A Machine Learning Approach to Forecasting NYC EMS Call Volume

Forecasting Borough-Level EMS Demand for Resource Optimization

MD A. Rahman Ruhel

Data Science Fellow – Springboard
Burning Wheel – Data Analytics Consultant

Introduction

New York City operates one of the busiest emergency medical services systems in the world. The **NYC 911 ambulance service** responds to hundreds of thousands of Emergency calls each year across its five boroughs — Bronx, Brooklyn, Manhattan, Queens, and Staten Island. The vast majority of these calls are handled **by FDNY Emergency Medical Services**, with the remainder covered by private or volunteer ambulance providers. Call volume fluctuates daily due to seasonal patterns, weather conditions, population density, and special events, creating operational challenges for timely response and resource allocation.

This project aims to **forecast daily 911 ambulance call volumes** to support effective staffing, ambulance deployment, and operational planning. Accurate predictions help ensure faster response times and better preparedness for high-demand periods such as weekends, holidays, and extreme weather events.

Two predictive modeling strategies are evaluated:

1. **Citywide model** — a single XGBoost model using borough indicators to capture geographic variation
2. **Borough-wise models** — separate XGBoost models trained for each borough to better reflect localized patterns in EMS demand

By comparing these approaches, the project identifies where granular, borough-level modeling provides operational advantage, and how machine learning can support FDNY in improving response time reliability across New York City.

Data Collection and Wrangling

The dataset used in this project was derived from publicly available NYC Open Data on daily ambulance call volumes, ranging from July 1, 2024 to April 25, 2025.

URL = https://data.cityofnewyork.us/resource/76xm-jjuj.json

Additional features were merged from external sources such as:

- Weather data (temperature and conditions)
- Holiday calendar
- Temporal features (day of week, lagged call volumes, rolling averages)

Data wrangling steps included:

- Cleaning null or inconsistent entries,
- Converting timestamps to daily resolution,
- Feature engineering (Creating lag-based features - *lag_1, lag_2, week_diff, roll_3, roll_7*),
- Encoding categorical variables like boroughs and severity level using one-hot encoding.

This prepared dataset was then used for both statewide and borough-wise model training.

Exploratory Data Analysis (EDA)

A comprehensive exploratory analysis was performed to understand the temporal, spatial, and environmental factors influencing daily ambulance call volumes across New York City.

EDA revealed distinct temporal patterns:

Ambulance call volumes tend to be higher in weekdays than weekends and holidays.

Each borough displays a unique daily call distribution — Manhattan and Brooklyn generally report higher volumes, while Staten Island remains the lowest.

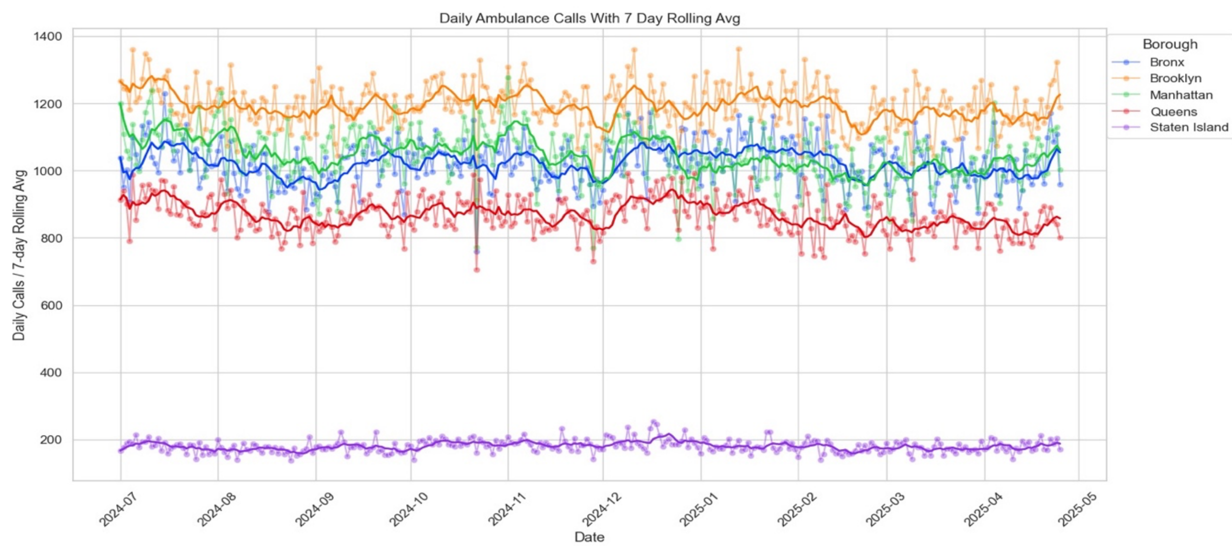


Fig: *Daily Ambulance Calls With 7 Day Rolling Avg*

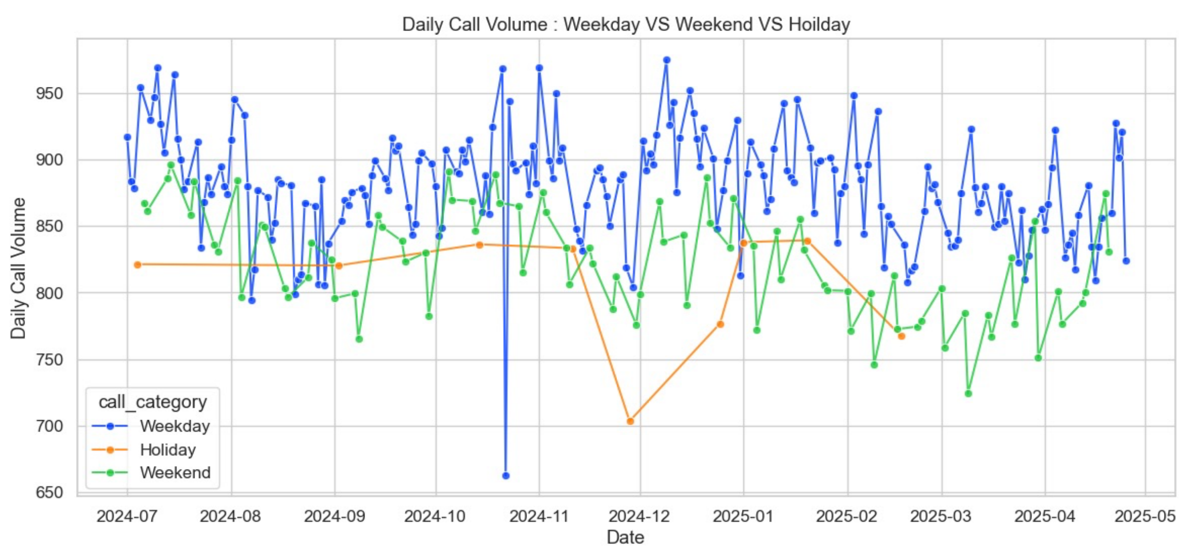


Fig: *Daily Call Volume: Weekday VS Weekend VS Holiday*

At this stage, temperature (temp) appeared to have little to no linear correlation with daily call volume. This initially suggested weather might not play a major role in emergency call fluctuations. However, as discussed later, feature importance analysis provided deeper insights into this assumption.

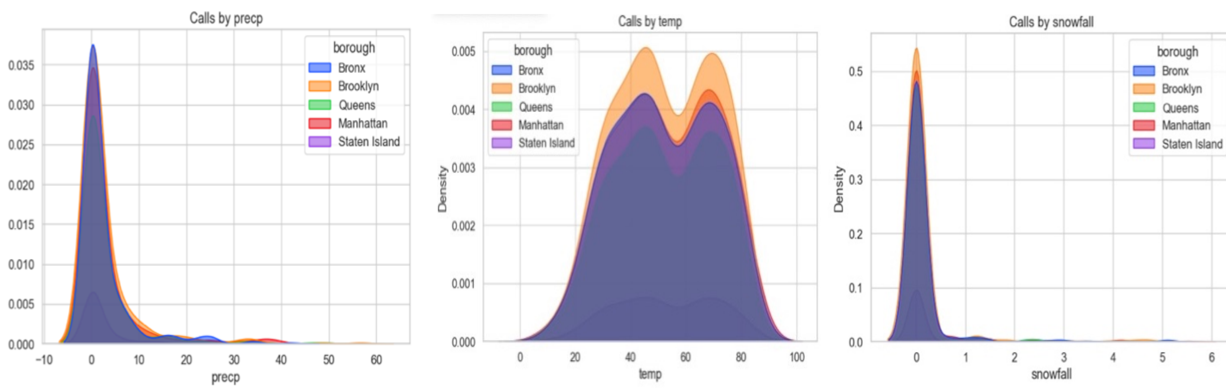


Fig: *KDE (Kernel Density Estimate) Plots*

The boxplots show that the data is well-distributed within the interquartile range (IQR) for all boroughs. a few outliers are visible in weekday call volumes for the boroughs Bronx, Queens, Manhattan, and Staten Island.

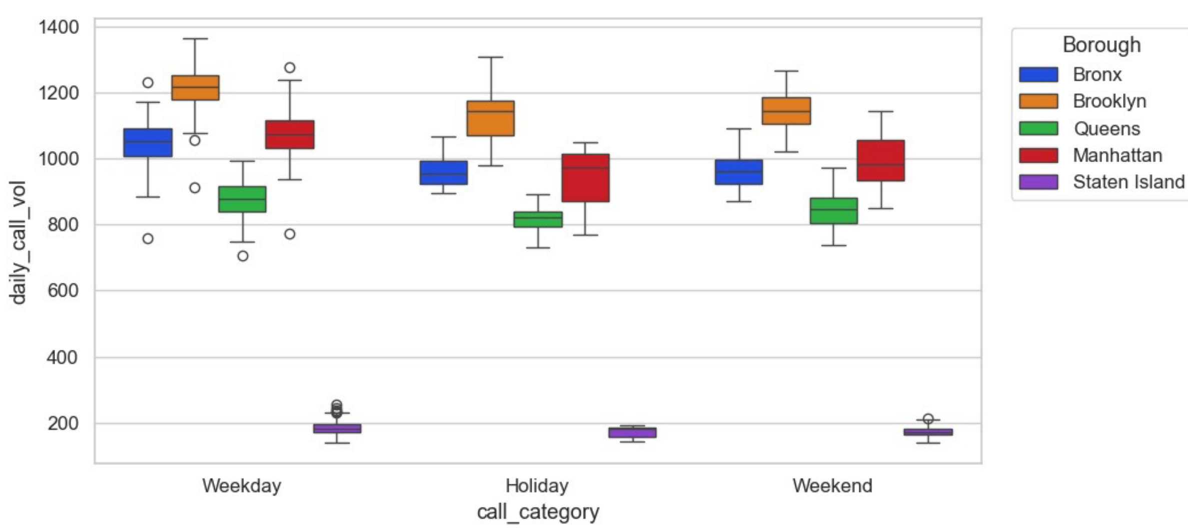


Fig: *Boxplot – Weekday, Weekend, Holiday*

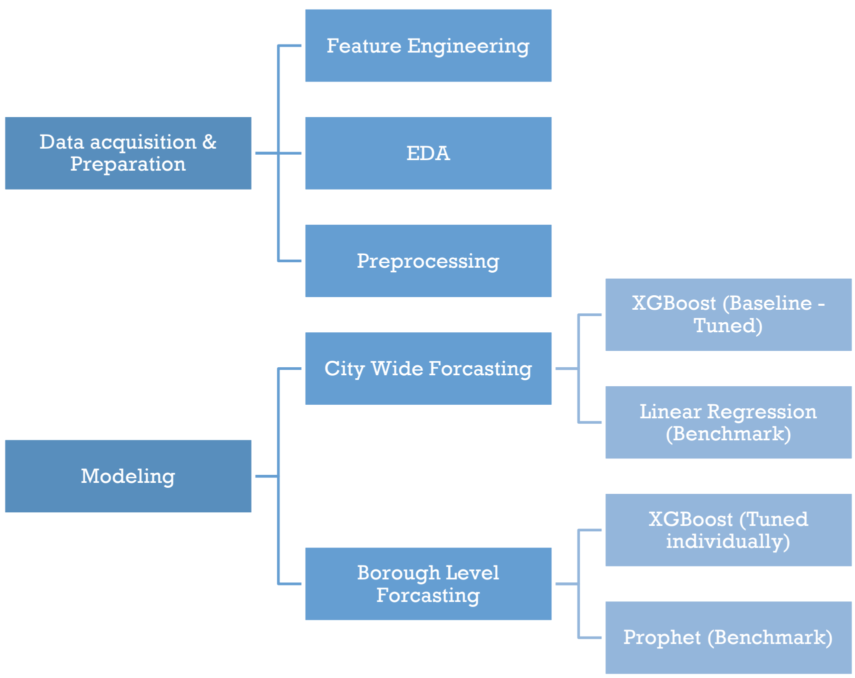
Modeling Approach

In this study, we implemented two complementary modeling strategies to forecast ambulance demand in New York City.

First, we built a single citywide model, representing borough identity with one-hot encoding. This approach allowed the model to learn general patterns across the city. We trained an XGBoost model and improved its performance through Randomized Search hyperparameter tuning, using Linear Regression as a simple benchmark.

Second, we developed individual borough-level models to better capture local variations in ambulance activity. For each borough, we trained and tuned XGBoost separately and used Prophet as a reference model due to its strength in modeling seasonality and trends.

These two approaches together provide both a high-level citywide view and detailed borough-specific insights for resource planning.



Citywide Model

Linear Regression (Benchmark model):

As a baseline benchmark, a regularized Linear Regression model was evaluated using cross-validation. The model achieved RMSE scores ranging from approximately **48.9 to 78.1** across folds, with an average CV RMSE of 62.90. While the linear baseline captured general seasonal trends, performance variability across folds indicates that ambulance demand exhibits nonlinear behavior and borough-specific fluctuations that a simple linear approach struggles to model effectively. Therefore, a more sophisticated method like XGBoost was pursued to better learn complex temporal interactions.

XGBoost Model:

A single model was trained using data from all boroughs combined, with boroughs represented through one-hot encoding. This approach allowed the model to learn both shared and borough-specific patterns within one predictive framework. The final citywide forecasting model was developed using **XGBoost**, as it demonstrated strong ability to capture nonlinear temporal patterns in ambulance demand. After establishing a baseline model, we performed **hyperparameter tuning using Randomized Search**, which led to significant performance improvements. The optimal configuration included a reduced tree depth and feature sampling strategy, which helped control overfitting while maintaining high predictive accuracy:

Hyperparameters were tuned through RandomizedSearchCV, optimizing for RMSE.

Key tuned parameters included:

Parameters	Value
n_estimators	100
max_depth	3
learning_rate	0.05
subsample	0.6
colsample_bytree	0.6

Final statewide model performance:

Metrics	Value
Test RMSE	47.84
Test MAE	36.93
Test R^2	0.98
MAPE	5.09%

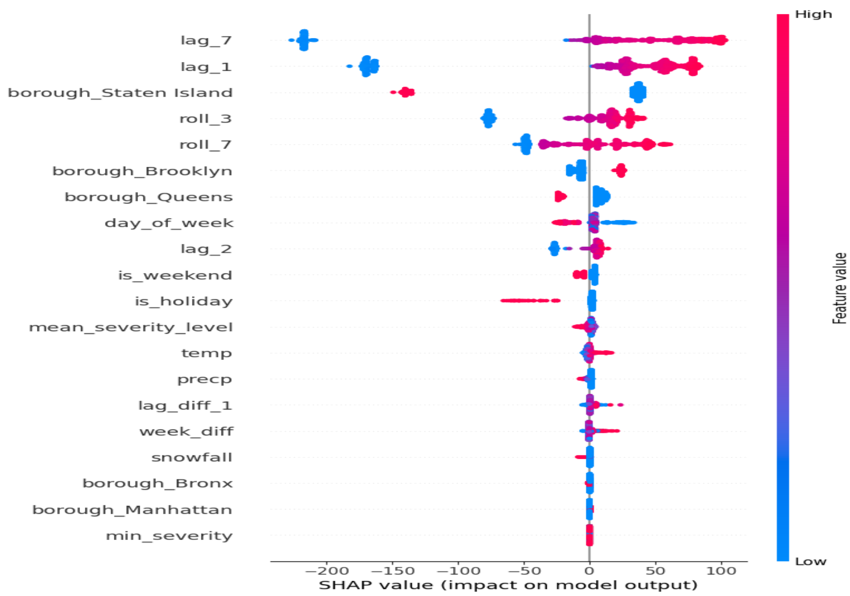
Summary table:

Model	Avg CV RMSE	Notes
Linear Regression	62.90	Baseline: Struggles with nonlinear patterns
XGBoost	47.84	Final selected model

The model demonstrates strong generalization performance with low forecast error **(RMSE ≈ 48, MAE ≈ 37).** The **MAPE of 5.09%** indicates that predictions deviate only slightly from actual ambulance demand on average. A high **R^2** value **(0.98)** confirms that the model effectively captures the key drivers of call volume, supported by residual analysis showing no noticeable bias. Collectively, these metrics validate the model's reliability for short-term operational forecasting.

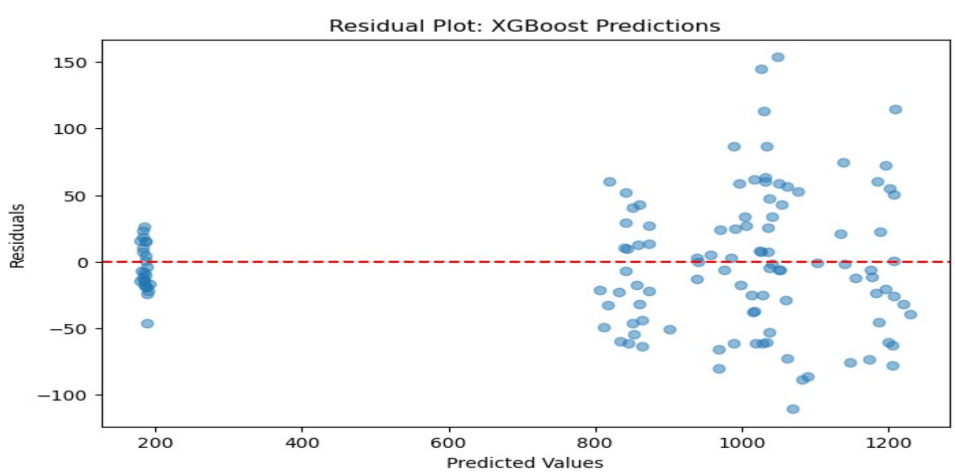
Citywide SHAP Analysis:

To interpret the citywide XGBoost model, SHAP (SHapley Additive exPlanations) values were computed to quantify feature contributions to predictions. The analysis revealed that lag features and rolling features were the most influential predictors. These results confirm that ambulance demand is highly dependent on recent trends and recurring temporal patterns, providing actionable insights for operational planning.



Residual Plot:

Residual analysis shows that errors are centered around zero, indicating no systematic bias. However, as predicted call volumes increase, the dispersion of residuals grows, reflecting higher uncertainty on high-demand days. This heteroscedasticity is typical for EMS call data and highlights limitations in predicting peak volumes, even with strong overall model performance.



Actual vs Predicted Line Plot:

The model successfully captures the overall demand patterns across boroughs, though some short-term fluctuations remain challenging to predict precisely.

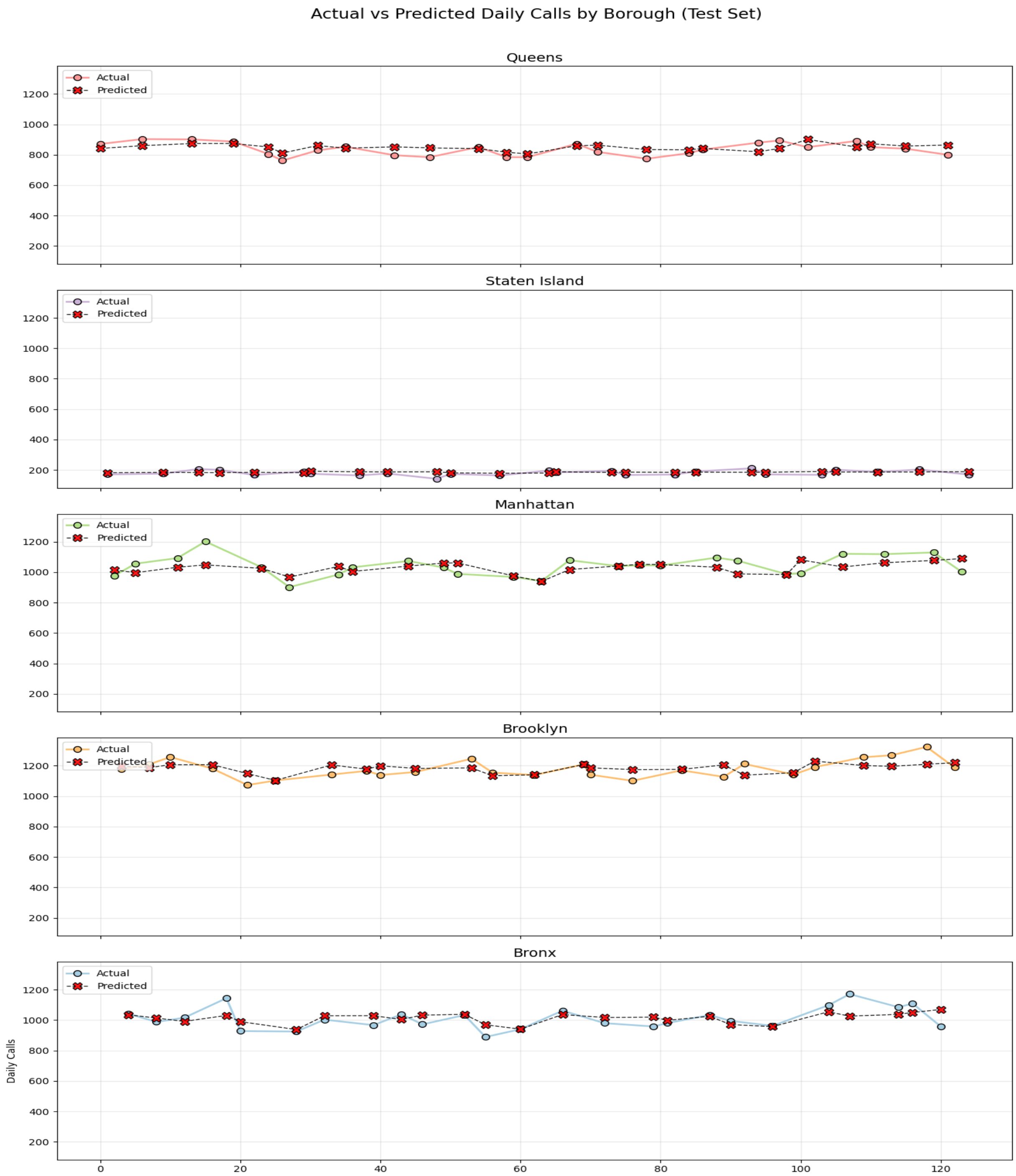


Fig: Actual vs Predicted Daily Calls by Borough.

Borough-wise Modeling

Ambulance demand varies significantly across New York City's boroughs due to differences in population density, local events, and temporal patterns. To better capture these localized trends, separate models were developed for each borough using XGBoost, allowing the algorithm to learn nonlinear interactions and temporal dependencies specific to each area. For comparison and interpretability, Prophet models were also trained as benchmarks to capture seasonal and trend components. Feature engineering for the borough-level models followed the same approach as the citywide model, including date-based features, lag features, and rolling averages, ensuring that each model had access to both recent demand patterns and seasonal signals relevant to its respective borough.

Performance Metrics (XGBoost Vs Prophet):

Borough	RMSE (XGBoost)	R^2 (XGBoost)	RMSE (Prophet)	R^2 (Prophet)
Bronx	55.32	0.38	58.01	0.29
Brooklyn	58.59	0.18	62.91	-0.09
Manhattan	72.94	-0.09	72.42	-0.19
Queens	44.07	-0.13	43.79	-0.15
Staten Island	16.88	-0.10	20.25	-0.53

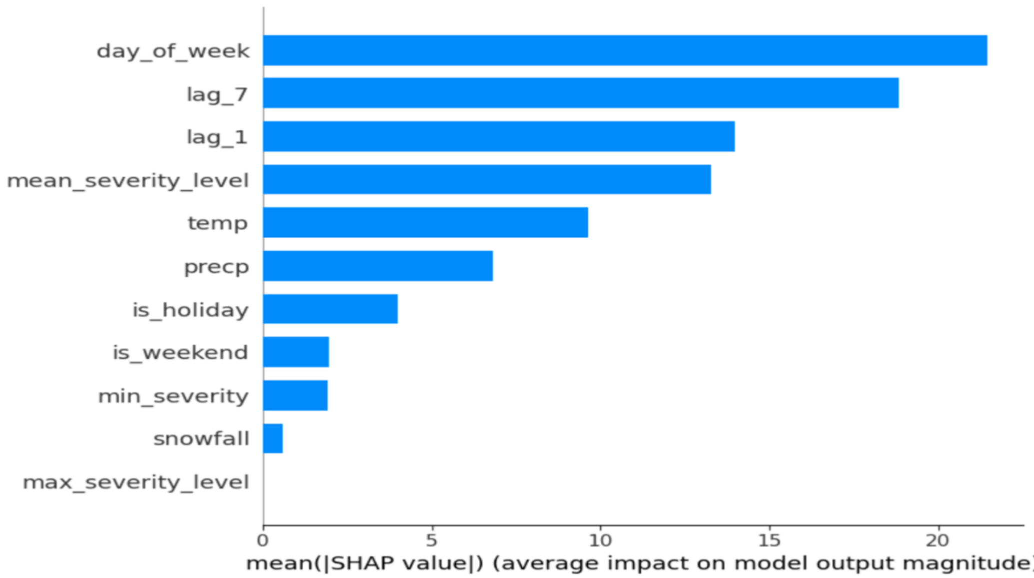
Borough-Level Takeaways:

Both Prophet and XGBoost models performed similarly, with XGBoost showing slightly lower RMSEs in most boroughs. Prophet performed comparably but had difficulty capturing borough-specific fluctuations due to its global seasonality assumptions.

- **Bronx:** *The model captures overall demand trends, with moderate accuracy, though short-term spikes occasionally deviate from predictions.*
- **Brooklyn:** *Predictions follow general patterns, but high variability in daily calls reduces model precision.*
- **Manhattan:** *The model struggles with highly volatile demand, resulting in lower predictive reliability.*
- **Queens:** *Forecasts capture broad trends, but sparse or irregular demand makes short-term predictions less accurate.*
- **Staten Island:** *Lower call volume allows the model to predict trends more consistently, though some fluctuations remain.*

Feature Importance Insights (SHAP Analysis):

Based on SHAP analysis, several features consistently influence the predictions of daily ambulance call volumes across all boroughs. **Day of the week**, **temperature**, **lag-1 call volume**, and **mean severity level** emerge as the most important predictors, indicating that both temporal patterns and weather conditions strongly affect demand.



Actual VS Predicted plot:

The actual vs predicted plots show that the XGBoost models generally follow observed demand trends across most boroughs, capturing day-to-day fluctuations and seasonal patterns. Bronx, Brooklyn, Queens, and Staten Island display good alignment between predictions and actuals, consistent with their moderate to low RMSE values. Manhattan, however, exhibits larger deviations, reflecting its higher RMSE of 72.94, indicating that the model struggles to predict the highly volatile and dense call patterns in this borough. Overall, the plots confirm that the borough-specific models capture broad trends while short-term spikes, particularly in high-variance areas like Manhattan, remain challenging.

Fig: *Actual vs Predicted Daily Calls by Borough.*

Conclusion

This study developed predictive models for NYC ambulance demand at both citywide and borough levels. Both the citywide and borough-wise modeling approaches provided valuable insights into NYC's ambulance demand.

- The citywide XGBoost model demonstrated strong overall performance (RMSE 47.84, R^2 0.98, MAPE 5.09%), significantly outperforming the linear regression baseline.
- Borough-level XGBoost models captured local trends and outperformed Prophet benchmarks, though performance varied due to differences in data volume and variability, with Manhattan showing the highest prediction error.

Feature interpretation using SHAP revealed that temperature, despite weak correlation in EDA, plays a meaningful role in certain boroughs — underscoring the importance of nonlinear feature interactions in predictive modeling.

Future improvements

Additional Features: Incorporate external factors such as public events, demographics, population density or traffic conditions to improve prediction of short-term demand spikes.

Advanced Temporal Models: Explore deep learning approaches like LSTM or Temporal Fusion Transformers to better capture complex sequential patterns, particularly in high-variance boroughs like Manhattan.

Data Granularity: Use finer-grained spatial or temporal data (e.g., neighborhood-level or hourly call data) to improve local accuracy.

Visualization & Dashboard: Develop an interactive dashboard for real-time monitoring of forecasts, residuals, and feature contributions to support operational decision-making.