Data Analytics Capstone
Executive Summary and Implications
January 16th, 2024
Aaron Balke
S011005116

Presentation of Findings

https://youtu.be/wMvnZnSEaKI

Problem and Hypothesis

When Winter Storm Elliot crippled the Northeastern United States, one of the major complications was PJM Interconnection LLC's (PJM) inability to provide energy to all or parts of 13 states. This complication led to PJM receiving 2 Billion USD in regulatory penalties, a massive loss of revenue, and the worsening of maintenance and repair problems (Howland, 2023). To avoid these effects, an analysis of PJM regional max energy consumption to compare and forecast energy load between regions, enables effective resource allocation, maintenance and repair estimation, and generalized executive decision-making. Using PJM's energy load dataset, paired t-testing for comparing the Eastern and Western regions, and the advanced TBATS Time Series Analysis for forecasting complex seasonal components, effective analysis of regional max energy consumption is possible.

For this capstone, my research question is "Is the difference between PJM regional max energy consumption means statistically significant? And if so, can PJM Max Energy Load per Region be forecasted using TBATS time series modeling?".

Null hypothesis: The difference between PJM regional max energy consumption means is not statistically significant.

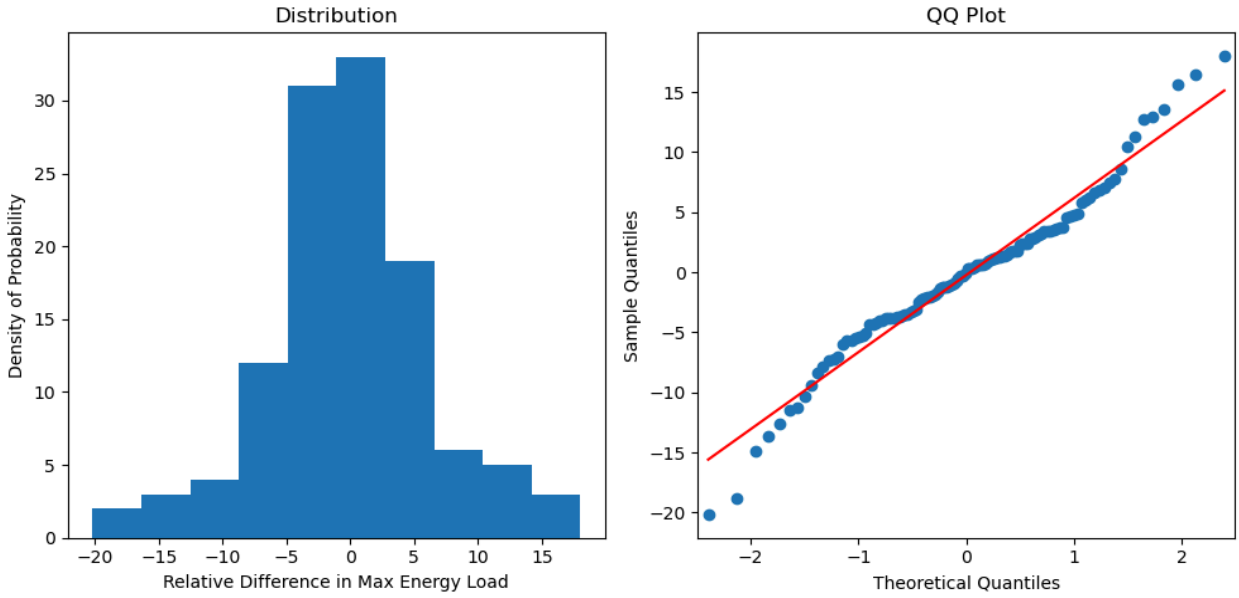
Alternate Hypothesis: The difference between PJM regional max energy consumption means is statistically significant.

A paired t-test has been chosen to evaluate if the difference in max regional energy consumption is statistically significant since it provides the ability to compare mean values of different regions, without requiring the removal of seasonal and trend components - components required for accurate differentiating in a business context.

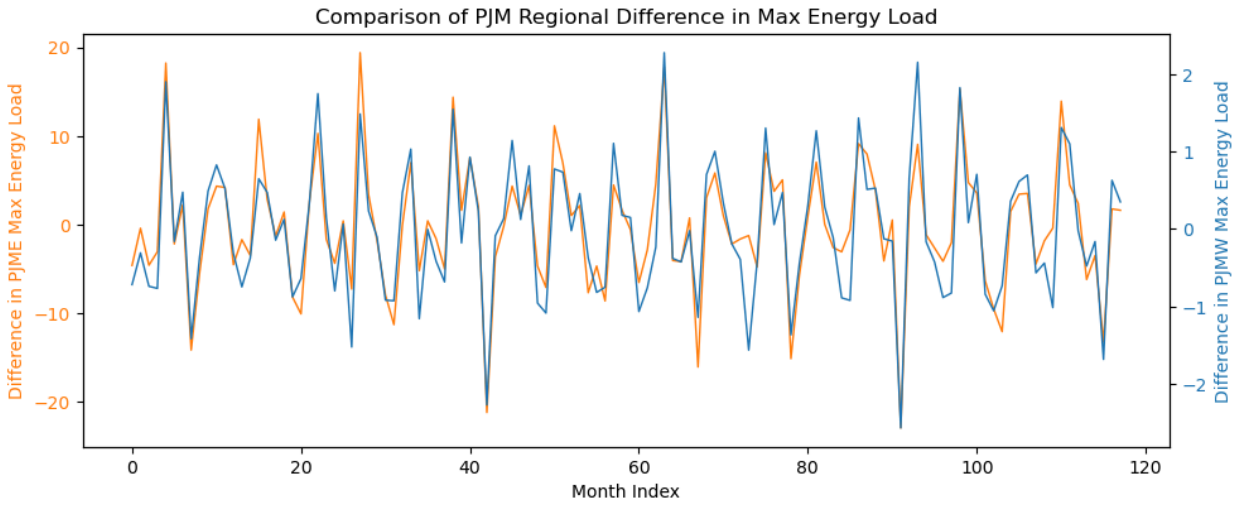
Summary of Data Analysis

Data was preprocessed for paired t-testing through a conversion from Magewatts to Gigawatts, and the aggregation of hourly data into monthly data. Since max load was required, only that month's highest consumption value was kept for each region. A monthly aggregation was chosen since there was 10+ years' worth of data, and too small of an aggregation period would be affected by randomness and the weekly seasonal component. For paired t-testing on time series data, the data must be stationary (Schein, 2020), outliers removed, and a normal distribution (Statistics Solutions, 2021). The PJM data was not stationary, so a relative difference value replaced the absolute energy consumption value. For outlier removal and normality testing, since a paired t-test is taking place, the difference between Western and Eastern regions is evaluated. Outliers were removed using the Z-score method, only outliers with a Z-score greater than 3 were removed. To verify normality, the D'Agostino-Pearson Normality Test was used which returned a p-value of 0.044, insufficient for the alpha value of 0.05. The solution was to apply data transformations. The shallow distribution could be fixed

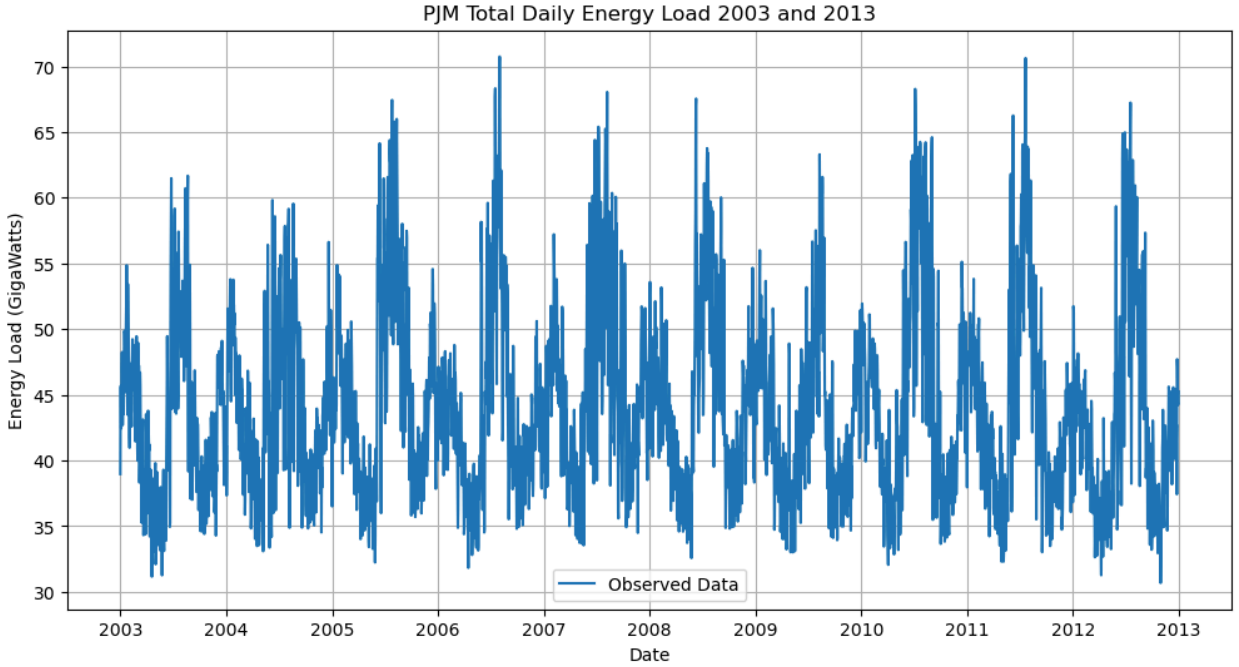
with either Box-Cox transformations or Yeo-Johnson transformations. Yeo-Johnson was chosen since it can be applied to our dataset's positive and negative values, which Box-Cox would be incapable of (Schendzielorz, 2021). After data transformations, a new p-value of 0.068 was achieved, meeting the requirements of our normality null hypothesis. To parallel the previous changes, identical transformations were applied to the original Western and Eastern data being inputted into the paired t-test.



With a Paired T-test, the null hypothesis is that the mean difference between the two groups is zero, there is no significant difference between the means of the groups. The alternative hypothesis is that there is a significant difference, and the mean difference between the two groups is not zero (Biology For Life, 2009). If the null hypothesis is met, observed differences can be explained as random variation (Chugh, 2023). An alpha value of 0.05 will be used, to evaluate the paired t-test. The paired t-test returned a p-value of 0.6512, greater than our alpha. We can conclude that the two groups are not significantly different, and any perceived differences between PJM East and West regional max energy load can be explained as random variation. Practically, this means trends and seasonal components are similar between regions, for example, the annual seasonal components are shared between regions, causing both to have similar increases in max energy consumption in the winter and summer. To visualize this, the following graph shows the relative difference values between PJME and PJMW, with scales adjusted.



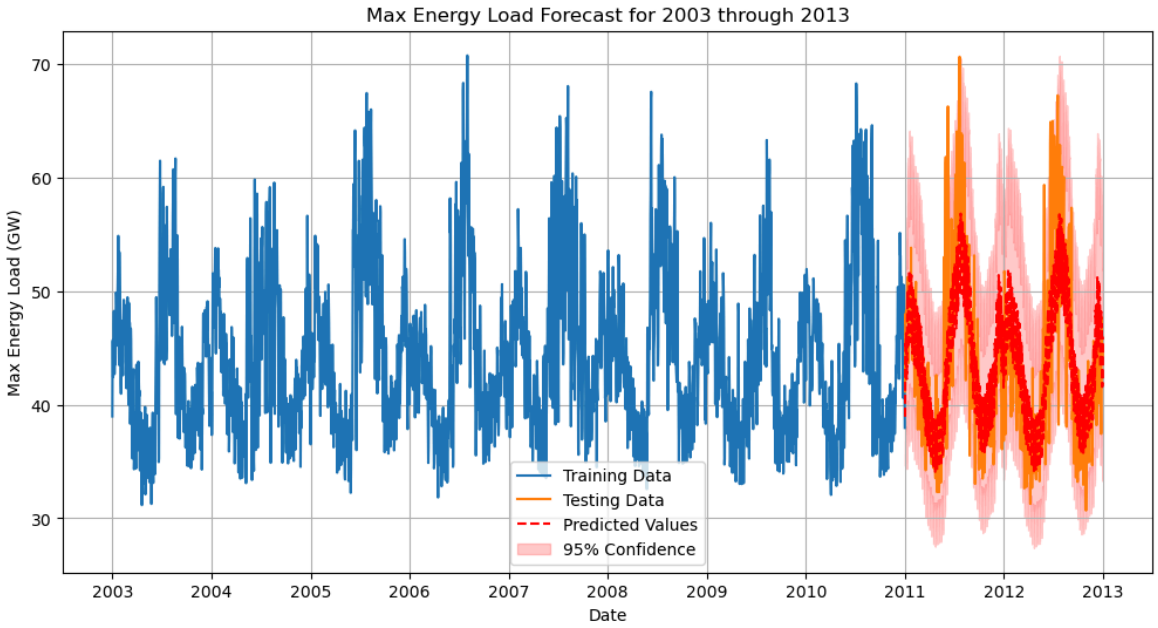
Since the Western and Eastern regions were not significantly different, the data was combined to create a Total PJM max energy consumption dataset. The creation of a single model to forecast PJM energy load was applied to both regions and allows for resource allocation, equipment maintenance/repair, and executive decision-making.

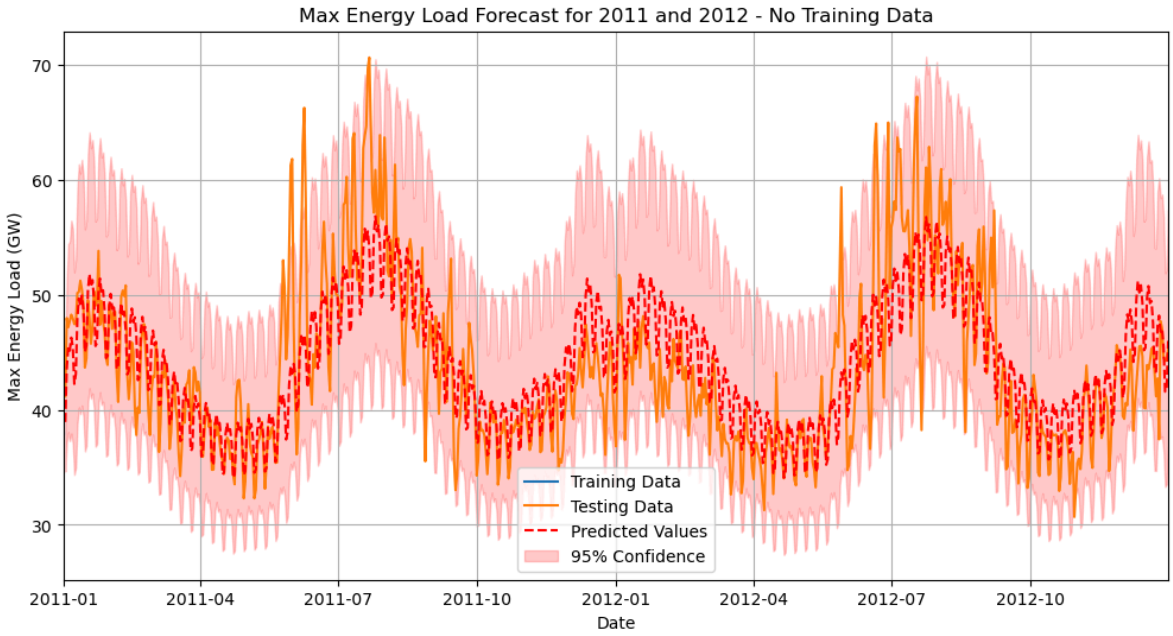


To prepare time series forecasting on the combined data, stationarity was evaluated using a Dickey-Fuller Test, which confirmed stationarity, and decomposition, which allowed for autocorrelation analysis to be completed on both the seasonal component and the observed data. This analysis rendered an interesting finding, the max energy load data had two seasonal components, a weekly and an annual seasonality. Originally, an ARIMA model was planned to be used for the analysis, however, the revelation of two seasonal components made this impossible. The solution was to use a TBATS time series model. TBATS is a forecasting method

that uses a combination of Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend components, and Seasonal components (Nadeem, 2021). The combination of these techniques creates a model that can fit multiple and complex seasonal components, particularly when compared to ARIMA models. A training and test split was created to validate the model's accuracy. The test size was two year's worth of data, equivalent to the 2011 and 2012 calendar years. All data available prior: the beginning of 2003 through the end of 2010, was used to train the model. This is roughly ~75%. The TBATS model was created with the seasonal periods as parameters, and the training data fitted. Then a prediction and confidence bounds were created using the model for 2011 and 2012, equal to the test data.

Findings





As previously mentioned, the paired test comparing PJM Western and Eastern regional max energy consumption ended with the conclusion that the regional values are not statistically different. The created TBATS model to forecast the combined max energy consumption was able to predict with a mean absolute percentage error of 7.67% and a root mean square error of 4.61 Gigawatts. This means our model can effectively forecast future max energy consumption for the combined dataset. Additionally, while not a standard metric, it is also observed that for a majority of the test/predict range, the actual data sits below our upper bound confidence value. From my limited understanding of equipment requirements, I believe we would want our minimum max consumption load to be our predicted values, and our hopeful value to be the upper bound of our confidence value. This would mean our equipment can always handle the maximum energy demand.

Limitations

A limitation of this analysis is how delicate the balance of stationarity and normality is. Depending on the aggreagation period and time range, it can become very difficult to remove stationarity from this dataset, and still meet a normal distribution. When the time range was set to be one year instead of ten, the data no longer was normally distributed, and when the aggregation period was set to one week instead of one month, the data was no longer able to be stationary and normally distributed. Choosing a good aggregation period, and time range while keeping as much data as possible requires a lot of trial and error.

Proposed Actions

I believe the next course of action for PJM will be to diagnose grid equipment to check it meets the max load values provided by the model. Since the model forecasted from the start of the year, if this was a hypothetical point of analysis, that would give PJM 7-8 months to assess the quality of equipment before the months of major power draw. Knowing this gives executives time to handle project management and get foresight into decision outcomes.

Expected Benefits

As mentioned previously, PJM Interconnection LLC was hit with 2 Billion Dollars in penalties for equipment failures in the 2022 Winter Storm Elliot (Howland, 2023). This does not include loss of business, or maintenance costs. The prevention of these financial hits starts with this comparison and forecast and it will assist PJM in catching equipment and resource problems before they become service problems for clients and regulatory bodies. ‘An ounce of prevention is worth a pound of cure’.

Sources

BATS and TBATS time series forecasting. (2022, December 22). GitHub. https://github.com/intive-DataScience/tbats

Biology For Life. (2009). T-test. BIOLOGY for LIFE. https://www.biologyforlife.com/t-test.html

Chugh, V. (2023, March 30). An Introduction to Python T-Tests. Datacamp. Retrieved January 14th, 2024, from https://www.datacamp.com/tutorial/an-introduction-to-python-t-tests

Hourly Energy Consumption. (2018, August 30). Www.kaggle.com. Retrieved January 8th, 2024, from https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption

Howland, E. (2023, January 12). PJM generators face up to $2B in penalties for failing to run during December's Winter Storm Elliott. Utility Dive. Retrieved January 16, 2024, https://www.utilitydive.com/news/pjm-generators-penalties-power-winter-storm-elliott/640242/

Nadeem. (2021, November 30). Time Series Forecasting using TBATS Model. Analytics Vidhya. Retrieved January 8th, 2024, from https://medium.com/analytics-vidhya/time-series-forecasting-using-tbats-model-ce8c429442a9

PJM - Home. (2024). Pjm.com. Retrieved January 8th, 2024, from https://pjm.com/

Schein, A., Bogiatzis-Gibbons, D., & Hardy, T. (2020). Guidance on conducting energy consumption analysis Created by the Behavioural Insights Team on behalf of the Department for Business, Energy and Industrial Strategy. In bi.team. The Behavioral Insights Team. Retrieved January 8th, 2024, from https://www.bi.team/wp-content/uploads/2020/12/Guidance-on-conducting-energy-consumption-analysis.pdf

Schendzielorz, T. (2020, January 15). A guide to Data Transformation. Medium. Retrieved January 8th, 2024, from https://medium.com/analytics-vidhya/a-guide-to-data-transformation-9e5fa9ae1ca3

Statistics Solutions. (2021, August 18). Paired Sample T-Test. Statistics Solutions. Retrieved January 14, 2024, https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/