Improving prediction of demand of electricity on holidays in New England, United States

Shivani Kohli

Table of Contents

EXECUTIVE SUMMARY **3**
TECHNICAL ABSTRACT **4**
INTRODUCTION **5**
METHODS AND MATERIALS **7**
THE MODEL **11**
RESULTS **21**
SUMMARY PAGE **22**
REFRENCES AND APPENDIX **23**

EXECUTIVE SUMMARY

The electric industry faces the problem of providing sufficient energy to consumers and therefore, strives to make accurate predictions of demand. Making predictions on holidays is different from the norm as residential consumption increases while industrial reduces. Thus, I explored a method of improving predictions on holidays using a multiple linear regression model with the variable holiday. As I believed that predictions on holidays can be improved by taking into account the variable, holiday. To test my hypothesis, I created two separate models one with and the other without the variable- holiday. The dataset I used for my models was the GEFCOM2014-E (Hong et. al), which contained 9 years of hourly demand. Additionally, to check the performance of my models I calculated the difference between the predictions and the actual values.

I noticed that my models did not perform differently in general. However, the model with the variable holiday performed better on holidays. Thus, my hypothesis was true that including a parameter holiday improves the model's predictions on holidays. Therefore, it is beneficial to include the variable to the model as improved predictions allow energy supply companies to better meet their customers' demands.

TECHNICAL ABSTRACT

This paper aims to find an approach to improve the predictions for electricity demand on national holidays for the LDCs using a multiple linear regression model.

$$\hat{y}=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

The dataset utilized to develop the models was the GEFCOM2014-E (Hong et. al), which only contained temperature, date, and load for 9 years of hourly data. To make the models more accurate, I found the parameters day of the week, hour, and month, and then undertook the industry standards of calculating HDD_k, CDD_k, and performing Fourier transformations on the day of the week, month, and the hour (Vuitillo).

After the data was prepared, two models were developed for comparison purposes. One model was trained with the variable indicating it was a holiday as a boolean value and one without it. To calculate the prediction accuracy the relative residuals for each model were found and plotted.

Upon comparison of the relative residuals, it was found that the model that contained the holiday variable performed better on holidays. As it had a lower relative residual value by approximately .1, indicating it is beneficial to include the variable holiday to improve predictions.

INTRODUCTION

Electricity has a high demand (demand is interchangeably used with flow, consumption, and load) because of its commercial and residential uses: it is used to run industrial processes, cook, run life support machines in hospitals, and run cars. Local distribution companies (LDCs) have the responsibility to provide the required amount of electricity to their customers. To provide this electricity they need to make purchases beforehand as spot purchases are expensive; therefore, they require accurate predictions of demand.

Much research has been done on developing better models for electric demand forecasting. Some of the commonly used techniques to forecast are artificial neural networks (Lee, Cha, and Park), multiple linear regression models (Moghram, Ibrahim, and Saifur Rahman.), auto-regressive models (Mbamalu, and El-Hawary), and deep neural networks (Marino, Amarasinghe, and Milos). As accurate predictions can help LDCs save millions of dollars, which in turn at times lead to residents' electricity bills going down if the LDC is run by a public-sector agency (Content). Therefore, accurate predictions have a large impact.

In contrast to using complex models for creating predictions, research has also been done to find more useful parameters for demand prediction. A relationship between temperature, precipitation, solar radiation, wind speed, wind direction, and electricity consumption has been found by Hernández et al. Although they did not use any modeling techniques, just statistical tools, their research is extremely useful as it validates the assumption held that there exists a relationship between demand and temperature.

Similarly, individuals have investigated climate's effects on electricity demand. They developed models for both summer and winter using days, months, and holidays as variables. Through their investigations, they found that a model with the aforementioned variables is the

best fit model, with a single parameter for weather. They additionally found that an increase in demand was related to an increase in temperature (Crowley, Christian, and Joutz).

Crowley, Christian, Joutz, and Hernández's proved that there existed a relationship between temperature, day of the week, time of the year i.e. month and demand for electricity. Intrigued by their research, I was curious to see if there was a role of holidays, i.e. federal holidays, on demand for electricity. Demand on national holidays varies from demand on regular days as for instance on Christmas people use more lights in comparison to a regular day in June. Therefore, it is necessary to take this variability into account as it improves predictions.

Furthermore, taking the variability added by holidays on demand into account is important because not all holidays are on the same day every year and this too can affect demand. For instance, this year 4th of July was on a Tuesday, therefore, the demand of electricity on the first Tuesday of July this year would vary from the demand on the first Tuesday of July last year. Thus, taking into account just the day of the week does not meet all requirements. This is an interesting factor to look at as it gives the LDCs more insight into demand of their customers and allows them to make better decisions, thereby, reduces spot purchases and in turn save money.

Thus, in the larger domain of energy prediction, I decided to investigate a smaller problem, i.e. could prediction of demand on holidays be improved by adding the parameter holiday to a multiple linear regression model? My hypothesis was that the predictions could be improved by including the parameter.

METHODS AND MATERIALS

To answer my research question, I developed two multiple linear regression models with and without the variables holiday. A multiple linear regression model is similar to the linear regression models we developed in class. A multiple linear regression model differs in that instead of modeling a relationship between a dependent variable and an independent variable, it models a relationship between a dependent variable and multiple independent variables (*Multiple Linear Regression*).

A multiple linear regression model's equation is:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + .. + \beta_n x_{in}$$

Here $\hat{y}$ is the predicted values, β_0 is the base load, i.e. the minimum demand in my situation as it is believed that there is constant minimum electricity consumption, and the other βs represent how the output is related to the input. Thus, the βs are similar to the 'a' coefficient we found in class for least square analysis.

$$\beta = X \backslash Y$$

where Y is my demand from the training dataset

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

and X is the matrix which contains my different independent variables from the training dataset such as day of the week, month, hour i.e. time of the day where hour 1 is one am and so forth, holiday, and temperature.

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

The dataset I used for predictions is from ISO's operating area, New England, United States (Hong et. al). The original data set contained 9 years of hourly temperature, date, flow, and the hour, i.e. time of day. I needed more parameters such as day of the week, month, and is it a US holiday, therefore, I used pandas' data frame tools in python to add these additional parameters to my data. Using pandas, I received the holidays as boolean values where 1 indicated it was a holiday and all other days were attributed 0. Furthermore, the day of the week, month, and hour where values ranging from 1 to 7, 1 to 12, and 1 to 24 respectively which are not extremely useful as they do not indicate much about the data. Therefore, I undertook Fourier transformations as suggested by Vitullo et. al because "periodic phenomena's can be represented using Fourier series" (Vitullo et al).

For the day of the week, month, and the hour I calculated:

$$cos(\frac{2\pi * DOW}{7})$$

$$sin(\frac{2\pi * DOW}{7})$$

$$sin(\frac{2\pi * Month}{12})$$

$$cos(\frac{2\pi * Month}{12})$$

$$sin(\frac{2\pi * Hour}{24})$$

$$cos(\frac{2\pi * Hour}{24})$$

As the week has 7 days it is divided by 7, month by 12, while the hour was divided by 24.

Another factor that was taken into account is cooling degree days (CDD) and heating degree days (HDD). As the initial data was not linear and made fitting a best fit line difficult. By calculating the HDD_k and CDD_k values I was able to make the data more linear. In industry, these transformations are undertaken because it is believed that when it is 65 degrees Fahrenheit outside we require neither heating nor cooling to remain comfortable. Thus, degree days are calculated.

Muller states when the temperature is above 65 Fahrenheit we do not require space heating and thus consumption is constant implying a need to calculate HDD values which can be calculated as

$$HDD_k = max(0,65 - T_k)$$

Additionally, CDD values are calculated because when the temperature is below 65 Fahrenheit we do not require cooling and there is a certain average baseload (Büyükalaca). CDD is calculated as

$$CDD_k = max(0,T_k - 65)$$

Thus, using the aforementioned steps, I prepared my data. Following which I divided my dataset into an eighty-twenty ratio, 80% of the dataset was used for training, i.e., 7 years of data and 20% was used for testing, i.e., 2 years of data. As this is the industry standard for dividing the dataset for time series analysis (*Splitting the data into training and evaluation data*). This prevents overfitting as the model has never been exposed to the testing values.

Lastly, to find the accuracy of the predictions residuals and relative residuals were calculated using

$$residuals = \hat{y} - y$$

$$relativeResiduals = \frac{\hat{y} - y}{y}$$

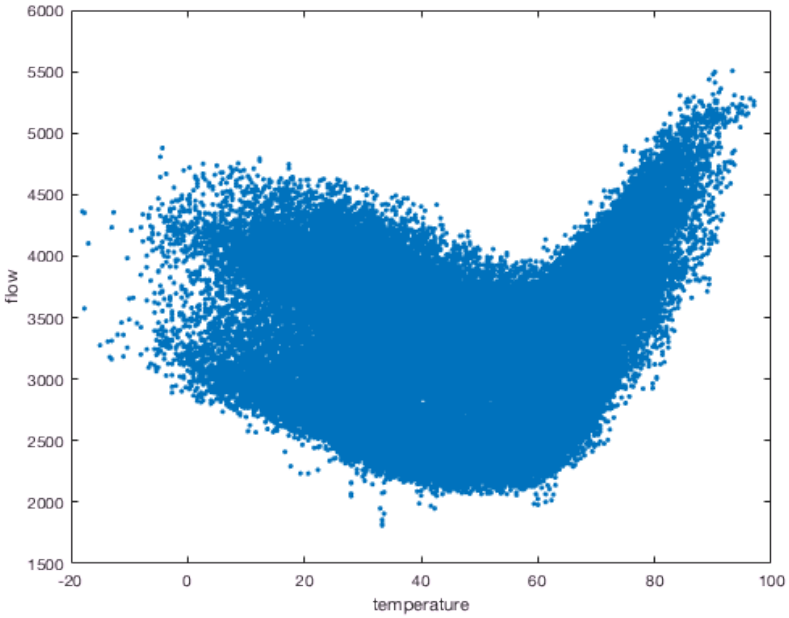
where $\hat{y}$ denote predicted values and y denote the actuals.

THE MODEL

The modeling approach I undertook to create my models is the deterministic approach, i. e, an approach in which the outcomes are determined by the parameters and initial conditions. Additionally, the hypothesis I made for the model is that it would improve the predictions on holidays by including data regarding a particular day being a holiday. I also assumed that by removing just the variable indicating that the given day is a holiday (is_holiday variable) and keeping all the other parameters which are the transformed day of the week, hour, month, HDD and CDD values; I would see the changes in the model as an effect of the is_holiday variable independently. I also, assumed that the aforementioned variables did have a correlation with electricity demand.

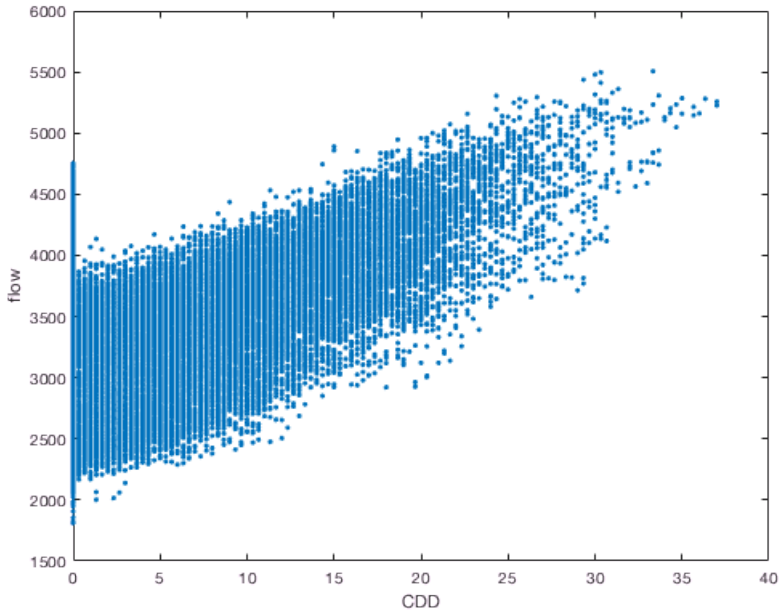
To test my hypothesis, I developed multiple linear regression models one the with is_holiday variable taken into account and another without the variable. To check the accuracy of the models the residuals and relative for both the models were found which were relatively small. Thus, my models were performing well overall and could be used to study demand on holidays.

Before modeling my data, the initial graph for temperature versus flow was plotted, in which it was noticed that the data was not linear and therefore it was difficult to fit a linear best-fit line.

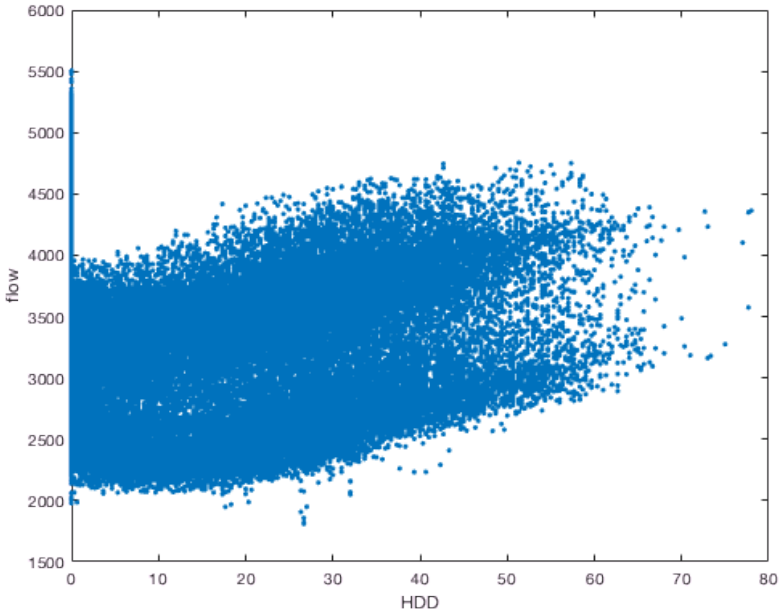


Additionally, in the graph we can see, around 65 Fahrenheit the demand was initially decreasing and after which it is increasing. Therefore, it was required to calculate CDD_k values and HDD_k values to make the data more linear.

CDD versus flow (demand) plot:



HDD versus flow (demand) plot:



To calculate my coefficients for the first model, I first set up a matrix X,

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

where x_{n1} = CDD_k, x_{n2} = HDD_k, x_{n3} =is_holiday(first model), x_{n4} = sin hour, x_{n5} =cosine hour, x_{n6} =the day of the week sine, x_{n7} = the day of the week cosine, x_{n8} = month sin, and x_{n9} = month cosine. And a 1 dimensional matrix Y where y_n is the flow at a particular hour. Using the matrixs I called the coefficients for the models, b = X\Y.

The equation I received for the first model was:

$$\hat{y} = 2939 + 50x_1 + 15x_2 - 142x_3 - 456x_4 - 369x_5 + 68x_6 - 140x_7 - 141x_8 - 77x_9$$

To calculate my coefficients for the second model, I set up a matrix X,

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

where $x_{n1} = CDD_k$, $x_{n2} = HDD_k$, x_{n3} = sin hour, x_{n4} =cosine hour, x_{n5} =the day of the week sine, x_{n6} = the day of the week cosine, x_{n7} = sin month, and x_{n8} = cosine month for the first model.

The equation for the second model:

$$\hat{y}_2 = 2607 + 73x_1 + 5.7x_2 + 143x_3 + 63x_4 + 138x_5 + 10x_6 + 124x_7$$

To calculate the residuals on days it was a holiday, I calculated

$$residuals = (\hat{y}(holiday == 1) - y(holiday == 1))$$

This equation subtracts the product of the predicted with the is_holiday vector and the actuals with the is_holiday vector and return the residuals only on days that it was a holiday, or the boolean value was equal to 1.

To calculate the residuals on days it was a holiday in the model that did not take into account whether it was a holiday I calculated

$$residuals = (\hat{y}_2(holiday == 1) - y(holiday == 1))$$

Using these values, I calculated the relative residuals for both the models using the equation:

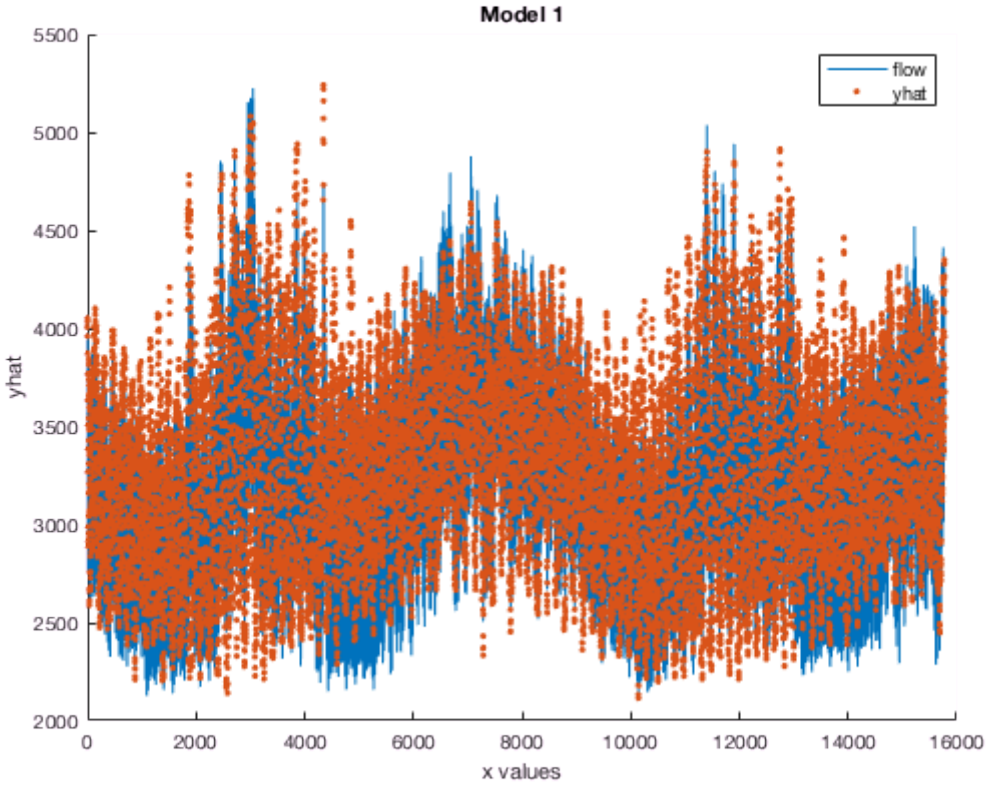
$$RelativeResiduals = \frac{(\hat{y}(holiday == 1) - y(holiday == 1))}{y(holiday == 1)}$$

ANALYSIS OF MODEL

The first model's equation was

$$\hat{y} = 2939 + 50x_1 + 15x_2 - 142x_3 - 456x_4 - 369x_5 + 68x_6 - 140x_7 - 141x_8 - 77x_9$$

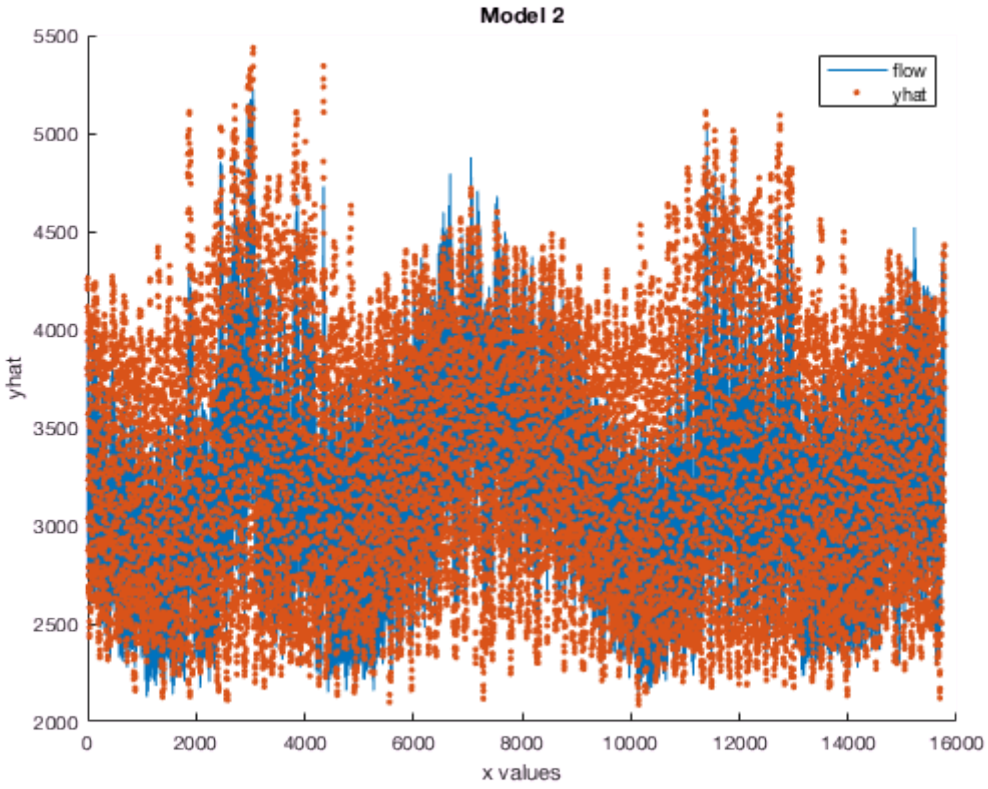
The model was then plotted against the flow from the test set:



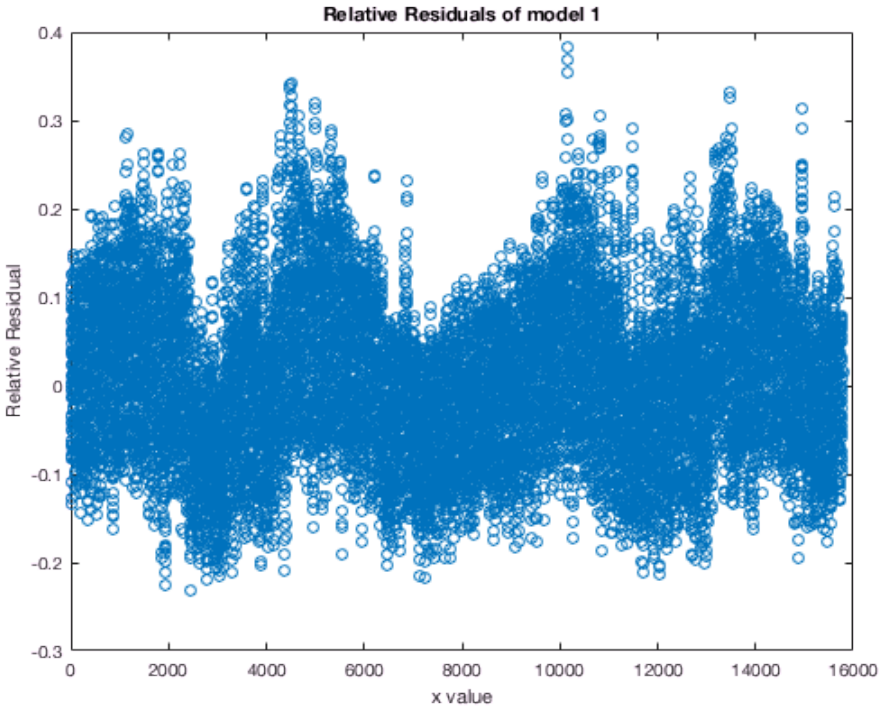
Similarly, the second model was found, without the holiday parameter. The equation for it was

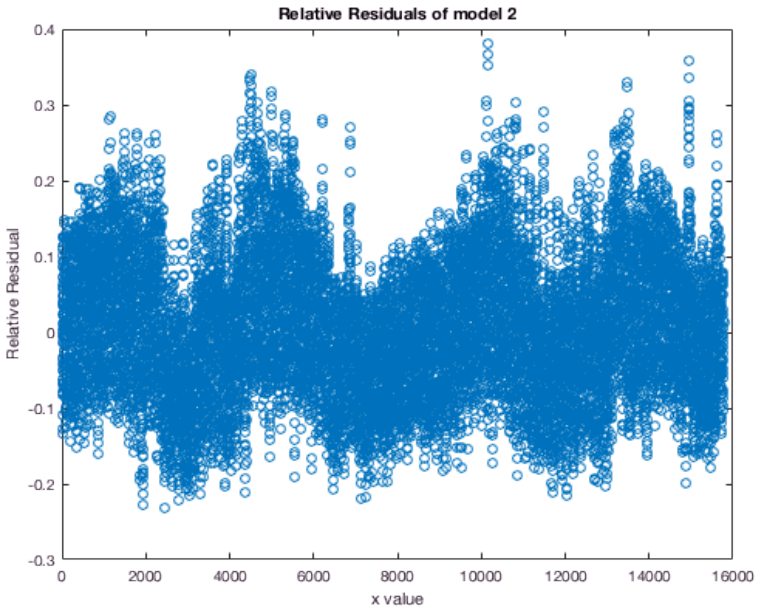
$$\hat{y}_2 = 2935 + 50x_1 + 15x_2 - 456x_3 - 370x_4 + 65x_5 - 143x_6 - 141x_7 - 81x_8$$

The model was then plotted against the flow from the test set:



To check the accuracy of predictions, the relative residuals were calculated.



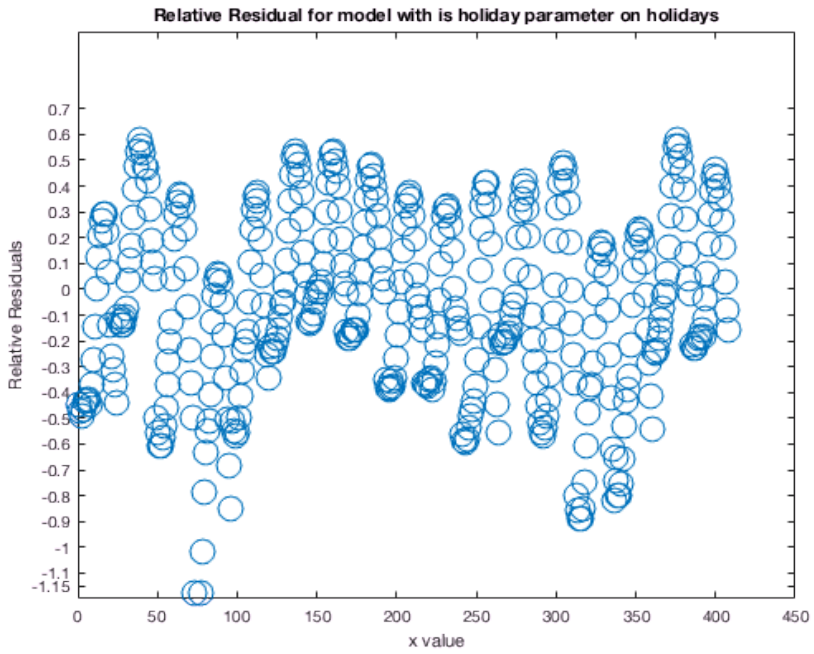


We notice that the relative residuals for both the models are similar and neither of the relative residuals is extremely high. Therefore, we can say that the models are performing well, allowing us to use the models to calculate the relative residuals for holidays using the equations' we found for both the $\hat{y}s$.

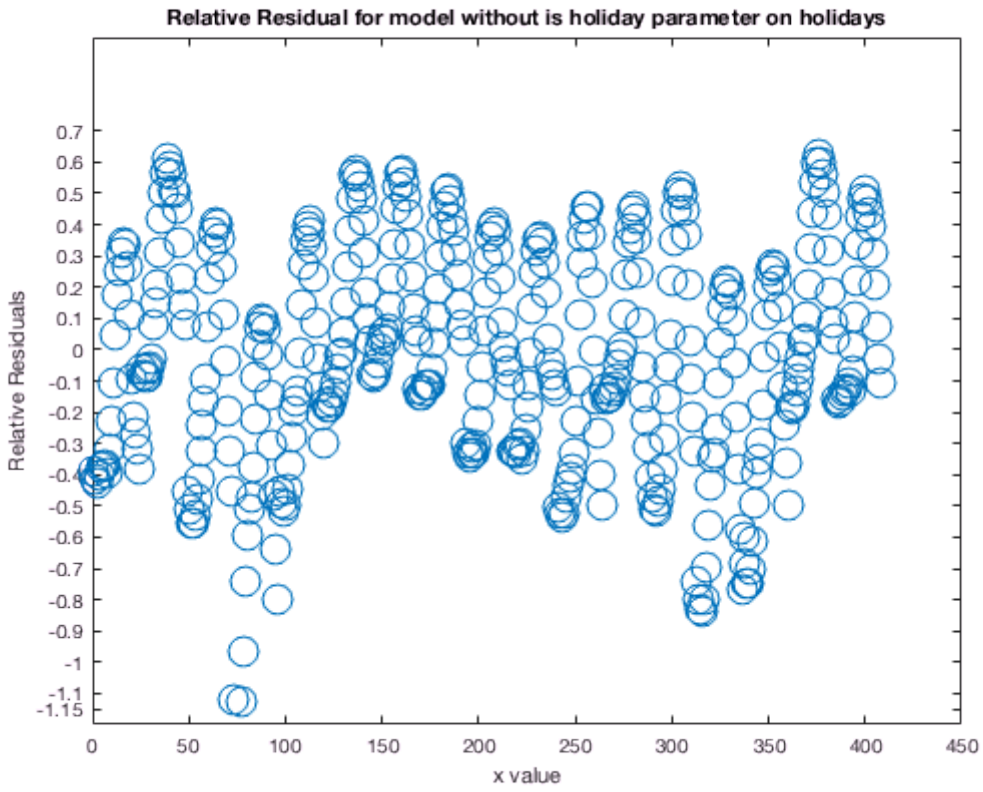
After analysis of the models using their relative residuals, the models appear to perform well and therefore, can be used to validate the initial hypothesis. To do so, I used the $\hat{y}$ and y values to calculate the relative residuals on the days it was a holiday, i.e., the boolean value was true.

Using the models, I received the following relative residuals of the models on just the holidays.

The relative residual for model one on holidays:



The relative residual for model two on holidays:



The relative residuals for holidays is lower for model one by approximately .1 compared to model 2. This confirms the initial hypothesis that the model does perform better on holidays if we include a is_holiday variable.

RESULTS

In the above section, we saw that the relative residuals for the days it was a holiday using the multiple linear regression model with the parameter indicating it is a holiday we received better predictions. Thus, our initial hypothesis was found to be true that adding the parameter to a multiple linear regression model improves the predictions. Additionally, using the equations I found for my models we could also look into the demand of electricity on particular holidays, for instance, the demand on 4^{th} of July.

Furthermore, the same process can be undertaken to find the improvement of the demand in other regions of the world using different datasets. Although my model has only been trained on the dataset for New England, United States we could use the exact same model for locations with similar temperatures or retrain the model on a different dataset.

In my model, the difference in relative residuals was not significant it is still relevant because my models were rudimentary and yet they had a difference in their relative residuals. Thus, training complex models with the parameter, is it a holiday, will have a larger improvement in predictions as compared to my model. This improvement in predictions can lead to better understanding the demand and saving corporations money.

SUMMARY PAGE

To validate the initial hypothesis, two multiple linear regression models were implemented. The first model contained the parameter is holiday while the second model did not contain it. The two models were separately trained on 80% of the GEFCom2014-E dataset that had undergone data manipulation while the other 20% was used for testing purposes.

To check the accuracy of the models, relative residuals were calculated and found to be extremely low. Thus, they were deemed fit to calculate the performance of the models on holidays specifically. It was found that the model with the parameter holidays, performed slightly better than the model without holidays. The models' relative residuals differed by approximately .1, indicating that taking into account whether it is a holiday does have an effect on predictions.

From, the results obtained, it can be concluded that to improve the accuracy of predictions on holidays, including holidays as a parameter is useful.

REFRENCES AND APPENDIX

Asbury, J. G., C. Maslowski, and R. O. Mueller. "*Solar availability for winter space heating: an analysis of SOLMET data, 1953 to 1975*." Science206.4419 (1979): 679-681.

Büyükalaca, Orhan, Hüsamettin Bulut, and Tuncay Yılmaz. "Analysis of variable-base heating and cooling degree-days for Turkey." *Applied Energy*69.4 (2001): 269-283.

Brown, Scott H. "Multiple linear regression analysis: a matrix approach with MATLAB." *Alabama Journal of Mathematics* 34 (2009): 1-3.

Content Thomas, "*Lab learns to predict gas demand"* http://archive.jsonline.com/business/81051892.htm (2010/01/05)

Crowley, Christian, and Frederick L. Joutz. "*Weather effects on electricity loads: Modeling and forecasting 12 December 2005*." Final report for US EPA on weather effects on electricity loads(2005).

Marino, Daniel L., Kasun Amarasinghe, and Milos Manic. "Building energy load forecasting using deep neural networks." *Industrial Electronics Society, IECON 2016-42nd Annual Conference of the IEEE*. IEEE, 2016.

Mbamalu, G. A. N., and M. E. El-Hawary. "Load forecasting via suboptimal seasonal autoregressive models and iteratively reweighted least squares estimation." *IEEE Transactions on Power Systems* 8.1 (1993): 343-348.

Moghram, Ibrahim, and Saifur Rahman. "Analysis and evaluation of five short-term load forecasting techniques." *IEEE Transactions on power systems*4.4 (1989): 1484-1491.

Giordano, *An introduction to Mathematical modelling,* Edition 5

Hernández, Luis, et al. "*A Study of the Relationship between Weather Variables and*

Electric Power Demand inside a Smart Grid/Smart World Framework." Sensors, vol. 12, no. 12, 2012, pp. 11571–11591., doi:10.3390/s120911571.

Lee, K. Y., Y. T. Cha, and J. H. Park. "Short-term load forecasting using an artificial neural network." *IEEE Transactions on Power Systems* 7.1 (1992): 124-132.

Multiple Linear Regression, www.stat.yale.edu/Courses/1997-98/101/linmult.htm.

Splitting the data into training and evaluation data, http://docs.aws.amazon.com/machine-learning/latest/dg/splitting-the-data-into-training-and-evaluation-data.html

Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli and Rob J. Hyndman, "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond", International Journal of Forecasting, vol.32, no.3, pp 896-913, July-September 2016.

Vitullo, Steven R., et al. "Mathematical models for natural gas forecasting." *Canadian applied mathematics quarterly* 17.4 (2009): 807-827.