Report of the Hourly load Forecasting for New Hampshire (NH) Electricity Consumption for January 2023

By Team 6

Siva Sushmitha Meduri

Janani Krishnamurthy

Divyasree Vammigari

Dept. of Data Science

1. Introduction

The purpose of this white paper is to present the methodologies, analysis, and results of the project of forecasting hourly load of the electricity consumption in New Hampshire. The three data sets utilized in the project is the historical hourly data from January 2020 to December 2022. Using machine learning models, specifically Linear Regression and Random Forest, we predicted the hourly load for December 2022 for testing the models.

2. Model Statement

Electricity load forecasting is a critical task in the energy industry, enabling traders to anticipate power prices and aiding in the efficient management of energy resources. Accurate load forecasting contributes to cost reduction, resource optimization, and overall grid stability. The goal of this project is to develop robust models for forecasting hourly load in New Hampshire.

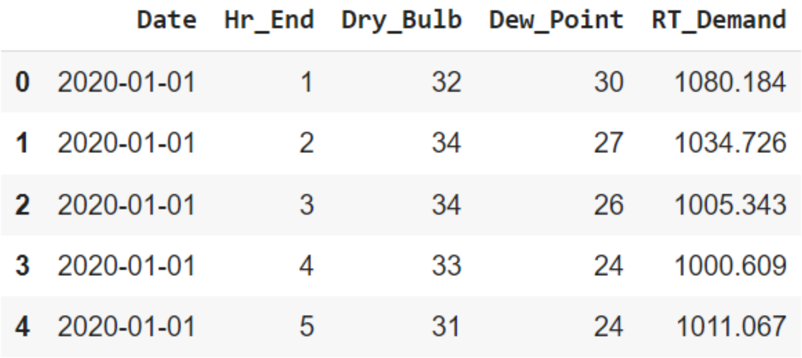
3. Analysis:

3.1 Data Collection and Preprocessing

Three years of historical hourly data (2020-2022) were obtained from the ISO New England website - “https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info”.

- 2022 SMD Hourly Data
- 2021 SMD Hourly Data
- 2020 SMD Hourly Data

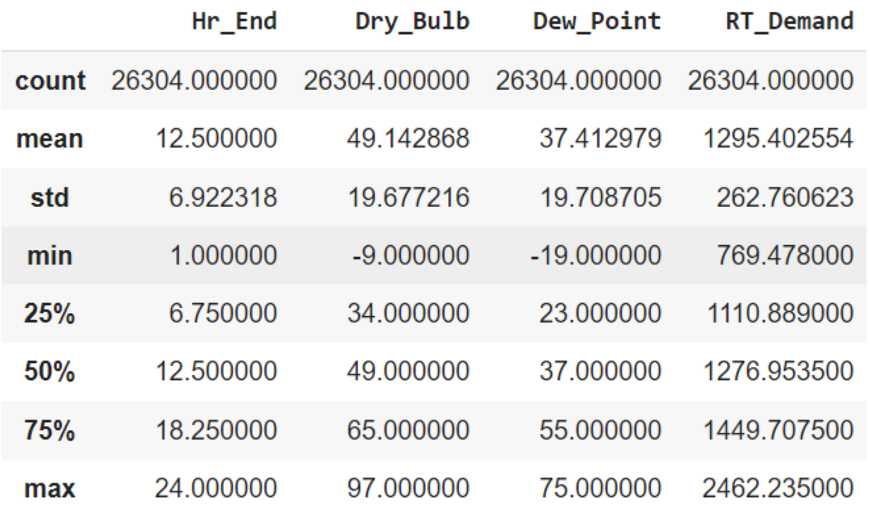
The data included information on temperature, humidity, and date-time attributes. Features such as Dry_Bulb, Dew_Point, and Day_of_Week were engineered to enhance model performance. We have removed the pricing related data from the data sets and combined all three data sets in to single data set named “hourly_data_NH”. The below table is a sample of top five data columns of the final data.



	Date	Hr_End	Dry_Bulb	Dew_Point	RT_Demand
0	2020-01-01	1	32	30	1080.184
1	2020-01-01	2	34	27	1034.726
2	2020-01-01	3	34	26	1005.343
3	2020-01-01	4	33	24	1000.609
4	2020-01-01	5	31	24	1011.067

3.2 Summary Statistics:

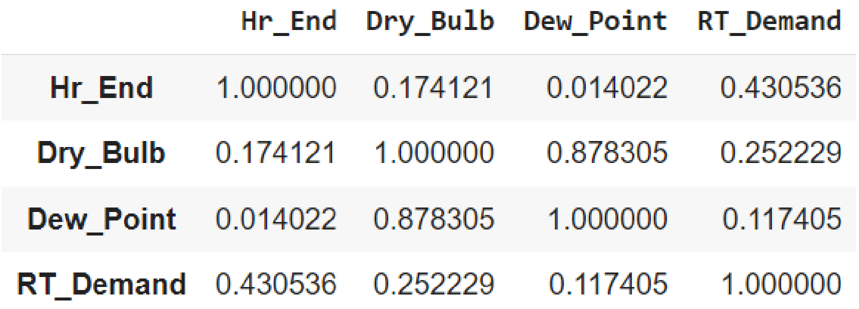
The below table shows the summary statistics of the data we have utilised for the model prediction.



	Hr_End	Dry_Bulb	Dew_Point	RT_Demand
count	26304.000000	26304.000000	26304.000000	26304.000000
mean	12.500000	49.142868	37.412979	1295.402554
std	6.922318	19.677216	19.708705	262.760623
min	1.000000	-9.000000	-19.000000	769.478000
25%	6.750000	34.000000	23.000000	1110.889000
50%	12.500000	49.000000	37.000000	1276.953500
75%	18.250000	65.000000	55.000000	1449.707500
max	24.000000	97.000000	75.000000	2462.235000

3.3 Correlation:

We have calculated the correlation between the variables and the table below holds all the information.



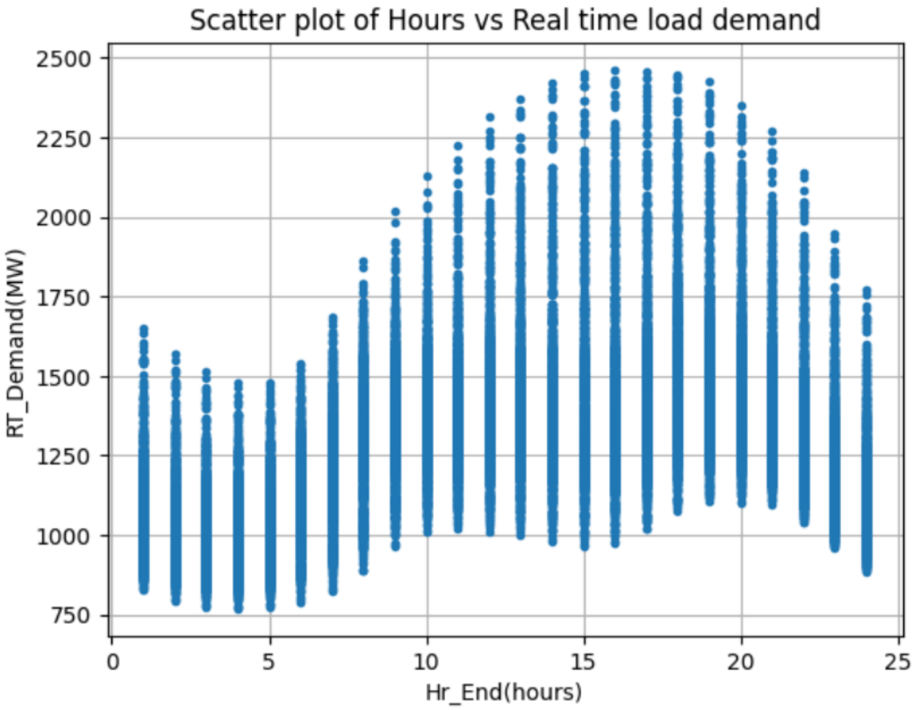
	Hr_End	Dry_Bulb	Dew_Point	RT_Demand
Hr_End	1.000000	0.174121	0.014022	0.430536
Dry_Bulb	0.174121	1.000000	0.878305	0.252229
Dew_Point	0.014022	0.878305	1.000000	0.117405
RT_Demand	0.430536	0.252229	0.117405	1.000000

4. Data Visualization:

In this section we have shown the relationship between the variables utilized and inferences drawn from them.

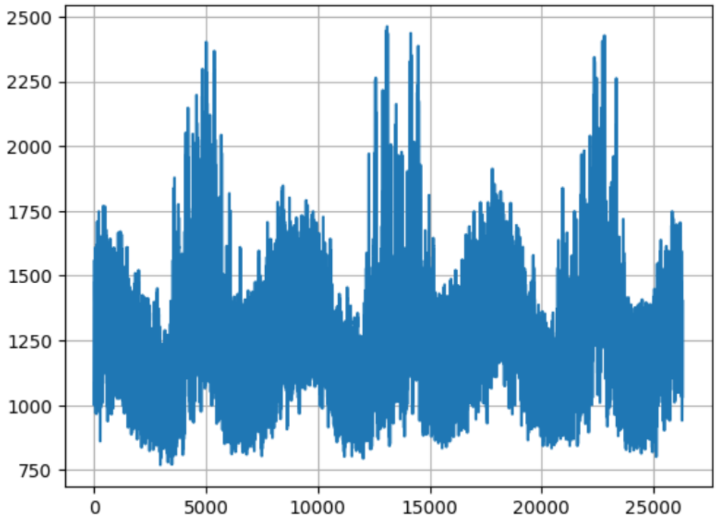
The graph below is a scatterplot of Real time Load(MW) along the hours of a day.

- It is evident that the electricity load is higher during the noon than the rest of the day.
- The peak of the load is during around 15:00 hours



Seasonality:

Similarly, we have the plot for Load demand for all three years 2020, 2021, 2022. The three peaks in the graph shows the highest load demand is during summer seasons of all three years. This is a time-series plot of the load demand against number of hours.

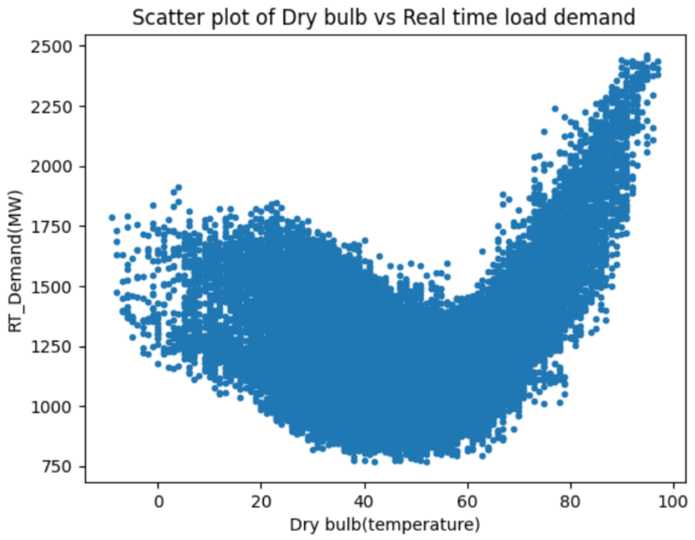


Hours

The three-year seasonality graph demonstrates that the seasonal pattern of RT demand is very stable year after year. However, there are some year-to-year fluctuations in demand. This is most likely due to a mix of factors such as seasonal trends, holidays and special events, and weather conditions.

Relationship between Temperature and Load:

As, it is evident from the seasonality that there is an effect on load demand due to temperature, we have plotted a scatterplot for load demand against temperature.



The scatter plot shows that the real time load demand is higher when the Dry Bulb temperature is higher. This is because people tend to use more electricity to cool their homes and businesses when the weather is hot.

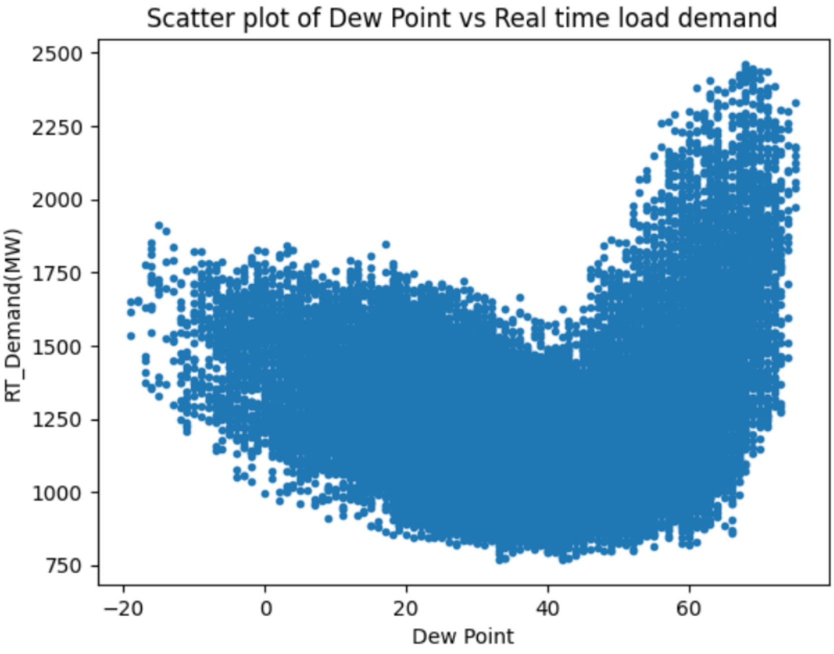
The relationship between Dry Bulb temperature and Real time load demand can be explained by the following factors:

- When the Dry Bulb temperature is higher, the air feels hotter. This is because our bodies cool down by evaporating sweat. When the air is hot and dry, the sweat evaporates more quickly, which helps us to cool down. However, when the air is hot and humid, the sweat evaporates more slowly, which makes us feel hotter. To compensate, we turn up the thermostat on our air conditioners, which increases electricity consumption.
- Hot air also contains more energy. This is because hot air molecules are moving more quickly than cold air molecules. When the air is hot, the electrons in the air molecules are also moving more quickly. This makes it easier for electricity to flow through the air, which can lead to increased power transmission losses. These further increases electricity consumption.

The scatter plot also shows that there is a lot of variation in the Real time load demand for any given Dry Bulb temperature. This is because there are many other factors that affect electricity consumption, such as the time of day, the day of the week, the weather, and the price of electricity. Overall, the scatter plot shows that there is a positive relationship between Dry Bulb temperature and Real time load demand. However, there is also a lot of variation in the Real time load demand for any given Dry Bulb temperature, due to other factors.

Relationship between Humidity and Load:

Similarly, to draw a relation between humidity and load demand.



The scatter plot shows that the real time load demand is higher when the Dew Point is higher. This is because people tend to use more electricity to cool their homes and businesses when the weather is humid.

The relationship between Dew Point and Real time load demand can be explained by the following factors:

- When the Dew Point is higher, the air is more humid. Humid air feels hotter than dry air, even at the same temperature. This is because humid air prevents our sweat from evaporating, which is how our bodies cool down. To compensate, we turn up the thermostat on our air conditioners, which increases electricity consumption.
- Humid air also contains more water vapor. Water vapor is a good conductor of electricity, so humid air increases the conductivity of the air. This can lead to increased power transmission losses, which further increases electricity consumption.

The scatter plot also shows that there is a lot of variation in the Real time load demand for any given Dew Point. This is because there are many other factors that affect electricity consumption, such as the time of day, the day of the week, the weather, and the price of electricity.

Distribution of Load Demand:

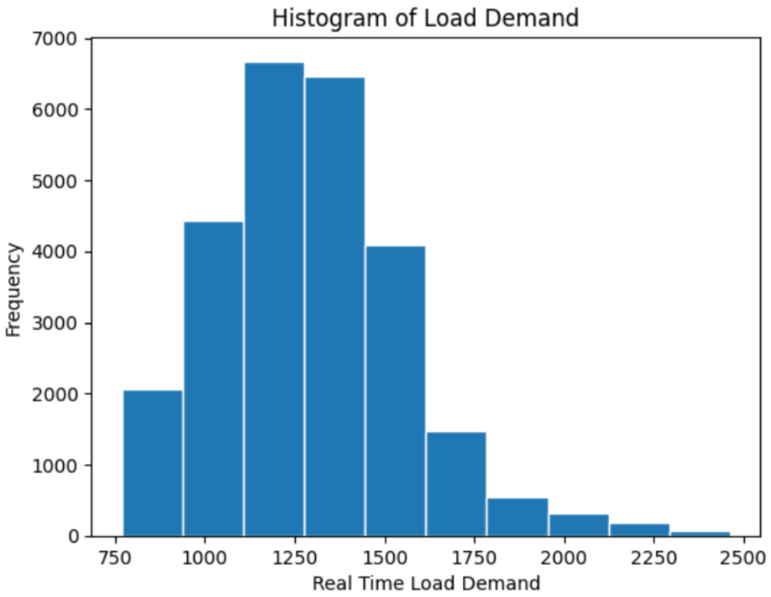
The histogram shows that the load demand is most frequently between 1100 and 1400. This means that most of the time, the electricity demand is within this range.

There are a few reasons why the load demand might be most frequently between 1100 and 1400. One reason is that this is the range of load demand that is required to meet the basic needs of most households and businesses. For example, this range of load demand is typically sufficient to power lights, appliances, and heating and cooling systems.

Another reason why the load demand might be most frequently between 1100 and 1400 is that this is the range of load demand that is most efficient for the electricity grid to operate. When the load demand is within this range, the electricity grid can operate without having to rely on expensive and inefficient peaking power plants.

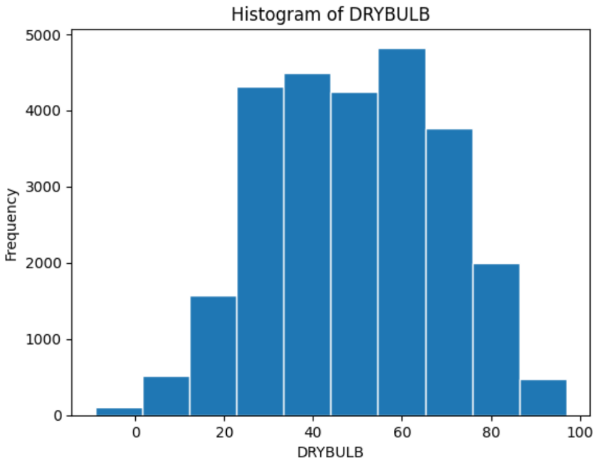
The histogram also shows that there are a few instances of load demand that are below 1100 and above 1400. These instances might represent times when the electricity demand is either unusually low or unusually high. For example, the load demand might be unusually low during the middle of the night when most people are asleep, and businesses are closed. The load demand might be unusually high during a heat wave or a cold snap, when people are using more electricity to cool or heat their homes and businesses.

Overall, the histogram of load demand over time provides a useful snapshot of the electricity demand at a given point in time. The histogram can be used to identify the most frequent load demand, as well as instances of unusually low or unusually high load demand.



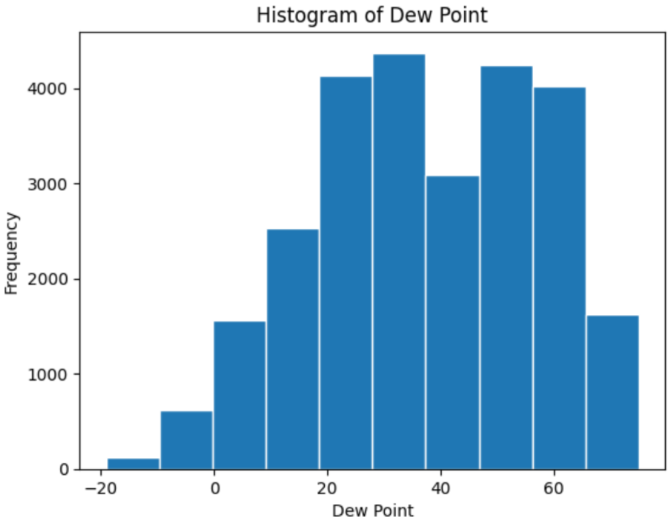
Distribution of Dry Bulb:

The histogram shows that the dry bulb temperature is most frequently between 40 and 70 degrees Fahrenheit. This means that the majority of the time, the air temperature is within this range. And the distribution is the distribution is normal.



Distribution of Dew Point:

The histogram shows that the dew point temperature is most frequently between 35 and 50 degrees Fahrenheit. This means that the majority of the time, the air is humid. And the distribution is the distribution is normal but slightly left-Skewed.



5. MODEL PREDICTION :

To predict right models, we have started to build linear model by adding features one by one. We felt this can be part of a systematic approach to model development and evaluation. The idea is to start with a simple model and iteratively add features to assess their impact on the model's performance. We built a total of 9 regression models and observed the R square values and their residual plots.

We shall see the models and the observations made in this section.

5.1 Model 1:

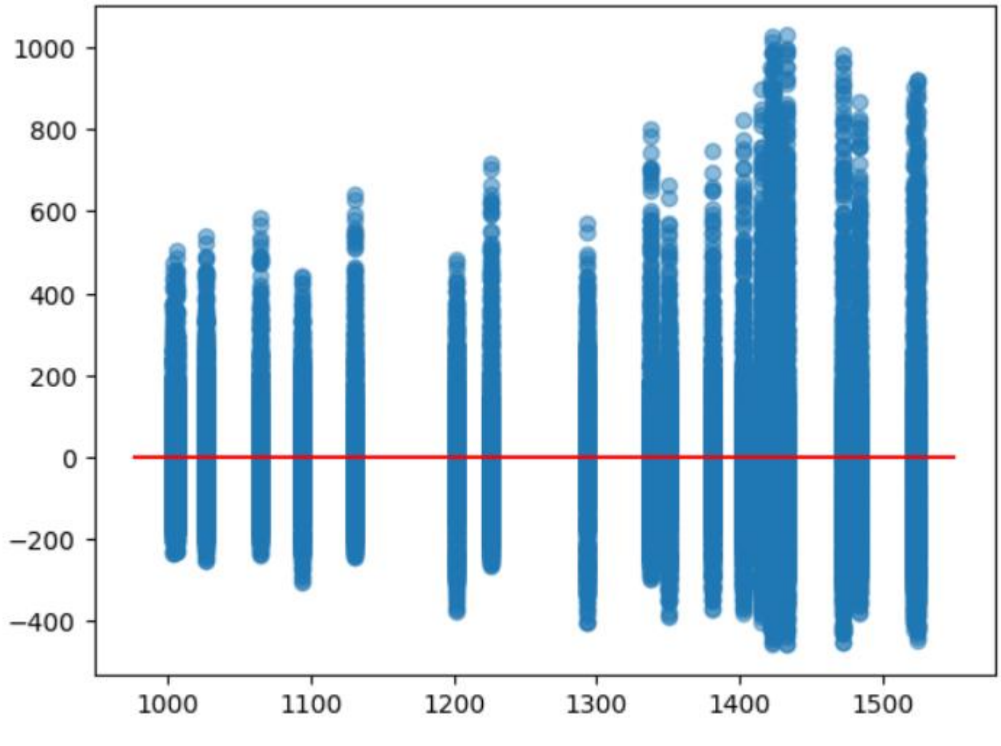
The dependent variable RT_Demand and the categorical variable Hr_End are used in a linear regression analysis. The formula here is as below

'RT_Demand ~ C(Hr_End) '

It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.4461998744343396
2. Adjusted R-squared: 0.44571519395914905

The below plot is between fitted values and residual values. The model is overly linear. In other words, the model assumes that the connection between the independent and dependent variables is exactly linear, although it may be more non-linear in reality.



5.2 Model 2:

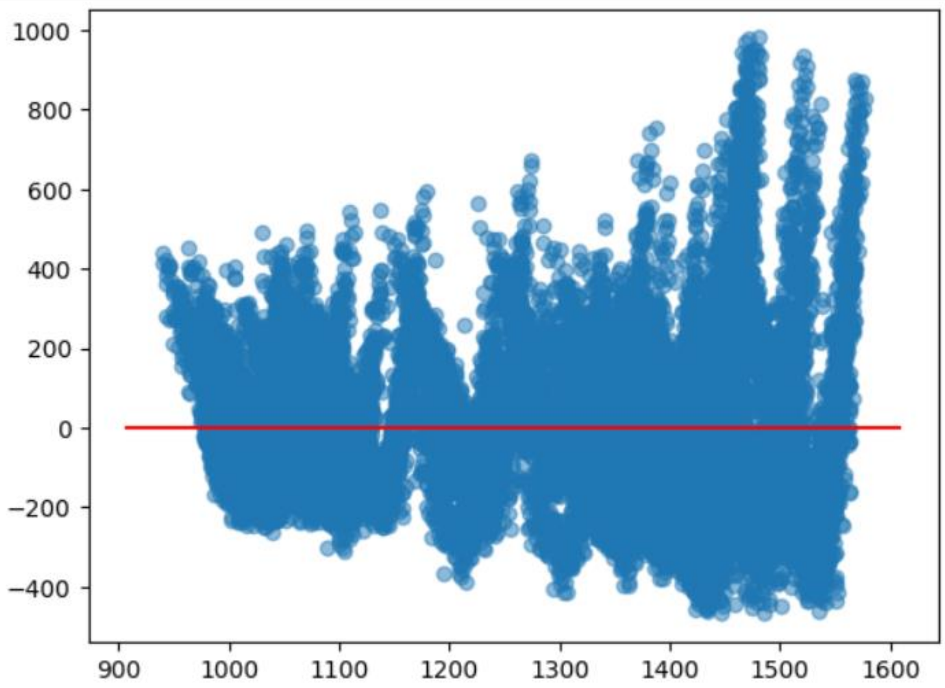
The RT_Demand is the dependent variable, and the two independent variables we considered in this model are the categorical variable Hr_End and the continuous variable Dry_Bulb. It generates a residual plot and publishes a summary of the regression findings for diagnostic purposes. The inclusion of Dry_Bulb indicates that the link between RT_Demand, categorical time periods (Hr_End), and temperature (Dry_Bulb) be investigated. The formula is as below.

'RT_Demand ~ C(Hr_End)+ Dry_Bulb'

It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.455
2. Adjusted R-squared: 0.455

Residual plot clearly indicates a linear relationship between the fitted residual values and the residual values. This shows that at higher fitted values, the model is underestimating the actual values.



5.3 Model 3:

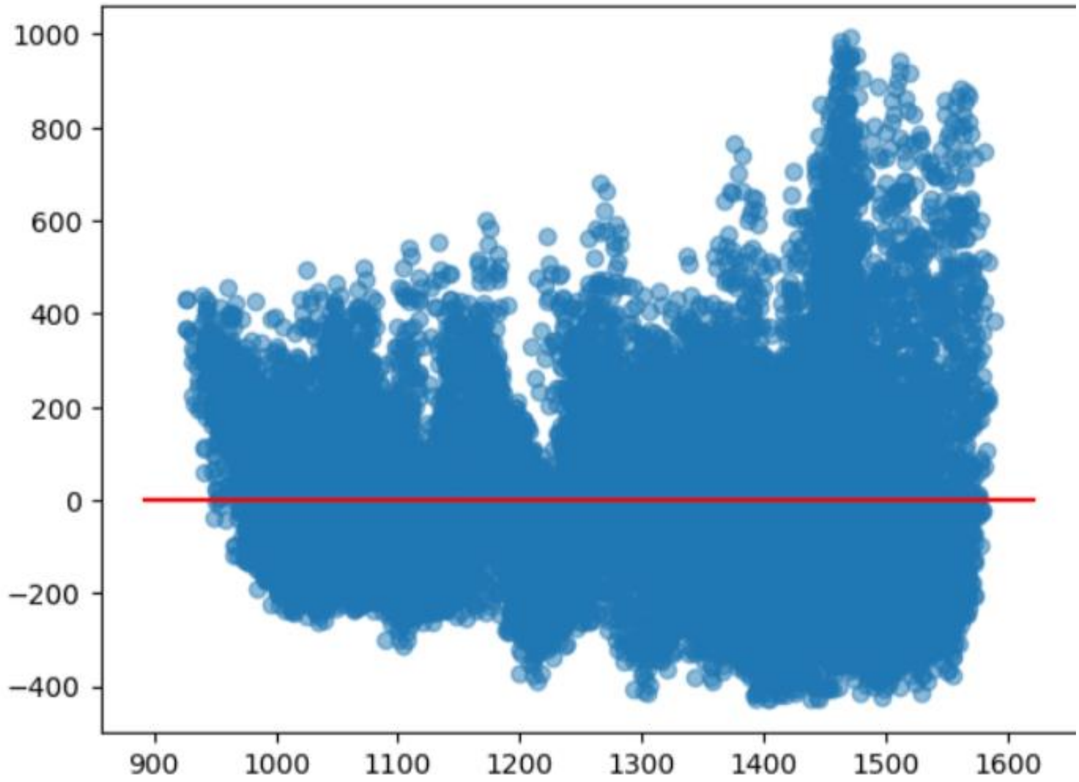
This code runs a linear regression analysis on the dependent variable RT_Demand and three independent factors: Hr_End, a categorical variable, and the continuous variables Dry_Bulb and Dew_Point.

'RT_Demand ~ C(Hr_End)+ Dry_Bulb+Dew_Point'

It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.459
2. Adjusted R-squared: 0.459

The model is overly linear. In other words, the model assumes that the connection between the independent and dependent variables is exactly linear, although it may be more non-linear in reality.



<u>5.4 Model 4:</u>

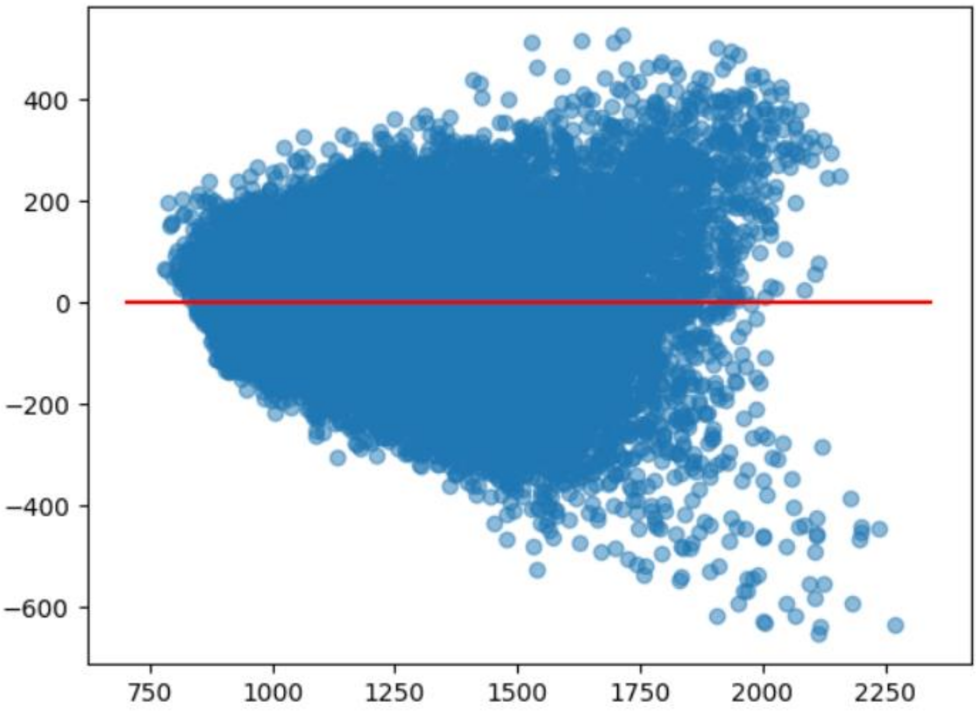
A linear regression analysis on the dependent variable RT_Demand and four independent variables: the categorical variable Hr_End, as well as the continuous variables Dry_Bulb, Dew_Point, and the squared term of Dry_Bulb. The formula is as below.

'RT_Demand ~ C(Hr_End) + Dry_Bulb + Dew_Point + I(Dry_Bulb**2)'

It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.775
2. Adjusted R-squared: 0.775

The residual plot clearly indicates a linear relationship between the fitted residual values and the residual values. This shows that with higher fitted residuals, the model is underestimating the real residuals.



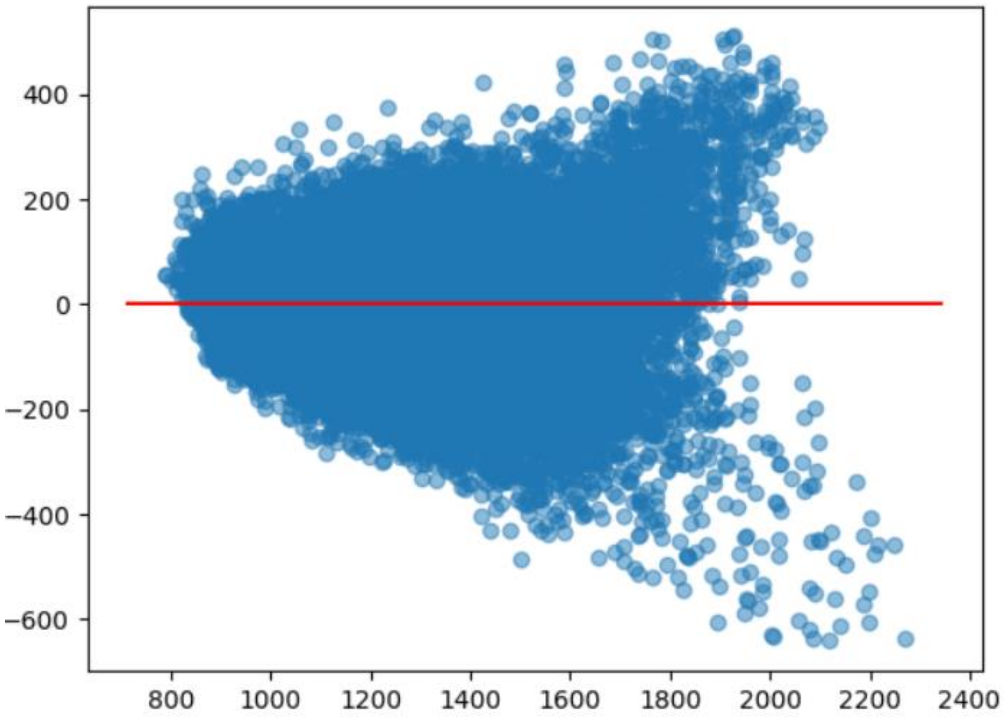
5.5 Model 5:

By introducing quadratic terms in the linear regression model for both Dry_Bulb and Dew_Point. The quadratic terms enable the model to account for hypothetical nonlinear interactions between RT_Demand and temperature variables.The formula is
'RT_Demand ~ C(Hr_End) + Dry_Bulb + Dew_Point + I(Dry_Bulb**2) + I(Dew_Point**2)'

It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.784
2. Adjusted R-squared: 0.783

The residual plot clearly indicates a linear relationship between the fitted residual values and the residual values. This shows that with higher fitted residuals, the model is underestimating the real residuals.



5.6 Model 6:

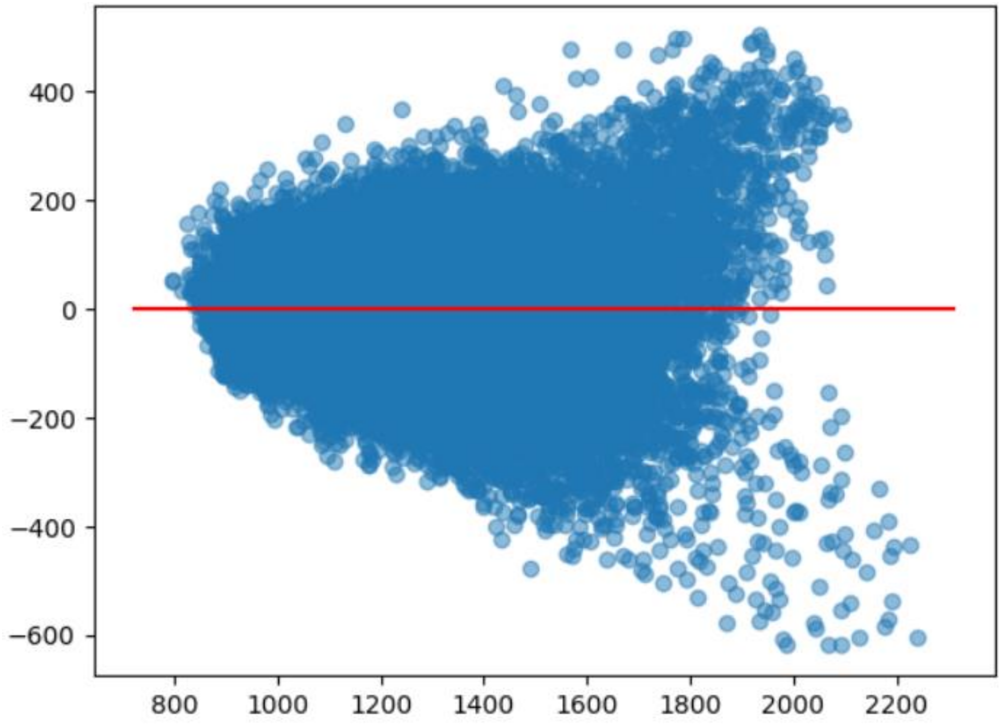
A linear regression analysis with the dependent variable RT_Demand that contains categorical variables (C(Hr_End)), continuous variables (Dry_Bulb and Dew_Point), their squared terms, and an interaction term. The used formula is as below.

'RT_Demand ~ C(Hr_End) + Dry_Bulb + Dew_Point + I(Dry_Bulb**2) + I(Dew_Point**2) + Dry_Bulb:Dew_Point'

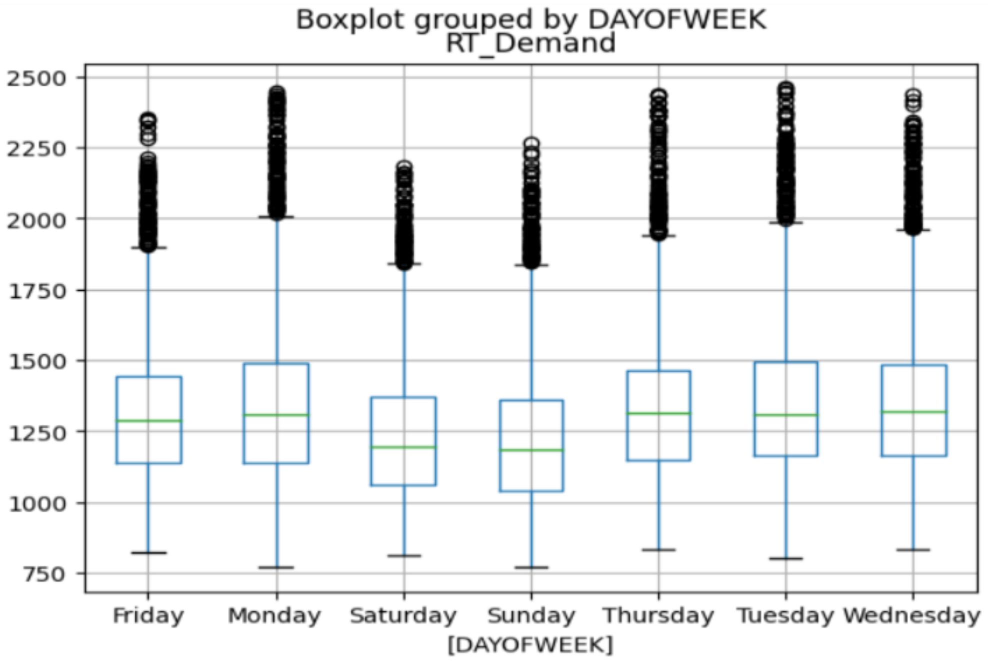
It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.789
2. Adjusted R-squared: 0.789

The residual plot clearly indicates a linear relationship between the fitted residual values and the residual values. This shows that with higher fitted residuals, the model is underestimating the real residuals.



Since we have tried all variations of models we have created a day of week column for accuracy and finding best model using datetime function. For data exploration, we have plotted the boxplot :



We can observe from the plot that Saturday and Sunday's have low Rt_demand since the all the offices and work places are closed on the days. Whereas on the weekdays the Rt_demand is comparitively high .

7. **Model 7:**

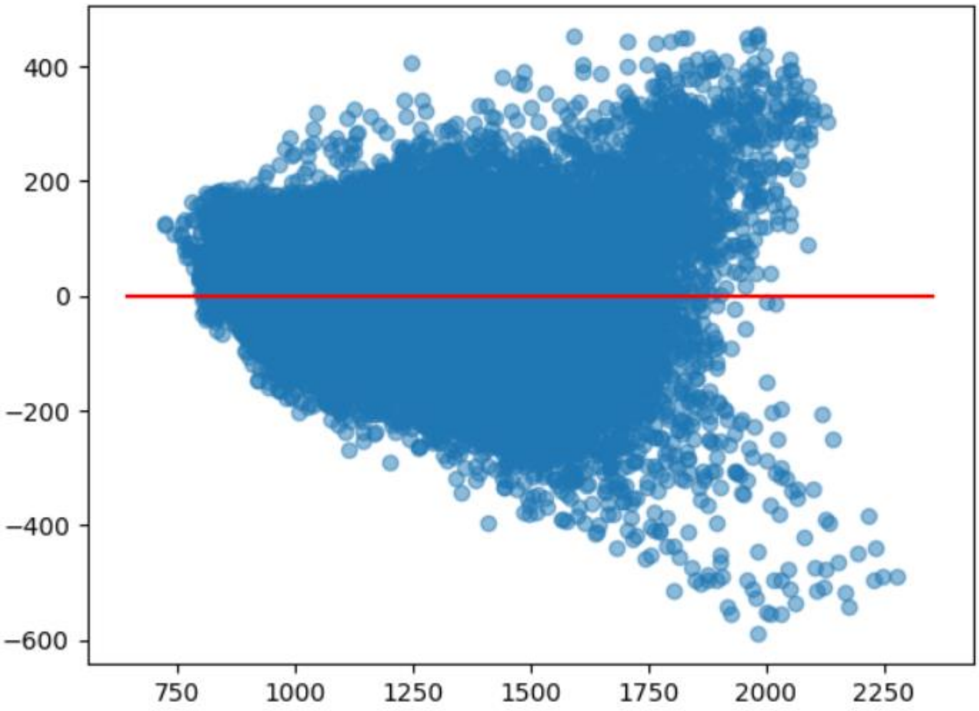
RT_Demand is the dependent variable, Hr_End and DAYOFWEEK are categorical variables (C(Hr_End) and C(DAYOFWEEK), and there are continuous variables (Dry_Bulb and Dew_Point), their squared terms, and an interaction term (Dry_Bulb:Dew_Point). The formula is as below.

'RT_Demand ~ C(Hr_End) + Dry_Bulb + Dew_Point + I(Dry_Bulb**2) + I(Dew_Point**2) + Dry_Bulb:Dew_Point + C(DAYOFWEEK)'

It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.827
2. Adjusted R-squared: 0.827

The residual plot demonstrates that the model does not match the data well. The evident pattern in the residual plot indicates that the model has a problem, misspecification. This issue must be addressed in order to increase the model's accuracy.



5.8 Model 8:

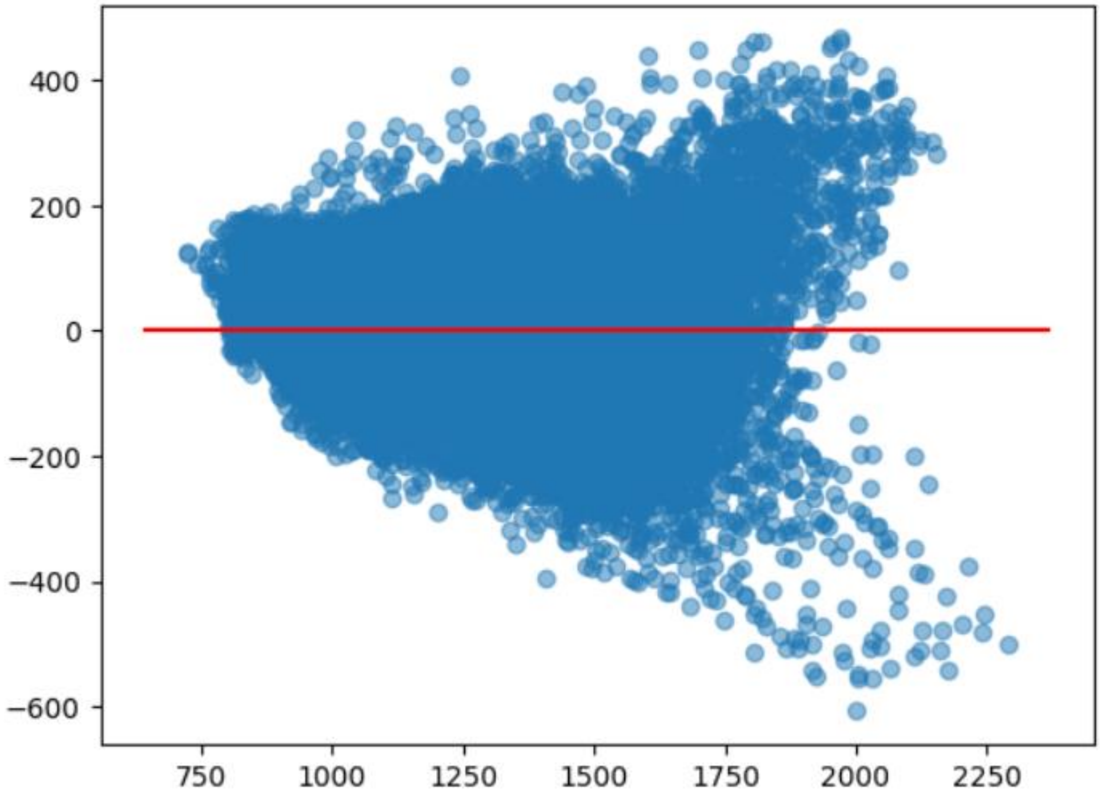
A linear regression analysis with the dependent variable RT_Demand and multiple independent variables such as categorical variables (C(Hr_End), C(DAYOFWEEK), continuous variables (Dry_Bulb and Dew_Point), squared terms, and interaction terms (Dry_Bulb:Dew_Point, Dry_Bulb:C(DAYOFWEEK)). The formula is as below

'RT_Demand ~ C(Hr_End) + Dry_Bulb + Dew_Point + I(Dry_Bulb**2) + I(Dew_Point**2) + Dry_Bulb:Dew_Point + C(DAYOFWEEK) + Dry_Bulb:C(DAYOFWEEK)'

It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.827
2. Adjusted R-squared: 0.827

Our residual plot clearly indicates a linear relationship between the fitted residual values and the residual values.

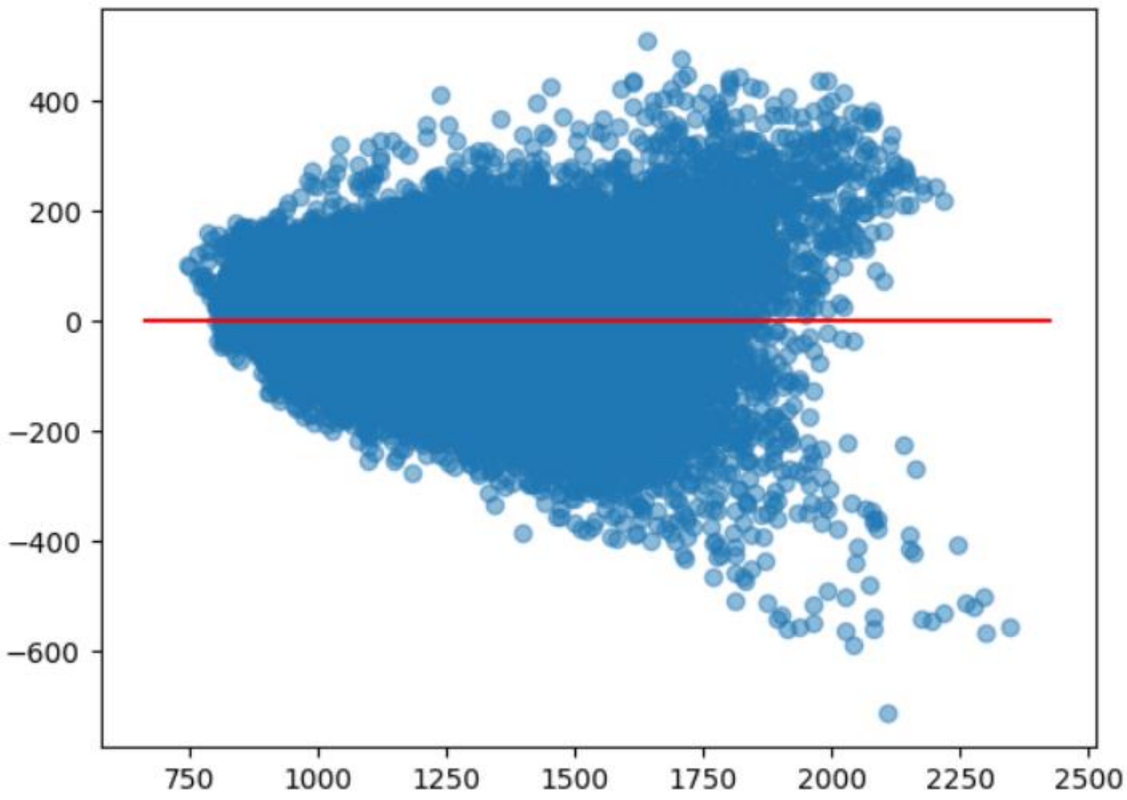


5.9 Model 9:

A linear regression analysis with the dependent variable RT_Demand and several independent factors, including categorical variables (C(Hr_End), C(DAYOFWEEK), continuous variables (Dry_Bulb and Dew_Point), squared terms, and different interaction terms. The interaction terms with DAYOFWEEK imply an investigation of the relationship between RT_Demand, temperature variables, and their interactions with various days of the week.
Used formula is as below
'RT_Demand ~ C(Hr_End) + Dry_Bulb + Dew_Point + I(Dry_Bulb**2) + I(Dew_Point**2) + Dry_Bulb:Dew_Point + C(DAYOFWEEK) + Dry_Bulb:DAYOFWEEK + I(Dry_Bulb**2):DAYOFWEEK'
It outputs a summary of the regression findings with R-squared and modified R-squared values as follows:

1. R-squared: 0.83
2. Adjusted R-squared: 0.83

The plot is non linear which shows a good fit to the model and are also fairly scattered and we do not see any clear patterns from the residual plot.

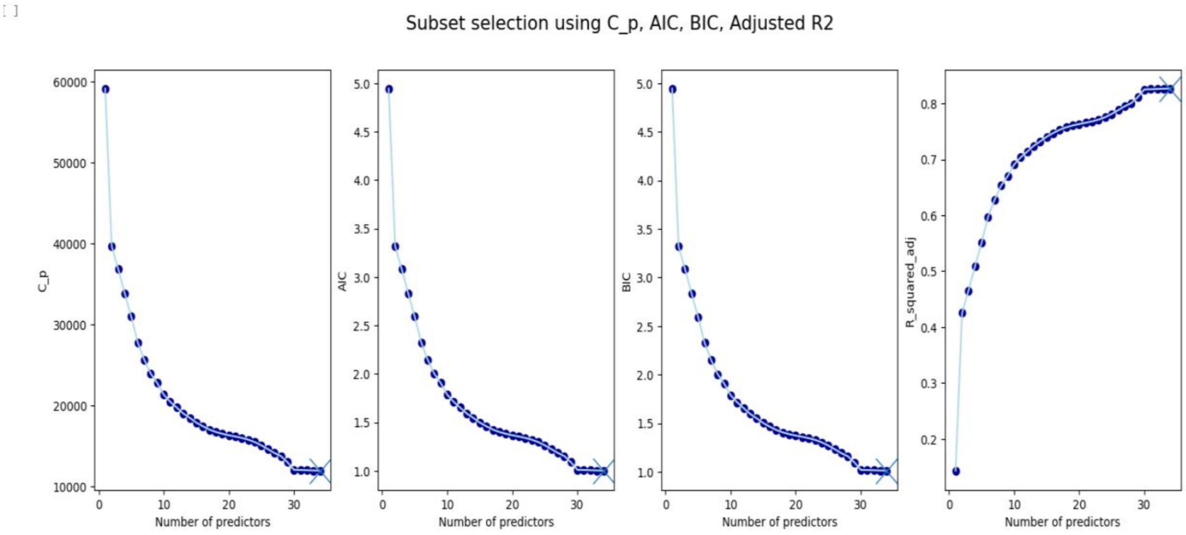


This approach gave us a baseline understanding of the relationship between the dependent variable and each individual predictor. By adding features incrementally, you can observe how each feature contributes to the model's explanatory power. Adding features one by one allows you to interpret the impact of each feature on the model. Incremental addition of features helps you monitor the risk of overfitting.

6. Feature Engineering:

6.1 Model Selection

To avoid the overfitting problem, we have implemented Step wise model Selection (Forward Selection) to pick the best features out of our full model mentioned above. There are 35 covariates in our full model omitting Date and RT_Demand features. Since the covariates are greater than 30, Exhaustive Search cannot be used due to huge computing workload.



We have plotted a subplot of line graphs and scatter plots to display the performance of various information criteria (C_p, AIC, BIC, and Adjusted R-squared) in the context of subset selection for better understanding.

A graphical illustration of how various information requirements behave as the number of predictors in the model grows. The plots aid in selecting optimal subsets of predictors for each criterion. On each plot, the 'x' marker denotes the subset with the highest criterion value. This visualization assists in making educated model selection selections based on many criteria.

This shows that,
'Dry_Bulb_squared','Dry_Bulb','Hr_End_4','Hr_End_3','Hr_End_5','Hr_End_2','Hr_End_6','Hr_End_24','Hr_End_7','Hr_End_23','Hr_End_19','Hr_End_18''Hr_End_20','Hr_End_17','Hr_End_21','Hr_End_16','Hr_End_14','Hr_End_15','Hr_End_13','Hr_End_12','Hr_End_11','Hr_End_10','Hr_End_9','Hr_End_22','Hr_End_8''DAYOFWEEK_Sunday','DAYOFWEEK_Saturday','DAYOFWEEK_Wednesday','DAYOFWEEK_Tuesday','DAYOFWEEK_Thursday','DAYOFWEEK_Monday''Interactive_Term','Dew_Point','Dew_Point_squared' - with all these 34 features are considered as best model for the data.

7. Model Training and Testing:

Two machine learning models were employed for load forecasting:

Linear Regression: Linear regression involves training a model on a training set, evaluating its performance on both the training and testing sets, interpreting coefficients, and making necessary adjustments to achieve a well-performing and generalizable model. It is a foundational technique widely used in regression analysis and serves as a basis for more complex models in machine learning.
These are the values

Random Forest: Random Forest Classifier:The code creates and trains a Random Forest Regressor model with 100 trees based on the characteristics (X_train) and target variable (y_train) supplied. Random Forest models are a collection of decision trees, with each tree trained on a distinct sample of the data and making its own predictions. The final forecast is frequently an average of all the trees or a vote process. The random_state parameter is set to allow others to duplicate the exact same model if they use the same random seed.

The rf_model instance holds information on the ensemble of decision trees after training, and this trained model can be used to make predictions on new data (X_test), evaluate performance, or examine feature importance.

After training the Random Forest Regressor model, the predict technique is used to generate predictions using new, previously unseen data (in this case, the test set).The projected values for the target variable are stored in rf_predictions, which is a NumPy array or array-like object based on the characteristics in X_test.

The dataset was split into a training set (January 2020 to November 2022) and a testing set (December 2022). Models were trained on the training set and evaluated on the testing set.

The goal of time-series forecasting is often to build a model that generalizes well to future, unseen data. Training the model on past data and testing it on future data provides a more realistic evaluation of its ability to generalize to new, unseen observations.

7.1 Linear Regression:

For Linear Regression models:

The MSE and RMSE values are calculated:
Mean Squared Error(MSE) = 11663.605360351292

Root Mean Squared Error = 107.99817294913508

The r2 score value is calculated:
r2 score= 0.654714560140025

7.2 Random Forest Model:

To evaluate the model's performance, these predictions can be further assessed and compared to the actual target values (y_test). Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared are all common measurements.
The MSE and RMSE values are calculated:
Mean Squared Error(MSE) = 6426.511885828212
Root Mean Squared Error = 80.1655280393525

8. Conclusion

Based on this MSE and RMSE values Random Forest Classifier is the best model.This project successfully developed and evaluated machine learning models for hourly load forecasting in New Hampshire. The results can guide decision-making processes in the energy industry, helping stakeholders make informed choices based on accurate load predictions.

9. Future Scope

We created models with different predictors and interactive variables in this model. We can extend this and improve our models by introducing the below interactive variables.

- Between Hr_end and Day of week to improve the model.
- Between Dry bulb and DAYOFWEEK to improve the model.
- Between Dry_bulb_square and Dayofweek to improve the model.

We can also perform second level of the cross validation by training data sets till October 2022 and test it with November 2022.

We can also try to consider public holidays and normal days and introduce them into our model.