ISYE 6740: Computational Data Analysis

7/27/2025

Group 252

Electric Load Forecasting and Congestion Modeling

By Tanvir Bakther, Gabriel Castaneda and Jayce Schwartz

Problem Statements

Power companies face a complex challenge when they need to generate and deliver electricity efficiently across their entire service area. To do this well, they strategically place power plants throughout the grid. This geographic planning matters because electricity travels through physical transmission lines and having all generation in one central location can create bottlenecks.

These bottlenecks happen when electricity demand exceeds what the transmission cables can handle. When this occurs, you get increased resistance, lines start heating up, and energy gets lost in the process. To address this, power companies use congestion pricing, essentially charging penalties when electricity flows exceed transmission limits. These penalties hit hardest during peak demand times when the grid is already stretched thin.

While engineers design grids with transmission capacity and power plant locations in mind, the system constantly changes. Demand shifts for all sorts of reasons time of day, weather, economic activity, and sometimes ideal locations for new plants just aren't available or cost too much. Usage patterns change daily, hourly, and seasonally, making it hard to predict exactly where problems will pop up.

Congestion is unpredictable and intermittent. It might happen in one area but not another or disappear completely without warning. This makes it nearly impossible to spot clear patterns using traditional analysis methods.

That's where machine learning comes in. By analyzing historical congestion data alongside publicly available grid information, like the data from the New York Independent System Operator (NYISO), we can build models that predict where congestion is most likely to happen. These insights can help power companies make smarter decisions about infrastructure investments, improve system reliability, and run more efficient electricity markets.

Data Source

The data source of this project is the New York Independent System Operator (NYISO) [1], which is the official website containing comprehensive electricity market data, real-time dispatch information, and historical energy statistics for New York State's power grid operations. Our dataset consists of 3 parts: (1) the Real-Time Dispatch Zonal Locational Based Marginal Pricing (LBMP) data for all NYC zones, including energy prices, congestion components, and transmission losses for each 5-minute dispatch interval; (2) the Real-Time Weighted Integrated Actual Load data representing electricity demand across Central NY, Genese NY, Long Island regions, capital regions, Millwood regions ,north regions ,west regions, Hudson regions and NYC measured in megawatts (MW) and used as the primary indicator of regional electricity consumption patterns; (3) the meteorological data from NYC weather stations, including temperature and humidity measurements that significantly influence electricity demand patterns. We collect all the statistics from 2021 to 2024, covering four complete years of electricity market operations, as our dataset. The LBMP data serves as our primary pricing variable, while actual load data provides demand metrics, and weather data offers explanatory variables for both demand and price forecasting models. Besides, we also utilize the most recent 2024 data for out-of-sample validation and forecasting performance evaluation of our XGBoost models for electricity demand, pricing, and congestion prediction across the different New York ISO regions.

Methodology

We developed a comprehensive machine learning framework employing two distinct XGBoost regression models to predict electricity demand and identify congestion issues across 9 of 11 operational regions of the New York Independent System Operator (NYISO). The congestion prediction model estimates transmission constraint levels using comprehensive market variables. The data acquisition process utilizes publicly available datasets from NYISO.com, which provides comprehensive electricity market information at no cost, enabling independent analysis of New York State's power grid operations. Model validations employ an 80-20 temporal split with performance evaluation based on R-squared coefficients, MAPE, and RMSE metrics, generating comparative visualization plots where original values serve as x-coordinates and predicted values as y-coordinates for direct assessment of model accuracy across all forecasting objectives.

We considered these factors in our models:

Time Period

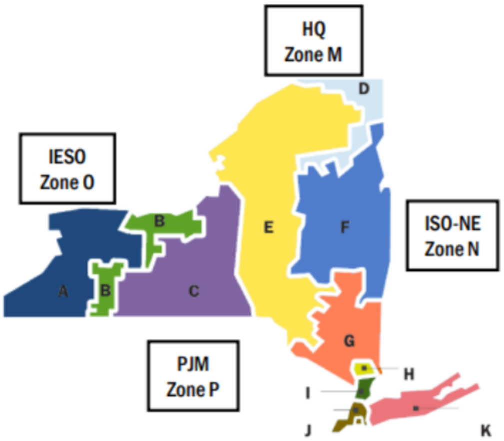
We chose to get data from 2021 to 2024 inclusive. We didn’t take data from 2020 as is was pandemic time, and changes in the people behavior would affect electricity consumption. The data was organized to have hourly data to be studied. Some datasets contained data every 5, 10 or 15 minutes, which were converted into hours. Most electricity markets and data are on hourly basis

Location

The first way we divided our algorithms was based on location, that is, on the areas identified by NYISO and where it provides electricity. We chose 9 locations to take data from so that each team member had 3 datasets each. The zones chosen were West, Genesse, Central, North, Capital, Hudson Valley, Millwood, NYC and Long Island.

NYCA Load Zones

A- West
B- Genesee
C- Central
D- North
E- Mohawk Valley
F- Capital
G- Hudson Valley
H- Millwood
I- Dunwoodie
J- NYC
K- Long Island



Electricity demand depends on weather, temperature and relative humidity, as it influences the use of either air conditioning or heating and temperature depends on location, Some areas of the state of New York can be quite hot during the summer while others will be in the 60s. In the winter, some areas will be full of snow and quite cold, while others will still be above freezing. Also, the terrain of New York State is not flat, there are mountains which make weather temperatures to differ significantly from one site to another. Lake Ontario serves as snow generator over the winter while the Atlantic Ocean influences its coastal region.

Power plants and population are unevenly distributed all over the state of New York, some areas have a lot of population, while others have power plants. Also, population growth changes with time. Some urban centers increase in size while others decrease. As for power plants, many factors, including solar radiation, snow precipitation, real state availability and pricing are considered to build new locations, so sometimes electricity will have to travel for longer distances to where the demand is, which will mean that the existing infrastructure needs to be updated from time to time to make sure the power lines will be able to effectively transmit the energy.

People activity

Activity of people varies on an hourly basis, daily basis, holidays and months. During some months, people go on vacations, during holidays many people take the day off, people typically work from 8 to 5pm and have a defined set of activities on each day of the week. Saturdays and Sunday's people work less and enjoy family activities.

Power plants

Power plants may be on or off depending on seasons, periods or months. Hydraulic power plants may not be used during rainy seasons but then used when there's need to deplete their loads. Solar plants will not be used during rainy season or seasons where overcast is common. Eolic plants will not be useful during seasons when the wind is minimal, gas turbines will be used less when all the other resources can be used, as they are a more expensive source of electricity as they use natural gas as fuel. These power plants will influence the capacity of the power lines, pricing and congestion.

Machine Learning Algorithm

The major consumer of electricity is the air conditioners people use. Their electricity demand depends on temperature and relative humidity. The electricity load of air conditioners depends on the efficacy of moving energy out of the rooms to be cooled down, the air enthalpy, which is quite nonlinear. The capacity to handle nonlinear variables is a major requirement on the algorithm to be used. Another major requirement is that it needs to be very detail oriented. The energy requirements used by people do follow a certain shape each day of the week, but they also instantly change if there's a weekend or a holiday. The algorithm needs to work on 365 days without much tuning and be able to differentiate between a normal day of the week to a holiday to a hot day whether if it happens during the weekday or weekend. The efficacy of gradient boosting algorithm to go over little details because of its boosting capability (takes the errors of an initial tree and goes again at it to try to model it right) makes this algorithm among the best available for this type of use. Also, gradient boosting can handle very nonlinear behaviors, such as air thermodynamics and electric load changes of air conditioners.

Evaluation

For the congestion and demand models, we applied a XGBoost model across the different NYISO zones. Each model used the same set of variables and relied on a single temperature and relative humidity (RH) reading from one weather station within each zone. What differentiates the results are the characteristics of each zone, including geographic size, temperature variation, population density (urban versus rural), power infrastructure, and the layout of the territory. Some zones are compact, while others have irregular shapes and scattered population centers.

We will first review the regression models to predict the demand of electricity on 9 NYISO zones, following by the review of classification models to predict congestion within each zone.

Below are the results for each territory for each model, followed by conclusions and insights based on the data.

Forecast Demand Models Results

Demand Model Results for the Different NYISO Zones

NYC

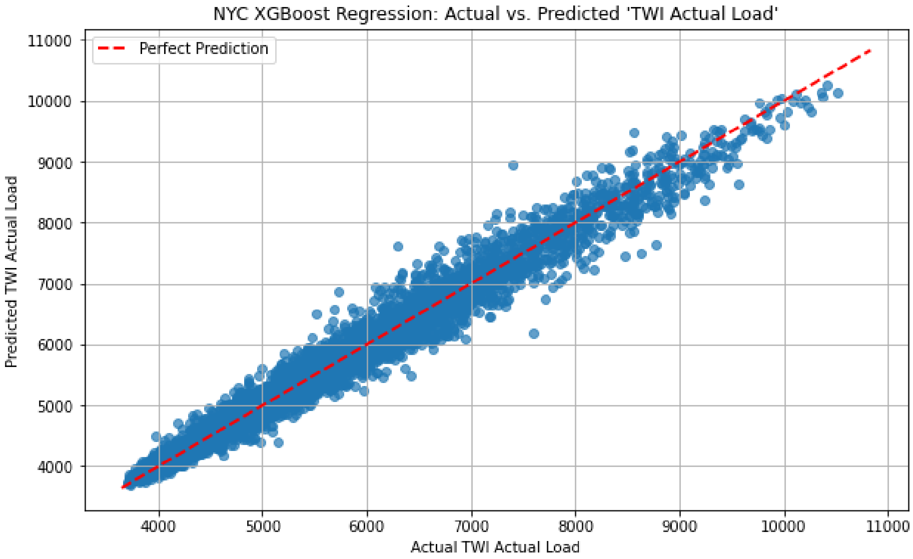


Figure 1: Actual vs Predicted Demand Comparison Graph for NYC Territory

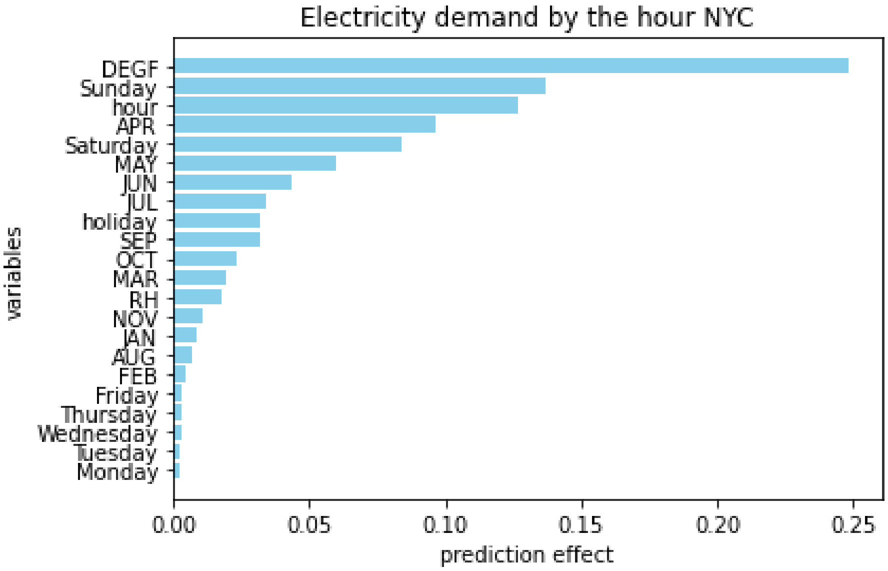


Figure 2: Top Predictors of XGBoost for NYC Territory

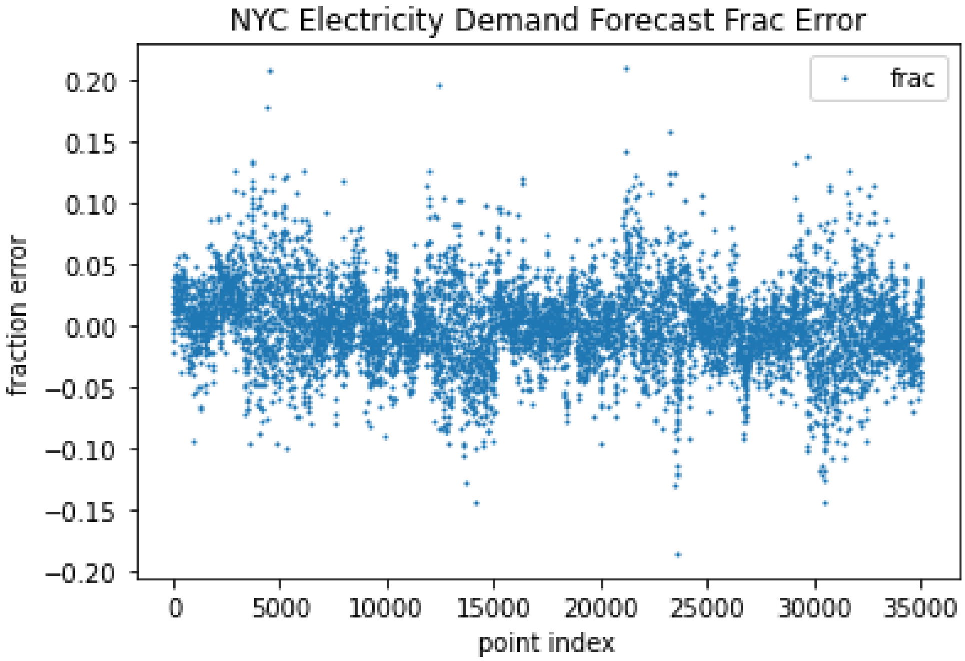


Figure 3: Residuals vs Point Index Graph (4 years) for NYC Territory

Mean Absolute Percentage Error: 0.02

R-squared: 0.97

The NYC demand model demonstrates exceptional performance with the lowest MAPE of 2% and highest R-squared of 0.97. The scatter plot shows tight clustering around the 45-degree line, indicating highly accurate predictions. Compact urban geography and consistent temperature patterns contribute to the model's superior predictive capability.

The point index in the frac/residual graphs is related to each data point across the 4 years of data. For instance, the total number of data points are 365 days x 24 hours x 4 years=35,040 data points

Hudson

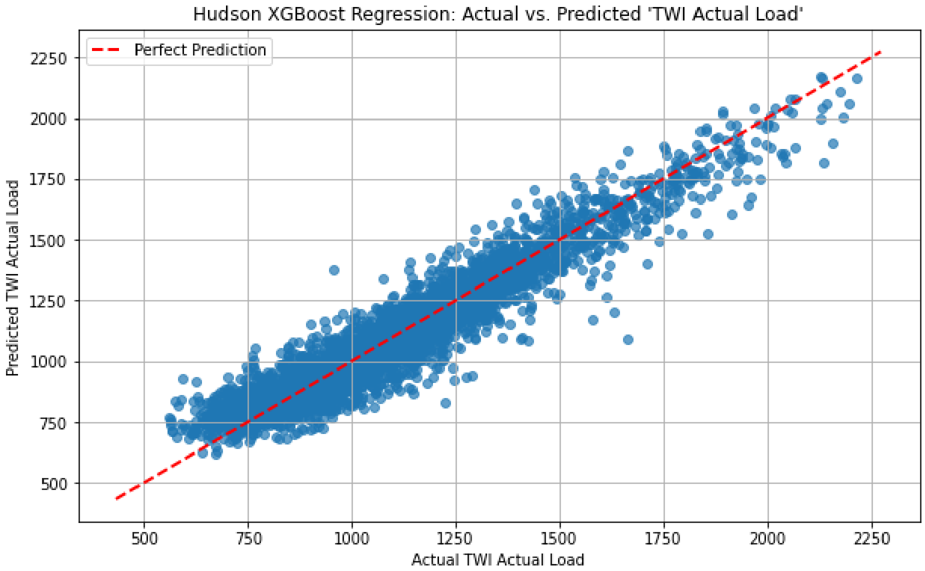


Figure 4: Actual vs Predicted Demand Comparison Graph for Hudson Territory

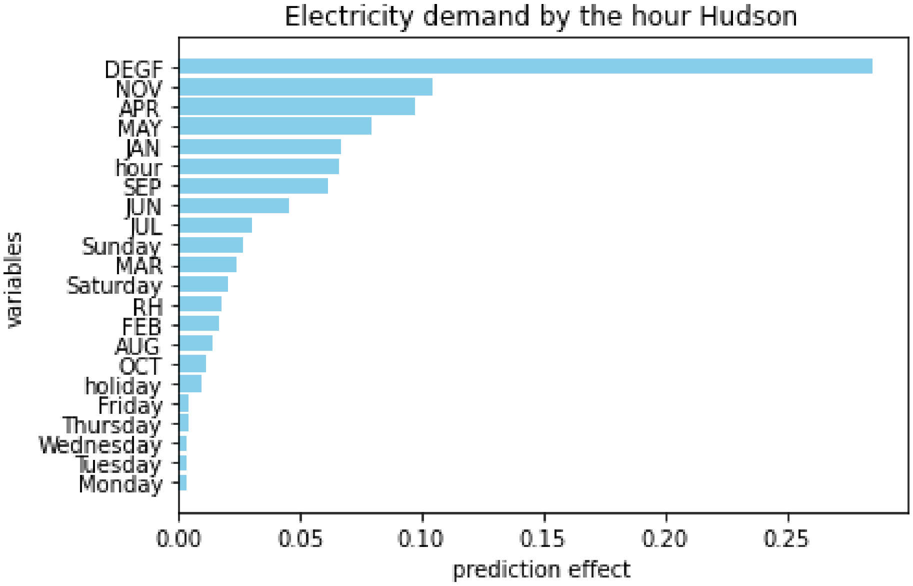


Figure 5: Top Predictors of XGBoost for Hudson Territory

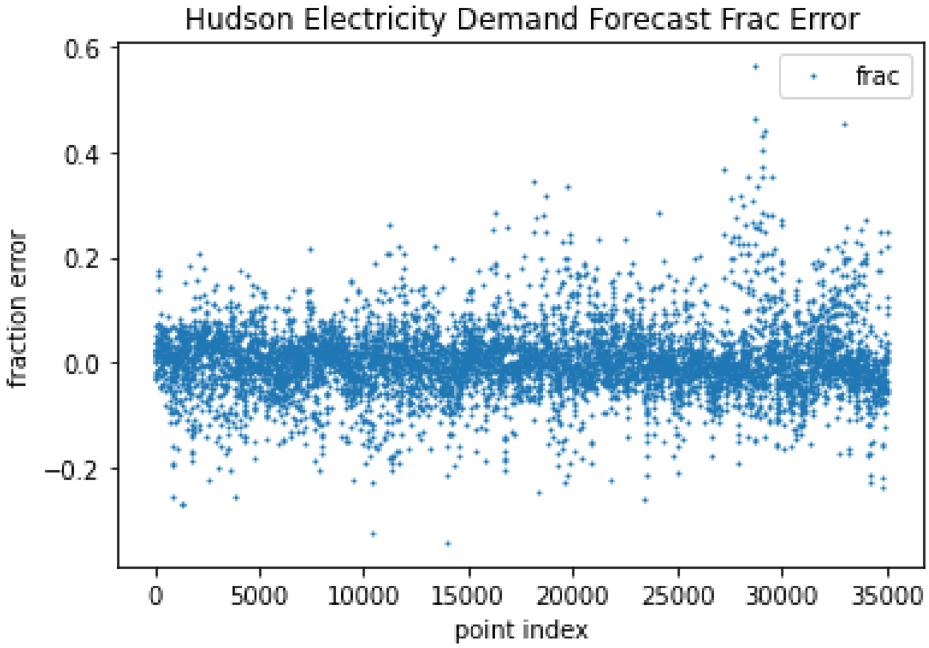


Figure 6: Residuals vs Point Index Graph (4 years) for Hudson Territory

Mean Absolute Percentage Error: 0.05

R-squared: 0.92

The Hudson model shows good performance with a MAPE of 5% and R-squared of 0.92. While slightly less accurate than NYC, the model maintains strong predictive power. The residuals plot reveals some seasonal variance, likely due to the region's mixed urban-rural characteristics and varied topography.

West

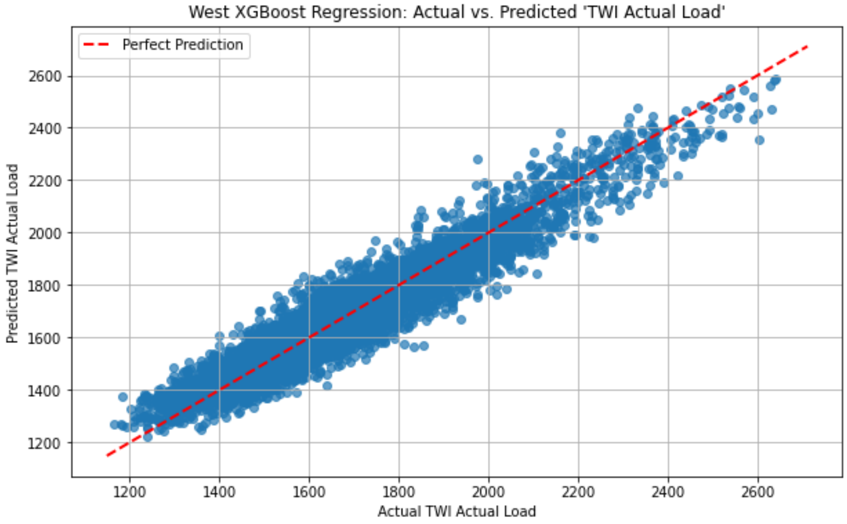


Figure 7: Actual vs Predicted Demand Comparison Graph for West Territory

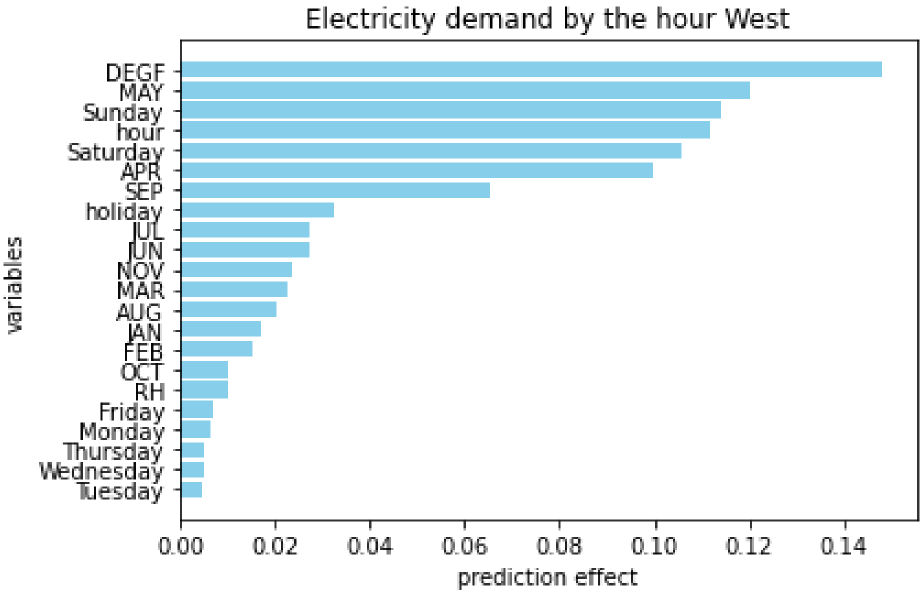


Figure 8: Top Predictors of XGBoost for West Territory

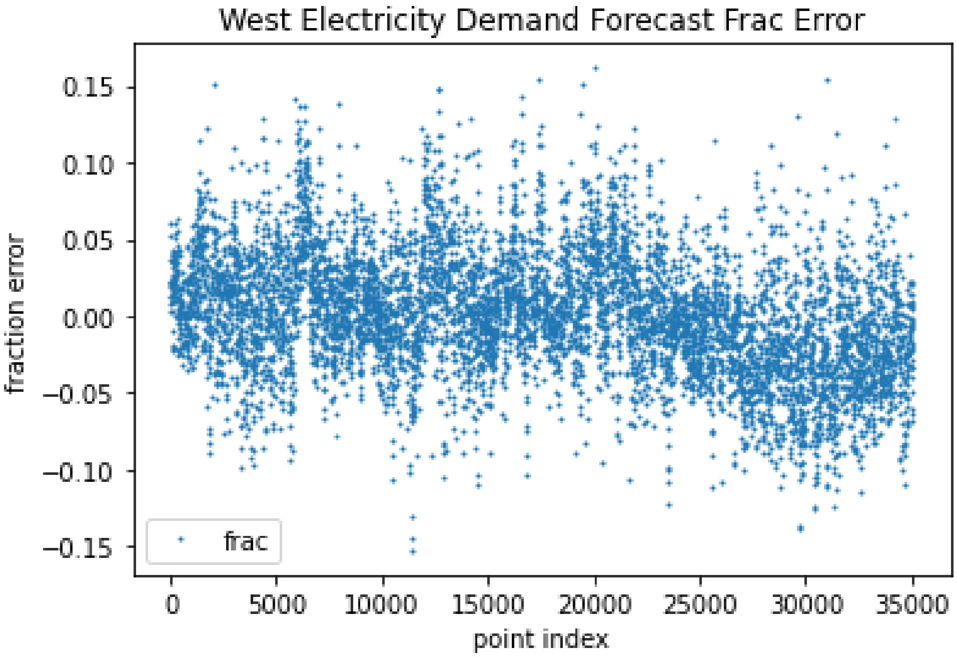


Figure 9: Residuals vs Point Index Graph (4 years) for West Territory

Mean Absolute Percentage Error: 0.03

R-squared: 0.91

The West region has a MAPE of 3% and an R-squared value of 91%, indicating that the model fits the Millwood territory very well. The top predictor was temperature. However, the residual graph shows some drift, with an increase in errors appearing around the fourth-year mark. Evidently, something during year four happened (365 days x 24 hr x 3=26,280)

Capital

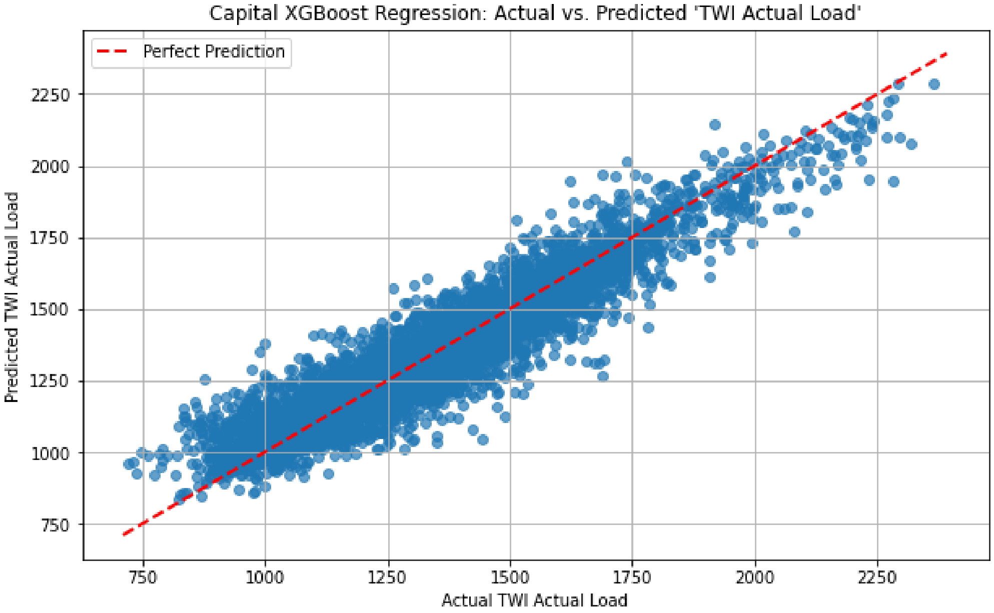


Figure 10: Actual vs Predicted Demand Comparison Graph for Capital Territory

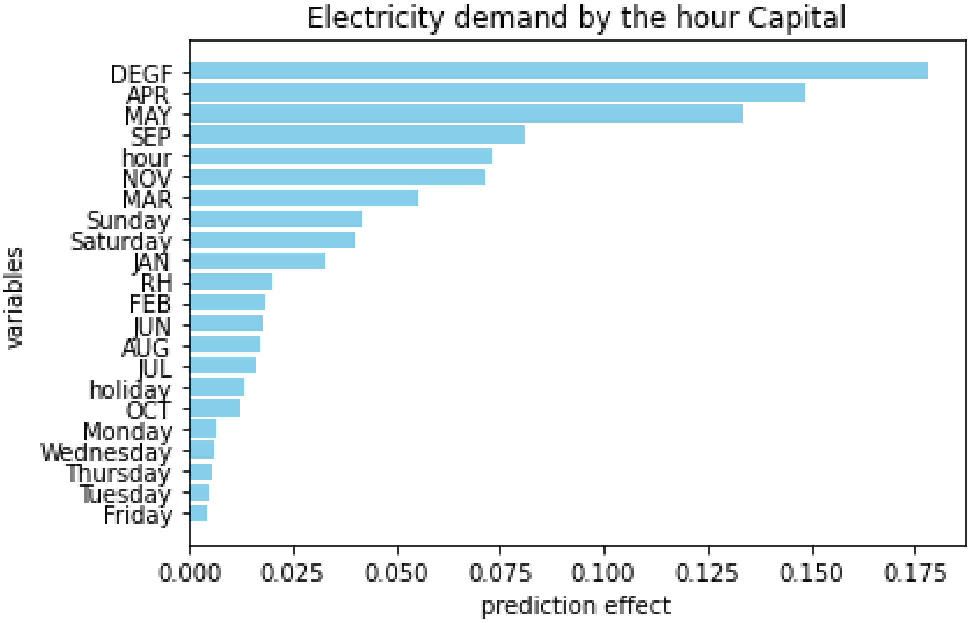


Figure 11: Top Predictors of XGBoost for Capital Territory

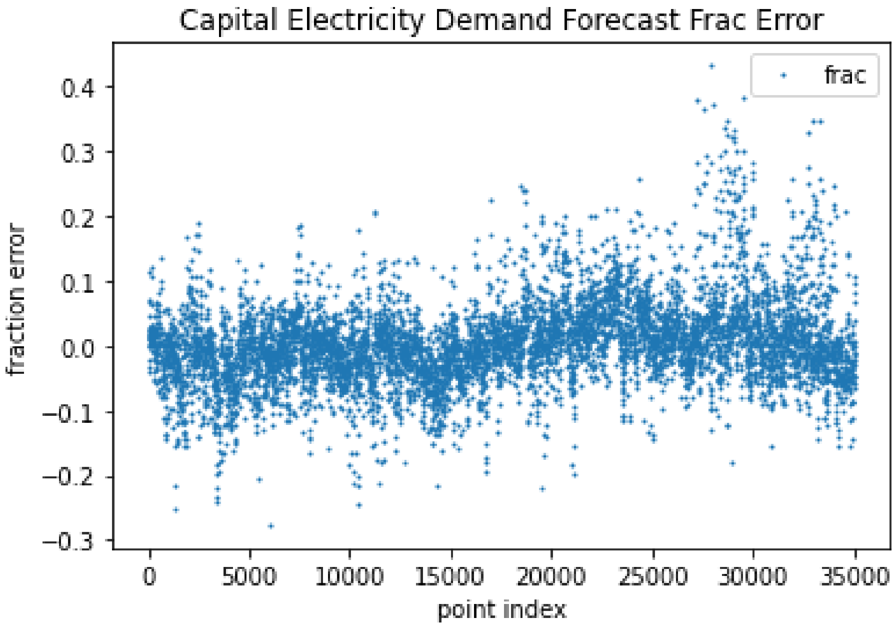


Figure 12: Residuals vs Point Index Graph (4 years) for Capital Territory

Mean Absolute Percentage Error: 0.05

R-squared: 0.88

The Capital region achieves a MAPE of 5% and R-squared of 0.88, indicating solid predictive performance. The model captures the general demand patterns effectively, though the residuals suggest some systematic drift over time, possibly reflecting changing demographic or economic conditions in the region.

Millwood

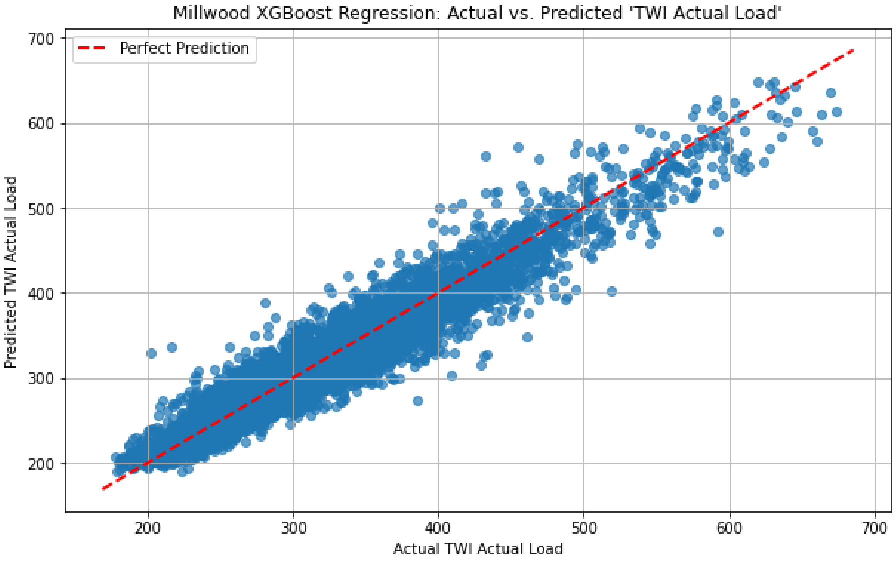


Figure 13: Actual vs Predicted Demand Comparison Graph for Millwood Territory

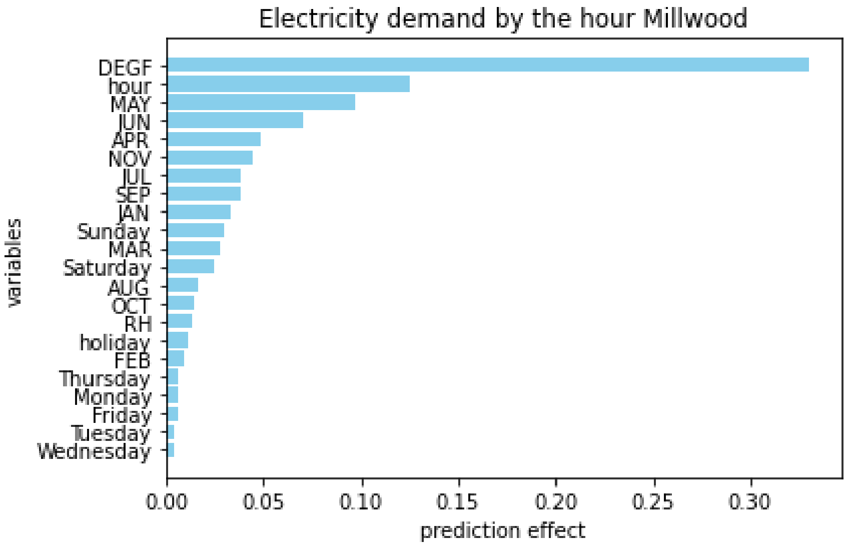


Figure 14: Top Predictors of XGBoost for Millwood Territory

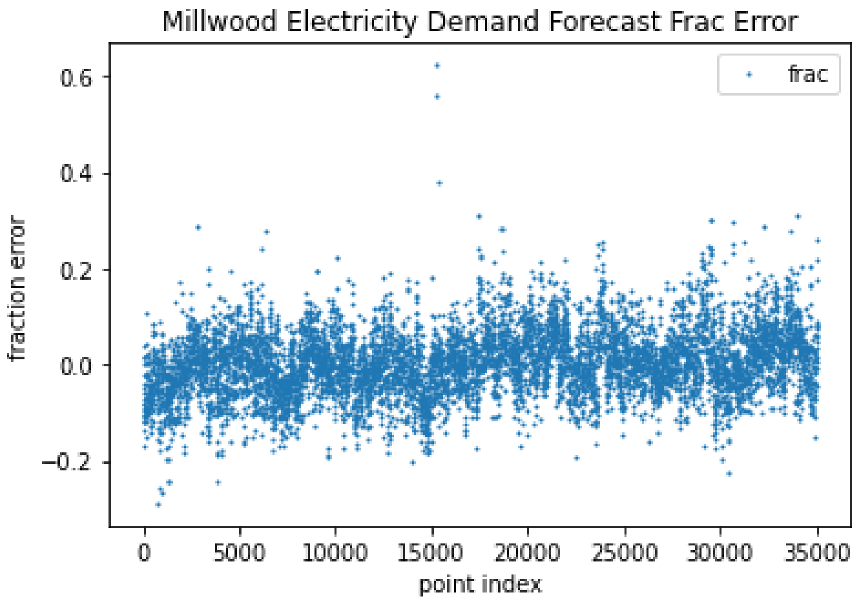


Figure 15: Residuals vs Point Index Graph (4 years) for Millwood Territory

Mean Absolute Percentage Error: 0.05

R-squared: 0.91

The Millwood model delivers reliable results with a MAPE of 5% and R-squared of 0.91. The predictions align well with actual demand, showing the model's ability to capture the load patterns in this smaller geographic area with consistent accuracy across seasons.

North

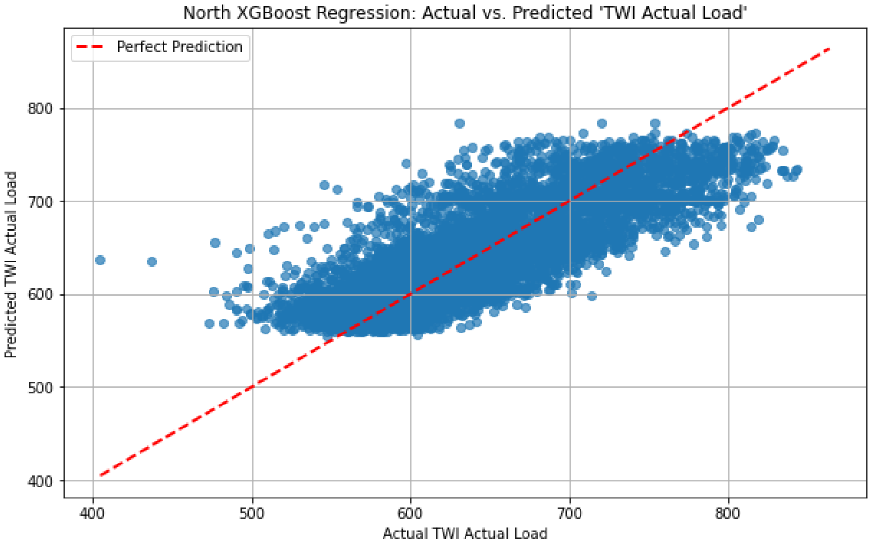


Figure 16: Actual vs Predicted Demand Comparison Graph for North Territory

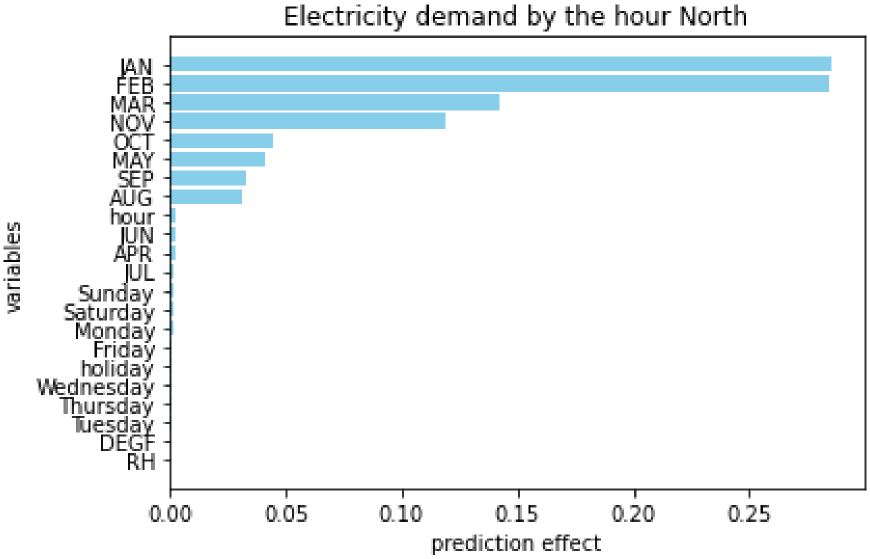


Figure 17: Top Predictors of XGBoost for North Territory

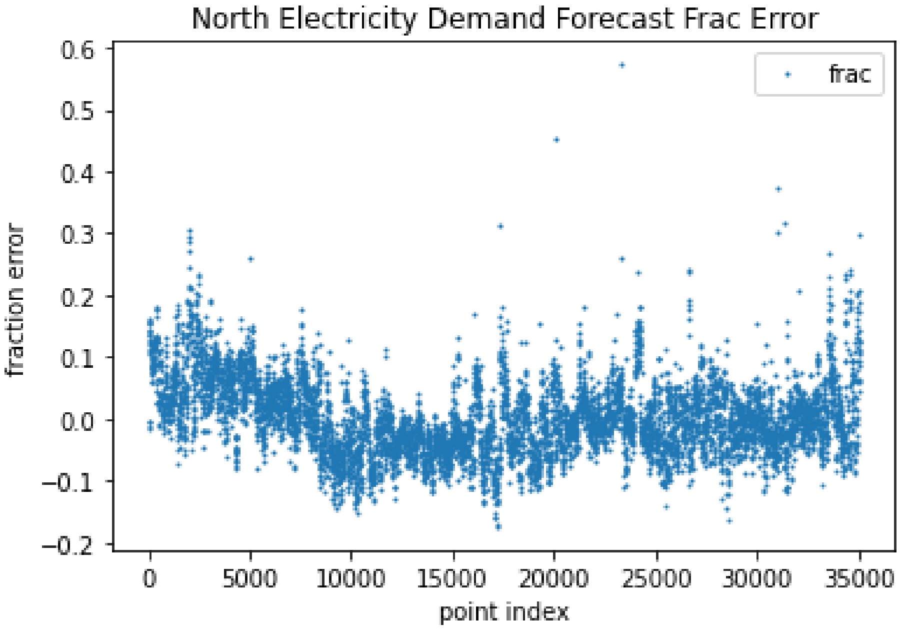


Figure 18: Residuals vs Point Index Graph (4 years) for North Territory

Mean Absolute Percentage Error: 0.05

R-squared: 0.63

The North zone shows the weakest performance among all regions with a MAPE of 5% but significantly lower R-squared of 0.63. The scatter plot reveals greater dispersion, and residuals show more pronounced variance. This is likely due to the region's irregular shape, sparse population distribution, and significant temperature variations across the territory. The north zone does have an area that is bordering Canada, while another area that ventures south within the state, extremely irregular, this zone has mountains as well, which create a lot of temperature differences within its territory.

Central

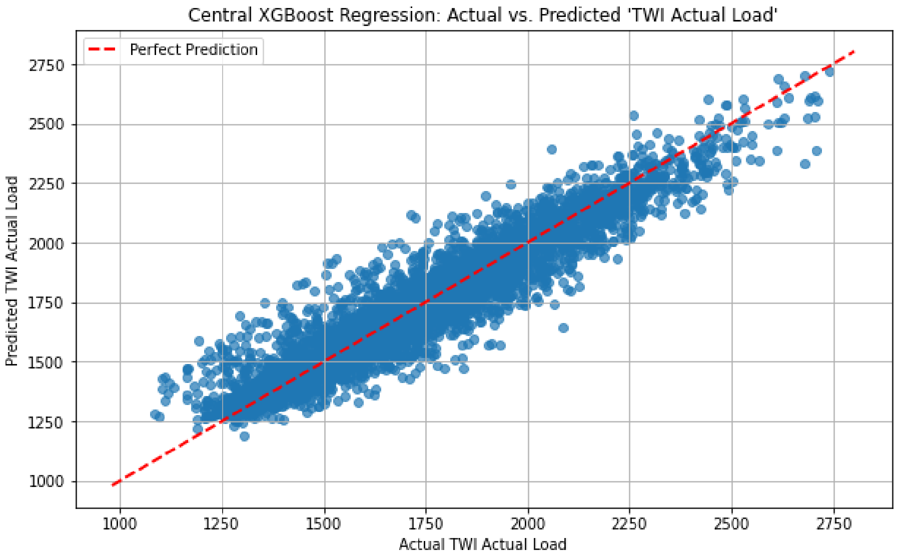


Figure 19: Actual vs Predicted Demand Comparison Graph for Central Territory

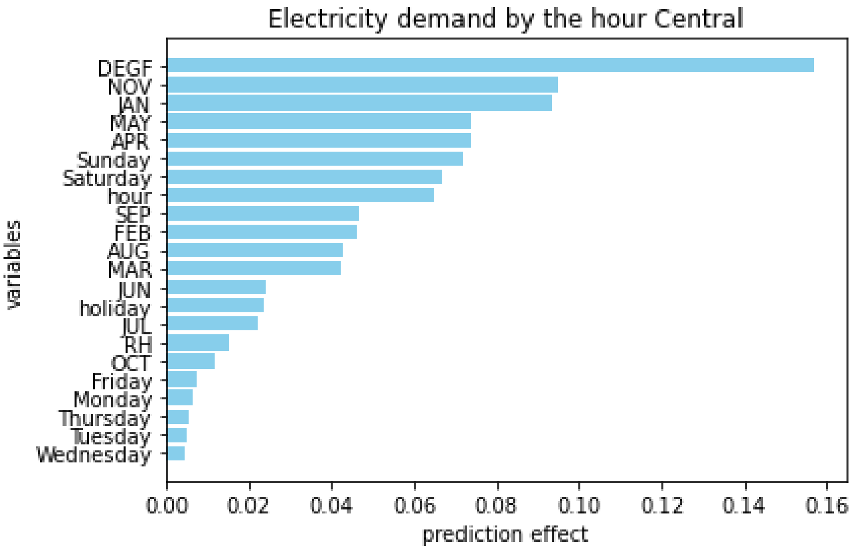


Figure 20: Top Predictors of XGBoost for Central Territory

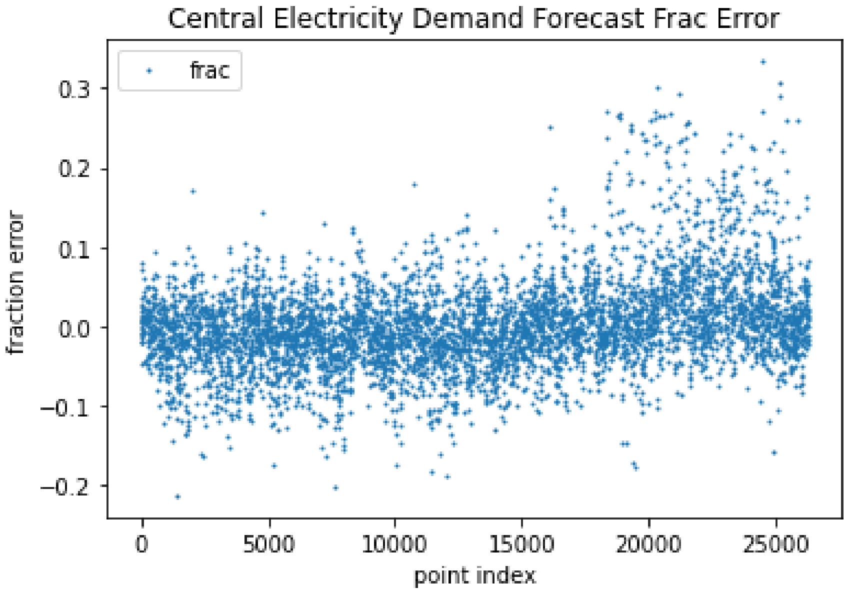


Figure 21: Residuals vs Point Index Graph (4 years) for Central Territory

Mean Absolute Percentage Error: 0.04

R-squared: 0.89

Central zone demonstrates good performance with a MAPE of 4% and R-squared of 0.89. The model effectively captures demand patterns, though residuals indicate some systematic bias over time, suggesting potential changes in regional electricity consumption patterns during the study period.

Genesse

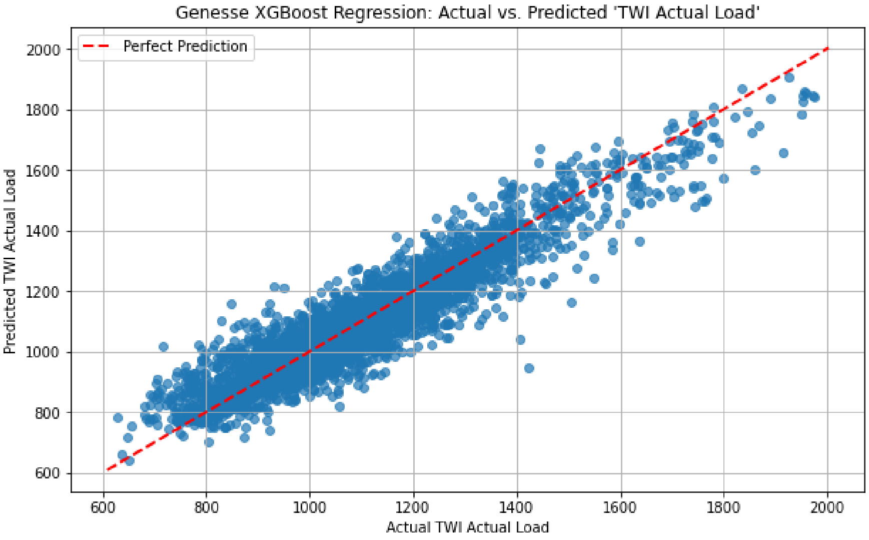


Figure 22: Actual vs Predicted Demand Comparison Graph for Genesse Territory

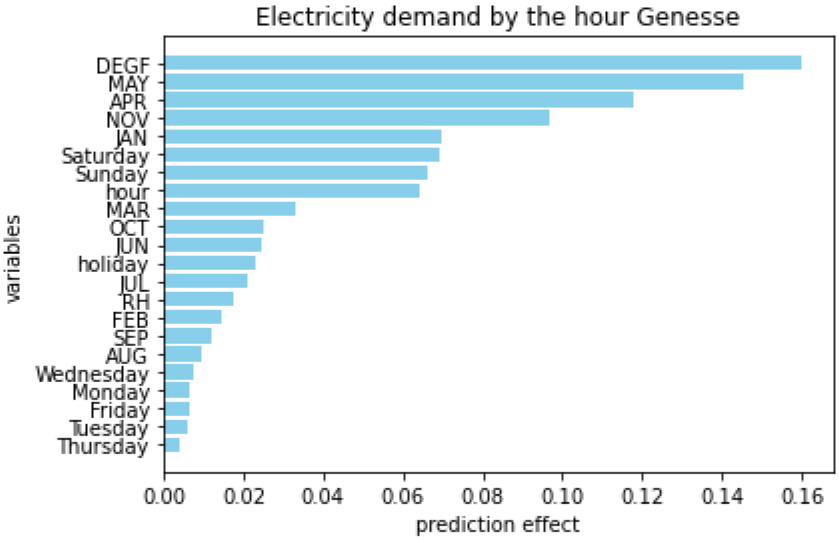


Figure 23: Top Predictors of XGBoost for Genesse Territory

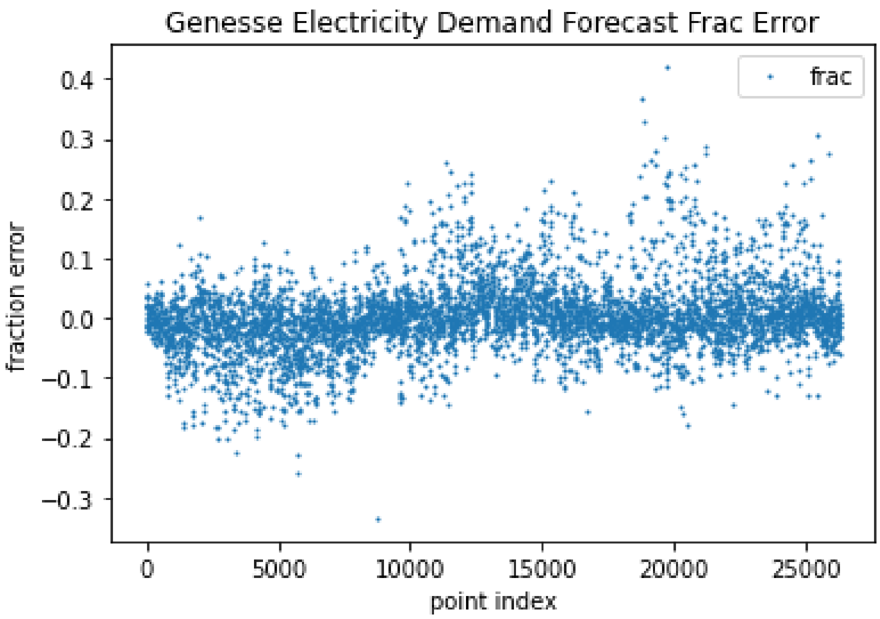


Figure 24: Residuals vs Point Index Graph (4 years) for Genesse Territory

Mean Absolute Percentage Error: 0.04

R-squared: 0.89

The Genesse model achieves solid results with a MAPE of 4% and R-squared of 0.89. The predictions show good agreement with actual demand, and the residuals remain relatively stable across the four-year period, indicating consistent model performance.

Long Island

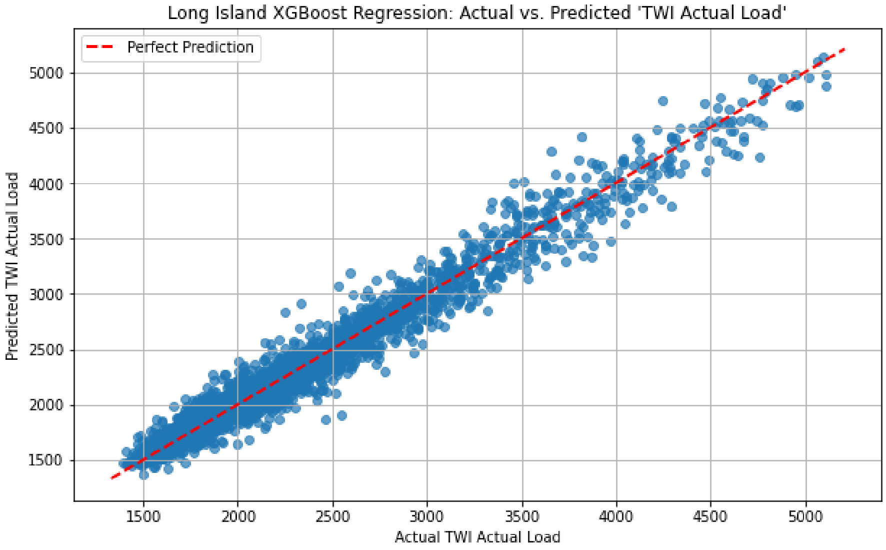


Figure 25: Actual vs Predicted Demand Comparison Graph for Long Island Territory

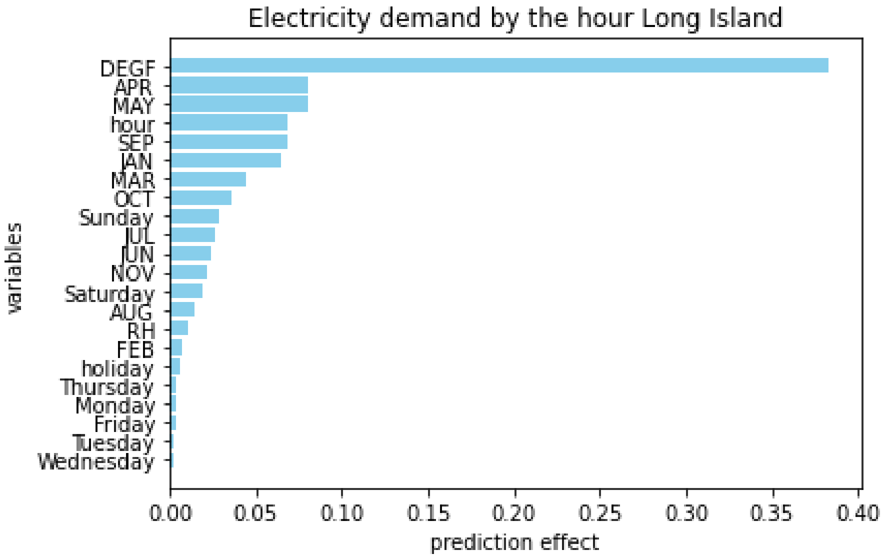


Figure 26: Top Predictors of XGBoost for Long Island Territory

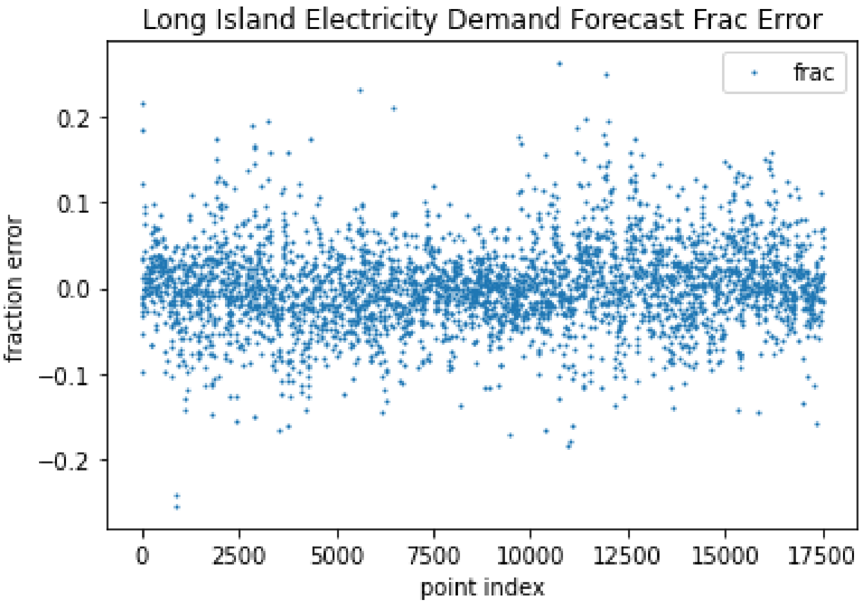


Figure 27: Residuals vs Point Index Graph (4 years) for Long Island Territory

Mean Absolute Percentage Error: 0.04

R-squared: 0.97

Long Island delivers excellent performance with a MAPE of 4% and exceptional R-squared of 0.97, matching NYC's accuracy. The island's geographic isolation and relatively homogeneous climate conditions contribute to highly predictable demand patterns, resulting in tight clustering around the prediction line.

Congestion Model Results for the Different NYISO Zones

To predict congestion we are using a classification model yes/no instead of a regression model which would predict a congestion price because the data required to calculate price was not available for free or was not of the free domain. Congestion is very dependent to the location of the power plants and their proximity to the demand of electricity.

NYC

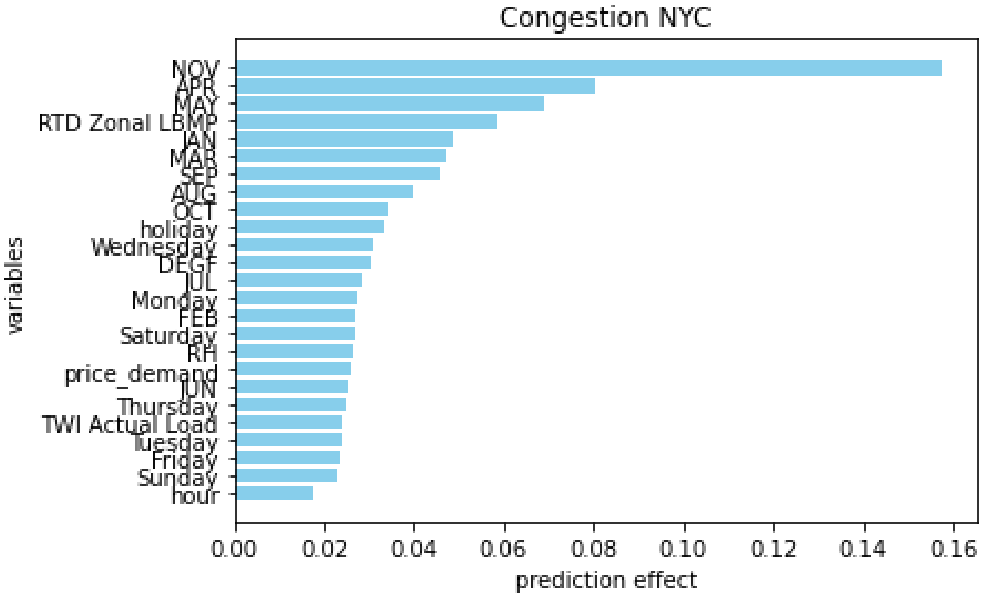


Figure 28: Top Predictors of XGBoost for NYC Territory

NYC	Actually Positive	Actually Negative
Predicted Positive	2426	926
Predicted Negative	1075	2566

Figure 29: Confusion Matrix for NYC Territory

Accuracy Score: 0.71

We can see that for the NYC territory, the model achieved an accuracy score of 0.71, indicating that it was reasonably effective at predicting congestion. Notably, November had the largest predictive effect among all the variables. As shown in the confusion matrix, the model correctly predicted a large number of true positives and true negatives. However, it did not classify all instances correctly, which explains why the overall accuracy is 71%. November likely stands out because it marks the beginning of the heating season and brings sudden increases in electricity demand, which can stress the grid. The combination of colder temperatures, shorter daylight hours, and pre-winter grid conditions makes congestion more likely during this month.

Hudson

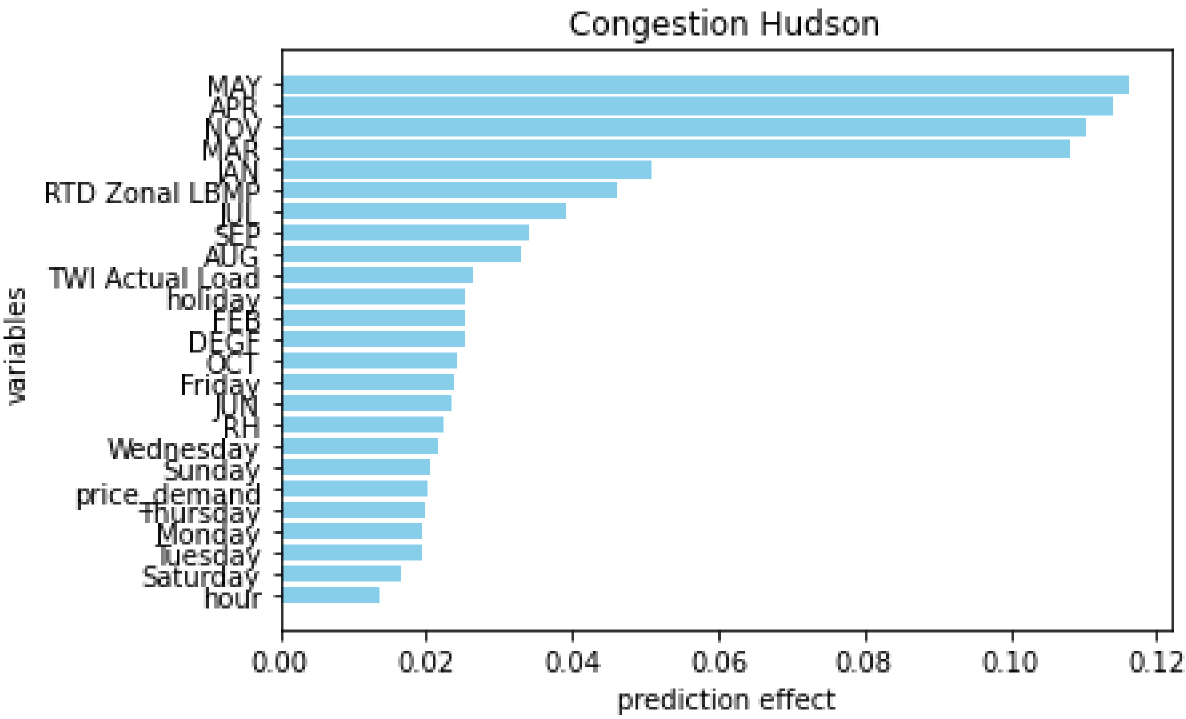


Figure 30: Top Predictors of XGBoost for Hudson Territory

Hudson	Actually Positive	Actually Negative
Predicted Positive	3237	751
Predicted Negative	1141	1884

Figure 31: Confusion Matrix for Hudson

Accuracy Score: 0.73

For the Hudson territory, the model's accuracy remained largely consistent, with a 73 percent accuracy score. However, the model predicted more true positives than true negatives compared to the NYC data. Unlike NYC, where November was the dominant predictor, the Hudson model did not have a single clear variable standing out. Instead, the top predictor was May, followed closely by April and then November. These months likely stand out because they represent transitional periods when temperatures shift, such as the change from spring to summer and from fall to winter. These shifts can cause fluctuations in electricity demand that contribute to congestion.

West

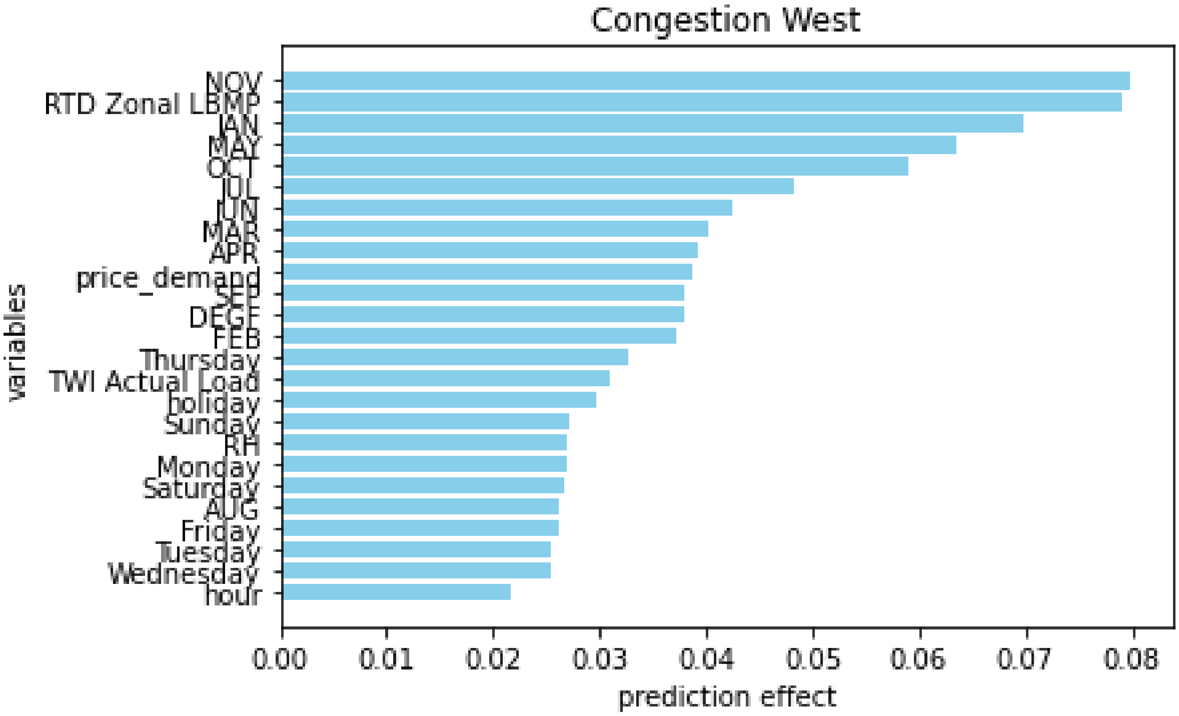


Figure 32: Top Predictors of XGBoost for West Territory

West	**Actually Positive**	**Actually Negative**
Predicted Positive	3941	465
Predicted Negative	1164	1443

Figure 33: Confusion Matrix for West Territory

Accuracy Score: 0.77

The congestion model performed better for the West territory, with the accuracy score increasing to 77 percent. When reviewing the confusion matrix, the model performed very well at identifying true positive values but had more difficulty predicting true negatives. The top predictors for this zone were November and LBMP (Locational Based Marginal Pricing).

November and demand stand out because this month marks the transition from fall to winter, which can bring temperature drops and increased electricity usage. A rise in demand during this period can overload the grid, leading to higher congestion.

Capital

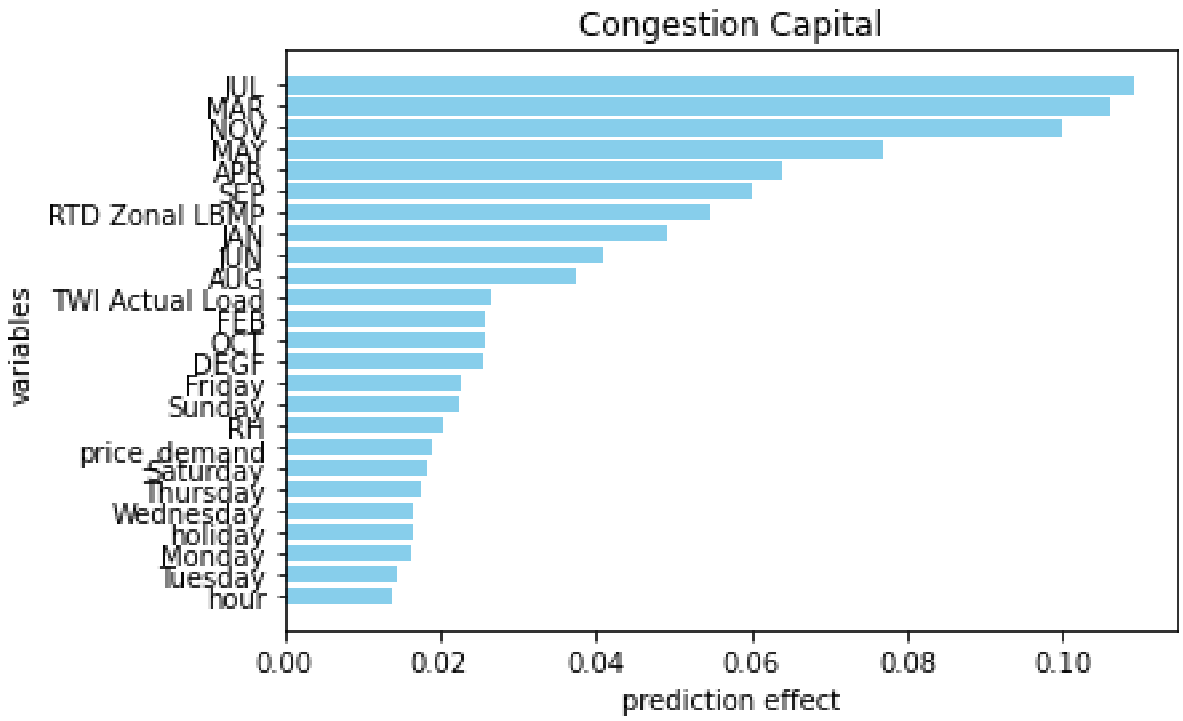


Figure 34: Top Predictors of XGBoost for Capital Territory

Capital	**Actually Positive**	**Actually Negative**
Predicted Positive	3601	596
Predicted Negative	1042	1774

Figure 35: Confusion Matrix for Capital Territory

Accuracy Score: 0.77

For Capital Territory, we can see that this model was also able to predict true positives much better than true negatives. The accuracy remained the same as in the West Territory, at 77%. The top predictors for this model were July, March, and November.

Millwood

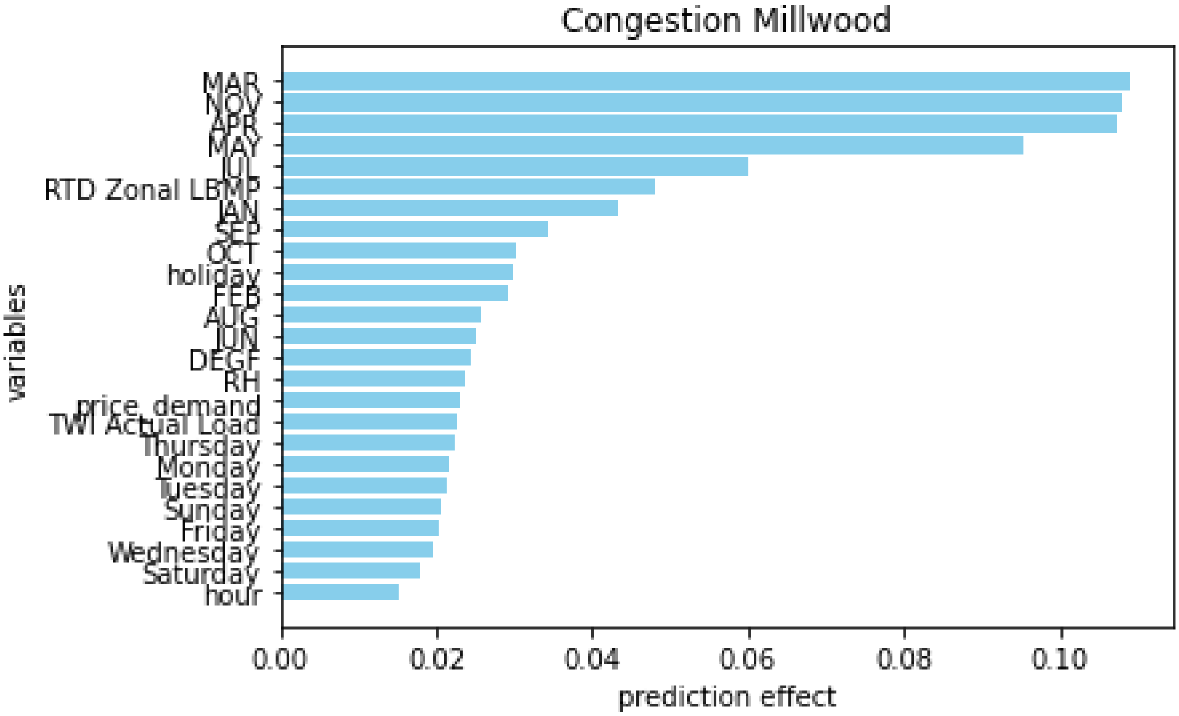


Figure 36: Top Predictors of XGBoost for Millwood Territory

Millwood	**Actually Positive**	**Actually Negative**
Predicted Positive	3052	770
Predicted Negative	1182	2009

Figure 37: Confusion Matrix for Millwood Territory

Accuracy Score Millwood 0.72

The congestion model for Millwood showed a slight improvement in identifying true negatives. However, the overall accuracy decreased to 72%, compared to the higher-performing models which were at 77%. The top predictors were November and March.

North

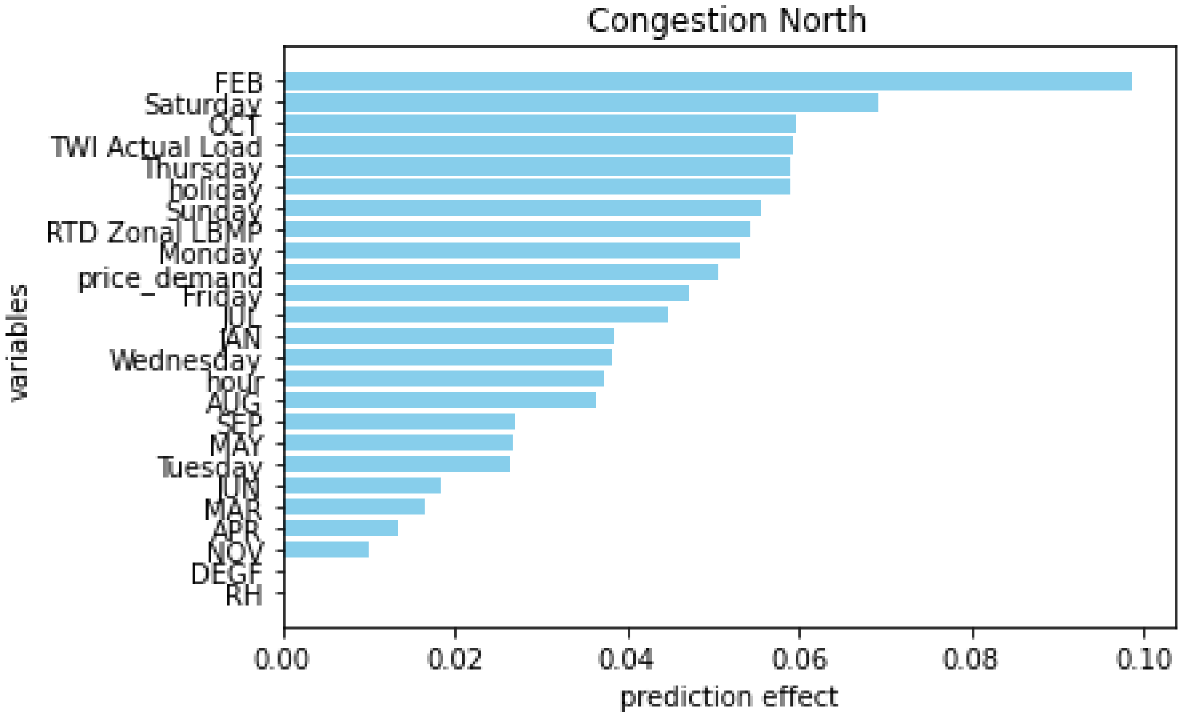


Figure 38: Top Predictors of XGBoost for North Territory

North	Actually Positive	Actually Negative
Predicted Positive	6843	13
Predicted Negative	153	4

Figure 39: Confusion Matrix for North Territory

Accuracy Score 0.98

The North Territory congestion model had the highest accuracy score at 98%. It also performed very well in predicting true positive values. However, the dataset contained very few actual negative values, which may have contributed to the high accuracy score. The top predictor for this model was the month of February, which occurs in winter and could lead to grid system overloads due to increased heater usage.

Central

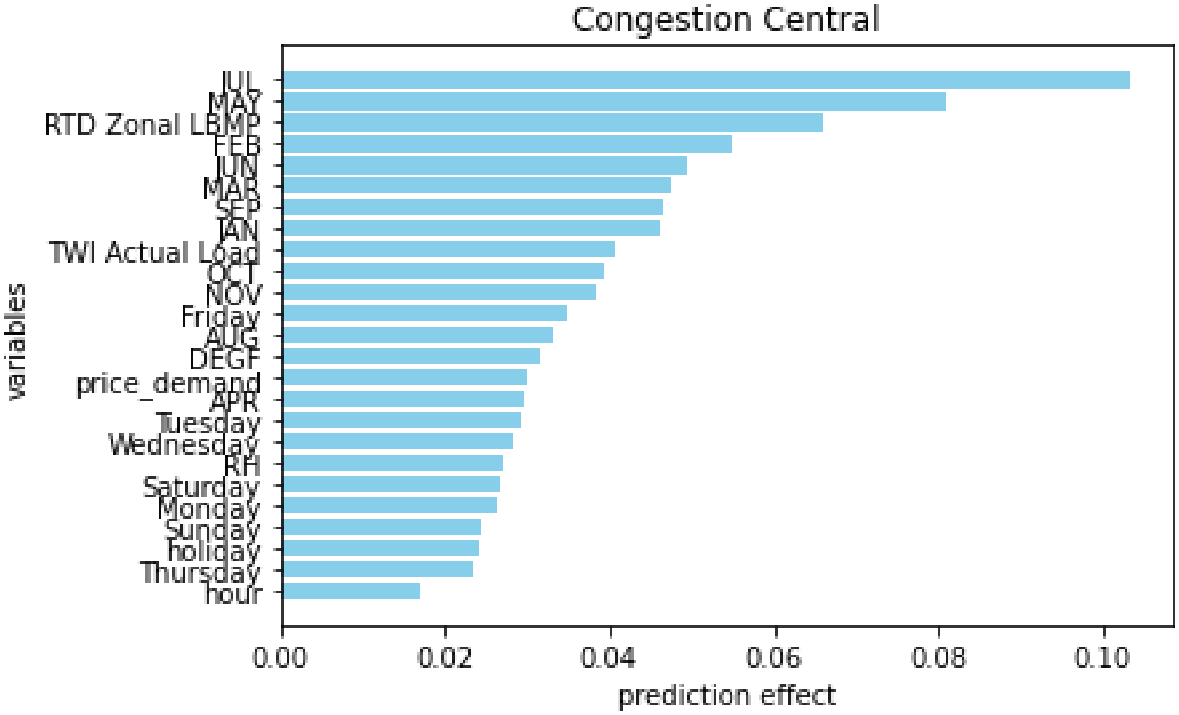


Figure 40: Top Predictors of XGBoost for Central Territory

Central	**Actually Positive**	**Actually Negative**
Predicted Positive	2352	491
Predicted Negative	861	1557

Figure 41: Confusion Matrix for Central Territory

Accuracy Score: 0.74

For the Central congestion model, we had an accuracy score of 74%. Similar to the other models except for the North model, we can see that it performed better at identifying true positives than true negatives. The top predictor for this model was July, which makes sense since it's the middle of summer, when heavy AC usage can strain the grid system.

Genesse

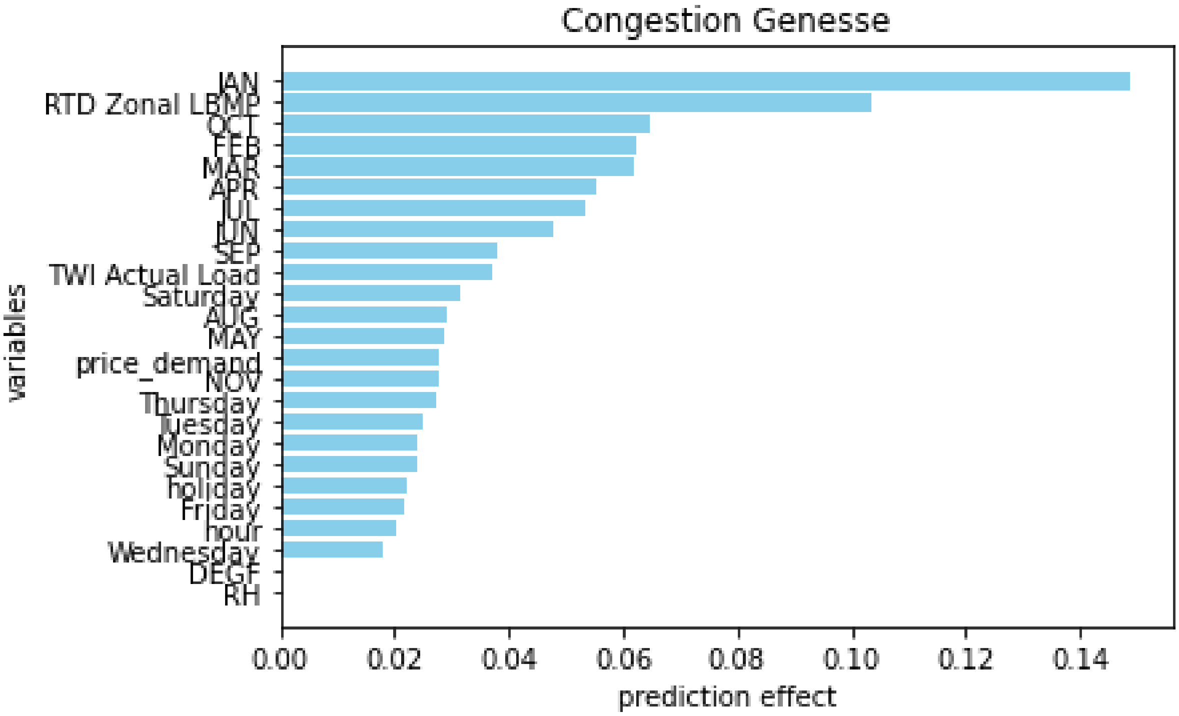


Figure 42: Top Predictors of XGBoost for Genesse Territory

Genesse	**Actually Positive**	**Actually Negative**
Predicted Positive	3571	224
Predicted Negative	704	762

Figure 43: Confusion Matrix for Genesse Territory

Accuracy 0.82

For the Genesee Territory, the congestion model performed very well, with an accuracy score of 82%. The model was able to predict true positives effectively but still struggled with true negatives. There were also fewer true negatives in this dataset compared to other territories, which could explain the higher accuracy score. The main predictor for this model was the month of January. This makes sense, as January is cold in New York and increased heater usage can lead to more congestion.

Long Island

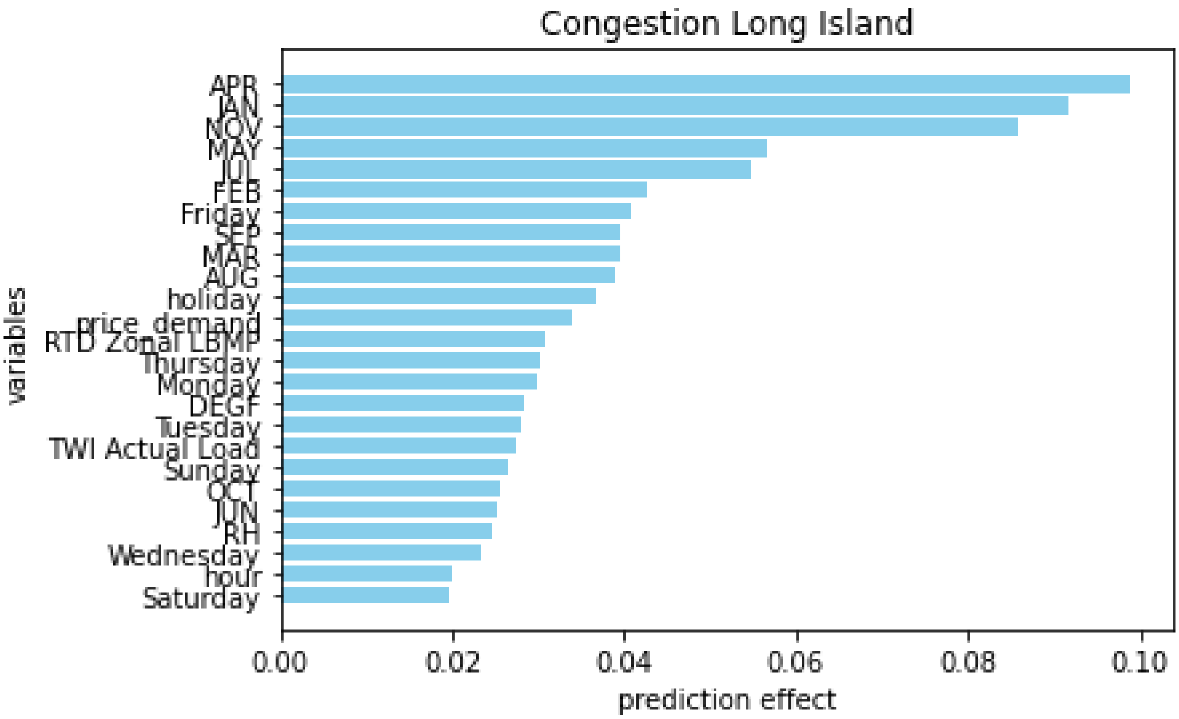


Figure 44: Top Predictors of XGBoost for Long Island Territory

Long Island	Actually Positive	Actually Negative
Predicted Positive	197	504
Predicted Negative	118	2685

Figure 45: Confusion Matrix for Long Island Territory

Accuracy Score: 0.82

For the Long Island congestion model, the accuracy score was 82%. Interestingly, the model was able to identify more true negative values than true positive values, the opposite of most models shown previously. The top predictors for this model were April and January.

Final Results

After running all models, we found that most of the demand models performed well. The XY plots show that the predicted load aligns closely with the actual load, falling along the 45 degree reference line. This alignment is also supported by the low Mean Absolute Percentage Error (MAPE) values, ranging from approximately 2 percent for NYC to 5 percent for Hudson. In these plots, the narrower the blue band, the better the model's performance. While some models, such as the one for NYC, have a tight band, others like the North model exhibit a wider, more scattered distribution. None of the models shows bumps or big irregularities, which may mean they were lacking variables in the model.

The North region presents a particular challenge. Unlike a single city, it is a geographically broad and irregularly shaped area that likely experiences more dramatic temperature variation across its territory, which may contribute to the model's higher error.

Across most demand models, the most important predictors were temperature, hour of the day, and day or month of the year. Even though Holidays have a limited impact, only significantly affecting demand on the specific holiday itself, likely due to reduced workplace activity, these type of models could be used by traders, so they need to be added to the model to make them accurate on that specific trading day.

The residual or fractional error graphs provided further insight. These graphs display the residuals across the four year dataset. Many territories hovered around a residual of 0; the only territory that did not was North. While overall MAPE was low, between 2 and 5 percent, the fractional error graphs revealed periods of high error variance, particularly during the summer months each year. This suggests that although average errors are low, model performance fluctuates seasonally.

In some regions, such as West, Capital, North, and Central, the residuals show a systematic shift over time, with the error trending either above or below the zero line, indicating potential changes in demand patterns or model drift.

Turning to the congestion models, we observed more variability across zones. Some areas, like Long Island, exhibited low congestion, whereas others, particularly Zone North, experienced significant congestion.

Unlike the demand models where temperature was the dominant predictor, the congestion models highlighted the month of the year as a more significant factor. This shift implies that congestion may be driven less by direct temperature effects and more by seasonal shifts in electricity flow, perhaps due to infrastructure limitations or power plant distribution. For instance, high congestion during non-winter months might be explained by an area's role as an energy corridor supplying electricity to another region.

This theory is supported by the fact that months appeared as the top predictors in most congestion models. Different regions may be characterized by different generation sources such as wind farms, solar installations, nuclear plants, or gas turbines, leading to different patterns of electricity flow and congestion.

Another noteworthy observation is that the congestion models were effective at predicting true positives but struggled with true negatives. This imbalance suggests room for improvement in distinguishing between congested and non-congested conditions.

Future Steps

In future work, we plan to investigate two key areas:

1. Why the congestion models had difficulty predicting true negatives, and
2. Why the North region was such a significant outlier in the demand models.

Overall, we conclude that the XGBoost model was a strong fit for this dataset, providing high accuracy in both demand forecasting and congestion prediction, while also highlighting areas for targeted model improvement and further study.

References:

- www.NYISO.com
- https://www.youtube.com/watch?v=vV12dGe_Fho
- https://www.youtube.com/watch?v=74_bdCARmO4