New Jersey Energy Prediction Final Report

Kshitija Banala, Grace Wagner, Akhilesh Nimse

I. Abstract

For utility and energy production companies, predicting energy usage helps to accurately forecast and manage resources better. New Jersey includes four utilities: Jersey Central Power & Light, Rockland Electric Company, PSE&G and Atlantic City Electric. These utilities provide power to various sections of the state. They require constant knowledge of how much power is being utilized and how much they will need to provide. This project aims to forecast energy usage for New Jersey's utility companies using weather, date category, and energy usage data. The findings of our project showed the potential of using a LightGBM model for energy forecasting. Here we show that the most significant features of predicting energy usage are temperature, humidity and pressure. Originally, it was assumed that the winter and summer months would prove to be more significant than other features within the date category data. However, they were not as significant as weekday, which was the most significant from this data. Overall, we can hypothesize that energy usage is best predicted through temperature, pressure and humidity. This work can be used by future utilities and energy producers to estimate anticipated demand, ultimately leading to bettering environmental issues and decreasing energy overproduction.

II. Introduction

The goal of this project is to forecast energy usage for four utility companies in New Jersey by analyzing weather, economic, and energy consumption data. Forecasting energy patterns is vital for ensuring the efficient use of resources, maintaining reliable energy services, and supporting New Jersey's transition to renewable energy sources. Accurate energy forecasts enable utility companies to plan better, reduce costs, and prevent shortages or overproduction. Additionally, policymakers can use these insights for strategic planning, while residents can benefit from lower costs and improved service reliability.

We are working with various datasets, including historical energy consumption, weather patterns, air quality, economic indicators, and financial status of residents. Since some of these datasets have limited amounts of data, maintaining their quality and accuracy is a core focus. By ensuring the data is accurate and well-prepared, our models will better reflect real-world trends and lead to more reliable predictions.

This paper details our approach to forecasting energy usage, findings from model testing, and the feature importance analysis that highlights which factors—like pressure, temperature, and humidity—most significantly affect energy consumption. Through this study, we aim to provide actionable insights that can support utility companies, policymakers, and New Jersey residents alike by improving decision-making and enabling more efficient energy planning.

III. Beneficiaries

The beneficiary of this project is Constellation. Through the analysis of how weather, economics and current usage affect load, Constellation can use this project to gain a better understanding of the amount of energy that is being required by utility companies on the New Jersey Grid. This can help them determine how much energy to produce and where to supply that energy. The forecast of usage can also help them determine how much energy will need to be supplied in the future and how much usage growth they will be able to support. New Jersey residents will also be a beneficiary due to Constellation being able to make better business decisions. Improving the efficiency of Constellation's energy production will allow the company to provide better service for it's customers, and therefore better service to New Jersey's public.

IV. Literature Review

Our team conducted literature research on the topics of time series forecasting and boosted decision tree regression models. We also looked at literature that reviewed similar problems to the one we are solving with New Jersey energy forecasting. First, we looked at the Practical Time series Forecasting with R: A Hands-On Guide [3rd Edition] textbook which we used this paper as a guide for understanding time series data and forecasting problems as a whole. The paper provided us with information on the various models that can be used with our datasets including regression and neural networks. It gave us an idea regarding cleaning data in terms of handling missing values, outliers, and transformations to stabilize variance. We gained an overall understanding of how to manage and manipulate time series data effectively. Next, we looked at the Boosted Decision Tree Regression Model paper from the Journal of Animal Ecology titled "A working guide to boosted regression trees" which we will use this paper as a guide to developing a basic understanding of boosted regression trees. It discusses the boosted regression tree (BRT) approach which differs fundamentally from traditional regression methods that produce a single 'best' model. This approach uses the technique of boosting to combine large numbers of relatively simple tree models to optimize predictive performance. The authors also offer guidance on model building, parameter selection, and interpretation, highlighting BRTs' flexibility in dealing with different types of data and challenges common in ecology. Next, we looked a paper titled "Regression Model-Based Short-Term Load Forecasting for University Campus Load" which compares various regression models and their applications for load forecasting, particularly focusing on approaches such as Multiple Linear Regression (MLR), ensemble trees (such as "Boosted Decision Trees), and Artificial Neural Networks (ANN). Finally, we looked

into XGboost through a paper entitled Short-Term Load Forecasting Method Based on Feature Preference Strategy and LightGBM-XGboost. Building on Yao, Fu, and Zong's 2022 methodology, our project adopted a similar feature selection and optimization approach with LightGBM and XGBoost to improve the precision of energy load forecasting for New Jersey. Their use of Maximum Information Coefficient (MIC) for selecting critical features guided us in focusing on the most relevant data, reducing redundancy, and enhancing our model's learning efficiency. Additionally, their technique of error correction by leveraging both LightGBM and XGBoost helped us refine our predictions by balancing accuracy across the two algorithms. This approach could help us manage the nonlinear and seasonal changes in New Jersey's energy demand, allowing for a model that can accurately reflect the state's climate and economic conditions. Finally, we looked at a paper titled "Energy Consumption Forecasts by Gradient Boosting Regression Trees" to analyze how others developed a model for a similar problem and what sources they used. This article gave us a lot of valuable information and helped provide us with a good way to outline our final report and work from here on out. After viewing the python package that they used for their gradient boosting regression trees model, LightGBM, we decided to attempt to use this package and its documentation to develop our own model. In the introduction of the article, we gained a better understanding of some of the benefits that the development of their model provided. For example, the article mentions how their model, providing a better estimation for energy usage and production, would help to promote green energy. This example as well as others helped us reflect on some of the ways our model would be beneficial. The introduction also explained the breakdown of Italy's utilities, since Italy was the area they were creating their model for. We noted this description as a really useful explanation for readers and we realized the importance of explaining this in our final report. Additionally, they explained the need for a model for each of the utility zones. This was not something that we considered before but came apparent after reading the article and beginning to develop our model and datasets for the models. The article also talks about some of the preprocessing they completed before training their model. We did not do any preprocessing like they mentioned but we did note their preprocessing and may try something similar to better our model if needed. Later in the article, they describe some of the trends they found in their data. They found that weather conditions and calendar data were likely to influence energy usage. Their findings helped guide us in the direction of where to look for trends in our own data. We did this by making visuals showing trends in weather and calendar data. Our trends are highlighted in section VI. Looking at the data that they used in their model, weather and calendar data, helped us to also reflect on what to use for our very first initial model. They did not use any financial data for their model and were able to get good results without this data. So, we plan to develop an initial model with just the weather and calendar data and compare it to a second model with financial data to see if it benefits the model in any way. The article also addresses some of the ways they modified their data, like removing leap days and getting data from the first wave of Covid. We noted this while reading and are discussing whether we should apply changes like this as well. One example of how we plan to analyze whether this change, specifically for Covid, is needed is by looking at trends from that time and seeing if there is a large shift in the data correlating to Covid. If there is, this could show us that we need to remove this period of data due to it being an outlier and not representing the normal pattern of energy usage. The article also helped us understand from an educational perspective how models like this, once developed and deployed, are maintained and used. This was something that we discussed as a team, and it helped us to understand our responsibility for this project in a much better way. We noted this part of the article as a really good source to reference for further consideration and work in our final paper. Overall, this paper was very valuable for us to review. It helped guide us a lot in where to go with the project and helped us understand the problem much more. It also helped us understand how a real world model is developed and deployed.

V. Data Science Solution Framework

First, we'll gather a mix of data, including past and current energy consumption, weather data, economic indicators, air quality, and the financial status of residents. We'll clean and prepare this data to ensure it's accurate and ready for analysis. Next, we'll create and test different features and expand upon features with non linear transformation example having both x and x^2 to capture the important factors affecting energy usage. We'll experiment with various models, like time series and machine learning techniques, to see which ones provide the best predictions. In addition, figuring out which error metrics produce the largest error by experimentation can be a valuable metric that can help determine room for improvement in terms of figuring out why the error is so large and minimizing the errors with the datasets.

VI. Weather, Energy, and Economic Datasets Description

A. NJ Weather Datasets Description

In our data cleaning process for the New Jersey weather datasets, we found useful R code online that focuses on data wrangling techniques.Using R for cleaning the weather dataset was crucial, especially given the extensive duration of the data spanning several years.We discovered online resources that provided R code specifically for data wrangling, which allowed us to calculate the mean for each column effectively. By comparing these mean values with individual entries, we identified and eliminated anomalies that deviated significantly from the averages. In time series data, missing values create "holes" that can significantly impact forecasting accuracy. This is particularly critical for weather data, where models like ARIMA or smoothing methods, which rely on the relationship between consecutive periods, cannot be applied directly to series with missing values (Elith, 39). For instance, if weather data lacks temperature readings for certain days, forecasts relying on ARIMA would struggle since they require complete

data for accurate calculations. To address missing values in our weather dataset, we can utilize various imputation methods. Simple solutions include averaging neighboring values, which may work for daily temperature readings, while more sophisticated methods could involve forecasting the missing values based on historical trends or utilizing data from nearby weather stations. We encountered unequally spaced weather data, and we can employ interpolation techniques to create an evenly spaced series. This allows us to utilize forecasting models effectively, ensuring our predictions account for all relevant data points. Extreme values, or outliers, in weather data—such as an unusually high temperature reading during a cold spell—can skew our forecasts. Determining whether to remove these values requires careful consideration. We need to ask if the extreme value was a result of a data entry error or a genuine anomaly (e.g., a heatwave). In cases where no clear justification exists for exclusion, a best practice would involve generating forecasts both with and without the extreme values, allowing us to assess their impact on our predictions. Finally, the time span of the data is vital for accurate forecasting. A series that is too short may lack sufficient information, while an excessively long series might incorporate outdated patterns that are no longer relevant. For instance, if our weather data spans several decades, we need to consider changes in climate patterns or urban development that could influence current weather trends. In our project, we should focus on using a time span that reflects the current climate conditions, avoiding historical periods that might not be applicable to our forecast horizon. This targeted approach will enhance the accuracy of our weather predictions, ensuring that our models are trained on the most relevant data.

B. NJ Economic Datasets Description

The economic data sets in our project contain data about personal income (annually and quarterly), employment, CPI, personal consumption expenditures (PCE), and New Jersey resident population. Data under the topic of annual personal income and employment consist of datasets on the following information: wages and salaries by North American Industry classification System (NAICS) industry, personal income by major component, personal income by major component and earnings by NAICS industry and compensation of employees by NAICS industry. Quarterly personal income consist of datasets on the following information: personal income by major component, personal income by major component and earnings by NAICS industry, compensation of employees by NAICS industry, personal current transfer receipts and wages and salaries by NAICS industry. PCE consists of datasets on the following information: PCE by major type of product, per capita PCE by major type of product, PCE by state by type of product and PCE by state by function. The employment and CPI data each includes the monthly values from 2014 to 2024 (up until August). New Jersey resident population includes the yearly populations, in thousands of persons, as of 2023. A lot of the data within the economic section includes missing values. There are three different types of missing values present in these datasets. The first type is labeled as ”(D)” within the data. This data isn't shown so confidential information is not disclosed. The second type of missing data is labeled as ”(T)”. This label stands for an estimation of the suppressed wages and salaries where the earnings are also estimated. Estimates for both of the first two types of missing data are included in higher-level totals within the data. The third missing data type is ”(NM)”, which stands for data which is not meaningful for our analysis. There is also data labeled as ”(L)”, which represent incomes that are less than 50,000 dollars. Finally, our data also includes NA values. The economic datasets were relatively small, and were able to be cleaned initially by hand. This cleaning consisted of removing columns that had information that was unnecessary. The columns that were removed were GeoFIPS, GeoName, Region, TableName and LineCode. LineCode and TableName had information that was completely unnecessary for our analysis. GeoFIPS, GeoName and Region contained information that was repetitive and that was already understood. For example, GeoName was a column with entries the contained only ”New Jersey”. Since all of the data is only from New Jersey, there was no reason to include this column in our data.

C. NJ Energy Datasets Description

The energy datasets in our project consist of the following categories: Net generation for biomass, small-scale solar photovoltaic, utility-scale solar photovoltaic, and for wind organized monthly, as well as the number of customer accounts. The datasets only have values ranging across 18 years from 2006 - 2024 and consist of two columns, the Month and Year along the New Jersey: all sectors thousand megawatthours. The data in the net generation for biomass has data from January 1, 2024 to June 24, 2024 with values ranging from 50 to 90 thousand megawatt hours with no outliers. After removing missing values, the next dataset regarding the net generation for small-scale photovoltaic energy has data from January 14, 2024 to June 24, 2024 with values ranging from 75 to 410 thousand megawatt hours. In this case, I think some of the values below 100 can be removed, as they seem to outliers in this case. Next, the values in the dataset for the utility-scale photovoltaic energy are from January 8, 2024 to June 24, 2024 with values ranging from 0.03 to approximately 200, indicating that the values can be removed to avoid outliers skewing the data, especially visualizations. The wind dataset ranges from March 1, 2006 through June 1, 2024 whereas the dataset regarding the number of customer accounts from January 1, 2008 to June 1, 2024. The wind dataset had outliers of 0 and missing values from January 1, 20021 through February 1, 2006 that have been removed whereas the dataset regarding the number of monthly customer accounts has no missing values. Regarding missing values, most of these energy datasets had missing values in the second column, specifically from the first week of all 12 months in 2024. These values were removed in Python and the columns were adjusted accordingly. Looking at outliers, the utility-scale solar photovoltaic monthly dataset had values ranging from less than 1 thousand to approximately 200 thousand. The decision to remove the smaller values should be carefully

considered since removing them can significantly impact the data and visualization done on the remaining data. In addition, formatting the datasets in order to create appropriate plots, as shown in Section VI as challenging as I took variety of steps in order to create cohesive and clear plots. First, I changed dates format in Month column to m/dd/yy format instead of dd-m. Then, I found the arithmetic means of all the sectors for megawatt hours in NJ for each month for each energy dataset and created according plots for all the datasets. Finally, for further analysis, I combined the means for each month create 1 bar graph in order to compare utility and small scale photovoltaic energy.

VII. Model Testing

A. Preliminary

We know that methods like linear regression or neural networks can tolerate some missing values. In our project, we can either choose to impute missing temperature or precipitation values, or we can fit models to the existing data points. However, the trade-off here involves balancing the risk of introducing errors through imputation against the potential loss of valuable data points. We were provided with several key insights and recommendations regarding potential reforms to our models, we are looking into testing the boosted decision tree regression model. Boosted regression trees combine the strengths of regression trees and boosting, which is an adaptive method for combining many simple models to yield improved predictive performance (Leathwick, 2). In the context of analyzing energy data, BRT can handle non-linearity and complex Relationships, missing data/outliers, can handle feature importance, and its adaptability to time series data. Our energy datasets on energy generation (biomass, solar, wind) have complex relationships with factors like time and BRT is well-suited for capturing such non-linear interactions without the need for explicitly modeling them, as decision trees naturally handle such complexities. In terms of feature importance, BRT can provide insights into the relative importance of different features (e.g., time, energy type) by quantifying how much each feature contributes to the model. Given the temporal nature of our datasets, BRT can be adapted for time series analysis by applying boosting techniques specifically designed for sequential data.

B. Future Models Plan

When considering what models to include for this problem, we used the recommendation models from the feedback from our project proposal. We plan to create an initial machine learning model using boosted decision trees. We are currently focusing on creating a working model for one of the four utility companies. Once we have a working decision tree model, we will then include the data to predict forecasting for the rest of the utilities. After that, we plan to move onto improving our model by experimenting with neural networks. This is not only the process that was recommended in our feedback, but it is also corroborated with the information found in the literature we reviewed, specifically ”Practical Time series Forecasting with R: Hands-On Guide [3rd Edition]”. This paper touches on the progression from an initial model to the more complex model of neural networks.

C. How our work offers new contribution

Our project offers a novel contribution to load regression and forecasting by specifically addressing New Jersey’s energy landscape in collaboration with Constellation. While existing studies have effectively applied regression techniques in various regions, including the campus in Newfoundland (II. LITERATURE REVIEW, Section C.), our work distinguishes itself by considering New Jersey’s diverse climate and population dynamics. Both projects emphasize the integration of weather, economic, and energy usage data; however, the Newfoundland study focuses on a more localized consumption pattern, whereas our project aims to tackle broader variability in energy demand across an entire state. This comprehensive approach enhances forecasting accuracy and operational efficiency tailored to New Jersey’s unique demands.

D. Conclusive Model Evaluation

To develop our initial model, we decided to use LightGBM, the python pachage references in our literature review of *Energy Consumption Forecasts by Gradient Boosting Regression Trees*. We referenced the official LightGBM documentation to get started on the model in preparation for our team’s project feedback. We attempted to use this model and came to an issue at the end of our model development. When running the line of code to actually train the LightGBM model, we came across an error that said one of our data files was unable to be opened. We received feedback and now believe that the error was coded by possibly referring to a dataset that was used in the LightGBM documentation without realizing it. We believe that this should be a simple fix and just caused from our lack of experience with the package. We plan to address this and seek out further help if this does not fix the issue as we have already attempted to search for the error and fix it by our own means.We would like to also note that the dataset we used in this initial model development was not completed fully. We realized that we were missing some essential data needed for our model and are working on fixing this. As of now, we are completing this dataset before developing the model further so that we can finish development and continue with the rest of the utilities. Once we fix the problems listed above, and are able to get an initial model from our code, we plan to implement hyperparameter training to find the best parameters for our model. This will provide us with the best performing model. For there, we also plan to add the provided financial data and see if this benefits our model or not. We also plan to consider trying out cross-validation and Bayesian optimization if we feel we need to based on our model’s performance.

VIII. Additional Data Cleaning

When we initially began to develop our model, there was a lot of additional data cleaning that we realized we needed to do before being able to split our data into training testing and validation data. Some of that consisted of removing units, making sure date formats were the same, and splitting the column with time and date into two separate columns. We initially did this in excel but decided to compete it within the code itself as we believe that this would be more beneficial for Constellation in the end. With the changes in the code, they would be able to clearly see all the changes that were made to the datasets provided. After completing these changes, we began to look at how we should merge the data frames into one large data frame so we could then split it into training, testing and validation data for the model. We also knew that we wanted to start small, working on one utility's model, and then building the others. To do this, we started with the utility RC. For this utility, we needed weather from 2020-2023. To get all this data into one dataset, we concatenated the data frames together. We knew that we wanted to initially work with the weather and calendar data. The weather data has the date listed in the time column with the time. Since we know that we will be merging data frames together based on dates, we knew that we would need to separate the time and date into two separate columns. After completing this we went through each column and removed the units attached to each value. For example, Temperature had values that were listed as a number followed by "F". So, we removed the "F", leaving us with the integer for the temperature value. We also realized that the formatting of some of the dates within our weather datasets were different. To fix this, we reformatted all of the dates to match the format "yyyy-mm-dd" to match the formatting within the other data frames. With this completed, we merged the weather data with the calendar data on the Date column. At this time, we are working to complete the dataset by adding the load profile data to the weather and calendar dataset described above. Once this is done, we can split this data and complete our development of this first model for RC. One important aspect of the data that came apparent to us during this time was how to apply the weather data. We initially didn't think too much about why we were given two different weather files. After looking further into this, and getting more comfortable with the data, we realized where the weather location was within New Jersey itself. The one file is for Newark, located near the top of New Jersey, weather while the other if for Vineland, located near the bottom of New Jersey, weather. When we realized this, we looked at the locations of the utilities we have data for and compared them to the locations of the weather. For each of the models we make for each individual utility, we will use the weather data, from either Newark or Vineland, based on which location they are closest to.

A. Datasets Visualization Explanation

Now let's discuss the visualization of our data through this this report.We decided to incorporate line, bar, and boxplots to effectively visualize our data. We used line graphs as they are ideal for showing trends over time. They allow us to visualize changes and patterns in data points across a continuous range, such as days, months, or years, which is the overall idea behind our project, to predict energy usage over hours, months, or years for 4 utilities and 2 rate classes. Next, we used bar graphs to compare different categories. They are effective for displaying discrete data and making comparisons between different sets of data. Finally, we utilized boxplots or box-and-whisker plots, are used to summarize the distribution of a dataset. They show the median, quartiles, and potential outliers in the data. We used boxplots as they provide a comprehensive summary of the data's spread and central tendency. They are particularly useful for identifying outliers and understanding the variability within the data.

IX. Energy Datasets Visualization

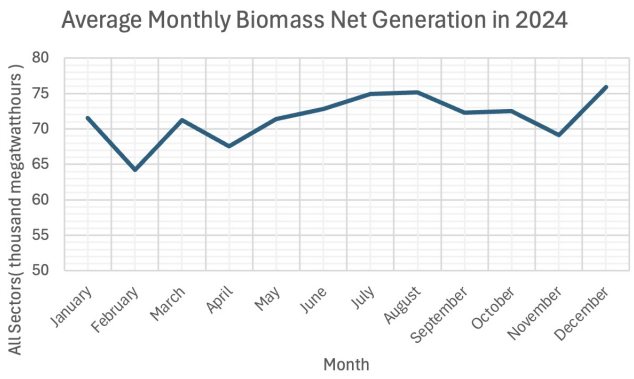


Fig. 1. This line plot above describes the correlation between the average monthly net generation of biomass energy in 2024. There tends to be a positive correlation for the overall trend of biomass energy as the months progress. It seems like the highest average biomass net generation was from June through August, likely due to the increased plant growth during these warmer months. This leads to a larger readily available supply of biomass material like wood, crops, and agricultural residues, which are the primary sources for generating biomass energy.

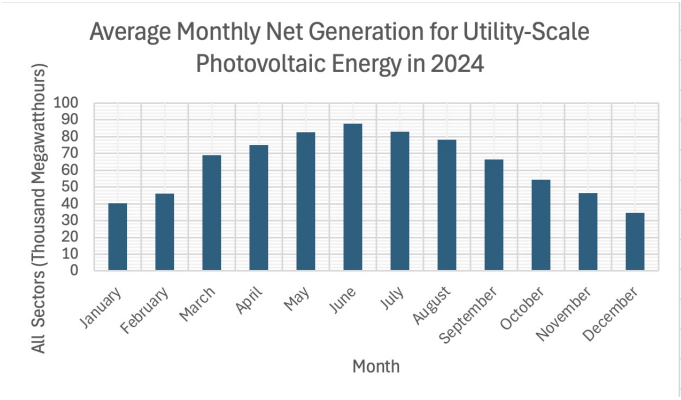


Fig. 2. This bar graph displays the average monthly net generation for utility scale photovoltaic energy. There seems to be a normal distribution among the average monthly values from January through December, with the peak values being from May through July and subsequent values in the late spring and early autumn months. Photovoltaic energy is essentially solar energy and utility-scale photovoltaic energy is higher in the warmer months primarily due to increased sunlight hours and intensity during the summer, leading to more solar radiation available to convert into electricity. Although the efficiency of individual solar panels slightly decreases at higher temperatures, the longer daylight hours and stronger sunlight outweigh the minor efficiency drop caused by heat.

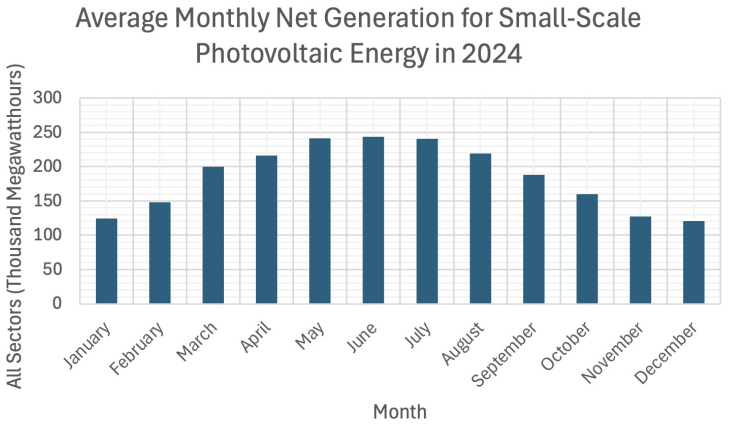


Fig. 3. This bar graph displays the average monthly net generation for small scale photovoltaic energy in 2024. There also seems to a relatively normal distribution with a slight right skew in January and February, with the peak values also being from May through July. While high temperatures, are not ideal for solar panels, they generally coincide with longer daylight hours and more direct sunlight in the summer which leads to a greater amount of solar radiation available for the panels to convert into electricity.

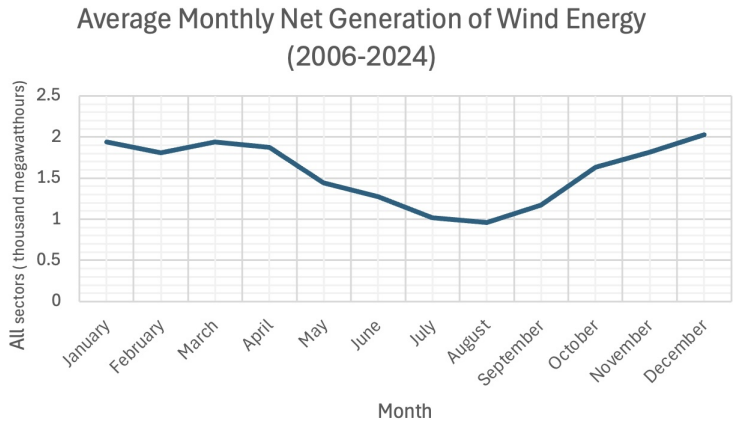


Fig. 4. This line graph displays the average monthly net generation of wind energy from 2006-2024. Unlike the previous plots, there seems to be a higher average monthly generation of wind energy during the colder months, such as January through Aril and peaking again from October through December. Wind energy is typically higher in colder months because cold air is denser than warm air, meaning the same wind speed can produce more power due to the closer air molecules.

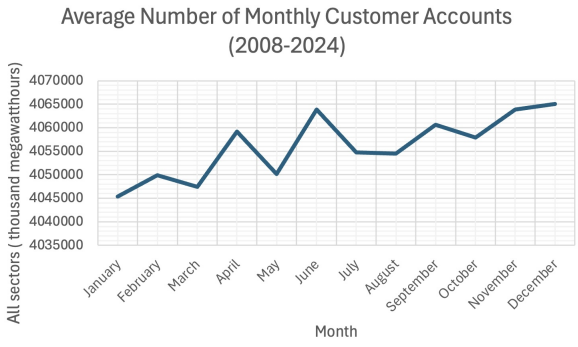


Fig. 5. This plot represents the average number of monthly customer accounts from 2008-2024. There are significant peaks in the customer accounts in the March and June with lower values in April and July. The highest value of customer accounts was in November and December. Since more people heat their homes in the winter months and cool them in the summer, there is increased demand for electricity or gas, causing an influx in price. There tends to be lower demand for electricity and gas during the spring and autumn months, leading to lower costs. Customer accounts is at a high in the early spring due to low demand and an influx of accounts is in the early summer and winter due to extreme temperatures.

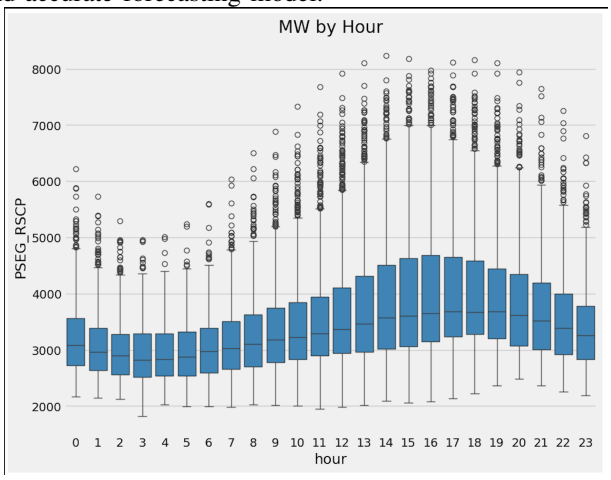
X. Model Performance and Findings

A. Preliminary Findings

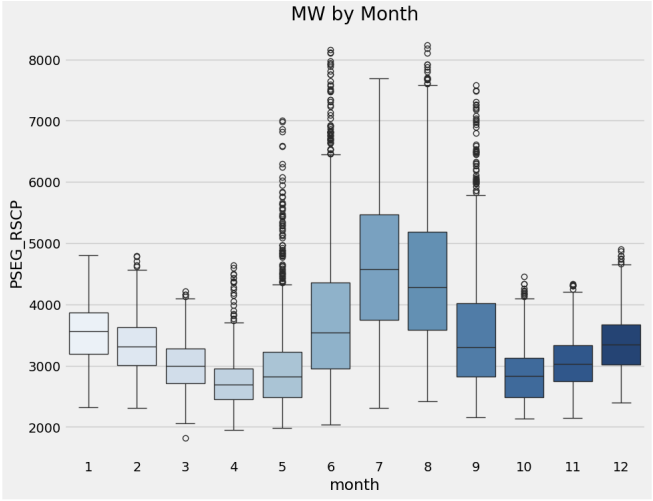
Our results demonstrate the key findings from our model, our model's overall performance, and a brief discussion of limitations in our overall analysis and future works.

B. Load Regression Analysis

To better understand the patterns in our load profile dataset, we visualized the energy generation data through two box plots. The first plot, "MW by Hour," illustrates the average energy generated at different hours of the day. From this graph, it is evident that energy generation follows a distinct diurnal pattern, with higher energy generation occurring in the late afternoon and early evening hours (approximately 3 PM to 8 PM). This trend aligns with typical energy consumption behaviors, where demand peaks during after-work hours due to increased residential and commercial activity, such as lighting, heating, and appliance usage. Conversely, the early morning hours (midnight to 5 AM) exhibit the lowest energy generation, reflecting reduced energy demand during sleeping hours. The second plot, "MW by Month," highlights the monthly variations in energy generation. This graph shows that the summer months, particularly June, July, and August, have the highest median energy generation. This is consistent with increased electricity demand during hot weather, driven by the extensive use of cooling systems like air conditioners. The winter months, such as December and January, show lower energy generation, possibly due to shorter daylight hours and less reliance on air conditioning. Spring and autumn months, like March and October, represent transitional periods with moderate energy generation. These analyses were crucial for interpreting the load data and gaining a deeper understanding of energy generation trends. They provided valuable insights into the temporal dynamics of energy usage, which helped us design a more comprehensive predictive model. By integrating these findings with additional parameters, such as weather features, we were able to capture the broader factors influencing energy generation and demand, paving the way for a robust and accurate forecasting model.



This screenshot shows the hourly megawatt energy usage for the PSEG utility and the RSCP rate class. This graph indicates that the 14th, 15th, and 16th hours of the day tended to have the highest energy usage. These are correlated with the afternoon timings of 2, 3, and 4pm indicating when the utility company might be doing the most work requiring energy generation. This boxplot seems to have some outliers at 1 and 2pm as well as 6pm with values of around 8000 megawatts.

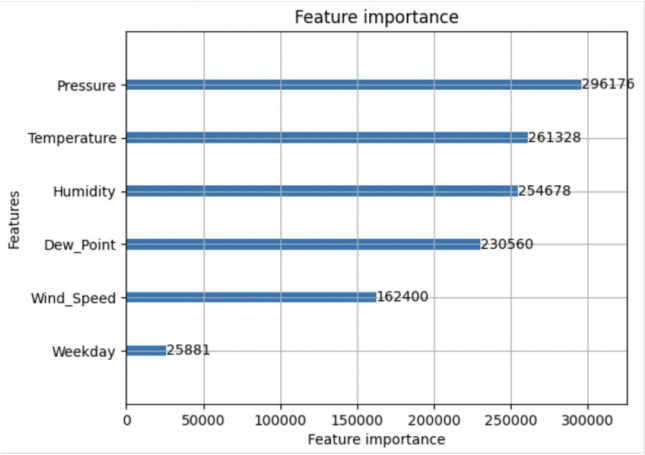


This screenshot shows the monthly megawatt energy usage for the PSEG utility and the RSCP rate class. This graph indicates that June, July, and August, and September had the highest monthly energy usage. A utility company typically uses the most energy in the summer months because of the widespread use of air conditioning, which significantly increases electricity demand as people try to cool their homes during hot weather, putting a strain on the power grid and requiring more electricity generation to meet the peak demand.

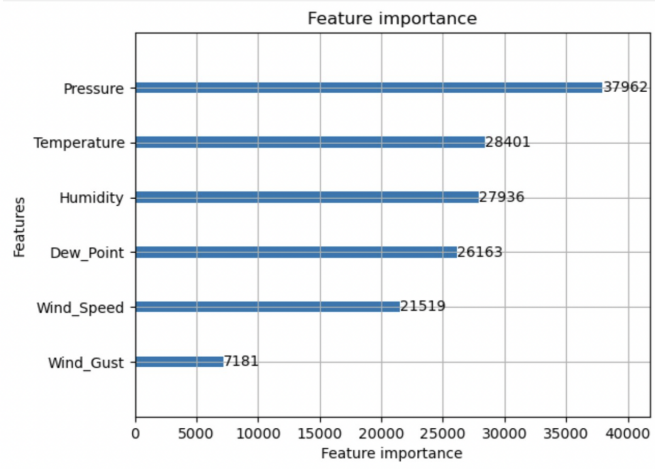
C. Feature Importance

Upon analyzing the feature importance results from our LightGBM load regression model, we observed intriguing trends across the eight graphs generated for the four utility companies and their two utility classes, RSCP and CIEP. In the majority of cases, atmospheric pressure emerged as the most significant factor influencing model performance. This finding aligns with real-world energy demand patterns, as variations in pressure often correspond to shifts in weather systems, such as the onset of storms or calm weather. For instance, low-pressure systems are typically associated with cloudy or stormy weather, which could increase heating or lighting demands. Conversely, high-pressure systems often indicate clear skies and more stable temperatures, potentially impacting energy usage patterns differently. Following pressure, temperature and humidity were the next most influential factors. Temperature directly correlates with heating and cooling needs, as extreme heat or cold drives energy consumption for HVAC systems. Humidity, on the other hand, can amplify perceived temperatures, influencing the operation of dehumidifiers or air conditioning systems. This synergy between temperature and humidity makes them critical factors in predicting energy loads. Interestingly, features like dew point, wind speed, and wind gust were not as influential in our models. Dew

point, though related to humidity, might be less relevant as a standalone metric since its effects are already encapsulated within temperature and humidity interactions. Wind speed and wind gust, while critical for wind energy generation, likely have limited direct impact on energy consumption for end users unless associated with extreme weather events, such as hurricanes, which are rare and seasonal in New Jersey. These insights provide valuable context for understanding the drivers of energy consumption and highlight the practical importance of selecting appropriate features for forecasting models. We categorized our features into 2 categories: weather and date where some weather features include temperature, dew point, humidity, and various types of precipitation whereas our date features include month, day, holiday, weekday, and weekends. Lets look at a feature importance graph below to gain a better understanding of some of the most valuable features for this model.



This screenshot shows all the features for the RECO utility's CIEP rate class. We can see that the most prevelant features include pressure, temperature, and humidity as they have the highest feature importance values. These features are so important because they directly influence the behavior of air, which in turn determines weather patterns, including wind direction, cloud formation, and precipitation. This implies that accurate measurements of pressure, humidity, and temperature are essential for predicting weather conditions.



This screenshot shows all the features for the RECO utility's RSCP rate class. We can see that the most prevelant features also include pressure, temperature, and humidity however there are other features such as wind speed and wind gust which play a role in this rate class. Wind speed and wind gust are important features because they impact various aspects like aviation, construction projects, and even the spread of air pollution, as they directly influence the movement of air masses and can cause significant damage when strong enough. Wind gust specifically represents sudden, rapid increases in wind speed which can be especially destructive.

XI. Future Works

In the next steps of our project, we plan to focus on improving the accuracy and usability of our energy forecasting models. One key area of improvement will be experimenting with neural networks, which can help us capture more complex relationships in the data that might be missed by our current LightGBM model. By exploring models like recurrent or convolutional neural networks, we hope to better handle the time-based patterns and interactions in the data. Additionally, we plan to fine-tune our LightGBM model by adjusting its parameters, such as learning rate and the number of trees, by using hyperparameter optimization techniques. To ensure that our models are reliable and not overfitting to the training data, we will also incorporate cross-validation as part of our evaluation process. Beyond technical improvements, we aim to ensure that this research provides tangible benefits for the people of New Jersey. By enabling utility companies to better predict and manage energy demand, our work can help reduce energy shortages and improve the reliability of energy services. Accurate forecasts can also support more efficient resource allocation, potentially lowering operational costs, which may translate to lower energy bills for consumers. Additionally, as New Jersey transitions to cleaner energy sources, our model could help promote the adoption of renewable energy by ensuring it is integrated effectively into the energy grid. We also plan to tailor our insights to specific utility classes, like RSCP and CIEP, to provide actionable recommendations for stakeholders. For example, scenario-based forecasts reflecting different weather or economic conditions could help companies and policymakers plan for extreme situations or seasonal variations. Ultimately, our goal is to deliver a solution that not only helps utility providers but also directly benefits New Jersey residents by making energy services more reliable, affordable, and sustainable.

XII. Conclusion

Our analysis of the feature importance results from the LightGBM load regression model has provided valuable insights into the key drivers of energy consumption for the four utility companies of PSEG, RECO, RCPL, and ACE along with their two utility classes, RSCP and CIEP. Atmospheric pressure emerged as the most significant factor, followed by temperature and humidity, highlighting the critical role of weather conditions in influencing energy demand. These findings emphasize the importance of selecting appropriate features for forecasting models to ensure accurate predictions. In conclusion, we aim to enhance the technical aspects of energy forecasting but also strive to deliver tangible benefits for the people of New Jersey, contributing to a more efficient and sustainable energy future.

XIII. Individual Contributions

The project begins with an Abstract, written by Grace. The introduction is next, contributed by all team members, setting the stage for the research and its objectives. The Literature Review is divided into five sections: Section A by Grace, Section B by Kshitija, Sections C and D by Akhilesh, and Section E by Grace. This review will cover the existing body of knowledge relevant to our study, providing a comprehensive background and identifying gaps that our research aims to fill. The Weather, Energy, and Economic Datasets Description is organized into three sections: Section A by Akhilesh, Section B by Grace, and Section C by Kshitija. This part of the report will detail the datasets used, including their sources, characteristics, and relevance to the study. Model Testing involves contributions from all team members in Section A, with Grace handling Section B, Akhilesh managing Section C, and Grace again taking on Section D. This section will describe the testing procedures, methodologies, and results of various models applied to the datasets. The additional Data Cleaning was conducted by Grace, ensuring the datasets are prepared for accurate analysis whereas the subsection datasets Visualization Explanation was done by Kshitija. Energy Dataset Visualization was handled by Kshitija, who created visual representations of the energy data to facilitate understanding and insights. The Model Performance and Findings section is a collaborative effort from all team members, divided into three parts (A, B, and C). This section will present the performance metrics of the models, interpret the results, and discuss the implications of the findings. The conclusion was done by Kshitija and finally, the Future Works section, contributed by all team members, outlines potential directions for further research, improvements to the models, and additional applications of the study's findings. In terms of the coding portion of the final project, Grace worked on the analysis of the RECO and ACE utility data whereas Kshitija worked on the JCPL utility analysis and Akhilesh focused on PSEG's data.

XIV. References

Literature Review Citations:

1) A: Shmueli, Galit, and Kenneth C. Lichtendahl. Practical Time Series Forecasting with R: A Hands-on Guide. Axelrod Schnall Publishers, 2018.
2) B. Elith, J., Leathwick, J.R. and Hastie, T. (2008), A working guide to boosted regression trees. Journal of Animal Ecology, 77: 802-813. https://doi.org/10.1111/j.1365- 2656.2008.01390.x
3) C. Madhukumar, M., Sebastian, A., Liang, X., Jamil, M., Shabbir, M. N. S. K. (2022). Regression model- based short-term load forecasting for university campus load. IEEE Access, 10, 8891-8905.
4) D. X. Yao, X. Fu and C. Zong, "Short-Term Load Forecasting Method Based on Feature Preference Strategy and LightGBM-XGboost," in IEEE Access, vol. 10, pp. 75257-75268, 2022, doi: 10.1109/ACCESS.2022.3192011.
5) E. Di Persio, L.; Fraccarolo, N. Energy Consumption Forecasts by Gradient Boosting Regression Trees. *Mathematics* 2023, *11*, 1068. https://doi.org/10.3390/math11051068