1.1 Data Source

1.2 Purpose

The purpose of this is to explore the possibility of using ridge regression from sci-kit learn to make predictions of power usage one hour in the future based on the past four hours and predict the average use of power in the next twenty-four hours based on rolling statistics of the previous twenty-four hours. Next we will take an aggregate of statistics over days and then try to predict maximum usage for the next calendar day using both ridge regression and an LSTM neural network. The results will be examined for increasingly long time steps.

1.2.1 Technology Choice

The data source for this a JSON provided obtained by direct download from the EIA and then uploaded to the Watson studio project directly. The pandas, numpy, matplotlib, sci-kit learn, and stasmodels libraries were used. In addition to this Keras was used for deep learning.

1.2.2 Justification

The data set is relatively small and complete for the needs of the project necessitating no further use of an API.

The project is intended to be an exploration of ridge and linear regression for predicting future power usage from previous power usage. The sets are relatively small, under 19000 data points and so it was not deemed necessary to use any kind of technologies that deal with large data sets. The libraries used are also open source and would allow any user with the data sets to explore them further.

1.3 Enterprise Data

1.3.1 Technology Choice

EIA sources were used.

1.3.2 Justification

Data was directly downloaded as this is past data and from a reputable source.

1.4 Streaming analytics

1.4.1 Technology Choice

No streaming data.

1.4.2 Justification

This data was from the past and the attempt was to form a model that could be used for streaming data later. This would be complicated and beyond the scope of the current project.

1.5 Data Integration

1.5.1 Technology Choice

NA

1.5.2 Justification

The data was not transformed for storage.

1.6 Data Repository

1.6.1 Technology Choice

The data was stored in the project directory on IBM Watson Studios.

1.6.2 Justification

This allowed direct access to the data and the ability to share access.

1.7 Discovery and Exploration

1.7.1 Technology Choice

The data was explored using the pandas, numpy, matplotlib, and statsmodels libraries. Ridge regression with Cvgridsearch for optimization were used from sklearn while an LSTM neural network from keras was used later for the aggregate days. Ridge regression was chosen for it's ability to deal with lower numbers of training data points, a feature only needed for the daily series. The LSTM network was chosen because it is most commonly used for time series. R^2 was used as a scoring metric.

1.7.2 Justification

Data was extracted from the main data frame. The power demand for the lower forty-eight states would be used to train and test and the power demand in the northwestern coastal states to serve as a validation set. The data was examined for NA values but none were found. Outliers were not removed as they might represent legitimate anomalies and were considered worth noting. Dates were converted to standard datetime using pandas. Pandas was used for all dataframes.

The data was graphed over the whole time period, one day, and one week for the contiguous forty-eight states and a Fast Fourier Transform (FFT) was performed and magnitude graphed against frequency. The series appear periodic and has some major frequencies but a great deal of noise. This used numpy.

A Dickey-Fuller test was performed on the series for the lower forty-eight states to test if it was stationary, I.E. if the mean and variance did not change substantially over time. That is, there is no significant upward or downward trend. The test returned that the series was stationary with 99% confidence. This required statsmodels.

All graphs were made using matplotlib.

1. Length four vectors were made from the hourly power demand data in the contiguous forty-eight states. A target column was made from the fifth value. The fifth value served as the target. Sklearn with ridge regression was used to form a model that was optimized via CVgridsearch to find the best alpha.

2. Rolling standard deviation, mean, minimum, and maximum were taken a season column for summer was added to account for summer increases in demand. The target column was a time shifted rolling average. The accuracy of predictions for shifts up to twenty-four hours were taken and evaluations

made for each shift NA values dropped. Once more a Ridge regression model using sklearn was emplyed.

3. Aggregates of the calendar days were made for the mean, minimum, maximum, and standard deviation of power demand for both data sets. The day of the week, if that day was a work day, and summer season were added. These latter three were dropped for the first analysis and kept for the second. In both cases ridge regression models were compared against LSTM model for up to twenty-one days in the future. Here sklearn was used to do a ridge regression model for the same shifts.

R^2 was chosen as a scoring method because it measures variance between predicted and real data and should remain consistent in stationary series regardless of it's place in the series.

1.8 Actionable Insights

1.8.1 Technology Choice

Sklearn and Keras were the technologies that generated insights for this project.

1.8.2 Justification

The creation and scoring of the models in each part of the project allowed for a reasonable basis on which more usable tools might be made. As the purpose of this project was to determine the viability of using these methods in making predictions about a time series.

1.9 Possible Applications / Data Products

1.9.1 Technology Choice

Sklearn with ridge regression and Cvgridsearch optimization seems to be a good tool for both the estimation of future use one hour ahead and the rolling statistics to predict average use. For the tests on daily aggregates it was found that both methods were effective for predictions one day ahead but faltered for later predictions. The LSTM neural network did better for longer term predictions when given more information.

1.9.2 Justification

For the first test the R^2 score was .996 for the test set and .977 for the validation set. These result appear unusual but for a highly regular time series this would not be surprising as the movement by one increment should not change the values much. This does validate the possible use of this model.

For the second part the ridge model predicted average use with an r^2 of about .8 for a twenty-four hour shift. This suggests that the model is useful but could be refined.

For the third part, the ridge regression was more accurate than the neural network for fewer variables. However for larger numbers of variable the LSTM network performed better over all and better for future predictions. This suggests that a neural network can be used but would need extensive training.