Summary

Process

Data Cleaning:

- **Reduced Dataset**: Trimmed from 37 columns and 9,240 rows to 10 columns and 6,300 rows.
- **Handled Null Values**: Removed rows with missing data.
- Eliminated **Non-Essential Columns**: Dropped columns with little impact or no selections by leads.
- **Dropped Redundant Variables**: Removed columns where one variable had a much higher percentage.
- **Addressed Outliers**: Treated extreme values in numeric columns to maintain data quality.

Data Processing:

- **Dummy Variable Creation**: Dummy variables were generated for categorical columns.
- **Data Scaling**: Numeric columns like **'TotalVisits', 'Page Views Per Visit'** and **'Total Time Spent on Website'** were scaled using Min-Max Scaler to normalize values and improve model performance.
- **Train-Test Split**: The dataset was split into training and testing sets with a ratio of 70:30, ensuring a robust evaluation of the model's performance.

Data modelling :

- **Post-Processing Columns**: The dataset was reduced to 55 columns after initial data processing.
- **Automated Feature Selection**: Recursive Feature Elimination (RFE) was used to identify the top 15 most important variables.
- Variables were further refined using **statsmodels**, focusing on those with a p-value less than 0.05 and a Variance Inflation Factor (VIF) not exceeding 5.
- This process resulted in the selection of 12 key features from the top 15.
- **Model Development**: Four models were prepared during the manual selection phase, ensuring a robust evaluation of the most critical features.

Model Evaluation:

- **Probability Prediction**: The model was used to predict probabilities on the training set.
- **Cutoff Threshold**: A cutoff value of 0.5 was applied to classify the predictions. Probabilities above 0.5 were considered as **converted (1)**, and those below 0.5 were classified as **not converted (0)**.

Finding Optimal CutOff:

- **Metrics Calculation**: Specificity, sensitivity, and accuracy were calculated over a range of cutoff values to evaluate the model's performance.
- **Plotting Results**: The results were plotted to visualize how these metrics varied with different cutoff thresholds.
- **Optimal Cutoff**: The optimal cutoff threshold was determined to be **0.42**, balancing the model's ability to correctly classify leads while minimizing errors.

Predicting Test set using model:

- **Accuracy**: The model achieved an accuracy of **0.788**, indicating the proportion of correctly classified leads out of the total leads.
- **Precision**: The precision was **0.789**, reflecting the proportion of true positives among the predicted positives.
- **Recall**: The recall was **0.783**, showing the proportion of true positives identified out of the actual positives.

Recommendations based on Insights:

- Enhance Visibility of the Lead Add Form: Improve placement and accessibility to capture more high-converting leads.
- Optimize Marketing on Key Sources: Allocate resources to enhance visibility on Google, Reference, and the Welingak Website.
- Focus on High-Interest Activities: Implement targeted follow-up strategies for leads with SMS Sent and Email Opened.
- Target Hot Specializations: Customize marketing and outreach efforts for the identified high-potential specializations.
- Address Working Professionals: Develop specialized programs and highlight benefits that align with the career needs of working professionals.