HOURLY PRICE AND DEMAND FORECASTING

ABSTRACT

In the rapidly evolving energy landscape, accurate forecasting is pivotal. Our project introduces an innovative approach to energy forecasting, leveraging regression models, machine learning, and time-series methods. Focused on hourly prices, demand, and weather conditions, our methodology enhances decision-making accuracy amid the dynamic energy sector. This report explores the challenges of the energy industry, outlines project objectives, and presents key findings from an extensive exploratory data analysis. Detailed methodologies underline the robustness and innovation of our approach. Results demonstrate the efficacy of our models in predicting energy variables, contributing to global efforts for a sustainable and resilient energy future.

1. INTRODUCTION

In the face of unprecedented transformations within the energy sector, driven by technological advancements, evolving consumer demands, and the urgent need to transition to sustainable practices, the challenges confronting the industry have never been more complex. The dynamics of the modern energy landscape demand accurate and sophisticated forecasting methods to navigate the uncertainties associated with supply, demand, and pricing. This report delves into our innovative energy forecasting methodology, designed to address the intricate challenges of the industry, accelerate the global transition to net-zero, and empower stakeholders with actionable insights for strategic decision-making.

As we stand at the intersection of tradition and transformation in the energy sector, the significance of precise forecasting cannot be overstated. Fluctuating demand patterns, the integration of renewable energy sources, and the inherent volatility in energy prices necessitate forward-thinking approaches. Our project aims to contribute to the solution space by developing forecasting techniques that leverage cutting-edge regression models, machine learning algorithms, and time-series forecasting methods. By embracing a holistic approach that considers variables such as hourly prices, demand, and weather conditions, our methodology is poised to elevate decision-making accuracy in this ever-evolving energy ecosystem.

This introduction sets the stage for an exploration into our energy forecasting journey, where we unravel the complexities of the problem statement, articulate the objectives and scope of our project, and delve into the nuances of the datasets that have fueled our analysis. As we progress, we will walk you through the key findings of our exploratory data analysis, unveil the methodologies employed in our forecasting models, and present the results that underscore the efficacy of our innovative approaches. Join us on this expedition as we unlock the potential of energy forecasting to shape a sustainable and resilient energy future.

2. PROBLEM STATEMENT

The energy industry is undergoing a transformative journey, marked by rapid modernization and technological advancements. Infrastructure upgrades, integration of intermittent renewable energy sources, and evolving consumer demands are reshaping the sector. However, this progress comes with its challenges. Supply, demand, and prices are increasingly volatile, rendering the future less predictable. Moreover, the industry's traditional business models are being fundamentally challenged. In this competitive and dynamic landscape, accurate decision-making is pivotal. The industry relies heavily on probabilistic forecasts to navigate this uncertain future, making innovative and precise forecasting methods essential that aids stakeholders in making strategic decisions amidst the shifting energy landscape.

The goal of this work is to create a machine learning-based method for precisely and effectively forecasting power use. Large data volumes, handling missing values and outliers, and extracting pertinent characteristics from the data should all be capabilities of the method. The method must to be able to decide which model performs the best and anticipate power use with accuracy. To ascertain the suggested approach's efficacy in forecasting price and demand, various assessment indicators should be used. The project seeks to advance energy management by offering a precise and effective approach for forecasting hourly price and demand.

3. METHODOLOGY

3.1 Data-Preprocessing

A dataset is a collection of data. With tabular data, each table row corresponds to a specific record of the data set, and each column to a single variable. A data set is related to one or more database tables.

The datasets contain 4 years of electrical consumption, pricing and weather data.

-Price Forecasting data upto December 24.csv:

Range Index: 35352 entries, 0 to 35351, Data columns (total 3 columns) – 3 Important parameters

	Delivery Day	Hours	Prices\n(EUR/MWh)
0	1/1/2020	H1	26.38
1	1/1/2020	H2	26.50
2	1/1/2020	H3	27.01
3	1/1/2020	H4	27.40
4	1/1/2020	H5	27.99

-Demand Forecasting Demand Data upto Feb 21.csv
-Demand Forecasting Weather Data upto Feb 28.csv

The Demand and Weather datasets were merged as they were related:

	datetime	Demand (MW)	Name	Temperature	feelslike	dewpoint	humidity	precipitation	precipprob	preciptype	...	visibility	solarradiation	uvindex	severerisk	conditions	Unnamed: 21	Unnamed: 22	Unnamed: 23	Unnamed: 24	Unnamed: 25
27547	2/21/2023 19:00	498.7	Pokhara	50.0	49.4	29.0	44.19	0.0	0	NaN	...	9.9	0.0	0	10.0	Partially cloudy	NaN	NaN	NaN	NaN	NaN
27548	2/21/2023 20:00	498.8	Pokhara	50.0	49.4	28.1	42.59	0.0	0	NaN	...	9.9	0.0	0	10.0	Partially cloudy	NaN	NaN	NaN	NaN	NaN
27549	2/21/2023 21:00	506.4	Pokhara	53.1	53.1	26.1	34.92	0.0	0	NaN	...	9.9	0.0	0	10.0	Overcast	NaN	NaN	NaN	NaN	NaN
27550	2/21/2023 22:00	478.8	Pokhara	53.1	53.1	28.1	38.02	0.0	0	NaN	...	9.9	0.0	0	10.0	Overcast	NaN	NaN	NaN	NaN	NaN
27551	2/21/2023 23:00	457.7	Pokhara	53.1	53.1	30.8	42.44	0.0	0	NaN	...	9.9	0.0	0	10.0	Overcast	NaN	NaN	NaN	NaN	NaN

Data Cleaning

Various data cleaning methods were applied to make data ready for training. The methods employed are explained below:

<u>For Price Dataset</u>

The rows with missing values were dropped from the price dataset.

```
df_price.dropna(how='all', inplace=True)
df_price.isna().sum()
```

Some inconsistent data (such as negative values in price) were handled using interpolation.

- Replacing Negative or Zero Prices with NaN

```
mask = df_price['Prices\n(EUR/MWh)'] <= 0
df_price.loc[mask, 'Prices\n(EUR/MWh)'] = np.nan
```

- Using linear interpolation to fill NaN Values

```
df_price['Prices\n(EUR/MWh)'].interpolate(inplace=True)
```

Z-score was implemented for removing outliers.

```
zscore = scipy.stats.zscore(df_price['Prices\n(EUR/MWh)'])
df_price = df_price[abs(zscore)<5]
```

Delivery day and hour columns were merged as DateTime to create date time object.

```
timeMap = {
    f'H{i+1}': f'{i:02d}' for i in range(0, 24)
}
```

```
df_price.loc[:, "Time"] = df_price["Hours"].apply(lambda x: timeMap[x] + ":00:00")
```

```
df_price.loc[:, 'Datetime'] = df_price['Delivery Day'] + ' ' + df_price['Time']
```

- Changing data type of Datetime values to Timestamp

```
df_price.loc[:, 'Datetime'] = pd.to_datetime(df_price['Datetime'])
```

<u>For Demand and Weather Dataset</u>

Unnamed columns and empty columns were removed.

```
columns_to_drop = ['Unnamed: 21', 'Unnamed: 22', 'Unnamed: 23', 'Unnamed: 24', 'Unnamed: 25']
df_merged.drop(columns_to_drop, inplace=True, axis=1)
```

Dropping redundant data e.g Dropping preciptype and precipprob as precipitation has more accurate and non-null data, similarly dropping windgust and keeping windspeed.

```
df_merged.drop(["precipprob", "preciptype" ], inplace=True, axis=1)
df_merged.drop(['windgust'], inplace=True, axis=1)
```

Interpolate () method was used to handle NaN values.

```
for column in df_merged.columns[3:17]:
    df_merged[column] = df_merged[column].interpolate(method='linear', limit_direction='forward', axis=0)
```

Feature Engineering

Normalization and Scaling - Data normalization is a process to standardize the range of features of the data as they may vary a lot. We used scikit-learn's MinMaxScaler () to normalize continuous values and avoid vanishing gradient problems to finalize our data before model training.

Encoding Categorical Values - We used pandas get dummies to handle categorical variables like condition creating new columns consisting of 0s and 1s for each column.

3.2 Data Exploration and Visualization

Data exploration and visualization is a critical phase in our project, serving as the initial lens through which we gain insights into the characteristics and patterns inherent in the hourly price, demand, and weather datasets. We used different data visualization techniques in our project.

i. Time Series Analysis of Daily Price

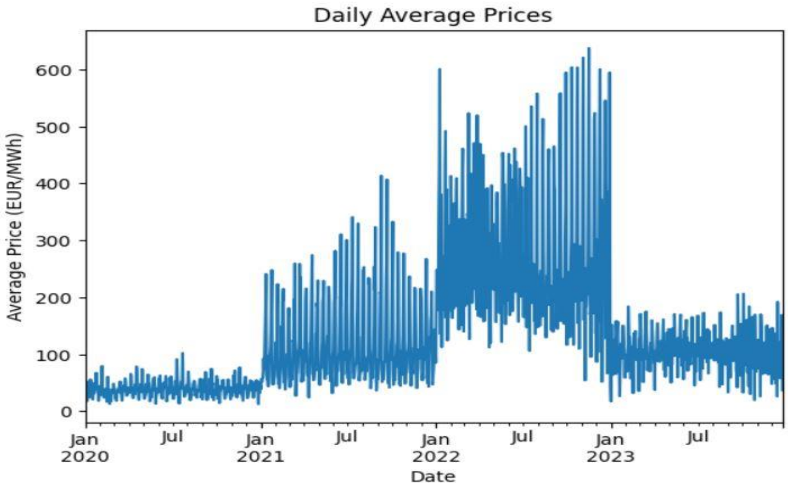


Fig: Plot of daily average price vs time

From the given line plot, we can observe that Prices have generally increased over the time period, with some fluctuations. There was a sharp increase in prices from July 2022 to January 2023, followed by a slight decrease in July 2023. Prices have decreased slightly from January 2023 to July 2023, but remain relatively high compared to earlier years.

ii. Average prices of electricity during each hour of the day

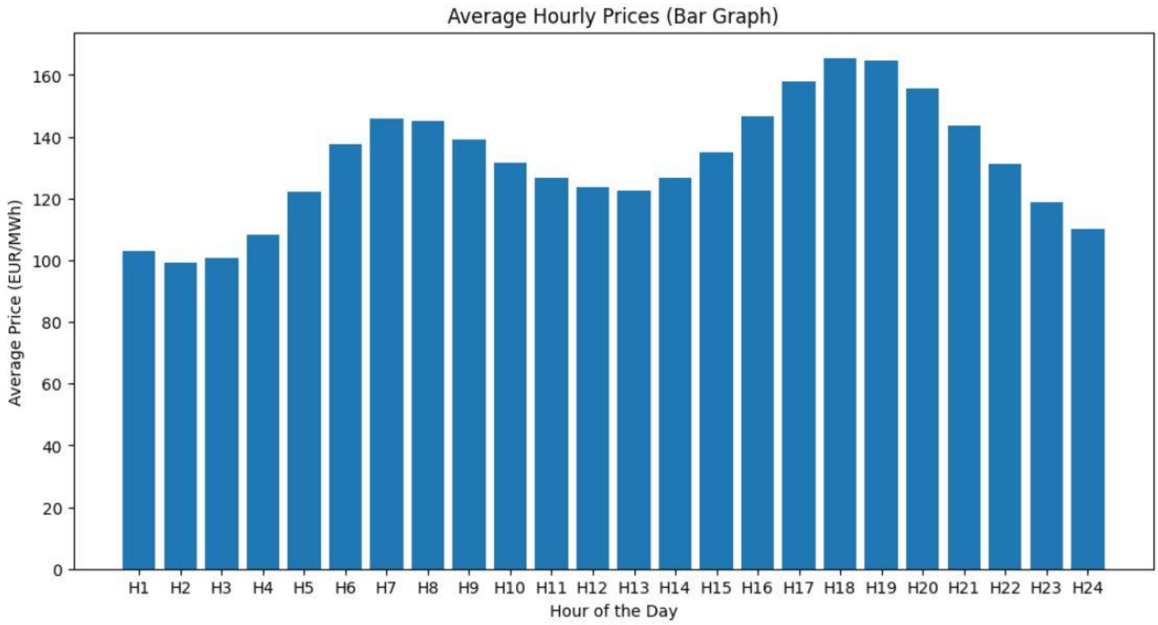
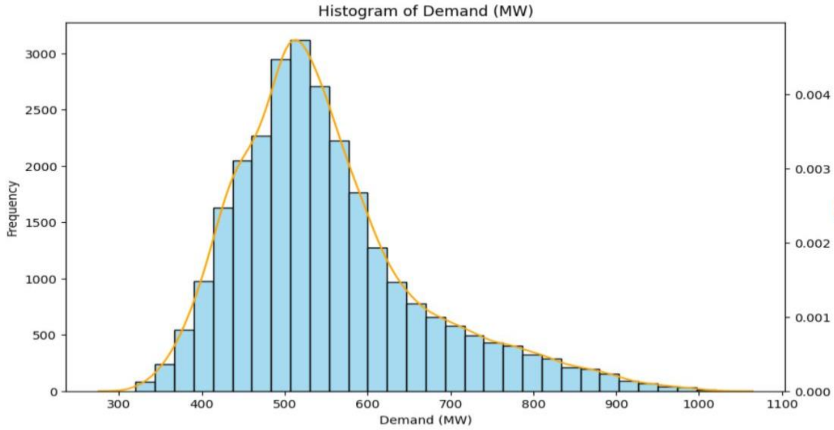


Fig: Average Prices on each hour of day

From the histogram we can see that the prices are highest during the peak hours of consumption, which are typically between 6am and 9am and between 4pm and 7pm. During these times, there is a lot of demand for electricity, so the prices go up. The prices are lowest during the off-peak hours, which are typically between 10am and 3pm and between 10pm and 5am. During these times, there is less demand for electricity, so the prices go down.

iii. Histogram and KDE of demand



Above histogram shows that the demand for electricity is concentrated between 500 MW and 1000 MW. There is a peak in demand around 700 MW. The demand is lower at night and on weekends, and higher during the day and on weekdays. The KDE curve shows that the distribution of demand is slightly skewed to the right.

iv. Heatmap

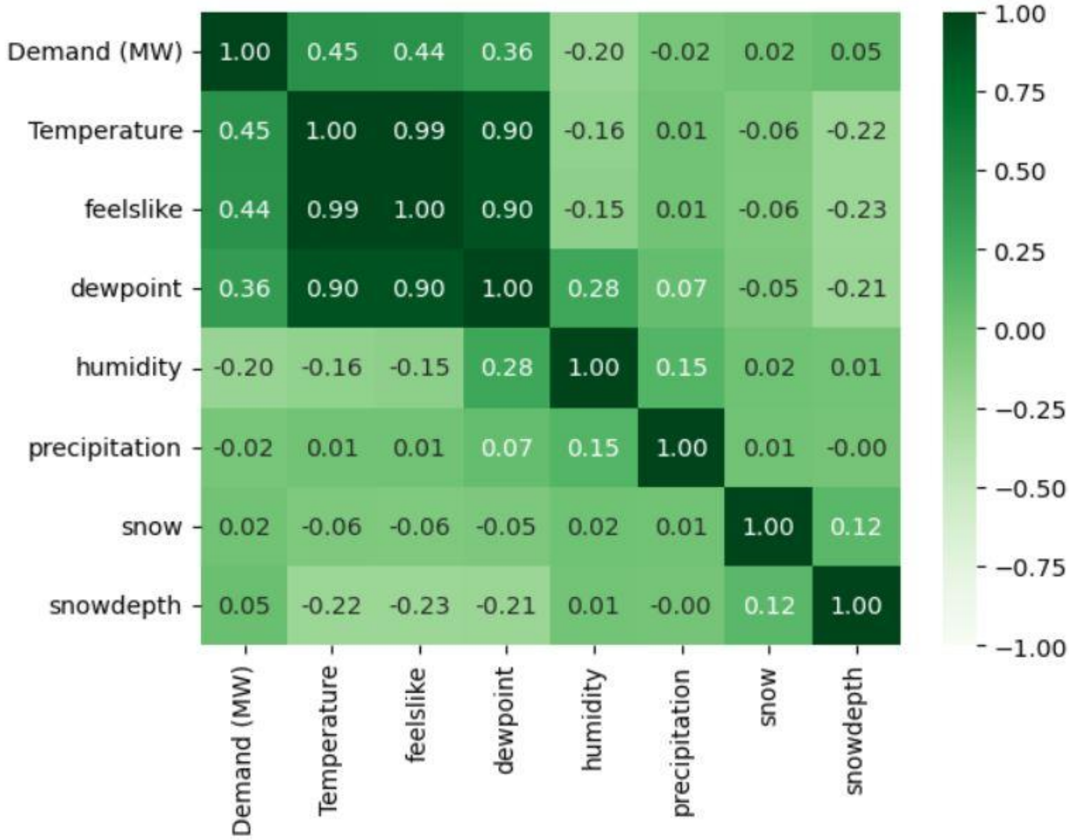


Fig: Correlation heatmap

From analysis regarding the most influencing factors in the data set demand it is found that temperature seems to have the highest importance, followed by dewpoint. We can ascertain the major influencing factors by plotting the correlation between variables in which correlation is high.

We observed from the given heatmap of the correlation there isa high correlation between:

Demand-Temperature = 0.45

Demand-Dew point = 0.36

Demand-Solar Radiation = 0.38

v. Regression Plot

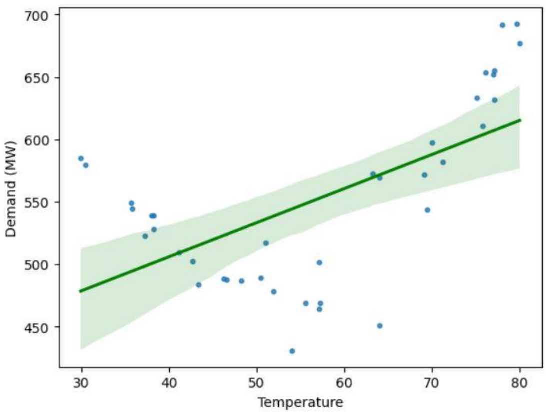


Fig: Regression plot of demand vs temperature

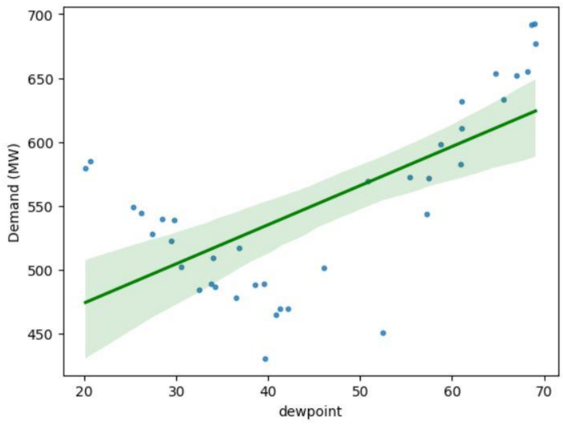


Fig: Regression plot of demand vs dewpoint

The plot of demand vs temperature is upward slopping confirming positive association. The data points are scattered around the regression line suggesting some variations in demand line for a given temperature. The second plot shows positive correlation between dew-point and demand.

Through this thorough exploration of the datasets, we lay the groundwork for subsequent model development. The insights gained from EDA inform our choice of features, model architectures, and preprocessing steps, ensuring a robust and informed approach to hourly price and demand forecasting.

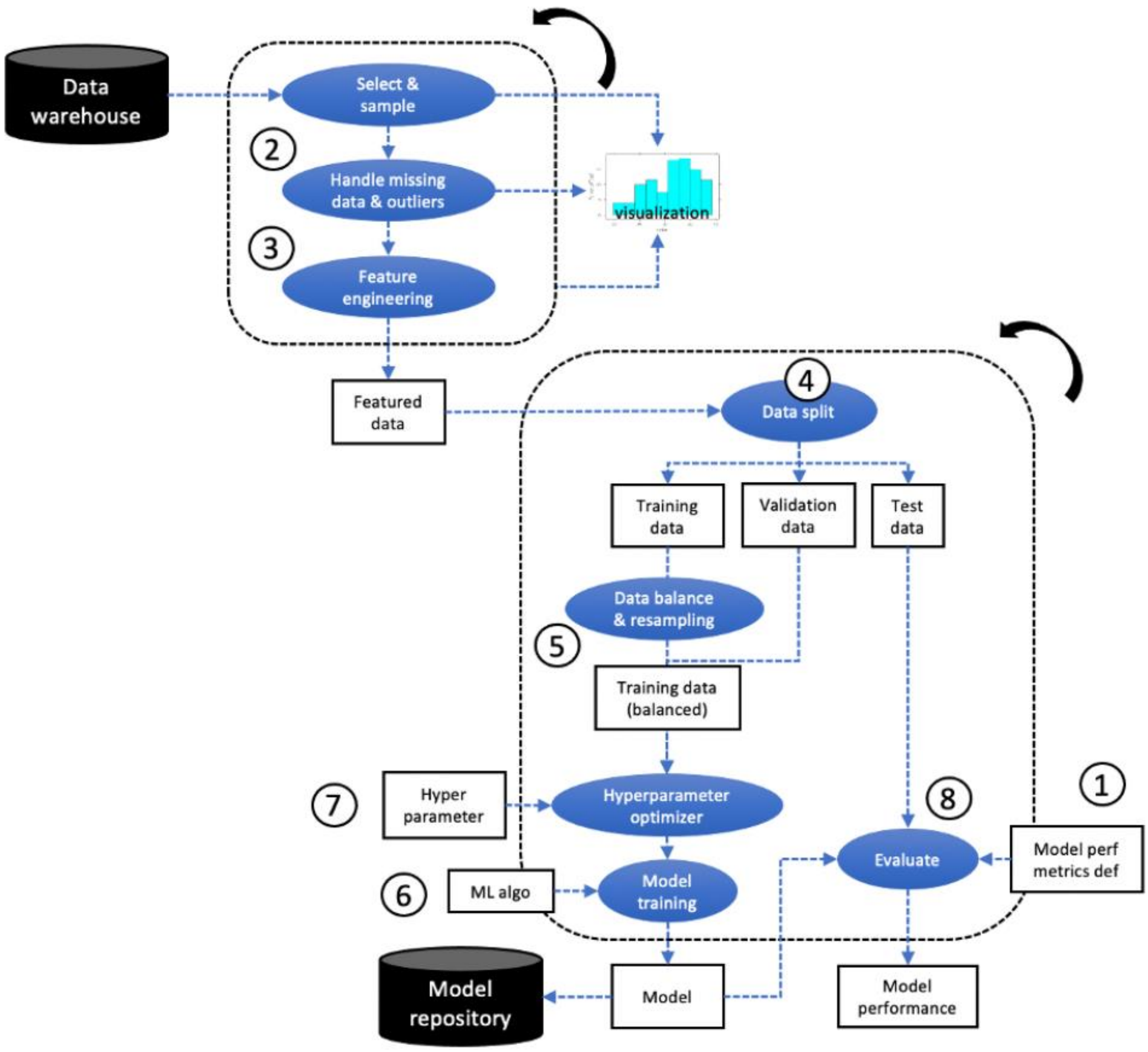


Fig: Model Architecture

3.3 Forecasting the Demand

The dataset is unique because it contains hourly data for electrical consumption and the weather data, the main focus is on predicting consumption a using MSE calculation using various methods. After pre-processing of the data, a new temporary dataset is made (for calculations and to make original data set unchanged) and the dataset is classified into Training set 80% and Testing set 20% with results as below: Training set: (22041), Testing set: (5511)

In our demand forecasting we implemented a diverse set of regression models to predict future demand. The selected models were as follows:

1. Linear Regression - Linear Regression is a simple and commonly used regression algorithm that models the relationship between the dependent variable and one or more independent variables by fitting a linear equation to the observed data. The MSE obtained using this model is:

2. Lasso Regression - Lasso Regression, or Least Absolute Shrinkage and Selection Operator, is an extension of linear regression that adds a penalty term for the absolute values of the coefficients. MSE obtained:

3. Ridge Regression - Ridge Regression is another extension of linear regression that adds a penalty term for the squared values of the coefficients. It helps to prevent multicollinearity in the dataset. MSE obtained:

4. K-Neighbors Regressor - K-Neighbors Regressor is a non-parametric algorithm that predicts the target variable by averaging the values of its k-nearest neighbors. MSE obtained:

5. Decision Tree Regressor - Decision Tree Regressor builds a tree structure to represent a set of decisions based on the input features. It predicts the target variable by traversing the tree. MSE obtained:

6. XGBoost Regressor - XGBoost (Extreme Gradient Boosting) is a powerful gradient boosting library that builds an ensemble of weak learners (typically decision trees). It focuses on boosting both speed and performance. MSE obtained:

7. CatBoost Regressor - CatBoost Regressor is a gradient boosting library that handles categorical features efficiently without the need for extensive preprocessing. It's designed for better performance on categorical data. MSE obtained:

8. AdaBoost Regressor - AdaBoost (Adaptive Boosting) is an ensemble learning method that combines multiple weak learners to create a strong learner. It assigns weights to misclassified data points to focus on improving their performance in subsequent iterations. MSE obtained:

The performance of each model was assessed using the Mean Square Error (MSE), Mean Absolute Error (MAE) and R2 Score as evaluation metric. After comprehensive evaluation, we identified the model with best performance metrices i.e. XG Boost Regressor as our model for price forecasting task.

3.4 Predicting the electricity price

For our hourly price forecasting task, we implemented SARIMA model. SARIMA stands for Seasonal AutoRegressive Integrated Moving Average. It is a variation of the ARIMA (Autoregressive Integrated Moving Average) time series forecasting model that incorporates seasonality. SARIMA models are particularly useful for predicting time series data that exhibit both trend and seasonality. As the dataset we are using has certain trends and seasonality we found this model best for our task. The dataset was split into Training Set (80%) and Test Set (20%).

3.5 Results and Evaluation

Various performance metrices were used to evaluate the trained model.

Table 1. Performance Evaluation of Demand Forecasting.

Mean Squared Error	0.006576619
Mean Absolute Error	0.06409737
R2 Score	0.75204408

Table 2. Performance Evaluation of Price Forecasting

Mean Squared Error	0.00411050
Mean Absolute Error	0.04669915
R2 Score	-0.00028715

4. CONCLUSION

From analysis regarding the most influencing factors in the demand and weather data set it is found that Temperature seems to have the highest importance, followed by dewpoint. On Forecasting the Energy Demand, it has been found that XGBoost Regressor model fits the most and for predicting the energy price SARIMA technique fits the best. Accurately projecting future power consumption and price is essential for effective energy management, cost savings, and environmental sustainability given the rising demand for energy. It is important to keep in mind that forecasting electricity consumption is a challenging process that calls for careful consideration of a number of variables, including seasonality, time of day, and weather. To make accurate forecasts, it is essential to choose the right characteristics and models.

5. References

[1] researchgate.net Electricity consumption prediction using machine learning

[2] researchgate.net Hourly Energy demand generation and weather Electrical demand, generation by type, prices and weather in Spain

[3] vmware.com Machine Learning Model-Development Lifecycle

[4] Wikipedia

[5] chat.openai.com (ChatGPT)