National College of Ireland

Forecasting of Power plants consumption using Machine Learning Techniques

MSc Research Project
Data Analytics

Mohit Kaushal Jain
Student ID: X21191514

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Mohit Kaushal Jain
Student ID:	X21191514
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	14/08/2023
Project Title:	Forecasting of Power plants consumption using Machine Learning Techniques.
Word Count:	3304
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Mohit Kaushal Jain
Date:	17th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Forecasting of Power plants consumption using Machine Learning Techniques

Mohit Kaushal Jain
X21191514

Abstract

A vital task in ensuring the effective use of energy resources and maintaining a steady power supply is to predict power plant consumption. The generation, distribution, and overall grid stability of energy can all be impacted by changes in power consumption. This study explores the use of machine learning algorithms to forecast power plant consumption with the goal of offering insightful information. This study creates a prediction model for energy consumption by using SARIMAX Time Series Modelling. The dataset used for this study was the Tetuan City power consumption dataset from Kaggle. To evaluate the performance of the model RMSE(Root Mean Square Error) and MSE(Mean Square Error) were used.

1 Introduction

In order to power both our economy and society as a whole, it is essential that we have access to a consistent and efficient energy supply. The energy industry continues to play a crucial role in meeting these needs despite obstacles like the rising demand for electricity. Power plant consumption forecasting is one strategy employed by this sector and is thought to be extremely valuable when managing resources efficiently to prevent scenarios involving unnecessary wastage.

1.1 Background

Forecasting power plant consumption involves predicting the energy usage of a power plant or an entire energy system. For the energy sector, it is crucial to forecast power plant consumption, and machine learning techniques have shown to have enormous potential in this regard. The prediction is based on data and other relevant factors. It assists energy companies in making decisions regarding resource management planning for capacity expansions, in the future and optimizing energy production and distribution.

1.2 Importance

To make wise choices about their energy usage, energy companies must have a thorough understanding of the trends and patterns of energy consumption. Companies are able to reduce waste and effectively manage their resources with the help of information. Additionally, some companies may need to comply with regulations related to energy usage. Accurate power plant consumption forecasting can help these companies meet their regulatory obligations and avoid fines.

1.3 Research Question

The research question for this study is:

"What are the most significant factors contributing to differences in power consumption across different neighborhoods?"

1.4 Objectives

The study aims to implement time series modelling. Exploring the ways to improve the accuracy of predicting energy consumption. Performing assumption tests for the model that was built. Lastly evaluating the results obtained after the analysis is done.

1.5 Report Structure

The Report contains the following sections: Section 2 will talk about the study done by previous authors and based on literature review what novelty could be added in this research project. The section 3 will talk about the approach of the machine learning technique involved in this research with KDD Methodology. The section 4 will talk about the hardware and software specification for conducting this research and why these machine learning models are being used in this research. Section 5 will talk about the models that will be implemented in the research. Section 6 will talk about the results obtained in the research done. Lastly, the section 7 will talk about the suggestions, the overall results obtained, the future scope and limitation of this research. [1]

2 Related Work

This section will talk about the researchers who have worked in the energy consumption prediction with referring the peer reviewed research papers till date.

2.1 Data-Driven Tools for Building Energy Consumption Prediction.

The research done by Olu-Ajayi et al. (2023) for Data-Driven Tools for Building Energy Consumption Prediction gives an emphasis on tools used for their study. The article discusses the value of data-driven models in energy planning and conservation. The authors evaluated the efficacy of data-driven tools by looking at 63 studies based on data properties, energy type, and building type. The evaluation's results show that Support Vector Machine (SVM) performed better in the majority of the review studies than other data-driven tools. In more studies, Artificial Neural Networks (ANN) and Random Forests (RF) outperformed statistical tools like Linear Regression (LR) and Autoregressive Integrated Moving Average (ARIMA). The authors did reach the conclusion that no data-driven tool is fundamentally better than others in all situations. Under various circumstances, the advantages and disadvantages of these tools frequently produce different results. In light of this, the article offers a suggested framework for choosing which tools to use by examining their benefits and drawbacks under various circumstances.

[1] https://www.iea.org/energy-system/electricity

2.2 A machine-learning ensemble model for predicting energy consumption in smart homes.

The research done by Priyadarshini et al. (2022) focused on forecasting energy consumption in smart homes with the help of IoT. The suggested ensemble model's evaluation is assessed using multiple evaluation parameters, such as mean square error (MSE), R-squared (R2), root mean square error (RMSE), and mean absolute error (MAE), Compared to other baseline algorithms, including decision trees, random forests, extreme gradient boosting, and k-nearest neighbor. The study reveals that the suggested model outperforms all other baseline algorithms for various datasets, exhibiting an R2 of approximately 0.99. The article leverages machine learning techniques to provide an exhaustive analysis of energy usage in smart homes.

2.3 Robust building energy consumption forecasting using an online learning approach with R ranger.

The research done by Moon et al. (2022) focused on a ranger approach to predict the energy consumption of a building. The proposed model used online learning to forecast a fast implementation of random forest in R. In the first phase, the researchers constructed three STLF(Short-Term Load Forecasting) models utilizing tree-based ensemble learning techniques. In the second phase, a ranger-based forecasting model was developed, wherein a sliding window size of seven days was employed with the predicted values of the STLF models acting as input variables along with external variables such as timestamp and temperature. The study performed data preprocessing for input variable configuration and utilized publicly available dataset for electricity consumption data of two office buildings to create training and test sets. The authors then proceeded to demonstrate the effectiveness of the suggested RABOLA model via comparative experiments.

2.4 Machine Learning-Based Ensemble Classifiers for Anomaly Handling in Smart Home Energy Consumption Data.

The researchers Kasaraneni et al. (2022) proposed a ensemble based classifier approach for handling anomalies in data on smart home energy consumption. The accuracy of analysis conducted on data pertaining to smart home energy consumption is influenced by data quality issues, specifically garbage data, outliers, redundant data, and missing data. This study presents a comparison of single classifier and ensemble classifiers using a variety of machine learning algorithms, including but not limited to random forest (RF), support vector machine (SVM), decision tree (DT), naive bayes, K-nearest neighbour, and neural networks.. The proposed methodology involves identifying and removing all anomalies prior to imputing missing or removed data. To evaluate classifier performance, metrics such as accuracy, precision, recall/sensitivity, specificity, and F1 score are employed. Results indicate that the ensemble classifier "RF+SVM+DT" outperforms other classifiers in terms of anomaly handling.

2.5 Stacking Deep learning and Machine learning models for short-term energy consumption forecasting.

The study proposed by A. et al. (2022) was related to short term forecasting of energy consumption. The study combined the predictions of base models using Gradient Boosting and Extreme Gradient Boosting (XG) in two different ensemble models. The proposed models are evaluated using a common dataset comprising approximately 500,000 electricity consumption values collected over nine years at regular intervals. The experimental validation reveals that the proposed ensemble model based on XGB is not only more accurate but also demonstrates a training time reduction of approximately ten times when compared to other models. The Root Mean Square Error (RMSE) was found to have decreased by about 39%. The study highlights that a single model may not be sufficient to address the linear and non-linear issues associated with predicting electricity consumption. Ensemble models, which integrate the predictions of multiple base models, can provide more precise and robust predictions.

2.6 Applying Machine Learning Methods for Power Plant Generation Time Series Forecasting.

The article entitled "Applying Machine Learning Methods for Power Plant Generation Time Series Forecasting" details a technique for predicting the generation of thermal power plants utilizing machine learning approaches. The authors, Shishkov and Pronichev (2022), initially constructed a model of the data using two recurrent neural network architectures while also deriving features from electrical and date-time values for the first stage. Subsequently, three-level ensembles of models founded on gradient boosting and linear regression over decision trees were created in the second stage. The proposed forecasting method for time series was assessed in the article utilizing quality metrics. The results indicate that the recommended technique has a high probability of being advantageous for resolving problems associated with time series forecasting.

2.7 Prediction of electricity generation from a combined cycle power plant based on a stacking ensemble and its hyperparameter optimization with a grid-search method.

The present article by Qu et al. (2021) introduces a method for predicting the full-load power generation of combined cycle power plants (CCPPs) through the utilization of a stacking ensemble hyperparameter optimization approach. The article posits that the precision of electricity planning and energy utilization relies heavily on the ability to accurately predict power generation. To develop the prediction method, the study utilized 9568 data items from a CCPP that had been operating at maximum capacity for a period of six years. The results of the study demonstrated that the proposed approach provides the power plant with a high degree of prediction accuracy, even under complex environmental variables.

2.8 A Stacking Multi-Learning Ensemble Model for Predicting Near Real Time Energy Consumption Demand of Residential Buildings.

In the paper entitled "A Stacking Multi-Learning Ensemble Model for Predicting Near Real Time Energy Consumption Demand of Residential Buildings" by Ves et al. (2019), implemented an energy consumption forecasting model for the purpose of predicting energy demand at the level of residential buildings. This proposed model was a stacked multi-learning ensemble model that combined Gradient Boosting Regression, Multi-Layer Neural Networks, and Long Short Term Memory Networks, and it offered a means of forecasting residential energy demands at both the individual and aggregate levels.

2.9 Comparison of Machine Learning Algorithms for the Power Consumption Prediction : - Case Study of Tetouan city.

The article by Salam and Hibaoui (2018) presents a comparative analysis of diverse machine learning models employed for predicting the electricity consumption in Tetouan, Morocco. The principal aim of this study was to determine the most optimal approach for forecasting power consumption at ten-minute intervals. The authors utilized historical data from the Supervisory Control and Data Acquisition (SCADA) system between the period of 2017-01-01 and 2017-12-31. The models used for the comparison included Support Vector Machine for Regression (SVR) with Radial Basis Function Kernel, Decision Tree, Random Forest, and Feedforward Neural Network with Backpropagation Algorithm. To achieve the most precise performance, the parameters associated with the comparative models were optimized using the Grid-search method.

2.10 Short-Term Forecasting for Energy Consumption through Stacking Heterogeneous Ensemble Learning Model.

The article proposed by Khairalla et al. (2018) titled "Short-Term Forecasting for Energy Consumption through Stacking Heterogeneous Ensemble Learning Model" introduced a novel approach for enhancing the precision of short-term energy consumption forecasts. The researchers combined support vector regression (SVR), backpropagation neural network (BPNN), and linear regression (LR) learners in their stacking multi-learning ensemble (SMLE) model to create a flexible ensemble framework. The model was tested using data from the Global Oil Consumption (GOC) dataset, which is intended for predicting nonlinear time series. The research compares the suggested SMLE model to several benchmark techniques and demonstrates its superior performance in terms of error rate, similarity, and directional accuracy. The proposed model is capable of producing precise short-term predictions for energy consumption. Furthermore, the authors argue that the ensemble model is a useful methodology for complex time series forecasting. The study concludes by highlighting that the SMLE model was the best-performing model.

3 Methodology

The study is divided into a number of stages that are organised in accordance with the Knowledge Discovery in Databases (KDD) process methodology. The reason for using

KDD Methodology are that: the dataset is very large consisting of multiple variables, it also involves in discovering useful and important insights about the data. The process involves Data Selection, Data Preprocessing, Exploratory Data Analysis, Data Transformation, Data Modelling, and Evaluation. The figure 1 shows the flow diagram of KDD.

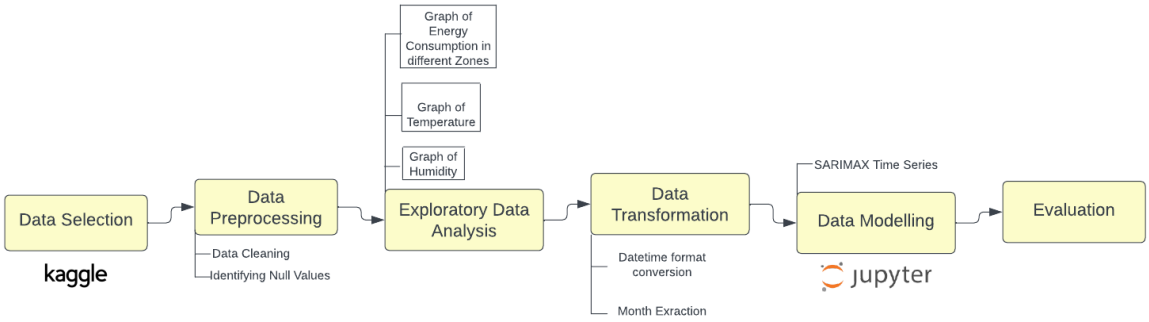


Figure 1: KDD Methodology.

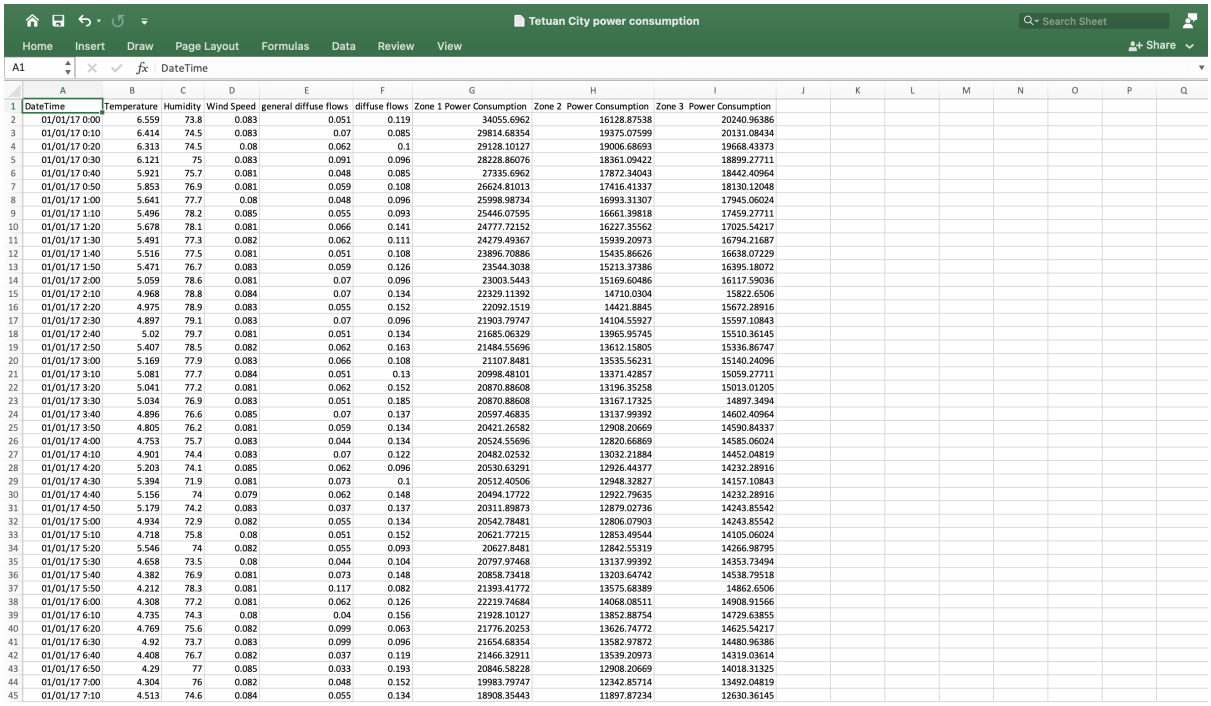
3.1 Data Selection

This is the first and crucial step toward the KDD Approach and it is a process to identify the important features from the dataset for doing the analysis. The dataset that was selected for doing the analysis is "Tetuan City power consumption" obtained from Kaggle [2]. The dataset contains 52,416 rows and 9 columns. The dataset description is as follows:

1. **DateTime**: Date with time intervals of 10 minutes.
2. **Temperature**: Temperature in the city of Tetouan.
3. **Humidity**: Humidity in the city of Tetouan.
4. **Wind Speed**: Wind Speed in the city of Tetouan.
5. **General Diffuse flows**
6. **diffuse flows**
7. **Zone 1 power consumption in the city of Tetouan**
8. **Zone 2 power consumption in the city of Tetouan**
9. **Zone 3 power consumption in the city of Tetouan**

The figure 2 shows the csv version of the dataset.

[2]https://www.kaggle.com/datasets/ashkanforootan/tetuan-city-power-consumption



Tetuan City power consumption

Search Sheet

Home Insert Draw Page Layout Formulas Data Review View Share

A1 DateTime

	A	B	C	D	E	F	G	H	I
1	DateTime	Temperature	Humidity	Wind Speed	general diffuse flows	diffuse flows	Zone 1 Power Consumption	Zone 2 Power Consumption	Zone 3 Power Consumption
2	01/01/17 0:00	6.559	73.8	0.083	0.051	0.119	34055.6962	16128.87538	20240.96386
3	01/01/17 0:10	6.414	74.5	0.083	0.07	0.085	29814.68354	19375.07599	20131.08434
4	01/01/17 0:20	6.313	74.5	0.08	0.062	0.1	29128.10127	19006.68693	19668.43373
5	01/01/17 0:30	6.121	75	0.083	0.091	0.096	28228.86076	18361.09422	18899.27711
6	01/01/17 0:40	5.921	75.7	0.081	0.048	0.085	27335.6962	17872.34043	18442.40964
7	01/01/17 0:50	5.853	76.9	0.081	0.059	0.108	26624.81013	17416.41337	18130.12048
8	01/01/17 1:00	5.641	77.7	0.08	0.048	0.096	25998.98734	16993.31307	17945.06024
9	01/01/17 1:10	5.496	78.2	0.085	0.055	0.093	25446.07595	16661.39818	17459.27711
10	01/01/17 1:20	5.678	78.1	0.081	0.066	0.141	24777.72152	16227.35562	17025.54217
11	01/01/17 1:30	5.491	77.3	0.082	0.062	0.111	24279.49367	15939.20973	16794.21687
12	01/01/17 1:40	5.516	77.5	0.081	0.051	0.108	23896.70886	15435.86626	16638.07229
13	01/01/17 1:50	5.471	76.7	0.083	0.059	0.126	23544.3038	15213.37386	16395.18072
14	01/01/17 2:00	5.059	78.6	0.081	0.07	0.096	23003.5443	15169.60486	16117.59036
15	01/01/17 2:10	4.968	78.8	0.084	0.07	0.134	22329.11392	14710.0304	15822.6506
16	01/01/17 2:20	4.975	78.9	0.083	0.055	0.152	22092.1519	14421.8845	15672.28916
17	01/01/17 2:30	4.897	79.1	0.083	0.07	0.096	21903.79747	14104.55927	15597.10843
18	01/01/17 2:40	5.02	79.7	0.081	0.051	0.134	21685.06329	13965.95745	15510.36145
19	01/01/17 2:50	5.407	78.5	0.082	0.062	0.163	21484.55696	13612.15805	15336.86747
20	01/01/17 3:00	5.169	77.9	0.083	0.066	0.108	21107.8481	13535.56231	15140.24096
21	01/01/17 3:10	5.081	77.7	0.084	0.051	0.13	20998.48101	13371.42857	15059.27711
22	01/01/17 3:20	5.041	77.2	0.081	0.062	0.152	20870.88608	13196.35258	15013.01205
23	01/01/17 3:30	5.034	76.9	0.083	0.051	0.185	20870.88608	13167.17325	14897.3494
24	01/01/17 3:40	4.896	76.6	0.085	0.07	0.137	20597.46835	13137.99392	14602.40964
25	01/01/17 3:50	4.805	76.2	0.081	0.059	0.134	20421.26582	12908.20669	14590.84337
26	01/01/17 4:00	4.753	75.7	0.083	0.044	0.134	20524.55696	12820.66869	14585.06024
27	01/01/17 4:10	4.901	74.4	0.083	0.07	0.122	20482.02532	13032.21884	14452.04819
28	01/01/17 4:20	5.203	74.1	0.085	0.062	0.096	20530.63291	12926.44377	14232.28916
29	01/01/17 4:30	5.394	71.9	0.081	0.073	0.1	20512.40506	12948.32827	14157.10843
30	01/01/17 4:40	5.156	74	0.079	0.062	0.148	20494.17722	12922.79635	14232.28916
31	01/01/17 4:50	5.179	74.2	0.083	0.037	0.137	20311.89873	12879.02736	14243.85542
32	01/01/17 5:00	4.934	72.9	0.082	0.055	0.134	20542.78481	12806.07903	14243.85542
33	01/01/17 5:10	4.718	75.8	0.08	0.051	0.152	20621.77215	12853.49544	14105.06024
34	01/01/17 5:20	5.546	74	0.082	0.055	0.093	20627.8481	12842.55319	14266.98795
35	01/01/17 5:30	4.658	73.5	0.08	0.044	0.104	20797.97468	13137.99392	14353.73494
36	01/01/17 5:40	4.382	76.9	0.081	0.073	0.148	20858.73418	13203.64742	14538.79518
37	01/01/17 5:50	4.212	78.3	0.081	0.117	0.082	21393.41772	13575.68389	14862.6506
38	01/01/17 6:00	4.308	77.2	0.081	0.062	0.126	22219.74684	14068.08511	14908.91566
39	01/01/17 6:10	4.735	74.3	0.08	0.04	0.156	21928.10127	13852.88754	14729.63855
40	01/01/17 6:20	4.769	75.6	0.082	0.099	0.063	21776.20253	13626.74772	14625.54217
41	01/01/17 6:30	4.92	73.7	0.083	0.099	0.096	21654.68354	13582.97872	14480.96386
42	01/01/17 6:40	4.408	76.7	0.082	0.037	0.119	21466.32911	13539.20973	14319.03614
43	01/01/17 6:50	4.29	77	0.085	0.033	0.193	20846.58228	12908.20669	14018.31325
44	01/01/17 7:00	4.304	76	0.082	0.048	0.152	19983.79747	12342.85714	13492.04819
45	01/01/17 7:10	4.513	74.6	0.084	0.055	0.134	18908.35443	11897.87234	12630.36145

Figure 2: Tetuan City power consumption dataset.

3.2 Data Preprocessing

The next step for this research project is preprocessing of the data. It involved checking of null values in the dataset, it was found that the the dataset didn't contain any null values. The figure 3 shows the information of the dataset.

3.3 Exploratory Data Analysis (EDA)

The third step of this research is the exploratory data analysis of power consumption. Exploratory Data Analysis is basically understand the trends, patterns and characteristics of the dataset.

The graph plotting for temperature is shown in figure 4. For the temperature plot, there was no seasonality observed from January 2017 to December 2017. The graph follows a slight decreasing trend from January 2017 to March 2017. The graph followed a increasing trend from march 2017 to August 2017. The maximum temperature was recorded for August 2017. From March 2017 to December 2017 a decreasing trend was observed for temperature. The graph plotting for humidity is shown in figure 5. There is a seasonality trend observed for this graph. The maximum humidity was observed in May 2017 and the lowest humidity was observed in August 2017. The graph plotting for Zone wise power consumption is shown in figure 6. It shows the monthly analysis of power consumption where zone 1 had the peak value of consumption in the end of August 2017, zone 2 had the peak value of consumption in September 2017 and zone 3 had a peak in the August 2017. EDA was applied to Zone 1 power consumption, zone 2 power consumption, zone 3 power consumption, temperature and humidity because the graphs were showing increasing or decreasing together but for wind speed, general diffuse flows

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 52416 entries, 2017-01-01 00:00:00 to 2017-12-30 23:50:00
Data columns (total 8 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Temperature                 52416 non-null  float64
 1   Humidity                    52416 non-null  float64
 2   Wind Speed                  52416 non-null  float64
 3   general diffuse flows       52416 non-null  float64
 4   diffuse flows               52416 non-null  float64
 5   Zone 1 Power Consumption    52416 non-null  float64
 6   Zone 2  Power Consumption   52416 non-null  float64
 7   Zone 3  Power Consumption   52416 non-null  float64
dtypes: float64(8)
memory usage: 3.6 MB
```

Figure 3: Dataset Information

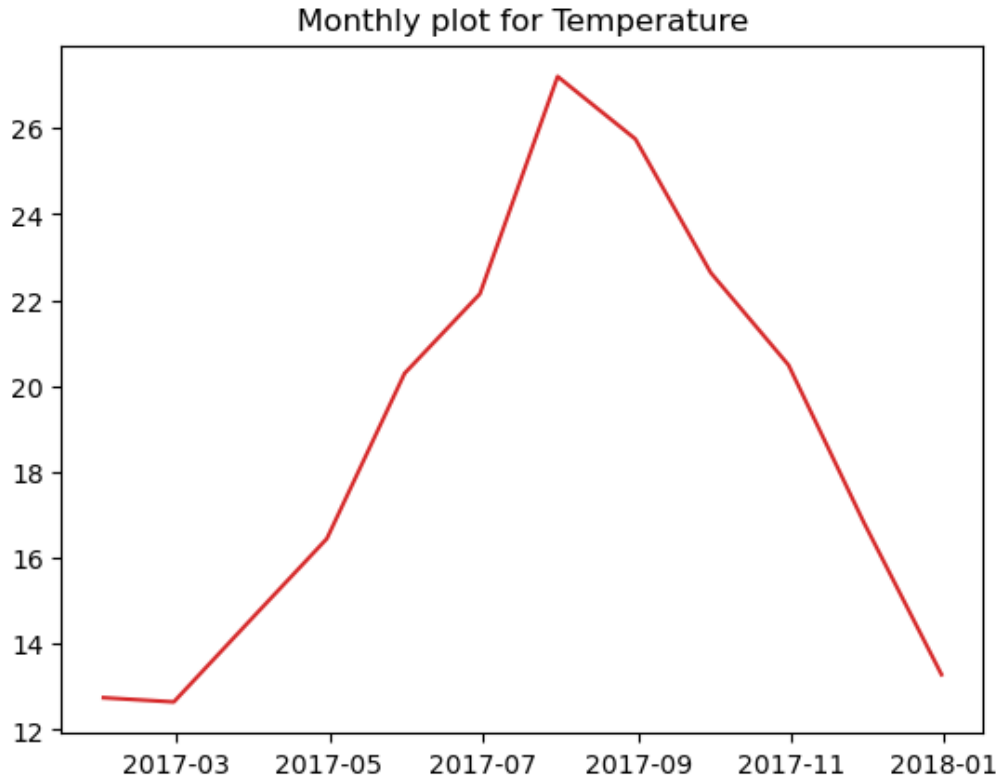


Figure 4: Plot for Temperature.

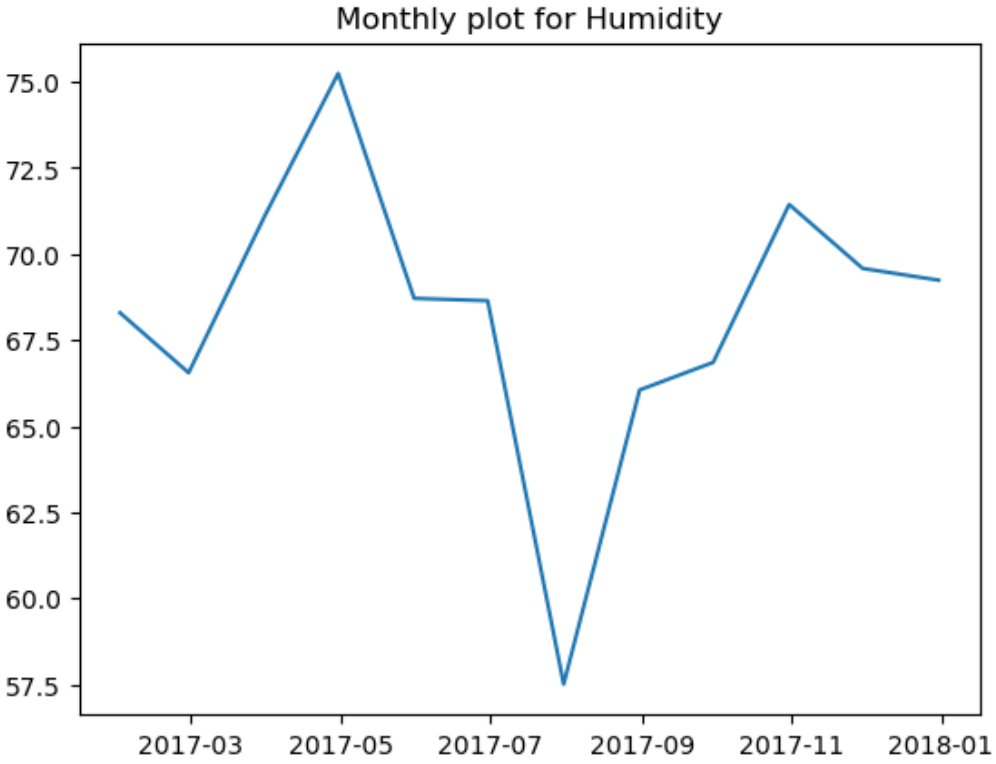


Figure 5: Plot for Humidity.

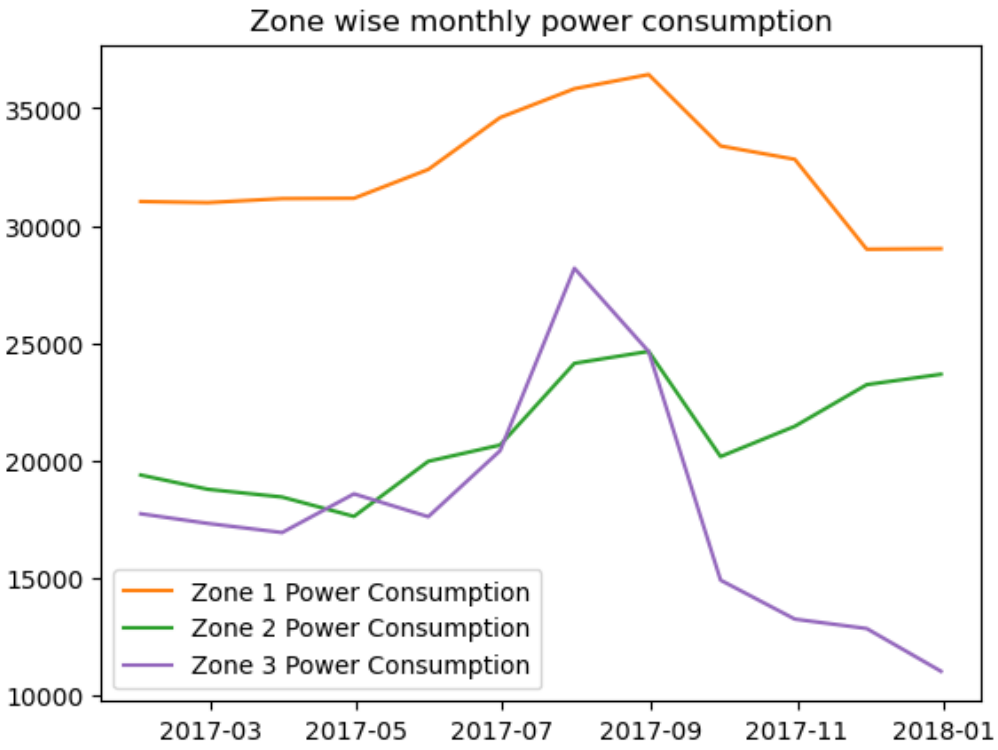


Figure 6: Plot for Power Consumption.

and diffuse flows the there were no fixed pattern observed hence I considered performing EDA on 5 variables in the tetuan power consumption dataset.

3.4 Data Transformation

In data mining, Data transformation is a systematic process of converting unprocessed data into a structured form for ease of doing the analysis and modeling. Additive decomposition was applied for the time series analysis of power consumption. It was done to identify the trend, seasonality, and residuals from the original data. The figure 7 shows the graph of the additive decomposition of time series.

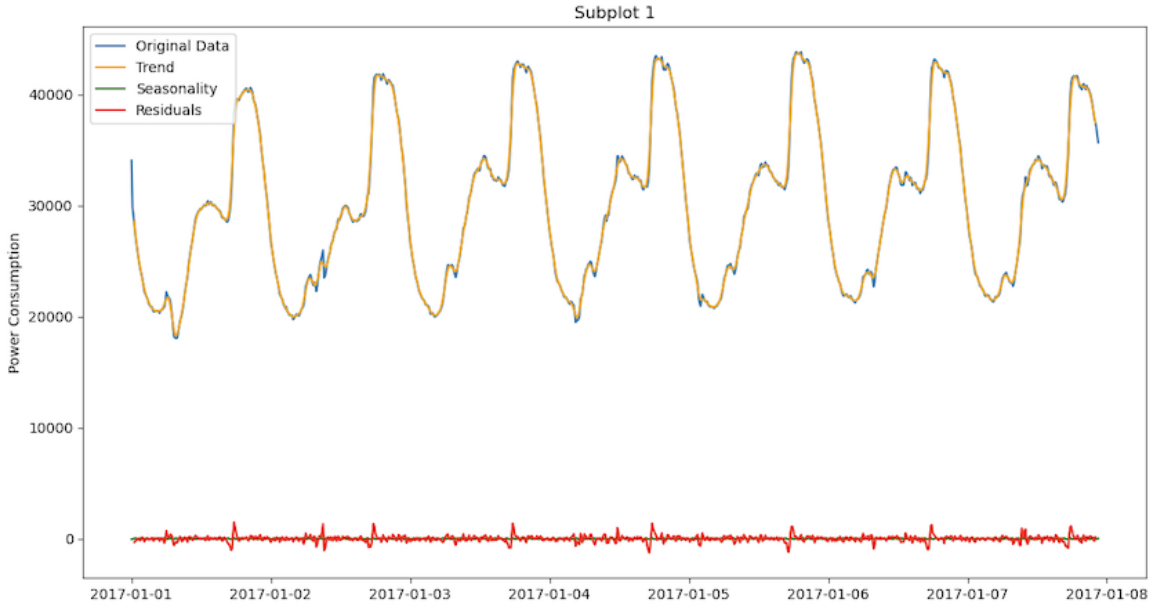


Figure 7: Additive decomposition of time series.

3.5 Data Modelling

The graph in figure 8 shows the plot for time series analysis of energy consumption in different zones. It was observed that zone 1 was having highest consumption of energy compared to zone 2 and zone 3. While zone 2 was having higher consumption than zone 3.

3.6 Model Evaluation

- **Augmented dickey fuller test**: It is a test to verify if the time series plot is stationary or not. For evaluating the stationarity a hypothesis test is performed. For this research project the p-value obtained for this research was greater than the significance value which means that the data is not stationary. The figure 9 below shows the results for augmented dickey fuller test.

- **Kwiatkowski Phillips Schmidt Shin(KPSS) test**: It is also a test for stationarity. If the p-value is greater than the significance value then the stationarity is

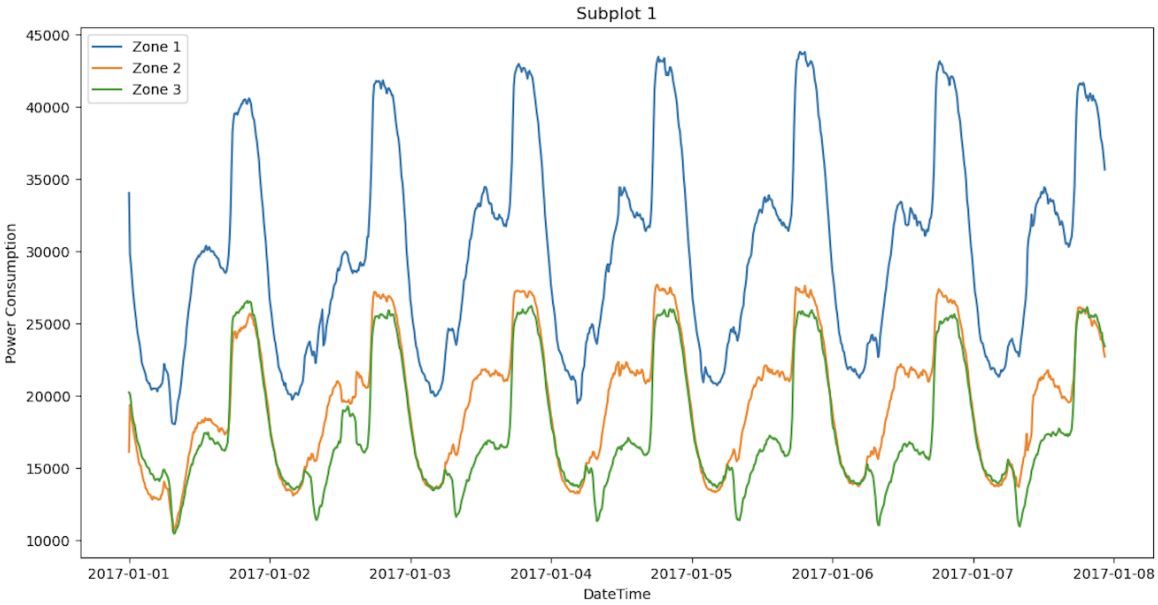


Figure 8: Plot for Time Series Modelling.

```
ADF Statistic for Zone 1 Power Consumption: -32.12127853462614
p-value for Zone 1 Power Consumption: 0.0
Critical Values for Zone 1 Power Consumption: {'1%': -3.4304749044184266, '5%': -2.861595205242518, '10%': -2.5667993
83915253}


ADF Statistic for Zone 2  Power Consumption: -25.22216377100368
p-value for Zone 2  Power Consumption: 0.0
Critical Values for Zone 2  Power Consumption: {'1%': -3.4304749044184266, '5%': -2.861595205242518, '10%': -2.566799
383915253}


ADF Statistic for Zone 3  Power Consumption: -16.36686797515679
p-value for Zone 3  Power Consumption: 2.835133086903964e-29
Critical Values for Zone 3  Power Consumption: {'1%': -3.4304749044184266, '5%': -2.861595205242518, '10%': -2.566799
383915253}


ADF Statistic for Humidity: -17.184247931293186
p-value for Humidity: 6.616075836617727e-30
Critical Values for Humidity: {'1%': -3.4304749044184266, '5%': -2.861595205242518, '10%': -2.566799383915253}


ADF Statistic for Temperature: -9.459827585705547
p-value for Temperature: 4.384185727809652e-16
Critical Values for Temperature: {'1%': -3.4304749044184266, '5%': -2.861595205242518, '10%': -2.566799383915253}
```

Figure 9: Augmented Dickey Fuller Test.

attained. For the results obtained for KPSS it showed that the time series is stationary. The figure 10 below shows the results for the Kwiatkowski Phillips Schmidt Shin test[3].

```
KPSS Statistic for Zone 1 Power Consumption: 6.018599702888408
p-value for Zone 1 Power Consumption: 0.01
Critical Values for Zone 1 Power Consumption: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}


KPSS Statistic for Zone 2  Power Consumption: 16.860093845930926
p-value for Zone 2  Power Consumption: 0.01
Critical Values for Zone 2  Power Consumption: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}


KPSS Statistic for Zone 3  Power Consumption: 8.92788914353952
p-value for Zone 3  Power Consumption: 0.01
Critical Values for Zone 3  Power Consumption: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}


KPSS Statistic for Humidity: 0.5777992157502823
p-value for Humidity: 0.024654616749974337
Critical Values for Humidity: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}


KPSS Statistic for Temperature: 11.643586215500964
p-value for Temperature: 0.01
Critical Values for Temperature: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
```

Figure 10: Kwiatkowski Phillips Schmidt Shin test.

- **Autocorrelation test**: This is also a test for analysing the stationarity in time series.The following was the graph obtained for ACF and PACF as shown in figure 11. Autocorrelation test is a test for calculating the degree of similarity between present and past values of lagged version of time intervals . Basically if the value of autocorrelation value is +1 then there is positive correlation and if the value of autocorrelation value is -1 then there is negative correlation. It is helpful in identifying future values based on past values. If the value is 0 then there is negative correlation.

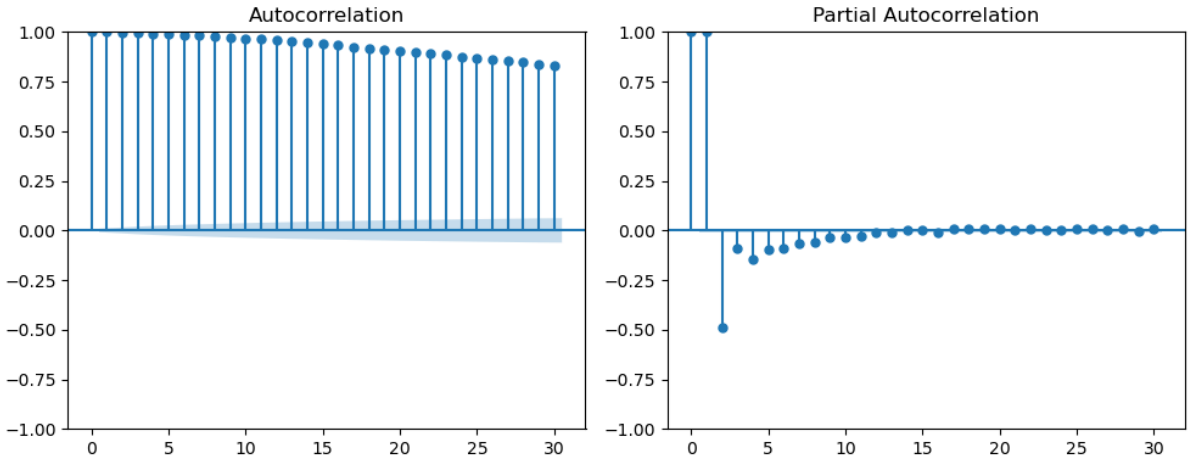


Figure 11: ACF and PACF test.

[3]https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity$_d$etrending$_a$df$_k$pss.html

4 Design Specification

Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors, or SARIMAX, is a powerful time series modeling technique that is frequently used for both forecasting and analyzing data that exhibit temporal patterns and possible relationships with outside variables. SARIMAX, which incorporates seasonality and exogenous predictors, is an advance version of the ARIMA (Autoregressive Integrated Moving Average) model. It contains the 2 components which non-seasonal and seasonal component. The non seasonal components contain p,q,d. The seasonal components contain P,Q,D, s and X.

5 Implementation

The programming was done using python programming language. For verifying the working of the model assumption tests were done for time series analysis. Time series analysis was implemented in this research project. The assumption tests that were carried out were Augmented Dickey-Fuller Test, Kwiatkowski-Phillips-Schmidt-Shin Test and Autocorrelation test.

6 Evaluation

The following are the evaluation metrics that were used when the project was done: Mean Squared error(MSE) and Root Mean Squared Error(RMSE). The Mean Squared Error (MSE) is a statistical metric that quantifies the level of error in models. It calculates the mean of the squared differences between the predicted and observed values. In cases where the model has no errors, the MSE will be zero. The Root Mean Square Error (RMSE) is defined as the square root of the Mean Squared Error (MSE). RMSE is a useful measure for evaluating the average error magnitude in the same unit as the orignal data. RMSE and MSE are the accuracy metrics used for evaluating the model forecasting that was implemented. The RMSE and MSE provides a details about how well the model worked for forecasting by comparing the actual values and the predicted values. If lower is the MSE and RMSE values the better is the performance of prediction model.

The formula for Mean Squared Error (MSE) is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

$$n = \text{Number of observations}$$
$$y_i = \text{Actual value for observation } i$$
$$\hat{y}_i = \text{Predicted value for observation } i$$

The formula for Root Mean Square Error (RMSE) is given by:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

$$n = \text{Number of observations}$$
$$y_i = \text{Actual value for observation } i$$
$$\hat{y}_i = \text{Predicted value for observation } i$$

6.1 Prediction of Energy Consumption in Zone 1

The figure 12 shows the predictions for energy consumption at zone 1. For this case the predictions obtained for this region was a constant value.

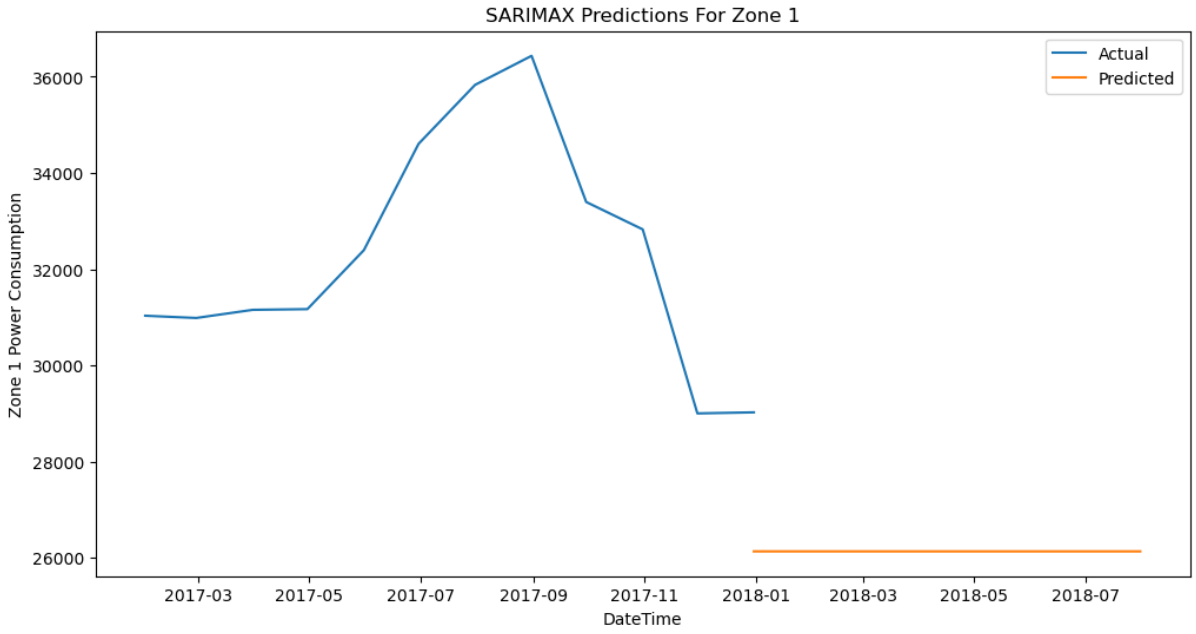


Figure 12: Prediction of Energy Consumption in Zone 1.

6.2 Prediction of Energy Consumption in Zone 2

The figure 13 shows the predictions for energy consumption at zone 2. For this case the prediction obtained for this region shows some amount of increasing trend end of July 2018.

6.3 Prediction of Energy Consumption in Zone 3

The figure 14 shows the predictions for energy consumption at zone 3. For this case it shows a decreasing trend till end of July 2018.

6.4 Predction of temperature

The figure 15 shows the predictions for temperature. A gradual decreasing trend is observed till the end of July 2018.

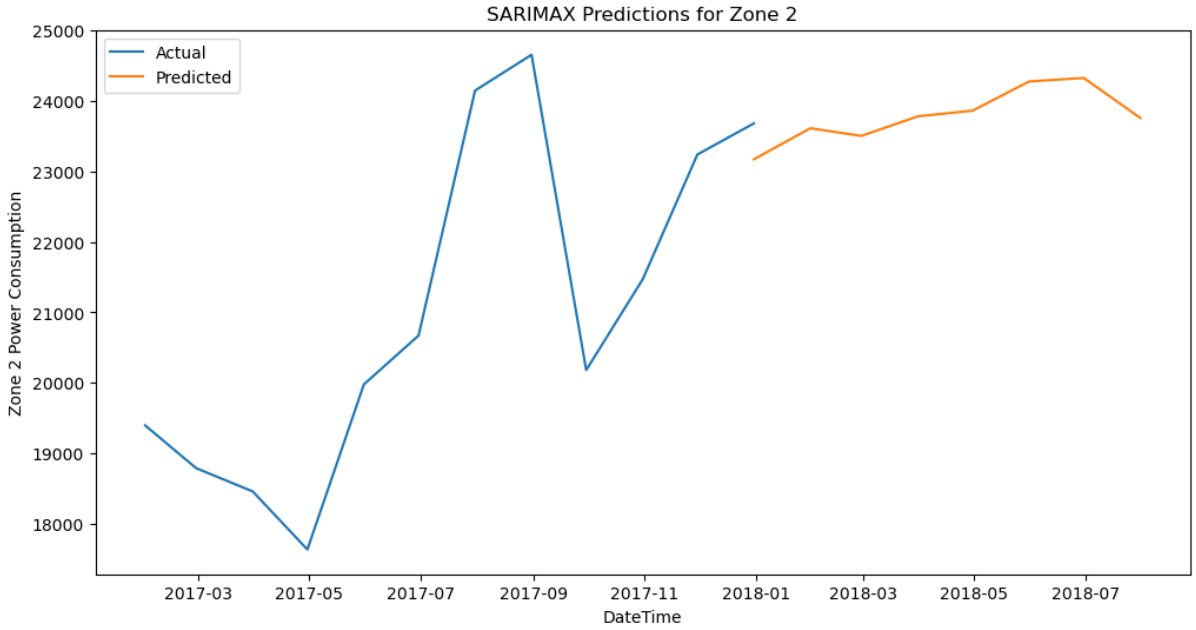


Figure 13: Prediction of Energy Consumption in Zone 2.

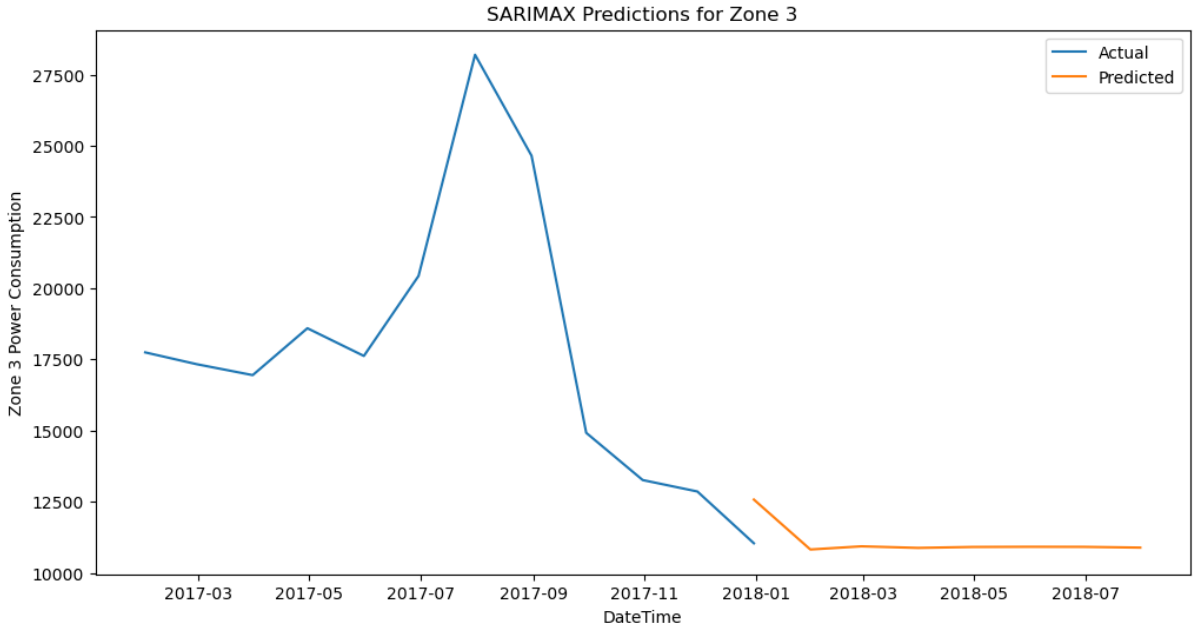


Figure 14: Prediction of Energy Consumption in Zone 3.

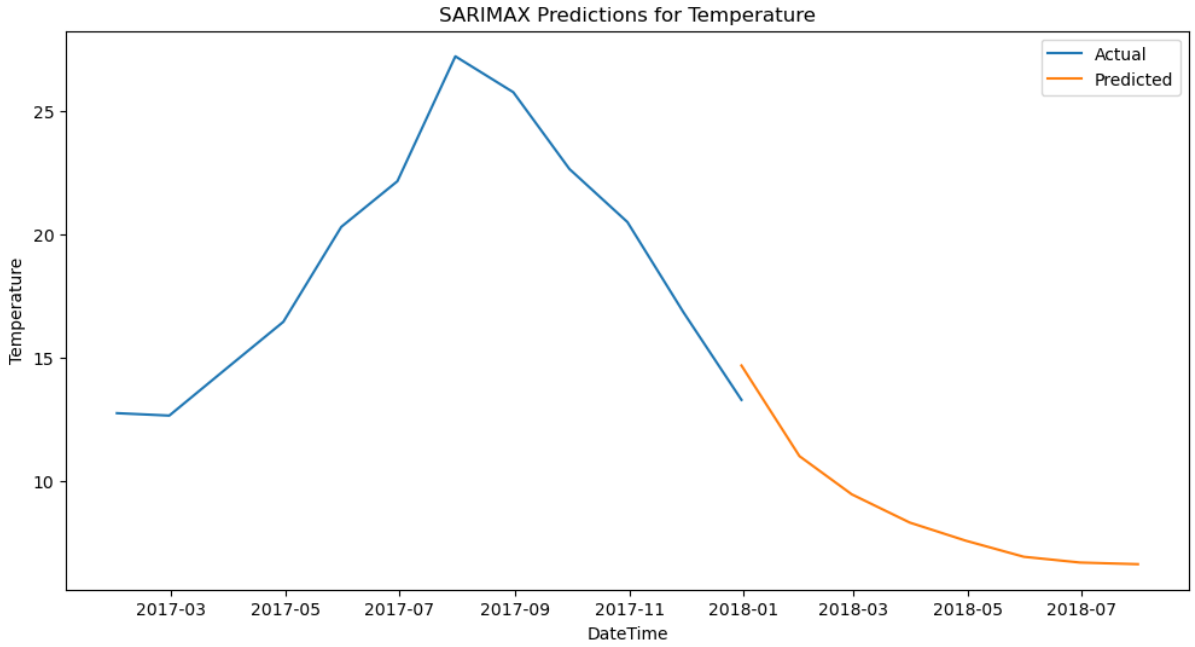


Figure 15: Prediction of Temperature.

6.5 Prediction of Humidity

The figure 16 shows the predictions for Humidity. A seasonal trend of prediction was obtained till the end of july 2018.

6.6 Discussion

In this section, the discussion related to The predictions were done for this research. The table 1 shows the errors related to different measurements. The mean squared error and root mean squared error obtained for zone 1 were 42747195.12 and 6957.96 respectively. The mean squared error and root mean squared error obtained for zone 2 were 13005467.29 and 3606.30 respectively. The mean squared error and root mean squared error obtained for zone 3 were 16661435.73 and 4081.84 respectively. The mean squared error and root mean squared error obtained for temperature were 1.11 and 1.05 respectively. The mean squared error and root mean squared error obtained for humidity were 23.17 and 4.81 respectively. The reason for getting these predictions are due to the dataset complexity.

Table 1: Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for Different Zones, Temperature, and Humidity.

Measurement	**MSE**	**RMSE**
Zone 1	42747195.12	6957.96
Zone 2	13005467.29	3606.30
Zone 3	16661435.73	4081.84
Temperature	1.11	1.05
Humidity	23.17	4.81

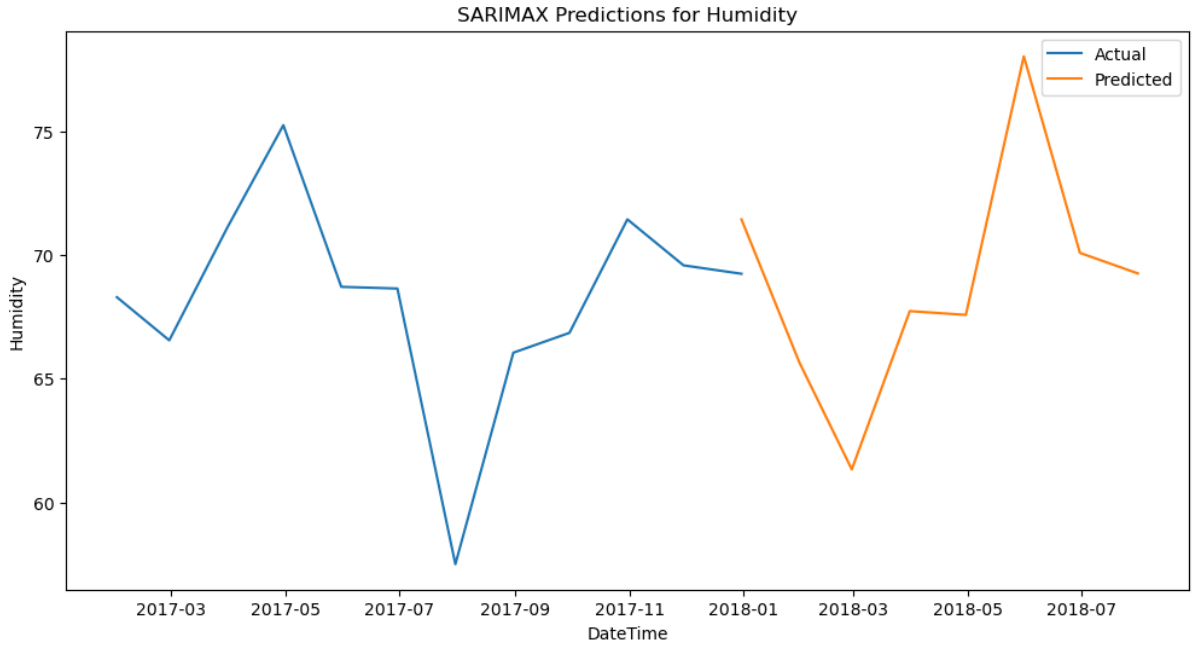


Figure 16: Prediction of Humidity.

7 Conclusion and Future Work

The research question has been achieved in this project which was identifying important factors for the differences in power consumption. SARIMAX time series analysis was implemented for this research project. The evaluation was done on the basis of RMSE(Root Mean Squared error) and MSE(Mean Squared Error). The prediction were done for first six months of 2018 using mean forecasting. Although satisfactory results for zone 2 and zone 3 were obtained for prediction. Better predictions were obtained for temperature and humidity.

The time series modelling could be improved by adding LSTM based approach to achieve better accuracy and results for this model. The LSTM approach is suitable because of the nature of the dataset, to identify the sequential patterns and large volume of data. LSTM is an abbreviation for long short-term memory networks, is employed in the domain of Deep Learning. These networks are a type of recurrent neural networks (RNNs) demonstrates their ability to acquire long-term dependencies, particularly in sequence prediction problems. Before implementing the LSTM based Time series model, the data must be normalized as a part of data transformation step of KDD. LSTM would be a good approach for further research in the energy consumption forecasting. Although the BiLSTM could be a good approach for the project but it would take more time for getting the prediction results hence LSTM would be better to get results as fast as possible for the complex dataset for this project. Normalization will help in scaling down the data in the range of 0 and 1. Since there was a problem in analysing the prediction of temperature and there was no seasonality observed for monthly analysis of temperature it would be beneficial to do normalization for that dataframe. Another approach that could be helpful in prediction of energy consumption would be Markov chain forecasting Meng et al. (2022). In statistics, probability theory, and data analysis, Markov chain

forecasting is a method for predicting future states of a system based on the present state and the Markov property. According to the Markov property, a system's future behaviour depends only on its current state, not on the series of events that led to it.

References

A., S. R., S., A., R., H. and S., S. K. (2022). Stacking deep learning and machine learning models for short-term energy consumption forecasting, *Advanced Engineering Informatics* **52**: 101542.
URL: *https://www.sciencedirect.com/science/article/pii/S1474034622000180*

Kasaraneni, P. P., Venkata Pavan Kumar, Y., Moganti, G. L. K. and Kannan, R. (2022). Machine learning-based ensemble classifiers for anomaly handling in smart home energy consumption data, *Sensors* **22**(23): 9323.
URL: *http://dx.doi.org/10.3390/s22239323*

Khairalla, M. A., Ning, X., AL-Jallad, N. T. and El-Faroug, M. O. (2018). Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model, *Energies* **11**(6): 1605.
URL: *http://dx.doi.org/10.3390/en11061605*

Meng, Z., Sun, H. and Wang, X. (2022). Forecasting energy consumption based on svr and markov model: A case study of china, *Frontiers in Environmental Science* **10**.
URL: *https://www.frontiersin.org/articles/10.3389/fenvs.2022.883711*

Moon, J., Park, S., Rho, S. and Hwang, E. (2022). Robust building energy consumption forecasting using an online learning approach with r ranger, *Journal of Building Engineering* **47**: 103851.
URL: *https://www.sciencedirect.com/science/article/pii/S2352710221017095*

Olu-Ajayi, R., Alaka, H., Owolabi, H., Akanbi, L. and Ganiyu, S. (2023). Data-driven tools for building energy consumption prediction: A review, *Energies* **16**(6).
URL: *https://www.mdpi.com/1996-1073/16/6/2574*

Priyadarshini, I., Sahu, S., Kumar, R. and Taniar, D. (2022). A machine-learning ensemble model for predicting energy consumption in smart homes, *Internet of Things* **20**: 100636.
URL: *https://www.sciencedirect.com/science/article/pii/S2542660522001172*

Qu, Z., Xu, J., Wang, Z., Chi, R. and Liu, H. (2021). Prediction of electricity generation from a combined cycle power plant based on a stacking ensemble and its hyperparameter optimization with a grid-search method, *Energy* **227**: 120309.
URL: *https://www.sciencedirect.com/science/article/pii/S0360544221005582*

Salam, A. and Hibaoui, A. E. (2018). Comparison of machine learning algorithms for the power consumption prediction : - case study of tetouan city –, *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, pp. 1–5.

Shishkov, E. and Pronichev, A. (2022). Applying machine learning methods for power plant generation time series forecasting, *2022 International Conference on Industrial*

Engineering, Applications and Manufacturing (ICIEAM), pp. 67–71.
URL: *https://ieeexplore.ieee.org/document/9787271*

Ves, A. V., Ghitescu, N., Pop, C., Antal, M., Cioara, T., Anghel, I. and Salomie, I. (2019). A stacking multi-learning ensemble model for predicting near real time energy consumption demand of residential buildings, *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 183–189.
URL: *https://ieeexplore.ieee.org/document/8959572*