**Energy Analytics Candidate Test**

Q3) Once you have collected and converted data, please answer following questions:

**a)** How many missing hourly temperature observations are in the dataset? (Provide the dates and hours missing if there are any and interpolate any missing values)

**Ans) Missing hourly temperature observations: 1**

**Below is the record**

| **Index** | **Year** | **Month** | **Day** | **Hour(UTC)** | **Ept(Date and hour)** |
| --- | --- | --- | --- | --- | --- |
| **40096** | **2023** | **7** | **29** | **22** | * + 1. **8:00:00** |

**Interpolated missing hourly temperature is 23.045°C for that observation.**

**b)**What is the mean air temperature for all observations in July 2021?

**Ans) Mean air temperature for July 2021: 71.835 °F or 22.131°C**

Q4) Generate Gas Day temperatures

The gas day starts at 10AM (EPT) Eastern Prevailing Time and runs to 10AM EPT the following day. The gas day temperature for that day is the average of the 24 hourly air temperature observations. The gas day temperatures should be reported in Eastern Time and in Fahrenheit. You will be aggregating the 24 hourly observations into one average temperature for that gas day.

**a)**What is the Boston gas day temperature for July 4th, 2020?

**Ans) Gas day temperature for July 4th, 2020: 66.76 °F**

Q6) Build a model to forecast Total AGT demand (Res/Comm + Power Plant)

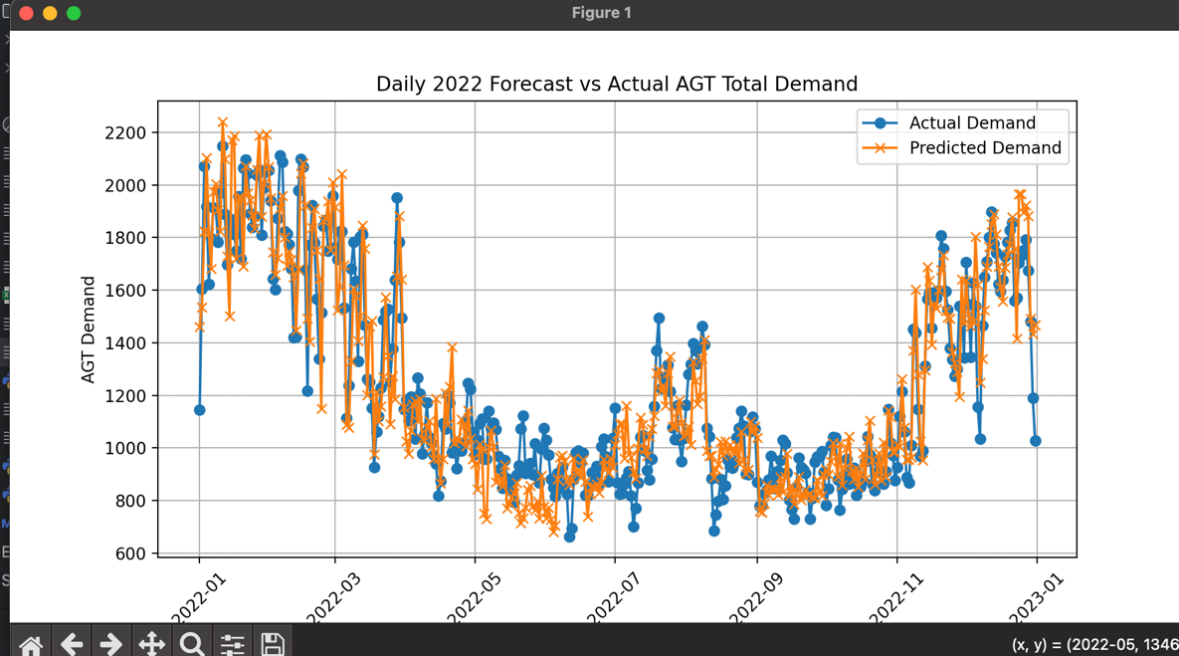
Provide the following:

1. Mean Squared Error (MSE) for your forecast vs actual AGT total demand for 2022

**Mean Squared Error (MSE): 20470.689112338347**

**Mean Absolute Error (MAE): 109.8610191985649**

1. **plot of your daily 2022 forecast vs actual AGT total demand**

****

**c)plot of your daily 2022 forecast vs the gas day temperature**

**A graph on a screen

Description automatically generated**

Q7) Provide a qualitative assessment of your work:

**a) What assumptions did you make and why?**

**Assumptions Made:**

* **Gas Day Temperature Calculation:** I assumed that for each gas day, the temperature recorded from 10 AM Eastern Prevailing Time (EPT) to the next 10 AM EPT would accurately represent the day's temperature for predicting gas demand.
* **Handling Missing Data:** In the NOAA dataset, missing values such as -9999 were treated as invalid data, and I used linear interpolation to fill in missing temperature records. This assumes that temperature changes between missing records follow a linear trend.
* **Features Selection:** I used the daily gas day temperature, the day of the week, and the month as features for forecasting gas demand. This assumes that these factors have a significant effect on AGT demand. No additional socio-economic or external data was included.

**b) Was there anything interesting about the dataset?**

**Interesting Aspects of the Dataset:**

* **Temperature Data:** The NOAA dataset provided detailed hourly data, and I had to aggregate this data into daily gas day temperatures to match the requirements of the problem. This revealed some gaps in the data, which had to be interpolated.
* **Demand Data:** The daily demand for Residential/Commercial and Power Plant sectors showed a cyclical pattern, which likely depends on seasonal and weekly trends (e.g., higher gas demand in colder months).

1. **Why did you choose to construct your model the way you did?**

**Model Construction Reasoning:**

* **Random Forest Regression:** I chose Random Forest because it is a robust and versatile model that handles both linear and non-linear relationships well. It can also handle interactions between variables like gas day temperature, day of the week, and month. Given the potential non-linearity of gas demand (due to factors such as weather and day of the week), Random Forest was a good choice for this task.
* **Feature Engineering:** I included the gas day temperature, day of the week, and month as features. I hypothesized that gas demand would be influenced by both temperature (e.g., colder days increasing demand) and day-related patterns (e.g., higher residential/commercial demand on weekends).

**d) What other data or information do you think would improve your forecast?**

**Suggestions for Improving the Forecast:**

* **Weather Forecast Data:** Integrating weather forecast data such as future temperature predictions, precipitation, and wind conditions might improve the accuracy of the forecast by giving a better picture of upcoming demand.
* **Economic and Social Data:** Variables such as population growth, gas prices, industrial activity, and socio-economic data could provide additional context for predicting demand. For example, economic downturns might lower demand in commercial sectors.
* **Time Series Modeling:** Incorporating a time series forecasting model like ARIMA, SARIMA, or LSTM (deep learning) could capture long-term trends, seasonality, and autocorrelation in gas demand more effectively.
* **Holidays and Events:** Including holiday data could improve predictions for certain spikes or drops in demand, such as increased residential demand during holiday seasons.

These improvements could enhance the model’s ability to capture external influences on demand, leading to more accurate predictions.

