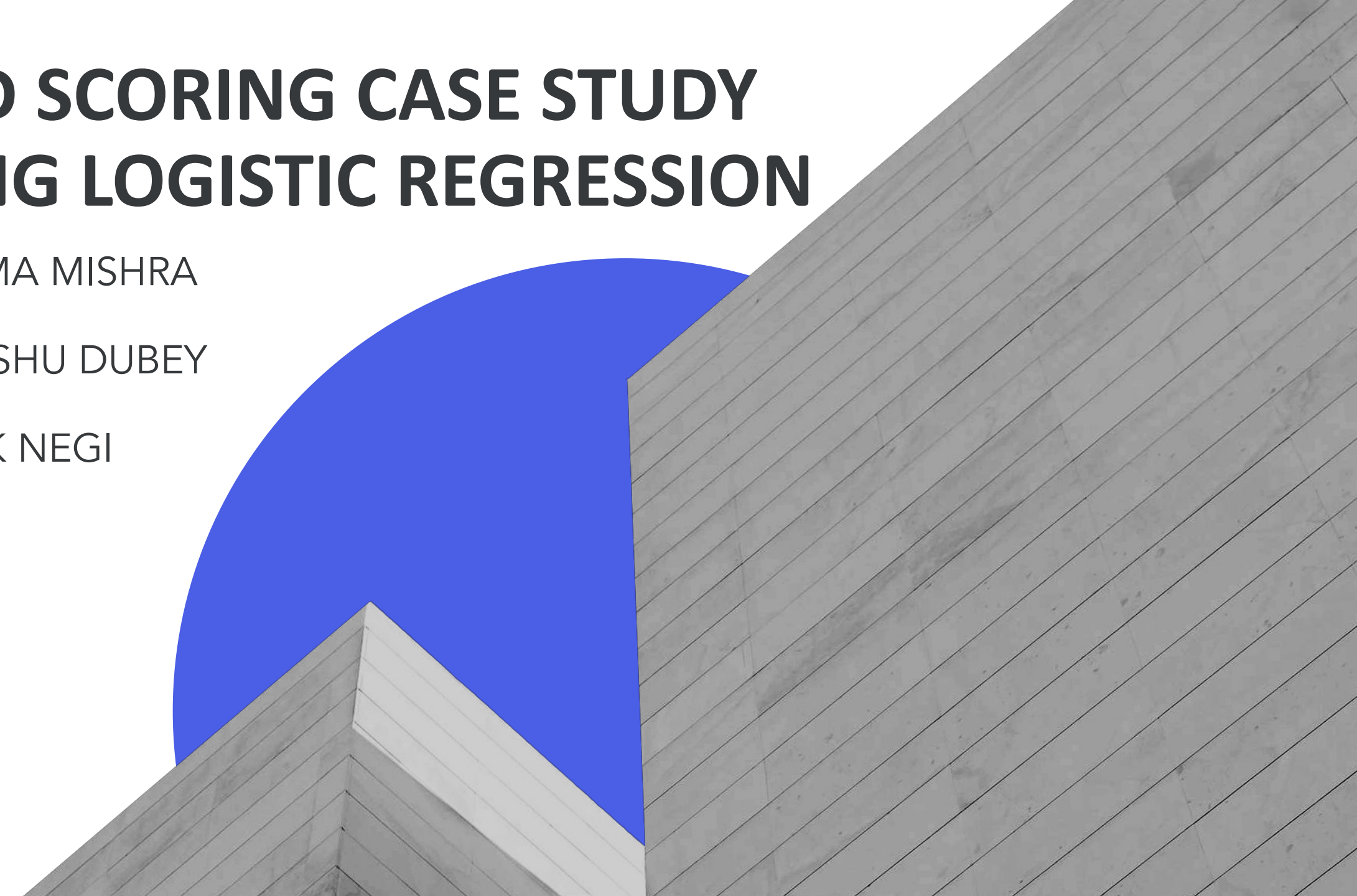
LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

ARUNIMA MISHRA

HIMANSHU DUBEY

HRITHIK NEGI



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

- Title: "Optimizing Lead Conversion for X Education"
- In line with X Education's mission to provide valuable online courses, our primary business objective is to significantly improve lead conversion rates.
- Our target, as set by the CEO, is to achieve an 80% lead conversion rate.
- This presentation will detail our analysis and recommendations to achieve this objective, providing actionable insights for the future success of X Education.

Problem Approach and Methodology

1. Data Import and Inspection

- Imported the dataset consisting of approximately 9,000 data points.
- Inspected the data frame to understand its structure and characteristics.

2. Data Preparation

- Performed data cleaning, addressing inconsistencies, duplicate records, and handling missing values.
- Resolved the issue of 'Select' levels in categorical variables, treating them as null values.
- Conducted feature selection and engineering to enhance model performance.

3. Exploratory Data Analysis (EDA)

- Utilized EDA techniques to gain insights into the dataset.
- Examined the distribution of the target variable 'Converted' and assessed correlations between features.
- Identified potential patterns and trends to guide model development.

4. Dummy Variable Creation

- Created dummy variables for categorical features to prepare the data for modeling.
- Ensured that the model could effectively use categorical variables in the analysis.

5

. Test-Train Split

- Split the dataset into training and testing subsets.
- The training set was used to build and train the model, while the testing set was reserved for model evaluation.

6. Feature Scaling

- Scaled the numerical features to ensure that they contribute to the model in a balanced way.
- Utilized standardization or normalization to improve model performance.

7. Model Building

- Employed a logistic regression model as the primary modeling technique.
- Implemented Recursive Feature Elimination (RFE) to select the most relevant features.
- Calculated Variance Inflation Factor (VIF) to assess multicollinearity among features.
- Analyzed p-values to identify significant features for the model.

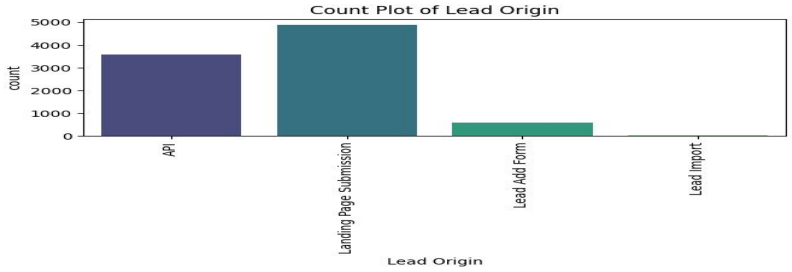
8. Model Evaluation

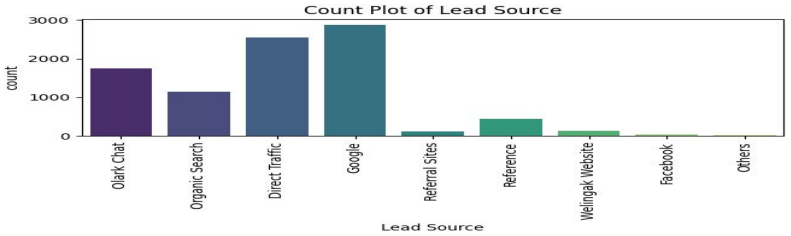
- Evaluated the logistic regression model using industry-standard metrics such as accuracy, precision, recall, and ROC-AUC.
- Assessed the model's ability to predict lead conversions accurately.

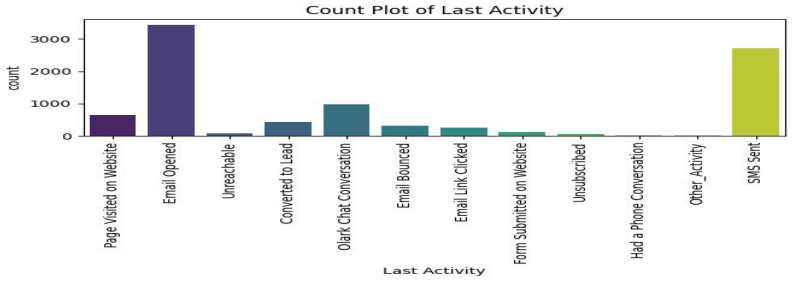
9. Making Predictions on the Test Set

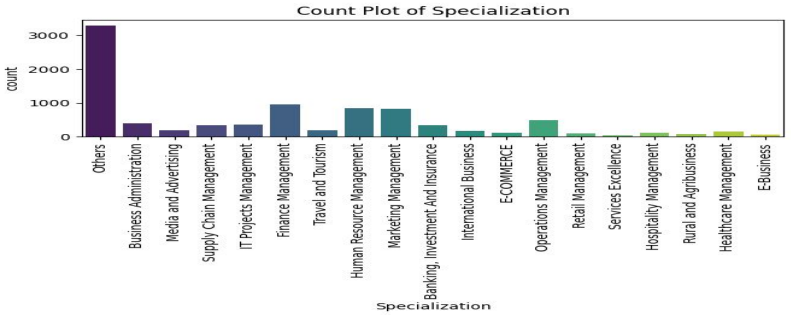
- Applied the trained model to make predictions on the test dataset.
- Evaluated the model's predictions to measure its effectiveness in identifying potential leads.

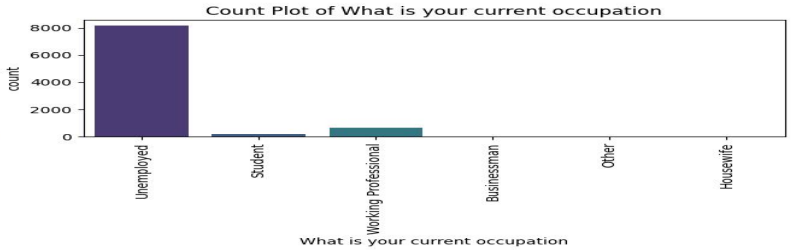
Univariate Categorical Columns

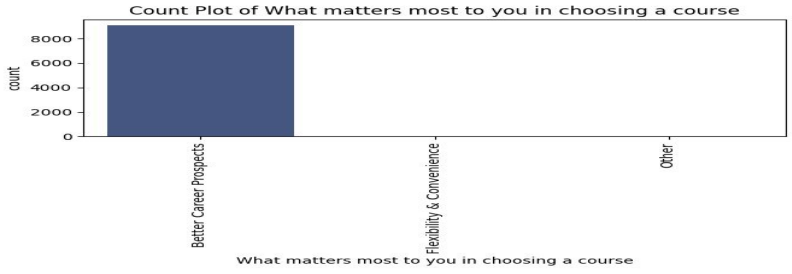


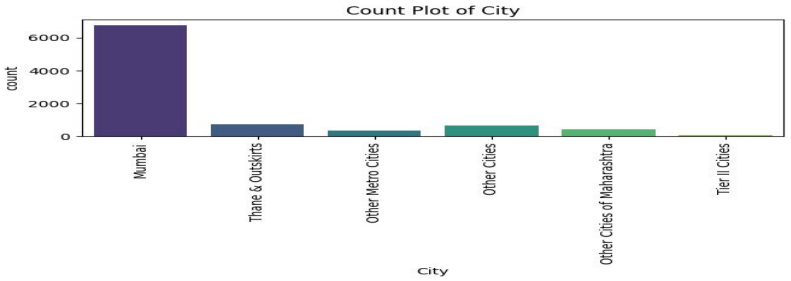


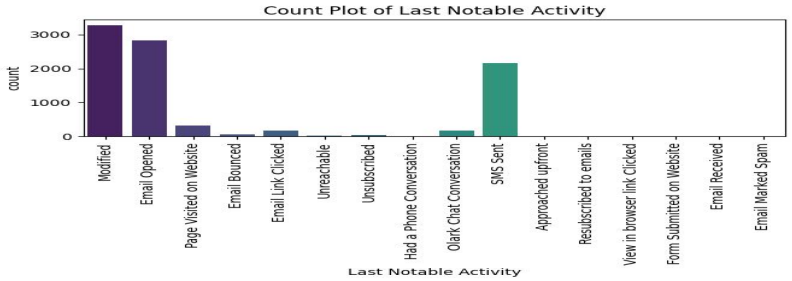


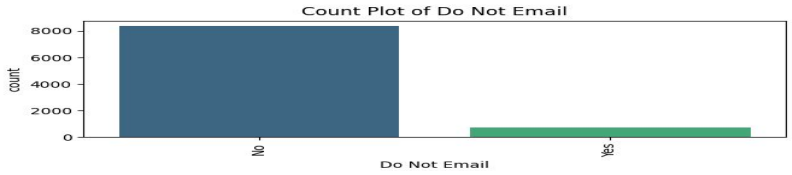


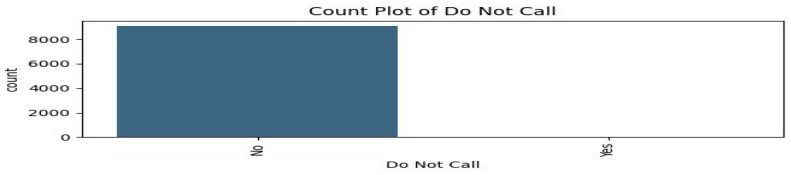


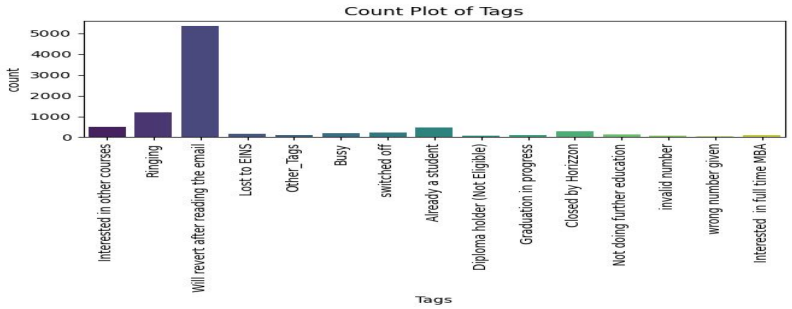












1. Lead Origin
- "Landing Page Submission" and "API" are the two most frequent lead sources.
- We should Allocate more resources and tailor strategies to nurture these high-potential leads.

2. Lead Source
- "Google" stands out as the most prevalent lead source, signifying its significant role.
- "Facebook" and "Others" have comparatively lower representation.

3.Last Activity
- "Email Opened" is the most frequent last activity, signifying high engagement.
- "SMS Sent" follows as the second most frequent activity, suggesting potential for lead nurturing through SMS.

4. Lead Quality
- "Might be" is the most frequently occurring lead quality category.
- Strategies should be adapted to engage and convert "Might be" leads effectively.

5. Specialization
- "Others" is the most frequent specialization among leads, indicating diversity in interests.
- "Finance" and "Management" is the second most common specializations.

6. What is your Current occupation
- "Unemployed" is the most prevalent occupation among leads, indicating a significant segment of the audience seeking opportunities.
- "Working Professional" ranks as the second most common occupation, showcasing potential interest from employed individuals seeking further education.

7. City
- "Mumbai" emerges as the most frequently mentioned city among leads, indicating a substantial regional presence.

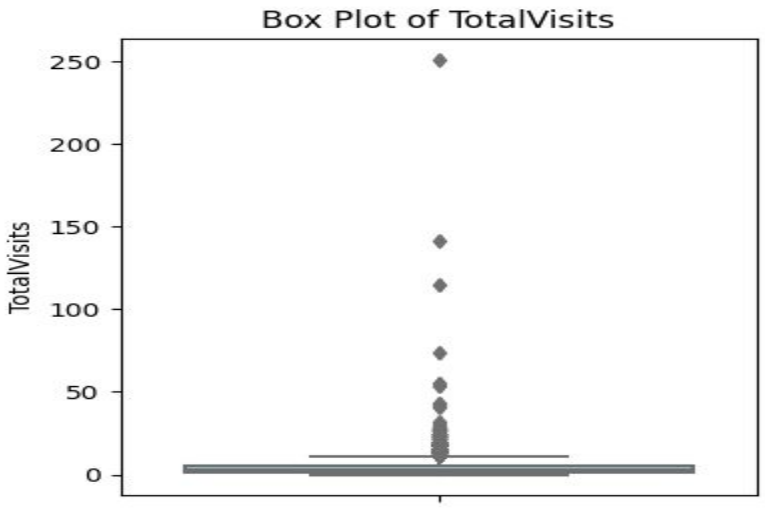
8. Last Notable Activity
- "Modified" in the "Last Notable Activity" category is the most prevalent, indicating significant lead engagement.
- "Email Opened" closely follows as the second most common activity, reflecting active interest in communication.

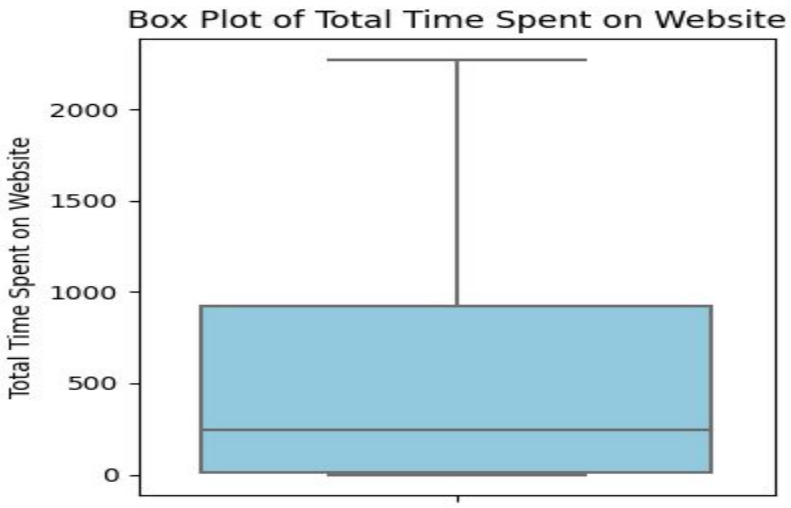
9. Do Not Email
- The majority of leads have "No" in the "Do Not Email" category, indicating a willingness to receive communication.

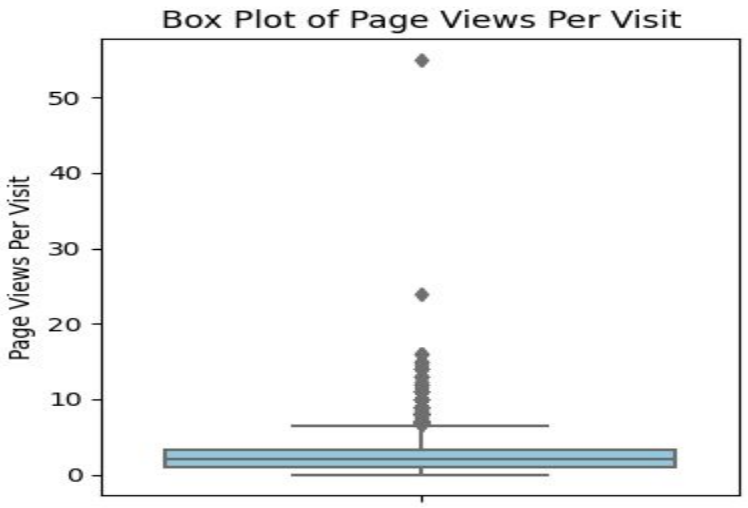
10. Do Not Call
- "No" is the common choice in the "Do Not Call" category, indicating that the majority of leads are open to receiving calls.

11. Tags
- "Will Revert After Reading Message" stands out as the most frequently used tag among leads, indicating a common response behavior.

Univariate Numerical Columns

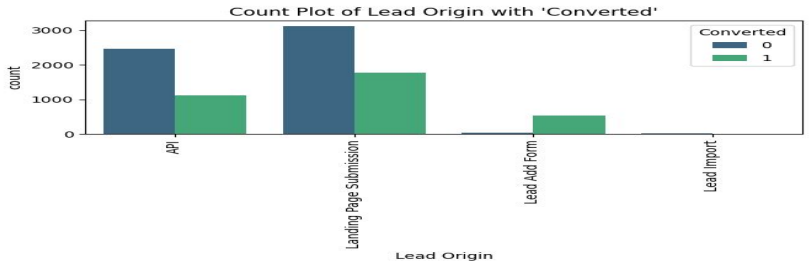


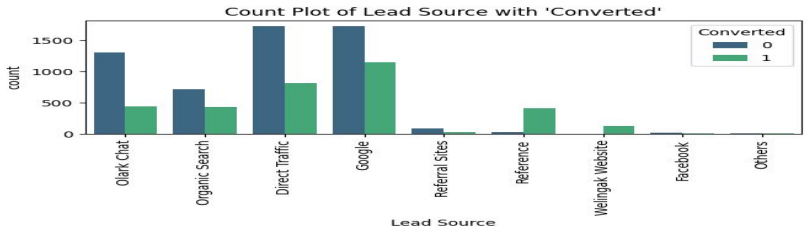


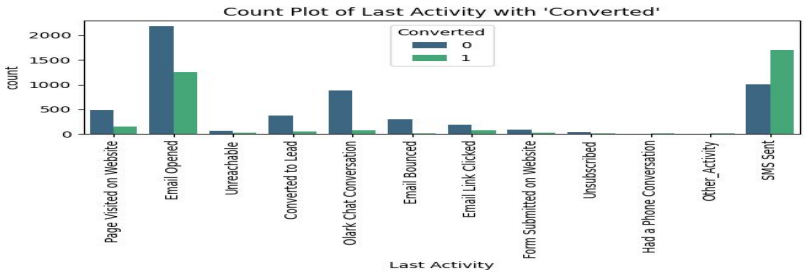


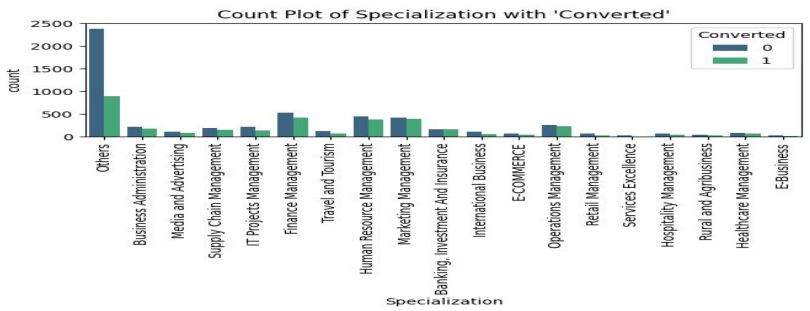
1. The "Total Visits" feature exhibits numerous outliers, which are values significantly different from the majority of observations.
2. The median time spent on the website is around 250, indicating that half of the leads spend less time, and half spend more.Notably, the upper quartile or "whisker" extends significantly higher, likely above 2000, showing the presence of outliers with extremely long website engagement.
3. The upper quartile, or "whisker," is observed around 8, indicating that a majority of leads typically view a few pages per visit.However, the presence of outliers with values significantly higher, above 50, indicates that a subset of leads exhibits extensive page views during their visits.

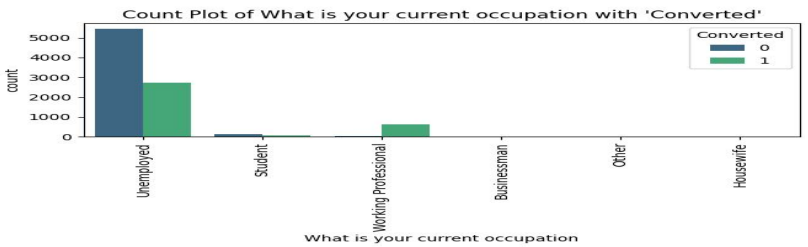
Bivariate Categorical Columns

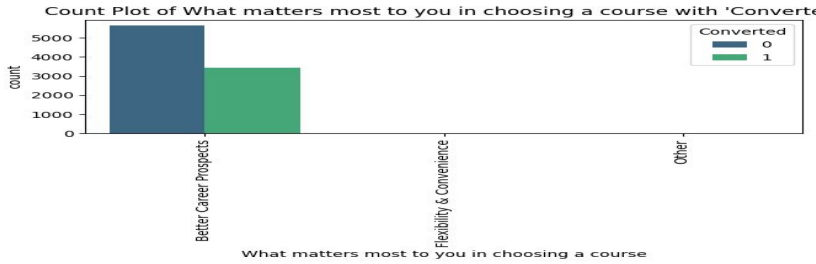


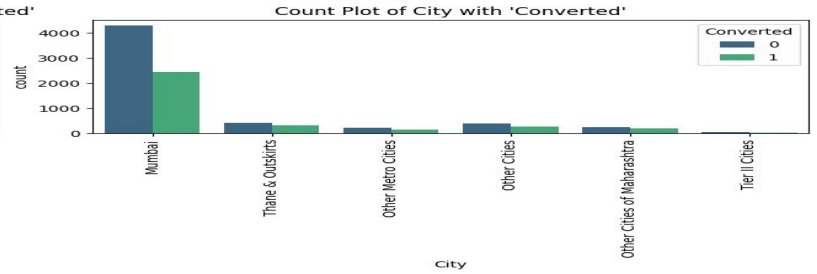


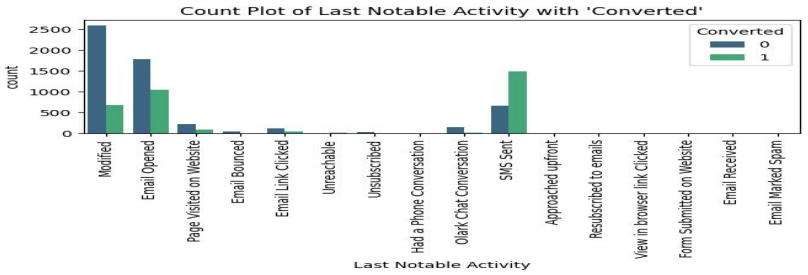


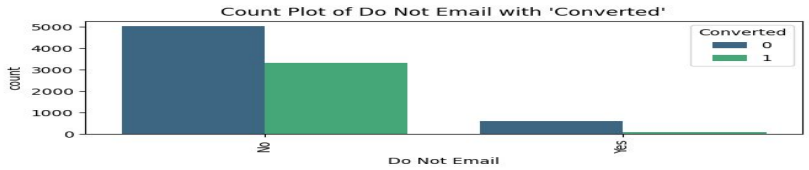


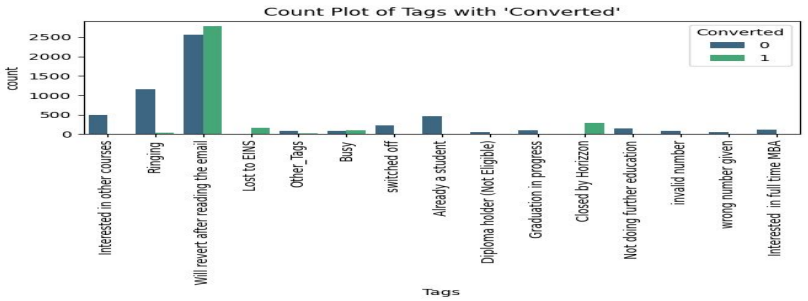






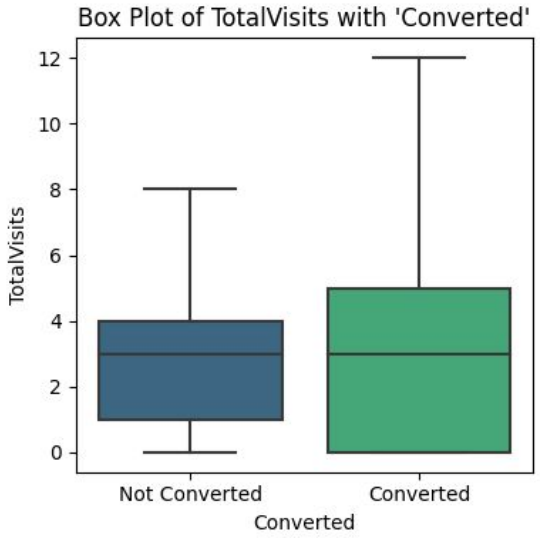


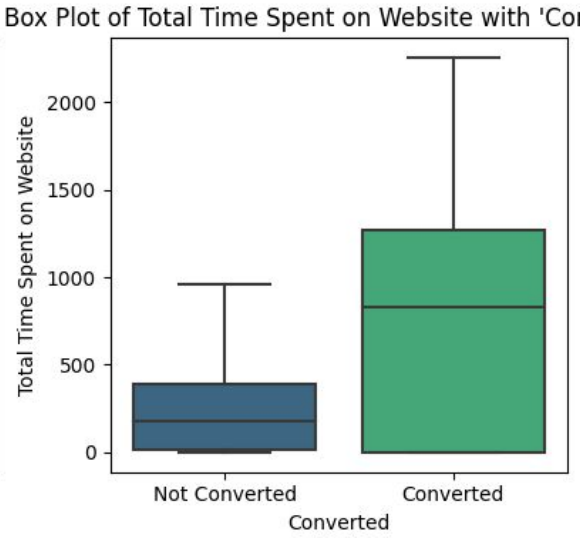


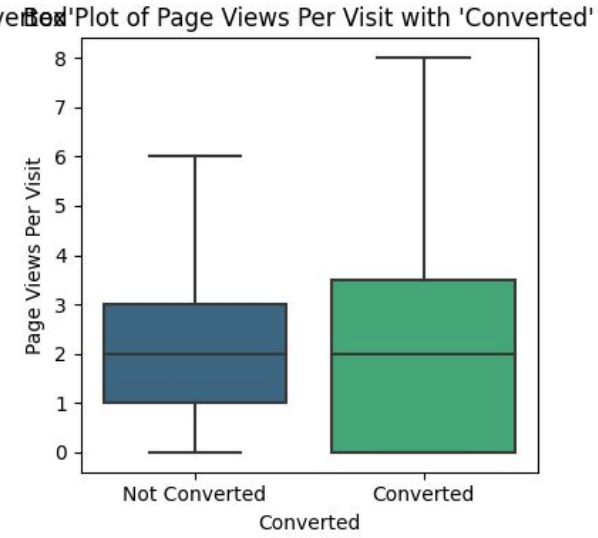


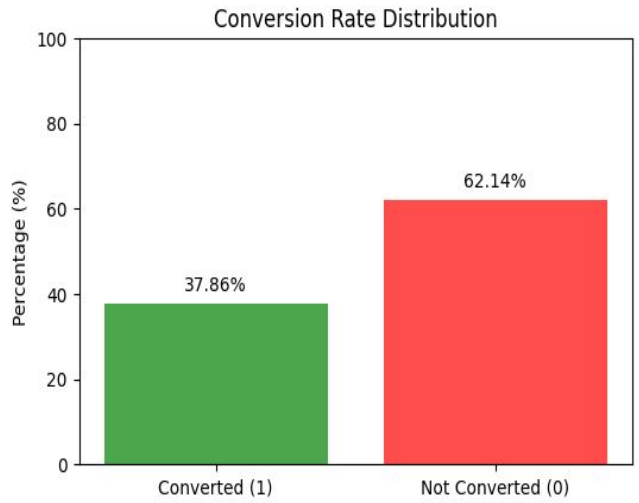
1. Lead Origin
   - In the Landing Page Submission category, leads marked as 0 (not converted) are more prevalent than leads marked as 1 (converted).
   - Similarly, in the API category, the number of leads marked as 0 is higher than those marked as 1, signifying a lower conversion rate for these leads.
2. Lead Source
   - Among the "Lead Source" categories, "Reference" stands out as the only source with more leads marked as 1 (converted) than 0 (not converted).
3. Last Activity
   - Among various lead activities, SMS Sent stands out as the sole activity with more leads marked as 1 (converted) than 0 (not converted).
   - This distinctive pattern suggests that SMS interactions are particularly effective in driving conversions.
4. Lead Quality
   - Among the Lead Quality categories, High in Relevance stands out as the sole category with more leads marked as 1 (converted) than 0 (not converted).
   - This distinct pattern indicates that leads categorized as High in Relevance consistently convert at a higher rate.
5. Specialization
   - In the Specialization categories, all categories exhibit more leads marked as 0 (not converted) than 1 (converted), indicating a common trend of lower conversion rates.
6. What is your current Occupation
   - Among the categories in the "What is Your Current Occupation" category, "Working Employee" is the only occupation with more leads marked as 1 (converted) than 0 (not converted).
   - This unique pattern implies that employed individuals, specifically "Working Employees," exhibit a higher likelihood of conversion.
7. City
   - Across all cities listed in the "City" category, there is a consistent trend of having more leads marked as 0 (not converted) than 1 (converted).
8. Leads Notable Activity
   - Within the "Leads - Notable Activity" category, "SMS Sent" is the sole activity with more leads marked as 1 (converted) than "0" (not converted).
   - This unique pattern suggests that SMS interactions are especially effective in nurturing and converting leads.
9. Do Not Email
   - "No" in the "Do Not Email" category has fewer leads marked as 1 (converted) than 0 (not converted), indicating a lower conversion rate for these leads.
   - In contrast, "Yes" in this category exhibits more leads marked as 1 than 0, highlighting a higher conversion rate among leads who opt for email communication.
10. Tags
    - The tag "Will Revert After Reading Email" demonstrates a higher number of leads marked as 1 (converted) compared to 0 (not converted).
    - This specific tag indicates a significant likelihood of conversion among leads who have been tagged with "Will Revert After Reading Email."

Bivariate Numerical Column









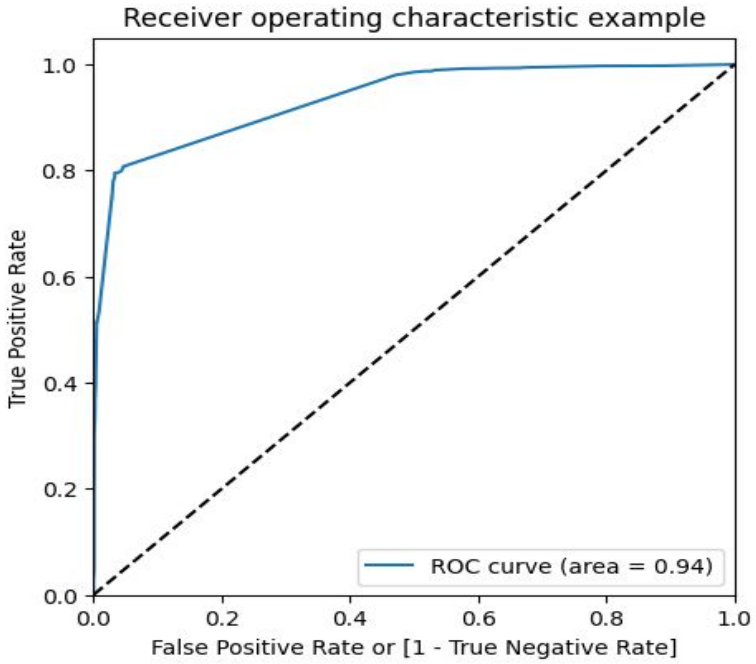
1. Total Visit
- The median number of total visits is identical for both converted and not converted leads, indicating a common midpoint value.
- However, the upper quartile, represented by the whisker and the 75th percentile, is notably higher for converted leads, showcasing a group of converted leads with exceptionally high total visits.

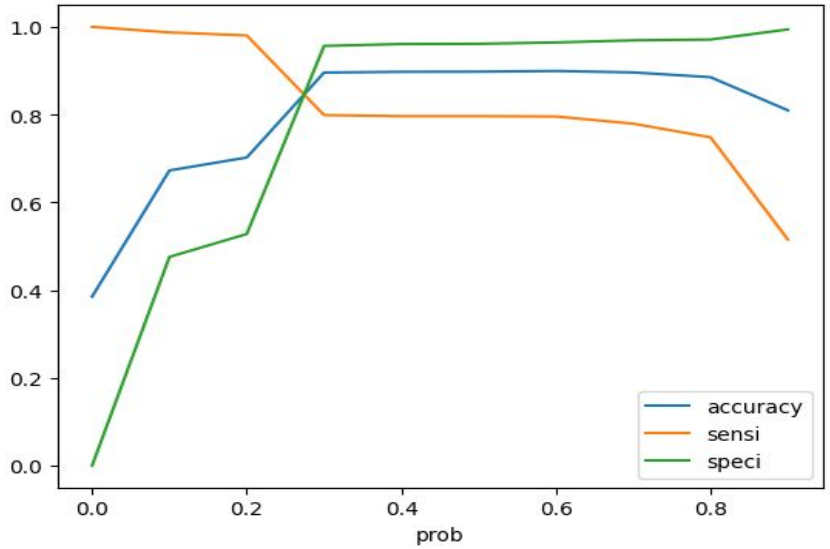
2. Total Time Spent on Website
- The total time spent on the website is considerably higher for converted leads in comparison to not converted leads, signifying a strong association between website engagement and conversion.

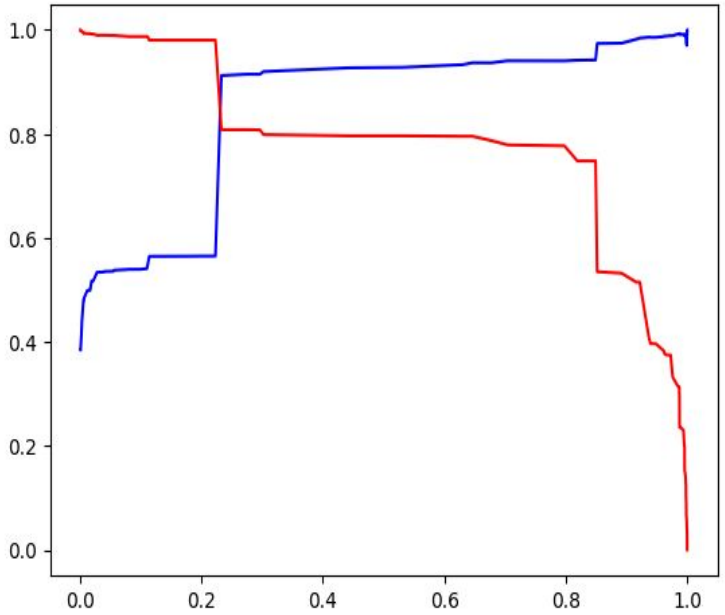
3. Page Views Per Visit
- The median number of page views per visit is identical for both converted and not converted leads, indicating a common midpoint value.
- However, the presence of more converted leads compared to not converted suggests that there is a subset of converted leads with a higher number of page views per visit.

4. Converted
- Within the Converted column, Not Converted accounts for 62.14% of the total leads, signifying a majority of leads that did not convert.
- Conversely, Converted constitutes 37.86% of the leads, reflecting a significant but smaller proportion of successfully converted leads.

MODEL EVALUATION







1. We receive ROC of 94 % which indicates our model has a strong ability to discriminate between the classes.
2. In Accuracy, Sensitivity and specificity curve we got an optimum cutoff probability as 0.3
3. In Precision,Recall curve we got optimum cutoff probability as 0.25

Recommendation

- Prioritize Lead Scoring: Implement a robust lead scoring system to identify high-potential leads, enabling the sales team to focus on those most likely to convert.
- Segment and Personalize: Segment leads based on engagement and personalize communication strategies to increase the relevance of interactions.
- Sales Team Training: Invest in continuous training for the sales team to enhance their skills and product knowledge, ensuring effective lead engagement.
- Automation and Efficiency: Integrate automation for follow-up emails and routine tasks to optimize time and resources.
- Content Marketing: Develop valuable content and content marketing strategies to educate and engage leads without over-reliance on phone calls.
- Social Media Engagement: Leverage social media for interaction and inquiries, building relationships and brand image.
- Data-Driven Decision-Making: Analyze data for insights and refine lead scoring, content strategies, and customer interactions.
- Stay Agile: Adapt strategies to business needs, whether it's aggressive lead conversion or minimizing unnecessary phone calls.
- Collaborative Efforts: Explore partnerships and collaborations to expand the customer base.
- Continuous Metrics Monitoring: Keep a close eye on key metrics to evaluate the effectiveness of strategies and make adjustments as needed.

SUMMARY

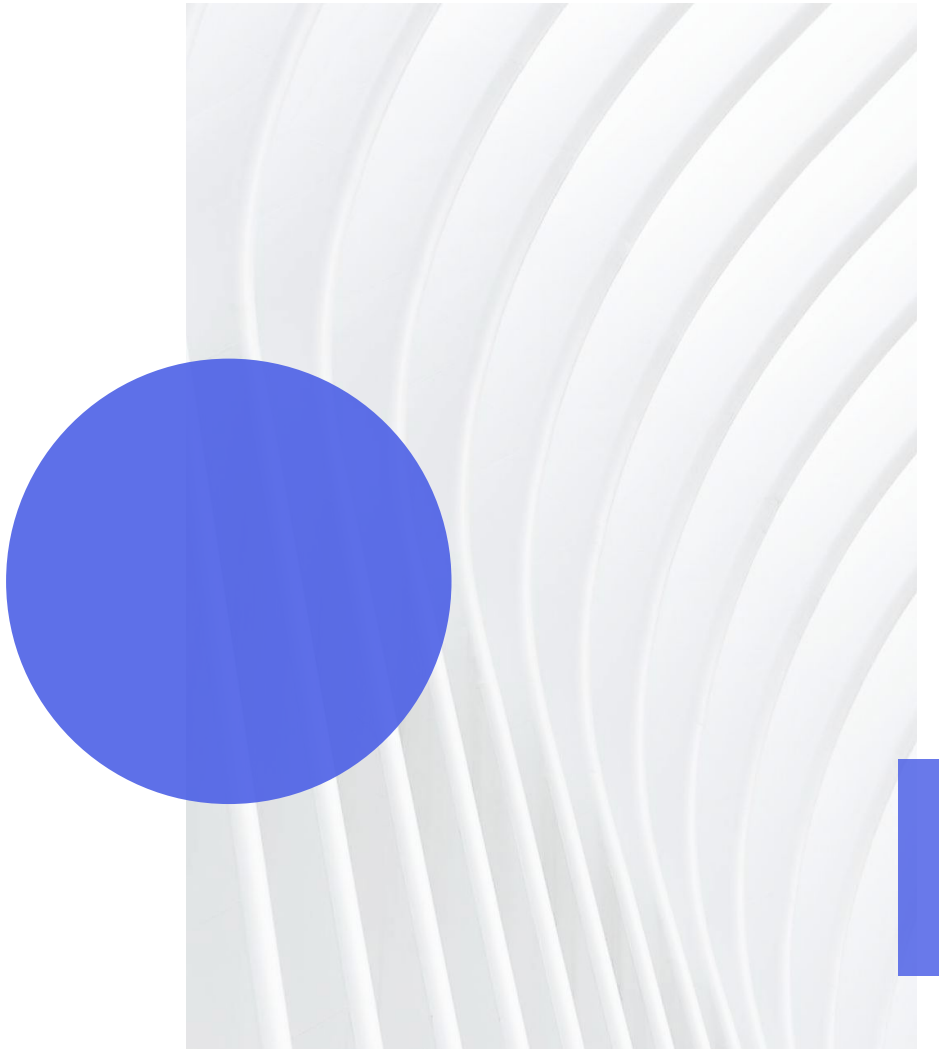
Train Data:

Accuracy : 89.78%
Sensitivity : 79.64%
Specificity : 96.1%

Test Data:

Accuracy : 89.4%
Sensitivity : 79.6%
Specificity : 96.1%

In conclusion, the logistic regression model we developed provides a lead score for potential customers, helping X Education identify "Hot Leads" with a higher likelihood of conversion. By focusing their efforts on these leads, the company can work towards achieving the CEO's target conversion rate of 80%.




THANK YOU