

Summary Report: Lead Score Case Study

Objective:

This assignment develops a predictive model to identify potential leads who are more likely to convert into customers, and aims to help the sales team prioritize leads.

Data Preprocessing:

1. Data Cleaning:

- **Dropping Irrelevant Columns:** dropped ID columns and columns with uniform data.
- **Handling Missing Values:** Features with a high percentage of missing values were dropped, while others were imputed with the mode or removed depending on their business significance.
- **Categorical Columns:** Categorical features were identified, and converted to dummy variables for inclusion in the logistic regression model.

2. Exploratory Data Analysis (EDA):

- **Correlation Analysis:** Heatmap visualisation was used to explore correlations between numerical features, highlighting features like TotalVisits, Total_Time_Spent etc had significant relationships with lead conversion.
- **Box Plots:** Box plots were used to detect outliers in the data. No outliers detected.
- **Bar Charts:** Bar charts plotted for categorical variables which provided insights like Lead Add Form and those with India as the country had higher etc have conversion rates.

Model Building:

We used a logistic regression model for this binary classification problem.

1. Feature Selection:

- **Recursive Feature Elimination (RFE):** RFE was employed to select the most significant features to reduce multicollinearity and retaining the most predictive features.
- **Handling Multicollinearity:** VIF was calculated for all features, and VIF above 5 were removed to avoid multicollinearity issues.

2. Model Evaluation:

- **Training Model:** The model was first trained on the training dataset, achieving an accuracy of approximately 85%, sensitivity 76%, specificity 90%, recall 75%, precision 83%, the area under the ROC curve 0.92.
- **Cutoff Selection:** Different probability cutoffs were explored to balance sensitivity and specificity. A cutoff of 0.4 was found to be optimal, optimizing the sensitivity to 81%, specificity to 87%, recall 81%, precision 80%.

- **Test Model:** The model was then evaluated on the test dataset, where it maintained similar performance metrics, indicating good generalizability.

Business Implications:

The logistic regression model revealed key features that influence lead conversion:

- **TotalVisits** and **Total Time Spent on Website** are the most influential positive predictors of lead conversion.
- **Lead Source - Welingak Website** and **Last Activity - Had a Phone Conversation** significantly boost the likelihood of conversion.
- **Lead Quality** is the strongest negative predictor, indicating that leads labeled as "worst" are highly unlikely to convert.
- Other negative predictors include **Lead Quality - Not Sure** and **Might be**.
- Less impactful negative factors include **Page Views Per Visit**, **Do Not Email**, and **Lead Source - Referral Sites**.
- **Adjust the Model's Threshold:** Lower/increase the cutoff probability threshold of the logistic regression model from the 0.4 to a lower or higher value depending on how aggressive the campaign should be. Lowering the threshold will classify more leads as "high potential".
- **Lead_Score:** Use the lead score (100 to 0) to prioritize sales efforts, focusing first on leads with higher scores.

This model guides the sales team to focus on leads with the highest conversion potential, optimizing resource allocation and campaign effectiveness.