Team - MetaMinds
Category: Energy Forecasting

Problem Statement:

We aim to develop forecasting models for both energy demand and pricing over the next 7 days. Leveraging historical demand data and comprehensive weather information, we seek to address the challenges posed by the volatile and autoregressive nature of the data. Our goal is to create a sophisticated predictive framework that considers the inherent variability, capturing the interplay between demand and weather factors. The outcome will empower stakeholders to make informed decisions, optimizing resource allocation and financial planning in a dynamically changing energy landscape.

Methodology:

1. Given Data Exploration

- **Demand Forecasting Demand Data upto Feb 21.csv :** 27552 entries from 1/1/2020 0:00 to 2/21/2023 23:00
- **Demand Forecasting Weather Data upto Feb 28.csv :** 27720 entries from 1/1/2020 0:00 to 2/28/2023 23:00 - Lots of missing values
- **Price Forecasting data upto December 24.csv :** 35352 entries from 1/1/2020 H24 (00:00) to 12/24/2023 H23 (23:00)

2. Basic Data Cleaning

- 5 columns in demand weather data - (Unnamed: 21 ... Unnamed : 25) NULL DATA removed.

3. Demand Dataset Formation

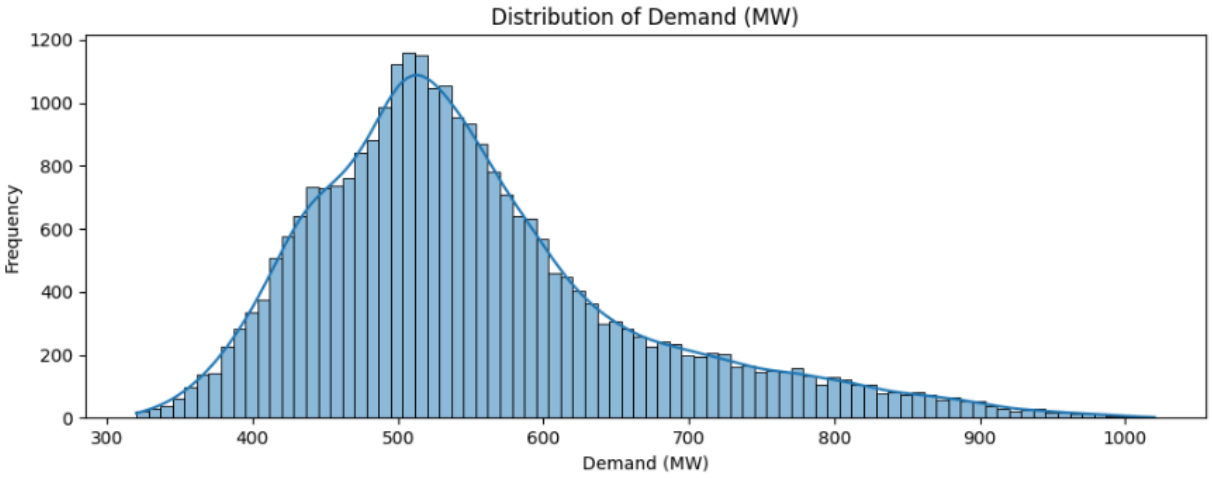
OUTER JOIN of Demand Data and Weather Data on datetime field - All datetime field in demand data has datetime matched with the weather data

4. Exploratory Data Analysis

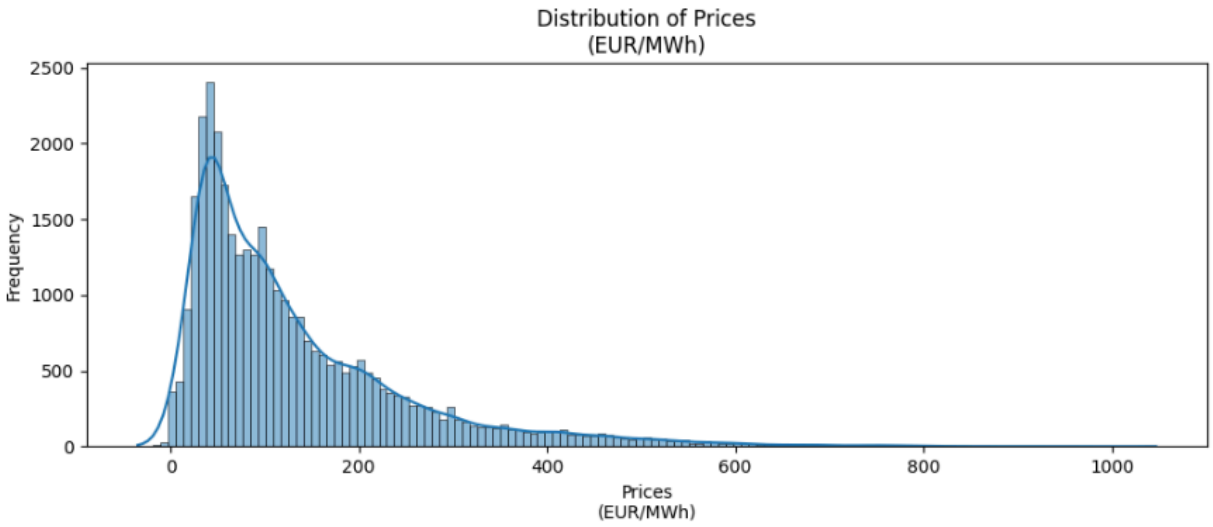
a. **Descriptive statistics** - Mean, standard deviation, minimum and maximum, first, second and third quartile

	Prices\n(EUR/MWh)	Demand (MW)	Temperature	feelslike	dewpoint	humidity	precipitation
count	34899.000000	27555.000000	27720.000000	27720.000000	27720.000000	27720.000000	27720.000000
mean	132.955080	551.811787	56.846847	55.505018	45.486089	68.485757	0.005618
std	120.246525	114.435250	18.734390	21.558142	18.198644	18.189507	0.044799
min	-35.000000	320.000000	-5.900000	-30.100000	-14.900000	17.100000	0.000000
25%	47.985000	474.500000	42.200000	38.000000	29.900000	55.172500	0.000000
50%	95.440000	529.500000	56.900000	56.900000	46.100000	70.540000	0.000000
75%	179.245000	601.900000	72.000000	72.000000	62.100000	83.960000	0.000000
max	1047.100000	1020.200000	99.000000	110.400000	79.100000	100.000000	2.376000

b. Distribution plot of the numerical variables:

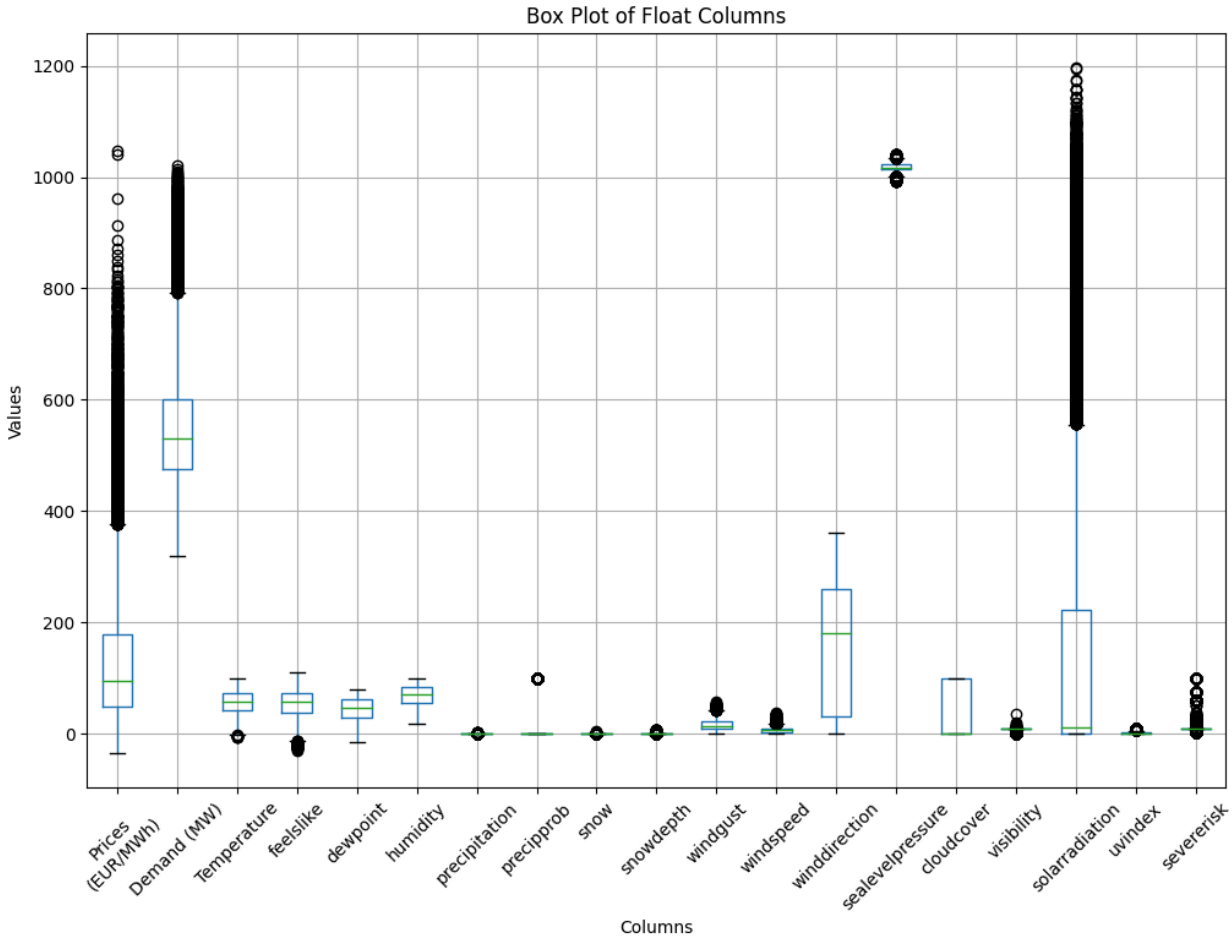


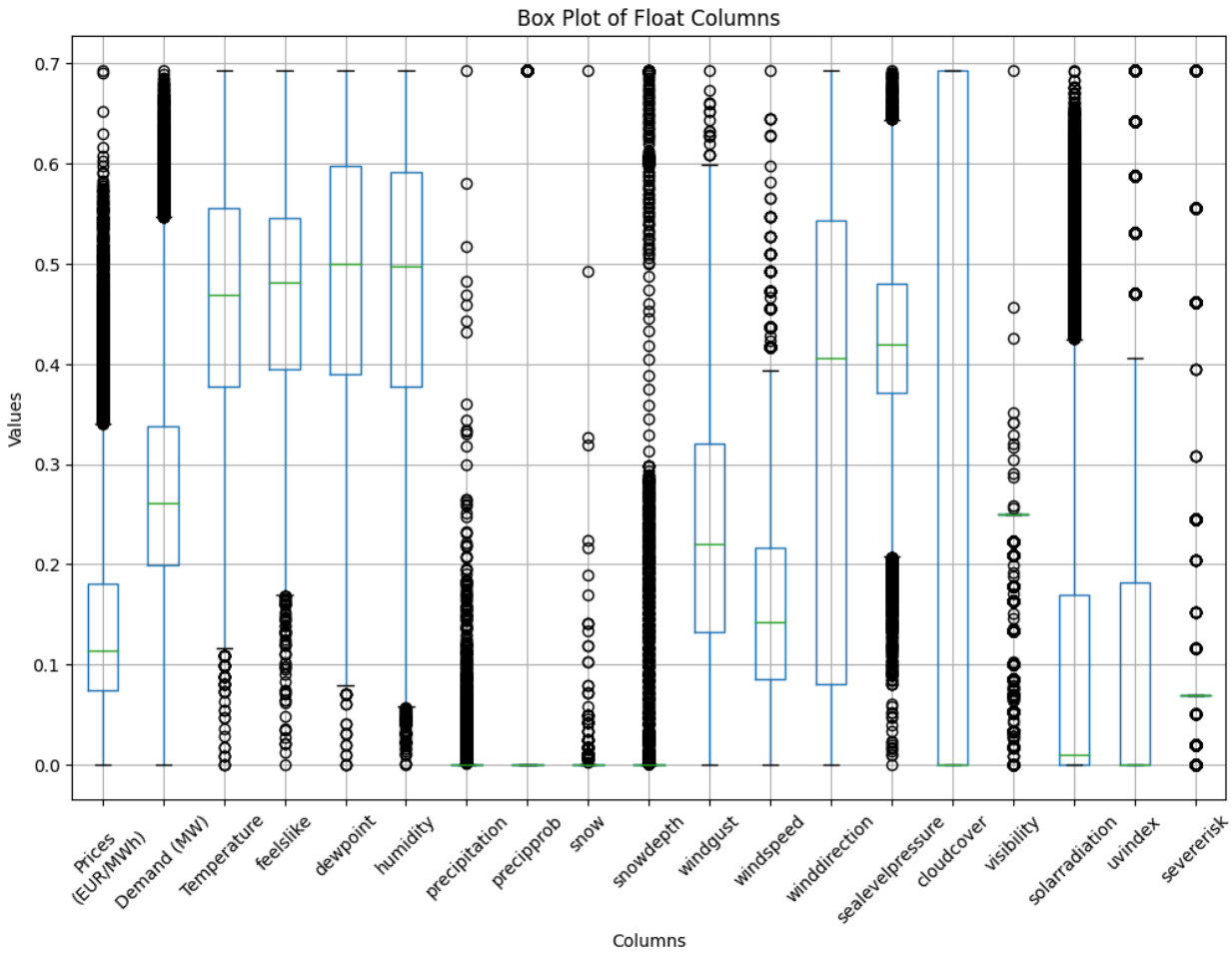
Nearly normal distribution for the demand data.



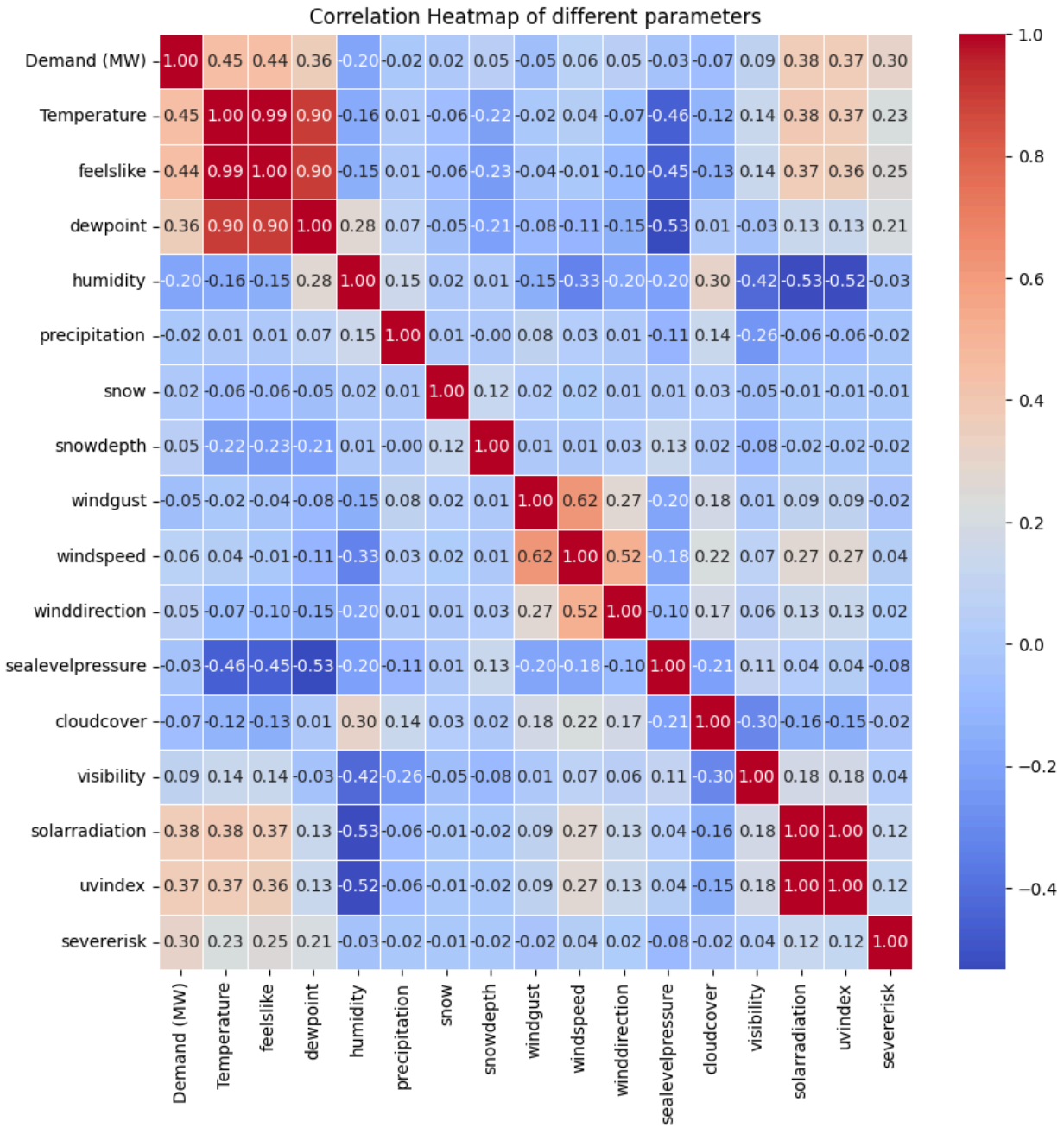
Left-skewed curve for the price data

c. Box plot for outliers detection:



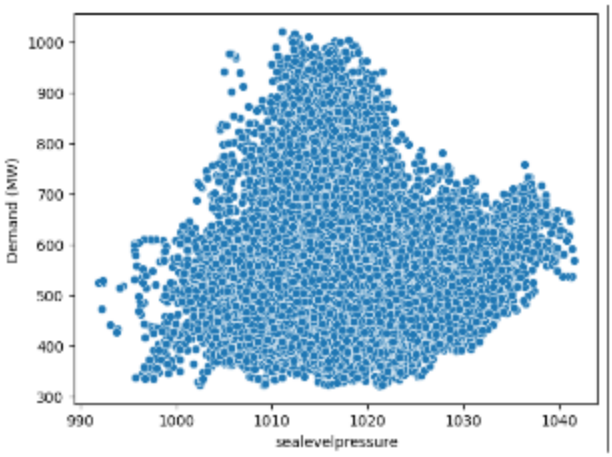

Box Plot of Float Columns
Values
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Prices (EUR/MWh)
Demand (MW)
Temperature
feelslike
dewpoint
humidity
precipitation
precipprob
snow
snowdepth
windgust
windspeed
winddirection
sealevelpressure
cloudcover
visibility
solarradiation
uvindex
severerisk
Columns

d. **Correlation analysis using the heatmap:**



The above heatmap shows high corelation between temperature and feelslike and perfect corelation between solarradiation and uvindex. While some other features have near 0 co-relation i.e. their corelation is not shown by linear co-relation so we move towards non-linear corelation with scatter plot.

e. **Scatter plot to capture non-linear co-relation:**



The above graph shown that demand decreases with increase in sealevelpressure.

f. Line Plot of DEMAND VS Different Continuous Parameters:

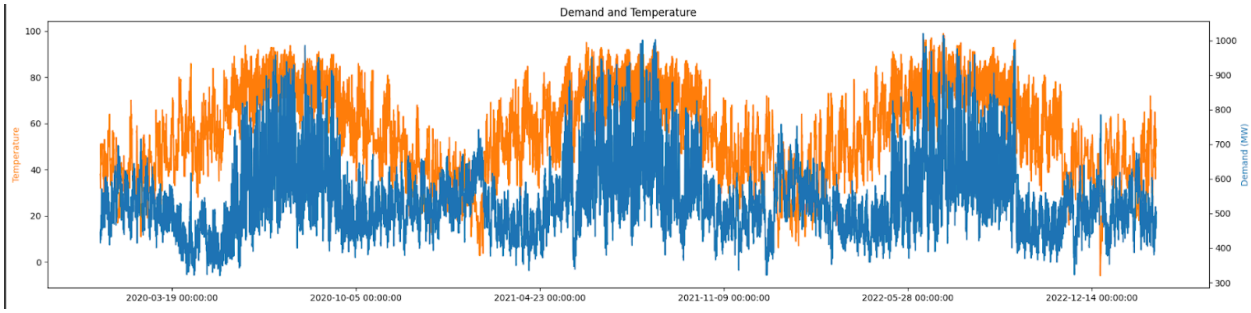


fig: Line plot of Demand vs Temperature to visualize their correlation

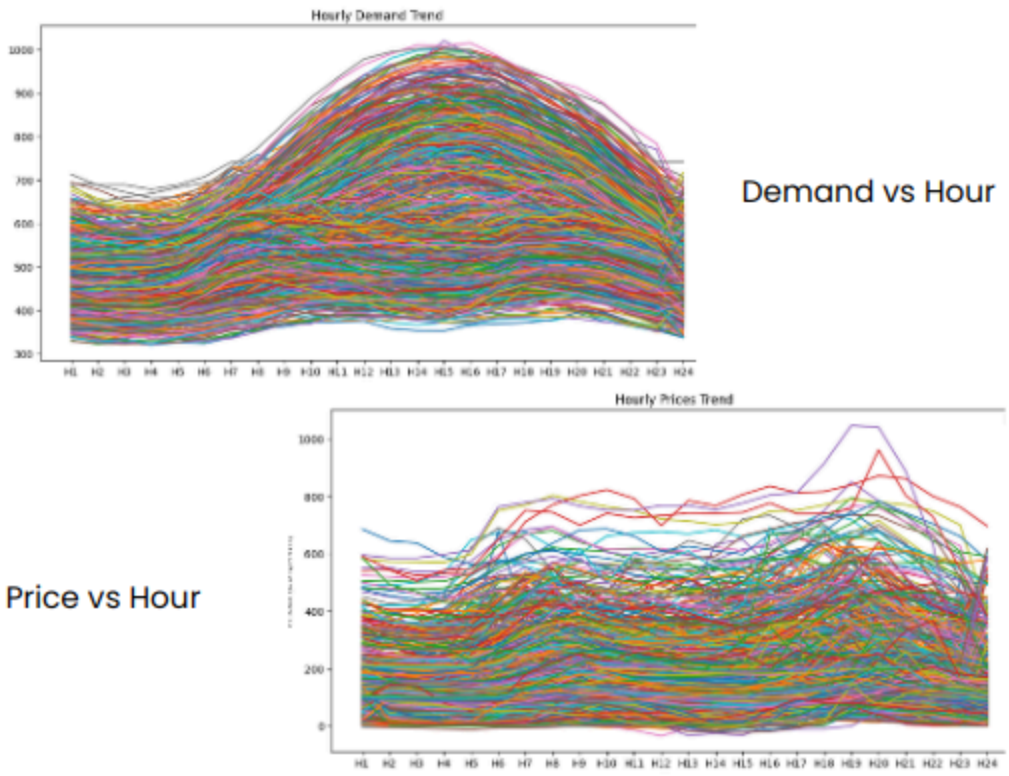
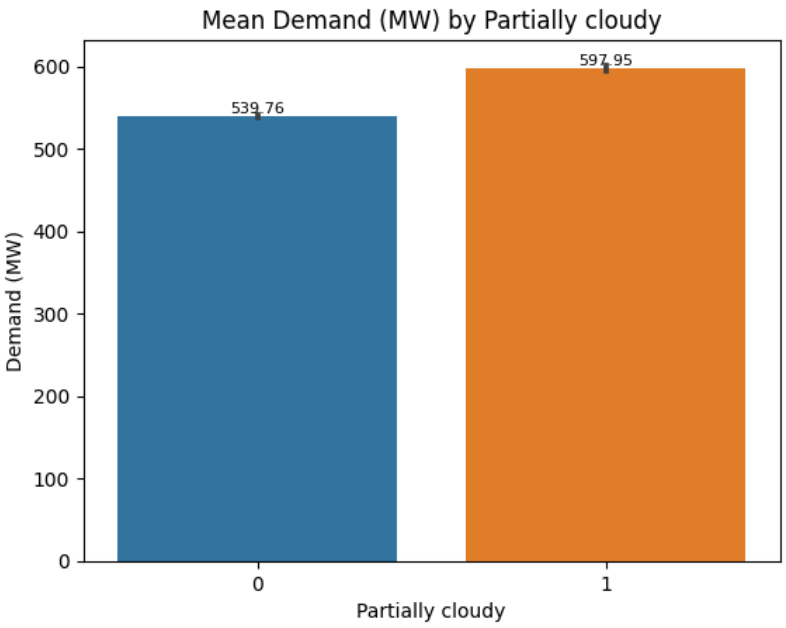


Fig: Hourly plot for each day

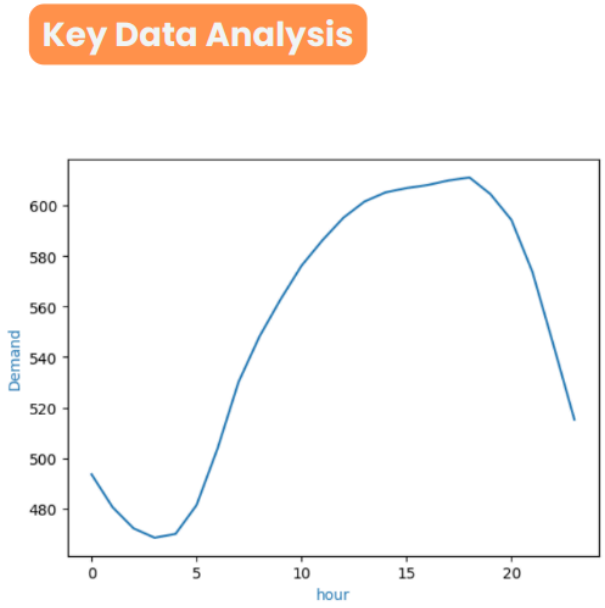
g. Bar graph for categorical Values:

For the categorical variables, the mean of the demand is calculated for the presence and absence of each category.

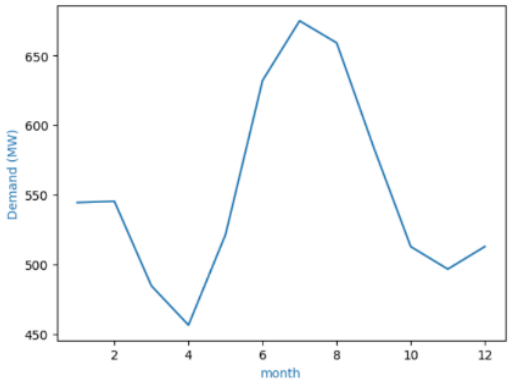


h. Hourly and daily analysis:

The below graph shows the peak energy demand for a day and for a month. We found that the peak hour for a day is 6-8 hours. Also, while analyzing the demand for a year, peak energy consumption is between June-August.



Peak hour: 6-8 PM



Peak Energy Consumption on June-July-August

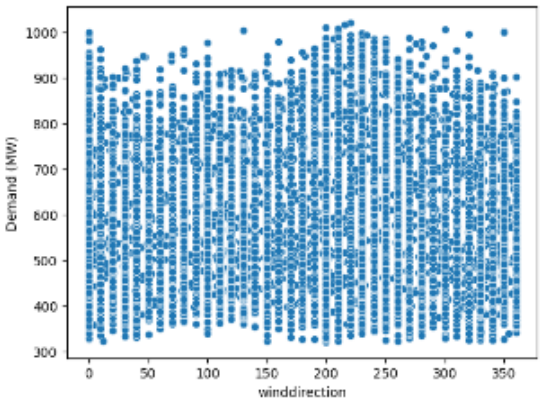
5. Data Transformation

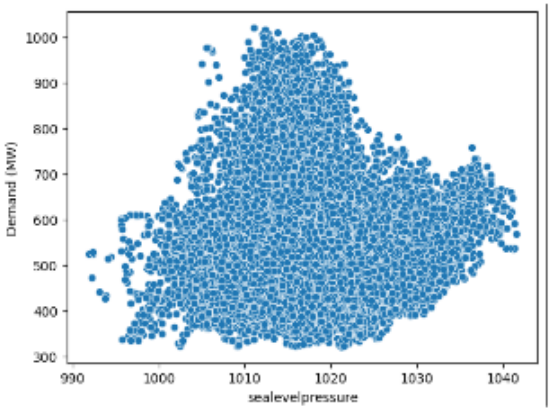
- Data Normalization using min-max normalization
- Performed log transformation in the min-max normalization data for outliers handling
- Convert categorical value i.e precipitation type and conditions into the one hot encoding format

6. FEATURE EXTRACTION:

a. **Dropping feelslike,uvindex by observing linear co-relation.**
Observing heatmap, we saw high co-relation between temperature and feelslike, and between temperature and uvindex, so we dropped feelslike and uvindex.

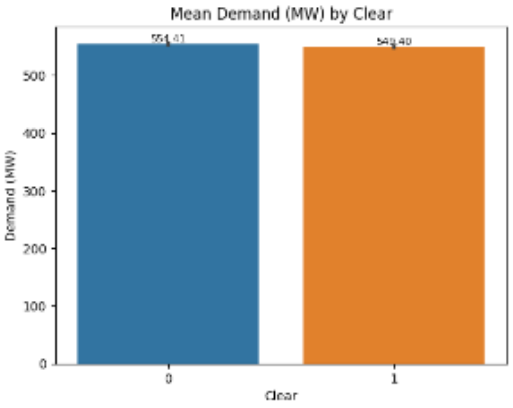
b. **Non-linear co-relation of features with demand:**



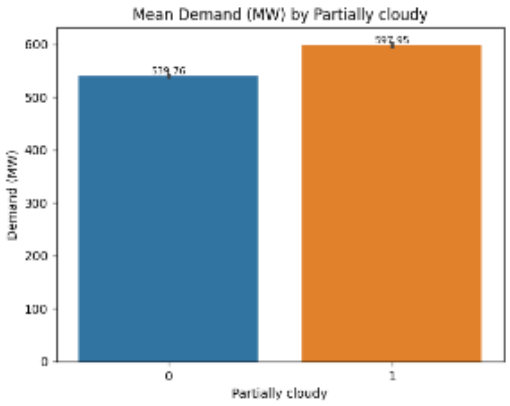


In the above figure, wind-direction has no co-relation with demand, so we dropped wind-direction; while kept sealevelpressure as feature as it has co-relation with demand. We selected features similarly for other continuous features too.

c. **Feature Selection for Categorical Value:**



Difference ~ 6MW



Difference ~ 58MW

While working with categorical features, we plotted its bar graph with demand. In the above graph clear has less difference while partially cloudy has high difference. However, 6MW is also a significant change. We observed similar results for other categorical features too. So, we selected all the categorical features.

5. DATA IMPUTATION

- **Mean and Interpolation**

Distribution of 'windgust', 'severerisk',concentrated at a point so **fill NaN with mean**

Interpolation in remaining columns

6. TRAIN-TEST SPLIT

- The first 20% of the data used for testing
- The remaining 80% of the data used for training

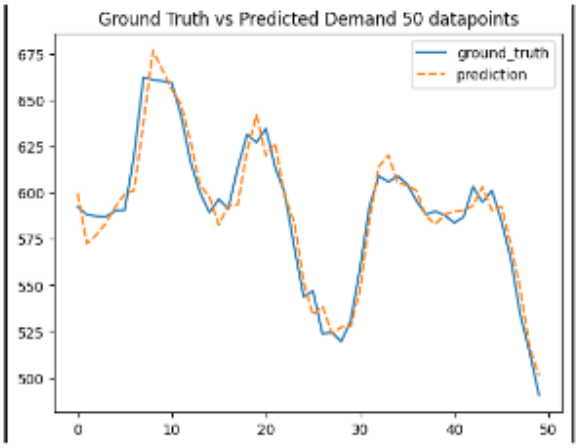
7. Model Building

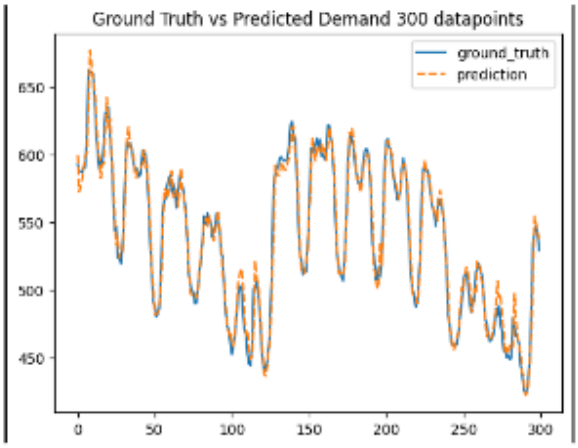
A. UNIVARIATE RANDOM FOREST REGRESSOR - Unit step prediction

a. Model Building:

- Input prior 3 weeks demand data = 24 * 7 * 3
- Output = Next Hour Data

Univariate Demand Random Forest Regressor

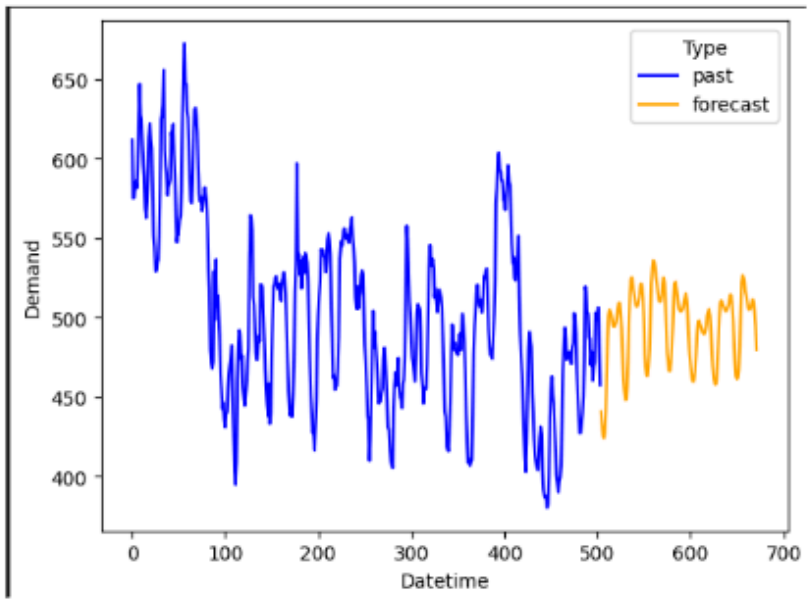




Mean squared Error = 133.1053
Mean Absolute Percentage Error= 1.651%,
Mean Absolute Error= 8.72

b. Demand Forecasting:

Univariate Demand Random Forest Regressor Test Data Prediction

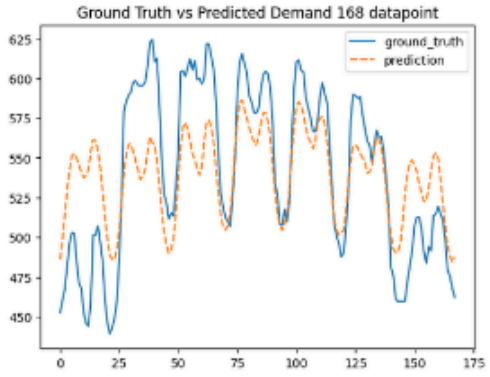


The above graph shows the demand forecasting for demand data taking a univariate model i.e. taking into consideration date-time only.

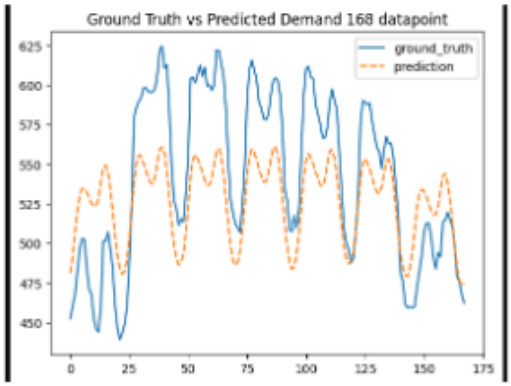
B. MULTIVARIATE RANDOM FOREST REGRESSOR - Multi-Step Prediction (1 Week Prediction)

- Input prior 3 weeks demand data = 24 * 7 * 3
- Output = 24 * 7

Multivariate Random Forest Regressor :



n_estimators = 30
Mean squared Error = 1405.814,
Mean Absolute Percentage Error=0.058979,
Mean Absolute Error=31.288



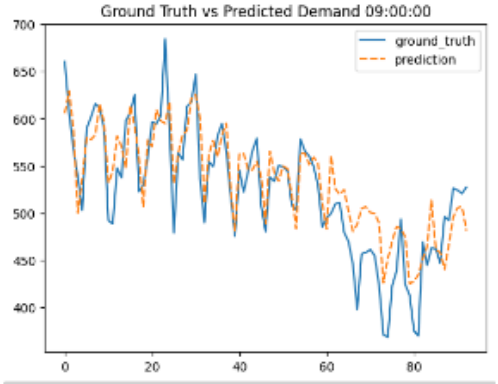
n_estimators = 200
Mean squared Error = 1621.5309
Mean Absolute Percentage Error=0.06597,
Mean Absolute Error=35.9930

-

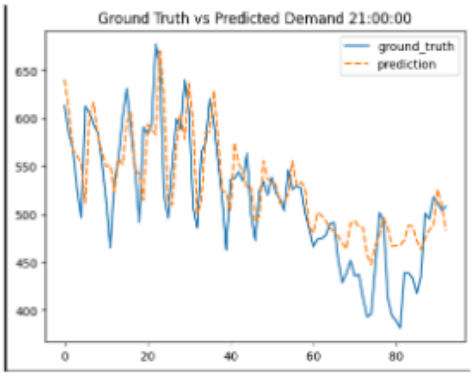
C. Hourly Multivariate Random Forest Regressor - Multi-Step Prediction

- 24 Models for all hours -> H1 … H24
- Prior 3 week data = 1 * 7 * 3
- Prediction = 7 Data

Hourly Demand Forecasting using Random Forest Regressor



MSE = 1219.906
MAPE = 5.7688%
MAE = 28.431

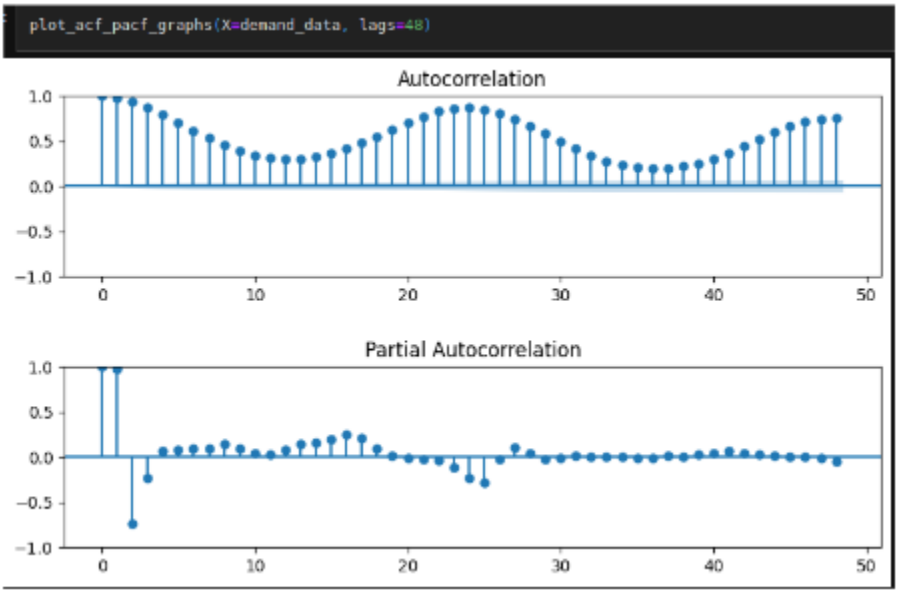


MSE = 1657.369
MAPE = 0.06457
MAE = 31.799

-

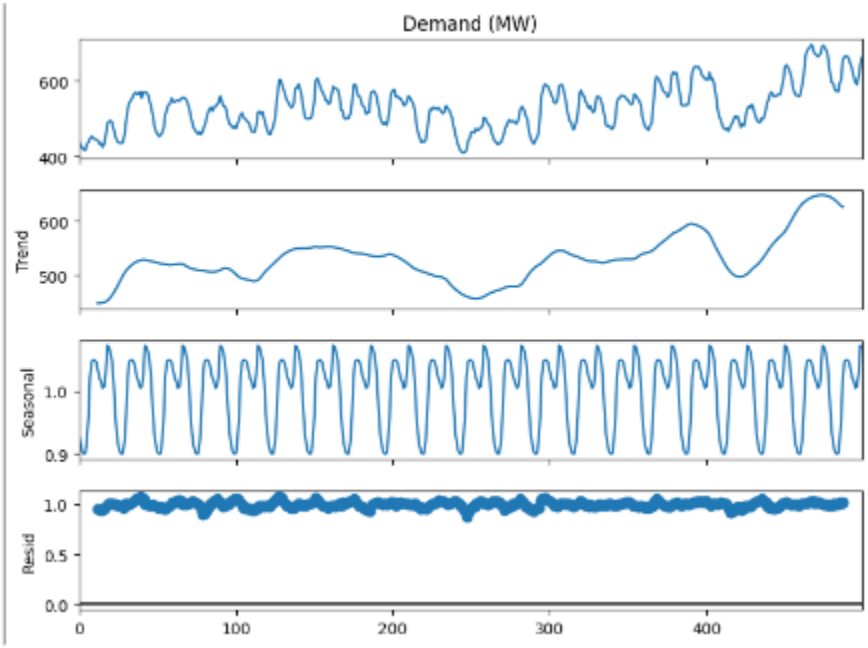
D. SARIMA Model:

a. ACF AND PACF PLOT FOR THE HOURLY DATA EXPLANATION:



Through the ACF graph, we can see that the data is seasonal. And PACF plot shows that auto-correlation moving to zero after 3rd lag.

b. Multiplicative Seasonal Decomposition of Demand Series:



The above decomposition shows the upward trend and seasonality in the data, and the residue.

c. ADF Test:

We found out that the data is stationarity through ADF Test.

d. SARIMA Model Building:

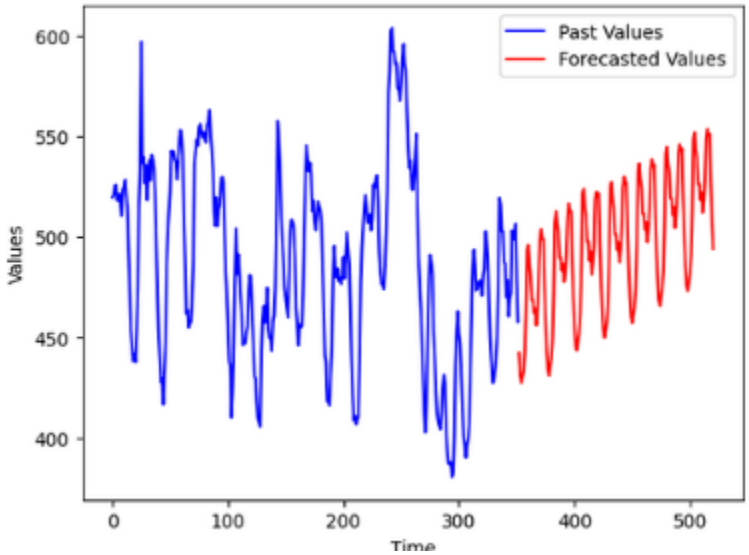
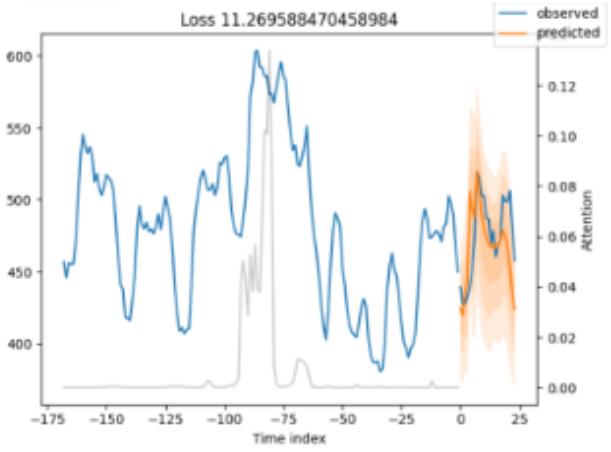
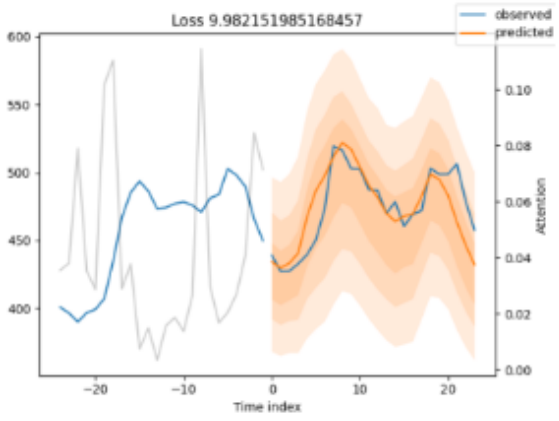


Fig: Forecasted value with SARIMA Model

E. TFT (Temporal Fusion Transformer):



Previous data = 7 days
Prediction = 1day



Previous data = 1 days
Prediction = 1day

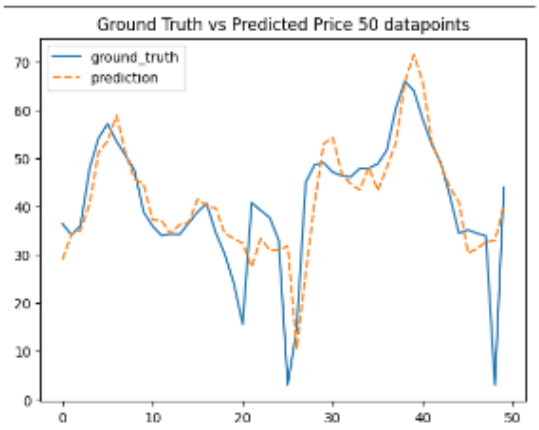
Loss is higher when forecasted data with reference to the previous 7 days compared to that forecasted taking in reference of previous one day.

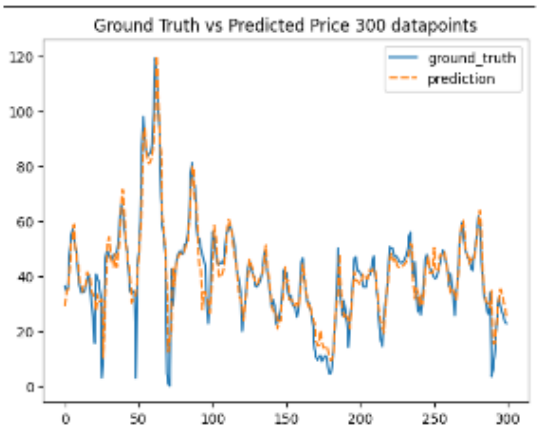
The gray plot shows the magnitude of attention shown by transformer for previous demand sequences. The condition variables were used as categorical known types where as rest of the features were used as real known type of input. TFT showed promising result but we were a bit short on time to perform more hyperparameter tuning.

8. Price Forecasting

a. Model Building:

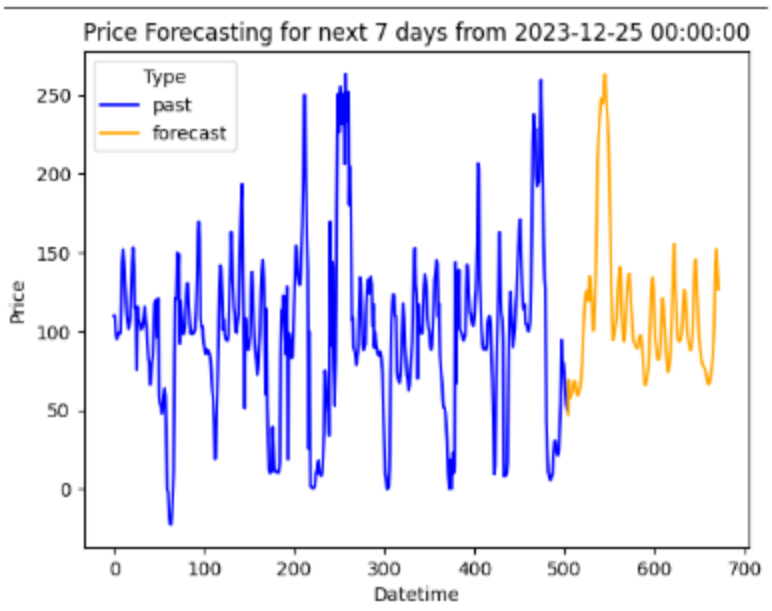
We build model for price forecasting using Random Forest Regressor.





Mean Squared Error = 37.1401
Mean Absolute Error = 3.9510 EUR/MWh

b. Price forecasting for next 7 days:



The above graph shows the price forecasting for the next 7 days using a random forest regressor.