LEAD SCORE SUMMARY REPORT

This report highlights the main insights and procedures followed to create a lead scoring model intended to improve the company's lead conversion rate to 80%. The model was built using logistic regression for predictive insights.

1. Objective

The CEO tasked the team with developing a system to assign a score to each lead. A higher score would indicate a greater likelihood of conversion, while a lower score would suggest a lower chance of converting. The primary aim was to optimize resource allocation by concentrating efforts on leads with the highest conversion potential, thereby enhancing the overall lead conversion rate.

2. Dataset Description

The analysis utilized two datasets:

- **Leads.csv**: This dataset contains roughly 9000 entries with various attributes, such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, and a 'Converted' column, where '1' indicates a successful conversion and '0' indicates no conversion.
- **Leads Data Dictionary.csv**: This file provided definitions and details for the columns in the Leads.csv dataset.

3. Methodology

The analysis was structured into the following phases:

- **Data Preprocessing**: The first step was identifying and handling missing data. Columns with over 3000 missing values, such as 'City' and 'Country,' were eliminated due to their irrelevance. Attributes like 'Lead

Profile' and 'How did you hear about X Education' were discarded because they contained insignificant data marked as 'Select.'

- **Exclusion of Low-Variance Columns**: Columns that displayed little variation, where all entries were the same (e.g., 'Do Not Call,' 'Magazine,' 'Newspaper Article'), were excluded from the model.
- **Dummy Variable Creation**: Dummy variables were generated for categorical features, such as 'Specialization,' and irrelevant categories were explicitly dropped.

4. Model Development

- The dataset was divided into a training set (70%) and a test set (30%). Three numerical variables—'Total Visits,' 'Total Time Spent on Website,' and 'Page Views Per Visit'—were scaled.
- The most relevant 15 variables were selected using Recursive Feature Elimination (RFE) for building the logistic regression model.
- The model was evaluated based on p-values (all <0.05) and Variance Inflation Factors (VIFs <5), indicating a strong model performance.

5. Model Performance

- **Initial Results**: The model achieved an accuracy of 76.65%, sensitivity of 81.12%, and specificity of 72.18%. The area under the ROC curve was 0.86, suggesting strong predictive performance.
- **Optimizing Threshold**: To enhance the model further, the probability threshold was adjusted to 0.44, providing an optimal balance between accuracy, sensitivity, and specificity.
- **Final Results**: With the optimized threshold, the model achieved a test accuracy of 71.94%, sensitivity of 73.30%, and specificity of 71.16%. A later adjustment to a 0.46 cutoff gave a final accuracy of 72.08%, precision of 59.45%, and recall of 72.80%.

6. Key Insights

- **Total Time Spent on Website**: This variable had the most significant positive impact on lead conversion, underscoring the importance of website engagement in increasing conversions.
- **Current Occupation (Working Professional)**: This group showed a higher likelihood of converting, identifying them as a key target market.
- **Lead Origin (Lead Add Form)**: Leads generated from website forms had a higher conversion rate, emphasizing the effectiveness of online lead generation strategies.

7. Conclusion

The model successfully identified the key factors influencing lead conversion, enabling the company to prioritize high-potential leads. With a final accuracy of 72.08% and a high recall of 72.80%, the model is well-suited for predicting lead conversions and helping the company meet the CEO's target. By optimizing thresholds and focusing on customer engagement through the website, the company can drive higher conversion rates and boost marketing efficiency.