HLOOV Assignment Project

Name: Naveen A

Email ID: nasajeager2001@gmail.com

GitHub profile: https://github.com/navi1910

Introduction

The built environment contributes 40% of global greenhouse gas (GHG) emissions. If left unchecked, they are set to double by 2050. Here, energy efficiency plays a vital role in lowering energy demands, reducing carbon emissions, and driving the clean energy transition. In built environment managing the heating, ventilation, and air conditioning systems (HVAC) through effective operation and maintenance is a best practice for conserving energy.

HVAC includes the different equipment like Compressor, Condenser, Evaporator, Cooling tower, Chiller, Air Handling Unit (AHU), Roof Top Unit (RTU), Pumps, Heater, etc. These equipments will consume more energy if not maintained and control efficiently.

Objective

To predict "HVAC system: electricity" based on other features/columns of data given from “RTU.csv”. Drop out the columns and rows having "NA". Also forecast "RTU: electricity" for the next 7 days from the latest time. Mention the insights and findings from the exploratory data analysis (EDA).

Data Source:

https://figshare.com/articles/dataset/LBNLDataSynthesisInventory_pdf/11752740/3

Data obtained from Lawrence Berkeley National Laboratory.

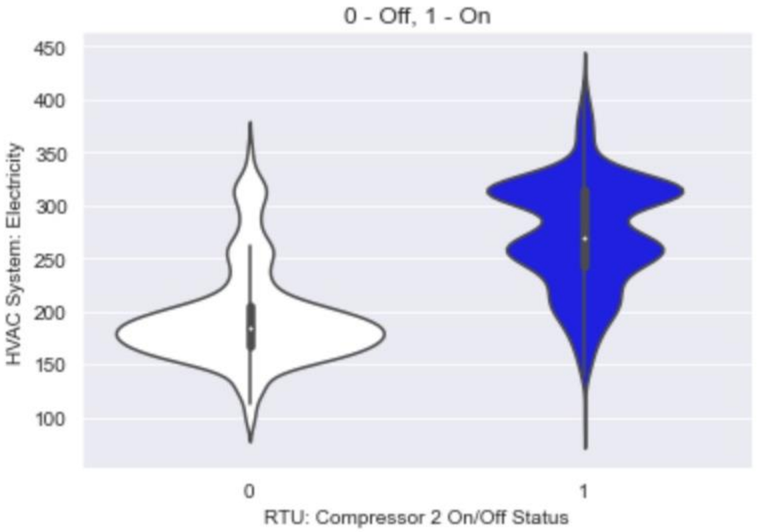
Note: Python and Jupyter-notebook was used.

Assumptions

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

Data Preprocessing

- Timestamp is parsed.
- Null values are removed.
- Data columns with low variation are removed.
- Highly correlated columns are removed to avoid multicollinearity.
- Extreme outliers are capped.
- Pre-processed data is saved as ‘csv’ file
- Data Visualization
- Visualization was done using ‘matplotlib’ and ‘seaborn’ python packages.
- Heatmap of the entire data is plotted.
- Boxplot of “HVAC System: Electricity” is plotted.
- Countplots for discrete variables are plotted. Discrete variables with low variability are removed.



- Paired T-test is conducted to check if the intervention is causing significant difference among the samples. Corresponding Violin-plots are plotted.
- Jointplots are created to check for correlation among the Variables.
- Histograms are created to check the distribution of the Variables.
- Average Room Temperature and Average Room Humidity is calculated from all the Rooms.
- The processed data in this step is saved as ‘csv’ file.

Findings from Visualizations

- “HVAC Systems: Electricity” has multiple outliers.
- Compressor 2 causes a significant difference in the Electricity consumption.
- There is some amount of correlation between “Supply Air Temperature” and “HVAC Systems: Electricity”. Correlation coefficient – r: -0.568313
- There is slight correlation between Average Room Temperature and Average Room Humidity.
- Correlation coefficient – r: -0.422744.
- There is significant difference in Average Room Temperature because of Compressor 2.
- There is significant difference in Average Room Humidity because of Compressor 2.
- Most of the Variables are normally distributed.
- Compressor 2 causes significant increase in Electricity consumption (observed from Violin-plot and Paired T-test) and also has significant effect on average Room temperature and average room humidity.

Prediction and Model Building

- The problem is identified to be supervised regression type, because we have continuous Target Variable.
- Scikit-learn was used.
- The data was scaled using StandardScaler.
- The dependent and independent variables were separated.
- The Training set and testing set were separated.
- Regression Models were created.
- **Linear Regression - Mean Absolute Error: 0.07096158314277902**
- **Lasso Regression – Mean Absolute Error: 0.09480500921591219**
- **RandomForestRegressor - Mean Absolute Error: 0.06713512777671357**
- **RandomForestRegressor with GridSearchCV used to select parameters – Mean Absolute Error: 0.06146472499614632**

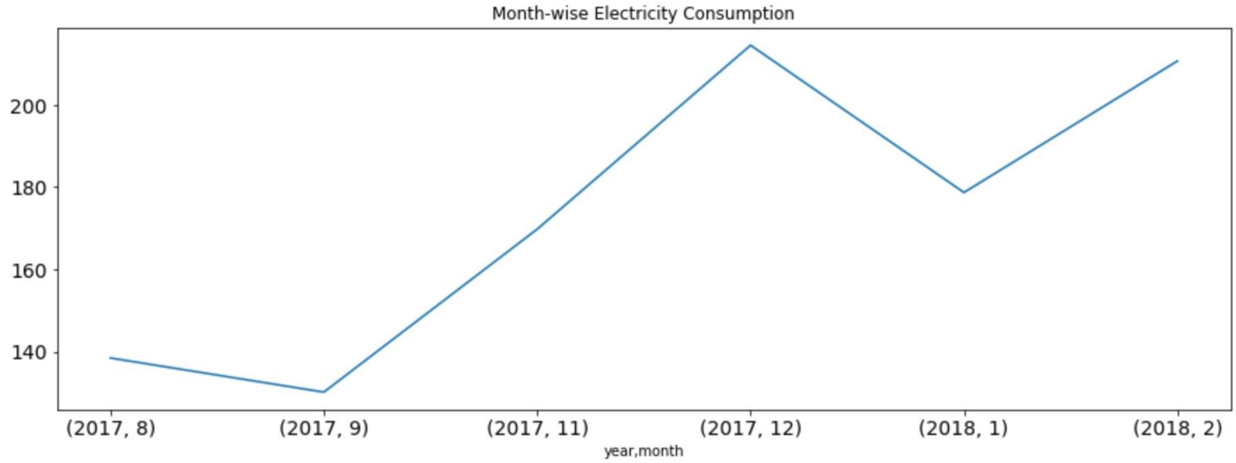
Conclusion

The RandomForestRegressor with GridSearchCV outperformed all other models.

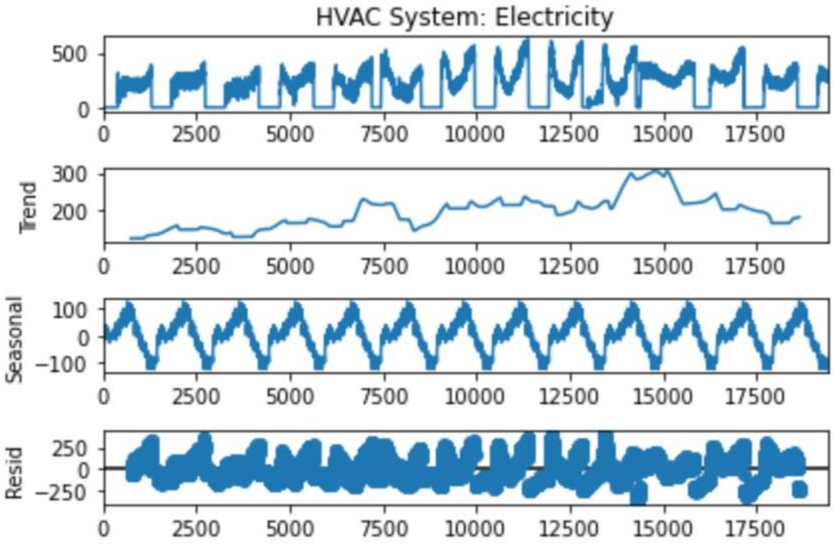
RandomForestRegressor Model 'best' - MAE: 0.06146472499614632

Forecasting

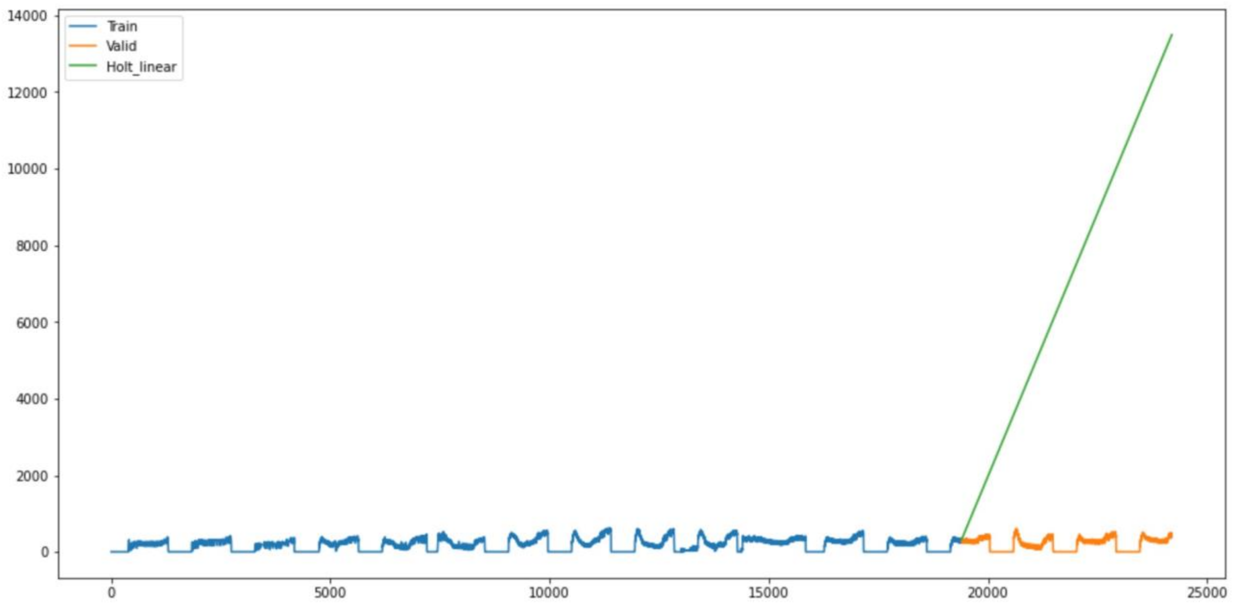
- Feature Extraction from “Timestamp”. ‘Year’, ‘month’, ‘day’, and ‘hour’ was extracted from “Timestamp”.
- “HVAC Systems: Electricity” was plotted.



- Data exploration was done on the newly extracted features and plots were created.
- “HVAC Systems: Electricity” was divided into Train set, validation set and Test set.
- Decomposition was also done for “HVAC Systems: Electricity”.



- We can see seasonality and trend in “HVAC Systems: Electricity”
- Holts-winters forecasting and Exponential smoothing were used for forecasting.
- The forecasting results were not satisfactory because of the high variations and inconsistency in the data.


Train
Valid
Holt_linear
14000
12000
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000