City 311 Insight Challenge: Forecasting and Anomaly Detection Report

Name: Chaw Thiri San (chawthiri177@gmail.com)

The report aims to explain the step-by-step decision-making processes behind the chosen models.

* Steps involved (The steps are not exactly serial.)
  + Data Retrieval (The corresponding codes are removed from the notebook, as they take ~20 mins to run.)
  + EDA
  + Optimize data using log transformation and winsorizing.
  + Look for time series patterns using periodograms and seasonal plots.
  + Model training and evaluation using training, validation, and test sets.
  + Find the optimal time-series features using feature importance.
  + Visualize trends and feature correlation on the Power BI dashboard.
* Model Used:
  + XGBoostRegressor for time series forecasting.
  + Statistical model for anomaly detection.
* XGBoost Regressor:
  + Temporal indicators: day-of-week, month, and holiday flags.
  + Lag variables (e.g., previous day call volume, last week call volume, rolling mean).
  + Trend and rolling statistics
  + The model was trained on pre-processed temporal data, and early stopping with validation loss was used to prevent overfitting.
* Standard deviation-based statistical model for anomaly detection.
  + After generating forecasts on unseen data (y\_pred), anomalies were flagged based on deviations from expected call volumes.
* Anomaly = predicted call volume is below or above a threshold, determined from standard deviation results for every month. Months with large SD are assumed to be having extreme ups and downs in daily call counts.
* The extreme anomalies obtained are used to retrain the forecasting model for optimal results.