Electric Power Load Forecasting

Alexey B. Stepanov

16 April 2016

1 Introduction

Accurate models for electric power load forecasting are very important for the electric industry. Load forecasts are used by energy suppliers, independent system operators (ISOs), financial institutions, and other participants in electric energy generation, transmission, and distribution. Many methods have been proposed for load forecasting, including time series models, regression, neural networks, support vector machines, etc. [1, 2, 3, 4, 5].

Power demand pattern is usually very complex, particularly because of the deregulation of the energy industries. Thus finding an appropriate forecasting model for a specific electricity network is not an easy task. In this project we apply and compare several machine learning algorithms for developing medium-term forecasting models for the energy grid of the state of New York. We perform most of our data wrangling and statistical analysis in Python.

2 Data

We use the real-time actual load data for the energy grid of the state of New York (http://mis.nyiso.com/public/) for the years 2008-2014. (The data for 2015 was not used because of the large number of missing values).

Weather is an important factor that influences the load. Some important predictors in power load forecasting include the air temperature and the dew point. We accordingly also use the relevant weather data from the weather stations in the state of New York. This weather data is provided by the National Climatic Data Center: https://www.ncdc.noaa.gov. We aggregate multiple files containing the load data and multiple files containing the weather data. We then clean and reformat the data and merge the electric load data and weather data together.

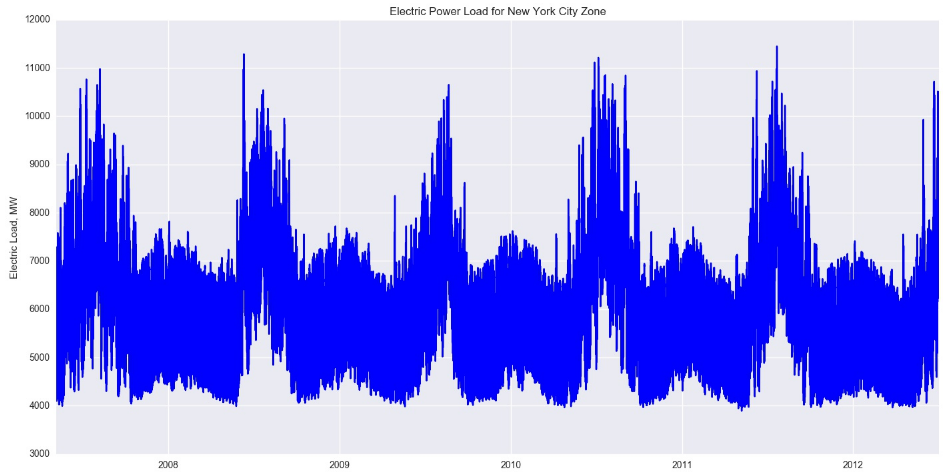

Figure 1: Electric power load for the New York City zone showing a complex almost periodic pattern

3 Models

We focus on medium-term forecasting models that aim to give accurate predictions over the period from one week to one year.

We break the date into separate parts to create the following temporal predictors: hour, day of the week, month, year, and an indicator variable corresponding to weekends. Since we expect the load data to be strongly autocorrelated, we also use the following predictors: load from the same hour on the previous day (prior day load), and load from the same hour and same day from the previous week. In addition, we use two weather-related predictors: air temperature and dew point (which is a measure of atmospheric moisture).

We split the data into training and testing sets based on the date. The observations before 31 December 2012 constitute the training set, and those after 31 December 2012 constitute the testing set. (A more sophisticated alternative to that is cross validation).

For the sake of illustration we will evaluate our models on the New York City zone. The weather station at Laguardia airport covers most of the zone. Fig. 1 shows the fluctuations of the actual electric power load for the New York City zone. Fig. 2 shows scatter plots of the actual load versus the following predictors: month, hour, air temperature, and prior day load. The scatter plot of the load versus the month predictor displays seasonal variations of the load: the power demand is higher during the hotter time of the year (air-conditioning generally requires more power than heating). The plot of the load versus the hour predictor displays daily variations of the load: the power demand

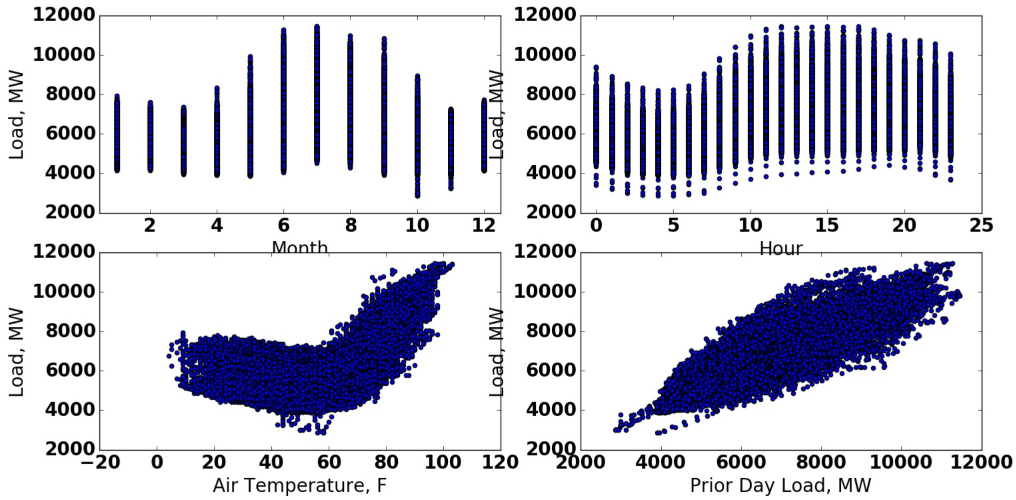

Figure 2: Scatter plots of the load versus the following predictors: month, hour, air temperature, prior day load

is smallest in the early morning between 2 a.m. and 5 a.m. and highest during the day between 12 noon and 6 p.m. The plot of the load versus air temperature displays a strong positive relationship between these variables when the temperature is above 60 $\circ F$.

Linear regression model. The linear regression model performs reasonably well. The (adjusted) R-Squared is 0.88. Fig. 3 shows a plot of the residuals versus the fitted values and a histogram of the residuals for our model. The histogram implies that the assumption that the residuals are normally distributed is not unreasonable. The plot of the residuals versus the fitted values displays no distinct pattern, implying that the other assumptions of the linear regression model (homoscedasticity, independence of errors, mean zero of errors) are not unreasonable.

Gradient Boosting. Fig. 4 shows the prediction of the gradient boosting model on the testing set and the actual value of the load.

Tree bagging. We train a bagged regression tree model that involves building an ensemble of trees, each trained on a resampled subset of the training set. Each tree creates a forecast and the average of all forecasts is taken. One of the benefits of using an ensemble is that each tree has some portion of the dataset it hasn't seen yet (out of bag observations). These can be used to perform validation of the model and also measure relative feature importance (see Fig. 3). As remarked in [5] forecasted weather parameters are usually the most important features in short-term load forecasts. Fig. 3 shows that indeed temperature and dew point are by the most important features.

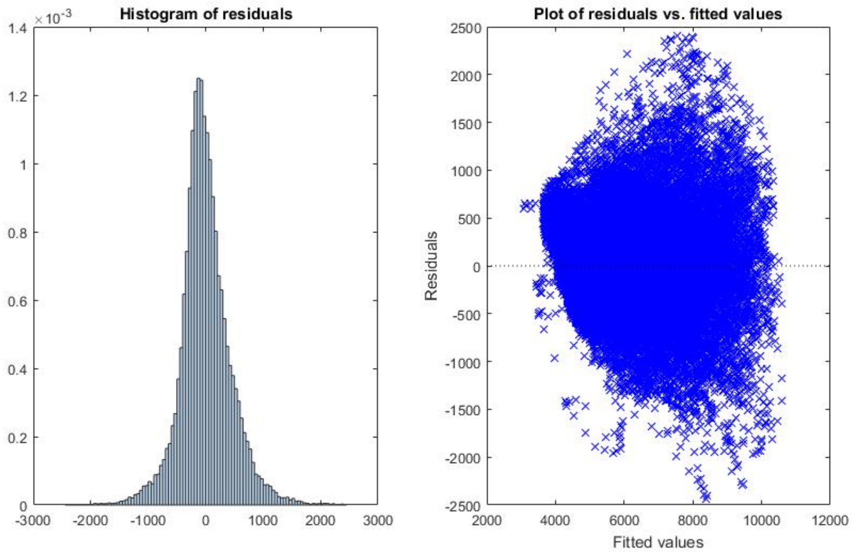

Figure 3: Linear Regression Model: Plot of Residuals vs. Fitted Values, and Histogram of the Residuals

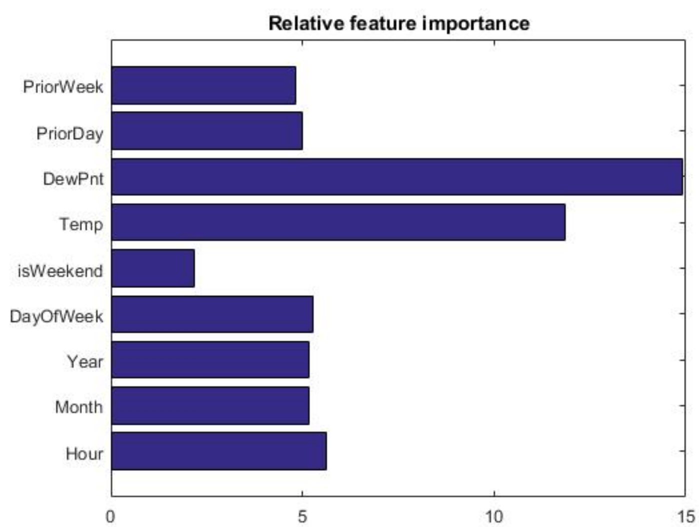

Figure 4: Tree Bagging: Relative Feature Importance

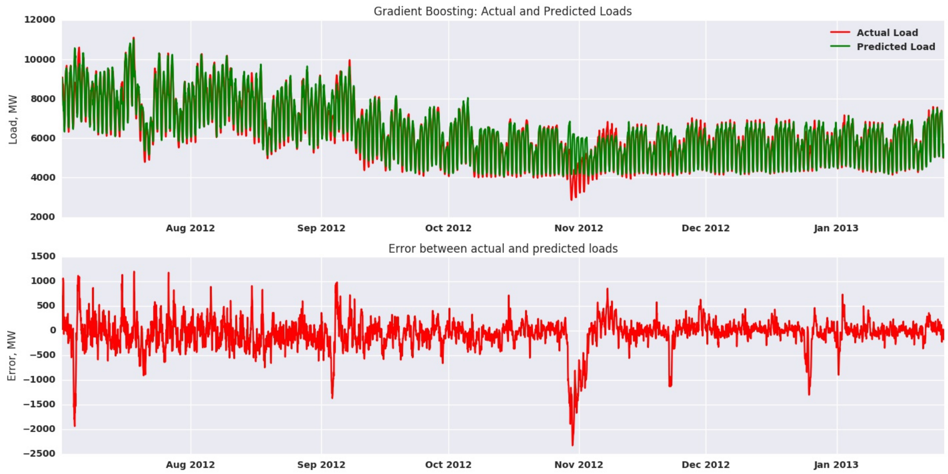

Figure 5: Above: Measured (actual) load and load predicted by the gradient boosting model. Below: Error between the actual and predicted loads.

4 Comparison of Model Performances

Fig. 5 compares the mean squared errors (MSEs) for our models on the training and testing sets. The linear regression model has the largest MSEs both on the training set and testing set. The tree bagging model has by far the lowest MSE on the training.

The performance of our models can be further improved by adding more useful predictors. For example, we could add a predictor corresponding to holidays and additional predictors related to weather. Moreover, besides the weather and temporal factors, other important factors for load forecasting include the number of customers in different categories, the appliances in the area, the economic and demographic data, the appliance sales data, etc. A more significant improvement in the performance could be achieved by using multiple weather forecasts as multiple inputs for load forecasts.

References

[1] Song, K., Baek, Y., Hong, D., Jang, G., *Short-Term Load Forecasting for the Holidays Using Fuzzy Linear Regression Method.* IEEE Transactions on Power Systems 20(1), 96101 (2005)

[2] Charytoniuk, W., Chen, M.S., Van Olinda, P., *Nonparametric Regression Based Short-Term Load Forecasting.* IEEE Transactions on Power Systems 13(3), 725730 (1998)

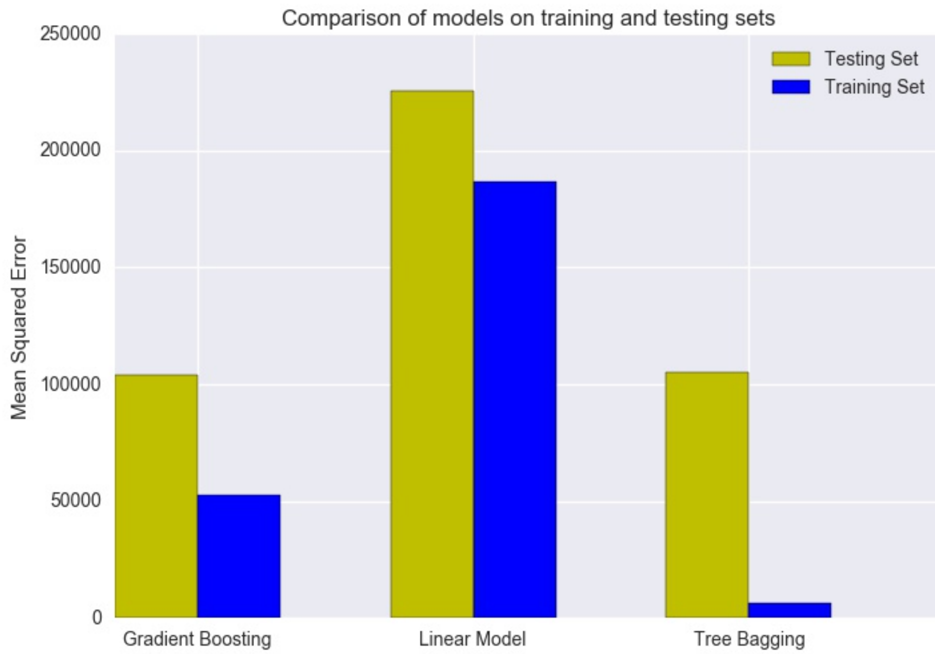

Figure 6: Performance comparison: mean squared errors for various models.

[3] Fan, J.Y., McDonald, J.D., *A Real-Time Implementation of Short-Term Load Forecasting for Distribution Power Systems*. IEEE Transactions on Power Systems 9(2), 988994 (1994)

[4] H. S. Hippert, C. E. Pedreira, R. C. Souza, *Neural networks for short-term load forecasting: a review and evaluation*. IEEE Transactions on Power Systems 16 (1), 44–55 (2001)

[5] E. Feinberg, D. Genethliou, Load Forecasting, *Applied Mathematics for Restructured Electric Power Systems: Optimization, Control, and Computational Intelligence* (J. H. Chow, F.F. Wu, and J.J. Momoh, eds.), Spinger, 269–285, (2005)