Model and Features Selection by Data Driven Analysis in Linear Prediction

Renfei Sun

April 29, 2019

1 Abstract

With the high speed development of Machine Learning technology, many of models – from linear regression to Bayes probability, from one layer perceptron to neural network – were discovered or produced by Computer, Math and Engineering Scientists. Especially, after people started to use neural network models, the speed of model developing get a critical increasing. Based on analysis of the development of Convolutional Neural Networks, Recurrent Neural Networks, AutoEncoders, and Deep Learning, we could say that nowadays, Scientists tried to find more powerful model driven by data, instead of driven by the math or statistic logic. Linear Model versus neural network is a good example. When people use linear model, such as linear regression, Bayes probability, or perceptron, they knew the details of the models. People understand that the models are using less square error, max likely-hood, or perceptron to find the appropriate weights. Due to people know the details of the model, they pay much attention on tuning the model by decided a loss function, kernel function, and regularization, but people did not get high accuracy prediction result. Compared with CNN, people do not 100% focus on the logic of model (they call the model a black box), but pay much more attention on the relations in each feature with other features. Therefore, when we select a model to predict our data, we could try to analyze much more on training data. By analysis of the relationship between features and labels, features and features, we could find out a appropriate prediction model.

2 Introduction

The increasing number of Machine Learning models could be choice by user, and many library of Machine Learning could be used as sketch. It will be much more helpful if user could pay more attention on understanding of the relationship between features and features, features and labels in order to quick find out a appropriate model, and try to develop it in order to make the model be much more accuracy.

In this paper, it will introduce a doable method to analyze data set, and select a appropriate model based on the analysis. The analysis method in this paper is called "Data Driven Analysis".

In Section 3, this paper will show the result of prediction without any data analysis.

In Section 4 & 5, this paper will show the details about how to do the analysis, and the model selection decision by analysis.

In Section 6, this paper will analyze the reason

Table 1: Models performance without data analysis

Model	RMSE	R^2	MAPE
Multiple Linear Regression	84.44	0.09	52.60
SVM Radial	94.63	-0.15	41.67
Gradient Boosting Machine	86.23	0.05	52.65
Random Forest	113.25	-0.65	74.95

why the prediction in Section 3 got a high RMSE (root of mean square error).
In section 7, this paper will show the result of prediction after the data analysis. The result shows that "Data Driven Analysis" gains the RMSE (root of mean square error) decrements 80%, and Gradient Boosting Machine is the best model.

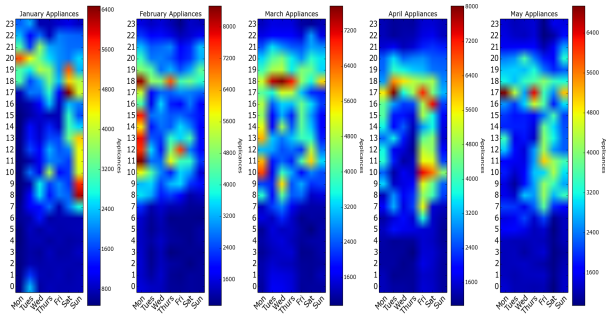


Figure 1: Residuals Plot Cases

3 Prediction without Data Driven

In Luis' paper, they tried to use Multiple Linear Regression, SVM, Gradient Boosting Machine and Random Forest to predict the energy usage by given temperature and humidity in house (each room) and outside, light, wind speed, pressure, visibility, and time. Table 1 presents the performance of the trained models in the testing set. From this table, RMSE of Multiple Linear Regression has the highest performance. It is obviously that using Multiple Linear Regression or SVM Radial to do the prediction is the best option, since the RMSE and MAPE is the lowest compared other methods.

4 Data Distribution Analysis

4.1 Data Distribution by Month

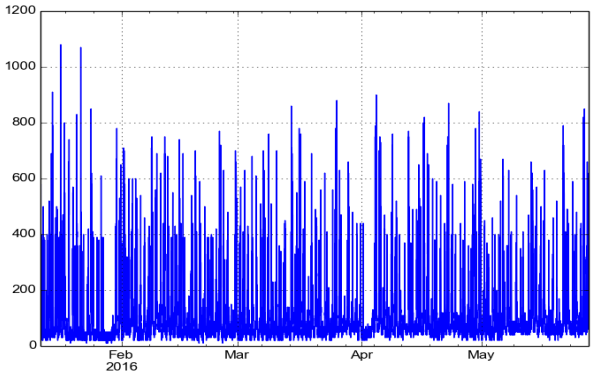


Figure 2: Residuals Plot Cases

Figure 1, Figure 2 shows that the energy usage is distributed eventually from February to May. The data does not have bias between each month. It could be fair that treat each month as same. The prediction could be accuracy without bias.

4.2 Data Distribution by Frequency

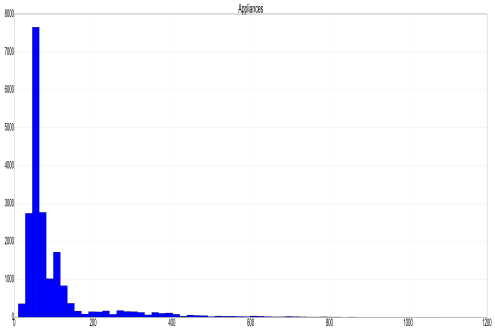


Figure 3: Residuals Plot Cases

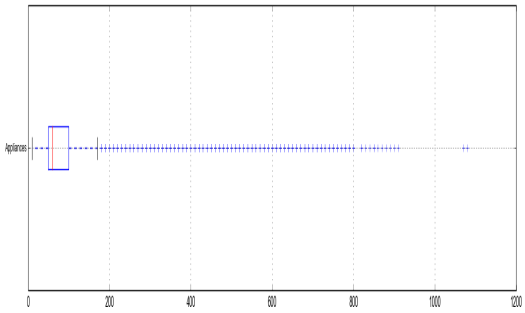


Figure 4: Residuals Plot Cases

Figure 3, Figure 4 shows that the energy usage per hour in each day is not eventually. Most of hours, the energy usage lays on 60w. So, this data is valuable to do the prediction, since if the energy usage is same at any time and any date, it is non-useful to do a prediction. The next step is to say, how the energy usage related to the time.

4.3 Data Distribution by Hour

Figure 1 represents that the energy was used mostly around 7am to 9 pm. Between these 14 hours, the most energy was used around 6 pm.

4.4 Effect by Energy Usage Distribution

Based on the energy usage analysis, the time of the day is critical effect on the energy usage. In Luis paper, figure 15 (page 91) has already showed that the NSM (number of seconds to midnight) is the most important and valuable feature used for prediction energy usage [1]. However, the point is that although the energy usage influenced by time, is it the linear relationship between time and energy. People start to use more energy from 7 am, and start to stop to use it after 9 pm. People do not turn on the device smoothly or turn off the device smoothly, but they start to use energy immediately after they wake up, and turn off immediately when to decide to sleep. Therefore, In Section 3, the conclusion of using Multiple Linear Regression should be the best model might be wrong.

5 Data Features Analysis

Appendix A-D represents the correlation between 1. each feature and label, 2. features and features. From the figure, the temperature in each room is positive correlation property with other temperature in the other rooms, and humidity has same property. The correlation between energy and other features are not very high, which cannot say positive or negative correlation. However, the energy usage was effected by human activity too much. It is not fair to check the influence of temperature to the energy usage in different human

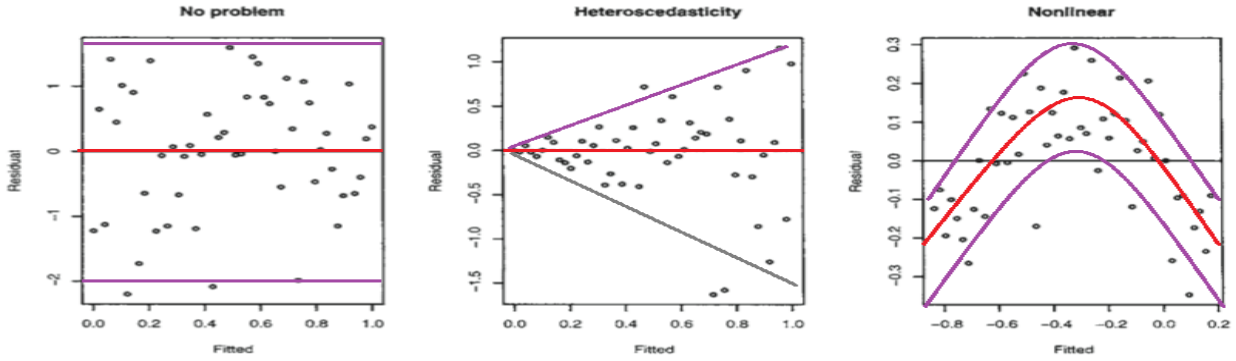


Figure 5: Residuals Plot Cases

activity time. For example, one case is that at 3am in the morning, the outside temperature is really low, but energy usage is also low, but at noon, the outside temperature is high, and energy usage is high, thus the conclusion is energy usage is positive correlation with outside temperature; however, the main reason that lower energy usage happened since people are sleeping, and they do not use many devices, but at noon, people are cooking or do other activities, then the energy usage is higher at noon. In contrast, check the energy usage at 3am with different outside temperature. Since the outside temperature is low, people would like to turn on the air-conditioner for sleeping, so outside temperature is negative correlation with outside temperature. At midnight, people do not do many other activities but sleeping, it is much confident to say outside temperature effect the energy usage. Therefore, it is not accuracy to mix the data between human activity and human non-activity. In order to get accuracy prediction model, the data should be split into two parts by human activity time, and human non-activity time.

6 Evaluation of Prediction without Data Driven

Confusion matrix is popular used for evaluation in classification problems, but it is not useful in regression problems. In the regression problems, the efficient method for evaluation is "Residuals Plot". The examples of Residuals Plot are shed in figure 5. There are 3 differences cases, which are "No problem", "Heteroscedasticity", and "Nonlinear". "No problem" means that the model picked by user is correct no matter the RMSE (mean square error) value. "Nonlinear" means that the model was picked wrong, i.e. current model could not describe (or predict) the features of the data no mater how many iteration the user tunes. "Heteroscedasticity" is a litter hard, it cannot say that the model was wrong, but the data was not good. The main reason that data lays into "Heteroscedasticity" is that the standard deviation of data is not constant. In another word, the current features (temperature, pressure, etc.) picked did not present data very well. The energy usage is not constant on the regression line drew by these features. The solution of "Heteroscedasticity" is pick more other

Table 2: Models performance after data analysis / Human Activity Period

Model	RMSE	R^2	MAPE
Multiple Linear Regression	103.83	0.04	67.48
SVM Radial	112.76	-0.13	56.39
Gradient Boosting Machine	156.812	-1.18	132.91
Random Forest	113.25	-0.65	74.95

Table 3: Models performance after data analysis / Human Non-Activity Period

Model	RMSE	R^2	MAPE
Multiple Linear Regression	24.86	-0.17	13.96
SVM Radial	24.75	-0.12	11.27
Gradient Boosting Machine	23.34	0.01	10.53
Random Forest	28.21	-0.45	13.96

relevant features.
Appendix E-H represents the Residuals Plot of linear regression, GBM, SVM, and random forest. From these four figures, they all showed that the data should be split into different groups, since they have different standard deviation.

7 Prediction with Data Driven

Based on the discussion on the above, the data features are not present the energy usage every well. Although it is impossible to pick up new data, but try split the data into 2 groups, since the discussion above showed that human activity effect on energy usage too much than other features. Split the data into 2 groups, and remove the time feature. Group one is the data from human-activity period, i.e. 7 am to 9 pm, and another group is human-nonactivity period. Since the standard deviation of data in these two groups is different, mix them together to predict, which could not get reasonable result.
table 2, table 3 represents the error analysis of two group prediction. From the 2 tables, they make much more sense than previous. In the people non-active period, the main energy usage is from emergency alert, refrigerator, and air conditioner. Emergency alert and Refrigerator are running 7/24, they should be ignored. The only thing left effect on energy usage is air conditioner, and air conditioner usage is very relevant with temperature. Therefore, prediction on hum non-activity period get high score making sense. In opposite, the prediction in human activity period get low score. The reason is that in the human activity period, there are lots of other device increasing the energy usage than air conditioner. At this period, the temperature is not relevant with energy usage very much.

8 Conclusion

Data driven model or features selection could give a significant improvement. Plot the analyse the

distribution of data to make sure that data distribution is eventually (not bias) and repeatably (less noise). Before choose the model or change the features, analyze the data, and find out the correlation between features and label to make sure that the features picked up are valuable, and labels are relevant to the features. Next, use simple standard machine learning model library to do simple prediction, then draw the Residuals Plot. By focus on the Residuals Plot, check if the model or features chose are appropriate to the data. If the Residuals Plot got "Heteroscedasticity" error, try to select other features to decrease the data standard deviation (i.e. features are more related to labels). If the Residuals Plot got "non-linear" error, try to find another more appropriate model. Flowed by the "Data Driven method" could increase the pertinence in model selection and features selection.

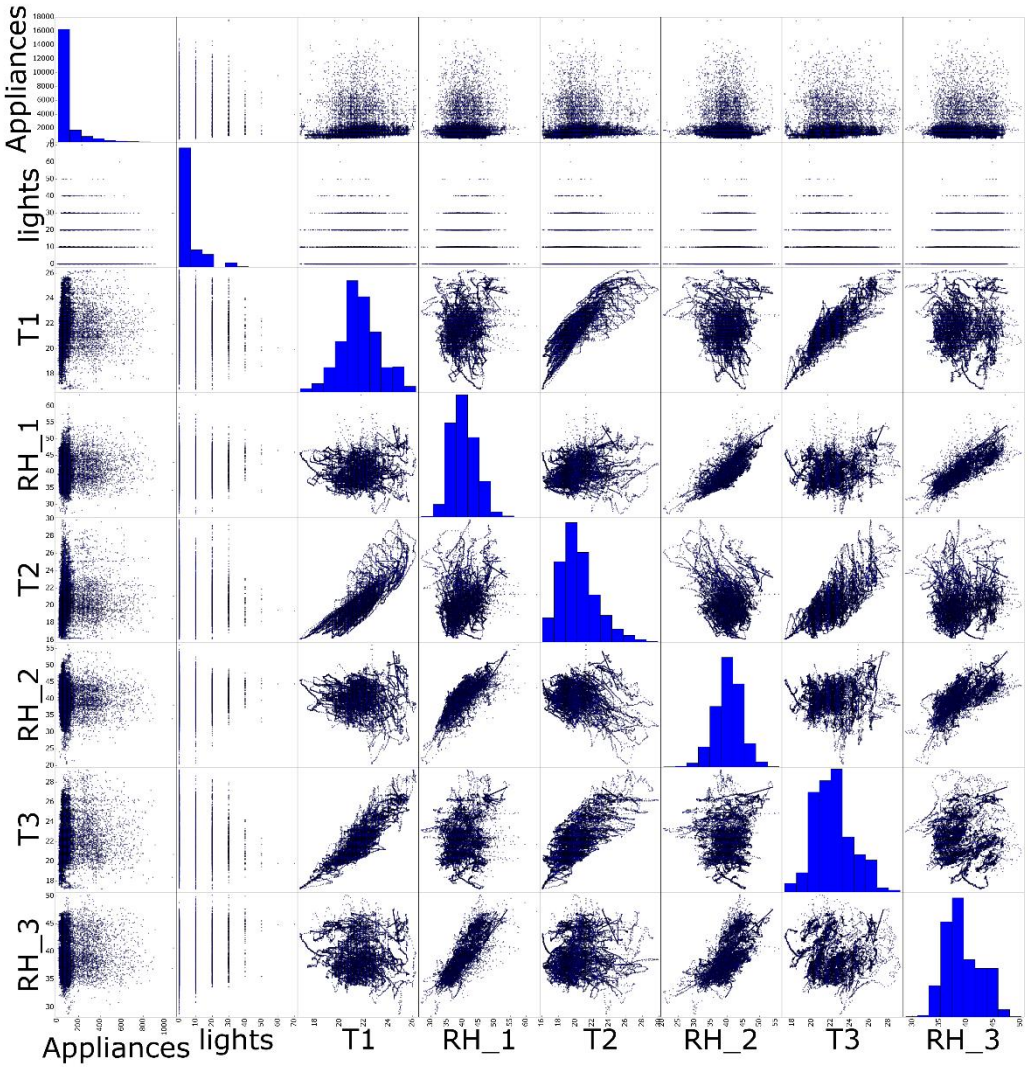
9 References

[1] Luis, Candanedo M. Data Driven Prediction Models of Energy Use of Appliances in a Low-Energy House. Energy and Buildings, Elsevier, 31 Jan. 2017

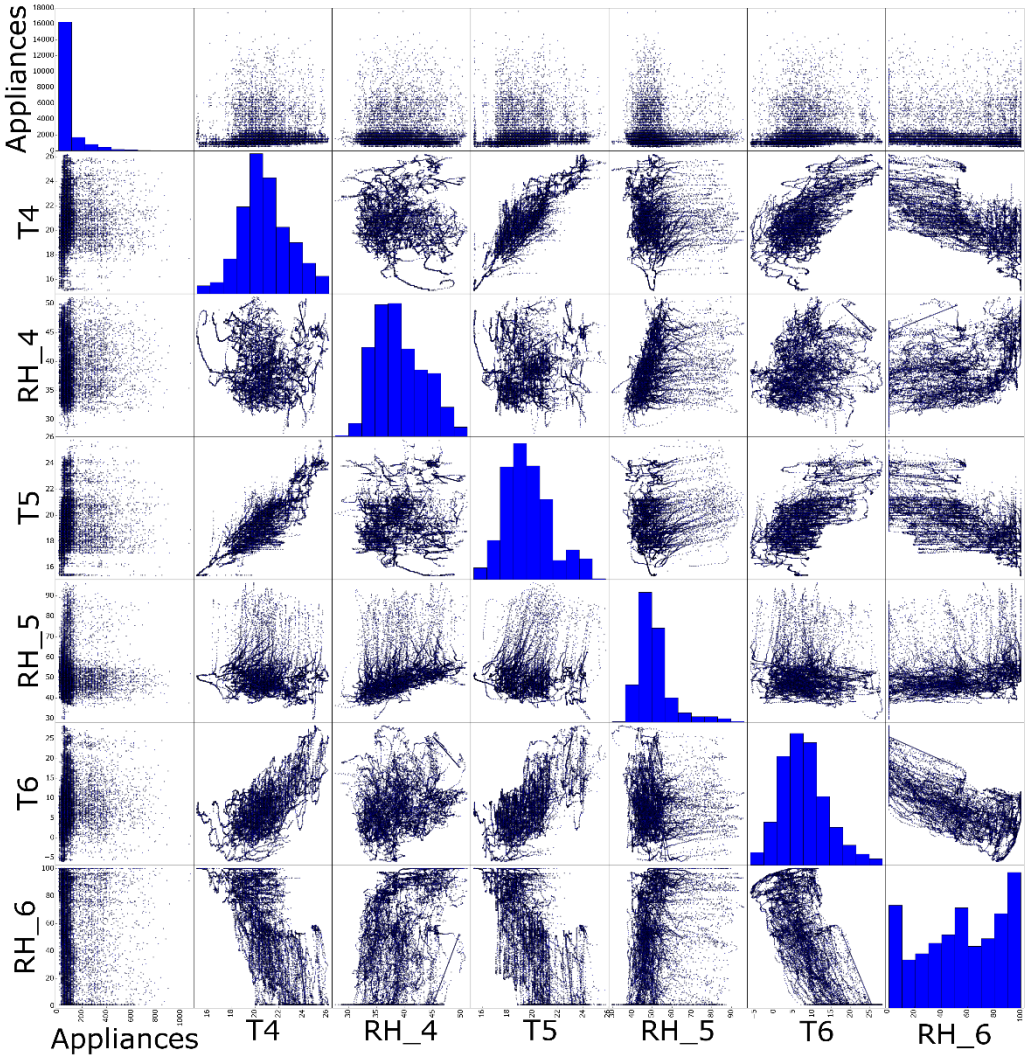
[2] THEO GASSER, LOTHAR SROKA, CHRISTINE JENNEN-STEINMETZ, Residual variance and residual pattern in nonlinear regression, Biometrika, Volume 73, Issue 3, December 1986, Pages 625633, https://doi.org/10.1093/biomet/73.3.625

[3] Hamidieh, Kam. A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor. Computational Materials Science, Elsevier, 10 Aug. 2018.

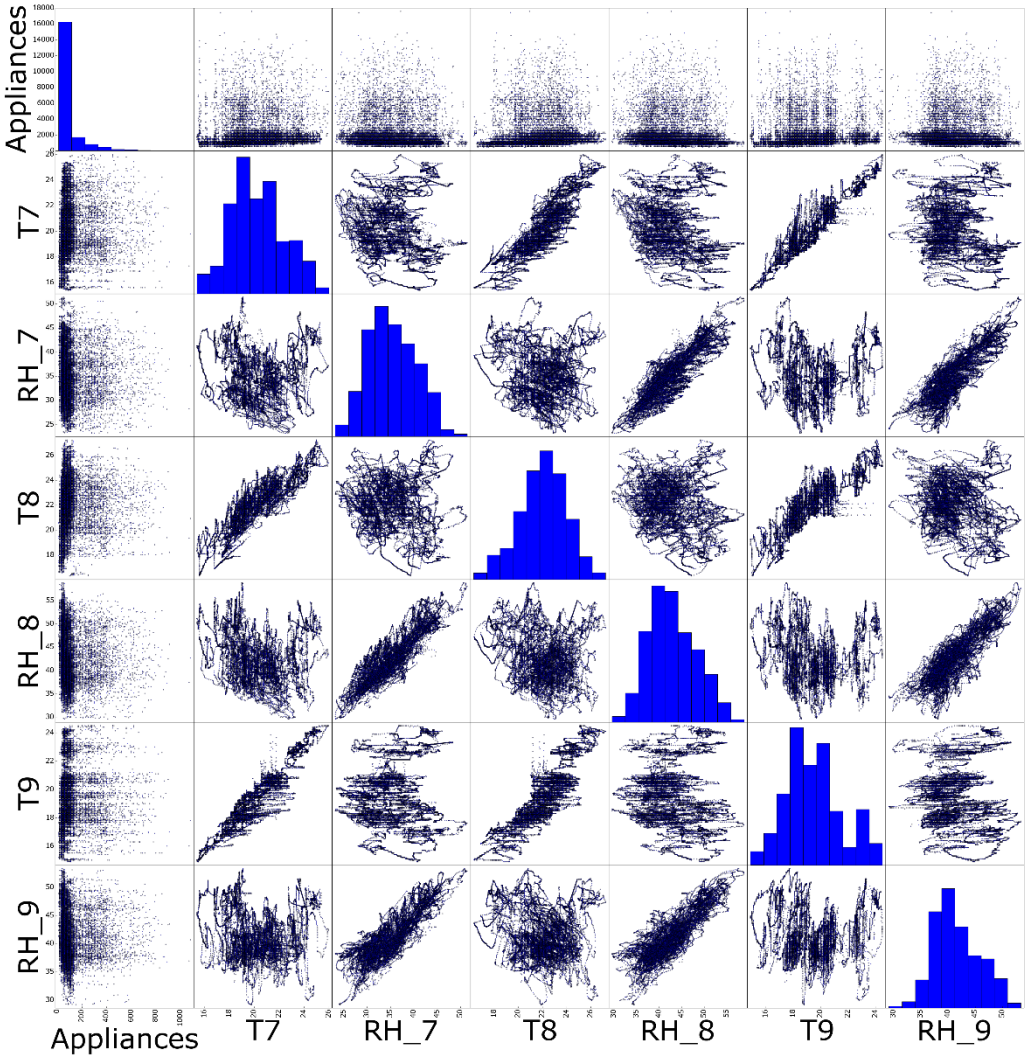
[4] Wang, Zeyu. A Review of Artificial Intelligence Based Building Energy Use Prediction: Contrasting the Capabilities of Single and Ensemble Prediction Models. Renewable and Sustainable Energy Reviews, Pergamon, 10 Nov. 2016.



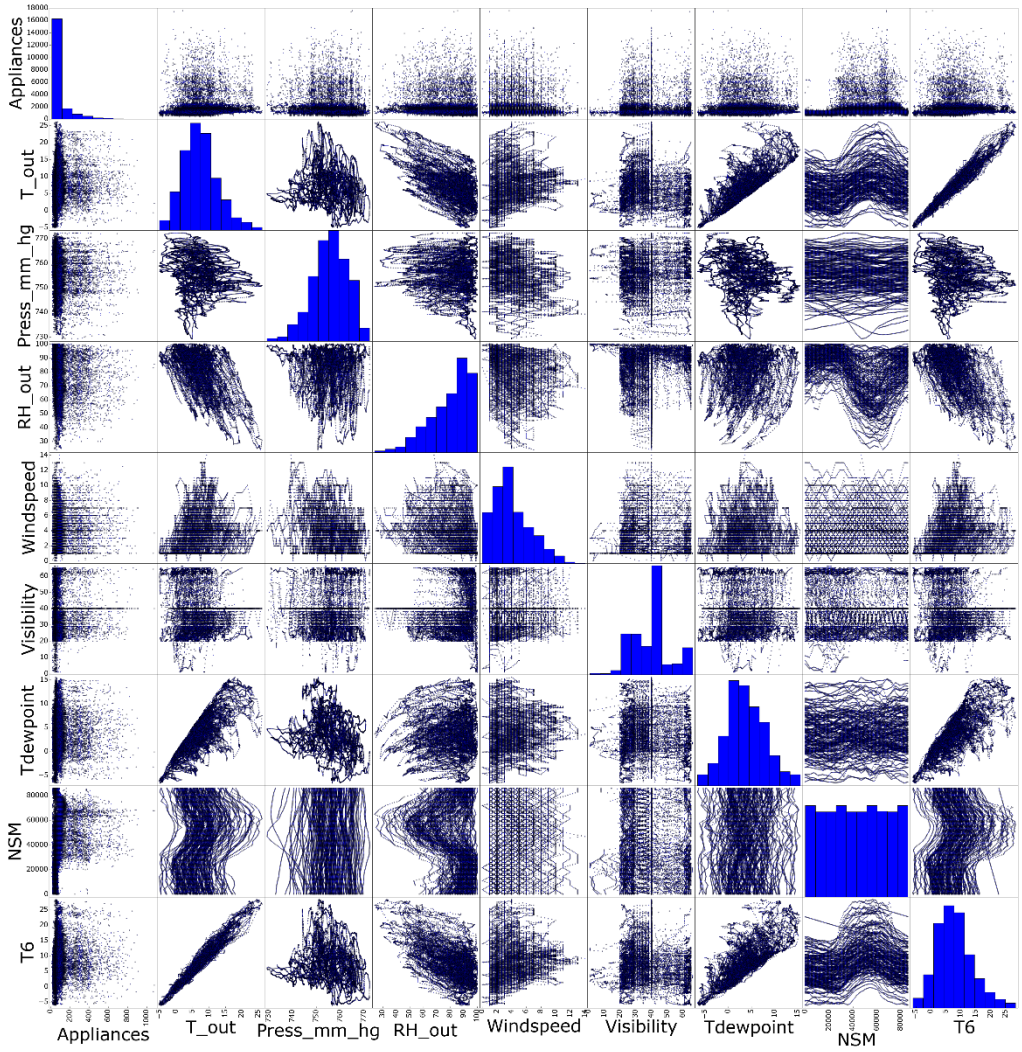
Appendix A Correlation (Appliances, Light, T1, RH1, T2, RH_2, T3, RH_3)



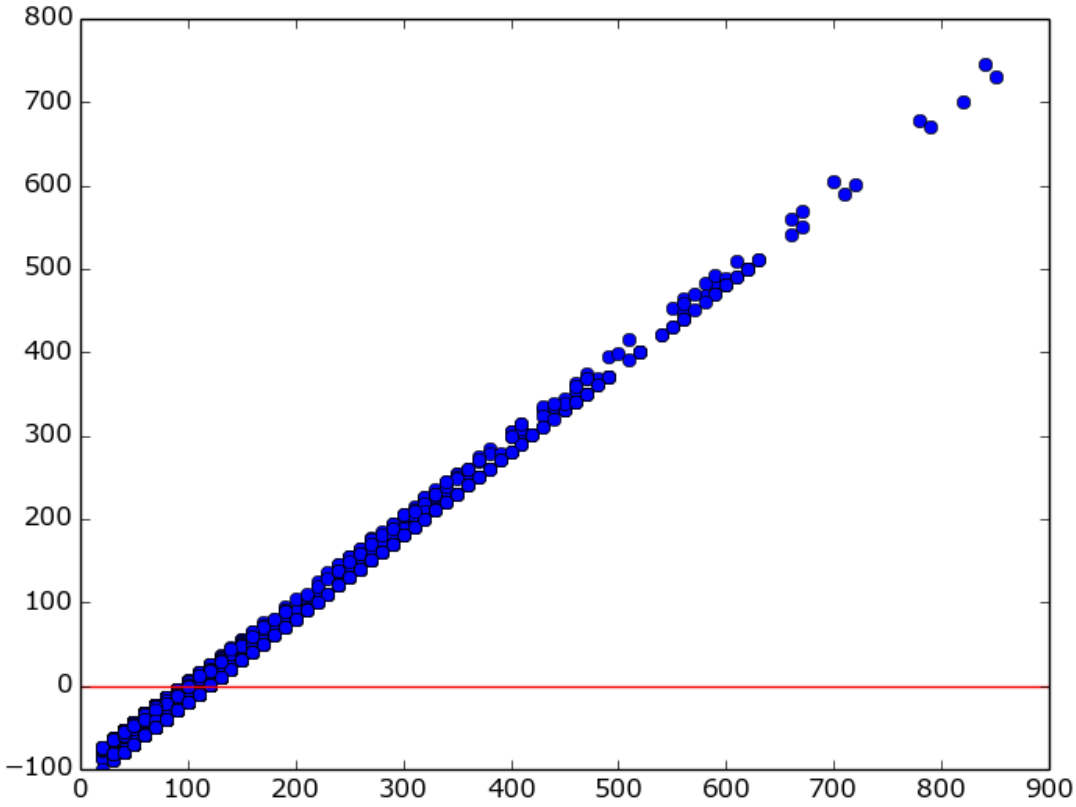
Appendix B Correlation (Appliances, T4, RH4, T5, RH_5, T6, RH_6)



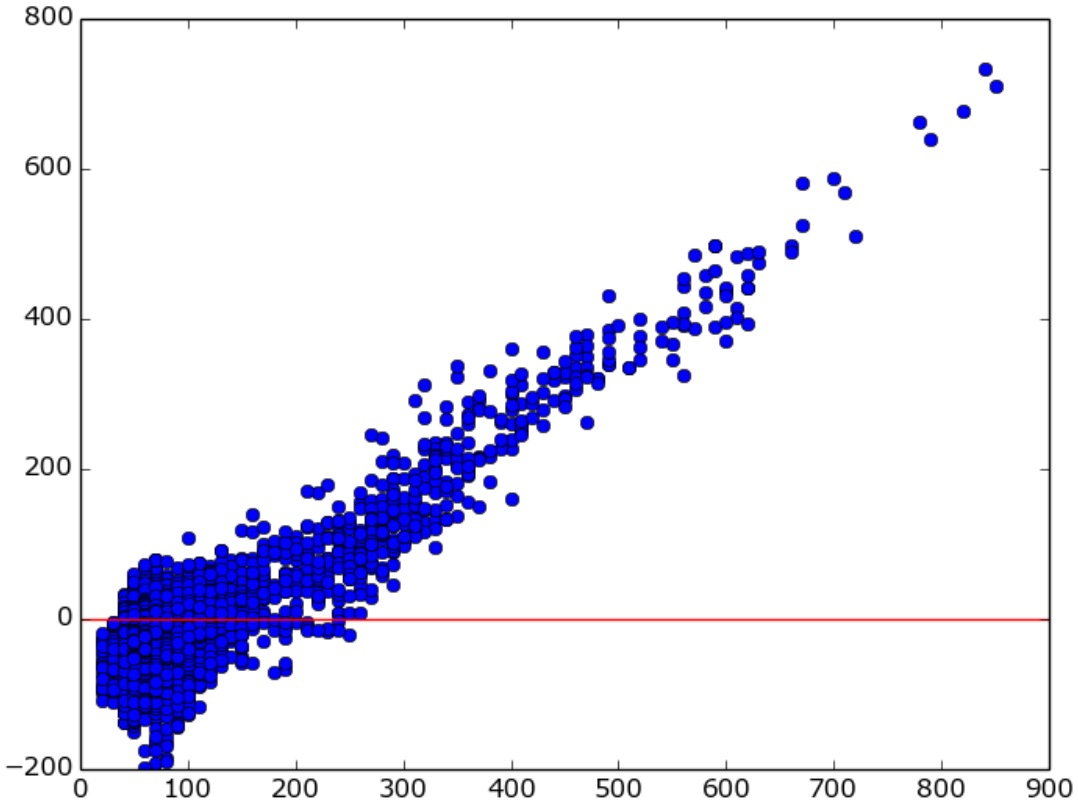
Appendix C Correlation (Appliances, T7, RH7, T8, RH_8, T9, RH_9)



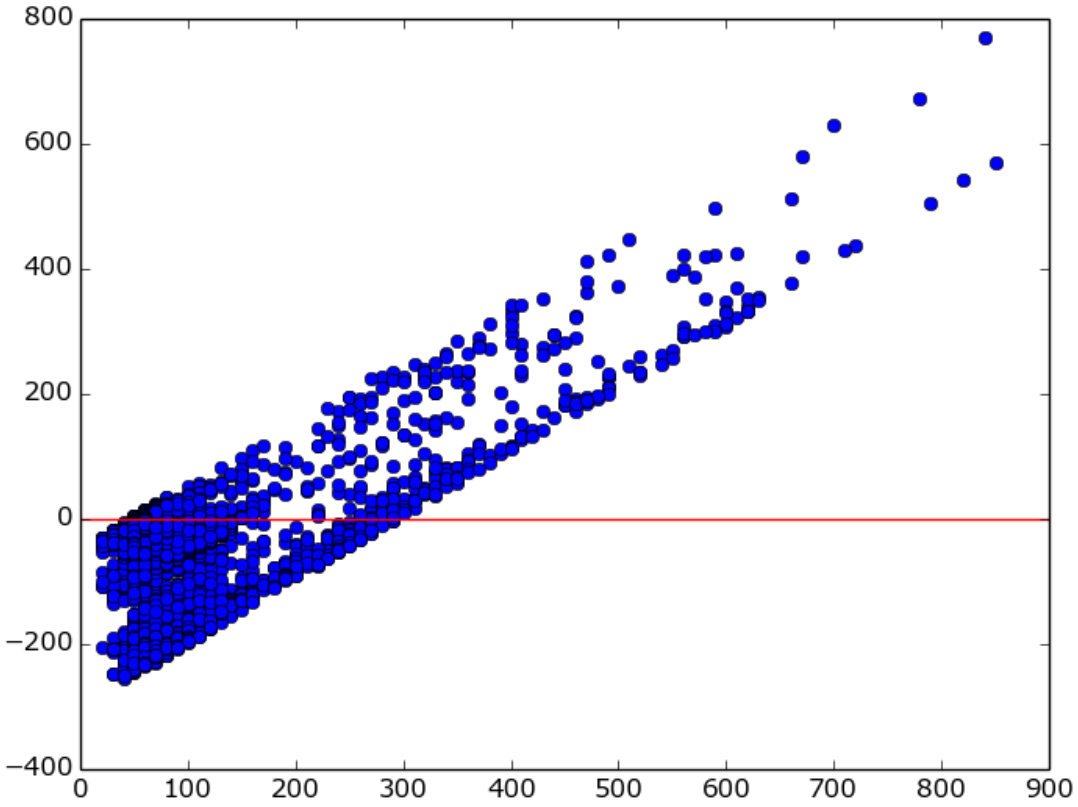
Appendix D Correlation (Appliances, T_out, Press_mm_hg, RH_out, Windspeed, Visibility, Tdwpoint, NSM, T6)



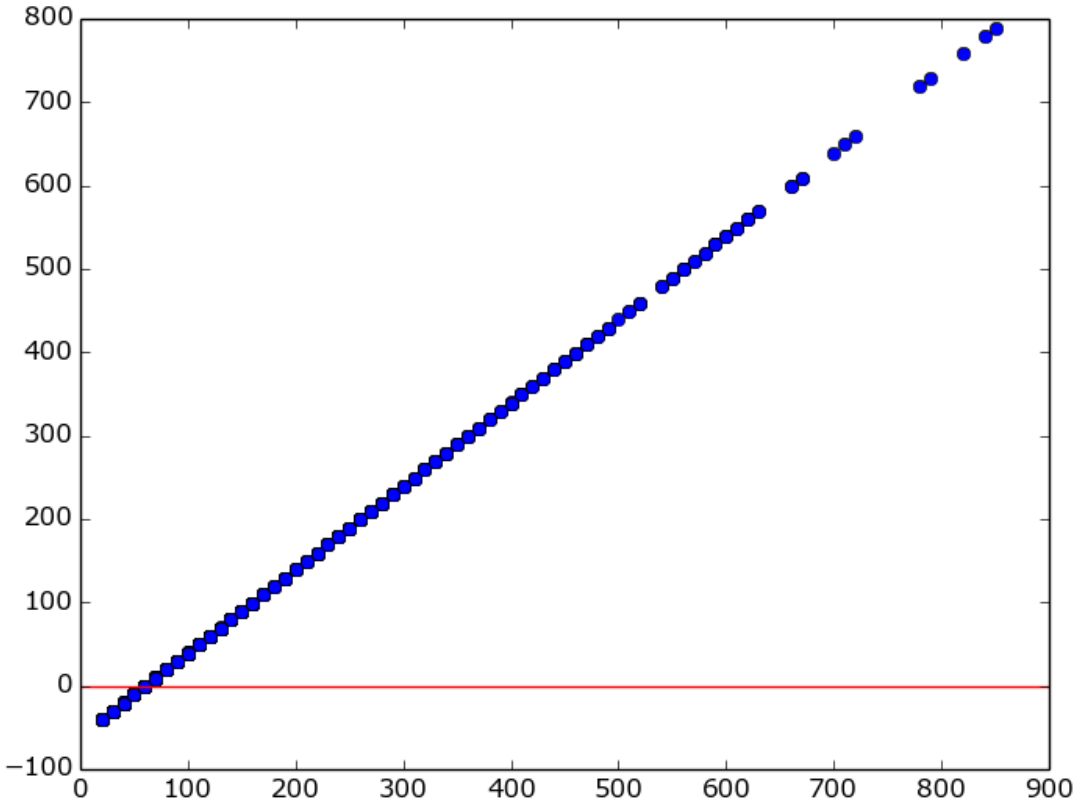
Appendix E Gradient Boosting Regression Residuals



Appendix F Multiple Linear Regression Residuals



Appendix G Random Forest Regression Residuals



Appendix H SVM Radial Regression Residuals