A Brief Exploration of Prediction Power Consumption Using Sci-kit Learn

Introduction

Section 1.1 The Problem.

We will attempt to use sci-kit learn to assess if it could be used as a tool for forecasting a future event in a time series. We will due this by performing linear and ridge regression on a vector of previous values in a time series and attempting to predict the next value. We will use the power demand for the lower forty-eight state of the U.S. for training and the north eastern states for testing.

Section 1.2 Background

We will use a data set from the U.S. Energy Information Administration [1]. This data set will contain energy demand per hour measured for the lower forty-eight states in the U.S. as well as data one hourly demand in the north eastern states. The accurate prediction of energy demand, and general prediction of future events is of great interest for almost any field. This will be treated as a time series [2] and we will attempt to predict future values using tools in sci-kit learn. It should be noted that a time series differs from a normal series in that it is a assumed that the dependent variable has some relationship to past dependent variables [3].

Data

Section 2.1 Data Sources

As stated above all data is taken from the U.S. Energy Information Administration data on hourly power consumption. A link is provided in the citations.

Section 2.2 Data Cleaning

The data taken from this particular source, a U.S. government body, is assumed to be accurate and complete. When tested it did not have any NA values which was true in this case. However, as a time series, the data needed to be tested for to see if it was stationary or not. A stationary time series is one where ‘parameters such as mean and variance also do not change over time’ [4]. This condition can be tested for using the Dickey-Fuller test [5]. Data was broken into length four and then length two vectors of hourly consumption and the target value was the subsequent hour. When tested using the statsmodels package test it was found that the time series had a test value of -6.209709e+00, a p-value of 5.545942e-08, and a 99% confidence interval of value of -3.430700e+00.

Methodology

In order to complete this project the power usage data for the lower forty-eight states of the U.S. was divided into a training and test set. The training set would be length four vectors and length two vectors of the power demand preceding the target to predict. Two methods were chosen to attempt to predict the next hour’s demand, ridge regression and linear regression. The results of the linear regression would be measured using the absolute mean error while the results for the ridge regression were scored by the inbuilt scoring method [6]. These models would then be tested against the power consumption data for states classed as being in the north eastern part of the United States. It should be noted that before this could be attempted the time series has to be tested to see if it was stationary. This would mean some periodicity and no overwhelming upwards or downwards trends in the data over the span of this time period. The Dickey-Fuller test was used to determine this.

Results

The Dickey-Fuller test using the statsmodels library returned a test value of -6.209709e+00, a p-value of 5.545942e-08, and a 99% confidence interval of value of -3.430700e+00. This suggests that there is a 99% confidence that the series is stationary as the test value is less than the 99% confidence interval and the low p-value suggests that this is valid over the given time period. A rolling average of twelve hour periods was taken and a standard deviation for the same. Information on the moving average to be found here [7] and the rolling standard deviation here [8]. Below are some figures showing the total demand versus time and the rolling mean and rolling standard deviation are shown below.

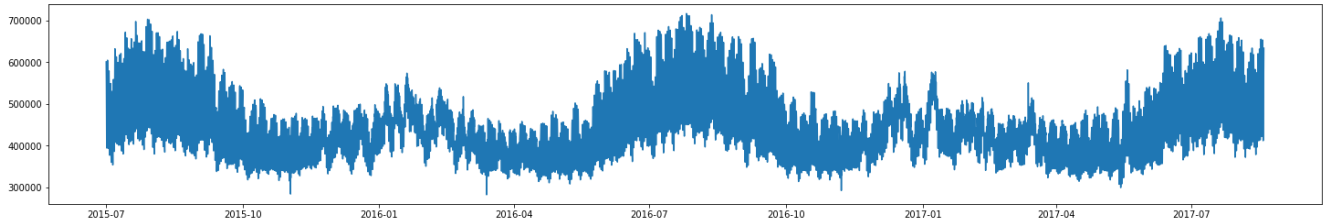

Fig 1. Power Consumption per Hour in the Lower 48 States

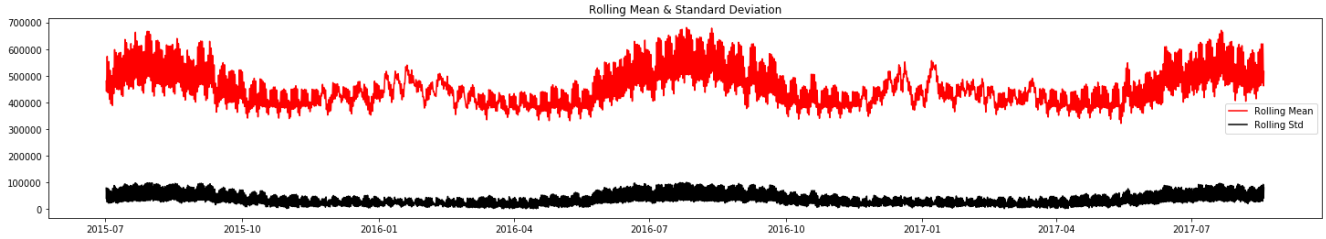

Fig 2. Rolling Mean and Standard Deviation, Twelve hour Periods

When analyzed using linear regression analysis with sci-kit learn the model showed for the length four vector coefficients of [0.1548783 -0.14050687 -1.02544636 1.98211599] and for length two [-0.89680341 1.85548955]. When tested against a validation set for length four, that was a withheld part of the lower forty-eight states data, it showed mean absolute error of 3452.468 which was a 0.756%. Respectively for length two it was 3895.422 and 0.853%. When tested against the north eastern states data it showed for length four 4178.133 and 0.915% of the average consumption per hour. For length two it showed 4514.384 and 0.988%

For the ridge regression models the R^2 value for the validation set for the length four was 0.997 while for the test set it was shown to be 0.994 both of which are extremely large values. For the length two vectors is was 0.996 and 0.994. A gridCV method was used to determine the best alpha for the regression from values [.001, .001, .01, .1, 1, 2, 5]. The best values found was 1.0.

Discussion

From the immediate analysis it seems both models performed very well in the short terms tests. In fact it seems almost too large. It is concluded that this is accurate as the linear model has given the last two hours the greatest weights. This was confirmed by the use of the length two vector models showing nearly identical results. What is happening in both models is that the next value is being approximated as close the previous. This is a common feature of a stationary series. It is therefore taken that these two models would work very well for short term predictions however their viability for a longer term prediction would be extremely dubious. However further modifications of these methods may yield good results, it is unknown. These results could also be compared to other methods such as LSTM neural networks or the statsmodels inbuilt libraries for time series.

Conclusion

While the methods above seem to be decent predictors of short term behavior they do not seem appropriate in the current form for longer term predictions. It is also recommended that only the preceding two hours are really needed for either model but four terms is slight better for ridge regression. The data predicts a high degree of periodicity and this is confirmed by the fast Fourier transform of the data.

It is recommended that both models be used at most in the short term and that any future usages be modified if used for forecasting future data. It is also recommended that they be compared in accuracy against the statsmodels library for decomposing time series or an LSTM neural network. It is also believed that the ridge regression method, which is recommended for regressions on a sinusoidal data set, may be better for long term predictions.

Citations

1. https://www.eia.gov
2. https://en.wikipedia.org/wiki/Time_series
3. https://en.wikipedia.org/wiki/Time_series
4. https://en.wikipedia.org/wiki/Stationary_process
5. https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_test
6. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
7. https://en.wikipedia.org/wiki/Moving_average
8. https://pandas.pydata.org/pandas-docs/version/0.17.0/generated/pandas.rolling_std.html