Forecasting Monthly Usage of İBB Public Toilets Using a Single Global Model

Student Name: Mehmet Aytuğ Yürük

Course Name: Introduction to Data Science

Instructor Name: Fatma Önay Koçoğlu

University Name: İstanbul University

Submission Date: January 2026

1 Abstract

Accurate demand forecasting is essential for the efficient operation of urban public services, enabling improved planning of staffing, cleaning frequency, and consumable replenishment. This study investigates monthly usage demand for public toilets operated by the Istanbul Metropolitan Municipality using openly available administrative records. The task is formulated as a **global multi-station time series regression problem**, where a single model jointly learns patterns across **26 stations** under **endogenous-only** constraints. The available dataset spans the period from 2020 to 2024, covering five consecutive years of monthly observations.

To mitigate distortions stemming from pandemic-related mobility disruptions, all data from 2020 are excluded, and data from 2021 are likewise not used for training but retained solely to construct lag-based features for early 2022 samples. The model is trained on **2022–2023** and evaluated on **2024**. Evaluation employs an operationally realistic **rolling one-step-ahead** protocol, where each month is predicted using the most recent observations available at prediction time.

A CatBoost gradient boosting regressor is employed due to its strong performance on tabular data and native handling of high-cardinality categorical features. After training the model, its performance is evaluated against three transparent baselines: Seasonal Naive (lag-12), Rolling Mean (rolling-3), and a Hybrid baseline combining both components. On the 2024 test period, Catboost achieves **MAE = 3986.18** and **RMSE = 5499.99**, outperforming the strongest baseline (Hybrid; MAE = 4389.13, RMSE = 6294.16). Beyond accuracy, the unified global architecture reduces operational overhead by replacing multiple station-specific models with a single deployable artifact.

Keywords: global time series forecasting, CatBoost, urban public services

2 Introduction

Urban public services must operate under fluctuating and often seasonal demand. For facilities such as public toilets, underestimation can cause service degradation, while overestimation can waste limited municipal resources. In a large and touristic city like Istanbul, usage demand varies substantially across locations leading to strong inter-station heterogeneity in both baseline volume and volatility.

This work focuses on forecasting monthly toilet usage at the station level for toilets operated by the Istanbul Metropolitan Municipality. Rather than building an independent forecasting model per station such as ARIMA, the study adopts a **global forecasting**

strategy. A global model learns shared temporal structure across stations while still capturing station-specific differences through a categorical station identifier. This approach is particularly suitable for settings with limited observations per station, where purely local models can overfit or become brittle.

A deliberate design choice in this study is the **endogenous-only constraint**: the model uses only historical usage signals and time-derived features. This isolates the predictive power of internal demand dynamics and establishes a strong baseline before the integration of exogenous drivers (e.g., weather, holidays, special events).

The primary contributions of this study are: (i) a preprocessing pipeline that aggregates gate-level and age-group counts to produce station-month totals and filters stations to ensure consistent temporal coverage; (ii) a feature engineering strategy that encodes short-term, medium-term, and annual seasonal patterns in a tabular regression format; and (iii) an evaluation protocol reflecting real municipal operations, where each month's forecast is generated using only the most recently observed data.

3 Related Work

The methodology of this study is situated at the intersection of global time series modeling, urban utility demand analysis, and gradient boosting regression.

Global vs. Local Forecasting Paradigms. The field of time series forecasting has undergone a significant paradigm shift from local models, which treat each series in isolation, to Global Forecasting Models (GFMs) that learn across a collective dataset [1]. While traditional statistical methods like ARIMA and Exponential Smoothing have been prominent for decades, they leave the potential for recognition of patterns across time series untapped. Recent research demonstrates that GFMs are theoretically robust to data heterogeneity and can achieve superior performance over a collection of local models, regardless of whether the underlying series are strictly related [1]. Furthermore, GFMs are particularly effective in scenarios characterized by data sparsity or short history, as shared model parameters allow for more complex non-linear modeling capabilities while maintaining generalization [1].

Forecasting Challenges in Urban Utility Demand. Accurate demand prediction is a critical requirement for the operational planning and efficiency of urban public services [2]. Similar to public toilet usage, urban electricity and water demand exhibit complex behaviors influenced by non-linear trends and multi-scale seasonal cycles [2, 3]. Research in electricity load forecasting emphasizes the difficulty of managing fluctuating demand patterns driven by urban activity and seasonal weather cycles [2]. In the context of water

demand, multi-scale assessments reveal that consumption behaviors are highly influenced by heterogeneous user profiles and institutional schedules, necessitating models that can adapt to high-resolution smart metering data [3]. These utility frameworks underscore the importance of distinguishing between regular workdays, weekends, and holidays to capture behavioral variations in usage.

Gradient Boosting and Feature Engineering. Gradient boosting algorithms have emerged as a powerful tool for time series regression due to their iterative greedy strategy and ability to capture non-linear relationships that traditional methods cannot satisfactorily model [4]. The use of window-based Gradient Boosting Regression Trees (GBRT) has been validated as highly effective for demand forecasting, particularly when the input structure is configured to learn the autocorrelation among target variables within a specified horizon [5]. This approach effectively utilizes endogenous lag features and time-derived covariates to construct a supervised learning table from raw series. Empirical evidence confirms that incorporating lagged consumption variables acts as a dominant driver of demand variability, substantially improving prediction accuracy relative to baseline models [3]. Such boosting architectures naturally handle the high-cardinality categorical features and non-linear representations common in multi-station urban service environments [4, 5].

4 Data Description

This section provides a comprehensive overview of the dataset utilized in this study, detailing the source of the administrative records and the structure of the raw variables. It further outlines the preprocessing pipeline, including the aggregation logic, station filtering criteria, and the temporal partitioning strategy used to ensure experimental integrity.

4.1 Data Source

The primary dataset for this study is sourced from the Istanbul Metropolitan Municipality Open Data Portal, a public repository of municipal administrative records [6]. It captures monthly usage of public toilets across Istanbul over multiple years, reflecting seasonal and social mobility patterns. Data are recorded at both the gate and age-group level; thus, each row corresponds to the total monthly passages for a specific gate–age-group combination within a station. Since stations often contain multiple gates, aggregating these counts is necessary to obtain total station-level usage.

4.2 Data Dictionary

The raw dataset consists of five primary variables. The details of these variables, including their data types and descriptions, are summarized in Table 1.

Variable	Type	Description
Station	Categorical	Unique station identifier (e.g., A, B, C).
Gate	Categorical	Gate identifier within a station (e.g., X7M, D1X, 12M).
Age Group	Categorical	User age group (e.g., 10–20, 21–40).
Date	Datetime	Monthly timestamp of observation (e.g., 2021-02-01).
Passage_Count	Numerical	Number of passages recorded for the station during the month.

Table 1: Variables in the raw dataset prior to preprocessing.

4.3 Target Variable and Aggregation

The forecasting target is the station-level monthly total usage, denoted as **Passage_Count**. As stations may have multiple gates and usage is recorded for different age groups, both gate and age-group records are aggregated by summing counts within each station and month. This aggregation aligns with the operational objective of planning resources for the entire station, rather than for individual gates or age segments.

4.4 Station Filtering

The raw dataset comprises a diverse set of stations, some of which exhibit discontinuous temporal coverage due to operational changes, closures, or recent openings. To ensure a consistent experimental framework and prevent evaluation artifacts arising from unequal history lengths, a strict filtering protocol was applied. Specifically, the analysis retains only those stations with complete, uninterrupted usage records spanning the entire modeling window. This approach eliminates the need for artificial imputation of missing values, which could otherwise distort the training signal. Consequently, after the filtering process final dataset consists of **26 distinct stations** with reliable long-term historical data.

4.5 Temporal Scope and Split

To mitigate pandemic-related distortions, all observations from 2020 are excluded. Data from 2021 are used only to construct lag-based features for early 2022 and are not included in model training or evaluation.

4.6 Data Split

To respect temporal causality and prevent information leakage, the dataset is partitioned chronologically as follows:

- **Training set:** January 2022 – December 2023 (24 months)
- **Test set:** January 2024 – December 2024 (12 months)

5 Exploratory Data Analysis (EDA)

Exploratory data analysis is conducted to (i) quantify heterogeneity across stations, (ii) examine seasonal patterns in aggregated demand, and (iii) assess variability among the stations.

5.1 Station-Level Demand Heterogeneity

Total usage volumes differ substantially across stations, reflecting location-specific demand drivers such as commuter traffic and tourism intensity. Figures 1 and 2 illustrate this heterogeneity.

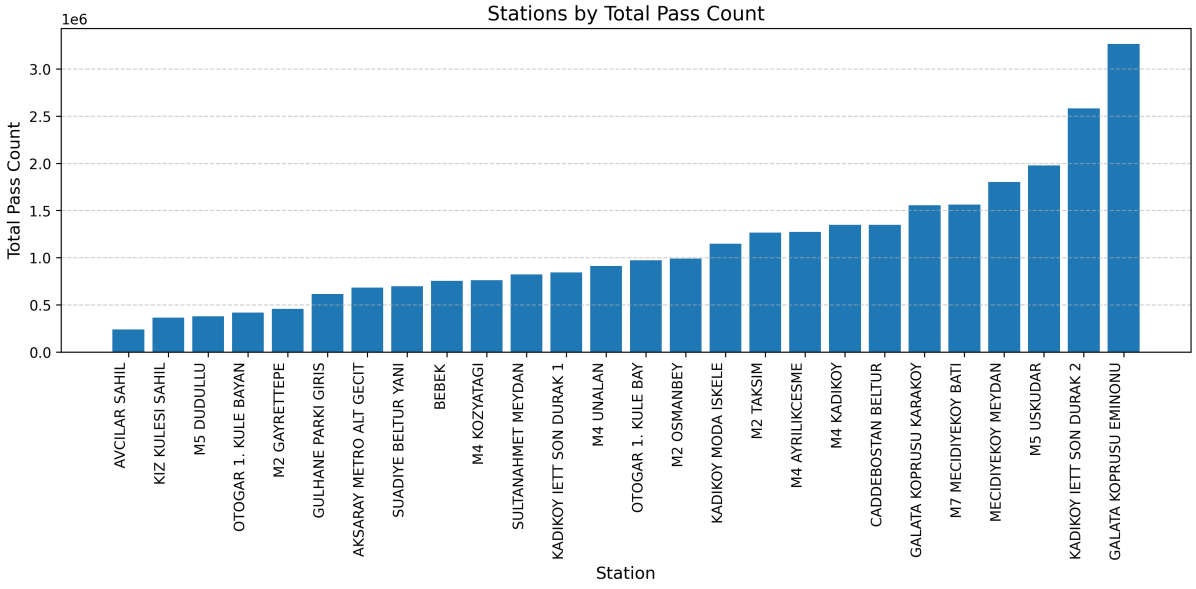


Figure 1: Total usage volume aggregated by station over the 2021-2024 period.

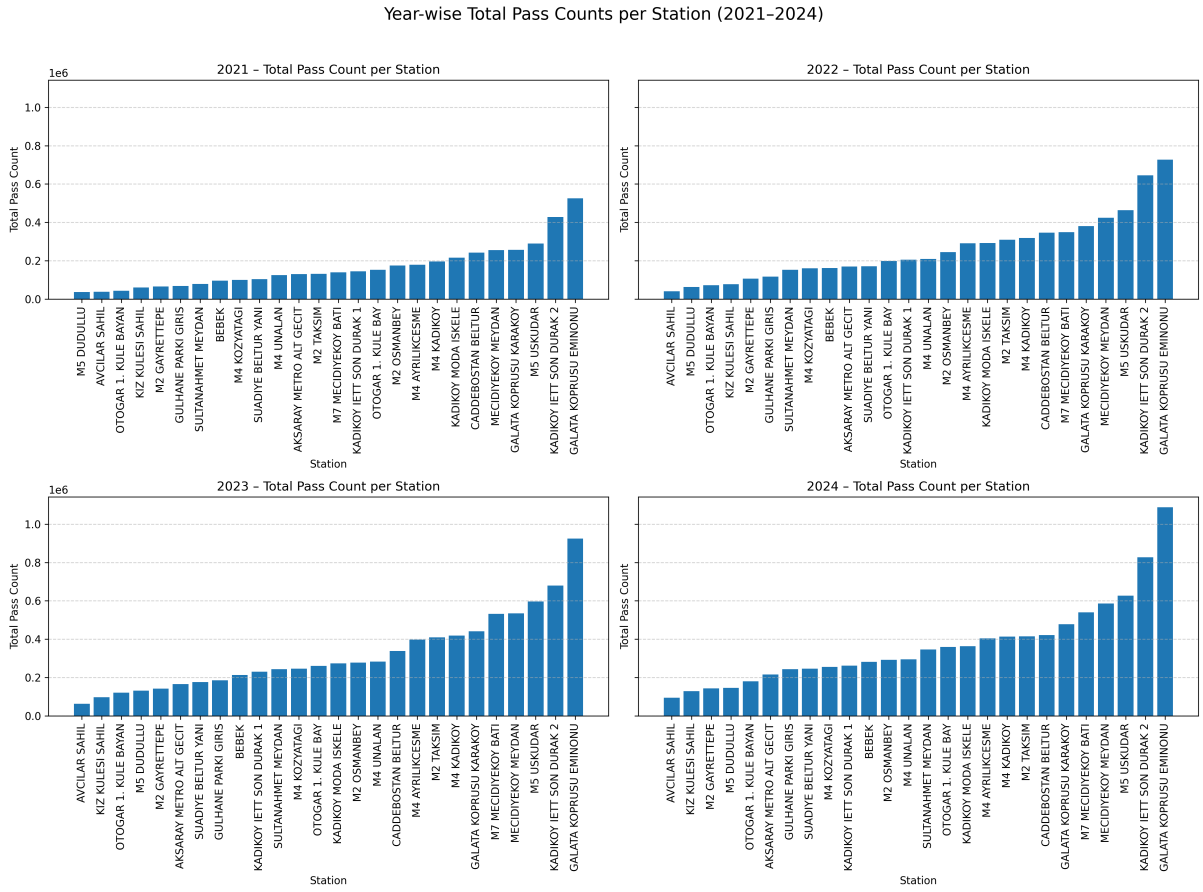


Figure 2: Total usage volume aggregated by station for each year separately (2021–2024).

5.2 Seasonality in Monthly Demand

Aggregating usage across stations and years reveals recurring monthly patterns. Figure 3 shows the total usage for all stations combined, distributed by month across the 2021–2024 period. The figure provides a clear indication of seasonality and motivates the inclusion of both annual lag features and cyclical month encoding in the model.

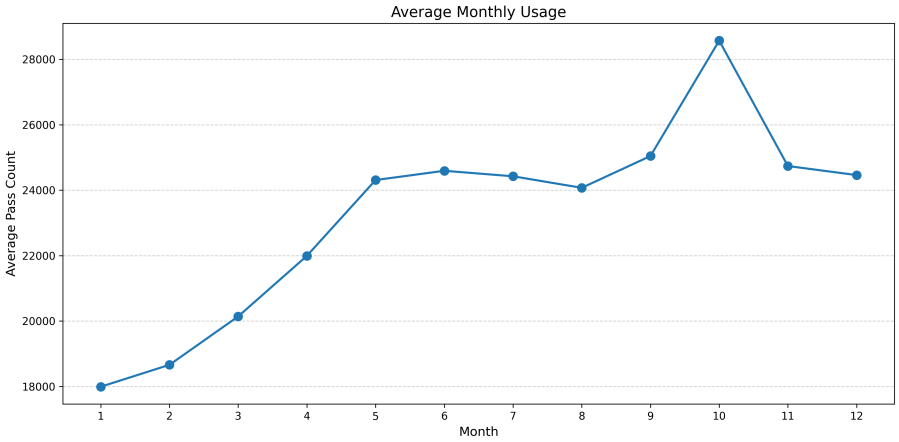


Figure 3: Total monthly usage across all stations aggregated for 2021–2024.

5.3 Station-Wise Usage Distribution

To systematically assess variability and detect potential outliers per station, station-level usage distributions are visualized using boxplots Figure 4. This analysis explicitly highlights the marked differences in volatility and extreme tail behavior across the heterogeneous network of stations.

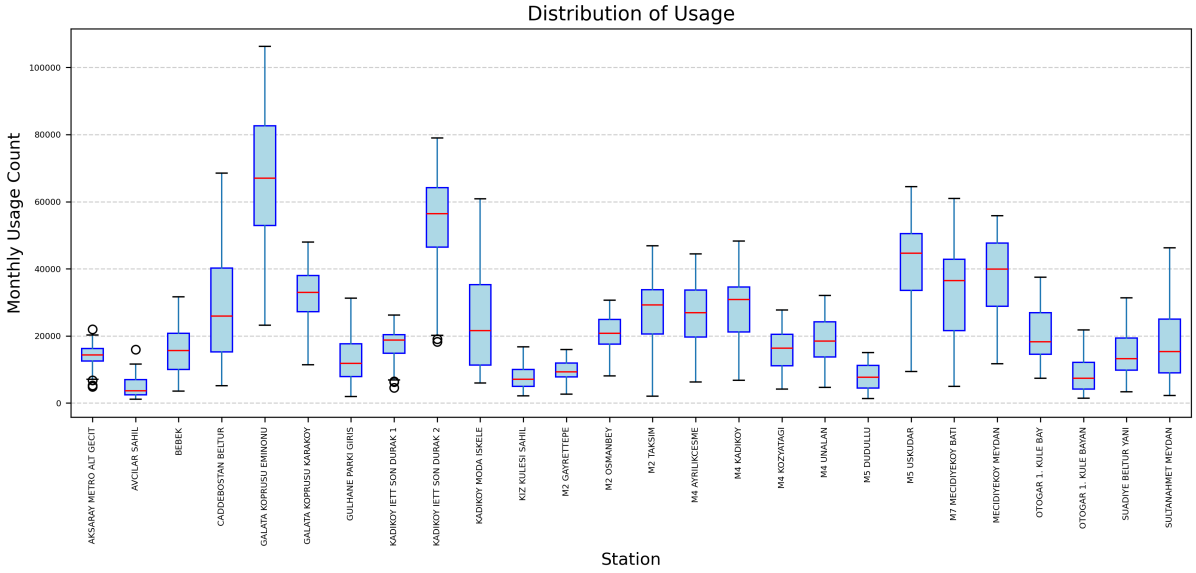


Figure 4: Station-wise distribution of monthly usage.

6 Data Preprocessing and Feature Engineering

The modeling pipeline converts the station-month usage series into a supervised learning table. Features are designed to capture (i) short-term autoregressive dependence, (ii) medium-term local level and volatility, and (iii) annual seasonality.

6.1 Lag Features

To capture temporal dependencies in station-level usage, lagged features are constructed for each station. These include:

- **Short-term lags:** lag_1, lag_2, lag_3, representing usage from the preceding 1, 2, and 3 months.
- **Annual lag:** lag_{12}, representing usage from the same month in the previous year to encode yearly seasonality.

Formally, each lag feature is defined as

$$\text{lag}_k(t) = y_{t-k}, \qquad k \in \{1, 2, 3, 12\},$$

where y_t denotes the observed usage at month t. These features enable the model to leverage both short-term persistence and recurring annual patterns in the demand series.

6.2 Rolling Statistics

To capture the local trend level while mitigating high-frequency noise, a 3-month rolling mean is calculated. Crucially, to prevent data leakage, this statistic is strictly computed using only historical observations available prior to the prediction time t:

$$\text{rolling\_3}(t) = \frac{1}{3} \sum_{i=1}^{3} y_{t-i}$$

Additionally, a rolling standard deviation feature (rolling_3_std) is derived over the same window to quantify local volatility and demand uncertainty.

6.3 Momentum and Relative Change

To explicitly encode the recent trajectory and velocity of demand, momentum features are derived from the first and third-order differences:

$$\text{mom\_1}(t) = y_{t-1} - y_{t-2}, \quad \text{mom\_3}(t) = y_{t-1} - y_{t-3}$$

Furthermore, to account for station heterogeneity, a relative change feature is introduced. This provides a scale-invariant measure of growth or decline, normalized by the baseline magnitude:

$$\text{pct\_3}(t) = \frac{y_{t-1} - y_{t-3}}{|y_{t-3}| + \epsilon}$$

where ϵ is a small constant added to ensure numerical stability and prevent division by zero.

6.4 Cyclical Month Encoding

The temporal nature of calendar months is inherently cyclical; however, standard ordinal encoding introduces an artificial numerical discontinuity between December ($m = 12$) and January ($m = 1$). To preserve the topological proximity of consecutive months across year

boundaries, the month variable $m \in \{1, \ldots, 12\}$ is projected onto a unit circle using sine and cosine transformations:

$$\text{month\_sin} = \sin\left(\frac{2\pi m}{12}\right), \qquad \text{month\_cos} = \cos\left(\frac{2\pi m}{12}\right)$$

This transformation allows the model to correctly interpret the continuity between the end of one year and the start of the next.

6.5 Feature Set

The final input vector provided to the model comprises **13 engineered features** covering categorical identifiers, temporal encodings, and historical dynamics. The complete feature set is summarized in Table 2.

Table 2: Engineered feature set used for model training.

Category	**Feature Variable**	**Description**
Identifiers	Station	Station identifier (categorical).
	Month	Calendar month (ordinal; 1 . . . 12).
Cyclical	*month_sin*, *month_cos*	Sine–cosine encoding of the month to preserve cyclical continuity.
Lags	*lag_1*, *lag_2*, *lag_3*	Short-term autoregressive lagged values.
	lag_12	Annual seasonal lag.
Rolling Stats	*rolling_3*	3-month moving average.
	rolling_3_std	3-month moving standard deviation.
Dynamics	*mom_1*, *mom_3*	Momentum over 1- and 3-month horizons.
	pct_3	3-month relative percentage change.

7 Model Development

This section details the methodology employed to construct and optimize the forecasting framework. It defines the heuristic baselines established for comparative benchmarking and describes the architecture and training protocol of the primary CatBoost gradient boosting regressor, including the time-aware cross-validation strategy used for hyperparameter tuning.

7.1 Baselines

To rigorously benchmark the proposed model, three transparent heuristic baselines are established:

- **Seasonal Naive (Lag-12):** Assumes demand strictly repeats on an annual cycle.

$$\hat{y}_t = y_{t-12}$$

- **Rolling Mean (Rolling-3):** Captures the local level using the average of the most recent quarter.

$$\hat{y}_t = \frac{1}{3}\sum_{i=1}^{3} y_{t-i}$$

- **Hybrid Baseline:** An equally weighted ensemble combining seasonal memory and local trend.

$$\hat{y}_t = 0.5 \cdot y_{t-12} + 0.5 \cdot \left(\frac{1}{3}\sum_{i=1}^{3} y_{t-i}\right)$$

7.2 CatBoost Regressor

A CatBoost gradient boosting regressor is used as the primary model because it performs strongly on tabular feature sets and can capture non-linear interactions between lag-based dynamics and seasonal structure. The model is trained with **RMSE** as the optimization objective:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}.$$

The station identifier is provided as a categorical feature, enabling the global model to learn station-specific offsets and seasonal behaviors while sharing parameters across all series.

7.3 Time-Aware Cross-Validation for Hyperparameter Selection

Hyperparameters are selected using a time-aware procedure. Unique monthly timestamps in the training window are split into three chronological folds using TimeSeriesSplit. For each fold, station-month samples matching the selected training dates form the training indices, while those for validation dates form the validation indices. This preserves temporal order and prevents mixing past and future information.

8 Model Evaluation

This section presents the empirical results of the forecasting experiments on the test set. It defines the quantitative metrics used for assessment, provides a comparative performance analysis against baselines, and offers a qualitative inspection of station-level forecast error distributions to assess model robustness.

8.1 Operational Evaluation Protocol

Evaluation mirrors real municipal usage: forecasts are generated monthly using only information available at prediction time, following a one-step-ahead setting.

8.2 Metrics

The regression performance is evaluated using MAE and RMSE, as defined in (1) and (2):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad (1) \qquad\qquad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (2)$$

where N is the sample size, and y_i and $\hat{y}_i$ are the actual and predicted values, respectively.

8.3 Overall Results

Table 3 presents the comparative evaluation on the 2024 test period. The CatBoost model demonstrates a clear superiority across both metrics.

Model	**MAE**	**RMSE**
Seasonal Naive (Lag-12)	5901.62	8020.03
Rolling Mean (Rolling-3)	5477.45	7446.98
Hybrid Baseline	4389.13	6294.16
CatBoost (Global Model)	**3986.18**	**5499.99**

Table 3: Predictive performance comparisons on the 2024 test set.

Relative to the strongest Hybrid baseline model, CatBoost reduces MAE by approximately **9.2%** and RMSE by approximately **12.6%**, indicating that the learned non-linear interactions between station identity, seasonal structure, and short-term dynamics provide additional predictive value beyond simple heuristics.

8.4 Station-Level Error Analysis

To examine heterogeneity in predictability, station-wise error metrics are computed on the test set. The error distribution illustrates how forecasting difficulty varies across different locations.

Figure 5 reveals significant variation in station-level RMSE. Error magnitudes largely correlate with volatility; stable stations yield lower errors than those with abrupt changes. This non-uniform predictability necessitates complementary qualitative inspection.

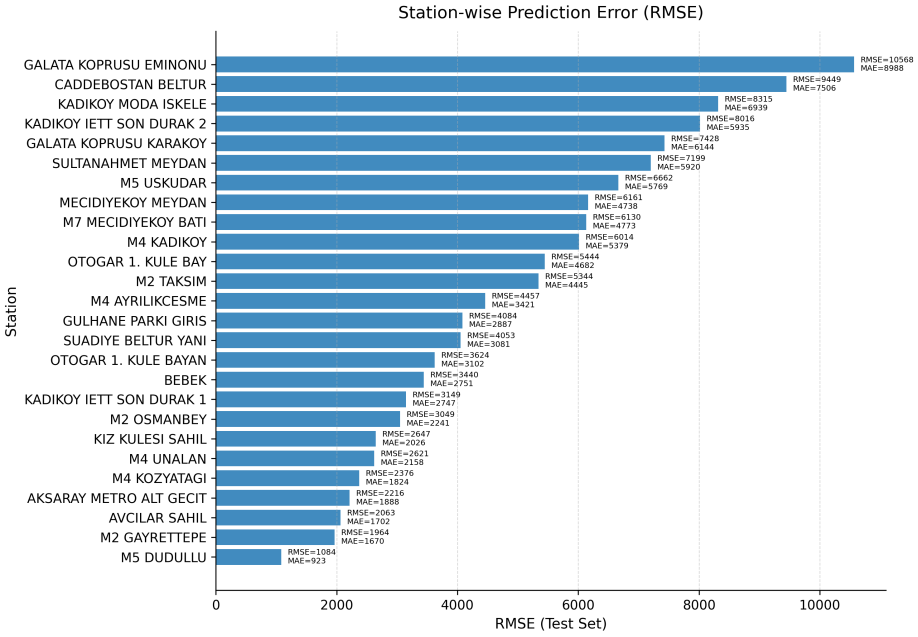


Figure 5: Distribution of station-level prediction errors on the 2024 test set.

8.5 Qualitative Forecast Diagnostics

While aggregate and station-level error metrics provide a summary of predictive accuracy, they do not fully capture the temporal alignment between forecasts and observed usage patterns. In particular, quantitative metrics alone may obscure systematic timing errors, peak mismatches, or deviations during periods of rapid demand change.

To complement the quantitative results, qualitative diagnostics are conducted through visual inspection of predicted and actual time series. This analysis examines model behavior for high-demand stations (Figure 6), low-demand stations (Figure 7), and the complete set of stations (Figure 8), allowing assessment of trend tracking, seasonal consistency, and cross-station generalization.

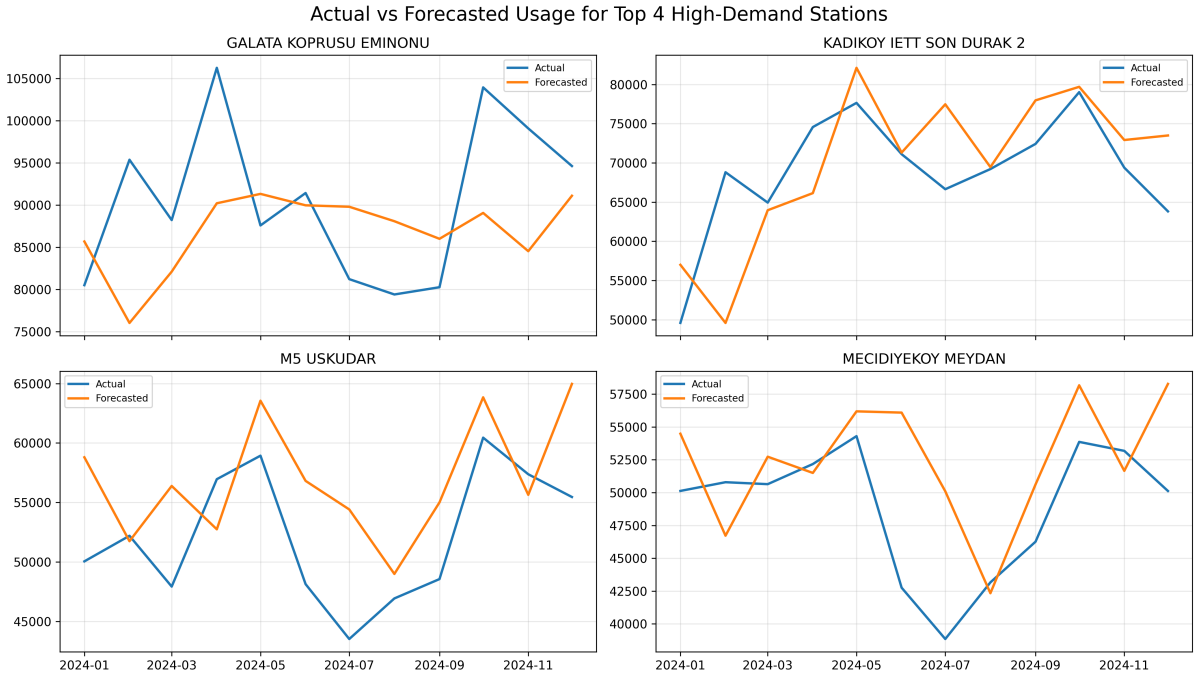


Figure 6: Actual vs. predicted usage for the four highest-demand stations (Test period).

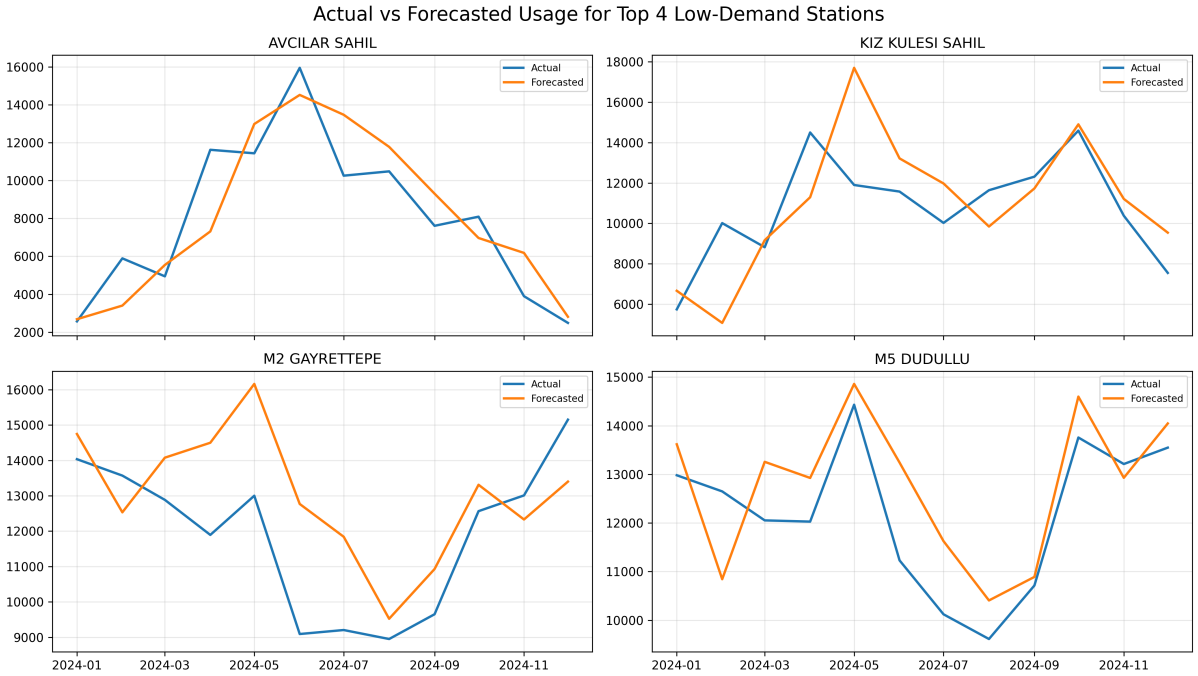
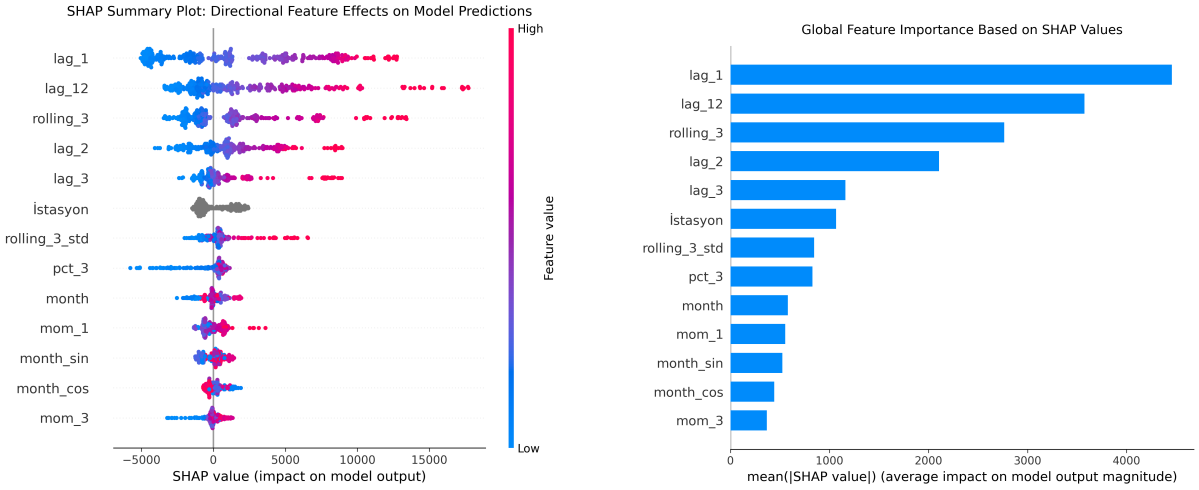


Figure 7: Actual vs. predicted usage for the four lowest-demand stations (Test period).

Figure 8: Actual vs. predicted usage for all 26 stations.

9 Interpretation and Explainability

To improve model interpretability, explainable AI analysis is conducted using SHAP to quantify feature-level contributions to model predictions. Figure 9a presents global feature importance, while Figure 9b illustrates the directional impact of features on the predicted demand.



(a) Global feature importance.

(b) Directional feature effects.

Figure 9: SHAP explanations for the CatBoost model.

10 Discussion

10.1 Interpretation of Results

The experimental results validate the efficacy of a **Global Forecasting Model (GFM)** approach for municipal infrastructure. By training a single CatBoost regressor across 26 distinct stations, the model successfully decoupled station-specific volatility from shared temporal dynamics. The superior performance over the Hybrid baseline suggests that the model effectively learned non-linear interactions between annual seasonality and short-term momentum patterns that simple linear combinations of lag features cannot capture.

Critically, the successful application of the endogenous-only constraint implies that historical usage patterns contain a significant portion of the signal required for monthly planning. However, the station-level error analysis revealed that while the model excels at tracking trend and seasonality, it struggles with sporadic peak events. This behavior is consistent with the absence of external regressors; without explicit signals regarding holidays, tourism influxes, or extreme weather, the model is mathematically limited to autoregressive smoothing.

10.2 Operational Implications

From an operational perspective, the global architecture offers a distinct advantage over local modeling strategies (e.g., ARIMA per station). In a production environment, maintaining 26 separate models requires significantly more computational overhead and monitoring effort than a single unified artifact. The ability of CatBoost to handle new stations or stations with shorter histories further enhances the system's scalability for expanding municipal networks.

10.3 Alignment with Literature

These findings corroborate recent advances in time series forecasting which argue that cross-series information sharing improves generalization in heterogeneous datasets [1]. Furthermore, the dominance of lag-based features observed in the SHAP analysis aligns with similar studies on urban utility demand [5, 3], confirming that immediate past consumption remains the strongest predictor of near-future demand in human-centric systems.

11 Conclusion

This study presented a framework for forecasting monthly public toilet usage in Istanbul using a global supervised learning approach. By treating the network of 26 stations as a unified regression task, the developed CatBoost model demonstrated that robust predictions can be generated solely from administrative usage records, without reliance on complex external data pipelines.

The evaluation on the 2024 test set established the proposed model as the superior candidate against all benchmarks. It achieved a **MAE of 3986.18** and an **RMSE of 5499.99**, representing a performance gain of approximately **9.2%** and **12.6%** respectively over the strongest heuristic baseline.

The primary contribution of this work is the demonstration that a global model can effectively balance network-wide generalization with station-specific specificity. For municipal authorities, this translates into a scalable tool for resource allocation enabling data-driven decisions on staffing and supplies while minimizing the complexity of the machine learning lifecycle.

12 Future Work

To further enhance the operational utility and accuracy of the forecasting system, future research should focus on three key areas:

- **Integration of Exogenous Drivers:** To address the residual errors observed during peak demand, the feature space should be augmented with external datasets. Specifically, incorporating weather data and a calendar of public holidays and large-scale Istanbul events would allow the model to anticipate non-cyclical demand surges.
- **Enhanced Feature Engineering:** Additional derived features such as rolling quantiles, demand volatility indices, or interaction terms between station and temporal signals may improve robustness.
- **Exploration of Deep Learning Architectures:** While Gradient Boosting is effective for tabular data, deep learning models such as Long Short-Term Memory networks or Temporal Fusion Transformers may capture complex sequential dependencies and interactions across the global dataset.

References

[1] H. Hewamalage, C. Bergmeir, and K. Bandara, "Global models for time series forecasting: A simulation study," 2021.

[2] L. Hou, X. Shen, and L. Zhang, "Urban electricity demand forecasting with a hybrid machine learning model," in *2023 International Conference on Networking, Sensing and Control (ICNSC)*, vol. 1, pp. 1–6, 2023.

[3] E. Farah and I. Shahrour, "Forecasting urban water demand using multi-scale artificial neural networks with temporal lag optimization," *Water*, vol. 17, no. 19, 2025.

[4] V. Mayrink and H. S. Hippert, "A hybrid method using exponential smoothing and gradient boosting for electrical short-term load forecasting," in *2016 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pp. 1–6, 2016.

[5] L. Zhang, J. Wang, Q. Zeng, H. Zhou, M. Hu, Y. Guo, and Q. Wang, "Predicting the energy meter demand by a window-based gradient boosting regression tree model," in *2023 IEEE International Symposium on Product Compliance Engineering - Asia (ISPCE-ASIA)*, pp. 1–6, 2023.

[6] İstanbul Büyükşehir Belediyesi, "Şehir Tuvaletleri Kullanım İstatistikleri." İBB Açık Veri Portalı, 2025. Erişim Tarihi: 03 Ocak 2026.