Load Forecasting

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Electrical and Electronics Engineering

by

THAVANESH R

21BEE0080

Under the guidance of

Dr. S. Meikandasivam

School of Electrical Engineering

VIT, Vellore.

November, 2024

DECLARATION

I here by declare that the thesis entitled "**Load Forecasting**" submitted by me, for the award of the degree of Bachelor of Technology in Electrical and Electronics Engineering to VIT is a record of Bonafide work carried out by me under the supervision of **Dr. S Meikandasivam**, Professor, School of Electrical Engineering, VIT, Vellore.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date: 22-11-2024

Raghavane

Signature of the Candidate

CERTIFICATE

This is to certify that the thesis entitled **"Load Forecasting"** submitted by **THAVANESH R (21BEE0080)**, School of Electrical Engineering, VIT, Vellore, for the award of the degree of Bachelor of Technology in Electrical and Electronics Engineering, is a record of Bonafide work carried out by him under my supervision during the period, 13.07.2024 to 20.11.2024, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Vellore

Date:22-11-2024

Meikandasivam 25/11/24

Signature of the Guide

Dr. S Meikandasivam

24/11/24

Head of the department - EEE

Head of the Department
Department of Electrical Engineering
School of Electrical Engineering (SELECT)
Vellore Institute of Technology (VIT)
Deemed to be University under section 3 of the UGC Act, 1956
Vellore – 632 014, Tamil Nadu, India

Dean SELECT

Dean
School of Electrical Engineering (SELECT)
Vellore Institute of Technology (VIT)
(Deemed to be University under section 3 of UGC Act, 1956)
Vellore – 632 014, Tamil Nadu, India

ACKNOWLEDGEMENT

With immense pleasure and deep sense of gratitude, I wish to express my sincere thanks to my supervisor **Dr. S Meikandasivam**, Professor, School of Electrical Engineering, VIT, without his motivation and continuous encouragement, this project work would not have been successfully completed.

I am grateful to the Chancellor of VIT, **Dr.G.Viswanathan**, the Vice Presidents, the Vice Chancellor for motivating me to complete the project.

I express my sincere thanks to **Dr.Mathew Mithra Noel**, Dean, School of Electrical Engineering, VIT for his kind words of support and encouragement. I like to acknowledge the support rendered by my colleagues in several ways throughout my project.

I wish to extend my profound sense of gratitude to my parents for all the sacrifices they made during my project and also providing me with moral support and encouragement whenever required.

Place: Vellore

Date: 22-11-2024

Thavanesh R

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

B.Tech- Electrical and Electronics Engineering

Vision and Mission Institute:

Vision:

Transforming life through excellence in education and research.

Mission:

- World class Education: Excellence in education, grounded in ethics and critical thinking, for improvement of life.
- Cutting edge Research: An innovation ecosystem to extend knowledge and solve critical problems.
- Impactful People: Happy, accountable, caring, and effective workforce and students.
- Rewarding Co-creations: Active collaboration with national & international industries & universities for productivity and economic development.
- Service to Society: Service to the region and world through knowledge and compassion.

- **Department:**

Vision:

To offer an education in electrical engineering that provides strong fundamental knowledge, skills for employability, cross-disciplinary research and creates leaders who provide technological solutions to societal and industry problems.

Mission:

M1 Provide personalized experiential learning in industry sponsored laboratories to prepare students in electrical engineering with strong critical thinking and employability skills.

M2 Foster design thinking, creativity and cross-disciplinary research with highly qualified faculty to create innovators and entrepreneurs in the broad area of electrical engineering.

M3 Collaborate with national and international partners to provide innovative solutions to societal and industry challenges.

EXECUTIVE SUMMARY

This project focuses on the development of a machine learning-based predictive model for real-time power consumption analysis using transformer data recorded at one-minute intervals. The primary goal was to create an efficient and accurate system for short-term power prediction, which could assist in proactive energy management and load balancing. Through detailed analysis and experimentation, linear regression emerged as the most effective model for predicting power consumption, demonstrating high accuracy for minute-level predictions and decent performance for hourly predictions. This was attributed to the strong correlation observed between consecutive real power readings.

Traditional time series models such as ARIMA and SARIMAX were explored as part of the study but proved less suitable due to the dataset's unique characteristics, which relied heavily on immediate dependencies rather than long-term trends or seasonal patterns.

By leveraging the high correlation between consecutive power readings, the project utilized linear regression to predict real power for the next minute with exceptional precision. The model was also extended to predict hourly values, achieving reasonable accuracy. The results highlight the practical applicability of linear regression for short-term power forecasting, especially when dealing with real-time data streams.

This project serves as a foundation for further advancements, such as integrating external factors (e.g., weather conditions, time of day) or applying more complex machine learning models for longer time-range predictions. Additionally, the live integration of this model with a data server offers real-world applicability for monitoring and optimizing energy usage. Overall, this work demonstrates a robust, scalable, and highly accurate approach to real-time power prediction, paving the way for smarter energy management solutions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TERMS AND ABBREVIATIONS

ML	**-** Machine Learning
AI	**-** Artificial Intelligence
ARIMA	**-** Auto Regressive Integrated Moving Average
SARIMA	**-** Seasonal Auto Regressive Integrated Moving Average.

ABSTRACT

This project focuses on predicting real-time power consumption using minute-level data from transformer readings. The goal was to create an accurate and reliable model to forecast short-term power usage, aiding in efficient energy management and load forecasting. By analyzing the dataset, a high correlation was identified between consecutive power readings, leading to the implementation of a linear regression model. This approach proved highly accurate for minute-level predictions and satisfactory for hourly forecasts after resampling the data.

Traditional time series models like ARIMA and SARIMAX were explored as alternatives. However, ARIMA failed to fit the dataset effectively, and while SARIMAX captured seasonality, its predictions were not sufficiently accurate. Consequently, linear regression emerged as the optimal solution for this specific dataset.

The project successfully demonstrates the potential of using linear regression for real-time power prediction, especially for minute-level accuracy. The developed model has been integrated into a framework for continuous monitoring and forecasting, paving the way for improved energy utilization and informed decision-making in load management systems. This work highlights the importance of tailored predictive approaches for datasets with unique characteristics.

CHAPTER 1

INTRODUCTION

Load forecasting is the process of predicting how much electricity will be needed at a given time and how that demand will affect the utility grid. It is used to ensure that enough power is available to meet consumption needs while avoiding waste and inefficiency.

Electric load forecasting is key to the operational planning of power systems, and crucial for avoiding outages. Load forecasting predictions can range from short-term (hours or days ahead) to long-term (months or years ahead). The accuracy of these forecasts directly impacts the cost and reliability of the entire power system. Load forecasting is also a component of broader energy forecasting, which includes predictions for the availability and pricing of fuels such as oil and gas, as well as renewable energy sources.

Accurate load forecasting ensures there is enough electric power supply to meet demand at any given time, thereby maintaining the balance and stability of the power grid. With that reliability comes greater efficiency as well as cost savings. Load forecasting allows utilities to better manage their resources through demand response programs, which shift usage by incentivizing consumers to reduce their electricity use during high-usage times. And this kind of demand forecasting can help utilities avoid the extra costs associated with producing too much or too little electricity.

Load forecast data may also be used in strategic planning decisions such as capacity expansion, infrastructure development and maintenance scheduling. For example, this data can highlight the optimal location of new power plants or transmission lines, ensuring that future demand can be met. In deregulated electricity markets, load forecasting data can also help market participants make informed bidding strategies, manage energy contracts and mitigate risks.

There are several methods used in load forecasting, each of which analyzes historical load data and other relevant inputs to generate forecasts for different time horizons.

1.1 Short-term load forecasting

This covers a period up to a week and relies significantly on weather forecasts and recent load data. Short-term load forecasting, including day-ahead predictions, is particularly important for managing the power grid in real time, as it allows system operators to make decisions in the moment about how much power to generate and where to direct it. Accuracy is crucial in this context, as even small errors in forecasting can lead to wasted energy or overloaded power lines

1.2 Medium-term load forecasting

This ranges from a week to a year and is used for maintenance scheduling and fuel reserve management. It considers seasonal variations in electricity consumption as well as planned outages.

1.3 Long-term load forecasting

This typically covers a period of more than one year and considers factors such as demographic changes, economic growth and energy policy impacts. Long-term load forecasting focuses on system planning and optimization, helping utilities to make decisions about where to invest in new power generation capacity and how to balance different sources of energy, such as renewable energy and traditional fossil fuels.

Load forecasting methods begin with historical load data collection. This includes data from the many factors that can affect electricity use, including weather data (temperature, humidity, wind speed), time of day, calendar variables (seasons, holidays, weekday versus weekend) and demographic factors (population density, economic activity). Load forecasting takes all of these data sets into account to create a comprehensive picture of energy demand.

Once data is collected, a forecasting model is developed. Some examples of models used for load forecasting include:

- **Regression models**: Linear regression models are often used for long-term load forecasting. They relate the load demand to variables like weather conditions and economic indicators.
- **Time series models**: Autoregressive Integrated Moving Average (ARIMA) and similar models are popular for short-term load forecasting. They rely on past load data to predict future demand.
- **Artificial Intelligence (AI) models**: Neural networks and support vector machines are increasingly used due to their ability to model complex non-linear relationships. Deep learning model can further improve forecasting accuracy by automatically extracting relevant features from the dataset.

The forecasting model is trained using a portion of the historical data and tested for validation. Performance metrics such as Mean Absolute Percentage Error (MAPE) are used to evaluate the accuracy of the forecasts.

Once the model is validated and fine-tuned, it can generate future load forecasts. These forecasts can then be used for operational planning, energy management and other decision-making activities. This is an ongoing and adaptive process: As new data becomes available, the models usually require updates or retraining to remain accurate.

Load forecasting can be valuable, but it has its limitations. One major issue is the increasing complexity of the power grid, which now includes distributed energy resources (DERs) such as solar panels and electric vehicles. These resources can be difficult to predict and integrate into load forecasting models, requiring new methodologies and input features.

Another challenge is the need for accurate weather forecasting, as weather conditions can have a significant impact on energy demand. Improvements in weather forecasting technology have helped to address this issue, but there is still room for improvement.

1.4 Load Forecasting Challenges

- Forecasting the future needs for electricity is a difficult task.
- Electricity production and distribution are highly capital intensive.
- Projects are large and lead times are long.
- Forecasting can not be an isolated activity.
- Role of electrical energy in the society should be reflected Government policy and strategic decisions taken by utility are important factors.
- Forecasting should view that the future is open to the effects of many human actions.
- Uncertainties arise from the impact of the changes in public perceptions, viewpoints and policies.
- Demand Side Management and conservation policies give additional requirements on load forecasting.
- Precise forecasting is impossible.
- To tie future plans too rigidly to a single load forecast projection is too risky.
- By incorporating the role of uncertainty into the analysis techniques, the emphasis of planning moves from making an accurate forecast to constructing a system that can adapt readily to changes.

1.5 Factors Affecting Load Forecasting:

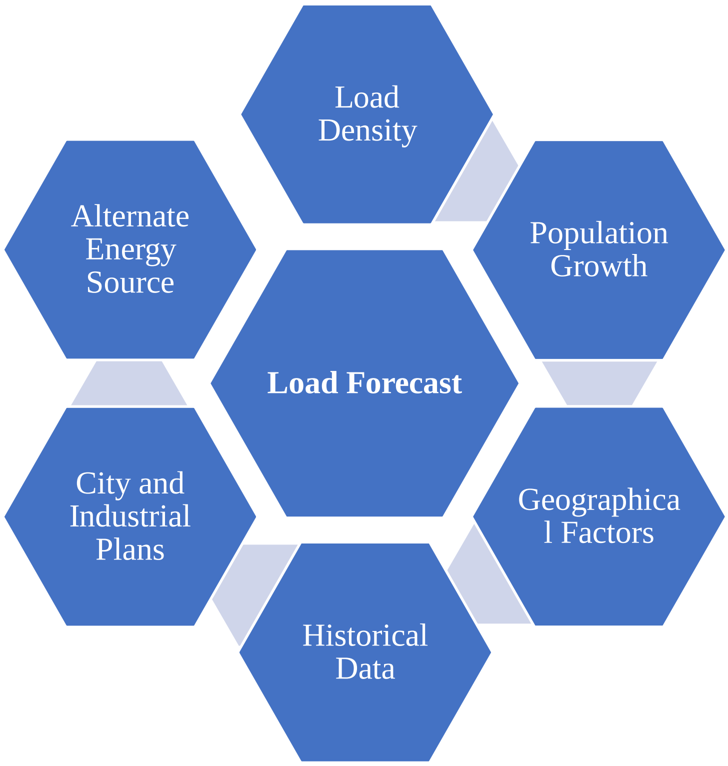

Figure 1: Factors affecting load forecasting

By enabling more efficient, flexible and intelligent power system operations, load forecasting is a critical sustainability tool. It can contribute to sustainability in several ways:

1.6 Renewable energy transitions

Accurate load forecasting is essential for integrating renewable energy sources like wind and solar power into the grid. These sources are intermittent, meaning their output depends on weather conditions and time of day. By accurately predicting electricity demand, utilities can better plan for fluctuations and maximize use. This can help reduce overall greenhouse gas emission by minimizing reliance on fossil fuel-based power generation.

1.7 Energy efficiency

Accurate forecasts allow electric utilities to operate their distribution systems more efficiently, based on daily or hourly load, which reduces energy waste and optimizes the overall energy supply. For instance, companies can use the information to schedule maintenance or other downtime for periods of lower demand.

1.8 Demand response programs

These programs incentivize people to reduce or shift their energy consumption during peak load times, helping to balance supply and demand without needing to bring additional, potentially less sustainable, generation sources online.

1.9 Grid modernization

Accurate load forecasting is crucial for smart, more flexible, and future energy systems. It will enable more sophisticated grid management strategies that can accommodate distributed energy resources, electric vehicles and other new technologies.

CHAPTER 2

DATASET

The dataset used in this project is a rich source of time series data derived from the readings of a specific transformer (referred to as Transformer-4) located on the college campus through Smart Meter. This dataset provides a minute-by-minute recording of electrical parameters over every month(from January 2024). The detailed data collection over this extended period offers a broad view into the transformer's performance and provides a valuable foundation for developing predictive models for load forecasting

Transformers are critical components in the electrical distribution system, converting high-voltage electricity from power stations into lower voltages suitable for home or industrial use. Monitoring transformer performance is crucial because it directly impacts the reliability of the power supply. Through precise data collection, utility companies can optimize their operations, enhance transformer efficiency, and prevent failures by predicting loads more accurately.

The dataset we are working with includes plenty of entries(expected 60*24*30=432000) every month(representing a full month of minute-level data), covering a variety of electrical parameters essential for understanding the behaviour of the transformer under different load conditions. However, it also includes missing entries, which could be filled using interpolation techniques or skipped to ensure continuity or less error/outliers in the time series analysis. These missing entries were identified and handled cautiously during the preprocessing phase, which is crucial for maintaining the integrity of the dataset.

The dataset was collected using advanced monitoring systems installed on Transformer. These systems were set up to automatically log electrical readings every minute. This high-resolution data collection allows for precise analysis and the development of models that can predict power usage on a minute-by-minute basis. The data was collected throughout every month in a year, a period chosen to capture typical variations in power consumption over the course of a month, including weekdays, weekends, midnights and potential anomalies related to special events or weather conditions and power outages.

The monitoring equipment used for data collection was calibrated to ensure accurate readings. Despite this, certain challenges were encountered, such as occasional connectivity issues, which resulted in gaps in the recorded data. These gaps can be addressed using interpolation methods to ensure the dataset remained continuous or skip it and usable for predictive modeling.

2.1 Structure of the Dataset:

The dataset is organized into a structured format with 28 columns, each representing a distinct parameter recorded by the monitoring system. Below is an in-depth explanation of each column, its relevance, and how it contributes to the overall analysis.

2.1.1 Datetime Column:

- **Description**: The Datetime column represents the timestamp at which each reading was taken. The format used is a combination of date and time (DD-MM-YYYY HH:MM:SS).
- **Importance**: This column is critical for time series analysis as it helps establish the chronological order of the data. By using the datetime information, we can analyze patterns, trends, and cycles within the dataset.
- **Usage**: In the predictive modeling, the Datetime column is used to create lag features, extract time-based patterns (such as hourly, daily, and weekly trends), and segment the data for training and testing the model.

The remaining columns in the dataset represent various electrical parameters measured by the transformer. These parameters include energy, voltage, current, power, and power quality metrics. Each of these columns provides critical insights into the transformer's operation.

2.1.2 RealEnergyWH (Real Energy in Watt-Hours):

- **Description**: This column measures the real energy consumption in watt-hours. It represents the amount of electrical energy actually consumed by the connected loads.
- **Relevance**: Real energy is a direct measure of the actual power usage over time. It helps in understanding consumption patterns and identifying peak load times.
- **Insights**: Analyzing real energy readings can help determine the efficiency of the transformer and identify periods of high demand. This data is useful for energy management and optimizing transformer performance.

2.1.3 ApparentEnergyVAH (Apparent Energy in Volt-Ampere Hours):

- **Description**: The apparent energy is the total energy drawn from the supply, measured in volt-ampere hours.
- **Significance**: Apparent energy includes both real power and reactive power components. It reflects the total electrical demand on the transformer, which is essential for understanding the overall load.
- **Impact on Load Forecasting**: While real energy measures actual consumption, apparent energy captures the full load on the system, including reactive components that do not contribute to work but still affect the transformer's capacity.

2.1.4 ReactiveEnergyVARHP and ReactiveEnergyVARHN:

- **Description**: These columns represent the reactive energy, measured in volt-ampere reactive hours (VARH). ReactiveEnergyVARHP captures positive reactive power, while ReactiveEnergyVARHN records negative reactive power.
- **Role in Electrical Systems**: Reactive energy is associated with inductive and capacitive loads that create magnetic and electric fields. While this type of energy does not perform useful work, it is essential for maintaining voltage levels and system stability.
- **Use Case**: Understanding reactive power flows helps optimize power factor and reduce energy losses. High reactive power can indicate inefficiencies that may need correction to avoid penalties from utility companies.

2.1.5 Voltage Measurements (Line-to-Line and Phase-to-Neutral):

The dataset includes several columns related to voltage measurements:

- **LineVoltageVRY, LineVoltageVYB, LineVoltageVBR**: These columns represent line-to-line voltages between different phases.
- **PhaseVoltageVRN, PhaseVoltageVYN, PhaseVoltageVBN**: These columns capture the voltage between each phase and neutral.
- **Importance**: Voltage stability is crucial for the proper functioning of electrical equipment. Deviations from nominal voltage levels can indicate potential issues such as overloads, phase imbalances, or faults in the transformer.

- **Insights for Analysis**: By monitoring voltage levels, we can identify trends that correlate with power consumption spikes. For instance, a drop in voltage may correspond to increased load, while surges could indicate potential faults.

2.1.6 Current Measurements (Line Currents):

- **Columns**: LineCurrentIR, LineCurrentIY, LineCurrentIB
- **Description**: These columns record the current flowing through each line of the transformer.
- **Relevance**: High currents can indicate potential overloading of the transformer. Monitoring current is essential for ensuring the transformer operates within its rated capacity.
- **Predictive Value**: Current data can be used alongside voltage measurements to calculate real, reactive, and apparent power. Analyzing current trends helps in understanding the load distribution across different phases.

2.1.7 Power Measurements (Real, Reactive, and Apparent Power):

The dataset includes columns that capture the different types of power:

- **Real Power**: Represents the active power consumed by the load.
- **Reactive Power**: The power associated with the magnetic and electric fields in inductive and capacitive loads.
- **Apparent Power**: The total power drawn from the supply, combining both real and reactive components.
- **Importance for Load Forecasting**: The Real Power column is the primary target for our predictive model. Real power is crucial for load management and predicting future consumption, while reactive and apparent power help in understanding power quality and efficiency.

2.1.8 Power Factor:

- **Column**: PowerFactor
- **Description**: The power factor is the ratio of real power to apparent power. It indicates how effectively the transformer is converting electrical power into useful work.

- **Significance**: A low power factor indicates inefficiencies, leading to higher operational costs and potential penalties from utility companies. Improving the power factor can reduce energy losses and enhance system performance.

2.1.9 Frequency:

- **Column**: Frequency
- **Description**: This column captures the frequency of the electrical supply, typically around 50 Hz or 60 Hz depending on the region.
- **Role in Stability**: Maintaining a stable frequency is crucial for the reliability of electrical systems. Deviations from the nominal frequency can indicate instability in the grid or transformer load issues.
- **Insights**: Frequency variations can be an early indicator of potential faults or load imbalances.

2.2 Snapshot of Dataset:

Datetime	RealEnergyWH	ApparentEnergyVAH	ReactiveEnergyVARHP	ReactiveEnergyVARHN	LineVoltageVRY	LineVoltageVYB	LineVoltageVBR	LineCurrentIR	LineCurrentIY	LineCurrentIB	RealPower	Rea
01-03-2024 00:00	803720	856868	13718	186357	417.38	413.95	423.2	232.18	238.15	247.81	150171.2	
01-03-2024 00:01	803723	856871	13718	186359	417.72	414.96	422.98	226.06	226.04	245.44	143366.31	
01-03-2024 00:02	803726	856874	13718	186360	416.94	416.18	422.29	229.96	235.24	252.48	151830.94	
01-03-2024 00:03	803727	856876	13718	186361	416.94	416.37	422.64	238.08	240.07	247.12	154914.55	
01-03-2024 00:04	803730	856879	13718	186363	416.85	415.41	423.34	237.23	246.5	249.04	157115.88	
01-03-2024 00:05	803733	856882	13718	186364	417.26	416.34	423.01	263.34	260.68	276.51	176814.27	
01-03-2024 00:06	803735	856885	13718	186365	417	416.66	423.86	239.53	237.54	247.73	154731.11	
01-03-2024 00:07	803738	856888	13718	186367	417.21	416.47	423.26	236.62	228.03	246.66	151210.73	
01-03-2024 00:08	803741	856891	13718	186368	416.86	414.75	422.57	236.16	238.99	253.24	156565.55	
01-03-2024 00:09	803743	856894	13718	186369	417.15	415.01	422.5	238.54	239.3	258.14	157849.66	
01-03-2024 00:10	803746	856897	13718	186371	416.53	415.09	422.35	238.31	240.99	254.85	154844.67	
01-03-2024 00:11	803749	856900	13718	186372	416.73	415.69	421.79	235.7	236.92	245.36	151009.81	
01-03-2024 00:12	803751	856903	13718	186374	416.19	415.55	421.93	235.32	231.02	239.7	147751.5	
01-03-2024 00:13	803754	856906	13718	186375	416.15	415.63	421.93	235.86	227.34	238.78	147096.34	
01-03-2024 00:14	803757	856909	13718	186377	416.8	415.98	422.13	227.97	229.87	243.22	146947.84	
01-03-2024 00:15	803758	856911	13718	186378	416.61	413.96	422.08	236.54	232.02	244.75	150887.52	
01-03-2024 00:16	803761	856914	13718	186379	416.04	412.49	421.9	240.68	239.76	248.5	156085.09	
01-03-2024 00:17	803764	856917	13718	186380	416.2	412.89	422.13	241.6	243.36	243.6	155639.59	
01-03-2024 00:18	803766	856920	13718	186382	415.58	414.15	421.15	237.39	241.29	246.43	155246.5	
01-03-2024 00:19	803769	856923	13718	186383	415.67	415.83	421.48	238.15	233.17	241.77	150494.42	
01-03-2024 00:20	803771	856926	13718	186385	415.93	415.31	421.75	244.74	243.44	235.72	154608.81	
01-03-2024 00:21	803774	856929	13718	186386	416.33	413.39	422.2	243.74	240.45	239.01	154722.36	
01-03-2024 00:22	803777	856932	13718	186387	416.39	416.48	422.27	232.79	238.53	231.43	149603.41	
01-03-2024 00:23	803779	856934	13718	186388	416.66	415.81	422.4	236.16	235.7	239.7	152110.48	
01-03-2024 00:24	803781	856937	13718	186390	415.74	414.61	421.19	237.46	235.62	250.26	156111.3	
01-03-2024 00:25	803784	856940	13718	186391	415.8	414.73	421.54	244.51	250.11	248.81	162278.52	

Figure 2: Dataset

With this high quality and informative dataset, there are many possible predictions that can be performed with different set of features as follows:

2.3 Energy Predictions:

- **RealEnergyWH**: Represents the real (active) energy consumption measured in watt-hours (WH). This is the energy actually consumed by the load.
- **ApparentEnergyVAH**: Indicates the apparent energy, measured in volt-ampere-hours (VAH). It represents the combination of real and reactive power in the system.
- **ReactiveEnergyVARHP**: Measures the positive reactive energy in volt-ampere reactive hours (VARH). Reactive energy is associated with the magnetic and electric fields in inductive loads.
- **ReactiveEnergyVARHN**: Represents the negative reactive energy measured in volt-ampere reactive hours (VARH), which can occur in systems with capacitive loads.

2.4 Voltage and Current Predictions:

- **LineVoltageVRY, LineVoltageVYB, LineVoltageVBR**: These columns provide the line-to-line voltages between the phases R-Y, Y-B, and B-R, respectively. Line voltage readings are crucial for assessing the health and stability of the transformer.
- **LineCurrentIR, LineCurrentIY, LineCurrentIB**: Represents the current flowing through each of the lines R, Y, and B. Monitoring line currents helps in identifying potential overloads and ensuring balanced load distribution.

2.5 Power Predictions:

- **RealPower**: This is the active power consumed by the system, measured in watts (W). Real power is the primary metric used in this project for prediction due to its direct impact on load management.
- **ReactivePower**: The reactive power in the system, measured in volt-amperes reactive (VAR). It reflects the power required to maintain the electric and magnetic fields in inductive loads.
- **ApparentPower**: Represents the total apparent power, measured in volt-amperes (VA). It combines both the real and reactive components of power.

2.6 Phase-Specific Predictions:

- **PhaseVoltageVRN, PhaseVoltageVYN, PhaseVoltageVBN**: These columns measure the phase-to-neutral voltages for each of the phases R, Y, and B. Phase voltages are critical for assessing the quality of the power supply.
- **RealPowerR, RealPowerY, RealPowerB**: The real power measured individually for each of the phases R, Y, and B.
- **ReactivePowerR, ReactivePowerY, ReactivePowerB**: The reactive power for each phase.
- **ApparentPowerR, ApparentPowerY, ApparentPowerB**: The apparent power measurements per phase.

For the load forecasting task, only the Datetime and RealPower columns were utilized. The focus was on predicting the Real Power for the next minute based on historical minute-level data. This was motivated by the high correlation found between consecutive minute readings of Real Power.

The decision to use Real Power for prediction was driven by its significance in real-world applications of load forecasting. Real power, being the actual power consumed by the load, is critical for operational decision-making in power systems. Accurate predictions of Real Power enable more efficient load management and optimization of power distribution.

2.7 The Characteristic of the Real Power:

	RealPower
count	42813.000000
mean	172384.239080
std	76569.310159
min	0.000000
25%	123545.630000
50%	159282.270000
75%	229078.310000
max	437426.970000

Figure 3: Characteristic of the Real Power

2.8 The Plot of the Real Power:

- For March 2024:

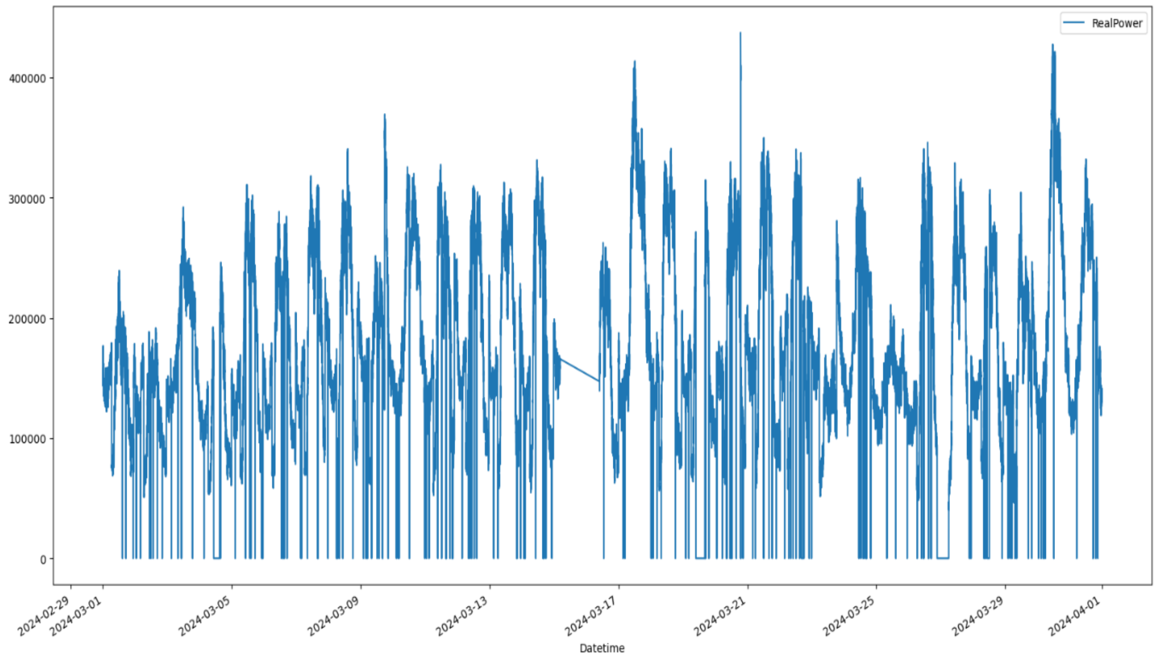

Figure 4: Real power for march 2024

CHAPTER 3

PROCESS AND METHODS

3.1 The Flow chart of the process:

Figure 5: Process of ML

The entire process can be broken down into various major steps:

3.1.1 Collecting Data:

The data collection process is fundamental to any machine learning project, as it lays the foundation for model training and accurate predictions. To access the dataset, I utilized pandas' pd.read_csv() function to read the files into Data Frames for further analysis. Each file represented a month's worth of minute-level data, capturing key electrical parameters like real power, voltage, and current readings. By leveraging pandas, I was able to efficiently handle these large datasets and seamlessly import them into my Python environment. This approach ensured that I could access, view, and manipulate the data quickly while preserving the accuracy of the recorded minute-level timestamps. Thus, using pandas not only streamlined the data collection process but also laid the groundwork for subsequent preprocessing and analysis steps.

To obtain only march month data:

```
df = pd.read_csv("C:/Users/Thavanesh/Downloads/TT-TFR-4.csv/TT-TFR-4-MAR.csv")
```

To obtain data from different dataset:

```
df1 = pd.read_csv("C:/Users/Thavanesh/Downloads/TT-TFR-4.csv/TT-TFR-4-MAR.csv")
df2 = pd.read_csv("C:/Users/Thavanesh/Downloads/TT-TFR-4.csv/TT-TFR-4-APR.csv")
df3 = pd.read_csv("C:/Users/Thavanesh/Downloads/TT-TFR-4.csv/TT-TFR-4-MAY.csv")
df=pd.concat([df1,df2,df3])
```

3.1.2 Preparing the Data and Selecting the Model:

After loading the data, it's essential to inspect its structure and understand its contents.

- head() method shows the first five rows of the dataset, giving an initial look at the data to verify its structure.
- The info() method provides a summary of the dataset, including column names, data types, and the presence of any missing values. This helps to understand the structure and completeness of the data.

  This step is crucial for identifying data types and potential data quality issues, such as missing or incorrect entries, which can affect further analysis.

In our case, we are interested in the columns Datetime and RealPower, so we filtered the DataFrame to keep only these columns.

- By focusing only on the Datetime and RealPower columns, we eliminate irrelevant data, reducing memory usage and making subsequent processing faster.
- Streamlining the dataset ensures that only relevant data is used in the predictive model, improving its efficiency and accuracy.

The Datetime column is currently stored as a string, so to convert it to a proper datetime object for easier manipulation.

- pd.to_datetime() converts the Datetime column into a datetime format, allowing to perform time-based operations.
- Setting the Datetime column as the index with set_index() makes it easier to resample the data and perform time series analysis.
- Time series analysis requires the data to be indexed by time. Converting strings to datetime objects and setting them as the index is crucial for using time-based functions like resampling, rolling windows, and seasonal decomposition.

Before proceeding with resampling or filling in missing data, it's helpful to explore the distribution of RealPower.

- df.describe() provides statistical insights, such as mean, median, standard deviation, and range, which help identify any anomalies or outliers in the dataset.
- Plotting the data helps to visually inspect trends, seasonality, and any irregularities.
- Understanding the distribution and behaviour of the data helps to make informed decisions about handling missing values, outliers, and transformations.

To have enough data for predicting hourly power consumption, we decide to combine data from multiple months (February, March, April, and May).

- pd.concat() is used to combine multiple DataFrames into a single one.
- This step is necessary to ensure that we have a sufficiently large dataset for training the model.
- Time series models perform better when trained on larger datasets, especially for capturing patterns over longer periods.

After combining data from multiple months, some timestamps might be missing. To address this, we reindex the DataFrame and fill in the gaps using forward fill.

- pd.date_range() creates a complete list of timestamps, ensuring there are no gaps.
- reindex() aligns the dataset with this complete list, introducing NaNs for missing entries.
- fillna(method='ffill') fills missing values by carrying forward the last known value.
- Handling missing data is crucial for maintaining the integrity of time series models, as gaps in data can lead to incorrect predictions or model failures.

The initial dataset is recorded at a minute-level frequency, but we want to predict power consumption on an hourly basis. Resampling the data helps in reducing noise and making predictions more manageable.

- resample('H') changes the frequency of the data to hourly intervals.
- df.boxplot() visualize range of outliers to remove those from data, reducing the noise of the data and indicates the skewness of the data, in our case it is positively skewed data and so we transformed the original to square root of the original data to reduce the skewness of the data and after prediction we can re-transform it to original scale by squaring it.
- Taking the mean of RealPower for each hour smoothens short-term fluctuations, providing a more stable signal for modeling.

- Hourly data is often more suitable for forecasting tasks, especially for applications like load prediction, where minute-level noise might obscure the underlying trend.

Finally, we perform seasonal decomposition to analyze the patterns in our data.

- Seasonal decomposition helps identify the trend, seasonal, and residual components, which can improve forecasting accuracy.

After all the preprocessing, it's essential to inspect the final processed dataset to ensure everything is correct

- This step confirms that the dataset is ready for the next stage, which is model building and training.
- A clean and well-prepared dataset is essential for building reliable and accurate predictive models.

With the dataset ready, we can move on to model training. we might start by experimenting with different models, such as linear regression, ARIMA, or machine learning models like decision trees and neural networks, depending on the complexity of the dataset.

Time series forecasting is a powerful tool for predicting future values based on historical data patterns. Among the collection of techniques, ARIMA (Autoregressive Integrated Moving Average) stands out as a fundamental model for capturing temporal dependencies. SARIMA (Seasonal ARIMA) and SARIMAX extends ARIMA to handle seasonal variations and to incorporate exogenous variables, respectively.

- **ARIMA (AutoRegressive Integrated Moving Average):**

A classic time series forecasting model that captures non-seasonal patterns and trends in the data. ARIMA models are particularly used for stationary time series data, where the mean and variance remain constant over time. It consists of 3 main components:

- *AutoRegressive (AR)* — captures the relationship between the current observation and its lagged values (previous time steps)
- *Integrated (I)* — involves differencing, which makes the time series data stationary by removing trends and seasonality

- *Moving Average (MA)* — represents the relationship between the current observation and residual errors from a moving average model applied to lagged observations, helping capture short-term fluctuations or noise in the data

- **SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables)**

This model extends SARIMA by allowing for the inclusion of external variables (exogenous variables) that may influence the time series data but are not affected by it. SARIMAX is valuable in situations where additional explanatory variables improve the predictive power of the model, allowing a more comprehensive analysis and improving forecast accuracy. For example, holiday dates, and if the date is a weekday or a weekend or the time is a less busy time(like midmights).

Before applying ARIMA or SARIMA models, it's essential to check if the time series data is stationary. The Augmented Dickey-Fuller (ADF) test is a statistical test commonly used for this purpose. It evaluates whether a unit root is present in a time series, indicating non-stationarity. If the null hypothesis of the ADF test is rejected, it suggests that the data is stationary.

If the time series data is found to be non-stationary according to the ADF test, differencing can be employed to make it stationary. The order of differencing depends on how many times this process needs to be repeated until stationarity is achieved. It's important to strike a balance and avoid over-differencing, which can lead to loss of valuable information. After each differencing step, it's crucial to inspect the differenced series for stationarity using the ADF test or visual inspection of plots. Once stationarity is achieved, modelling with ARIMA or SARIMA can proceed.

Non-seasonal ARIMA are generally denoted as ARIMA(p,d,q). Seasonal Arima (SARIMA) accepts an additional set of parameters (P,D,Q) that specifically describe the seasonal components of the model.

- *p* — the number of autoregressive terms
- *d* — number of differencing needed to make time series data stationary
- *q* — the number of lagged forecast errors in the prediction equation.

The parameters p and q can be be found manually by plotting the AutoCorrelation (ACF) and the Partial AutoCorrelation (PACF) plots. The ACF plot helps in identifying the MA

component (parameter *q*), while the PACF plot helps in identifying the AR component (parameter *p*). The choice of *p* and *q* is often guided by where the plots "cut off" or become insignificant.
If there's a sharp drop-off in the AC plot after a certain lag, it indicates the number of lagged terms needed to capture the autocorrelation in the data, which will be the number for *q*. If there's no clear pattern or shows a slow decrease, it suggests that the series may not need a MA component, and *q* can be set to zero.
Similar to the ACF, a sharp cutoff in the PACF plot indicates the number of lagged terms and represents the order of the autoregressive (AR) process. Additionally, if there's no clear pattern or the plot shows a slow decay to zero, it suggests that the series may not need an AR component, and *p* can be set to zero.

ARIMA, SARIMA, and SARIMAX models are powerful tools for time series analysis and forecasting. ARIMA serves as a foundation for modelling non-seasonal data, while SARIMA extends its capabilities to handle seasonal patterns. SARIMAX further enhances the model's predictive accuracy by incorporating exogenous variables. The models provide valuable insights into the underlying patterns of time series data. Their versatility and effectiveness make them indispensable tools in the field of time series analysis. The incorporation of statistical tests like the ADF test ensures the data's stationarity, a crucial step in model preparation, is the selection of optimal (S)ARIMA model parameters.

Linear regression is one of the most widely used algorithms in machine learning due to its simplicity, interpretability, and efficiency. It is particularly well-suited for understanding the relationship between variables and making predictions based on past data.
Linear regression is a type of supervised learning algorithm used for predictive analysis. It models the relationship between a dependent variable y (the target or output) and one or more independent variables x (the features or inputs). The algorithm aims to find a linear relationship between these variables, which is represented by a straight line.
The simplest form of linear regression is known as Simple Linear Regression, which involves a single feature and a single target variable.
The relationship is modeled using the equation: $y=wx + b$.

- y is the predicted value (target).
- x is the independent variable (feature).
- w is the weight (slope of the line).

- b is the bias (intercept).

The objective of linear regression is to find the optimal values for w (weight) and b (bias) that minimize the difference between the predicted values (ŷ) and the actual values (y). This difference is measured using a loss function, commonly the Mean Squared Error (MSE).

The goal is to minimize the MSE by adjusting the values of w and b using optimization techniques like Gradient Descent.

Time series data consists of observations recorded at regular time intervals (per minute). In time series forecasting, the goal is to predict future values based on historical data. This type of data often exhibits patterns such as trends, seasonality, and correlations between consecutive time points.

In our case, we are using minute-level power readings to predict the next minute's real power value and also extended this approach to hourly-level predictions.

The reason linear regression works well in our case is due to the high correlation between consecutive power readings. In time series data, especially in scenarios like power consumption, the value at time t is often highly influenced by the value at t−n (in our case n values are 1 to 60). The value at time t (next minute) is predicted by the value at t-1 (previous minute) and also predicted by the values at from t-60 to t-1 (previous 60 minute).

Linear regression is a powerful and efficient algorithm for time series forecasting, especially when the data shows a strong correlation between consecutive readings. Our approach of predicting power consumption using previous values is a classic application of this method. The simplicity of the model, combined with the strong correlation in the data, allows for accurate short-term predictions.

However, to extend the predictions to longer time horizons or the data starts to exhibit non-linear patterns, need to consider more sophisticated models such as ARIMA, SARIMA or LSTM neural networks.

3.1.3 Model Training:

Model training is one of the most crucial steps in building a machine learning model. It involves feeding data into a machine learning algorithm so that the model can learn patterns, relationships, and structures from that data to make accurate predictions.

Model training is the process of applying a machine learning algorithm to a dataset to enable the model to learn the underlying patterns and relationships within the data. The goal is to build a model that generalizes well, meaning it can make accurate predictions on unseen data, not just on the training data it was exposed to.

During training, the model learns by adjusting its internal parameters (like weights and biases in linear regression) to minimize the error between its predictions and the actual observed values. The process continues until the model reaches an acceptable level of accuracy.

In our case ,we trained the linear regression model with 70% of data and obtained the weight and bias value and trained the ARIMA and SARIMAX models with 90% of data with appropriate values of p,d,q and P,D,Q ,as the time series model is primarily used for short term prediction, testing the high percentage of data would result in huge error, so training the model with 90% of data would result in less error and high possibility to learn the pattern.

3.1.4 Model Evaluation:

After training the model, it is essential to evaluate its performance on the test set to ensure that it generalizes well to new data. This step is crucial for avoiding overfitting, where a model performs well on the training data but poorly on unseen data.

Model evaluation refers to the process of using metrics and techniques to assess how well a trained machine learning model performs on new, unseen data. It involves testing the model on a separate dataset (commonly called the test set) to see how accurately it makes predictions.

There are many methods to evaluate the trained model,in our case we used Root mean square error (RMSE) method which is simply the square root of (Mean Square Error)MSE, the average squared difference between the predicted and actual values. RMSE is one of the most commonly used metrics as it provides a more interpretable value by bringing the units back to the original scale of the target variable to evaluate.

The error percentage for the linear regression models to predict the real power for next minute using previous minute real power and previous 60 minute real power is 4.98% and 4.75% respectively. The error for the time series model (ARIMA and SARIMAX) is very vast which is visualized graphically, the model doesn’t fit the expected pattern. The error percentage for the linear regression model to predict the real power for next hour using previous hour real power is 8.90%.

On overall evaluation the simplest and efficient model for the dataset is linear regression model.

3.1.5 **Prediction**:

Prediction is one of the most crucial applications of machine learning (ML), involving the process of using data and algorithms to make informed guesses about future or unknown outcomes.

Prediction in machine learning refers to the process of using a trained model to generate output values based on input data. The goal is to make accurate predictions on new, unseen data by identifying patterns learned from historical data during the training phase.

In our case we predicted the values of real power for next minute using the previous values of real power and also the predicted for next hour using real power of previous hour and even predicted each minute in a particular day.

CHAPTER 4
RESULTS

4.1 Result of the Linear regression model based on previous minute value:

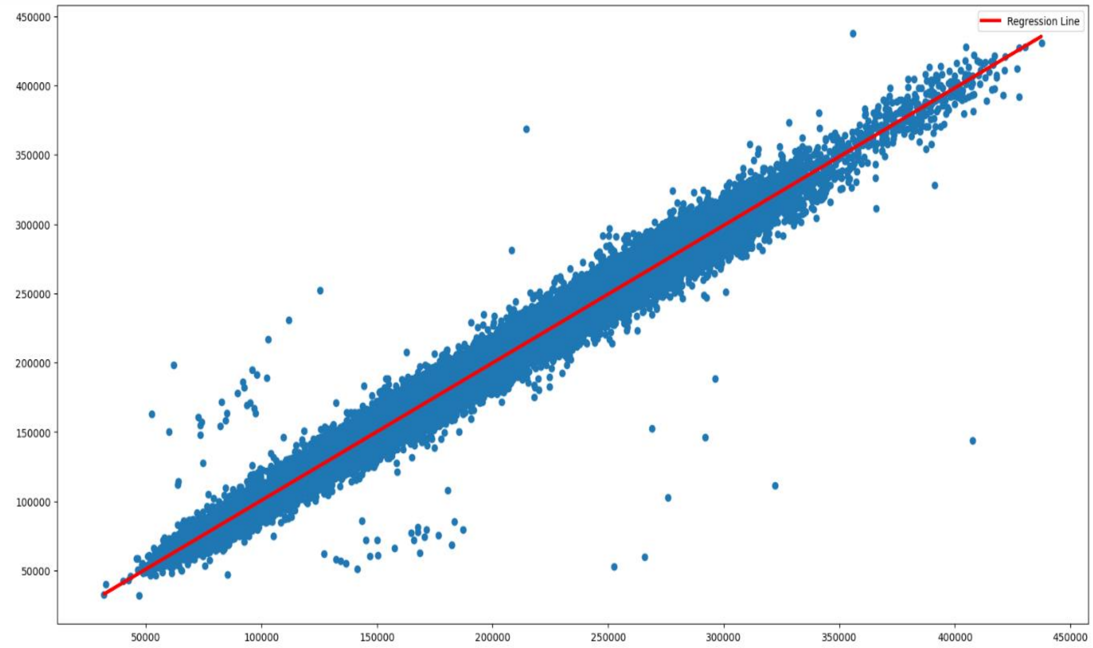

Figure 6: Next Minute Result based on previous minute value

Where x-axis is previous minute Real Power and y-axis is the next minute Real Power. The dots are the actual data points(i.e.,) real power and the linear line is the predicted value. The Linear Regression Model fits the data so well.

4.2 Result of the Linear regression model based on previous 60 minutes values:

```
next_minute_pred = linear_model.predict([df['RealPower'].iloc[-60:].values])
print("The last 60 minute's power:",[df['RealPower'].iloc[-60:].values])
print("Prediction for the next minute:", next_minute_pred)
```

```
The last 60 minute's power: [array([141427.05, 141322.22, 149839.27, 134954.11, 124506.52, 132071.41,
       127258.19, 128096.79, 139776.05, 132997.36, 127852.2 , 127328.07,
       131285.22, 121781.07, 120470.76, 125135.48, 136142.11, 119885.49,
       119815.6 , 119439.98, 124768.59, 132359.67, 119605.96, 119571.01,
       124960.77, 120287.32, 118837.24, 138168.73, 143305.17, 135749.03,
       135845.11, 139173.31, 137653.34, 138483.22, 143907.91, 133669.98,
       128367.59, 129634.22, 126926.24, 132455.77, 128760.68, 134508.59,
       132447.03, 127930.82, 132184.97, 128874.25, 132377.16, 137644.61,
       126183.73, 128411.27, 134997.78, 131695.78, 134840.53, 132333.47,
       132508.19, 129817.66, 138841.36, 141584.3 , 131285.22, 134526.06])]
Prediction for the next minute: [134077.29019987]
```

Figure 7: Next Minute Result based on previous 60 minutes values

Here the error for this model is less(4.75%), so the predictions are tend to be more accurate. Thus this Linear Regression Model fit the data so well.

4.3 Result of the ARIMA and SARIMAX model based on previous hour value:

4.3.1 ARIMA:

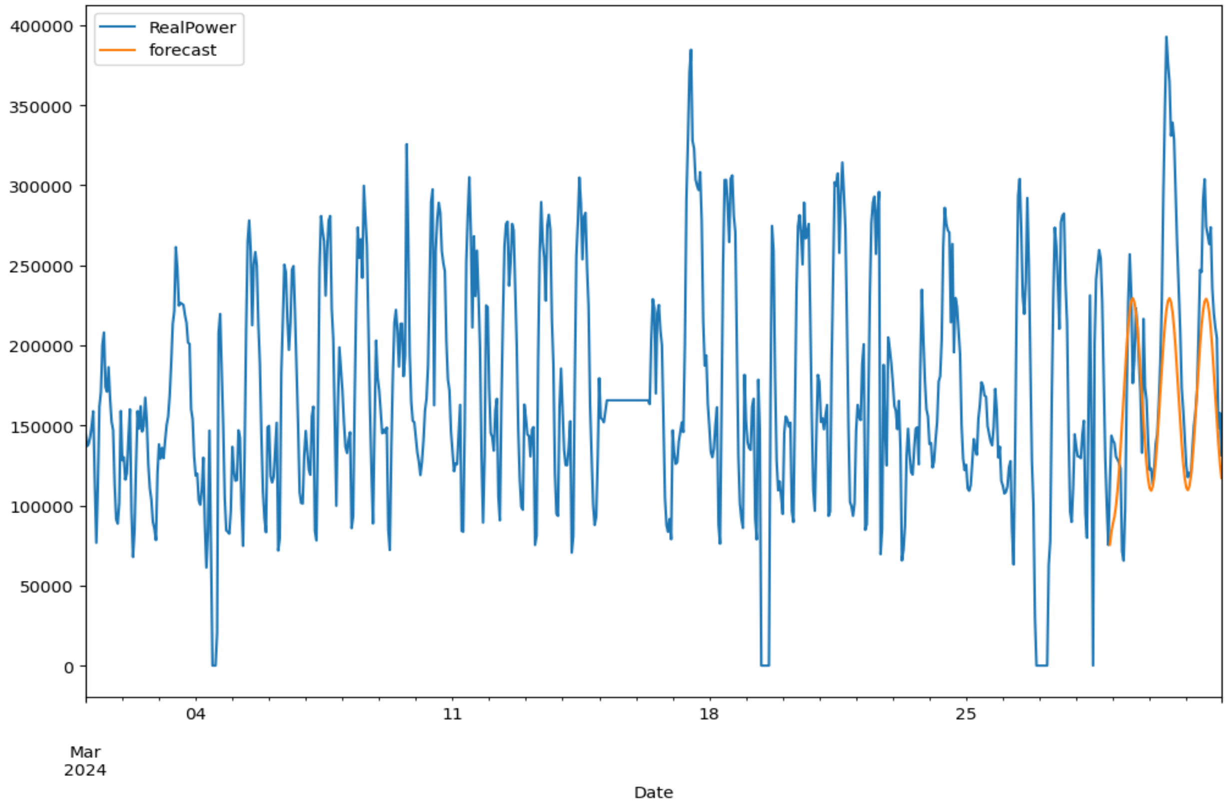

Figure 8: ARIMA

x-axis is date and time of march 2024 and y-axis is the values of real power.

The ARIMA model doesn't fit the tested dataset well.

4.3.2 SARIMAX:

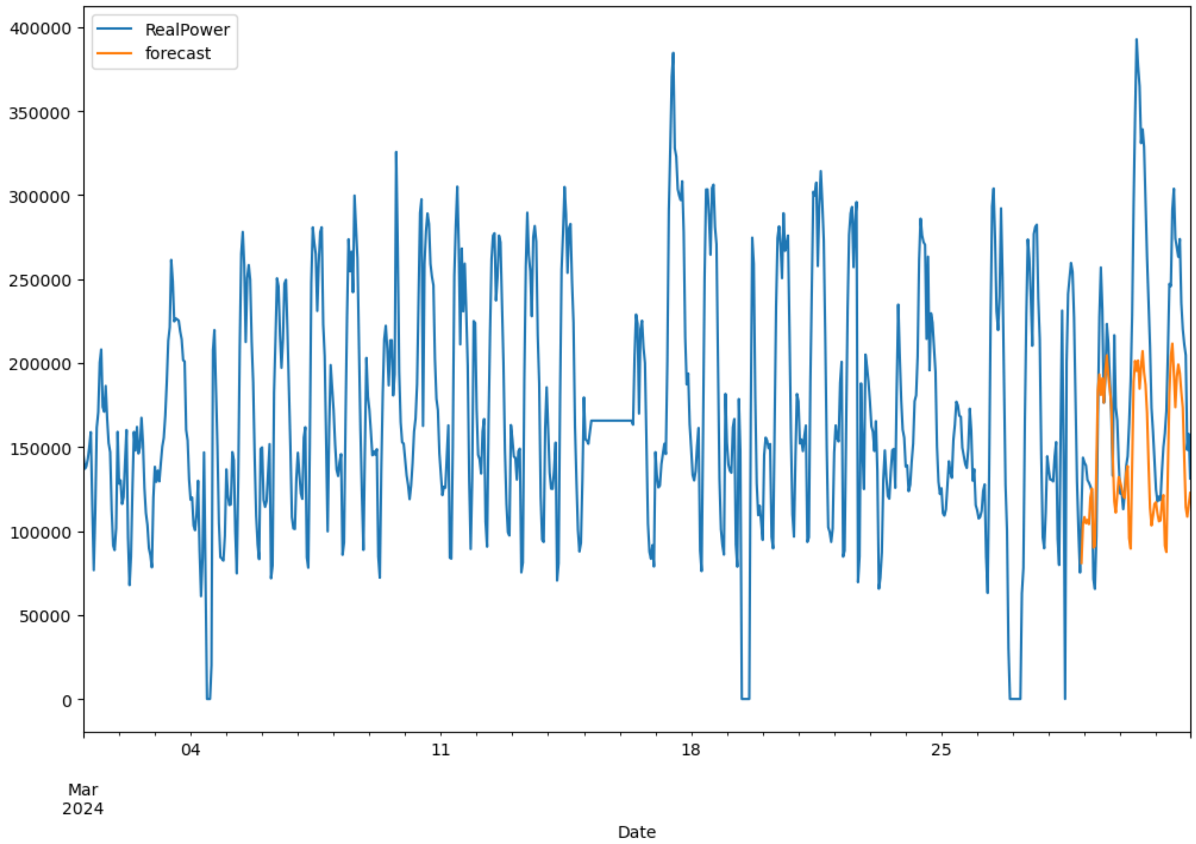

Figure 9: SARIMAX

x-axis is date and time of march 2024 and y-axis is the values of real power.

The SARIMAX model doesn't fit the tested dataset well even though it fits the seasonality it doesn't able to capture the values of it.

4.4 Result of the Linear regression model based on previous hour value:

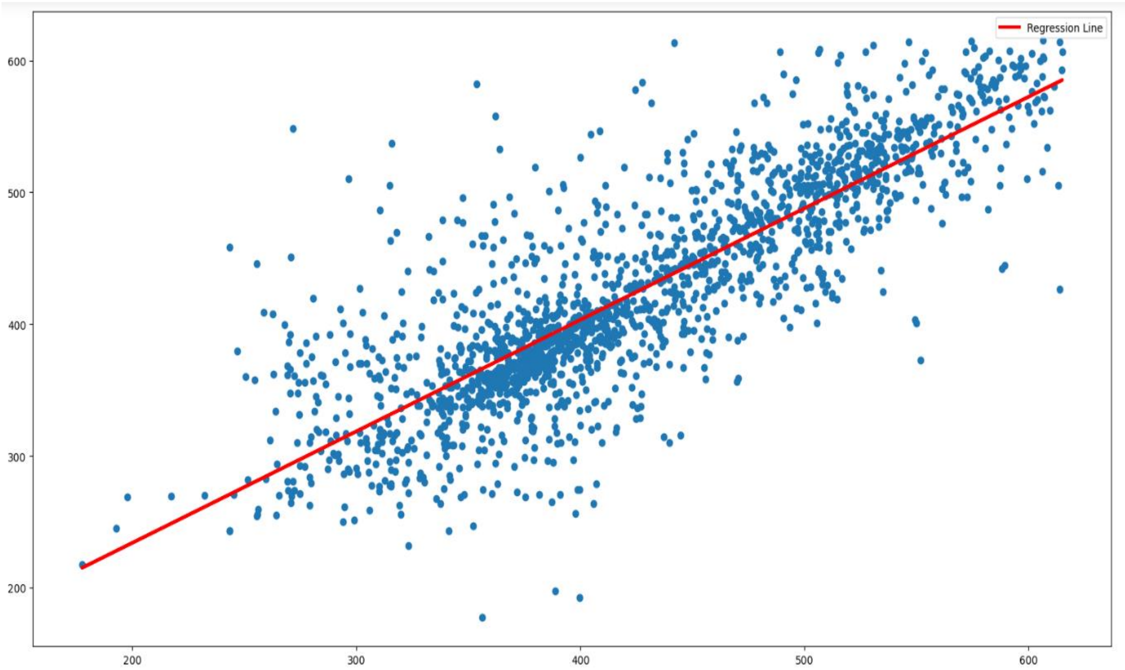

Figure 10: Per Hour Result based on previous hour value

Where x-axis is previous hour Real Power and y-axis is the next hour Real Power. The dots are the actual data points(i.e.,) real power and the linear line is the predicted value. The Linear Regression Model fits the data decently.

CHAPTER 5
CONCLUSION

In this project, we developed a predictive model to forecast real-time power consumption based on historical data from transformer readings. Our primary goal was to achieve accurate and efficient predictions for minute-level and hourly-level power usage to optimize load management. Throughout this project, we explored various predictive modeling techniques, including traditional time series analysis and machine learning methods, ultimately determining that linear regression was the most effective approach for our specific dataset.

Initially, time series models like ARIMA and SARIMAX were considered due to their popularity in handling temporal data. However, upon further analysis, these models did not fit well with the nature of our dataset. Instead, a simpler yet robust approach using linear regression was found to be far more effective. By shifting the Real Power data and using previous values to predict future ones, we were able to capture the inherent correlation in the data efficiently.

For minute-level predictions, the linear regression model demonstrated high accuracy, achieving minimal error rates. The correlation between consecutive minute-level readings was exceptionally high, which allowed the model to make precise forecasts based on the previous minute's power consumption. This resulted in accurate short-term predictions that can be useful for real-time monitoring and adaptive load management.

When predicting power consumption at an hourly frequency, linear regression also showed decent performance, though slightly less accurate than minute-level predictions. By resampling the data to an hourly frequency and training the model accordingly, we were able to smooth out short-term fluctuations while still capturing broader trends. This capability is especially valuable for applications where hourly predictions are sufficient for planning and resource allocation.

One of the significant achievements of this project was the successful integration of real-world transformer data with a machine learning model to make real-time predictions. The data was preprocessed cautiously to handle missing values, align timestamps, and resampled data to different intervals, which ensured that our model was trained on clean and consistent data. By leveraging the linear relationship between consecutive readings, our approach not only simplified the predictive process but also maintained a high level of accuracy, especially for short-term forecasts.

In conclusion, this project highlights the effectiveness of using a data-driven approach, specifically linear regression, for predicting power consumption in real-time. While more complex models exist, the simplicity and efficiency of linear regression made it the ideal choice for our dataset, especially given its high correlation between consecutive data points. Future work could explore enhancing model performance by integrating additional features such as weather conditions, transformer load characteristics, or leveraging ensemble models to capture non-linear patterns if needed.

Ultimately, the insights gained from this project demonstrate the potential for predictive analytics to optimize power usage, improve load management, and support decision-making in real-time energy systems. The successful implementation of this predictive model can serve as a foundation for further enhancements and integration into larger energy management systems, contributing to smarter and more sustainable energy consumption practices.

CHAPTER 5

FUTURE SCOPE

The completion of this project on power prediction using machine learning marks a significant milestone, providing a strong foundation for further research and applications. However, there remains considerable potential to expand and enhance the project in the following ways:

i. **Real-Time Implementation**: The predictive model can be integrated into live systems for real-time monitoring and forecasting. This would enable organizations to dynamically manage energy usage, address fluctuations in power demand, and optimize energy distribution efficiently.

ii. **Feature Enrichment**: The model's accuracy and robustness can be improved by incorporating additional features such as weather data, time-of-day effects, and holiday schedules. These factors can help capture complex dependencies and enhance the predictive capability.

iii. **Longer-Term Predictions**: While the current project focuses on minute- and hourly-level predictions, extending the model to predict power usage over daily, weekly, or monthly periods would provide utility companies and industries with strategic insights for long-term planning and resource allocation.

iv. **Exploration of Advanced Models**: Although linear regression performed well for minute-level predictions, exploring advanced machine learning models such as LSTM, or hybrid models combining machine learning with statistical approaches could uncover non-linear patterns and improve performance for more complex datasets.

v. **Scalability and Broader Applicability**: This project can be scaled to handle data from multiple transformers across diverse locations. Such an extension would enable regional or nationwide power consumption forecasts, benefiting utilities and grid operators in planning and load balancing.

By pursuing these future directions, the project's framework can be significantly extended, driving innovation and delivering valuable insights for efficient energy management and planning in various sectors. These advancements would ensure the sustainability and scalability of predictive systems in addressing real-world challenges.

REFERENCES

- “Short-Term Load Forecasting Methods: A Review” By A.K. Srivastava, Ajay Shekhar Pandey and Devender Singh in the Proceedings of the International Conference on Emerging Trends in Electrical, Electronics and Sustainable Energy Systems (ICETEESES-16) 2016.
- “Electric Load Forecasting: Literature Survey and Classification of Methods” By Hesham K. Alfares and Mohammad Nazeeruddin in the International Journal of Systems Science, April 2001 pp. 23-34.
- D. Papalexopoulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," in *IEEE Transactions on Power Systems*, vol. 5, no. 4, pp. 1535-1547, Nov. 1990, doi: 10.1109/59.99410.