# Bosch Industrial 4.0
# Monitoring of large amount of Devices

Chinmay D Hegde

SPOTSeven Lab,
COLOGNE UNIVERSITY OF APPLIED SCIENCES

31-01-2019

***Abstract***– The Bosch company has been installing huge number of Boilers. Every system cannot be checked manually for its correctness. The primary goal of this case study is to detect the faulty systems (outliers) and hence finding the variables which are contributing to the faulty behaviour. In order to achieve this, necessary data pre-processing like data standardisation, analysis of correlation, treating the NAs, feature extraction has been carried out. Several algorithms and packages has been implemented for Data preprocessing and for variables analysis. Leaps package, Relative importance, LASSO algorithm, Random forest algorithms helped in determining the important variables in order to determine the energy consumed by the boiler, ggplot package was used to find the correlated variables. Cook's Distance was applied in order to find the outliers and hence the variables contributing for the faulty behaviour. The next step was finding faulty systems by applying influence plot which uses Cook's distance for outlier detection. As a result five systems were detected as faulty by the influence plot. These systems were analysed to find out the reason for its fault.

## 1 Introduction

After the installations of the boiler systems by the Bosch company, several sensors are employed to monitor the conditions of systems. Through sensors, values of certain variables are recorded each day, so that the trend is observed. This is achieved by the Internet Of Things application. Then the dataset is formed for all the systems. It becomes tedious and inefficient to manually check the sensors' correctness daily. This rises a need of developing a method to find the characteristics of a faulty system and a proper system. This serves the main purpose of this project. In the dataset the data variables forms the columns and each individual systems form separate rows. This data set is analysed and interpreted. There are different factors need to be considered here like weather conditions, seasons, class of boiler and some more. Different methods can be

employed in order to achieve the desired goal. In order to reach the target we go through Cross Industry Prcoess for Data Mining (CRISP-DM process). This forms a systematic approach. The process is divided into different sections.

Firstly all the experiments here are performed in the software R studio. Section 2 explains the variable description of the data set. Section 3 explains the steps involved in data standardisation, treating the NAs and correlated variables. Section 4 explains the finding of important variables in determining energy consumptions. Section 5 explains the finding of variables in detection of faulty systems. All sections are provided with necessary plots for data visualisation. Different packages and algorithms were used in order to achieve the goal.

## 2 Data Set Description

The given data set consists of sensor information of more than 25,000 installed devices which is divided into 209 Json files. Each Json files consists of information of 120 systems for one year. As this is huge data of about more than 10 GB, we have picked the first Json file and studied in depth to get the required results. Most of the variables of this data set is of numeric data type, very few are character and one date variable. The frequency of variable measurements are taken on daily basis. The proportion of missing values are different for each variable, this is discussed in detail in the later section. So here managing the big data is a challenging task and also to detect anomalies present in the system.

### 2.1 Variables

In order to derive the variables for detecting faulty systems, the first step is to get into the variables exploring and knowing the inner sights of the given parameters. Each categories have subcatogories which incldues variables related to their respective fields. The variables are divided into groups and each variable falls in any one of these categories:

1. Based on appliance identification
2. Based on age of appliances
3. Based on connectivity metadata
4. Based on internal set points
5. Based on power consumption
6. Based on scaled consumed energy
7. Based on energy consumed
8. Based on system physical characteristics
9. Based on working times
10. Based on burner starts
11. Based on keys
12. Based on outdoor temperatures

As the variables are very large in number, the variable description is done along as the paper proceeds. The total number of variables provided are about 50. Some analysis could help in variable reduction. Going through first json file following observations were made:

1. The variables related to outdoor temperatures like outdoor_temp_max, outdoor_temp_min and outdoor_temp_mean are unrealistic. As these parameters does not contribute to the goal ,this is removed from the dataset.
2. prim_t_set_max - internal set states is constant for each appliances.
3. nom_max_pow_ch_kw( maximum power for central heating) and nom_max_pow_dhw_kw-(maximum power for domestic hot water) are always constant through out the model.

So the three temperature variables are removed and the variable size is now reduced to 47.

## 3 Data PreProcessing

### 3.1 Pattern Recognition

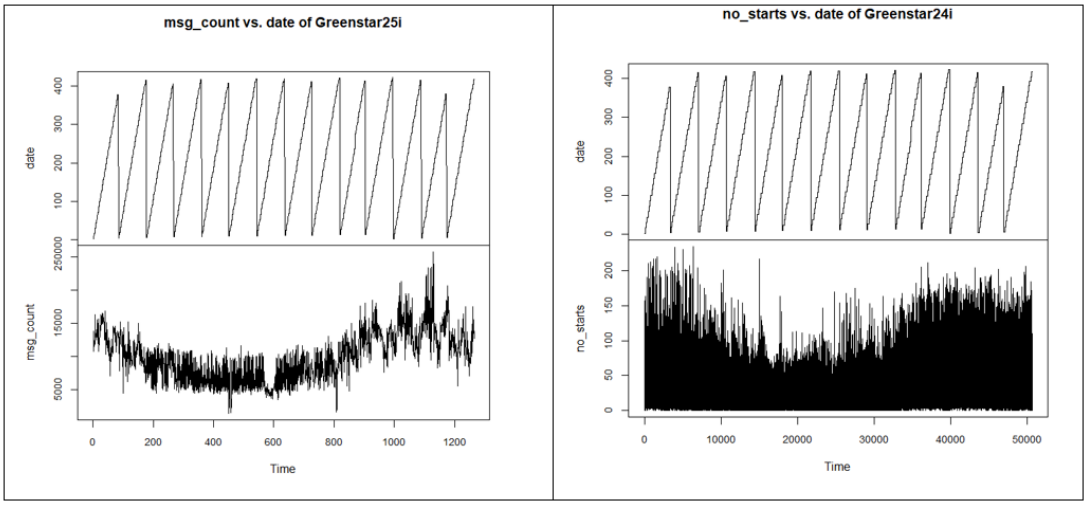


**Fig. 1.** Time Series shows between Greenstar25i and Greenstar24i system, these are plots between msg count vs date and burner starts vs date which shows the msg count and burner starts are less during summer season. The spikes in the date section corresponds to each month starting from February 2017 till March 2018

Here time series analysis is done to check the behaviour of different appliances and to get the impact of seasons on the system behaviour. Two different appliances Greenstar25i and Greenstar24i System are considered here. From the Fig.1 it can be observed that the trend or pattern between both devices. As we can observe from variable "Date", that during summer the number of burner starts and also message count is less where as during winter it is high. Hence

we can conclude that the dataset in summer season may be anomaly or outlier to winter season and dataset from winter can be anomaly to summer season. The seasons are named on basis of United Kingdom's(UK) yearly seasons as the dataset belongs to UK systems.

### 3.2 Scaling

As each variable in the data set is measured with different units and are bounded with different ranges, this data would not be a good feed to the training model. For this reason every variable is scaled between 0 and 1. This is done by the difference of the certain variable value and minimum value of that variable divided by the difference between maximum and minimum

$$z_i = (x_i - min(x))/(max(x) - min(x))$$

Now after the scaling every variable ranges between 0 and 1 and hence every variable is given equal importance in terms of an algorithm. This step also now decrease the variance of the variables[21].

### 3.3 Treating the NAs

NAs are the data that may be missing in the data set due to sensor failure or due to miscellaneous reasons. Some algorithms like random forest have the built in option to ignore NAs. Or NAs can be replaced by median of the particular variable, Or even sometimes by mean values. Working on missing values is one crucial task while doing data preprocessing. By treating NAs it is possible to reduce the bias of the model[20].

If any variables has many missing values, then that variable may not be such important. Sometimes these variables may even cause problem to its accuracy. In such cases dealing with those variables and removing them is better idea.

The function VIM aggr in R [1] calculates and represents the number of missing entries in each variable and for certain combinations of variables (which tend to be missing simultaneously). Here VIM package[22] was employed to work on missing values. The Fig.2 explains the VIM employment. The plot is useful in knowing the proportion of missing values in each variable.

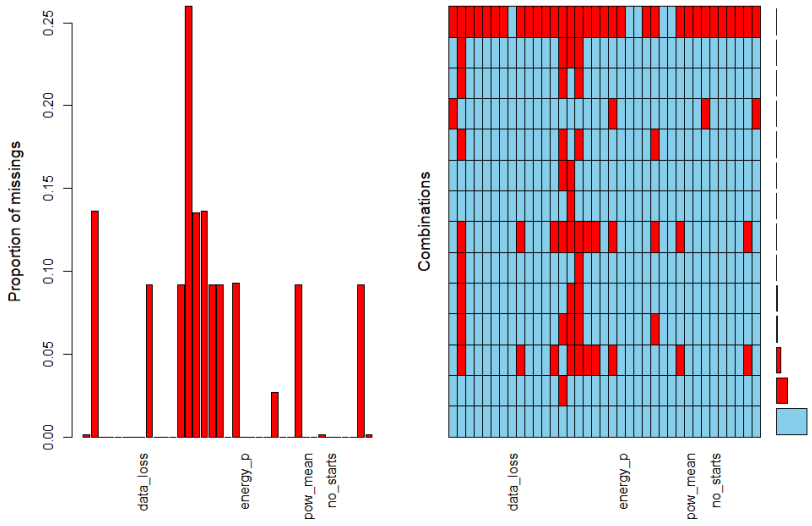


**Fig. 2.** In the Combinations plot from VIM package in R, the left plot shows the proportion of missing values of all variables in the order of the dataset given. In the plot on the right-hand side, the grid presents all combinations of missing (red) and observed (blue) values present in the data. Even though for our experiment the left plot is sufficient, the right plot gives information on the missing values of variables in combination.

#### Imputation with mean, median or mode

Amelia package in R also performs multiple imputation (generate imputed data sets) to deal with missing values[2]. Multiple imputation helps to reduce bias and increase efficiency. It is enabled with bootstrap based Expectaion Maximizing with Bootstrapping(EMB) algorithm which makes it faster and robust to impute many variables including cross sectional, time series data etc. Also, it is enabled with parallel imputation feature using multicore CPUs[13].
It makes the following assumptions:
All variables in a data set have Multivariate Normal Distribution (MVN). It uses means and covariances to summarize data. Missing data is random in nature (Missing at Random).

All the variables were imputed from amelia package. Hence amelia package was a simple and efficient step to treat NAs in this occasion,but only after converting the categorical values into numerical values during imputation.

### 3.4 Correlation of variables

It is important to know how the variables are correlated in a data set. This helps to categorize the correlated variables as a group. Correlation measures how strongly or weakly each variables are linearly related to each other. The

range of correlation lies between between -1 to +1. If the correlation is +1, then it is higly correlated or if it is -1, then it is negatively correlated. If the value approaches 0 ,the variables are weakly correlated. If it is 0, the variables are not correlated at all. The figure 3 is the correlation plot from ggplot in R package.

Merits of taking correlation into account are that it shows if there is relationship between different variables or not[3]. The demerit is that no cause and effect can be established in correlation as its not certain that one variable caused another to happen, it could be one or the other or it could even be an unknown variable that causes the correlation. Along with this if one variable is set as a target it is better to keep the correlated variables for its prediction, this would strengthen the model. As we don't know exactly which is the target variable, the elimination of correlated variables is neglected[14]. But it is crucial to know which variables are correlated, for better understanding of the model. Here, the plot in Fig. 3 describes correlations between each variables.

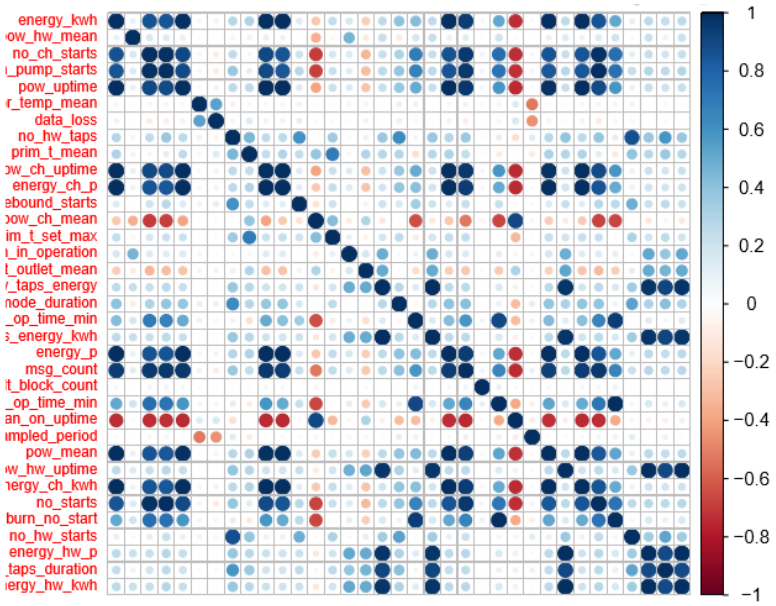


**Fig. 3.** The correlation plot in the form of square matrix represents the correlation between variables. The strength of correlation is measured for each variable with every other variable. The dark shaded colour indicates stronger correlation. It can be observed that there are considerable amount dark shaded box which means that there are considerable amount of correlated variables.

## 4 Important Variables Selection

In this section the variables which contributes to the energy_kwh output is drawn out. energy_kwh is chosen because of the reason that, energy_kwh variable shows

the energy consumed by the appliance, which is a very important measure to know the performance of a system. The important variables selection causes the reduction of variables in the data set. There are some advantages of using this technique[4]:

1. It can reduce the training times of the model
2. It can prevent overfitting increasing generalizability
3. It removes unwanted or redundant variables, which reduces data size.

Here few packages are used to perform this action, and then the results are combined to pick the common important variables.

### 4.1 Random Forest

"Random forests or random decision forests are an ensemble learning method for classification,regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees[18]. Random decision forests correct for decision trees' habit of overfitting to their training[19]"

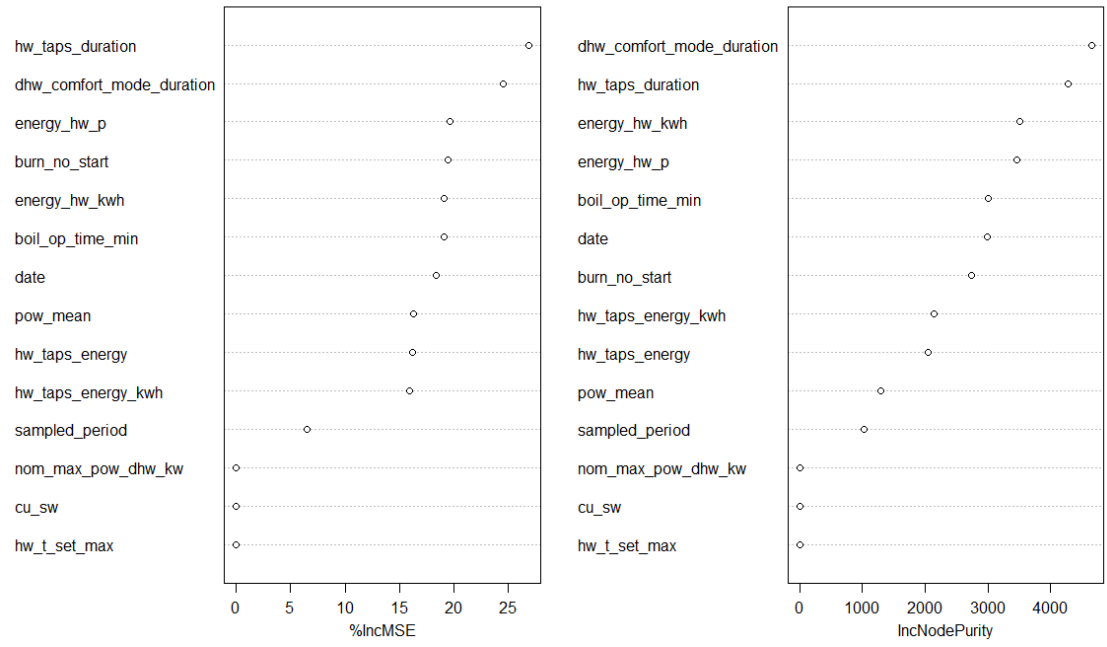


**Fig. 4.** The importance of variables is calculated by using Node impurity and MSE as shown in the diagram .The variables which appear in the top are more important.

After the data pre-processing has been done, now the data is ready to model by using a helpful algorithm. At first the random forests is applied by taking energy_kwh as the target variable against the rest of the variables. Random forest can be used to find important variables[4] by using "varImpPlot " attribute of the random forest package in R. The plot of this is given in Fig.4 along with the important variables listed in the top of plot. Mean squared error and node purity is used to find the important variables [5].

### 4.2 Leaps Package

Another good package used for variable selection is Leaps package. The function regsubsets() in the library leaps can be used for regression subset selection[7][15]. Thereafter one can view the ranked models according to different scoring criteria by plotting the results of reg subsets[8][16].

Here black indicates that variable is included more in the model, white indicates that they are not. So the variables which has more black marking is more important in creating model with energy_kwh as a response variable and the variables with less black marking is less important. The sample leaps package plot is shown in Fig.5.

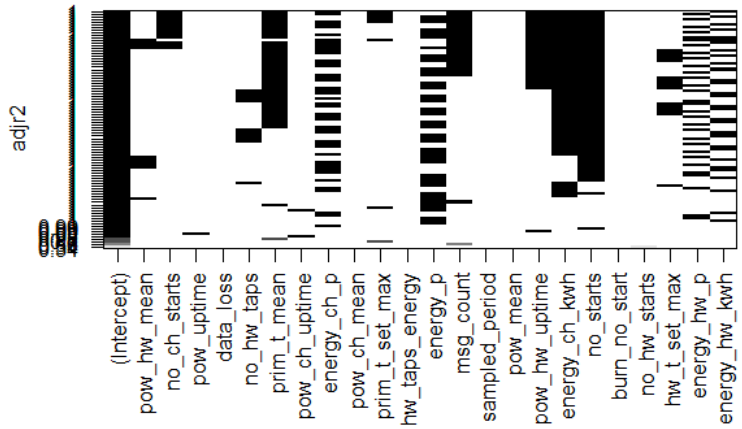


**Fig. 5.** Feature importance by using Leaps package. The variables which has more black markings are tend to involve more in the modelling, hence also more important. Here all the variables are not used because to increase readability. The purpose of this plot is to give the glimpse of the Leaps package usage.

### 4.3 LASSO Package

LASSO - Least Absolute Shrinkage and Selection Operator - was first formulated by Robert Tibshirani in 1996. It is a powerful method that perform two main tasks: regularization and feature selection[6]. The LASSO method applies a constraint on the sum of of the model parameters by taking their absolute values, the sum has to be less than a fixed value which is called as upper bound. In order to do so the method apply a shrinking process which is also known as regularization process, where it penalizes the coefficients of the regression variables shrinking some of them to zero[6]. While selecting important features, the variables that still have a non-zero coefficient after the shrinking process are selected to be part of the model. The aim of this process is to minimize the error in prediction. As the important variables diagram was unclear and LASSO did not perform as per to the expectations, the results were not considered. In our opinion the LASSO did not perform as expected because of huge data computaion required. As this is a complex algorithm, the time expected for computation is also large and the number of variables were also more in number such that the quality of image got distorted by reducing readability.

### 4.4 Relative Importance

Relative Importance is applied for the linear model. This comes with the package 'relaimpo'[23]. This is very simple step to observe the importance of the variables. The Relative Importance result is shown below in the Table 1. The important variables which are chosen in common from different algorithms of Leaps, Random Forest and Relative Importance are shown in Table 2. The variables are also picked based on our analysis manually. Here all variables picked is considered to be important as it is related to energy consumption of a system. This is shown in Table 1

**Table 1.** Table explains relative importance from decreasing order

| Variables | Relative Importance |
|---|---|
| pow_ch_uptime | 0.4794 |
| prim_t_set_ma | 0.1672 |
| burn_op_time_min | 0.0965 |
| pow_mean_on_uptime | 0.0795 |
| no_hw_rebound_statrs | 0.0473 |
| pow_hw_mean | 0.0359 |
| hw_t_outlet_min | 0.0282 |
| pow_hw_uptime | 0.0217 |
| dhw_comfort_mode_duration | 0.0149 |
| no_hw_starts | 0.0141 |

**Table 2.** Important Variables

| Important Variables | |
|---|---|
| energy_p | energy_hw_kwh |
| pow_mean | hw_taps_energy |
| sysid | date |
| pow_hw_uptime | burn_op_time_min- |
| pow_uptime | energy_hw_p |
| hw_taps_duration | pow_ch_uptime |
| nom_max_pow_ch_kw | pow_mean_on_uptime |

## 5 Fault Detection

In order to find the faulty systems and hence the variables that are important to find faulty systems, here it is assumed that the outliers are the indication of being fault. So when we find more outliers for a variable this can be an indication of fault. Firstly a univariate approach is taken, Fig.6 shows the boxplot of certain variables. But this univariate approach is not a good way in detecting outliers because the outlier may have the combination effect of group of variables.

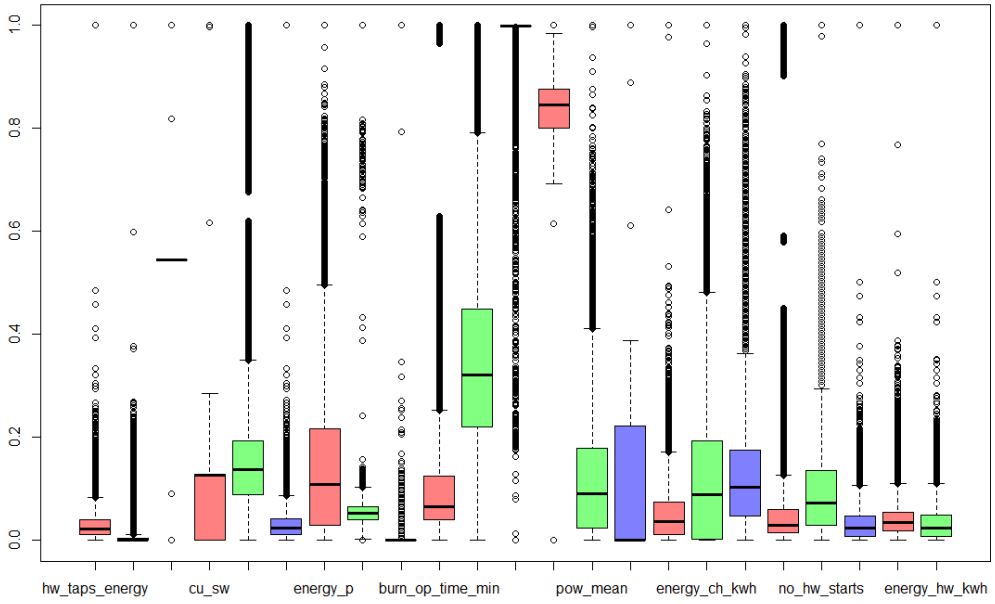


**Fig. 6.** By taking univariate approach, these boxplots are obtained which suggests to take multivariate approach

So it is better to take multivariate approach. This takes the combination of variables and find the outliers for combination. This can be considered as a anomaly or fault. So, multivariate approach is preferred over univariate approach[9]. So for this approach Cook's distance method was considered.

### 5.1 Cook's Distance

Cook's Distance method is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis[10][11]. In a practical ordinary least squares analysis, Cook's distance can be used in several ways to indicate influential data points that are particularly worth checking for validity, or to indicate regions of the design space where it would be good to be able to obtain more data points. After applying cook's distance, the variables with more outlier are said to be important variables for outlier detection and these are listed below in Table 3. The fault detection was also performed by taking energy_kwh as the target variable. And it was found that many of the important variables in shown in Table 2 also plays role in finding the faulty systems.

**Table 3.** Important Variables for Fault Detection

| Important Variables | |
|---|---|
| no_ch_pump_starts | pow_uptime |
| data_loss | fault_block_count |
| burn_op_time_min | pow_mean_on_uptime |
| fault_lock_count | msg_count |
| energy_ch_kwh | energy_hw_p |
| energy_hw_kwh | no_starts |

### 5.2 Influence Plot

An experiment is performed to detect faulty systems(outlier systems) by applying linear model and then applying Influence plot of Car package to that linear model to detect the systems(rows) which are deviating from the linear model. This is applied to original unprocessed data as the presence of NAs can be a important criteria for finding anomaly. The Fig.7 shows the influence plot. The influence plot uses Cook's distance to calculate the outliers[12][17]. After applying this, it predicted the following rows in the Table 4 as faulty or outlier systems. The sysid and appliance name is given in Table 5.

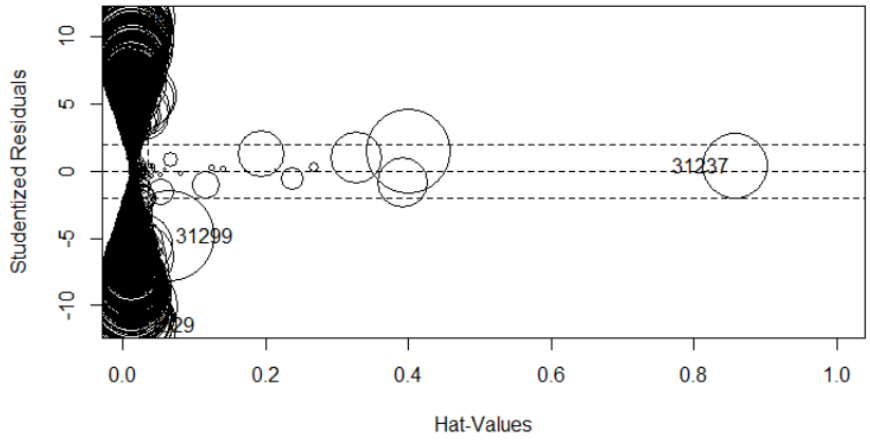


**Fig. 7.** The black marking shows that most of the devices follow the designed model but few is out of black marking which shows that it is outlier or faulty.

The system with sysid Ckf7JcEQ3knPUwBp_+5jsg== is predicted as faulty because most of the variables of it are N As, which certainly creates a doubt because as it is faulty it may not be able to obtain sensor values .

**Table 4.** After applying the influence plot for the original unprocessed data set, the algorithm detected these systems as faulty. The first column shows the row(appliance) which is faulty. The main measurement used to find anomaly is Cook's distance which is shown in third column. If the cook's distance is close to 0,it is more likely to be outlier.

| Row number | StudRes | Hat | CookD |
|---|---|---|---|
| 853 | 11.3393 | 0.0116 | 0.0026 |
| 21829 | -11.4738427 | 0.01217 | 0.0028 |
| 22185 | NaN | 1.0000 | NaN |
| 31237 | 0.3730 | 0.8577 | 0.0014 |
| 31299 | -4.8496 | 0.0648 | 0.0028 |

**Table 5.** Faulty systems with sysid and appliance name

| Sysid | Appliance name |
|---|---|
| YMVJ9zbXXBv0wjY9I5fYvA== | Greenstar 30Si Compact |
| nVHEXkG5YefqmilhpCzAWA== | Greenstar 30Si Compact ErP |
| Ckf7JcEQ3knPUwBp_+5jsg== | NA |
| QoGxpYDaaqVUQ_a6RgUeGw== | BG532/I Combi |
| QoGxpYDaaqVUQ_a6RgUeGw== | Spare part CU |

After the analysis, the system with sysid YMVJ9zbXXBv0wjY9I5fYvA== has minimum nom_max_pow_dhw_kw among all devices where as the other power utilization variables' value of this system is on the higher side. This would be a reason to be faulty. The other three systems are suspected to be faulty because some variables' values of these systems are near to that of mean value of those variables when considered overall systems. Meanwhile the value of other variables of those systems deviate from the overall mean of those variables compared with other systems. This hint that the systems may have partial problem with few sensors. Also the cook distance given in Fig.7 is relatively small for these systems which also suggests that these are faulty.

As per the influence plot and some analysis these 5 systems are produced as faulty systems. One more interesting thing was two systems with same sysid was found but for different appliance name. This can be seen in Table 5.

## 6 Comparision of Methods

For the data pre-processing steps the VIM package , feature correlation by ggplot and a simple scaling method were simple approaches and also gave the efficient results. For the next step of important variables selection, theoretically LASSO package was very promising but was not so helpful after implementing that. In further steps stronger algorithms like randomforest, leaps package , relative importance gave interesting results and helped in shortlisting the important variables. For the fault detection, outlier detection through the cook's distance method was a stronger approach rather than univariate boxplot approach. Hence the multivariate apoproach- cook's distance method along with influence plot applied to linear model was majorly helpful in detecting faulty systems and faulty variables.

## 7 Conclusion

After receiving the huge data from the Bosch, considerable amount of work was done on Data Preprocessing like Data scaling, Treating NAs with the help of packages like VIM and Amelia package, finding Correlations with the help of the package ggplot. Later algorithms and packages like Random Forest, Leaps, Relative Importance were applied to find the important variables. Due to time constraints the theoretical application of LASSO algorithm introduced by popular statistician Robert Tibshirani was not successfully implemented practically in this paper. Future work of this paper could have a scope on LASSO implementation. Finally Cook's Distance method and influence plot were used in order to find the outliers(faults) in the systems. Here mainly the multivariate approach was discussed. Random Forest with its complex structure and Leaps package with a good visualization of variable importance gave desirable results in important variables finding, where as both cook's distance with it's multivariate approach and Influence plot with its simple linear model deviation visualization combined with some systems analysis was helpful to find five faulty systems from the given Bosch dataset.

## References


1. Bernd Prantner,2011,Visualization of imputed values using the R-package VIM
2. James Honaker, Gary King, Matthew Blackwell,2011,Amelia II: A Program for Missing Data. Journal of Statistical Software
3. Meng, Xiao-Li, Robert Rosenthal, and Donald B. Rubin,1992, Comparing correlated correlation coefficients .
4. Isabelle Guyon,2003, An Introduction to Variable and Feature Selection
5. Brandon M. Greenwell, Wright State University and Bradley C. Boehmke University of Cincinnati and Andrew J. McCarthy The Perduco Group,2018,A Simple and Effective Model-Based Variable Importance Measure
6. Variable selection using Random Forests. Pattern Recognition Letters, Elsevier, 2010, 31 (14), pp.2225-2236.Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot.
7. Valeria Fonti and Eduard N. Belitser,2017,Paper in Business Analytics Feature Selection using LASSO
8. Graeme Hutcheson,2011,Variable selection: towards a restricted-set multi-model procedure
9. K. Senthamarai Kannan and K. Manoj,2015,Outlier Detection in Multivariate Data
10. Algur, S.P. and Biradar, J.G., 2017. Cooks Distance and Mahanabolis Distance Outlier Detection Methods to identify Review Spam. International Journal Of Engineering And Computer Science, 6(6).
11. T.Jagadeeswari , N.Harini ,2013, Identification of outliers by cook's distance in agriculture datasets.
12. https://rdrr.io/rforge/car/man/influencePlot.html Retrieved: 12.12.2018
13. Barnard, J Meng X. L, (1999-03-01), Applications of multiple imputation in medical studies: from AIDS to NHANES. Statistical Methods in Medical Research.
14. Mahdavi Damghani, Babak ,2012, The Misleading Value of Measured Correlation. Wilmott.
15. Waldron, L, Pintilie, M. Tsao, M. -S, Shepherd, F. A, Huttenhower, C.; Jurisica, I,2011,Optimized application of penalized regression methods to diverse genomic data
16. Vincent Calcagno,Claire de Mazancourt,2010, An R Package for Easy Automated Model Selection with (Generalized) Linear Models
17. R. Dennis Cook,1997, Detection of Influential Observation in Linear Regression
18. Ho, Tin Kam ,1995, Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC.
19. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome ,2008, The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5
20. Bennett, Derrick A. "How can I deal with missing data in my study?." Australian and New Zealand journal of public health 25.5 (2001): 464-469.
21. Sælensminde, Kjartan. "Inconsistent choices in stated choice data; use of the logit scaling approach to handle resulting variance increases." Transportation 28.3 (2001): 269-296.
22. Matthias Templ, Andreas Alfons, Alexander Kowarik, Bernd Prantner,VIM: Visualization and Imputation of Missing Values,2017-04-11
23. Ulrike Grömping 2006,Relative Importance for Linear Regression in R: The Package relaimpo