Project Name:	Siemens Industry Inc.
UIUC MSEE Capstone Project	
Sub-Project Name:	
Ensemble Learning and Reinforcement Learning for Energy Demand Prediction	

Author(s)

Name	E-Mail Address	Signature	Date
Hanliang Jiang	hj33@illinois.edu		

Reviewed by

Name	E-Mail Address	Signature	Date
Ashmin Mansingh	ashmin.mansingh@siemens.com		
Julio Massignan	julio.massignan@siemens.com		

Abstract

The growing technology of renewable resources makes it vitally important to accurately predict the power demands of electricity consumption, so that power management agencies can provide an appropriate amount of unrenewable energy and maintain the stability of power supply. Previous research work carried out by ISOs (Independent System Operators) took advantage of day ahead market and real time market statistical data to help build machine learning models to predict energy consumption.

In this paper, we collected data from CAISO (California Independent System Operator) in the past few months to experiment with a more systematic approach to predict power consumption at a higher accuracy. With the help of noise-based data augmentation and ensemble learning techniques, we achieved a higher accuracy rate than any methods CAISO used. We also explored the use of reinforcement learning to provide feedback for better model optimization. The prediction result we got, after being visualized, improved the accuracy rate in a decent amount compared to the existing models.

Table of Contents

1. Introduction

The technical-scientific methodology consists of exploring reinforcement learning methods alongside privacy preserving mechanisms inside the time series forecasting perspective. Initially, three theoretical background pillars are of interest in this project:

- **Time Series Forecasting for Electricity Markets.**
- **Privacy Preserving Algorithms and Emerging Data Markets.**
- **Uncertainty Quantification and Reinforcement Learning.**

For the purpose of accurately predicting and forecasting time series of data related to electricity demands and productivity, especially renewable energy resources like wind and solar power generators, it is necessary to construct handful tools and machine learning models to improve the accuracy of prediction. The work of prediction is of vital importance, as a small number of biases in prediction can result in a huge loss of industrial profit.

Exploring privacy preserving techniques is also important, for it ensures the customers' user data can be well preserved from leaking and malicious hacking while serving for computing and prediction model training purposes at the same time.

Leveraging Reinforcement Learning methods provides methods for taking advantage of different models to improve accuracy of predicting at the same time. Using ensemble learning techniques with the help of reinforcement learning, different models are used for prediction simultaneously, and the rewarding model provides feedback for these machine learning models. This feedback mechanism makes up for the deficiency of former methods lacking feedback, thus providing a better way for the models to iteratively improve themselves, and improve prediction accuracy step by step.

The software and systems for the envisioned tool will be developed using Python programming languages. The project development will be supervised by Siemens team of experts who will provide industry experience in software development for the energy sector. The current project has a strong potential for technological impact since it seeks to incorporate new concepts of Data Analytics for the energy sector while using state-of-the-art high-performance computing. This project will aim to reach a TRL level of 3 or 4, with a proof-of-concept and validation in the laboratory environment.

Bibliographic research will be conducted to select scalable algorithms with high performance. Then, a modular library will be developed in the Python programming language to provide a tool for Privacy-Preserving Reinforcement Learning Feedback for Time Series Forecasting Improvement. The software development will pursue software engineering concepts, object orientation, clean code, and design patterns.

2. Literature Review

A growing volume of data permeates the electricity sector in all distribution, transmission, and generation levels, aiming at the digitalization of modern electrical grids. In line with this new paradigm, modern energy markets rely on diverse information and data analysis methods, especially for two consolidated applications: uncertainty quantification and demand and generation forecasting. Forecasting capability is among the most important features of electricity markets and ISOs (Independent System Operators).

The market clearing process is extremely sensitive about the quality of its input forecasts, which may benefit a lot from reinforcement learning techniques. However, sharing sensitive operational data that could improve those methods can become a liability. This project aims to develop privacy-preserving techniques that can enable feedback of operational data to improve energy forecasting models, without disclosing protected information.

Modern electricity markets aim to provide an efficient and reliable supply of electricity. Multiple market levels encompass these objectives while accounting for characteristics of resource availability (e.g., generation type, resource limits, intertemporal temporal constraints, and capability), transmission network conditions (under normal or contingency conditions), demand level (both temporal and spatial), economic exposure (either in a pool market or purely cost-based) and regulation aspects, through an optimized energy procurement framework [1].

As in any forecast-dependent model, the market clearing is prone to forecasting errors, and hence, the forecasting accuracy has a high impact on energy procurement and prices in the day-ahead and real-time markets. Forecasting has always been at the forefront of decision-making and planning, with individuals and organizations seeking to minimize risks and maximize utilities. The large number of forecasting applications calls for diverse forecasting methods to tackle real-life challenges.

Thousands of energy forecasting papers have been published over the past few decades, including influential review articles and original research papers [2]. Besides the demand forecasting error, the electricity load has its random component, e.g., due to electricity consumption variations, different patterns in different day types, and weather impact [2,3, 4]. Moreover, the increasing introduction of renewable resources (such as solar and wind generation) increases the system-wide uncertainty, in this case, from the generation perspective.

Renewable resources forecasts are input to the market clearing, directly its results since the forecasting errors propagate their effects on dispatch and commitment instructions across different market levels. Many aspects affect renewables forecasting errors, such as seasonality, weather conditions, and outages, with increased influence according to their respective capacity in the energy mix [5, 6, 7].

The forecasting models can generally be categorized into three types: time-series, machine learning, and hybrid [2]. Several time-series forecasting methods, like the Autoregressive families and probabilistic methods, have already consolidated their positions in control and operation centers [2, 3, 4, 6]. Machine learning has also been adopted well since the evolution of computational performance, more prominently neural network-based methods [2, 3, 4, 5].

In recent years, Artificial Intelligence (AI) has experienced more hype, largely due to the advancement of computing technologies. Various advanced AI techniques, such as deep learning, reinforcement learning, and transfer learning, have been adopted in energy forecasting [3, 5]. Regarding the hybrid methods, that is joining time-series concepts together with machine learning and AI, the ensemble forecasts have been showing exceptional results, and their adoption has been more common [2, 3, 6], especially incorporating exogenous models to increase accuracy [7] or providing coherence among different levels of the system through hierarchical forecasting [3].

In this research project, energy forecasting will provide the starting point for understanding the role of information in operation centers. One or more forecasting techniques for demand and renewable generation will be explored to provide the initial setup for the reinforcement learning architecture to be developed, as discussed in the next sections.

3. Work Completed

Chapter 1: Research Plan

With the help of professional experts of Siemens, I have made progress for the first part of our project, time series forecasting for electricity markets. By diving deep into the CAISO (California Independent System Operator) information portal, I have gradually understood how power markets work. Consumers (Power demands) and electricity generators (Power producers) are interconnected with the help of market operators, market regulators and system operators, who take control of market prices and ensure the reliability of the power supply system.

With the help of these agencies, several short-term and long-term markets are maintained, including weekly markets, day-ahead markets, five-minutes markets and fifteen-minutes markets. The purpose of this part of our project is to make accurate predictions regarding these demand markets and the productivity of wind and solar power generation.

The overall plan for this project can be described as follows:

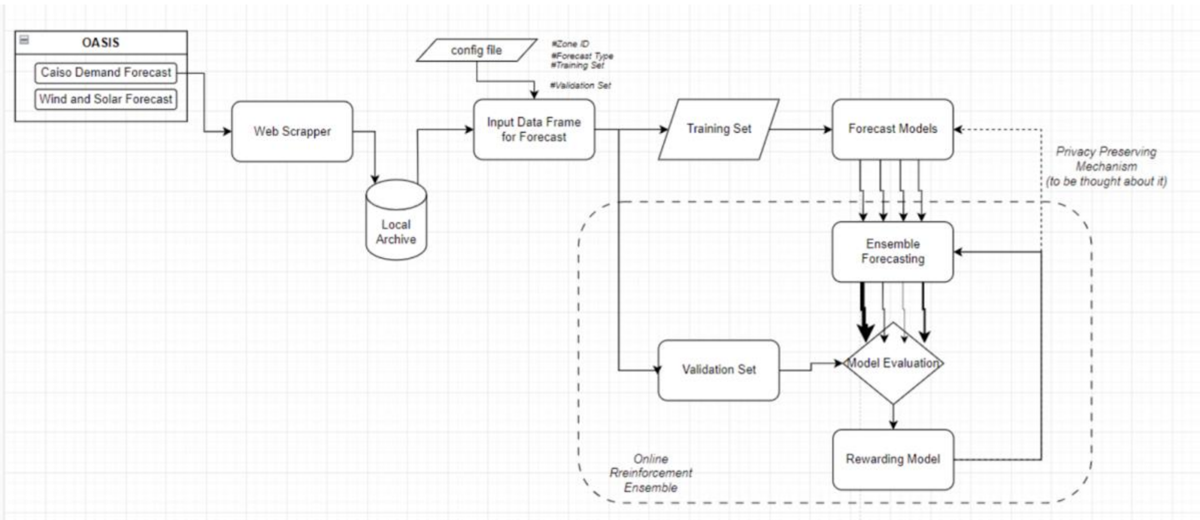

Figure 1: Overall Plan for the Project

In the diagram above, public data of CAISO (California Independent System Operator) is achieved by web scraper and stored locally, and the data is converted to the form of data frame with the use of config files. The input data is then split into training set and validation set, where different machine learning algorithms are used and ensemble together with the help of validation set. To make things better, we would use reinforcement learning methods to give reward and feedback to forecast and ensemble methods, thus increasing the accuracy of prediction.

Chapter 2: Data Collection and Visualization

After exploring the official portal of CAISO (California Independent System Operator), we achieved authorization of developer identification from CAISO. With the help of automatic python crawling tools, we get the data stored in csv file by taking use of configuration file and specifying the data range we want. An example of actual market demand during a time period is shown below:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	INTERVALSTA	INTERVALENI	LOAD_TYPE	OPR_DT	OPR_HR	OPR_INTERV	MARKET_RUI	TAC_AREA_N	LABEL	XML_DATA_IT	POS	MW	EXECUTION_	GROUP
2	2024-01-01T	2024-01-01T	0	1/1/24	3	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1087	ACTUAL	1
3	2024-01-01T	2024-01-02T	0	1/1/24	16	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1398	ACTUAL	1
4	2024-01-02T	2024-01-02T	0	1/1/24	23	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1228	ACTUAL	1
5	2024-01-02T	2024-01-02T	0	1/1/24	24	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1157	ACTUAL	1
6	2024-01-01T	2024-01-01T	0	1/1/24	5	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1109	ACTUAL	1
7	2024-01-01T	2024-01-01T	0	1/1/24	6	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1155	ACTUAL	1
8	2024-01-01T	2024-01-01T	0	1/1/24	10	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1345	ACTUAL	1
9	2024-01-01T	2024-01-01T	0	1/1/24	11	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1379	ACTUAL	1
10	2024-01-01T	2024-01-01T	0	1/1/24	12	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1391	ACTUAL	1
11	2024-01-02T	2024-01-02T	0	1/1/24	18	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1498	ACTUAL	1
12	2024-01-02T	2024-01-02T	0	1/1/24	21	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1417	ACTUAL	1
13	2024-01-02T	2024-01-02T	0	1/1/24	22	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1312	ACTUAL	1
14	2024-01-01T	2024-01-01T	0	1/1/24	2	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1101	ACTUAL	1
15	2024-01-01T	2024-01-01T	0	1/1/24	7	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1206	ACTUAL	1
16	2024-01-01T	2024-01-01T	0	1/1/24	8	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1251	ACTUAL	1
17	2024-01-01T	2024-01-01T	0	1/1/24	13	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1389	ACTUAL	1
18	2024-01-01T	2024-01-01T	0	1/1/24	4	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1088	ACTUAL	1
19	2024-01-01T	2024-01-01T	0	1/1/24	9	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1295	ACTUAL	1
20	2024-01-01T	2024-01-01T	0	1/1/24	14	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1379	ACTUAL	1
21	2024-01-02T	2024-01-02T	0	1/1/24	17	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1473	ACTUAL	1
22	2024-01-02T	2024-01-02T	0	1/1/24	20	0	ACTUAL	AVA	Total Actual	SYS_FCST_A	3.8	1434	ACTUAL	1

Figure 2: Data achieved from CAISO

For having a better understanding of the data structure and deciding which machine learning algorithm and optimization methods to use, we visualize the data of actual markets, day-ahead markets, five-minutes markets and fifteen-minutes markets. The experimental results are shown below:

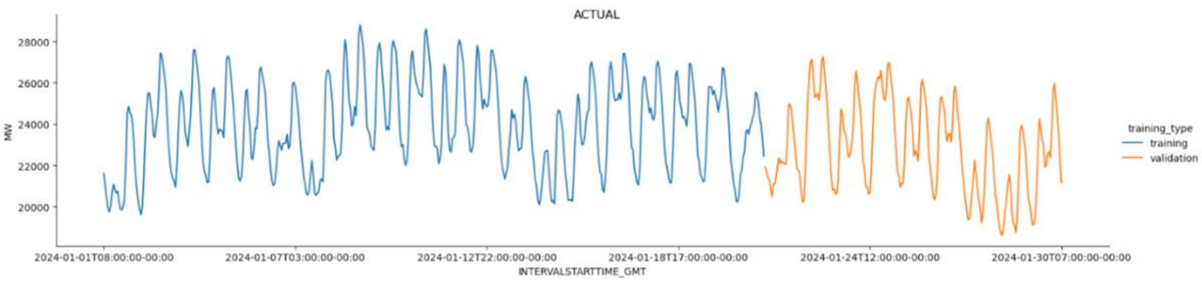

Figure 3: Ground Truth Visualization

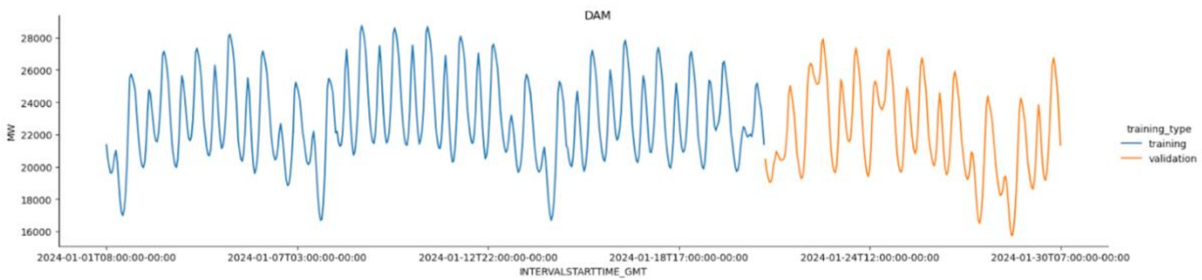

Figure 4: Day Ahead Market Visualization

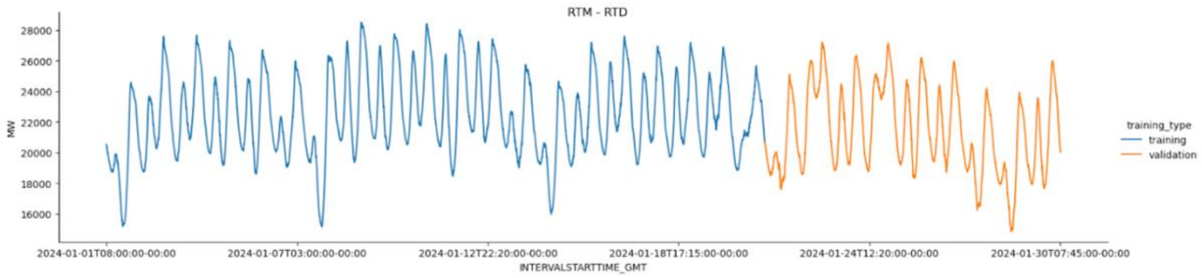

Figure 5: caption

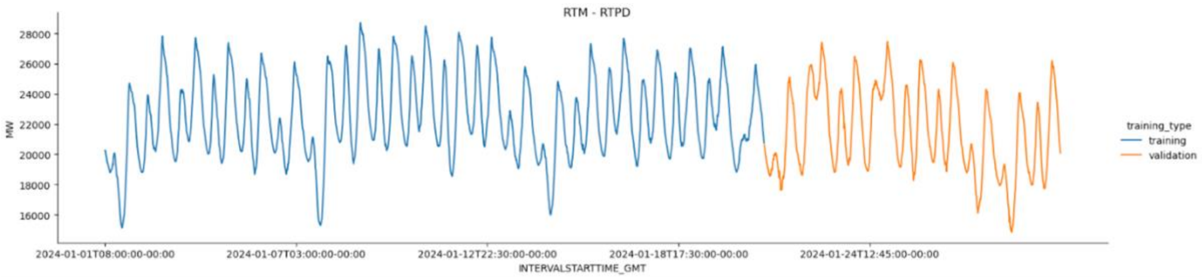

Figure 6

The visualization and figures above show the approximate trend for the demand market during the time period of January of 2024. Different colors in the pictures denote either training dataset or validation dataset for the purpose of our first training stage.

Learning about the current ability of CAISO (California Independent System Operator)'s prediction models are also necessary, for we want to know how well we can expect to help with the existing models, and further estimate how much more profit can our models contribute to with the help of bidding and market study. The current deficiency of the prediction models trained by CAISO is visualized below:

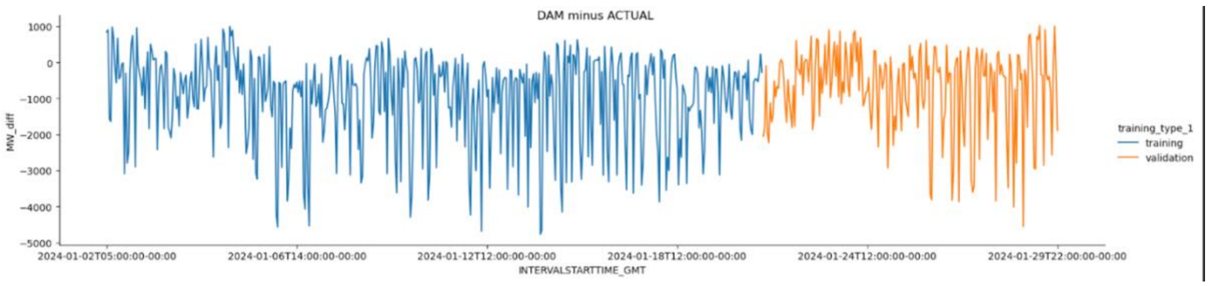

Figure 7: Day Ahead Market Error Visualization

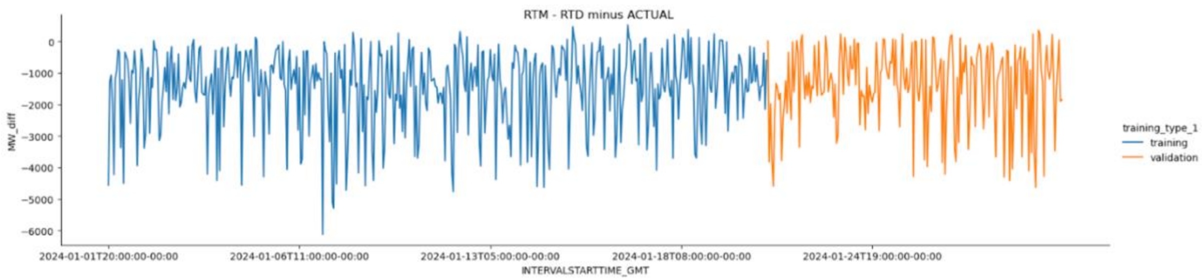

Figure 8: Real Time Market Error Visualization

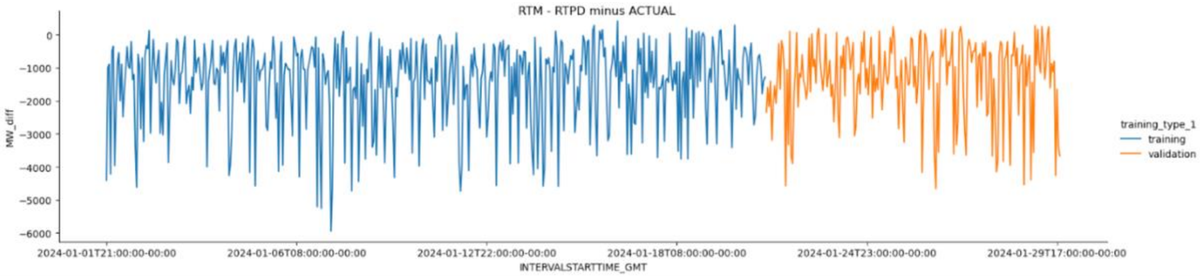

Figure 9: Real Time Market Error Visualization

From the visualizations above, we can conclude that none of three markets have done a good job predicting the demands. The biases of their predictions can go up to 6000 MWs, which is a huge amount of money. In a sense, using more advanced techniques like ensemble learning and reinforcement learning is of vital importance.

By analyzing data from CAISO (California Independent System Operator), we have a good understanding of the concepts and baseline of Time Series Forecasting for Electricity Markets. The purpose of the next stage is to actually explore a few machine learning methods for predicting demands and renewable energy productivity and using online reinforcement learning methods to provide feedback for ensemble forecasting.

Chapter 3: Data augmentation

Having Collected datasets from CAISO (California Independent System Operator) public website, we then focused on processing data and to overstep the existing models. We chose an official prediction model for the day ahead market as baseline, and selected ensemble learning as our major approach.

For the purposes of augmenting data for further training, we took advantage of ground truth data given by CAISO (California Independent System Operator) from Jan 1st, 2023, to July 1st, 2023. We augmented data by adding A wide variety of noises with different parameters, including Gaussian noise, Beta distributed noise, Weibull noise and Laplace noise. The plan of data augmentation is shown below:

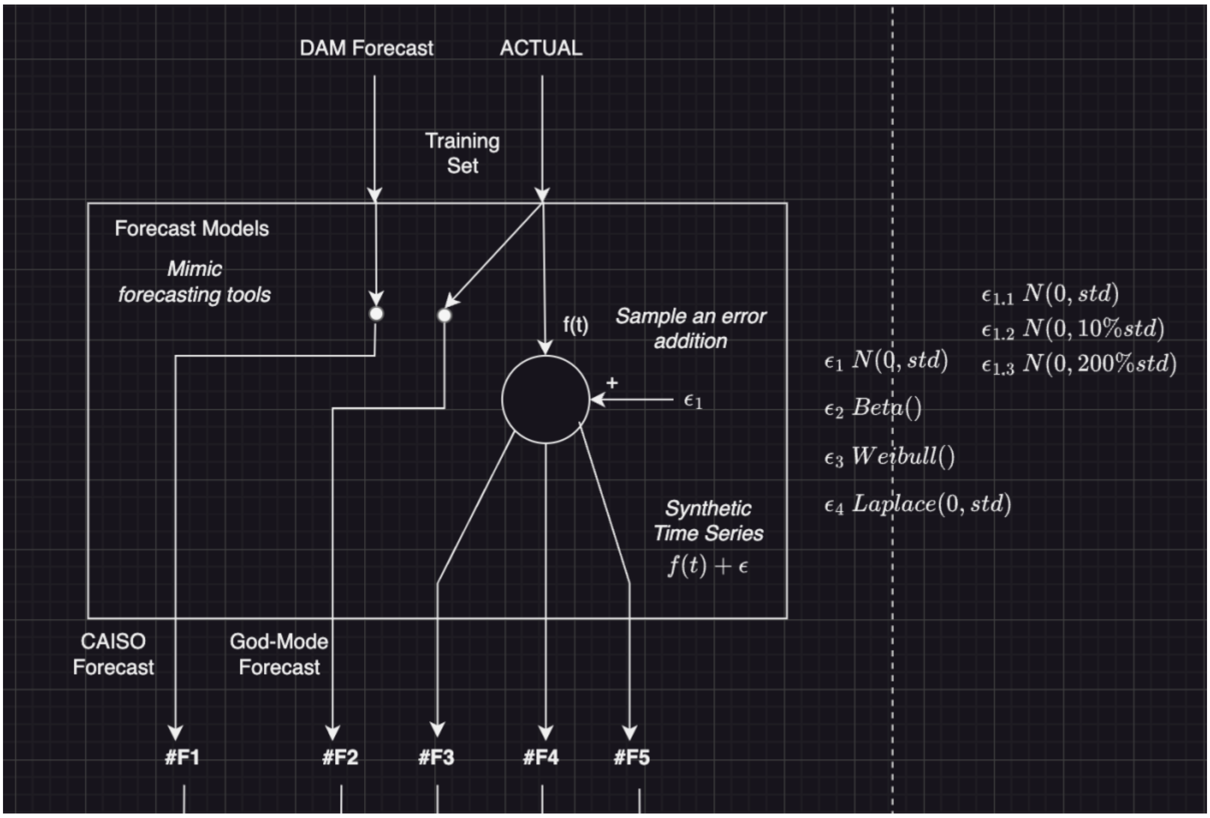

Figure 10: Research Plan for Data Augmentation

To keep the noise in an appropriate range so that the generated augmented data can reasonably reflect the real world condition, we normalized the noise with a zero mean and a standard deviation within the range of 0.02 of the ground truth data range.

The original ground truth data and augmented data are visualized below:

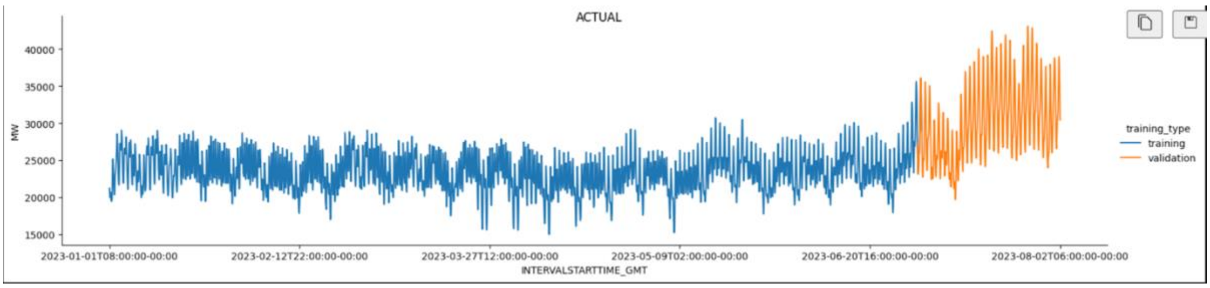

Figure 11: Original Ground Truth Data

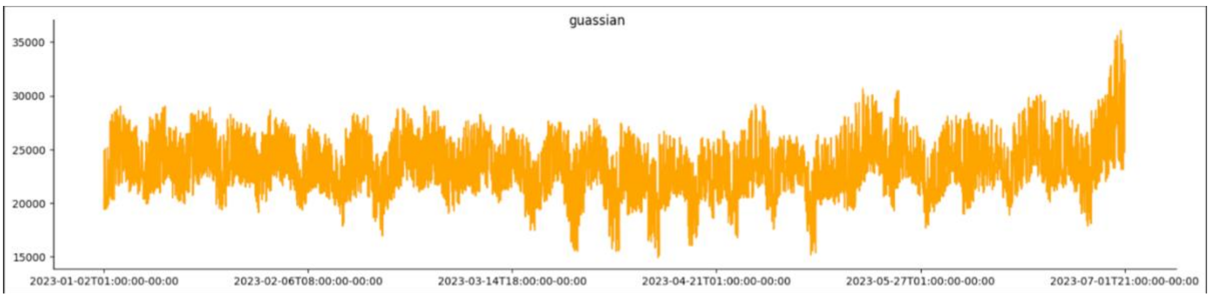

Figure 12: Ground Truth Data Augmented with Gaussian Noise

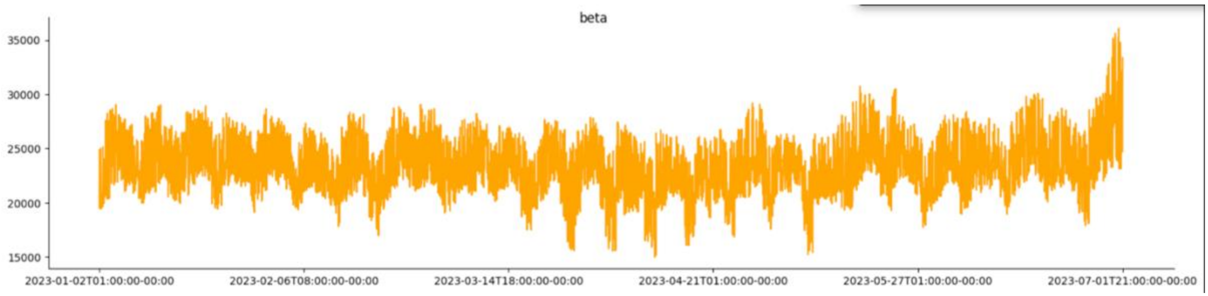

Figure 13: Ground Truth Data Augmented with Beta Noise

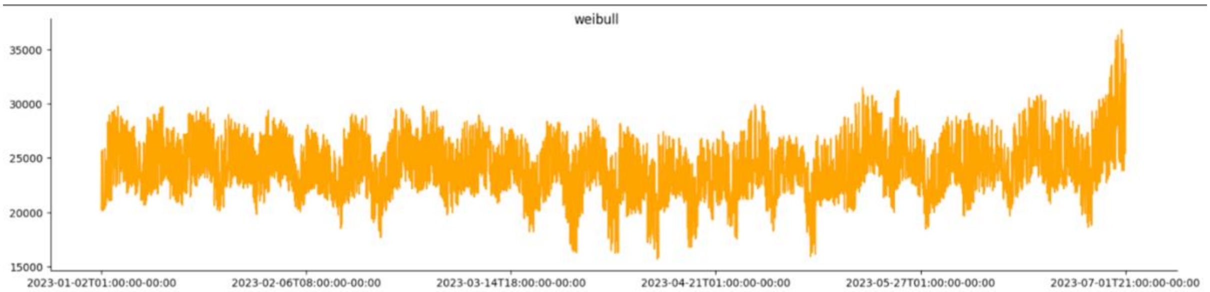

Figure 14: Ground Truth Data Augmented with Weibull Noise

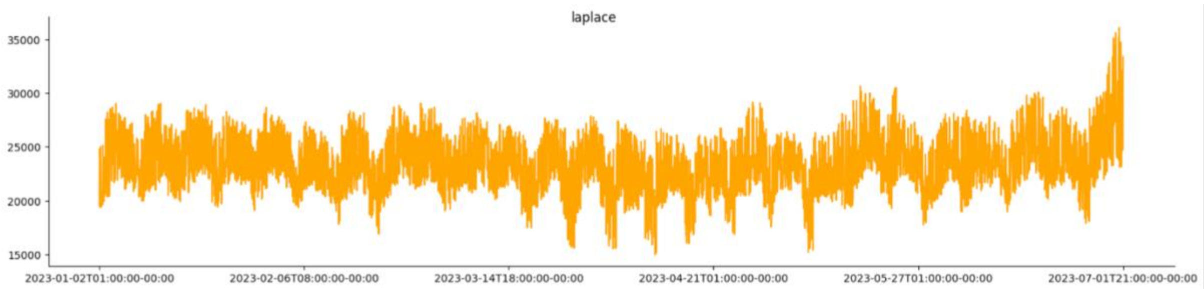

Figure 15: Ground Truth Data Augmented with Laplace Noise

From the visualizations above, we can conclude that data augmentation made the data more compelling while maintaining the time series feature of ground truth data.

With the help of augmented data of different flavors, we can figure out a way of ensembling data and accurately predict the power demand of consumers.

Chapter 4: Ensemble training

Ensemble learning is an important approach to improve the accuracy further when there are many models of similar performances. In this circumstance, we use ensemble learning on day ahead market prediction and models augmented by adding Gaussian noise, beta noise, Weibull noise and Laplace noise.

The research plan of ensemble forecasting is shown below:

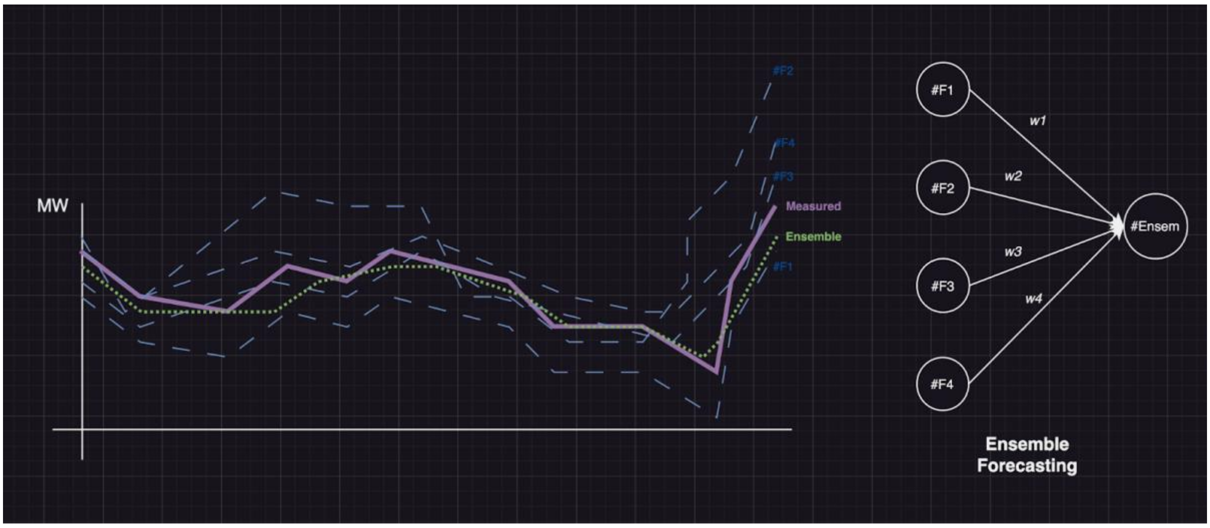

Figure 16: Plan for generating Ensemble Model

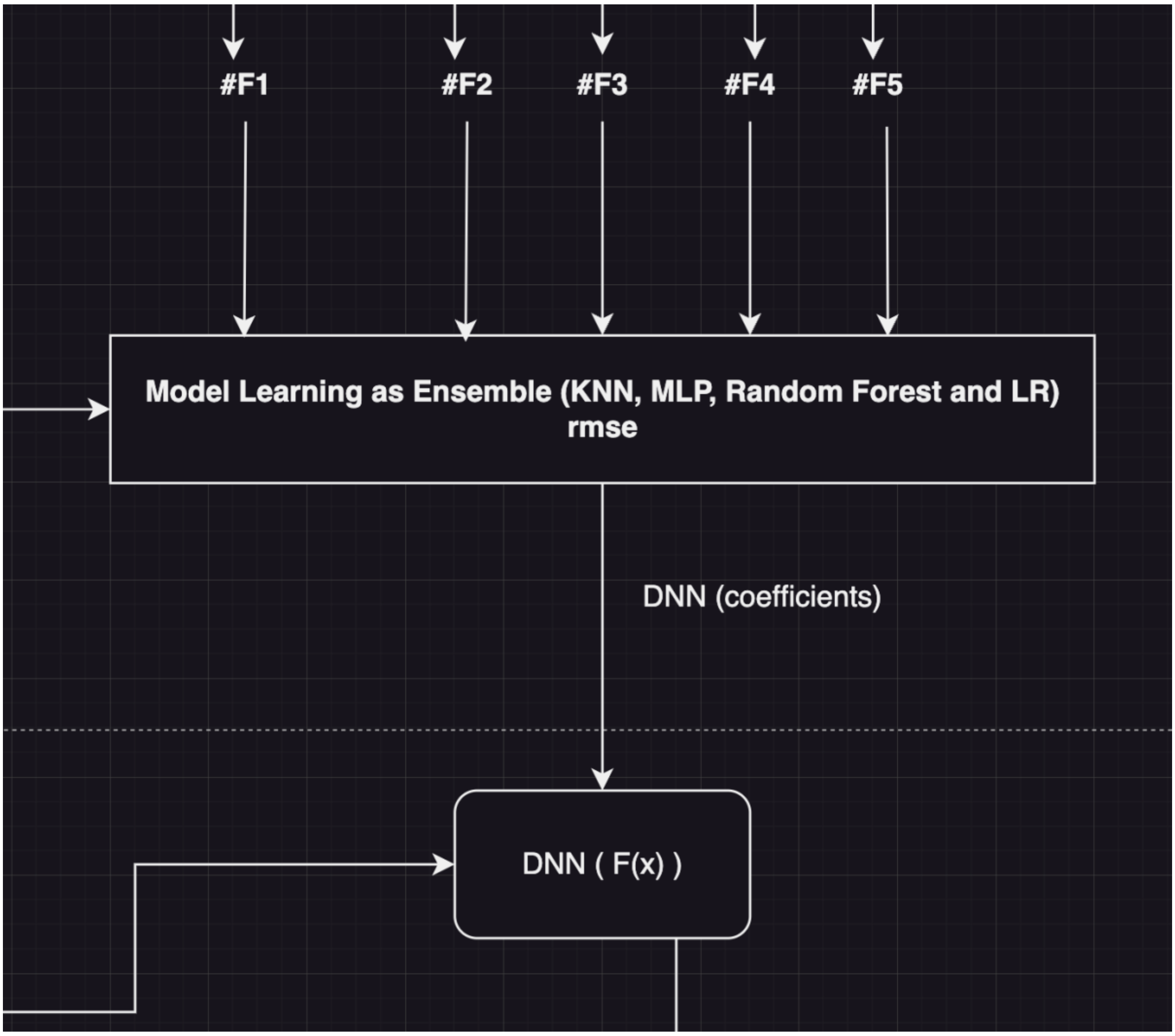

Figure 17: Plan for generating Ensemble Model

Given five prediction models, we used different methods to decide proper weights for each of the five models. As the training dataset and validation dataset have the real ground truth demand, we used K nearest neighbor algorithm, linear regression method, Random Forest method and we also experimented with a two layers neural network.

To get a better result, we tuned the hyperparameter of the number of nearest neighbors and epochs of training the Multilayer Perceptron. After training the model each time for a specific hyperparameter, we used the validation set, which included data from July 1st, 2023, to August 1st, 2023, to test the performance of the model's generalization.

The procedure of tuning K is shown below:

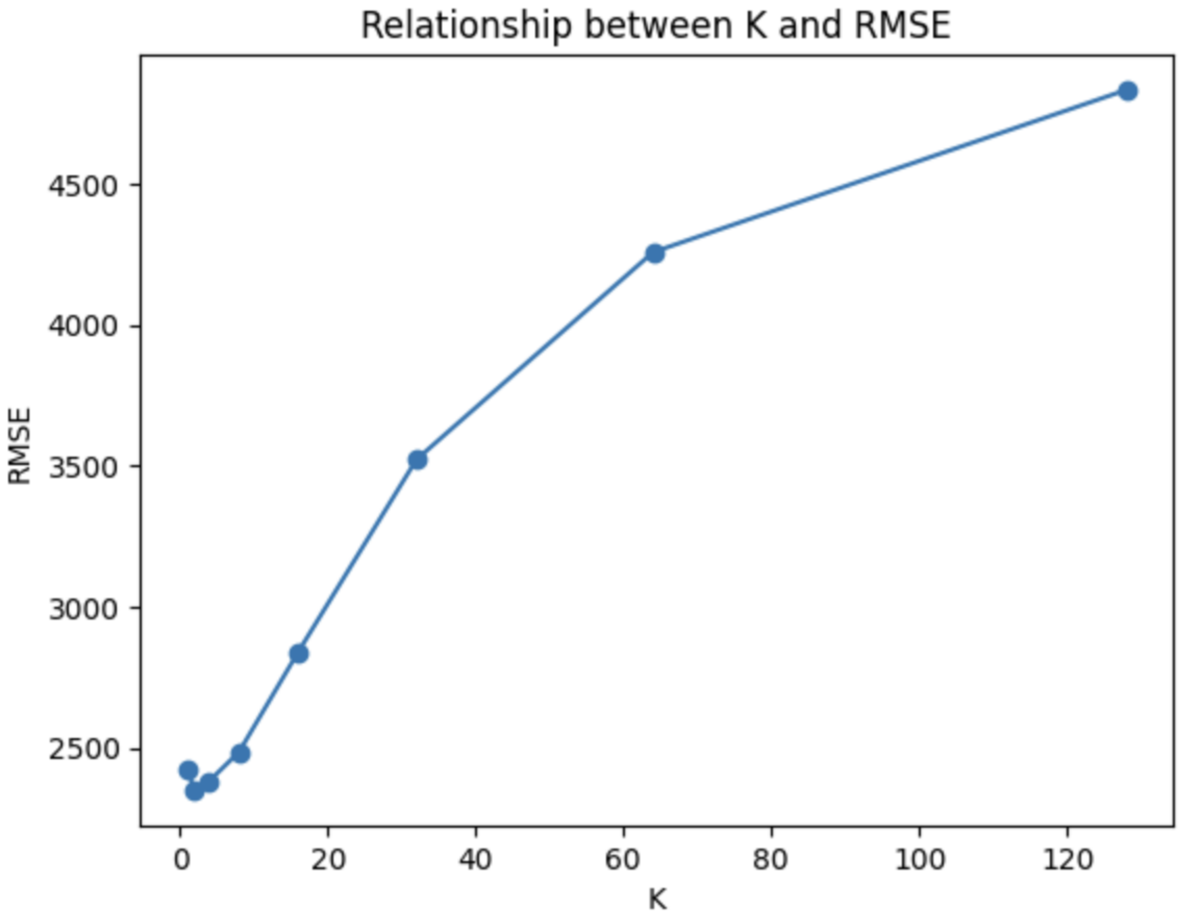

Figure 18: Tuning Hyperparameter for K Nearest Neighbor

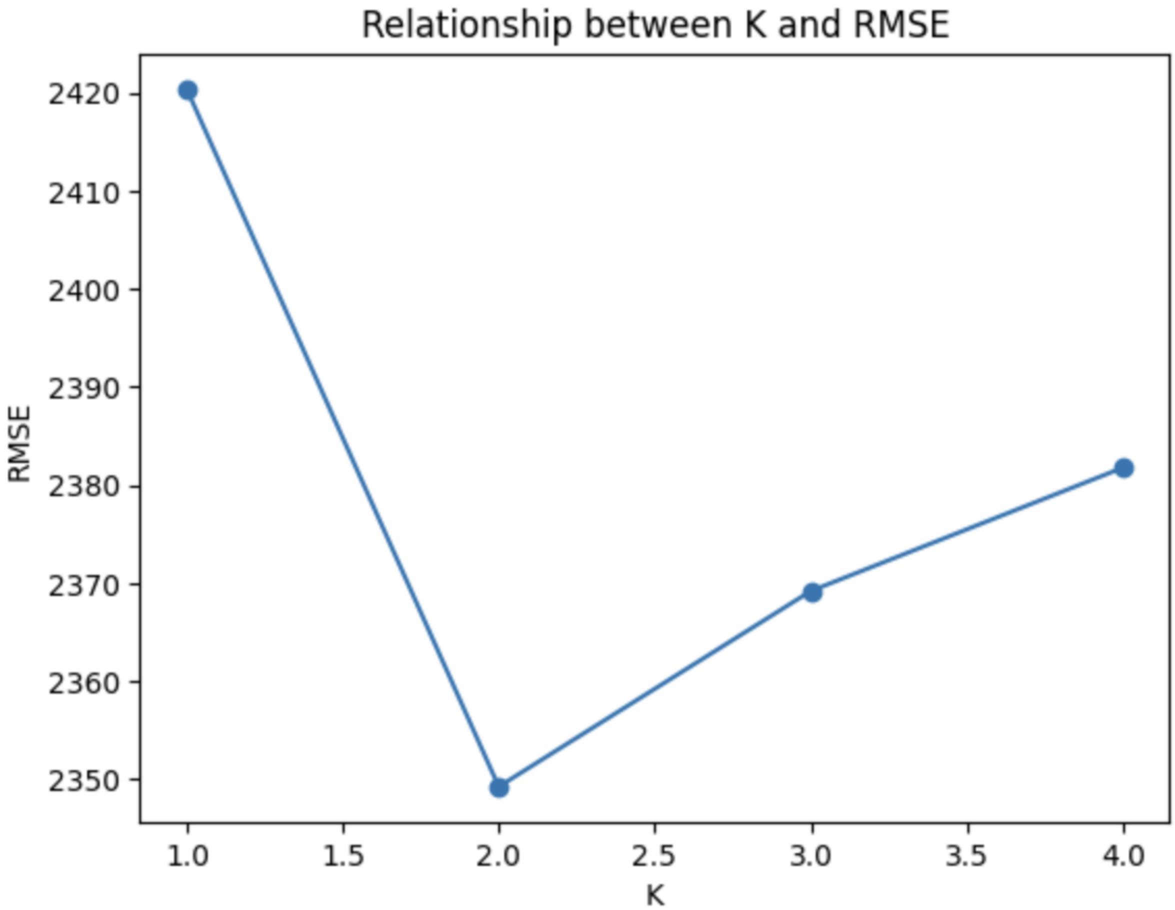

Figure 19: Tuning Hyperparameter for K Nearest Neighbor

From the visualizations above, we can see that error reaches its lowest when choosing two neighbors with an error of 2349.

The procedure of tuning epoch is shown below:

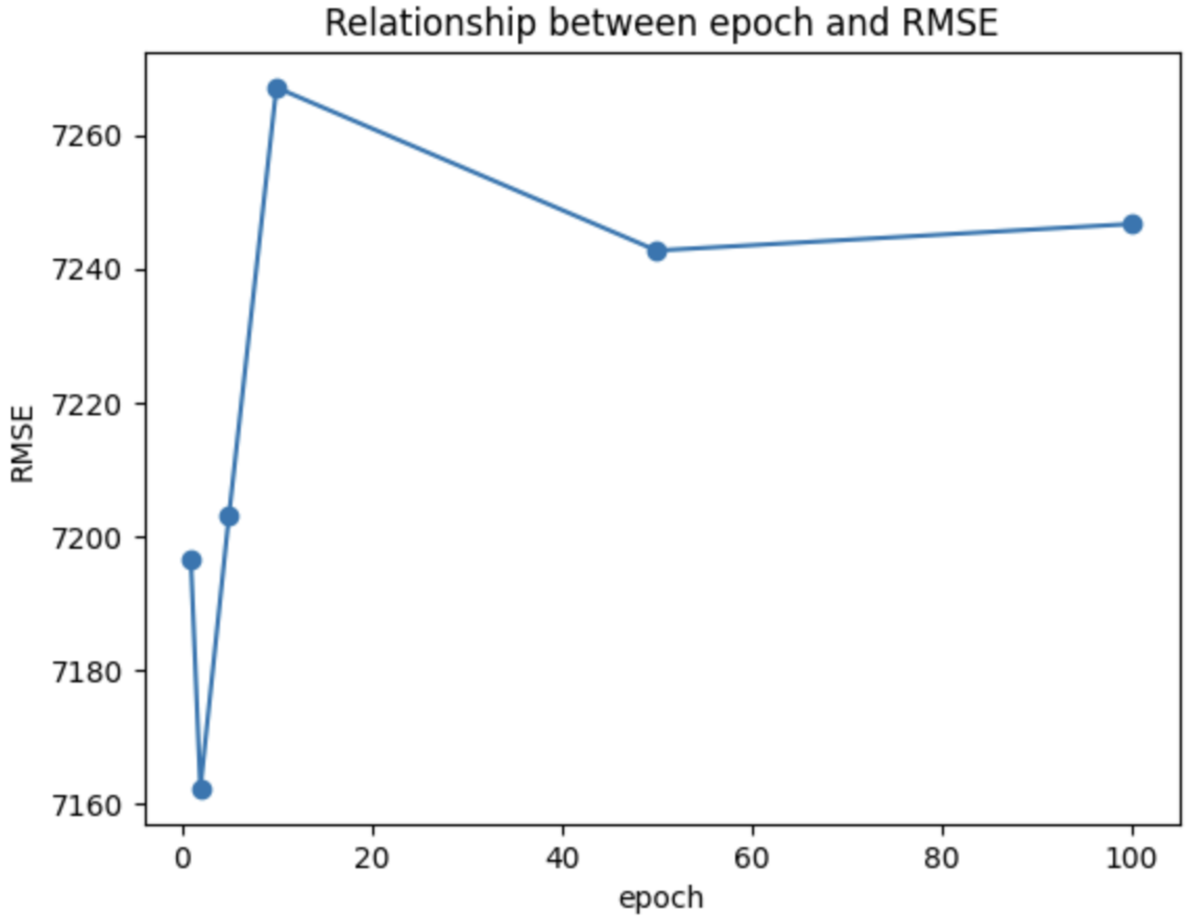

Figure 20: Tuning Hyperparameter for Multilayer Perceptron

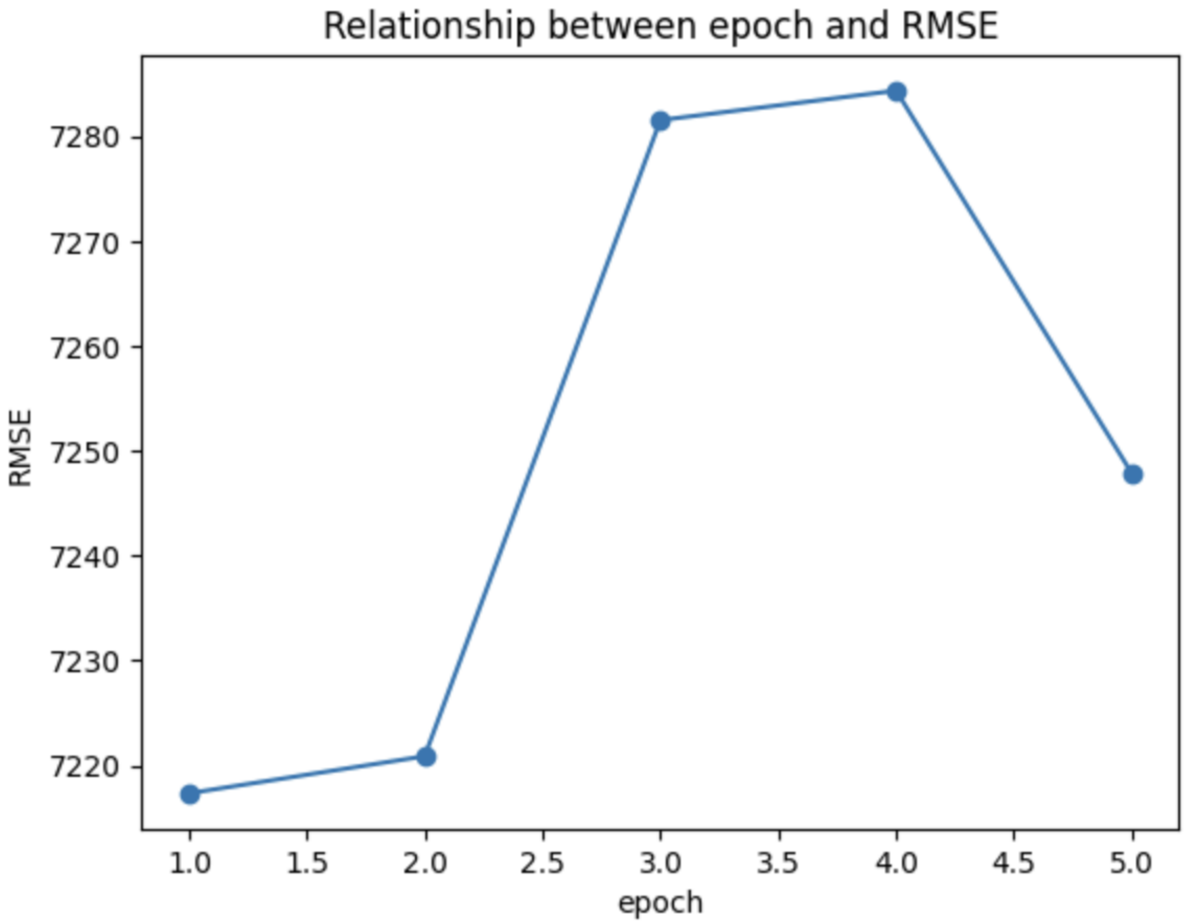

Figure 21: Tuning Hyperparameter for Multilayer Perceptron

From the visualizations above, we can see that error reaches its lowest when choosing one epoch with an error of 2420.

The procedure of tuning the number of estimators for Random Forest method is shown below:

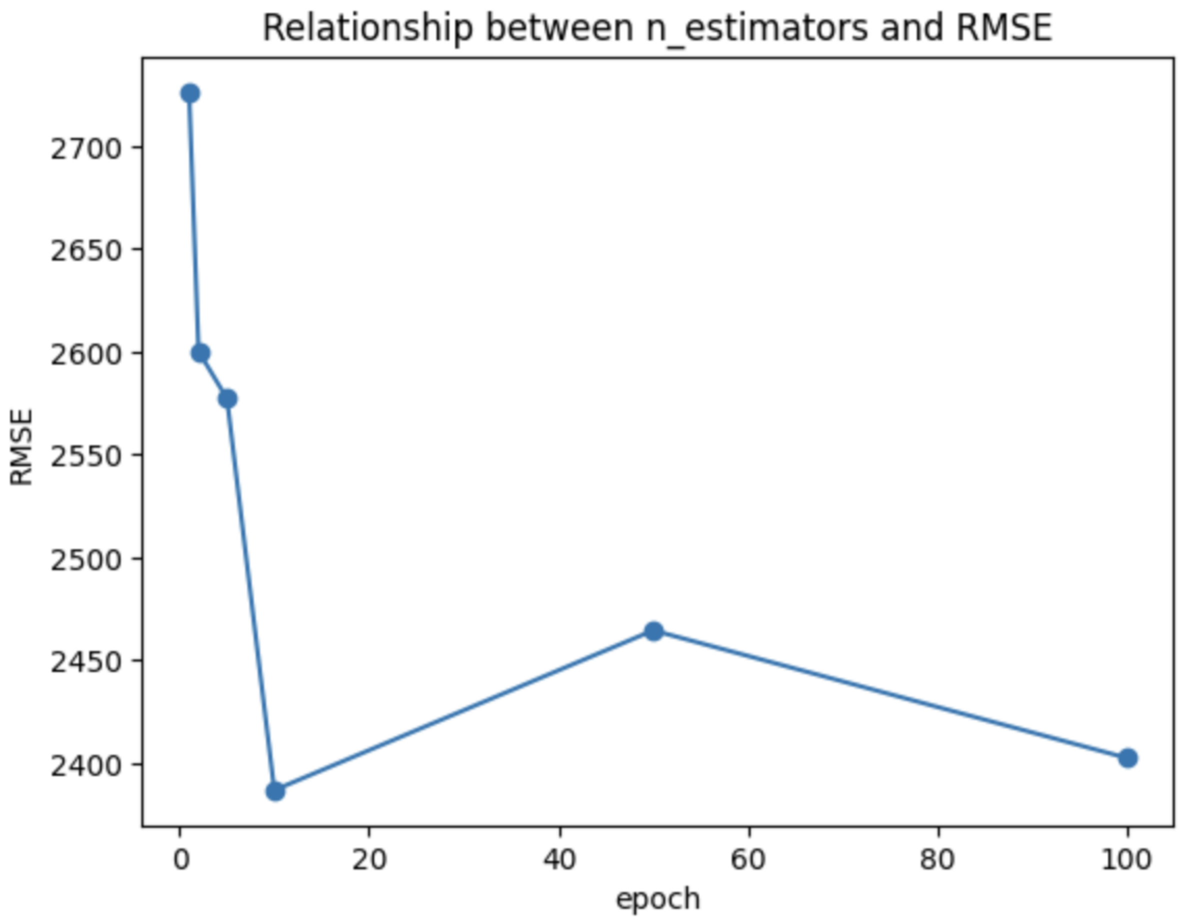

Figure 22: Tuning Hyperparameter for Random Forest

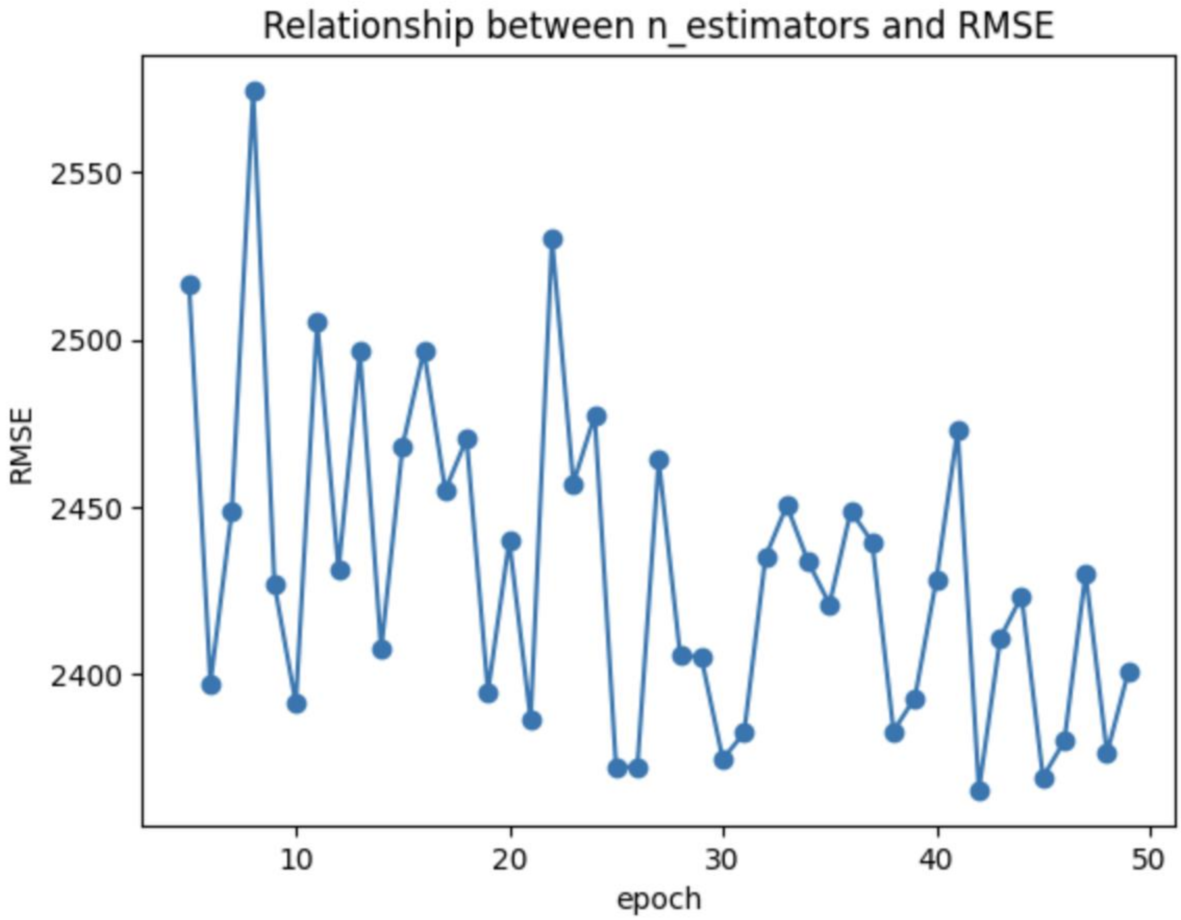

Figure 23: Tuning Hyperparameter for Random Forest

From the visualizations above, we can see that error reaches its lowest when choosing 42 estimators with an error of 2365.

After experimenting with linear regression method and analyzing parameters, we found that linearly taking advantage of all the submodels resulted in the optimal root mean squared error of 1940, which proves the effectiveness compared to the baseline day ahead market forecasting with an error of 6508.

With the help of hyperparameter tuning, we chose linear regression as our model to ensemble the five models we already have. We tested the model on the test dataset, which contained data from August 1st, 2023, to August 20th, 2023. The test result comparing the five submodels and the ensemble model is shown below:

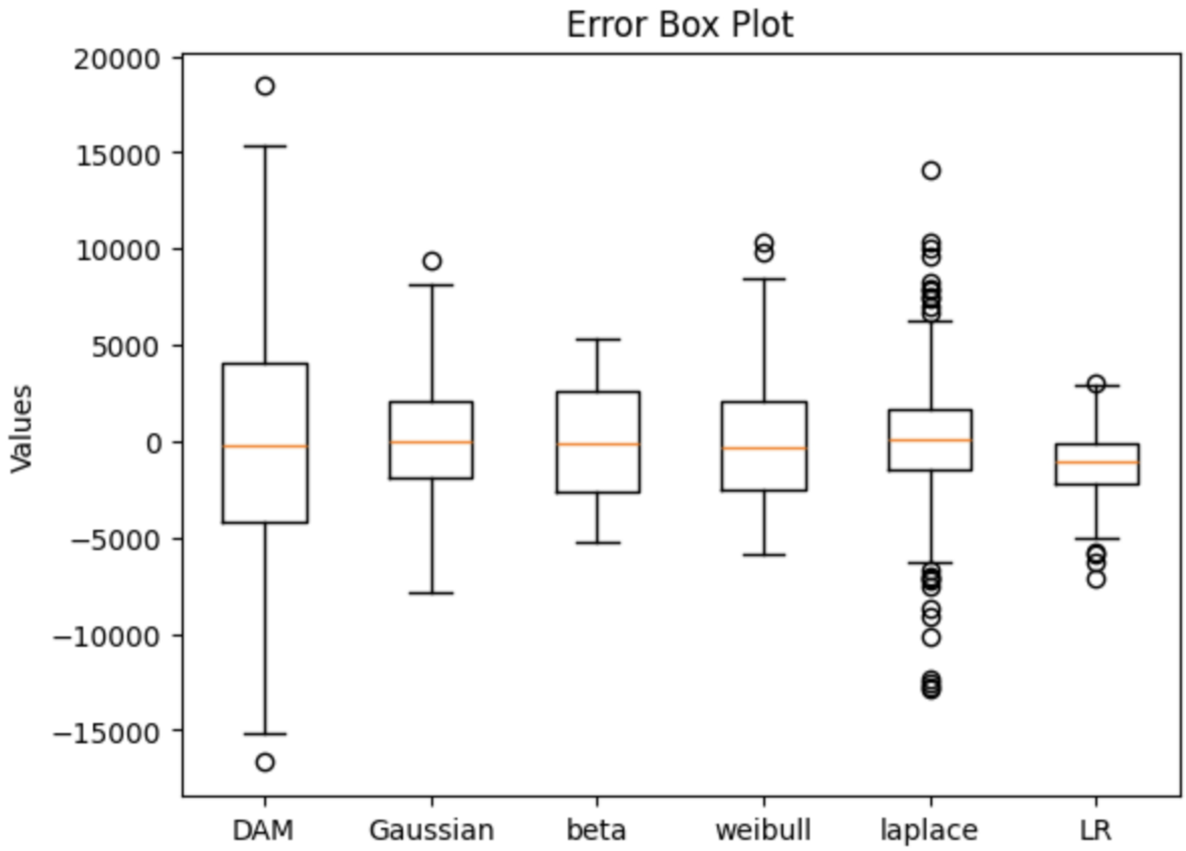

Figure 24: Error Plot for Linear Regression

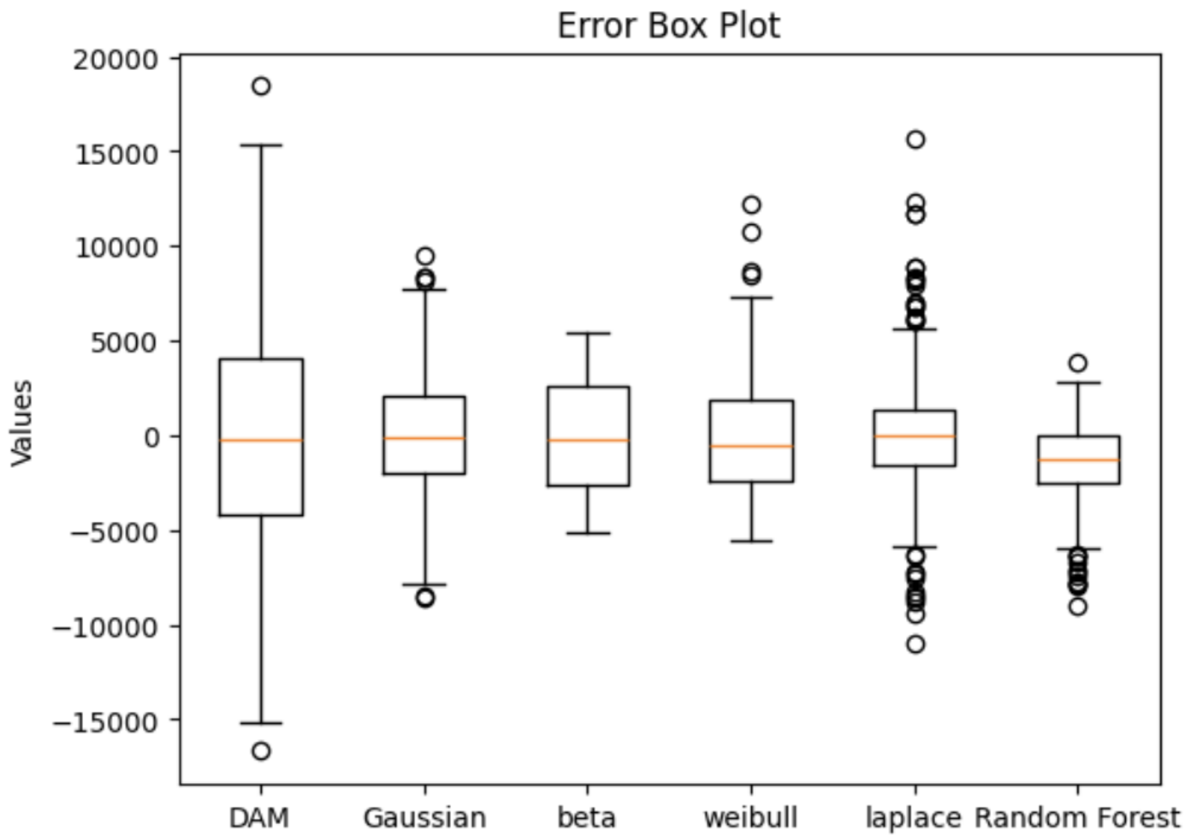

Figure 25: Error Plot for Random Forest

From the visualization above, we can conclude that using linear regression model to ensemble models decreased the root mean squared error, which also improves accuracy of prediction.

4. Summary and Conclusions

In this research project, we explore the whole procedure of deploying and optimizing machine learning models for increasing prediction accuracy.

We collected data from the public CAISO (California Independent System Operator) website and visualized them for a good control of the target prediction model, we also used data augmentation with different noise types to increase the data scale. Then we used different machine learning techniques like K Nearest Neighbor, Multilayer Perceptron, Random Forest and Linear Regression to ensemble the model. Finally, we compared the result of CAISO (California Independent System Operator) Forecasting model and the Ensemble model.

In conclusion, by experimenting with data processing and machine learning methods, we improved the accuracy of the power demand forecasting.

5. Suggestions for future work

For further improvement of the prediction model, experimenting with reinforcement learning methods will further decrease the error rate for power demand forecasting. Using the previous data for a certain time series to provide feedback for future predicting would provide help for a more accurate prediction.

Furthermore, to make the prediction more profitable and commercialize, combining with data markets is a possible method to add to meaningfulness for the prediction model.

As the experiment is set up in a controlled environment, it is also necessary to use the technique on real world CAISO (California Independent System Operator) models to help predicting demand forecast.

6. Bibliography

[1] D. S. Kirschen, G. Strbac, Fundamentals of Power System Economics, Willey, 2nd ed., July 2018
[2] F. Petropoulos, et. al. Forecasting: theory and practice. International Journal of Forecasting, vol. 38, no. 3 Sept. 2022, pp. 705-871.
[3] T. Hong, et. al., Energy Forecasting: A Review and Outlook, IEEE Open Access Journal of Power and Energy, vol. 7, Oct 2020, pp. 376-388
[4] T. Hong, S. Fan, Probabilistic electric load forecasting: A tutorial review, International Journal of Forecasting, vol. 32, no. 3, Sept. 2016, pp 914-938
[5] C. Sweeney, et. al. The Future of Forecasting for Renewable Energy, Wiley Interdisciplinary
Reviews: Energy and Environment, 2020
[6] J. Jox, et. al. Forecasting and Market Design Advances: Supporting an Increasing Share of Renewable Energy, IEEE Power and Energy Magazine, vol. 19, no. 6, Dec. 2021, pp. 77-85
[7] J. R. Andrade, R. J. Bessa, Improving Renewable Energy Forecasting With a Grid of Numerical Weather Predictions, IEEE Transactions on Sustainable Energy, vol. 8, no. 4, Oct. 2017, pp. 1571-1580