ELENE6690

Final Project Report

Data Driven Prediction Models of Energy Use of Appliances

Instructor: Prof.Predrag Jelenkovic
Prof.Aurel Lazar

Petar Barac pb2700| Cheng Sun cs3612 | Gan Jin gj2297

Abstract

This paper reproduces the results in the paper *Data Driven Prediction Models of Energy Use of Appliances in a low energy house* by Luis M. Candanedo. In addition, some other techniques were used to explore the data in other ways. PCA and PCR were used to reduce the dimension of the original data, making it faster to generate models. Neural networks are implemented to try to get a higher accuracy.

Keywords: energy consumption, multiple linear regression, support vector machine, gradient boosting machine

1 Introduction

1.1 Motivation

Since the appliance energy consumption is an significant part of the energy demand, there has been a lot research on in this area. Regression models can provide us with an insight about the relationship between appliance energy consumption and different variables. A thorough understanding about this relationship can be applied in several scenarios: detecting abnormal energy use, predicting future energy demand and being a part of the energy management system.

1.2 Challenges

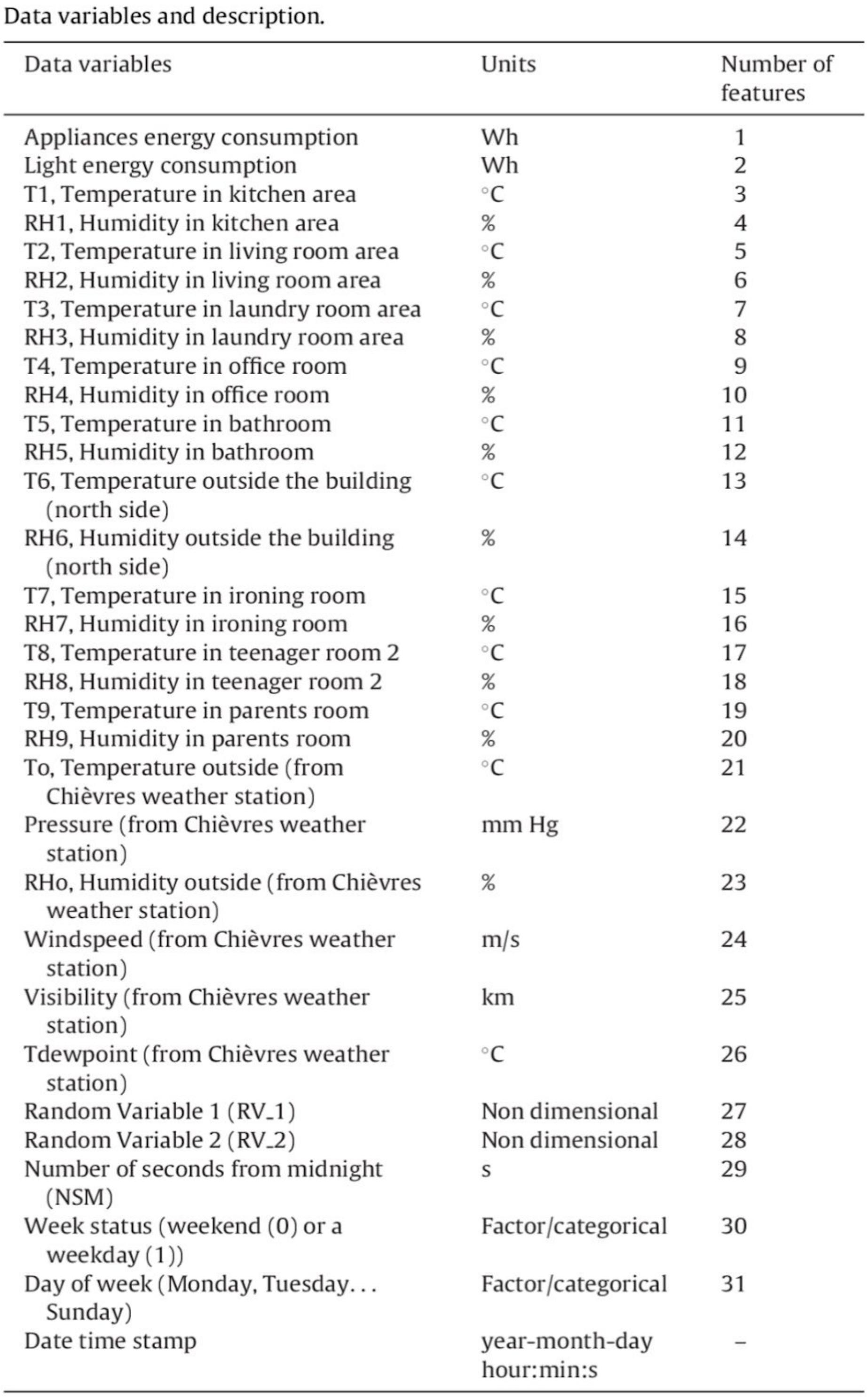
With the increasing number of appliances in the house, it is difficult to find out the main contribution to the energy consumption. Different appliances may have different energy use pattern. It is desirable to determine if the weather parameters help improve the prediction.

2 Data set and paper

2.1 Data set

2.1.1 Data set description

The data set records the energy consumption of appliances in a house along with the corresponding environment, weather and time statistics in a time span of 137 days. There are a total of 19735 pieces of data and 32 variables in the data set. A description of each variable is presented in Table 2.1.1.



Data variables and description.

Data variables	Units	Number of features
Appliances energy consumption	Wh	1
Light energy consumption	Wh	2
T1, Temperature in kitchen area	°C	3
RH1, Humidity in kitchen area	%	4
T2, Temperature in living room area	°C	5
RH2, Humidity in living room area	%	6
T3, Temperature in laundry room area	°C	7
RH3, Humidity in laundry room area	%	8
T4, Temperature in office room	°C	9
RH4, Humidity in office room	%	10
T5, Temperature in bathroom	°C	11
RH5, Humidity in bathroom	%	12
T6, Temperature outside the building (north side)	°C	13
RH6, Humidity outside the building (north side)	%	14
T7, Temperature in ironing room	°C	15
RH7, Humidity in ironing room	%	16
T8, Temperature in teenager room 2	°C	17
RH8, Humidity in teenager room 2	%	18
T9, Temperature in parents room	°C	19
RH9, Humidity in parents room	%	20
To, Temperature outside (from Chièvres weather station)	°C	21
Pressure (from Chièvres weather station)	mm Hg	22
RHo, Humidity outside (from Chièvres weather station)	%	23
Windspeed (from Chièvres weather station)	m/s	24
Visibility (from Chièvres weather station)	km	25
Tdewpoint (from Chièvres weather station)	°C	26
Random Variable 1 (RV_1)	Non dimensional	27
Random Variable 2 (RV_2)	Non dimensional	28
Number of seconds from midnight (NSM)	s	29
Week status (weekend (0) or a weekday (1))	Factor/categorical	30
Day of week (Monday, Tuesday... Sunday)	Factor/categorical	31
Date time stamp	year-month-day hour:min:s	–

Table 2.1.1

The energy consumption of appliances is recorded every 10 minutes in order to capture the quickly changing data. The energy consumption of lights and relative humidity recording are also used to predict whether a room is occupied. To study the impact of the weather condition, data from the nearest weather station is included in the data set.

The data set is split into two parts: the training data and the testing data. 75% of the data is used for training models and the rest is used for testing.

2.1.2 Data features filtering and importance

Considering the weather data is not exactly metered in the house and there are several variables in the data set, we want to choose the most relevant variables to do data training. The Boruta package is used to evaluate the variable importance. We can observe from figure 2.1.1 that NSM is the most important variable among the 32 variables.

Although the Boruta package provides us with the sense of the variable importance, we still want to know how many variables are suitable to make the prediction. The recursive feature elimination (RFE) is used to choose an appropriate number of variables to perform the prediction. The random forest method is used with 10-fold cross validation. From figure 2.1.2, we can observe that the optimal number of variables is 31. We also observe form the result that there is not much difference between 31 and 32 variables, so we select all the variables to do the prediction.

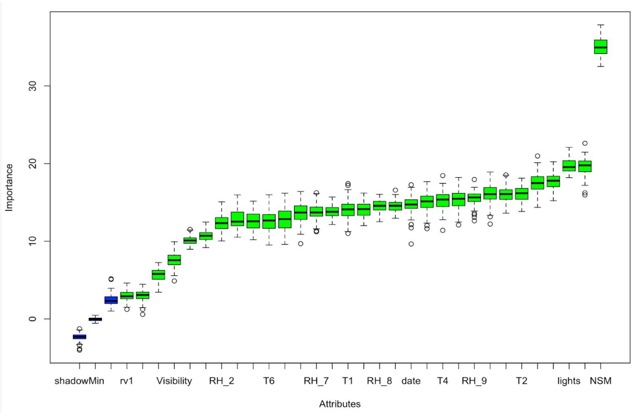


Figure 2.1.1

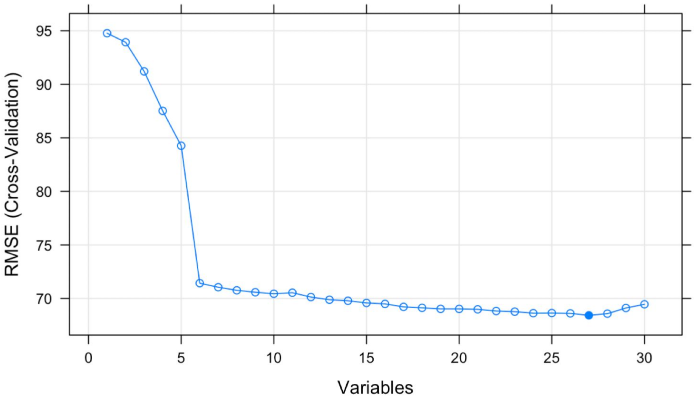


Figure 2.1.2

2.2.2 Paper

The paper explores the relationship between appliances energy consumption and different predictors.

Four regression models (linear regression, SVM-radial, random forest and gradient boosting machine) are used in this paper. The importance of predictors are ranked in each model. After building all the models, the evaluation of trained models in test sets is shown. In the end, the paper makes a comparison and conclusion.

In our report, we begin with the introduction of the application area. Then, we describe the data set in detail and briefly describe the work done in the paper. According the model used in the paper, we reproduce the result. Moreover, we choose some other methods that are not mentioned in the paper, including principal component regression and a neural network. The report ends with the comparison between different models and future work.

3 Results reproduction

3.1 Multiple linear regression

3.1.1 Principle

Multiple linear regression is widely used to explain the relationship between different features. We will try this method to see how variables in the dataset are correlated with each other.

The model for a multiple linear regression model that relates p-1 x variables and y is:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$

The following assumptions were used: (1) The regression residuals ε needs to be normally distributed. (2) The residuals are homoscedastic and approximately rectangular-shaped. (3) The model is linear in the parameters β. (4) Multicollinearity may happen in this model, and independent variables are not highly correlated.[8]

3.1.3 Original Model Selection

Table 3.1.2 shows the preference of multiple linear regression model from the original paper.

LR	RMSE	R^2	MAE	MAPE(%)
Training	93.21	0.18	53.13	61.32
Testing	93.18	0.16	51.97	59.93

Table 3.1.2

3.1.3 Model Reproduction

LR	RMSE	R^2	MAE	MAPE(%)
Training	92.31	0.17	52.77	61.13
Testing	96.17	0.16	52.92	61.14

Table 3.1.3

We can observe from the importance graph that the temperature in the laundry room, humidity in the kitchen area and light energy consumption are the most important factors of energy consumption of appliances.

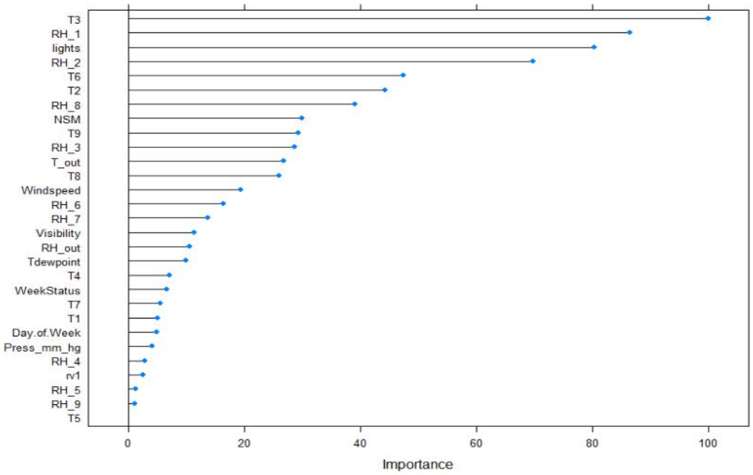


Figure 3.1.3

3.2 Support vector machine

3.2.1 Principle

The radial basis function kernel is a kernel that is in the form of a radial basis function, which takes the form

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

Radial kernel has very local behavior, in the sense that only nearby training data has an effect on the class label of a test data. Two tuning parameters, sigma and cost variable, are needed in addition to the predictors. We obtain the optimal parameters by a grid research. The result shows that when sigma is 0.35 and cost variable is 3, we can achieve the optimal performance.

3.2.1 Original Model Selection

Table 3.2.1 shows the preference of SVM-radial model in the training and testing data. The top five important variables are ranked starting with NSM, lights, RH_out, RH_6 and T6.

SVM-radial	RMSE	R^2	MAE	MAPE(%)
Training	39.35	0.85	15.08	15.60

Testing	70.74	0.52	31.36	29.76

Table 3.2.1

3.2.3 Reproduction of Model

The model was trained with 10-fold cross validation with 3 repetitions. The performance of the reproduction is shown in table 3.3.2 and the importance of variables is shown in figure 3.2.1. It is obvious that SVM-radial has a better performance than multiple linear regression.

SVM-radial	RMSE	R^2	MAE	MAPE(%)
Training	48.98	0.76	17.85	16.45
Testing	82.82	0.36	37.92	35.82

Table 3.2.2

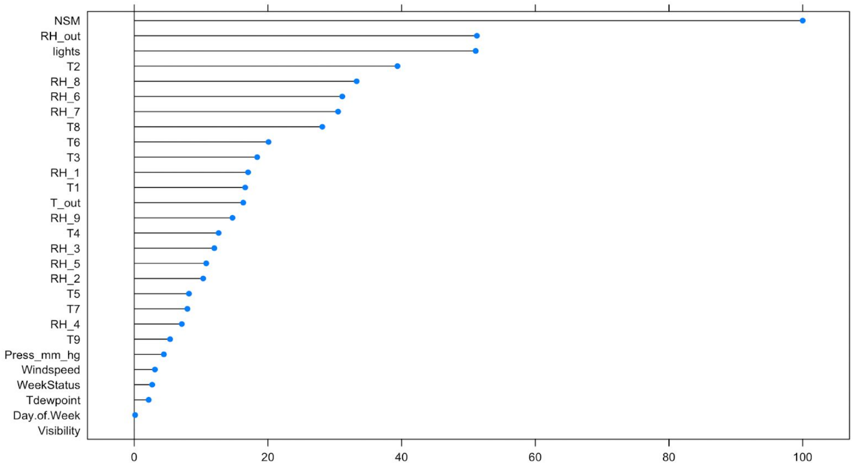


Figure 3.2.1

It can be observed that the top three important variables are identical with the paper, which are NSM, RH_out and lights.

3.3 Random forest

3.3.1 Principle

Random forest is a supervised machine learning algorithm can be used for both classification and regression problems. Random forest is very similar to an ordinary decision tree or bagging classifier. Unlike decision trees, the random forest grows the tree randomly. In the bagging method, the procedure consists of randomly sampling subsets for the training data and training each tree. The extra randomness of Random Forest is achieved by randomly selecting the subset of the training data and randomly selecting the subset of the features. Random forests do not suffer from overfitting like deep decision trees. Computationally, random forests can be quite intensive but easy and fast to train but slow to make the prediction.

3.3.2 Original Model Selection

In the Candanedo paper, three metrics were chosen to compare the performance of the models: root mean squared error (RMSE), R-squared (R^2) , mean absolute error (MAE) , and mean absolute error percentage (MAPE). In section section 2.1.2, it was discussed that the data set was analyzed with the recursive feature elimination (RFE) algorithm. This method used the random forest regression method and found that all features were relevant to minimize the RMSE. The random forest method was trained with all these features.

The models used in the paper were trained and generated with the CARET package. The model was trained with 10-fold cross validation with 3 repetitions. The performance of the model is measured by the lowest root mean squared error (RMSE) and highest R^2 . The results from the Candanedo paper are outlined in Table 3.3.1.

The random forest model in the original paper had the following 10 variables as the most important (in order of most important to least): NSM, lights, Press_mg_hg , RH_5, T3, Tdewpoint, RH_3, T5, Windspeed, RH_7.

RF	RMSE	R^2	MAE	MAPE(%)
Training	29.61	0.92	13.75	13.43
Testing	68.48	0.54	31.85	31.39

Table 3.3.1

3.3.3 Reproduction of Model

Using the same package and method from the Candanedo paper, a random forest model was trained using a smaller subset of the data. The results of the reproduced model are outlined in Table 3.3.2 and a graph of the important features in order of most important to least important is outlined in Figure 3.3.1.

RF	RMSE	R^2	MAE	MAPE(%)
Training	32.83	0.9234	16.063	16.36
Testing	81.52	0.3839	40.3	40.00

Table 3.3.2

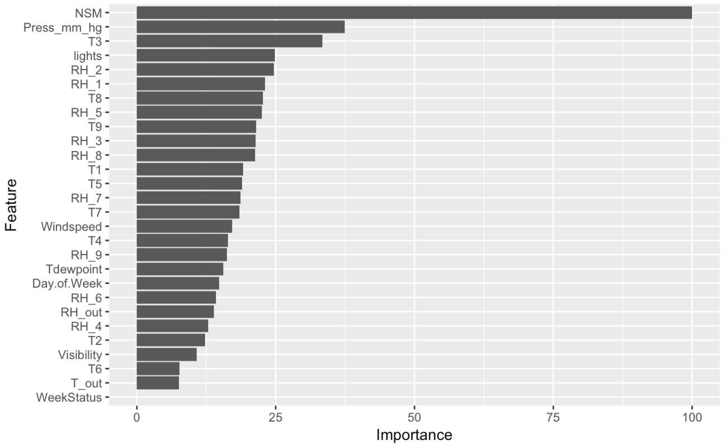


Figure 3.3.1

3.3.4 Discussion of Random Forest Results

The results of the reproduced testing results do not differ significantly from the those of the training model. The results of the testing results differ more significantly from the original paper. But the reproduced results still performed with reasonable improvement over the LM method and SVM method which was observed and consistent with results in the Candanedo work.

The most important takeaway from the reproducing the random forest is the important variables. From the model produced in this work, the top ten features (in order of importance) were : NSM, Press_mg_hg, T3, lights, RH_2, RH_1, T8,RH_5, T9,RH_3. This model was able to capture the most important features of NSM, Press_mg_hg, and lights.

3.4 Gradient boosting machine

3.4.1 Principle

The gradient boosting is the final model used to model the appliance energy output of the house. This method is capable of doing regression, classification and ranking. In the reproduction of our results it was used as a regression model to predict the energy output based on a set of features and also was used to rank the importance of these features. The GBM method is where simple models, in this case regression trees, are added together to create a more powerful model. Each weak learner is added to improve the larger model and minimize a metric. In this case, we are trying to minimize the root mean square error.

3.4.2 Original Model Selection

The GBM models were trained using the same method as the random forests. The models were generated with the CARET package. The model were trained with 10-fold cross validation with 3 repetitions. The performance of the model is measured by the lowest root mean squared error (RMSE) and highest R^2 . The data from the original is outlined in Table 3.4.1

GBM- All Features	RMSE	R^2	MAE	MAPE(%)
Training	17.56	0.97	11.97	16.27
Testing	66.65	0.57	35.22	38.29

Table 3.4.1

3.4.3 Reproduction of Model

Table 3.4.2 outlines the reproduced data used in this work. The GBM model's ranking of important features from this generated model are found in Figure 3.4.1.

GBM - All Features	RMSE	R^2	MAE	MAPE(%)
Training	5.77	0.9967	4.36	6.55
Testing	85.13	0.3412	45.26	48.33

Table 3.4.1

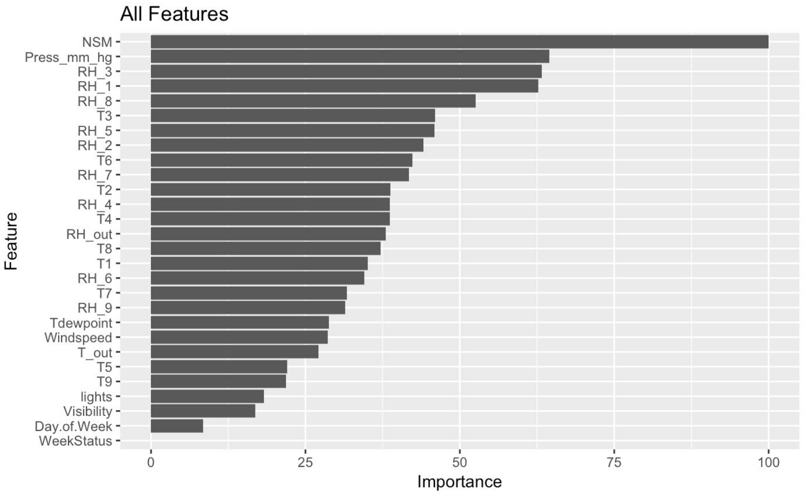


Figure 3.4.1

3.3.4 GBM with Different Data Subsets

Since the GBM performed well it was selected to assess the performance of different unique subsets of the data. The first data set contained all the same features with the exception of the light. This was done because the light variable can be positively correlated with other variables. Another data subset removed the lights and all weather features. The third data set only excluded all temperature and humidity data in the house. The final subset of data only contained weather and time information. The performance of these models is shown in Table 3.4.2.

GBM-Data Subset Testing (Data Driven)	RMSE	R^2	MAE	MAPE (%)
No Lights	66.21	0.58	35.24	38.65
No LIghts and No Weather	68.59	0.54	36.21	39.23
No Temp. and No Humidity	72.64	0.49	40.32	45.33
Only Weather and Time	72.45	0.49	40.73	46.53

GBM-Data Subset Training (Reproduced)	RMSE	R^2	MAE	MAPE (%)
No Lights	5.175	0.997	3.949	5.89
No LIghts and No Weather	4.63	0.9979	3.54	5.43
No Temp. and No Humidity	8.63	0.993	6.44	9.41
Only Weather and Time	14.31	0.98	10.31	14.37

GBM-Data Subset Testing (Data Driven)	RMSE	R^2	MAE	MAPE (%)
No Lights	66.21	0.58	35.24	38.65
No LIghts and No Weather	68.59	0.54	36.21	39.23
No Temp. and No Humidity	72.64	0.49	40.32	45.33
Only Weather and Time	72.45	0.49	40.73	46.53

GBM-Data Subset Testing (Reproduced)	RMSE	R^2	MAE	MAPE(%)
No Lights	84.95	0.3496	45.19	48.78
No LIghts and No Weather	86.52	0.3269	47.12	51.58
No Temp. and No Humidity	90.36	0.273	53.66	62.8
Only Weather and Time	88.95	0.299	51.43	59.12

Table 3.4.2

3.4.5 Discussion of GBM Results

In an interest to minimize computation time, the reproduced model was trained with 25% of the dataset. This led to the reproduced results performing slightly worse than the original work. In the context of the work presented here, the GBM performed very well and did significantly better than the linear models. Because the training data was smaller, the GBM performed marginally worse than the random forest and SVM models. It was expected that this model would outperform both models. This error is only marginal and it is still evident that the gradient boosting machine still is a powerful modeling tool and was able to predict accurate results.

The most important result reproduced in this work is the rankings of the most important features. In all the generated models, the time of day (NSM) ranked the most important feature. In applicable models, the pressure outside and room humidity were the most important features. This result agrees with what was observed in the original work.

All in all, it can be be seen that energy consumption of a house's appliances is heavily related to the time of data, outdoor pressure, and humidity and temperature of the rooms. The time of day is most likely explained by the fact that power consumption varies across the day with certain peak hours where more people are home and using electricity. The importance of pressure as a variable is probably best explained as when the weather outside forces people to spend their time inside. Low pressure corresponds to outside conditions that are more rainy and cloudy, where people would most likely spend their time in doors. High pressure corresponds to better weather, so the occupants are more likely spend less time in their home and therefore consume less energy. The final important variable is the temperature and humidity of individual rooms. This feature is most likely important because it corresponds to room occupancy. High humidity and temperature in rooms is attributed to human presence in the room. This increased presence means that more appliances are likely to be on when someone is in the room. This feature most likely ranks lower than the others because it only shows in which rooms the people are in and not their appliance output. For example, in a bedroom there might be a lot human presence but with low energy appliances running. While in a room with less presence might have more power hungry appliances.

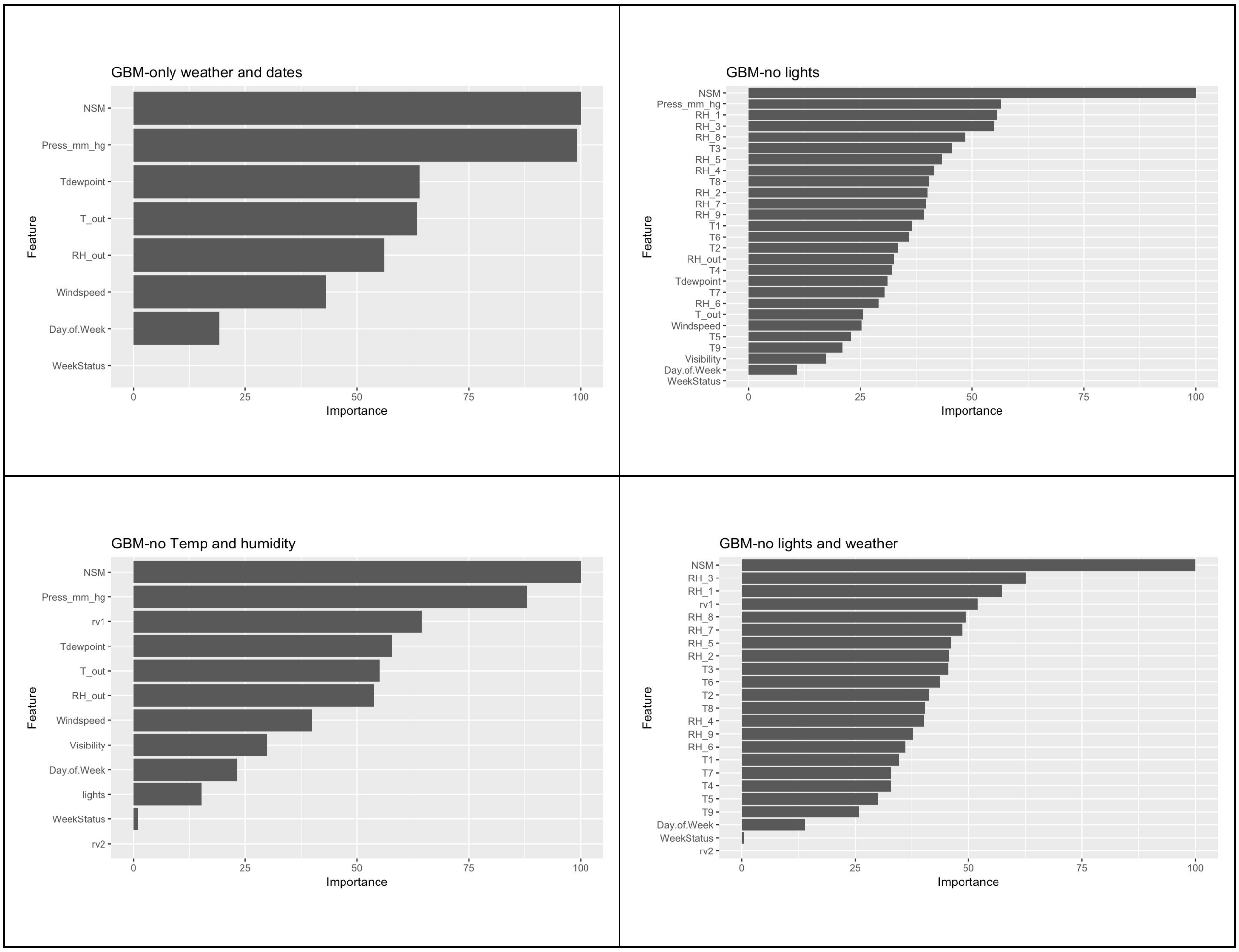


Table 3.4.3

4 Different Techniques

4.1 Principal component regression

It took ten hours or so to train the SVM and GBM models. So it comes to our mind that whether we can reduce the dimension of the data set and explain the data well at the same time. PCA and PCR seem to satisfy the demand.

4.1.1 Principle flow chart

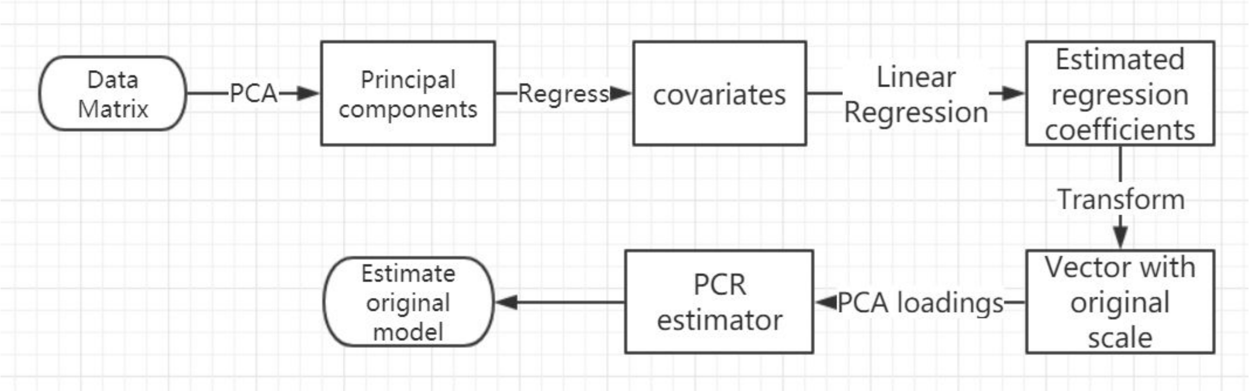
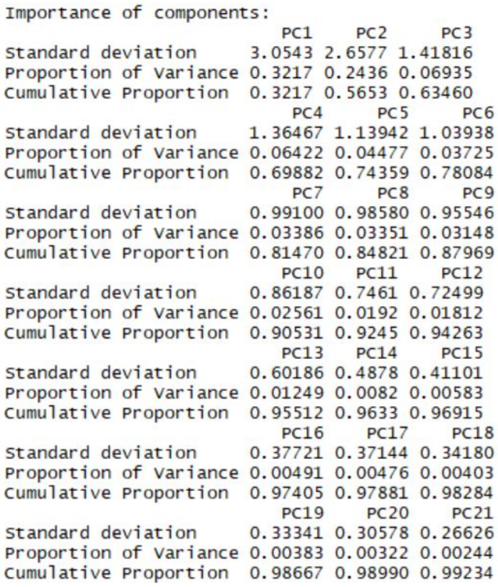


Figure 4.1.1[1]

4.1.2 Results



```
Importance of components:
                           PC1    PC2     PC3
Standard deviation     3.0543 2.6577 1.41816
Proportion of Variance 0.3217 0.2436 0.06935
Cumulative Proportion  0.3217 0.5653 0.63460
                           PC4     PC5     PC6
Standard deviation     1.36467 1.13942 1.03938
Proportion of Variance 0.06422 0.04477 0.03725
Cumulative Proportion  0.69882 0.74359 0.78084
                           PC7     PC8     PC9
Standard deviation     0.99100 0.98580 0.95546
Proportion of Variance 0.03386 0.03351 0.03148
Cumulative Proportion  0.81470 0.84821 0.87969
                          PC10   PC11    PC12
Standard deviation     0.86187 0.7461 0.72499
Proportion of Variance 0.02561 0.0192 0.01812
Cumulative Proportion  0.90531 0.9245 0.94263
                          PC13   PC14    PC15
Standard deviation     0.60186 0.4878 0.41101
Proportion of Variance 0.01249 0.0082 0.00583
Cumulative Proportion  0.95512 0.9633 0.96915
                          PC16    PC17    PC18
Standard deviation     0.37721 0.37144 0.34180
Proportion of Variance 0.00491 0.00476 0.00403
Cumulative Proportion  0.97405 0.97881 0.98284
                          PC19    PC20    PC21
Standard deviation     0.33341 0.30578 0.26626
Proportion of Variance 0.00383 0.00322 0.00244
Cumulative Proportion  0.98667 0.98990 0.99234
```

Figure 4.1.2

From all the 32 features, we need only 13 of the components to explain 95% of the variance. 21 components are needed to explain 99% of the variance.

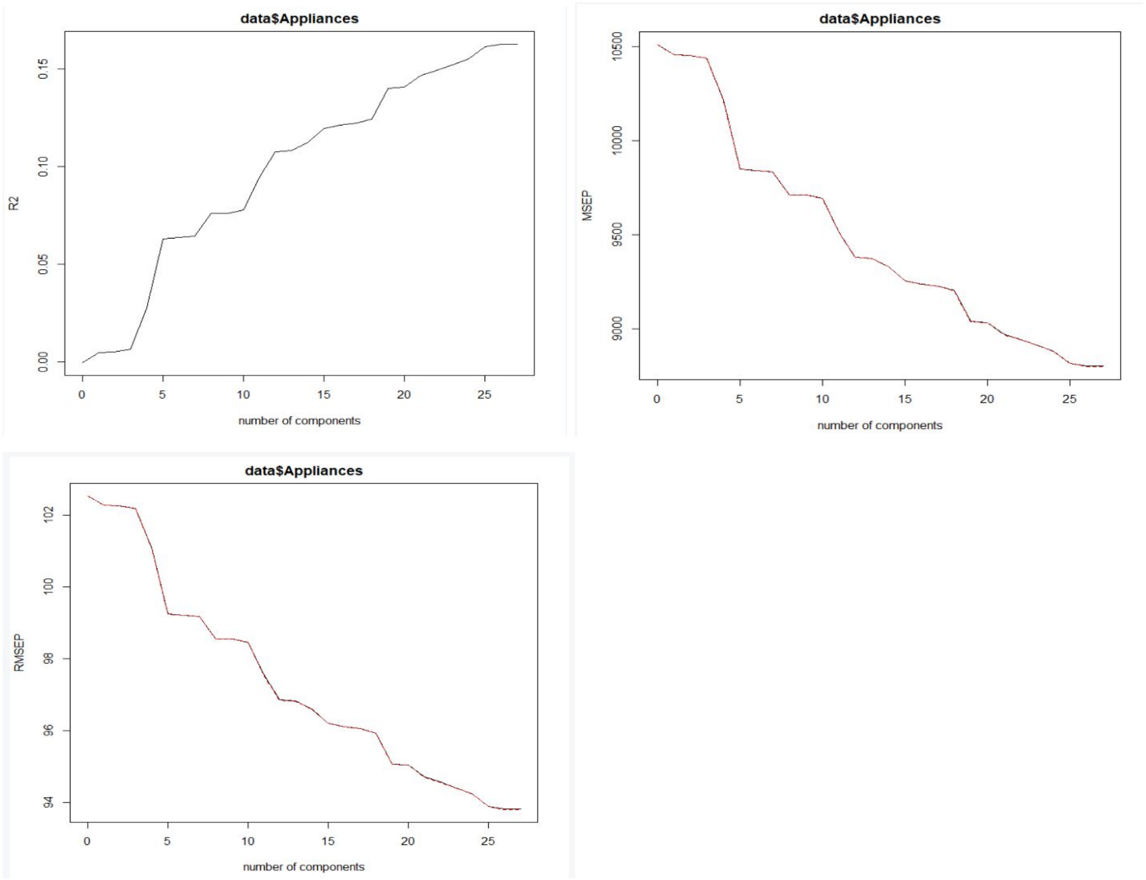


Figure 4.1.3

MSE, RMSE and R2 are ideal for PCR with 21 components. It is also acceptable for 13 components.

4.1.2 Discussion

There are some advantages to the PCR model. Dimension can be reduced effectively with a small amount of information lost. This method can avoid multicollinearity, which is a problem that we had to face in multiple linear regression. In addition, the overfitting problem is also mitigated.

However, there are also some drawbacks. Components are not the original features after PCR, so it is hard to explain the relationship between original features. Moreover, PCR can only deal with numerical data. To solve this problem, we used numbers from 1 to 7 instead of the “weekday” function in the R language.

4.2 Neural networks

4.2.1 Principle

In other papers and research neural networks were used to predict the electrical output given a set of features and data sets [7]. Neural networks were chosen because they might be able to provide some more powerful prediction models.

Neural networks are used in a deep learning method where the prediction models have very high performance but are nearly impossible to interpret. A neural network consists of the input layer where our features are the inputs, one to several hidden layers, and an output layer. Within the hidden layers are weighted neurons that take the information from the inputs and is passed to a summation and pass them to another hidden layer or directly to the output layer. The blue circles represent a bias applied to the weights that allow for shifts in the prediction model to better fit the model to the training data. As the model trains the weights in the layers are changed in an manner that reduces the error in prediction and makes the model more powerful.

In our work, a 13 input layer with the 13 most important features extracted from the previous work were used to fit a neural with two hidden layers with 5 and 2 weights respectively. This number of weights was chosen because it led to the best performance and the training was able to converge to a model while other layers and numbers of weights failed to converge into a working model. With a model of 21 components, the problem of multicollinearity becomes severe and it takes much longer to train the neural networks.

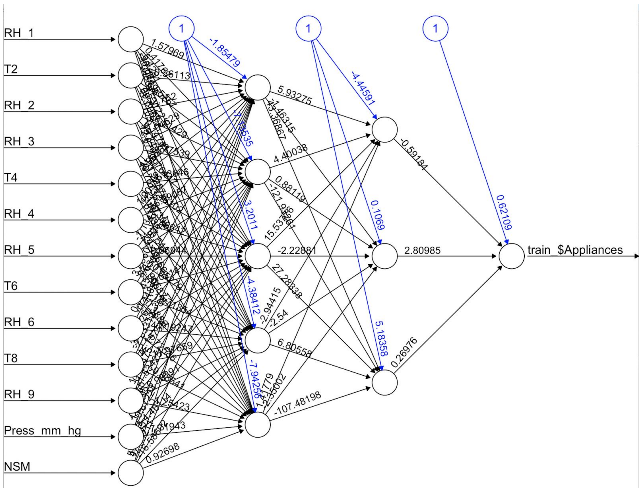


Figure 4.2.1

4.2.2 Results

Table 4.2.1 contains information about the training and testing results of the generated neural network when around 50% of the data set was used for training.

Neural Network Model	Training	Testing
RMSE	86.24	92.68
R^2	0.2864	0.18
MAE	47.14	49.133
MAPE%	43.28%	45.49%

Table 4.2.1

4.2.3 Discussion

The neural network in this work was able to achieve the accuracy of the linear regression models in the original work that were trained with a larger sample size. While there was no drastic improvement in model prediction, it can be seen that neural networks can be powerful tools in predicting appliance energy usage. In our work the input layer was chosen because it contained the most important features observed in other models and because each feature was not heavily linearly related to one another.

While there was not a great improvement in model performance with the neural network, there are definite benefits using this model. One benefit is that if this model were allowed to train on a larger data set and given more computing power, it is extremely an powerful tool in regression and classification. GIven more time and data, a model could be possibility be formed with all the features that outperforms the other methods.

This method is quite powerful but it does have many drawbacks. One issue that might have been present in our work is that some of the features might be more linearly related then we realized and this would hurt the performance of our neural network. This can be possibly be fixed by preprocessing the data using principal component analysis to remove the multicollinearity in the input features. Probably the biggest draw back in this model is that it more or less acts like a black box and while it's performance can be very powerful it is not intuitive and doesn't tell us about which features are important or why. This is troubling for this work because prediction is only half the goal, the other goal is to see which features contribute more to power consumption. In the context of this work, neural networks might accurately predict the power consumption but can't tell us how certain features influence the appliance power consumption. This does not allow us to learn how to minimize our ever growing electricity consumption. This model specifically identifies the main contributors and points of improvement. Maybe there are other ways to train neural networks so they can help identify the major sources of wasteful energy use in our homes.

4.3 Observations with different appliances

We separate the data into two groups, one with energy consumption of appliances below 100, one with energy consumption of appliances 100 or higher. We ran multiple linear regression model separately on the two groups of data and do some observations.

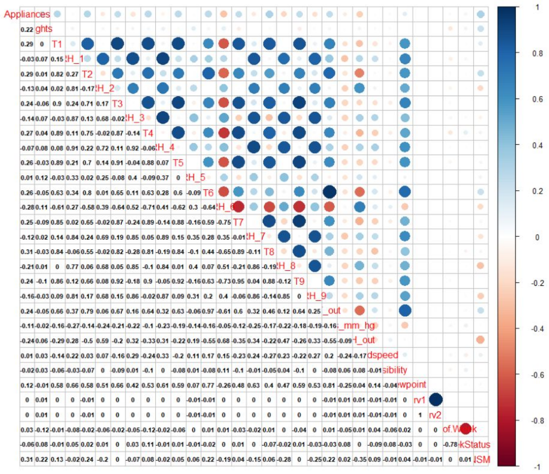


Figure 4.3.1. energy consumption below 100

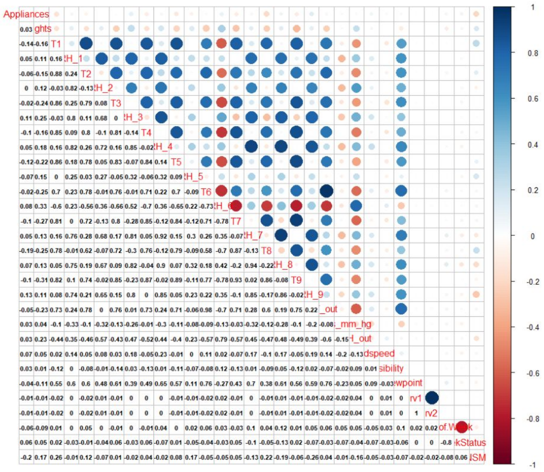


Figure 4.3.2. energy consumption over 100

We can see from the first row of the two pictures that energy consumption in lower energy consumption houses have a closer relationship with other factors. On the other hand, in the houses with higher energy consumption, the energy consumption is not greatly affected by other features.

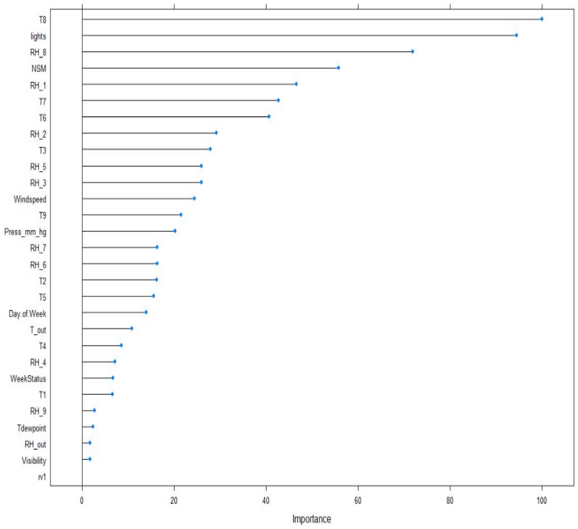


Figure 4.3.3. Importance with consumption below 100

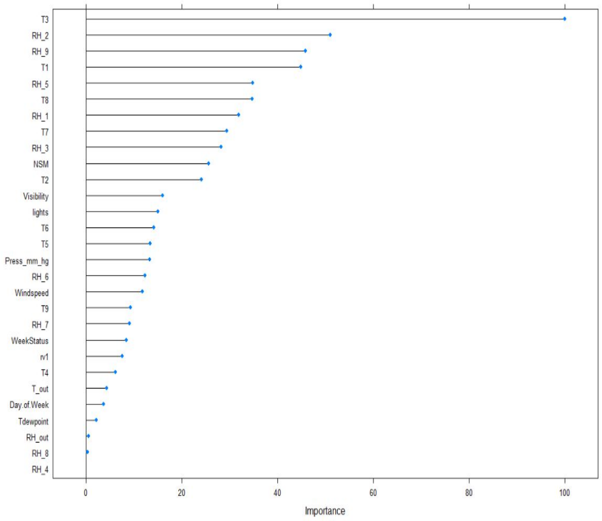


Figure 4.3.4. Important features with energy consumption over 100

5 Conclusion

5.1 Models efficiency

GBM has the best model performance among LM, SVM, GBM, NN and RF. GBM and SVM model require about 10 hours to be trained. So dimension reduction can be useful. PCA is a good way to reduce dimension and extract important features. But new components are harder to explain original relationship between features.

5.2 Observations

Times after midnight are the most important to predict appliances consumption. For houses with lower energy consumption, temperature and humidity in teenager rooms and lights contribute most to appliances consumption. For houses with higher energy consumption, temperature and humidity in the living room, temperature in the kitchen and parents room contribute most to energy consumption.

5.3 Future work

We did not reach a high accuracy with our GBM and SVM model. We will try to arrange the training partition to make the results closer to the paper. We also thought of using PCA and PCR to reach a higher training speed. PCR works well considering both speed and accuracy. But we did not apply PCA to other models, such as random forest, GBM and SVM. We will apply PCA to other models to increase their performance. One method that might be improved is the neural network. Preprocessing the data and training a neural network might reduce the dimensionality and the multicollinearity in the training data and that might lead to a better training model.

6 References

[1] https://en.wikipedia.org/wiki/Principal_component_regression

[2] https://www.r-bloggers.com/performing-principal-components-regression-pcr-in-r/

[3] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[4] https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/

[5]https://towardsdatascience.com/boosting-algorithm-gbm-97737c63daa3

[6] A. Kavousian. Ranking appliance energy efficiency in households: Utilizing smart meter data and energy efficiency frontiers to estimate and identify the determinants of appliance energy efficiency in residential buildings.Energy Build. 99 (2015) 220-230.

[7] S. H. Ling, F. H. F. Leung, H. K. Lam and P. K. S. Tam, "Short-term electric load forecasting based on a neural fuzzy network," in *IEEE Transactions on Industrial Electronics*, vol. 50, no. 6, pp. 1305-1316, Dec. 2003.

[8] https://en.wikipedia.org/wiki/Linear_regression

[9] Luis M. Candanedo. *Data Driven Prediction Models of Energy Use of Appliances in a low energy house.*