**PREDICT HOURLY ENERGY DEMAND WITH TIME SERIES FORECASTING**

Background

The world has recognized the immediate threat of climate change that can impact many crucial aspects of all lives. One focus is to reduce anthropogenic greenhouse gases such as carbon dioxide by pivoting to clean and renewable energy sources. In terms of planning and implementing green energy infrastructures, it is crucial to predict energy demands at an hourly frequency accurately. External factors such as weather could also be necessary for making reliable predictions.

Dataset

The energy datasets have been aggregated and posted on Kaggle by Rob Mulla.1 The data were derived from PJM Interconnection LLC (PJM), a regional transmission organization (RTO) in the United States. The energy data are hourly data points between 50K to 120K instances for different regions in the U.S. For simplicity, the Chicago region will be the focus of this work, where the data were obtained from Commonwealth Edison. Independently, the weather data of Chicago was also combined with the energy data. The weather data were obtained from Weather API posted on Kaggle by David Beniaguev.2

Data wrangling and preprocessing

Because the energy and weather data were derived from different sources, many NaN and duplicated data points were identified and removed. After cleaning, the energy and weather data were merged based on the DateTime index. In the preprocessing stage, the data between 2012 to 2016 are considered as training data, and the 2017 data are regarded as testing set. The categorical data of weather type were converted as the one-hot encoding format. Any weather types with occurrences less than 100 are labeled as others. For numerical data, standard scaler, mix-max scaler, and power transformer scaler from Scikit-learn were fitted and applied on the training data depending on the shape of their distributions. Then, the same scalers were used to transform the testing data.

Exploratory data analysis (EDA)

Based on the EDA, a few insights can be noted. First, about seven weather types make up most of the data (i.e., clear sky, broken clouds, overcast clouds, scattered clouds, mist, few clouds, and light rain). Other weather types are scarce.

A picture containing histogram

Description automatically generated

Second, the correlation between energy and temperature can be observed (i.e., the energies usages are high at extreme temperatures). Other than that, no clear correlations between energy, humidity, pressure, temperature, wind direction, and wind speed are noticeable.

Scatter chart

Description automatically generated

Third, the energy usage varies on an hourly, daily, and monthly basis, likely an indication of human activities and seasons. Additionally, the autocorrelations indicate that the energy data appear periodic with multiple timeframes from 24 hours, one week, and six months.

Chart, box and whisker chart

Description automatically generated

Modeling

It is necessary to use the modeling techniques that have memory components associated with frameworks to predict hourly energy demands. Here, two methods were considered: Autoregressive integrated moving average (ARIMA) and Long short-term memory (LSTM) techniques. The ARIMA model is based on linear combinations, and it contains three variables (p, q, and d). The p, q, and d are referred as auto regressive, moving average, and difference terms, respectively.3 The LSTM model is a type of recurrent neural network where an internal state is responsible for keeping information about the past inputs using cell and gate units.4 The ARIMA and LSTM models were built using Statsmodel and Keras.

Parameters

In the ARIMA model, the variables p, q, and d were optimized using the values between 0 and 5 based on the Akaike information criterion (AIC). Here, the optimal p, q, and d values are 4, 1, and 5., which means that the model requires four lags in the historical data, 1 degree of differencing to account for non-stationary, and a moving average window of 5. For the LSTM model, five lags of data were used as inputs. The LSTM model was optimized based on MSE using Adam optimizer. The network contains a fully connected one LSTM layer with 100 nodes. The data from 2013 were used as training data, and the year 2017 data were used as testing data.

Results

Between ARIMA models, the model with temperature data as an exogenous feature reduces the AIC from 106445 to 106071. Thus, adding temperature as an extra feature can improve the performance of the ARIMA model. The RMSE of these ARIMA models was reduced from 188.8 to 184.4 upon adding the temperature feature. The LSTM models have a slightly worse performance relative to ARIMA models. The LSTM without and with temperature features have RMSE of 210.9 and 202.3, respectively. Overall, ARIMA with temperature feature has the best performance.

Chart

Description automatically generated

A picture containing tool

Description automatically generated

|  |  |
| --- | --- |
| **Model** | **RMSE** |
| ARIMA | 188.8 |
| ARIMA + temperature | 184.4 |
| LSTM | 210.9 |
| LSTM + temperature | 202.3 |

Conclusions and recommendations

While both models perform relatively well for time-series energy usage data, the ARIMA has a slightly more superior performance relative to LSTM. Notably, the performance of both models did improve when the weather data (temperatures) were included.

Other strategies should be explored to improve these models:

* Both models might not be optimized at the global level. In the ARIMA model, the range of the parameters was limited between 0 and 5. For LSTM, the lags and network architecture were not broadly explored (e.g., number of layers and nodes). All of these factors could significantly affect the performance of the models.
* Only temperatures from the weather data are included at this stage. Including other weather data such as humidity, pressure, wind speed, wind direction, and weather condition could further boost the performance of the ARIMA and LSTM models.

References

1. <https://www.kaggle.com/robikscube/hourly-energy-consumption>
2. <https://www.kaggle.com/selfishgene/historical-hourly-weather-data>
3. <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>
4. <https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/>