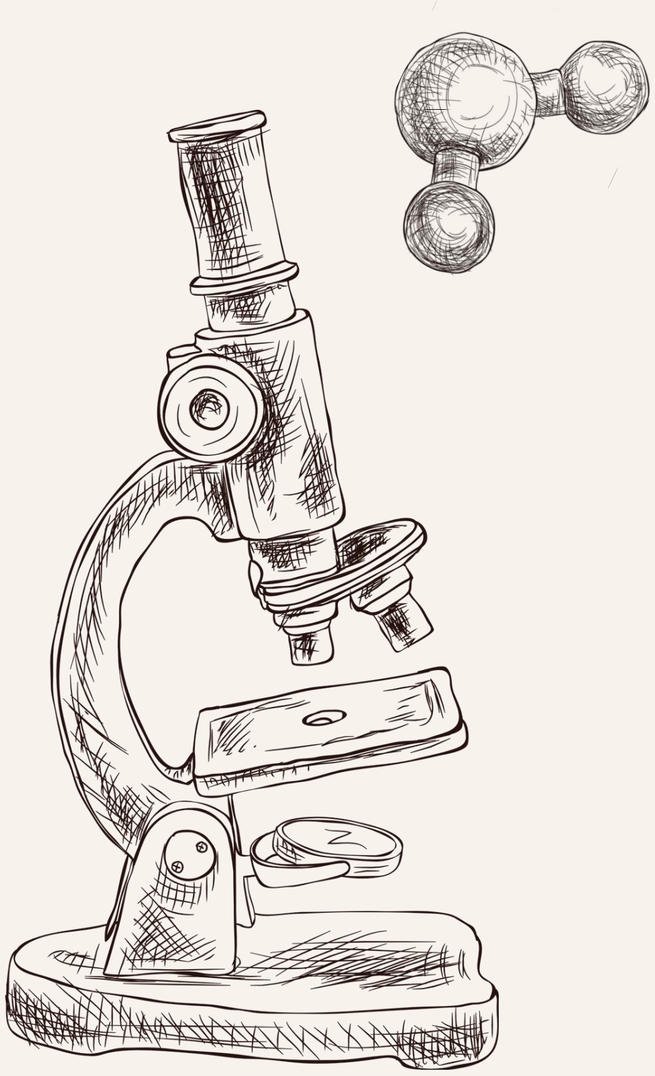
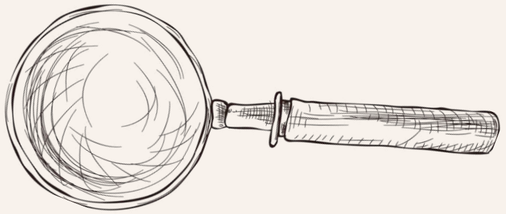
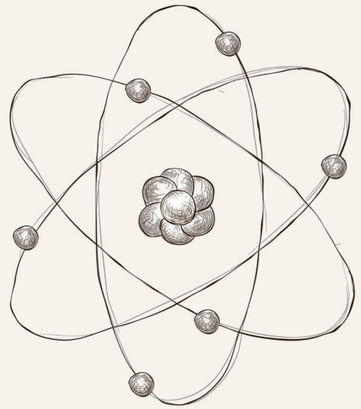


LEAD SCORING CASE STUDY X-EDUCATION





Optimizing Lead Conversion Using Logistic Regression

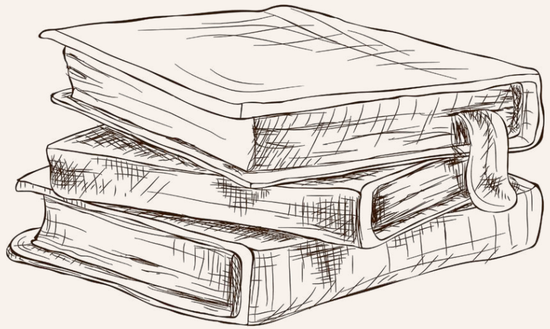
Presented by :

Rohini M.

Sagar Kumar

Prathamesh Salunke

Upgrad IIITB DS C67 batch



Problem Statement

- **Problem Description:**

  - X Education faces a lead conversion rate of 30%.
  - The goal is to improve the lead conversion rate to 80% by identifying hot leads.
  - Hot leads are defined as leads that are more likely to convert into paying customers.

OBJECTIVE:

BUILD A LOGISTIC REGRESSION MODEL THAT ASSIGNS LEAD SCORES BASED ON CONVERSION PROBABILITY.

Dataset Overview:

- Rows: 9,240 leads (data points)
- Features: 37 variables (both categorical and numerical)
- Target Variable: Converted
  - 1 for converted leads (paying customers)
  - 0 for non-converted leads

Key Features:

- Lead Source: Source of the lead (e.g., Google, Direct Traffic).
- Total Time Spent on Website: Measures engagement; more time correlates with higher conversions.
- Last Activity: Most recent lead interaction (e.g., Email Opened, Olark Chat).
- Specialization: Area of interest (e.g., Marketing, Finance); important for conversion rates.
- Lead Profile: Type of lead (e.g., Potential Lead, Student).

Dropped Columns:

- 'Tags' and 'Lead Quality': Dropped due to over 30% missing data to maintain dataset quality.

Missing Data Handling:

Columns with >30% Missing Data:

Dropped: ‘Tags’, ‘Lead Quality’ (due to excessive missing values).

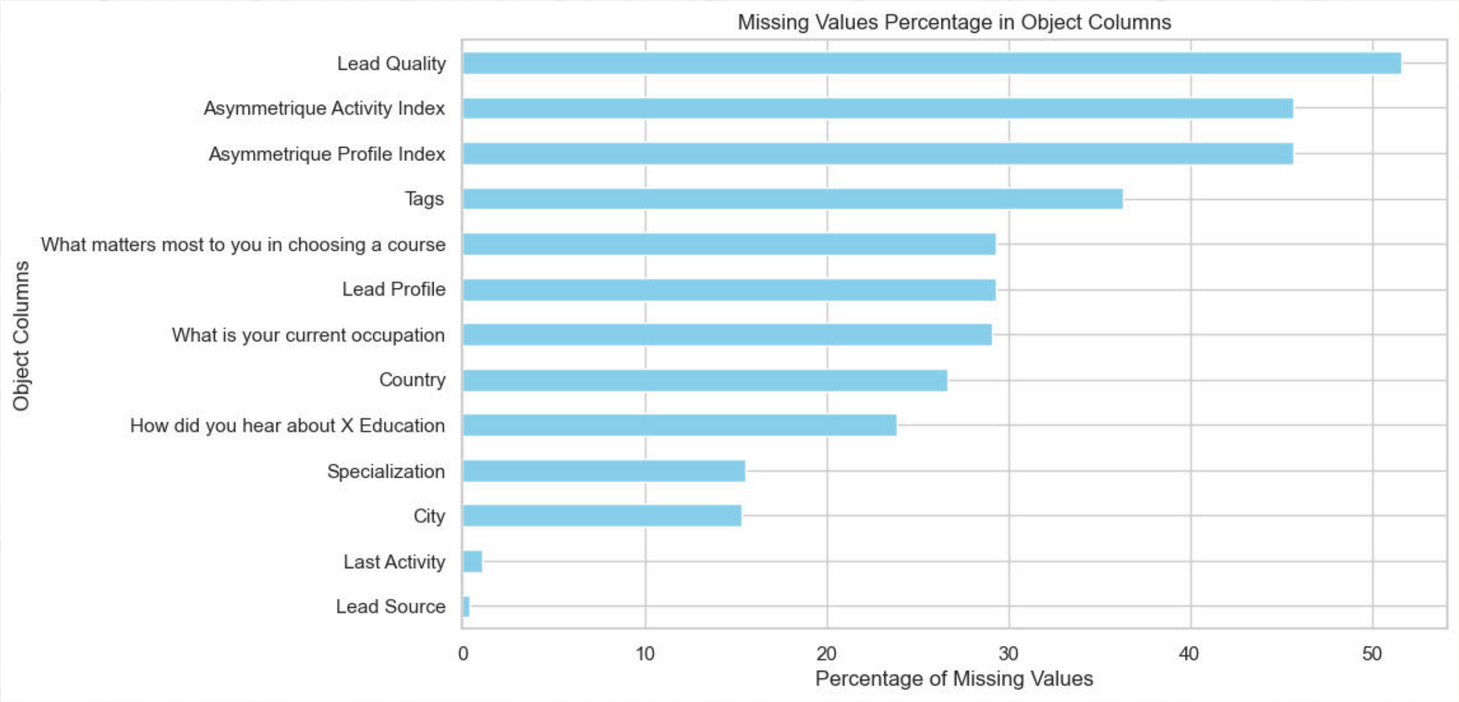
Imputation:

Mode Imputation: Applied to categorical variables like 'Specialization' and 'Lead Source'.

Median Imputation: Applied to numerical variables like 'Total Visits' and 'Page Views Per Visit'.

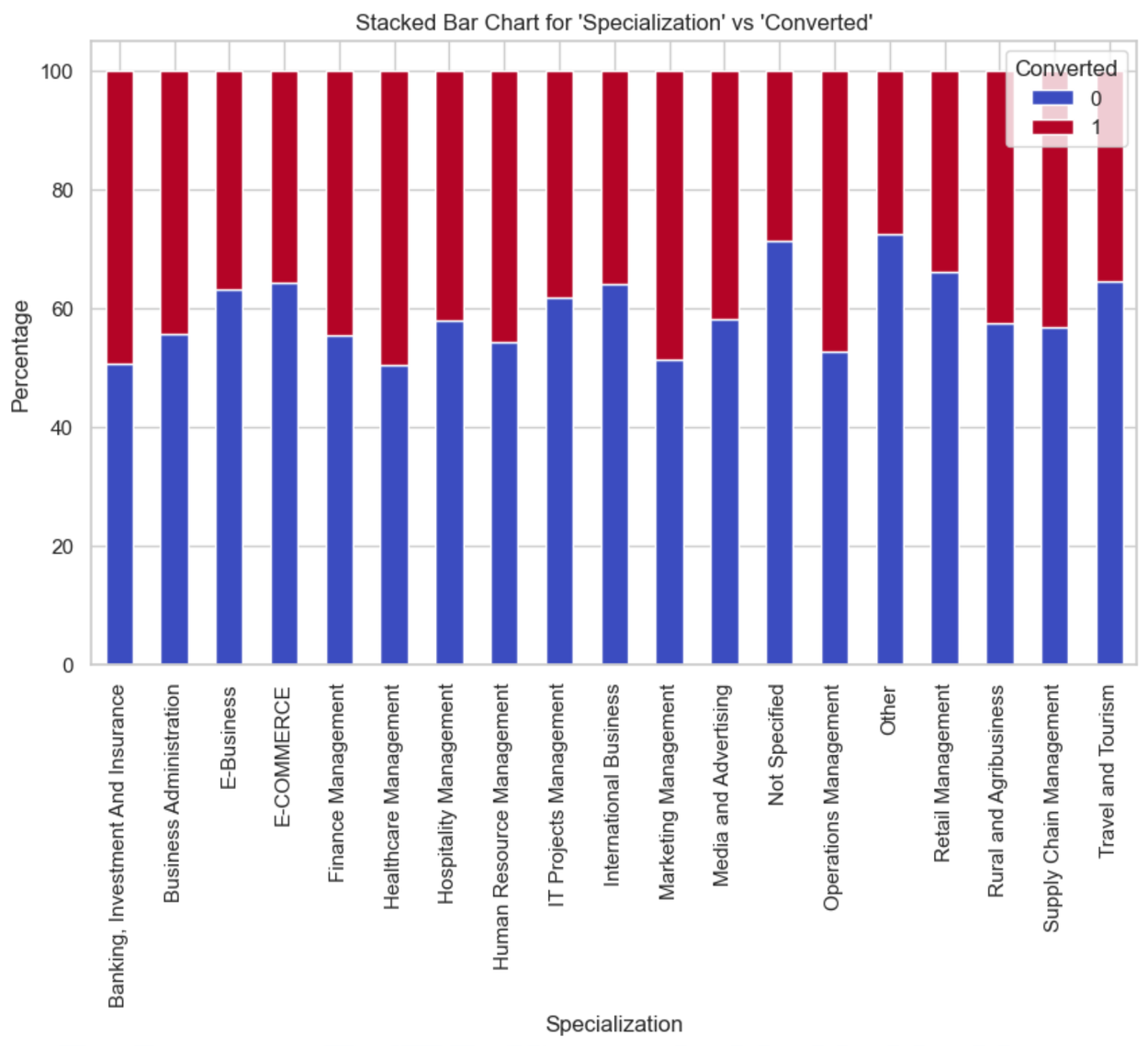
Outlier Treatment:

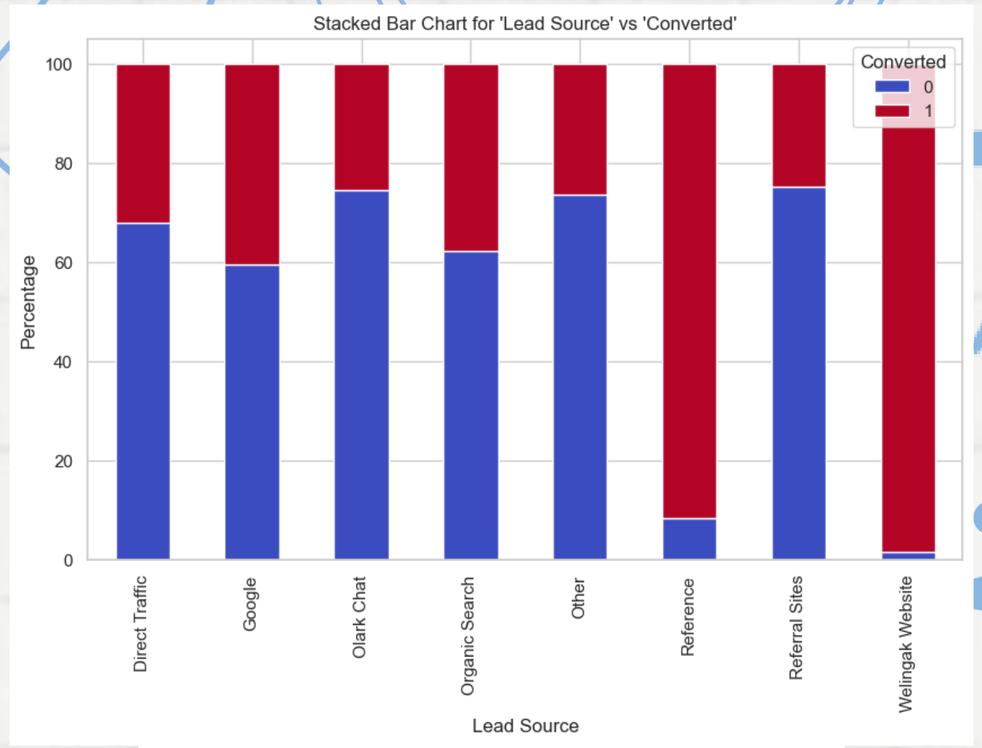
Log Transformation: Applied to 'Total Visits' and 'Page Views Per Visit' to reduce skewness and handle outliers.

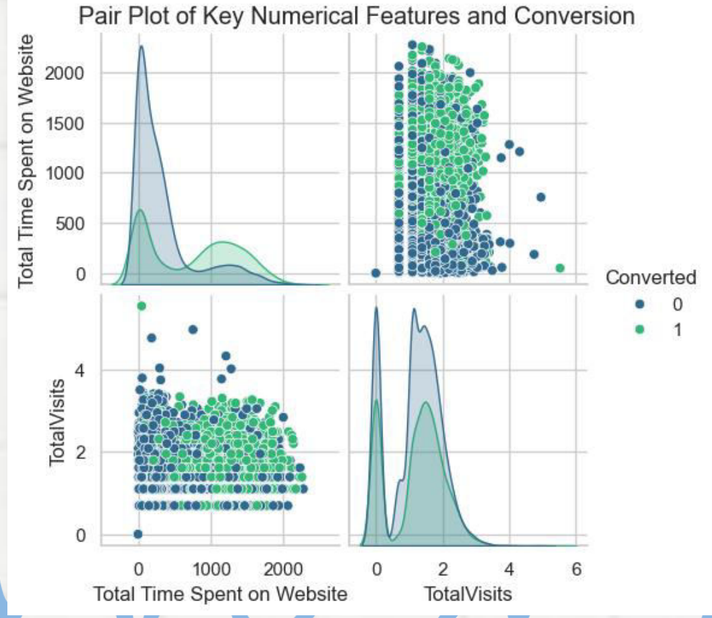


Dropped irrelevant or highly imbalanced categorical levels, e.g., categories with fewer than 10 leads.

DATA VISUALIZATION







Feature Transformations:

Log Transformations:

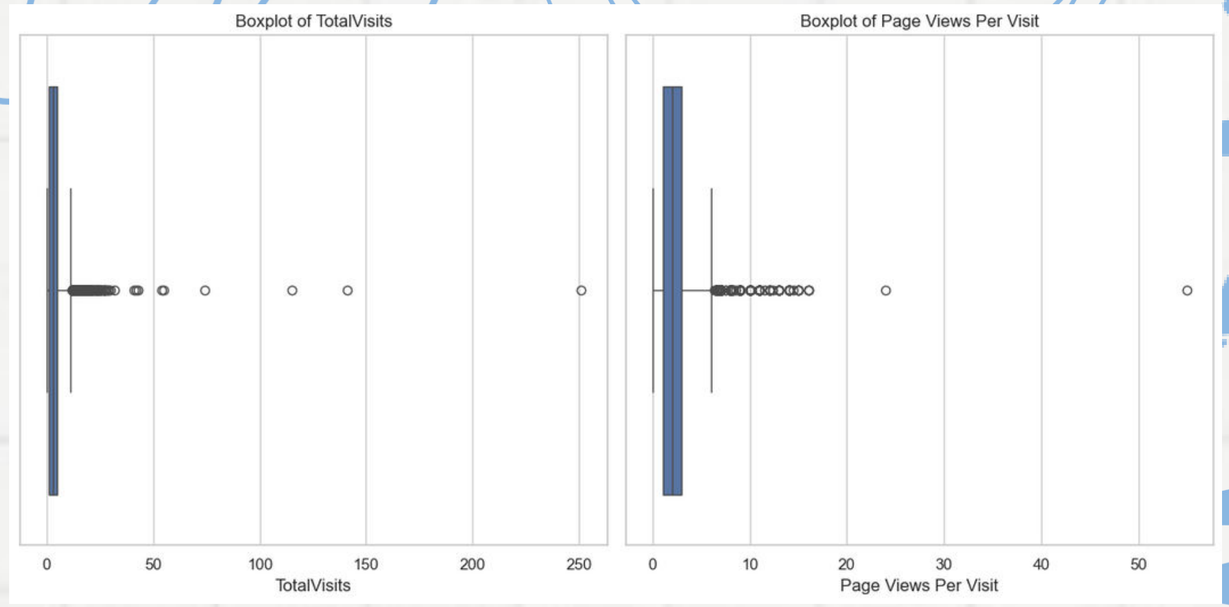
Applied to Total Visits and Page Views Per Visit to reduce skewness and handle outliers.
This transformation helps normalize the data and make it more suitable for logistic regression.

One-Hot Encoding:

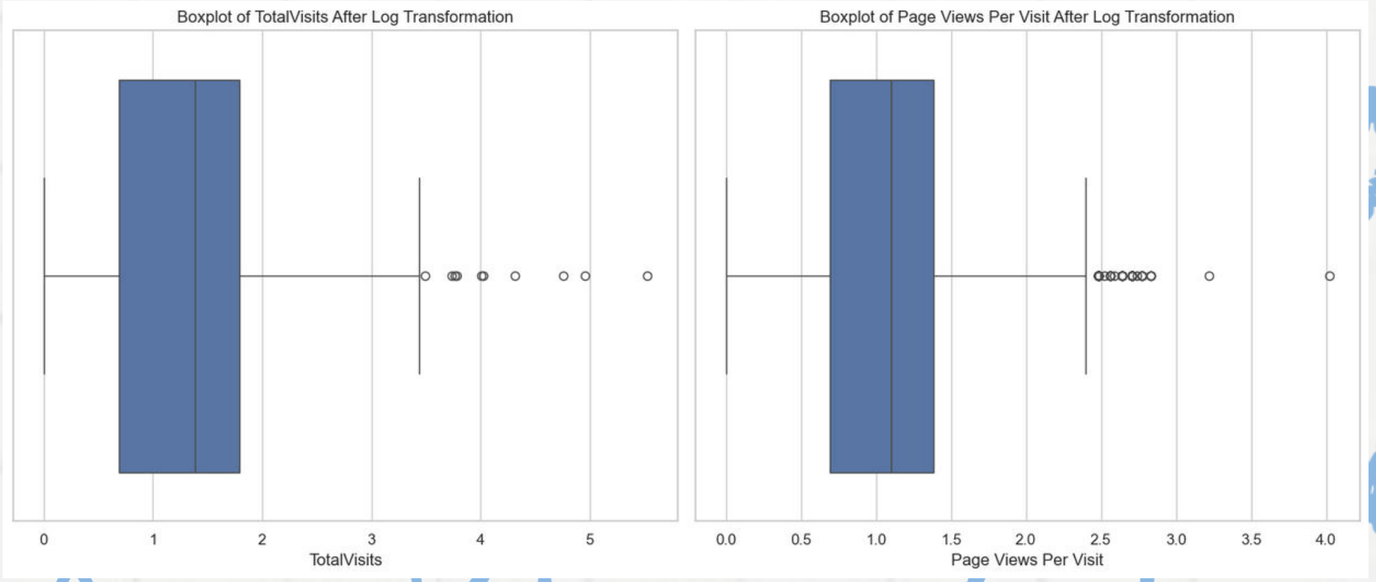
Categorical variables like Lead Source, Specialization, and Last Activity were transformed into numerical representations using one-hot encoding.
This transformation allows the logistic regression model to interpret these categorical features.

Missing Value Imputation:

Mode imputation was applied to categorical variables like Specialization and Lead Source to handle missing values.
Median imputation was applied to numerical variables like Total Visits to ensure robust handling of missing data without distorting the dataset.



BEFORE LOG TRANSFORMATION



AFTER LOG TRANSFORMATION

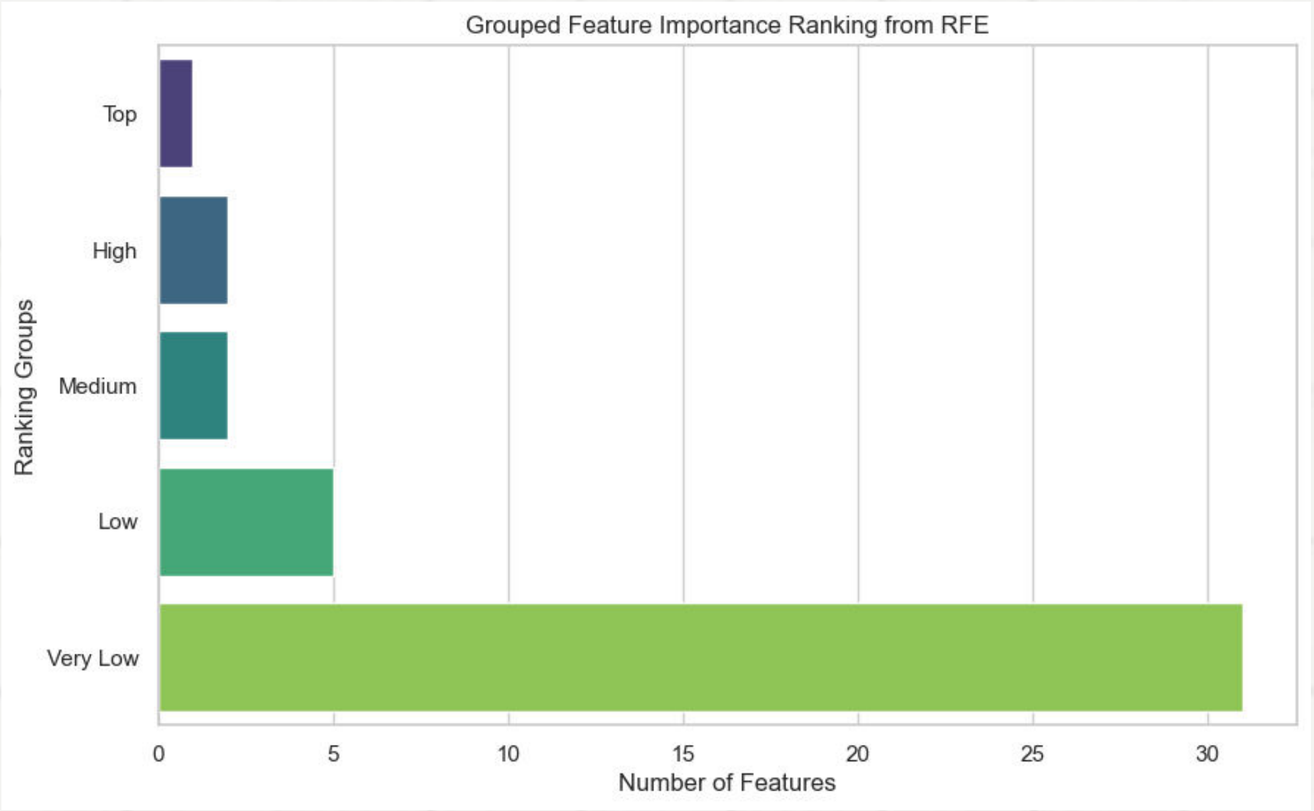
LOGISTIC REGRESSION MODEL BUILDING

Methodology:

Recursive Feature Elimination (RFE) was used to select the top 18 features based on their importance to the logistic regression model.

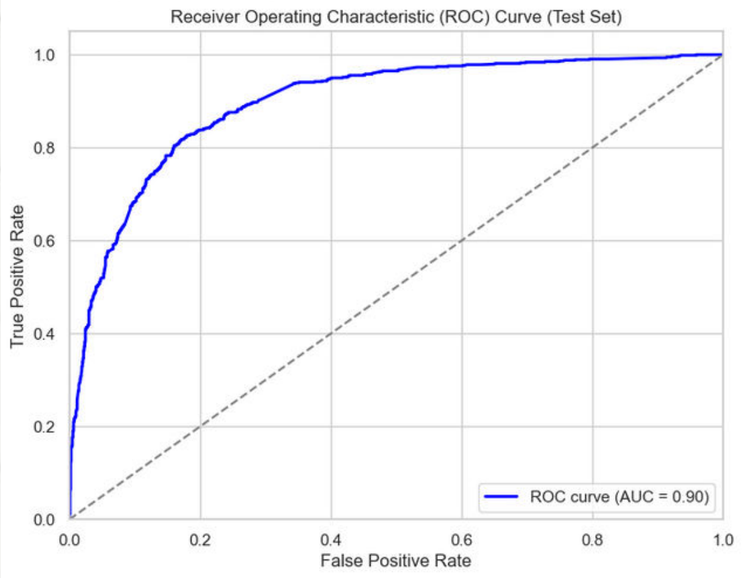
Logistic regression was chosen for its simplicity and interpretability, making it ideal for this task of assigning lead scores.

Two logistic regression models were built and tested:
Model 1: Initial model with the top 18 features selected via RFE.
Model 2: Refined model where certain features (like Specialization_Rural and Agribusiness) were dropped based on further analysis and p-value examination.



MODEL EVALUATION

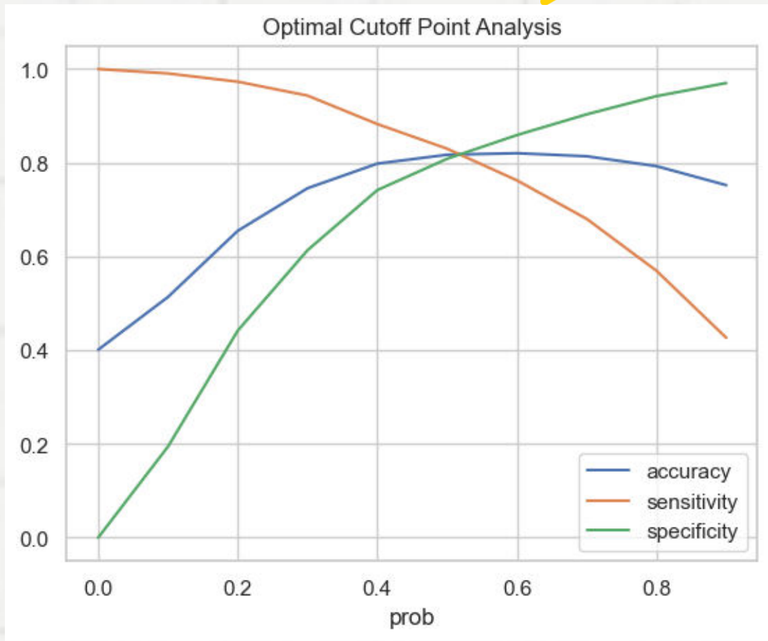
Metrics Used:
The model was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC.

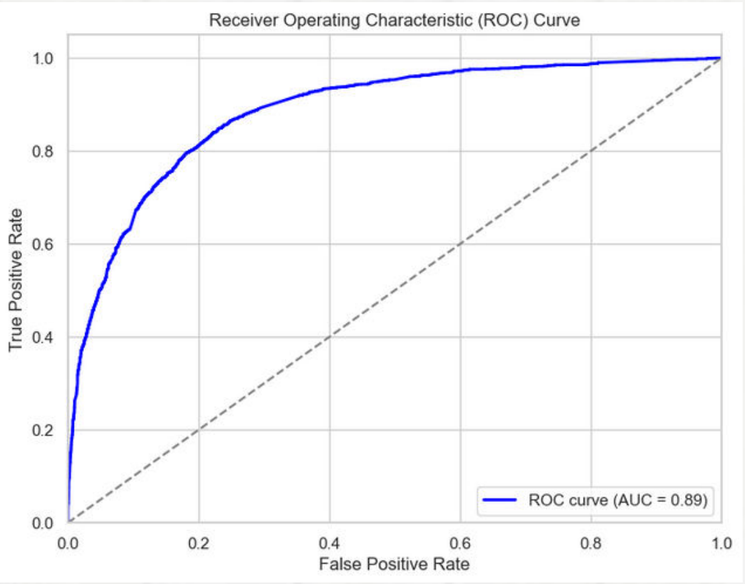


ROC of Test model

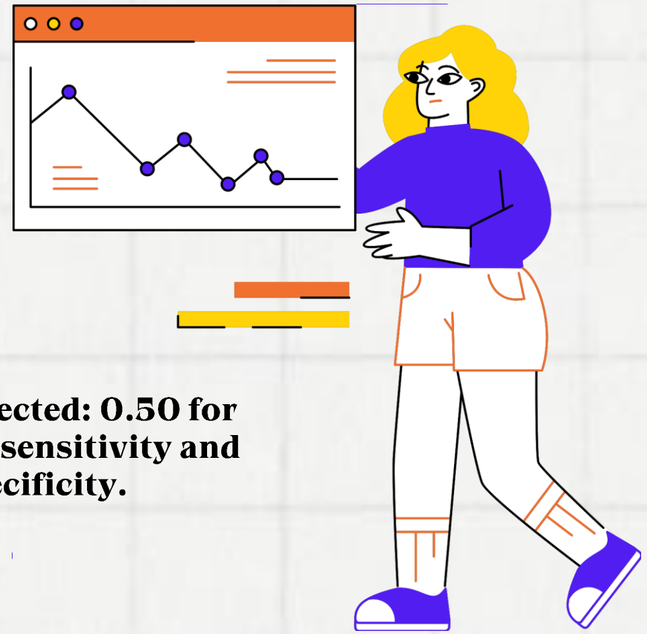
Model Performance:
Accuracy on Training Set: 80.19%
ROC-AUC on Training Set: 0.89
Test Set Performance:
Accuracy: 81.71%
Precision: 0.7437
Recall: 0.8300
F1-score: 0.7844
ROC-AUC: 0.90

Cutoff selected: 0.50 for balancing sensitivity and specificity.



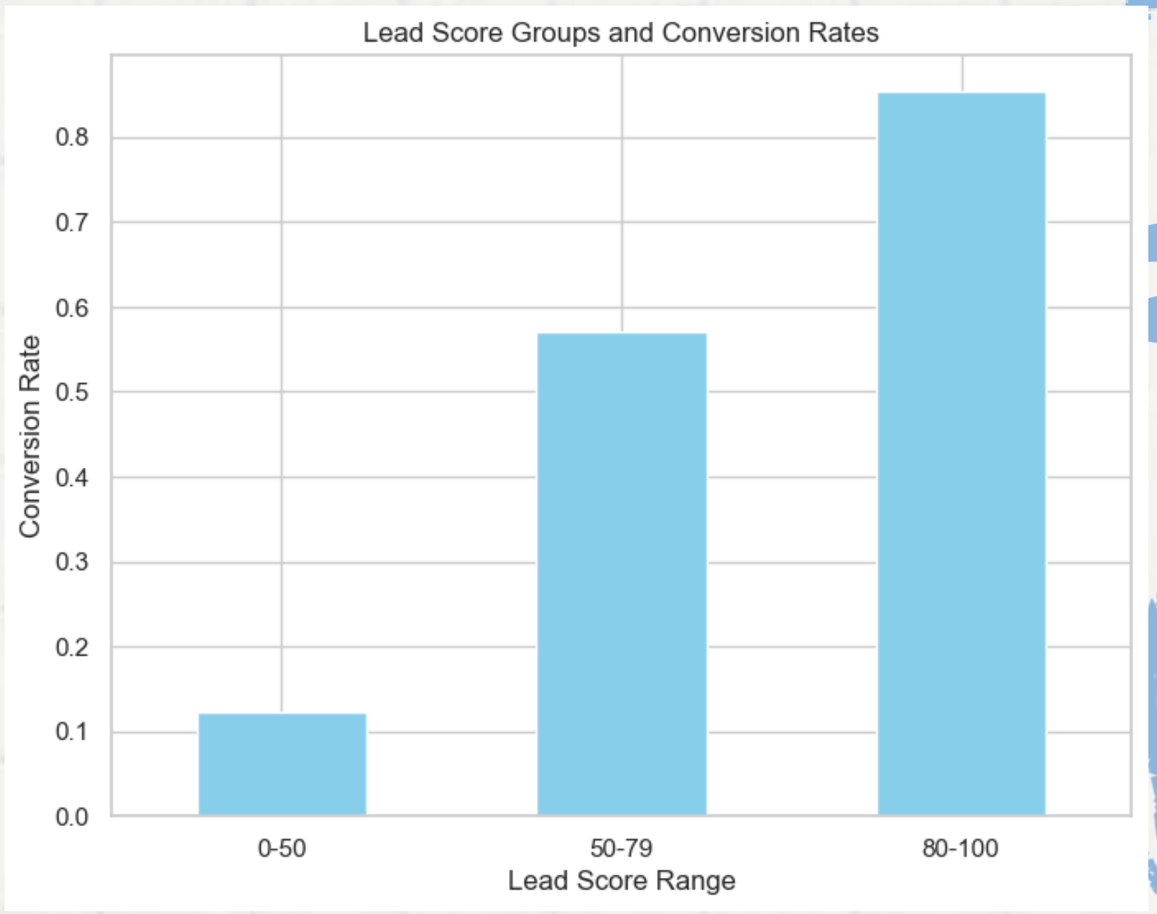


ROC of Train model



Impact of Lead Scoring:

The lead scoring system allows the sales team to focus on hot leads rather than spreading their efforts thinly across all leads. By concentrating on the leads with the highest probability of conversion, the team can increase efficiency, reduce costs, and boost revenue. The model helps streamline the sales process, ensuring the team spends more time with qualified leads and ultimately drives a higher conversion rate, meeting the company's goal of 80% lead conversion.



Insights and Recommendations

Key Insights:

- **Total Time Spent on Website: Higher time spent leads to more conversions.**
- **Focus on High Potential Lead Sources:**
- **Lead sources like Welingak Website and Reference consistently show a positive conversion effect. These lead sources should be targeted with more focused engagement.**
- **Lead Profile – Potential Lead: These leads have a much higher chance of conversion.**
- **Lead Profile – Student of SomeSchool: This profile shows a negative impact on conversion.**
- **SMS Activities: Sending SMS increases the chances of conversion.**
- **Lead Source – Olark Chat: Leads engaging with Olark Chat are more likely to convert.**
- **Specialization Fields (Marketing, Finance, Operations): Leads from these specializations convert more frequently.**
- **Lead Origin – Landing Page Submission: These leads have a lower conversion rate.**

Lead Source Recommendation:

- **Increase Investment in Welingak Website Campaigns.**
- **Enhance Engagement Features on the Website.**
- **Strengthen Referral Programs for More Leads from References.**
- **Prioritize Potential Leads for Sales Efforts.**
- **Continue Sending SMS Communications to Nurture Leads.**
- **Expand Real-Time Chat Services (Olark Chat).**
- **Focus Marketing on Specializations Like Marketing, Finance, and Operations.**
- **Reevaluate Marketing Focus for Leads from Landing Page Submissions (they convert less frequently).**

Business Impact of the Lead Scoring Model

Increased Revenue: Optimizing Lead Conversion: By focusing on high-probability leads (lead scores 80-100), the sales team can prioritize those most likely to convert.

Revenue Growth: With improved targeting, the company can expect an increase in revenue as more leads are converted into paying customers.

Improved Sales Team Efficiency: Prioritizing Sales Efforts: The model helps sales teams spend more time on leads with a higher chance of conversion, reducing wasted efforts on low-quality leads.

Faster Lead Follow-Up: By focusing on 'Hot Leads,' the team can improve the speed and efficiency of their sales process, resulting in quicker conversions.

Improved Marketing ROI: High Performing Sources: The model identified lead sources like Google Organic Search and Direct Traffic as having higher conversion rates.

Optimized Marketing Spend: By allocating more resources to these high-performing channels, the company can expect a better return on marketing investments, leading to cost savings and higher profitability.

Thank you very much!