Summary of Data Preparation and Model Building Process

1. Data Understanding and Preparation

- Data Import and Libraries: Utilized essential Python libraries like pandas, numpy, matplotlib, seaborn, and scikit-learn for data analysis and modeling.
- Missing Values: Identified and handled missing values by deleting columns with more than 40% missing data, dropping rows with less than 2%, and filling remaining missing values using median and mode for numerical and categorical data, respectively.
- Duplicate Removal: Ensured data integrity by removing duplicate entries.
- Outliers: Detected and removed outliers in numerical columns such as 'TotalVisits', 'Total Time Spent on Website', and 'Page Views Per Visit' using the Interquartile Range (IQR) method.
- Text Standardization: Standardized text data for 'Do Not Email' and 'Do Not Call' by converting them to lowercase and mapping them to binary values.
- Dummy Variables: Converted categorical variables into dummy variables to prepare the data for the logistic regression model.
- New Metrics: Derived new metrics like 'Time Per Visit' to enhance the analysis.

2. Exploratory Data Analysis (EDA)

- Univariate Analysis: Created histograms to visualize the distribution of variables.
- Bivariate Analysis: Generated scatter plots and box plots to analyze relationships between variables, focusing on key metrics like 'TotalVisits', 'Total Time Spent on Website', and 'Page Views Per Visit'.
- Correlation Analysis: Produced a correlation matrix heatmap to identify significant correlations between relevant variables.

3. Model Building

- Data Splitting: Split the dataset into training and testing sets (80:20 ratio) for model validation.
- Model Training: Trained a logistic regression model with a maximum of 1000 iterations to predict lead conversion.
- Model Evaluation: Assessed the model's performance using metrics such as accuracy (0.90), precision (0.91), recall (0.82), F1 score (0.87), and ROC-AUC score (0.96).

4. Model Interpretation

- Feature Importance: Identified the top five most important features contributing to lead conversion, including 'Tags_Will revert after reading the email', 'Tags_Lost to EINS', 'Tags_Closed by Horizzon', 'City_Select', and 'Lead Origin_Lead Add Form'.
- Key Dummy Variables: Highlighted the top five dummy variables impacting lead conversion, including 'Last Activity_Olark Chat Conversation', 'Specialization_Select', 'Tags_Interested in other courses', 'What is your current occupation_Unemployed', and 'Tags_Ringing'.

5. Recommendations

- Aggressive Conversion Strategy:
- Prioritize all leads predicted as high probability (1) by the model.
- Increase follow-up frequency and offer special promotions to encourage conversion.
- Provide additional training to interns for handling more leads efficiently.
- Minimized Call Strategy:
- Focus on leads with a high probability of conversion (>0.8) when targets are met early.
- Reduce follow-up calls and rely more on automated emails.
- Reallocate resources to other tasks like lead nurturing and content creation.

Graphs and Visualizations

- Target Variable Distribution: Visualized the distribution of the target variable in training and test sets using count plots.
- Feature Importance: Displayed bar plots of the top 10 most important features and dummy variables based on their importance scores.