Summary Report: Leads Conversion Using Logistic Regression

This report details the process of building a logistic regression model to predict lead conversion.

Steps Performed:

1. **Data Reading:** The lead conversion dataset was loaded.
2. **Basic Information:** An overview of the data was obtained.
3. **Missing Value Handling:** Missing values were imputed for categorical features (new category) and numerical features (mean value).
4. **Exploratory Data Analysis (EDA):**
    - The target variable was imbalanced, with more negative examples (non-conversions).
    - Certain categories within categorical features showed a correlation with conversion rates (e.g., higher conversion for reference source).
    - Continuous numerical features had left-skewed distributions with outliers. Time spent and total website visits showed a positive impact on conversion.
5. **Outlier Detection and Capping:** Outliers were identified in total visits and page views per visit. Capping was applied using the interquartile range (IQR) to limit their influence.
6. **Feature Selection:** Feature importance was analyzed using heatmaps, likely to identify redundant or irrelevant features and Select K beast score for classification model.
7. **Model Selection and Training:** A logistic regression model was trained. It achieved an accuracy of 72%, successfully classifying leads, and an F-score of 71, indicating good performance.

Next Steps:

- Hyperparameter tuning to potentially improve model accuracy.

Overall, this report demonstrates the development of a logistic regression model for lead conversion prediction. The model achieved promising initial results, and further optimization through hyperparameter tuning is recommended.