EAEE 4000
FINAL PROJECT
FALL 2022
12/23/2022

Daily Electricity Demand Prediction in Columbia Campus

Jianfeng Wang
jw4166@columba.edu

Introduction

Background

Nowadays, electricity is one of the essential parts of human society. New York City is the most populous city in the United States and is also described as the world's cultural, financial, and media capital. Due to environmental degradation and global warming, it is necessary to develop low-carbon technologies to reduce energy consumption and carbon emissions. [1] Most of the electricity in New York City is produced by fossil fuel combustion. The daily electricity consumption in New York City is 11,000 Megawatt-hours on average. There are nearly 50,000 buildings in Manhattan that consume a considerable amount of energy. Many factors influence the energy consumption of a commercial building, such as historical consumption of electricity, weather conditions, including temperature, humidity, precipitation, etc., working date, including season, weekends, holidays, etc., and economics and political factors. [2, 3] Electricity demand prediction is essential for reducing energy consumption and energy cost at Columbia University. With the development of machine learning, future electricity consumption prediction becomes possible. Columbia University, as a commercial building in New York City, pays its electricity bill in two parts. The first part will pay its energy consumption at a fixed rate using a demand-supply contract. Another aspect is going to pay the peak demand charge monthly. The peak demand rate is higher than the base demand rate. If it is possible to predict the daily market on campus, Columbia University is much easier to know about its energy consumption and will apply some solutions to reduce energy consumption. Take the example of installing batteries. Batteries could store energy and discharge whenever needed, but with machine learning, it would be easier to predict the peak demand period accurately and utilize batteries efficiently to reduce energy consumption and cost.

Objectives

This Project aimed to predict the daily demand in 2019 at the Columbia campus using machine learning based on historical electricity demand data, weather conditions, and calendar data. Created three prediction models and compared each other to find an optimal model. Three models are used in this Project: Linear Regression, Random Forest Regressor, and Multilayer Perceptron Regressor. In addition, feature importance is going to be figured out.

Data Collection

Input data (Predictors):

Daily electricity demand in 2018 and 2019: historical electricity demand is critical for electricity prediction using machine learning. Figure 1 shows the visualization of energy demand at Columbia University from 2018 to 2019, which was normalized. Data normalization reduced data redundancy and improved data integrity so that extremely high or low digits of electricity demand would not show on figures.

Daily weather data in 2018 and 2019 [5]:

Including Maximum Temperature, Minimum Temperature, Average Temperature, Snow Depth, Precipitation, and Humidity. This Project analyzed how important the temperature influences the electricity demand at Columbia University. However, there are some constant energy consumptions for a commercial building, such as emergency lighting, safety electricity, a massive proportion of energy consumed by heating/cooling systems, ventilation, and air-conditioning systems (HVAC) affected by weather conditions. Weather data were also normalized, and all the demand data in 2018 was used as train data to create prediction models.

Working date in 2018 and 2019 [4]:
Working data was considered as a factor influencing daily demand on campus. At weekends or holidays, the number of alumina in Columbia decreases, and the building would be empty, so some lighting energy consumption would also decrease. The calendar also shows when class begins and finishes seasonally. During spring, summer, and fall semesters, students at Columbia University are more than at other times. Furthermore, there would be more classrooms occupied for each building. This Project aimed to predict the daily demand in 2019 at the Columbia campus using machine learning based on historical electricity demand data, weather conditions, and calendar data. Created three prediction models and compared each other to find an optimal model. Three models are used in this Project: Linear Regression, Random Forest Regressor, and Multilayer Perceptron Regressor. In addition, feature importance is going to figure out.

Feature index: ‘Demand’, ‘Temp_max’, ‘Temp_min’, ‘Temp_avg’, ‘Temp_departure’, ‘HDD’, ‘CDD’, ‘Precipitation’, ‘new_snow’, ‘snow_depth’, ‘class’, ‘dayoff’, ‘Monday’, ‘Tuesday’, ‘Wednesday’, ‘Thursday’, ‘Friday’, ‘Saturday’, ‘Sunday’.

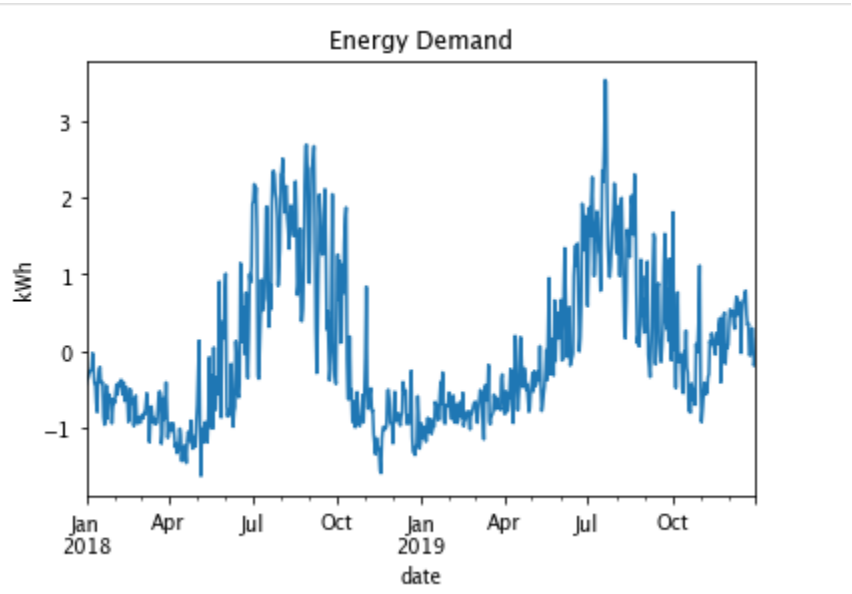

Figure 1: Energy demand in Columbia University in 2018-2019

Output data (Predictants):
Daily electricity demand in 2019:

The first part of demand data in 2019 is defined as train data and the rest of demand data was defined as test data. The daily electricity demand in 2019 as the test data can be predicted by models created.

Split data

Data splitting is widely used in data science, usually applied for creating models based on data. It is helpful for machine learning creating models based on train data and test model using rest of data. Data splitting could efficiently avoid overfitting. In this project, daily demand is divided into 80% of train data and 20% of test data. Figure 2 visualized splitting data. The blue part is train data and orange part are test data.

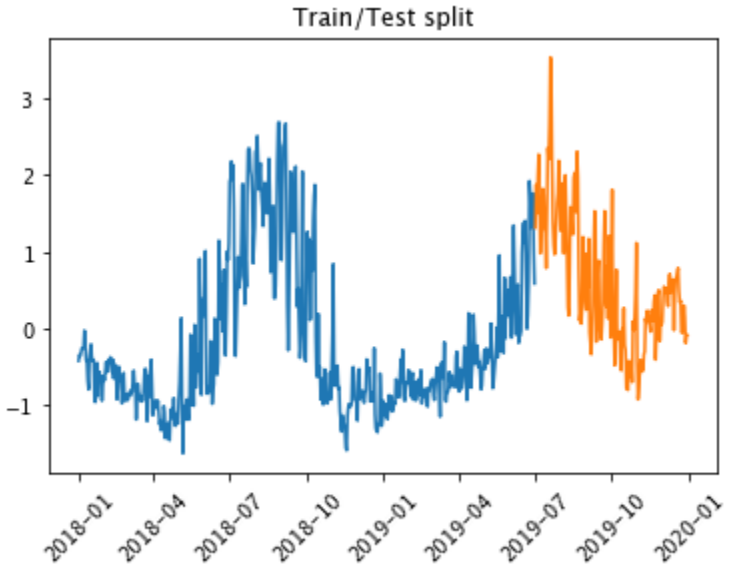

Figure 2: Train data and test data split

Exploratory Data analysis

After data splitting, exploratory data analysis (EDA) is necessary for daily demand at Columbia University in 2018 and 2019. Involving statistics and visualizations, EDA can identify demand data trends and help understand the datasets deeply. Moreover, it is easier to know the relationship between electricity demand data and various features. Three kinds of EDA are applied for forecasting energy demand: Target analysis, Volatility analysis, and Time series analysis.

For target analysis, Skewness and Kurtosis of two years' daily demand are calculated. Skewness is a result predicted how asymmetry the dataset is. Skewness is often used in probability theory and statistics. The skewness for daily energy demand is 0.9, which is a positive skew. According to Figure 3, the count of the number of energy demands was shown since the daily electricity demand can be defined as random variables. Besides, the mean value of the whole dataset was

presented as a red line, and a reasonable range of normal values are shown as two orange lines in figure 3. Smaller energy demand occurred more than larger energy demand. Therefore, the skewness of the dataset is positive. In other words, the mass of the distribution is concentrated on the left of the figure. The kurtosis of the normalized daily energy demand is 2.94. This means the dataset distributed is approximately Laplace distribution because the result is near 3. It can also be explained as the peak of the figure is relatively high, and the count of electricity demand data around -1 to -0.5 (normalized electricity demand) occurs frequently.

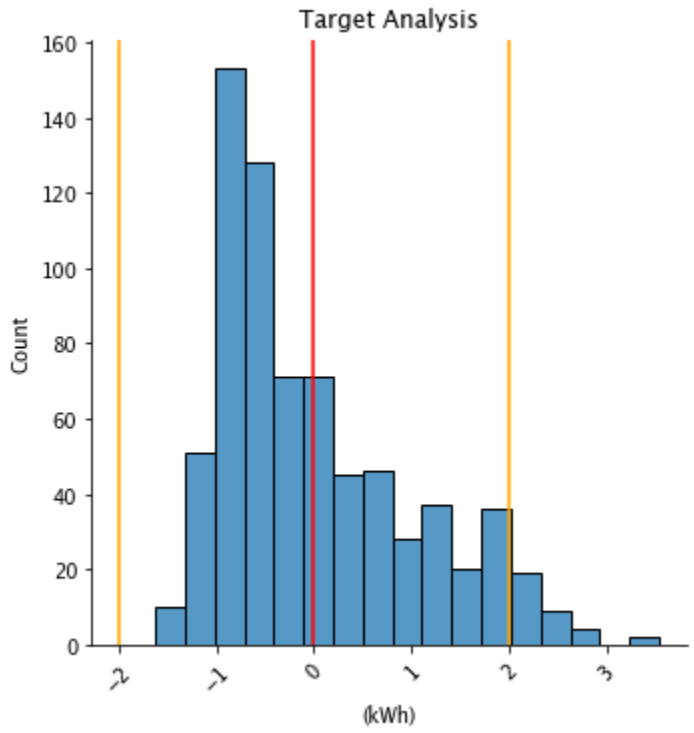

Figure 3: Target analysis

Volatility is a statistical measure of the dispersion of data around its mean over a certain period. It is very suitable for analyzing daily demand because it describes a tendency to fast and unpredictable changes. From figure 4, by using Volatility analysis method for daily energy demand, the trend of how daily demand changing would be better for visualization. Monthly-rolling percentiles are used as the method for volatility analysis. Blue line is 10 percentiles, orange line is 50 percentiles and green line presents 90 percentiles.

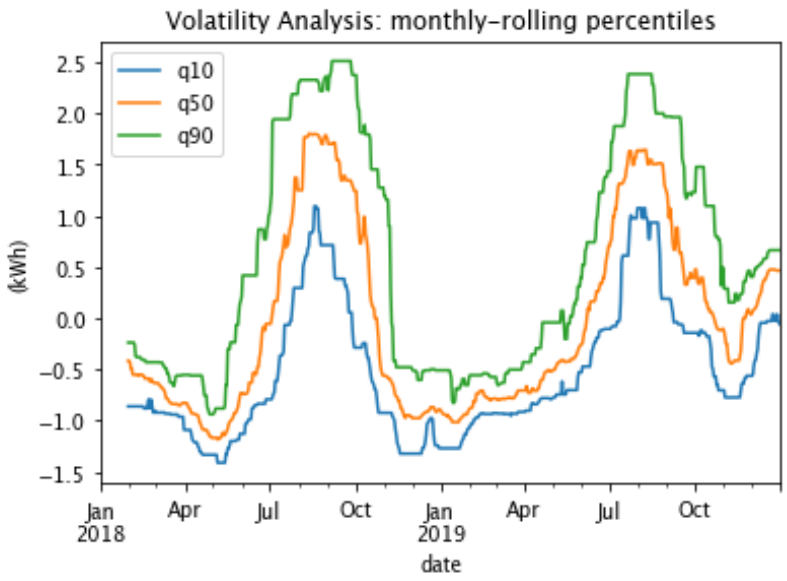

Figure 4: Volatility analysis for the energy demand in Columbia University in 2018-2019

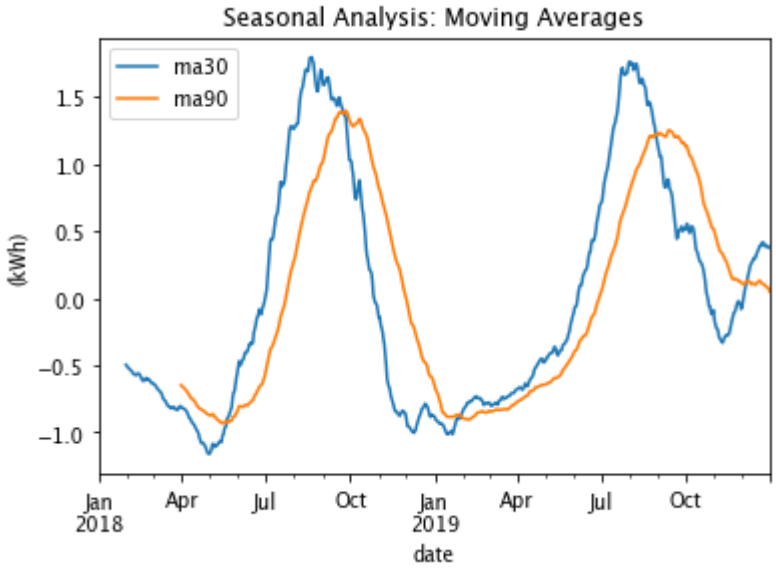

Figure 5: Seasonal analysis for the energy demand in Columbia University in 2018-2019

A time series is a series of data points indexed in time order. In this project, the electricity demand at Columbia University, weather conditions in New York City, and working date followed time series day by day. Time series analysis is needed for demand prediction due to all the datasets based on time series. Seasonality can be defined as the seasonal characteristics of the time series data, and it is predictable. [6] Figure 5 describes the seasonality of energy demand. The peak demand always occurs around August and September. The electricity demand on campus kept increasing from May to August 2018 and 2019. Then, the energy demand had a decreased tendency starting from August to November. The blue line was going to figure out the

monthly repeatability of the daily energy demand, and the orange line was going to find its quarterly repeatability.

Model 1: Linear regression

Linear regression is applied to investigate the relationship between one dependent variable and one or more independent variables. In this project, linear regression could help to estimate the dependence of daily electricity demand on weather data and calendar data. The package of "LinearRegression" is imported from "sklearn.linear_model". By using this model, the test data, named as the future energy might be predicted with some degree of accuracy.

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. Equation 1 is the formula for calculating RMSE in three models,

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad \text{Equation 1}$$

Where $\hat{y}_i$ are predicted values, y_i are observed values and n is the number of observations.
R- squared (R^2) measures how close the data point is to the fitted line. Equation 2 is used to calculate R^2 for three models:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \qquad \text{Equation 2}$$

Where $\hat{y}_i$ are predicted values, y_i are observed values and $\bar{y}$ is the mean value.
As a result, train RMSE for this model is 0.445, test RMSE is 0.663, and R^2 is 0.393.
From Figure 6, the result was visualized as Forecast daily energy demand versus Actual daily energy demand. The orange points were test data and blue points were train data. The test data was not good fit on the regression line since its RMSE is approximately 1.5 times larger than the train RMSE. With larger RMSE, the test data had a larger error than the train data. Observed from Figure 5 and 6, the test data were relatively above the regression line on account of the daily demand in 2019 from October to December is averagely higher than the daily demand in 2018 from October to December.

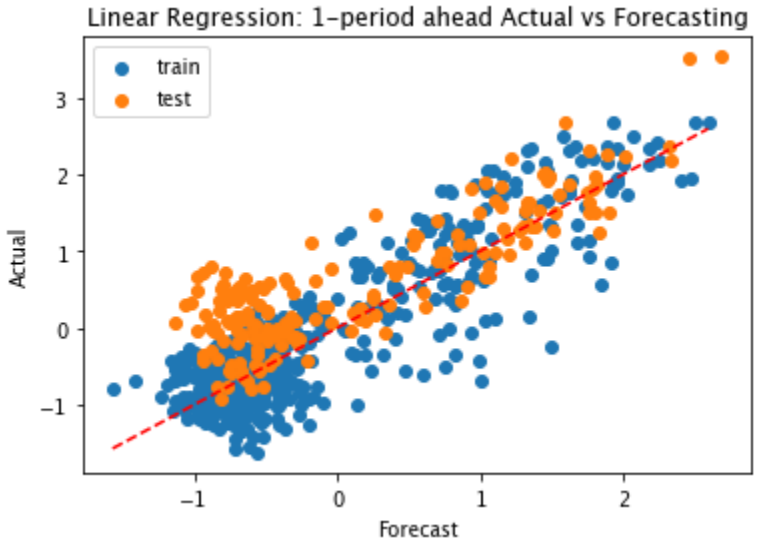

Figure 6: Linear Regression: Forecast data vs. Actual data

Model 2: Random Forest Regressor

Random forest is an ensemble of decision tree algorithms, and it is widely used for regression predictive modeling problems. Due to the daily electricity demand is a time series dataset, random forest was adopted to forecast energy demand. The package of "RandomForestRegressor" is imported from "sklearn.ensemble". Train RMSE for this model is 0.277, test RMSE is 0.717, and R^2 is 0.289. Figure 7 shown that the test data did not concentrate on the regression line as the train data were.

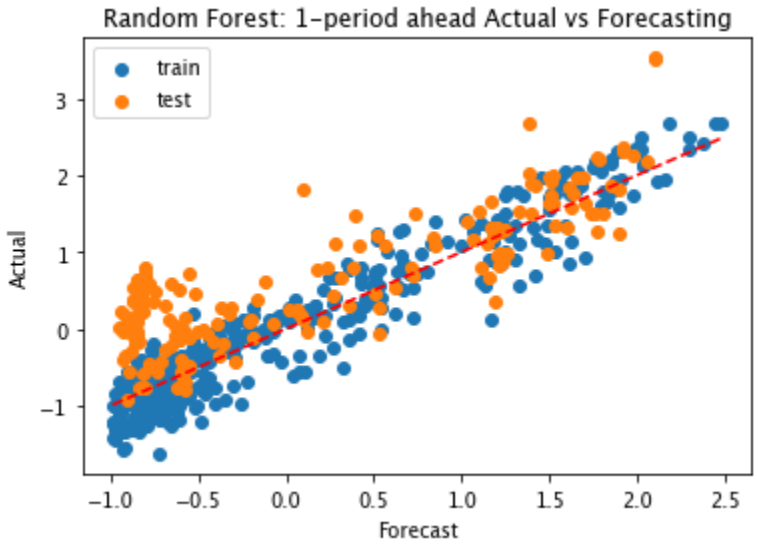

Figure 7: Random Forest Regressor: Forecast data vs. Actual data

Model 3: Multilayer perceptron regressor

Multilayer perceptron regressor is used to predict daily electricity demand as well, including input layers, hidden layers and an output layer. Unless the input nodes, other nodes are neuron that uses a nonlinear activation function. The package of "MLPRegressor" is imported from "sklearn.neural_network". As a result, train RMSE for this model is 0.292, test RMSE is 0.762, and R^2 is 0.198.

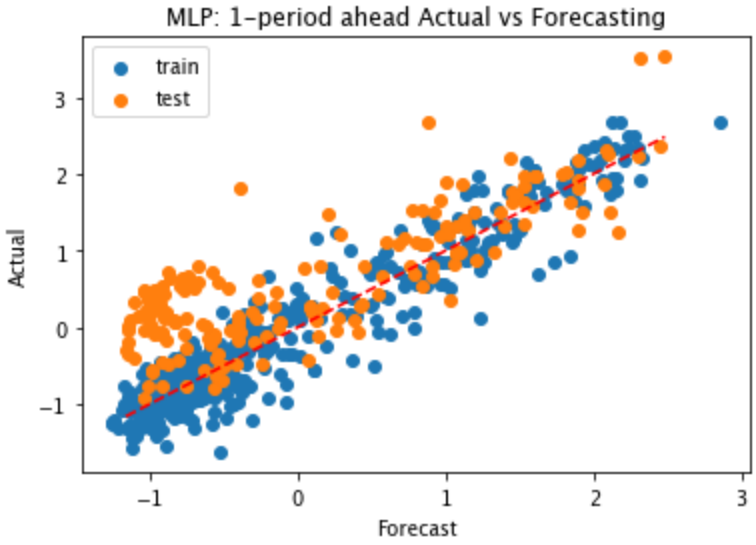

Figure 8: Multilayer Perceptron regressor: Forecast data vs. Actual data

Feature importance analysis

Feature importance analysis helped to understand the relationship between weather conditions, working dates, and daily energy demand. There are 19 features studied in this project. From figure 9, an essential feature for forecasting daily energy demand is the minimum temperature, and the second important feature is the average temperature. The academic calendar was not crucial for predicting the energy demand at Columbia University, which is shown to be of relatively low importance in the figure. Besides, the snow depth could have been more critical when forecasting.

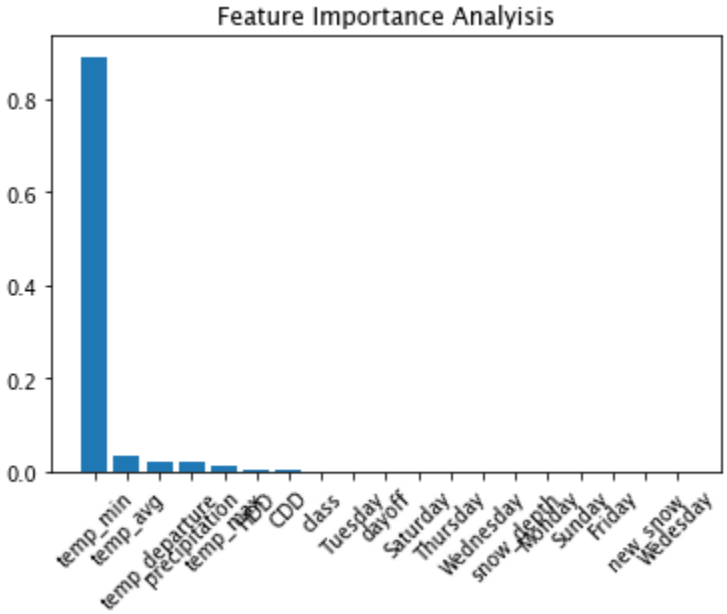

Figure 9: Feature importance analysis for daily demand forecasting

Result and Discussion

Model	Train RMSE	Test RMSE	R^2
Linear Regression	0.445	0.663	0.393
Random Forest Regressor	0.277	0.717	0.289
Multilayer Perceptron Regressor	0.292	0.762	0.198

Table 1: Train RMSE, Test RMSE, R^2 comparison within three models

Table 1 lists the Train RMSE, Test RMSE, R2 results in three built models. Random Forest Regressor had the most miniature train RMSE, Linear Regression had the lowest Test RMSE, and Multilayer Perceptron had the smallest R2. It is hard to find an optimal model since these three models' results have a high error and test data needed to fit the regression line. Combined with the result from importance analysis and seasonality analysis, RMSE and R2 are significant because the features had less importance for predicting daily peak demand. In other words, the daily demand was separate from some of the daily weather data and working date.

Conclusion

This project tried to predict the daily electricity demand at Columbia University in 2019 based on the historical daily electricity demand, weather data, and Columbia academic calendar in 2018 and 2019. Three models were built: Linear Regression, Random Forest Regressor, and Multilayer Perceptron Regressor. Train and Test RMSE and R2 were calculated for each model. For these three models, their Test RMSE was extensive, meaning that the daily demand as an output is not related to the selected features. All the test data were not fit well on the regression line. The most critical part was the daily minimum temperature, and the working date had less importance on daily demand forecasting. In the future, more models such as XGboost would be built for predicting the daily electricity demand, and there might be a better model with less error. Secondly, more features as input need to be investigated as well.

Reference

1. Zhao, H.-xiang, & Magoulès, F. (2012). A review on the prediction of Building Energy Consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586–3592. https://doi.org/10.1016/j.rser.2012.02.049
2. Chen, S., Ren, Y., Friedrich, D., Yu, Z., & Yu, J. (2021). Prediction of office building electricity demand using artificial neural network by splitting the time horizon for different occupancy rates. *Energy and AI*, *5*, 100093. https://doi.org/10.1016/j.egyai.2021.100093
3. Electricity demand forecasting using machine learning. (n.d.). Retrieved December 23, 2022, from https://www.neuraldesigner.com/blog/electricity_demand_forecasting
4. *2018-2019: Academic calendar: Academics: Teachers college, Columbia University*. Teachers College - Columbia University. (n.d.). Retrieved December 23, 2022, from https://www.tc.columbia.edu/academics/academic-calendar/2018-2019/
5. US Department of Commerce, N. O. A. A. (n.d.). National Weather Service. Retrieved December 23, 2022, from https://www.weather.gov/
6. Zhu, Z. (2021, September 3). *Taking seasonality into consideration for time series analysis*. Medium. Retrieved December 23, 2022, from https://towardsdatascience.com/taking-seasonality-into-consideration-for-time-series-analysis-4e1f4fbb768f