotalVisits' column, suggesting a notable association between these two variables.
- Moreover, 'Total Time Spent on Website' exhibits a high degree of correlation with the 'Converted' metric. This strong correlation underscores the critical importance of 'Total Time Spent on Website' in influencing lead conversion outcomes.



This pie plot shows the conversion rate and non-conversion rate. The conversion rate lies at 62.14 % compared to non-conversion rate which is at 37.86%



In this plot, we observe that Page Views Per Visit' is positively correlated with TotalVisits column.



BOX PLOT

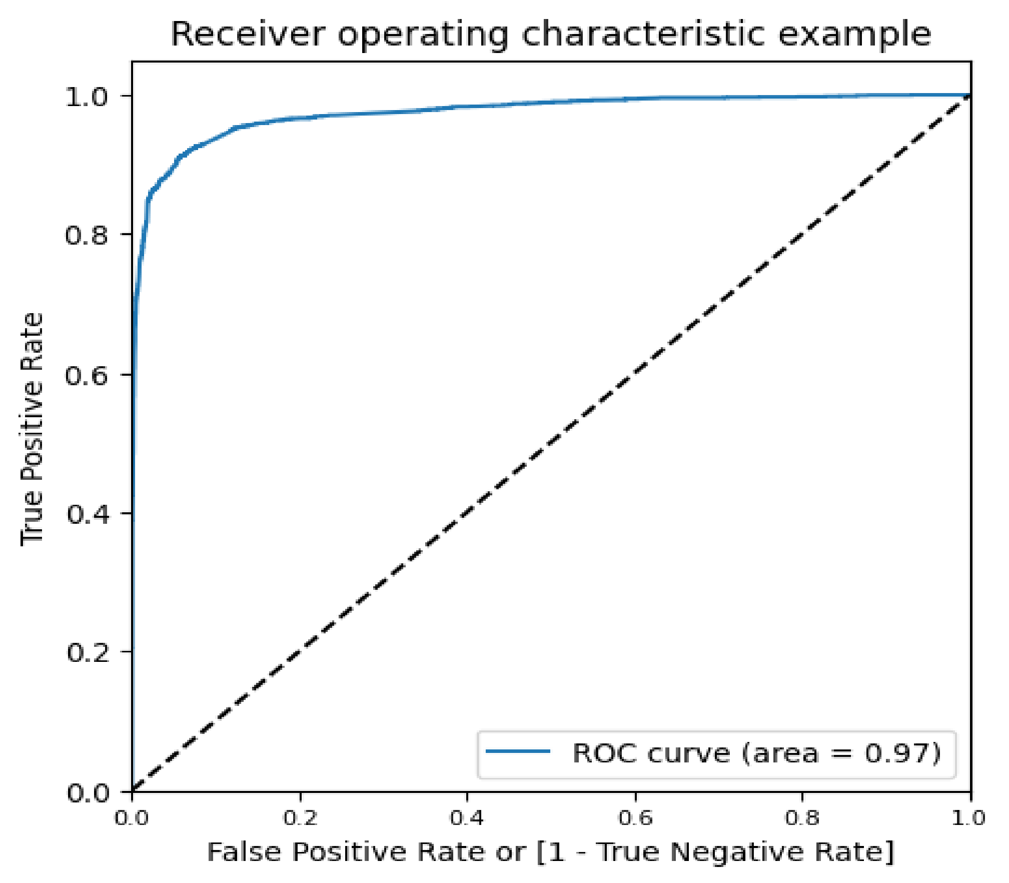
In this box plot, we check for any outlier in TotalVisits and Page Views Per Visit.

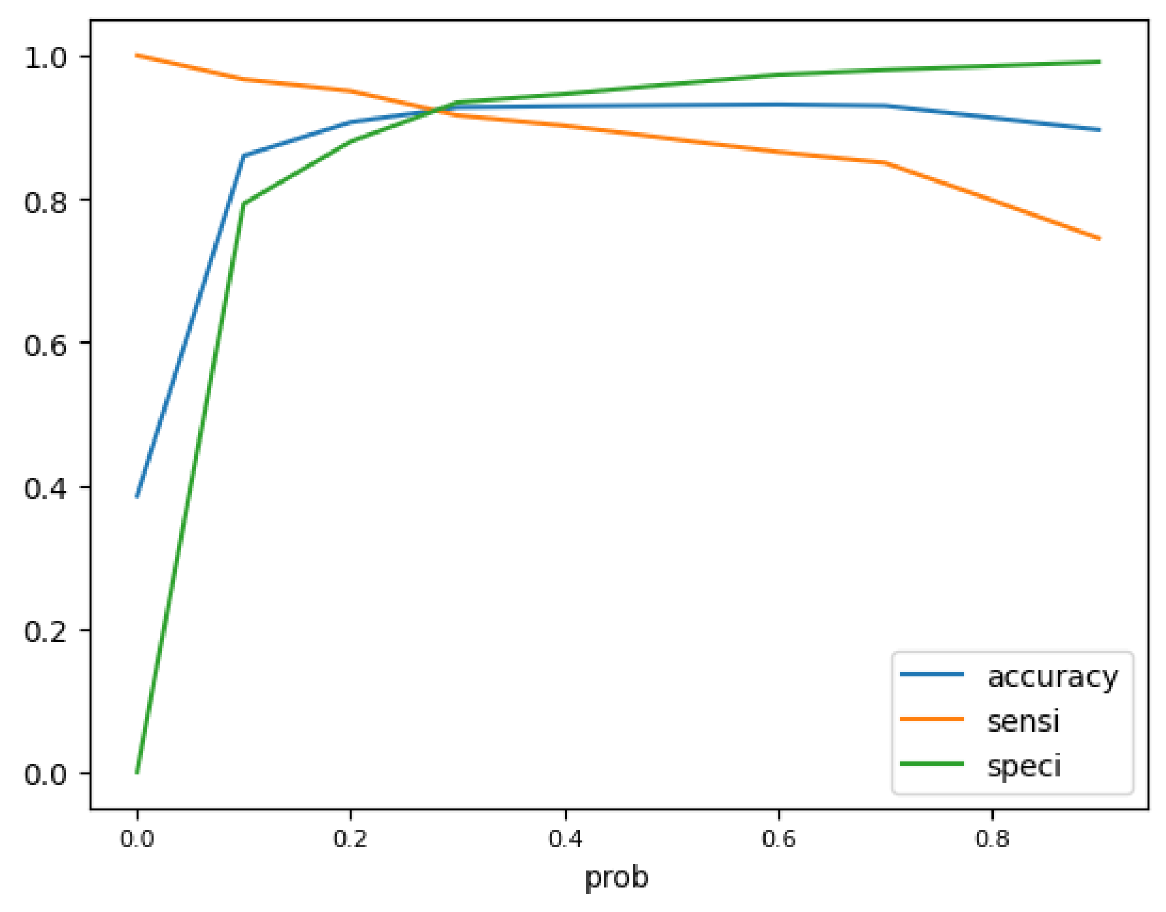
MODEL BUILDING

- The initial step in our regression analysis entails partitioning the dataset into distinct training and testing subsets. We have opted for a partition ratio of 70:30, where 70% of the data is allocated to the training set.
- Following the data split, we employ Recursive Feature Elimination (RFE) as a feature selection technique. RFE is executed with the objective of retaining the 15 most relevant variables.
- Subsequently, we construct the regression model by iteratively eliminating variables. All VIFs are < 2.0 and p values are all variables are 0.
- We apply this model to make predictions on the test dataset, ultimately achieving an overall predictive accuracy of 92.7%.

PLOTTING THE ROC CURVE

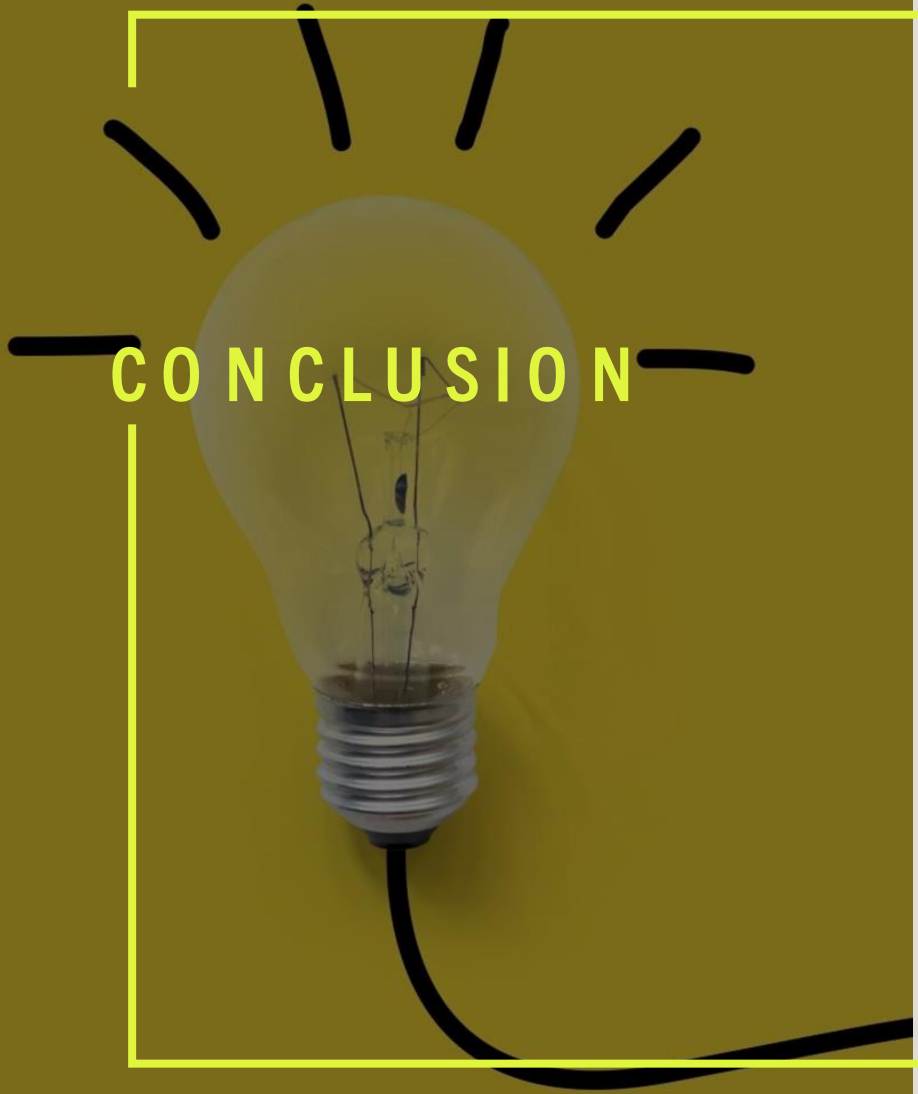
The Area under the Receiver Operating Characteristic (ROC) curve, denoted as 0.97, serves as an indicator of the model's predictive performance. In this context, an ROC AUC score of 0.97 signifies the model's capacity to make robust and accurate predictions, affirming its quality as a highly effective predictive model.





FINDING THE OPTIMAL CUT OFF POINT

If we observe, we find that 0.3 is the optimum point to take it as a cutoff probability.



CONCLUSION

- The model demonstrates a robust performance, with a sensitivity of 92% in the training dataset and 90% in the test dataset, employing a threshold value of 0.3.
- Sensitivity, in this context, signifies the model's ability to accurately identify and classify the relevant leads.
- It is noteworthy that the CEO of X Education had established a target sensitivity of approximately 80%, and the model has surpassed this threshold by a significant margin.
- Furthermore, the model attains an overall accuracy of 93%, surpassing the predefined accuracy threshold.
- This achievement underscores the model's effectiveness in correctly classifying leads, surpassing the CEO's expectations.