|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Detection of Outliers In Energy Usage  Analysis & Results |
| |  |  |  | | --- | --- | --- | | Brett Hawes (96016786) | 9/6/19 |  | |

# Abstract

A manufacturing firm has installed sensors on seven of its key machines which measure energy consumption in 15min increments with the firm’s hypothesis being that its operations are inefficient with energy usage on machines at times being higher than is optimal for a variety of reasons.

The objectives of this project are to establish the relationships of energy usage on each machine in relation to the other machines, production levels and climatic data and utilise those relationships to identify possible energy use outliers.

The process followed was to merge the different data sets into a single dataset and add extra indicator variables in order to help model recurring patterns. Data mining methods involving the application of different modelling approaches were then applied to this dataset for each machine to determine the optimal weights for each factor. In general, it was found that ensemble models consisting of a linear regression, decision tree and a neural network produced the best results.

The results indicated that it was possible to model the energy consumption of each machine in relation to energy consumption on the other machines, production levels and climatic data. By successfully establishing a relationship for each machine it has therefore been possible to identify data points were observed values are significantly higher than would be expected. These points can then be prioritised for investigation by the firm based on a combination of absolute and relative error to confirm whether they are in fact outliers or not. This feedback could then be used to improve the modelling process.

# Introduction

The purpose of this project has been to identify data points where recorded energy usage is significantly higher than expected. Such points are deemed to be ‘outliers’ and are indicative of potential energy waste by the company concerned. Successful identification of these outliers will allow the company to investigate the root-cause of this over usage, resolve any potential issues and therefore save money for the firm.

The firm has provided energy usage readings from seven key machines in 15-minute intervals along with daily production data and hourly climatic data.

The simplest way to detect possible outliers would be to plot energy usage over time and observe any potential usage spike.

The major disadvantage of this method is there is no ability to detect an outlier where a common (eg average) energy reading may have been observed but given the circumstances (eg energy usage on other machines, production levels) the reading should have been much lower.

The core hypothesis of the problem is that there are significant relationships within the system which if accurately modelled will greatly assist in the detection of outliers. Those relationships could be between machines, production volumes, climatic data or simply repeated patterns during the day or week or year.

A process is undertaken to apply a data mining process for each of the seven machines to determine which of the possible factors are most important for explaining energy usage on that machine. Each machine will therefore have a difference model which explains energy usage for that machine.

## Method / Approach

The general approach taken to model development was:

1. Data preparation
2. Modelling
3. Outlier identification

Step 1 was a general step for all machines while steps 2&3 were repeated for each of the seven machines. The following sections outline these steps in some detail.

**Data Preparation**

Four datasets were provided: 15min energy readings of the seven machines, daily production data, hourly climatic data and details around heating/cooling requirements through the period on a monthly basis.

The Python scripting language was used to combine and prepare the data for modelling and can be found in Appendix D. A high-level summary of processing steps is as follows:

* Read in the energy usage data
* Transform the data from ‘long’ format to ‘wide’ format
* Merge with production, climatic and heating/cooling data
* Create indicators for hour of day, day of week, week, year, public holidays and the Christmas / New Year period
* For the small of amount of missing temperature and humidity data, values were interpolated between the known values (deemed a valid approach as missing data was typically for only several hours at a time)
* Where production data was missing it was assumed that this was because there was no production on that day and therefore values were set to zero
* Removed data relating to 12/11/2018 which was incomplete

At this point a dataset at the 15min level granularity was created. The following steps were then applied to create daily and hourly datasets:

* Aggregations to either daily or hourly level.
* Additional indicators to indicate whether a machine reading was zero or not (see rationale under indicator list)
* Lagged variables for energy usage for machines. For daily data the lag was set at one day whereas for hourly data separate lagged values were created for up to six hours.
* Only compete data records between 01-Oct-2016 and 03-Dec-2018 were finally exported for modelling.

The table below outlines the additional variables which were created and the rationale.

|  |  |
| --- | --- |
| **Variable** | **Rationale** |
| Day of Week | Allows model to adjust prediction for repeated daily patterns |
| Hour of Day | Allows model to adjust prediction for repeated hourly patterns |
| Month of Year | Allows model to adjust prediction for repeated monthly patterns |
| Year | Allows model to adjust to broad trends in energy usage over time |
| Public Holidays | Stop model from overpredicting usage on public holidays |
| Xmas / New Year Period | Stops model from overpredicting usage through the Christmas / New Year holiday period |
| Zero energy usage indicators (set to 1 when reading was zero) for each machine | The rationale for adding these was the hypothesis that when readings were non-zero there maybe some linear relationship with other variables (say 0.25). However, inclusion of the zero points may force the model to fit a different line (say 0.5) to accommodate the zero values.  Adding the zero indicators allows a linear model to fit the right relationship to both zero and non-zero situations |
| Machine lags (1 to 6 hours) | Six variables for each machine indicating 1-hour, 2-hour lags and so on respectively. The purpose of creating lag variables is that there is a hypothesis that some machine’s energy usage would be related to other machines energy usage earlier in time. This is particularly relevant if a machine is a ‘downstream’ process. In these situations, it would be expected that energy usage would be related to energy usage on an ‘upstream’ machine at some early point in time. |

**Modelling**

SAS Enterprise Miner 14.1 was used for modelling. The high-level approach to model development was:

* Partition the data into 60% training, 20% validation and 20% test. The purpose of this step is to fit the model to the training data and avoid overfitting the model by referencing the validation data (by ensuring average squared errors is decreasing in the validation data set as model complexity increases) . The test data allows for a final check on the model performance with the characteristics of the test data having no influence on the overall model selection. We would be expecting model performance under the test data to be in line with that observed in the validation data.
* The following modelling approaches were then run:

|  |  |
| --- | --- |
| **Model** | **Setup Options** |
| Linear Regression | Stepwise selection with interaction effects for the time indicators (hour, weekday, month and year) |
| Decision (Regression) Tree | SAS EM default options, splits based on F-test(to minimise ASE) |
| Neural Networks | Multilayer perceptron with 1-6 hidden layers, hidden bias = ‘yes’ |
| Gradient Boosting | Shrinkage = 1%, Iterations =500 |
| Partial Least Squares | Iterations = 200, factors = 15 |
| Manual Ensemble | Combination of Linear Regression, Decision Tree and the best Neural Network |

* The best model would be selected by choosing the resulting model with the lower average squared error. This model would be the model with the least amount of unexplained variation.
* With the best model, the following diagnostic charts were run to test fit:
  + Fit vs Actual plots over time
  + Fit vs Actual plots only
  + Residual histogram
  + Residual vs target variable

**Outlier Identification**

From the selected model, potential outliers could be identified by looking at the residuals of actual values observed vs fitted values.

An outlier was defined as being a positive residual that exceeded three standard deviations with the three standard deviations being selected as an arbitrary point which gives a useful number of outlier for initial investigation. Only positive residuals were considered as outliers as the purpose of the exercise is to identify points where energy usage was higher than expected.

Outliers were then placed into a matrix according to:

* Absolute difference as measured by the standard deviation of the residual
* Relative differences defined as being the actual value observed vs the modelled value

Outliers with both high absolute and relative differences would be most likely to be true outliers and provide a mechanism for prioritising identified outliers for investigation.

## Results

The following results focus in detail on identifying outliers in energy usage in the by-product processing facility on an hourly basis.

The process for building a model and identifying outliers for the by-product processing facility was then repeated for the other six machines. The results for those machines can be found in Appendix C

Hourly data was chosen to work with as hourly data felt like it was granular enough to identify short term outliers. Once the model could be validated and refined with the customer then it would be appropriate to try and apply the model to the 15min daily data set.

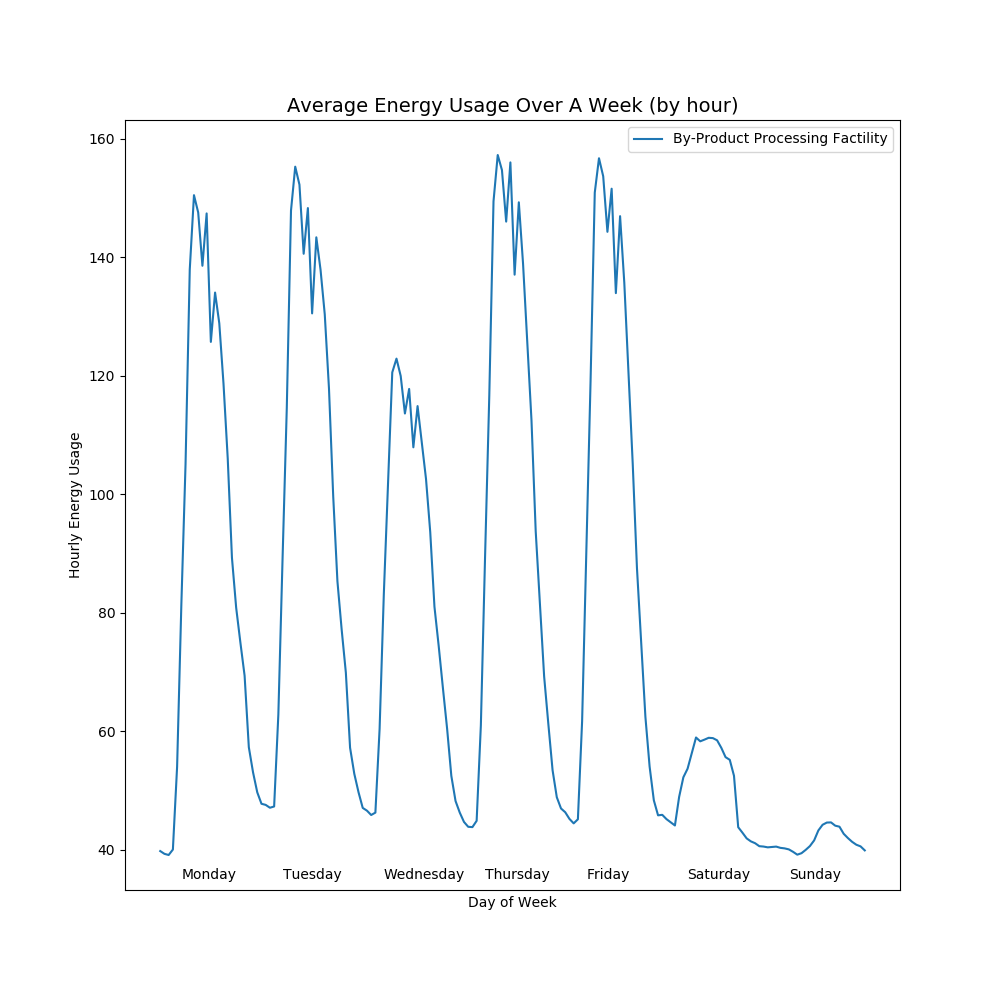
**Initial Exploratory Analysis**

Averaging energy usage for each hour across each day of the week and plotting gives a quick overview of the energy usage pattern for the by-product processing facility.

Observation of the plot show:

* A general Monday – Friday working hours pattern
* Drops at midday and in the afternoon, possibly due to breaks
* Wednesdays showing lower energy usage than other mid-week days
* Some, but limited usage on Saturday
* Very limited usage on Sundays

***Figure 1 – weekly pattern on the by-product processing machine***

****

These observations confirm that using day of week and hourly indicators would be useful to model these trends.

Reviewing the energy usage for all machines (figure 2) simultaneously highlights the following points:

* ‘Product B – Further Processing Facility’, ‘Air Compressor – Fixed Speed’, ‘Product B Processing Facility’, ‘Product M Processing Facility’ all show similar ‘working day’ patterns to the by-product processing facility.
* The ‘Air Compressor – Variable Speed’ simply seems to be ‘on’ during working hours and ‘off’ otherwise.
* The ‘Main Refrigeration Compressor’ has by far the highest usage of all machines. Usage patterns show a ‘working day’ pattern with a possible lag to the other machine. Out of hours, usage is highly erratic.

***Figure 2 – weekly pattern across all machines***

A picture containing text, map

Description automatically generated

Figure 3 below indicates the high level correllations between the seven machines. All machines show a high level of correllation except for the refrigeration compressor. Appendix A contains a full listing of correlations between machines. A pairwise matrix of scatter plots also gives a good indication of where clear linear relationships between machines exist or not (found in Appendix A).

|  |  |  |
| --- | --- | --- |
| **Machine** | **Correlation** | **Linearity** |
| Air Compressor - Variable Speed | 0.95 | Linear initially, then flat |
| Air Compressors - Fixed Speed | 0.97 | Flat initially, then linear |
| Product B - Further processing facility | 0.90 | Weak linearity |
| Product B Processing Facility | 0.98 | Strong linearity |
| Product M Processing Facility | 0.97 | Strong linearity |
| Main Refrigeration Compressor | 0.63 | Weak |

***Table 1 – correlation between by-product machine and other machines***

## Modelling Results

SAS Enterprise Miner was used to fit differing models to the data. The measure of model fit used was the Average Squared Error (ASE). Models with a lower ASE are indicative of a better fit. A better fit overall is likely to mean that any outliers observed will be true outliers.

For the by-product processing machine, the ASE results of the different modelling approaches, in order of best fit were as follows:

|  |  |  |
| --- | --- | --- |
| **Model** | **ASE - Validation** | **ASE - Test** |
| Ensemble (Neural Network 5 + Linear Regression + Decision Tree) | 55.97 | 55.31 |
| Neural Network (5 hidden layers) | 58.38 | 57.79 |
| Neural Network (3 hidden layers) | 59.86 | 58.81 |
| Neural Network (4 hidden layers) | 63.37 | 62.77 |
| Linear Regression | 64.4 | 64.68 |
| Neural Network (6 hidden layers) | 67.60 | 65.28 |
| Neural Network (1 hidden layers) | 70.15 | 71.69 |
| Partial Least Squares | 72.1 | 73.3 |
| Gradient Boosting | 77.06 | 76.1 |
| Decision Tree | 78.41 | 74.92 |
| Neural Network (2 hidden layers) | 128.42 | 122.76 |

The fact that the ASEs observed on the test data are in line (and rank the models in the same order) with those that are observed in the validation data helps to support the selected best model.

**Key drivers of energy usage**

The linear regression and decision tree modelling approaches give significant insight into the factors driving modelling results. Neural networks, gradient boosting and ensemble models by their nature contain numerous sub-models which make it more difficult to identify the drivers.

By looking at the results for the linear regression and decision tree we can get some insights into the likely drivers (technical details can be found in Appendix B) of the final model.

|  |  |  |
| --- | --- | --- |
| **Variable Importance\*** | **Linear Regression** | **Decision Tree** |
| 1 (most important) | Product B Processing Machine | Product B Processing Machine |
| 2 | Product M Processing Machine (1-hour lag) | Product M Processing Machine |
| 3 | By Product Machine Processing (1-hour lag) | By product machine processing (1-hour lag) |
| 4 | Air Compressor Fixed Speed | Hour |
| 5 | Air Compressor Fixed Speed (3-hour lag) | Year |
| 6 | Hour of day | Air Compressor Fixed Speed |
| 7 | Interaction effect of Hour of day and Weekday | Product M Processing Machine (1-hour lag) |
| 8 | By-Product Processing Machine (4-hour lag) | Product B Processing Machine (1-hour lag) |
| 9 | Year | By-Product Processing Machine (3-hour lag) |
| 10 | Air Compressor Fixed Speed (4-hour lag) | Month |

*\*Defined as being the order of entry in a stepwise selection for linear regression and variable importance from SAS EM for the Decision Tree*

Reviewing the variable importance for the linear regression and the decision tree show several similarities. Both models place the most importance on energy usage levels on product B, product M and by-product processing machines. Both models also use the hour of day and year indicators in order to model repeating (in the case of the hourly indicator) or trends (in the case of the year indicator). Air compressor fixed-speed and lags of the by-product processing machine also feature.

The results indicate the relation to energy usage on the by-product processing machine is linked to usage on product B and M machine (sometimes with a lag) and recurring daily patterns.

The neural network results and final ensemble model are likely to have these same key drivers explaining most of the variation however their added sophistication is likely to be using other variable combinations to explain small pockets of variation in the data.

## Model Diagnostics

All diagnostics are performed on the validation dataset. This is the appropriate dataset to perform diagnostics on as the model has been fitted to the training data meaning any actual model issues will be less visible in this dataset compared to the validation dataset.

**Modelled Values vs Actual Over Time**

Figure 4 displays a random selection of 150 hourly data points over the period. The plot indicates that the model has the general ability to predict a wide range of actual values.

***Figure 4 – fitted vs actual usage for by-product processing facility***

*A close up of a piece of paper

Description automatically generated*

**Modelled Values vs Actual**

Plotting all the fitted values from the training data set against their actual values gives the plot seen in figure 5. Ideally, we want to see a 1:1 relationship from bottom left to top right. The plot gives a good indication of the model’s ability to fit across the spectrum of observed values.

**Figure 5 – fitted vs actual usage for by-product processing machine**

**A screenshot of a cell phone

Description automatically generated**

**Residual vs Actual & Initial Outlier Classification**

Looking at figure 6 of residual values versus actual values indicates that residuals are well spread around zero for actual values under 150. However, for actual values over 150 residuals become biased to being consistently positive. This is indicating that the model tends to under predict values at higher levels of energy usage.

The standard deviation of residuals was 7.44. Applying an arbitrary cut-off level of three standard deviations would mean any point with a residual in excess of approximately 21 could be deemed a potential outlier. This is indicated by all points which lie above the red line in figure 6.

Noting the model tends to underpredict at higher levels means we needs to apply some caution to outlier identification at higher levels of energy usage. For instance, a residual of 25 at and energy ready of 150 isn’t as likely to be an outlier as a residual of 25 at a reading of 100 due to this observed bias. The outlier classification sections deal with this issue in more detail.

**Figure 6 – residuals vs actual usage for by-product processing facility**

**A screenshot of a cell phone

Description automatically generated**

**Residual Distribution**

Figure 7 below shows the distribution of residuals being generally normally distributed.

***Figure 7 - distribution of residuals***

*A screenshot of a cell phone

Description automatically generated*

## Outlier Identification

Table 2 below further classifies potential outliers. Within the validation data set a total of 26 hourly data points were identified that met the criteria of having a residual in excess of three standard deviations from the residual mean.

One approach for investigating and confirming outliers would to treat all 26 observations as equally likely to be an outlier. This is likely to be an inefficient way of investigating and prioritising investigations.

Recognising the value of a client’s time needed to investigate possible outliers the table below outlines these outliers on the following dimensions:

1. Absolute error. The number of standard deviations from the mean for the residual
2. Relative error. The percentage difference between the actual value observed and the fitted value.

It is expected that, all else being equal, points with higher absolute and higher relative errors are more likely to be true outliers. These points are highlighted in grey in table 1. For instance, the for points with absolutely errors of 3.0-3.5 standard deviation the three points with relative errors of 40% are more likely to be outliers than the other 19 points and should be prioritised for investigation.

**Table 2**

| **# of potential outliers** | | Relative Error | |
| --- | --- | --- | --- |
| **20%** | **40%** |
| **Absolute Error (St Deviations)** | **3.0** | 12 | 2 |
| **3.5** | 7 | 1 |
| **4.0** | 2 | 0 |
| **4.5** | 1 | 0 |
| **5.0** | 1 | 0 |

## Other Machines

The same modelling process was applied to the other six machines. The following table summarises the best model found, key drivers and number of outliers that were identified. Full diagnostic charts for each machine can be found in Appendix C

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Machine** | **Best Model\*** | **ASE - Validation** | **Key Drivers\*\*** | **Outliers @ 3 St. Dev.** |
| Air Compressor – Fixed Speed | Ensemble (LR+DT+NN6) | 32.06 | -Air Compressor – Fixed Speed (1-hour lag)  -Hour  -By Product Processing Facility | 33 |
| Air Compressor – Variable Speed | Ensemble (LR+DT+NN6) | 3.82 | -Air Compressor – Variable Speed (1-hour lag)  -Hour  -Air Compressor – Zero Reading (indicator variable) | 15 |
| Product B | Ensemble (LR+DT+NN6) | 27.67 | -Product B Processing (1-Hour lag)  -Hour  -Air Compressor Fixed Speed | 51 |
| Product M | Ensemble (LR+DT+NN3) | 10.71 | -Product M Processing (1-Hour lag)  -Hour  -Product B Processing | 35 |
| Product B -Further Processing Facility | Ensemble (LR+DT+NN6) | 103.24 | - Product B Further Processing Facility (1-Hour lag)  -Hour  -Product M Processing | 35 |
| Main Refrigeration Compressor | Decision Tree | 849 | -Refrigeration Compressor (1-Hour lag)  -Hour  -Refrigeration Compressor (3-Hour lag) | 42 |

*\*’LR’ denotes Linear Regression, ‘DT’ denotes Decision Tree, ‘NNx’ denotes a Neural Network with x hidden layers*

*\*\*Drivers determined from stepwise linear regression results*

## Discussion

The results above indicate that it is possible to build a model which can identify a small subset of observations that have the potential to be outliers where energy usage is higher than would be expected. No extra data appears to be required other than to prepare a number of supplementary indicator variables to help model recurring patterns in the data.

A variety of models were explored. Several them could potentially be used for the purpose of identifying outliers with each showing a reasonable ability to model the variability in the data. If the client would prefer to adopt a model with greater interpretability the linear regression (in particular) and decision tree models could still be effective. This report however recommends the use of an ensemble model utilising linear regression, decision trees and a neural network. The reason for this is that neural network can specifically model more subtleties in the data and the overall ensemble in this instance seems to allow for the strengths of each model to be utilised. The practical implications of this to the clients is that this model is expected to be less likely to product false positives which is likely to save the client time and money in terms of investigation of the outliers produced.

The general data mining approach employed can be applied to other machines efficiently. While this data mining process was specifically built around the by-product processing facility once the modelling approach was confirmed the amount of effort to apply to the other machines was very low. The diagnostics analysis in Appendix C indicates that each of these models has a reasonable fit.

Evolution of these models would require client feedback on the possible outliers. If the points identified are confirmed as not being an outlier, then it would be expected that some other form of explanation would become apparent that then be incorporated into the modelling for the next iteration.

The current model involved no transformation of any variables. Modelling was attempted with log transformations of the product B, product M and product B further processing facility (which generated the most normal like distribution), however modelling with transformed variables did not have the effect of improving the ASE. The likely explanation for this is that the energy usage on machine is multi-modal concentrations of values at different levels (either ‘high’/’low’ or ‘on’/’off’).

Input data into the data mining process only included data points where full data was available for all factors. This resulted in any data points before October 2016 being discarded. While this approach is correct for the linear regression and neural network the decision tree model could benefit from being able to be trained from the entire dataset given their ability to handle missing data. An improvement to the modelling process could be to allow the decision tree model to be trained off the entire period.

Several machines appear to be highly correlated to each other. It therefore could be possible to use principal components analysis to reduce the input factors into few dimensions, however interpretability of results could be harder.

## References

L Breiman, JH Friedman, RA Olshen, and CJ Stone. Classiﬁcation and Regression Trees. Wadsworth Inc, 1984

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression

trees. Journal of Animal Ecology, 77(4), 802-813

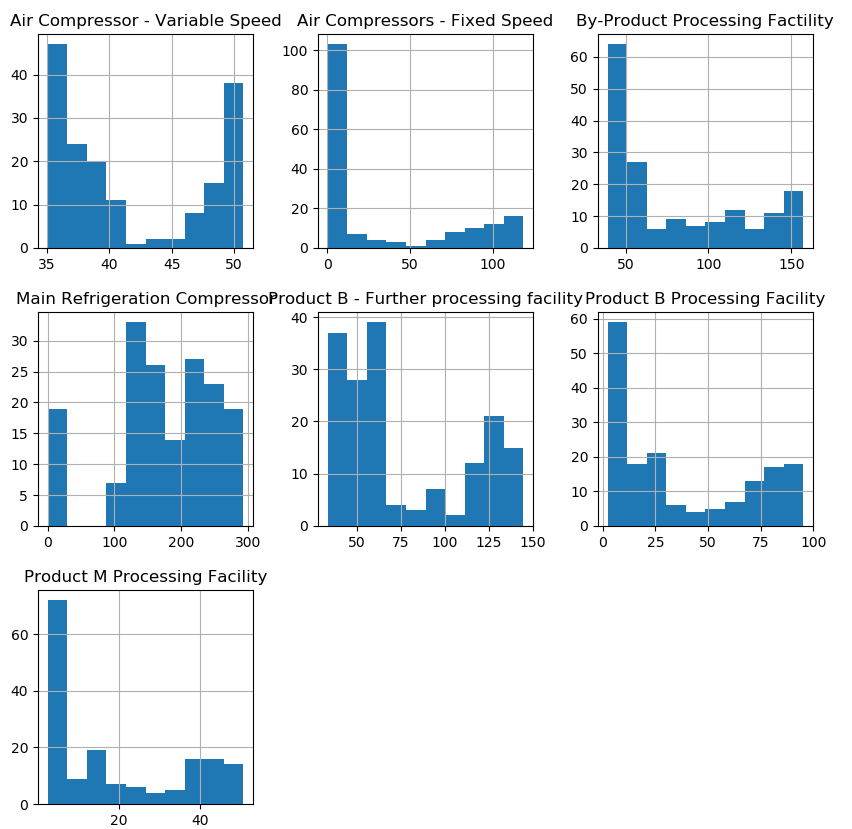
http://support.sas.com/documentation/cdl/en/stathpug/68163/HTML/default/viewer.htm#stathpug\_hpsplit\_details02.htm

## Appendix A – Initial Exploratory Analysis

**Correlation table across all machines**

|  | **Air Compressor - Variable Speed** | **Air Compressors - Fixed Speed** | **Product B - Further processing facility** | **By-Product Processing Facility** | **Product B Processing Facility** | **Product M Processing Facility** | **Main Refrigeration Compressor** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Air Compressor - Variable Speed** | 1.000000 | 0.914884 | 0.913414 | 0.950911 | 0.966582 | 0.953482 | 0.610422 |
| **Air Compressors - Fixed Speed** | 0.914884 | 1.000000 | 0.887903 | 0.965821 | 0.941346 | 0.947803 | 0.531883 |
| **Product B - Further processing facility** | 0.913414 | 0.887903 | 1.000000 | 0.895924 | 0.922025 | 0.951959 | 0.515629 |
| **By-Product Processing Facility** | 0.950911 | 0.965821 | 0.895924 | 1.000000 | 0.976001 | 0.969198 | 0.630830 |
| **Product B Processing Facility** | 0.966582 | 0.941346 | 0.922025 | 0.976001 | 1.000000 | 0.976508 | 0.613928 |
| **Product M Processing Facility** | 0.953482 | 0.947803 | 0.951959 | 0.969198 | 0.976508 | 1.000000 | 0.527322 |
| **Main Refrigeration Compressor** | 0.610422 | 0.531883 | 0.515629 | 0.630830 | 0.613928 | 0.527322 | 1.000000 |

**Distribution of energy usage across machines**



**Pairs plot between machines**

A close up of text on a white surface

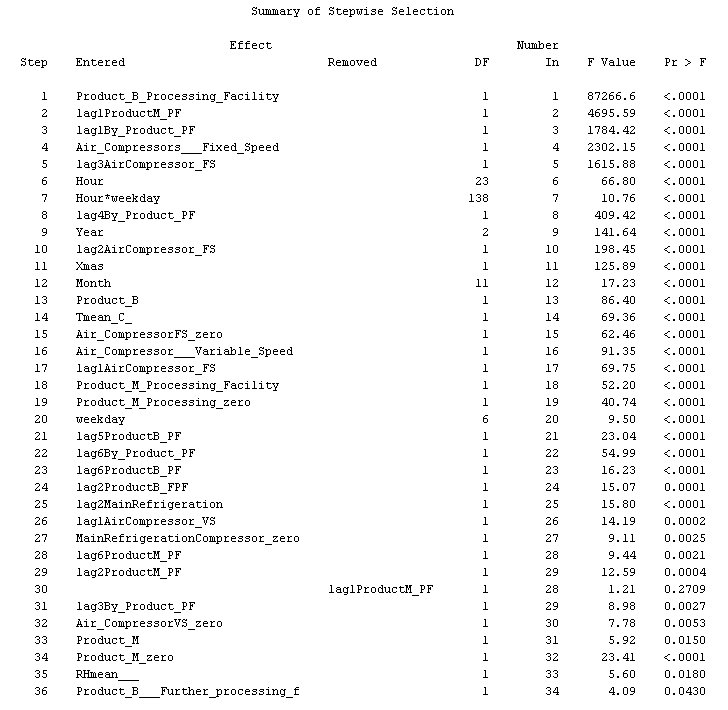
Description automatically generated

## Appendix B Input Variable Importance

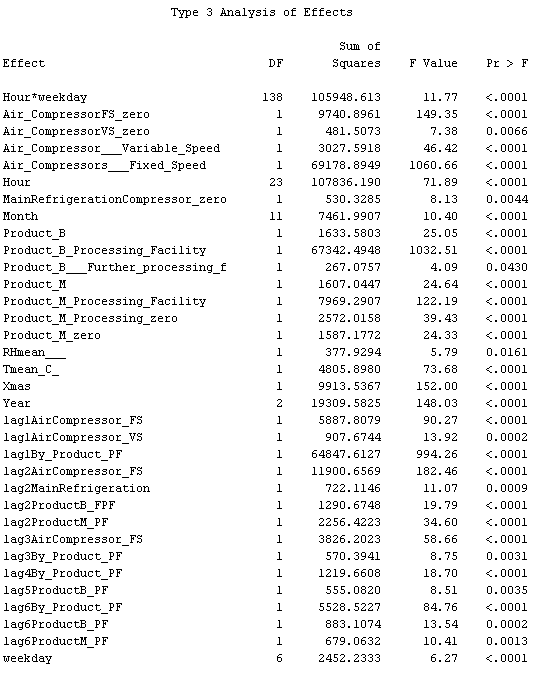
**Decision Tree Variable Importance**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Number | Variable | Splitting Rules | Validation Importance |
| 1 | Product B Processing Facility | 3 | 1 |
| 2 | Product M Processing Facility | 2 | 0.290037016 |
| 3 | By-Product Procssing Facility (1 hour lag) | 16 | 0.195535471 |
| 4 | Hour | 9 | 0.173855319 |
| 5 | Year | 7 | 0.050984606 |
| 6 | Product M Processing Facility (1-hour lag) | 1 | 0.04120923 |
| 7 | Air Compressor Fixed Speed | 3 | 0.036212384 |
| 8 | By Product Processing Facility (3-hour lag) | 1 | 0.030136084 |
| 9 | Month | 2 | 0.026213566 |
| 10 | Product B Processing Facility (1-hour lag) | 2 | 0.022219899 |
| 11 | Weekday | 1 | 0.019635002 |
| 12 | Air Compressor Fixed Speed (1-hour lag) | 1 | 0.014032593 |
| 13 | Product B Processing Facility (2-hour lag) | 2 | 0.011482678 |
| 14 | Product B Further Processing Facility | 1 | 0.009754126 |
| 15 | By-Product Processing Facility (4-hour lag) | 1 | 0.005262323 |

**Linear Regression Stepwise Order**



**Linear Regression Type 3 Errors**



## Appendix C: Diagnostics for Other Machines

**Air Compressor Fixed Speed**

**A close up of a logo

Description automatically generated**

| **# of potential outliers** | | Relative Error | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **20%** | **40%** | **60%** | **80%** | **100%** |
| **Absolute Error (St Deviations)** | **3.0** | 7 |  | 4 |  | 2 |
| **3.5** | 9 |  | 1 | 2 |  |
| **4.0** | 3 | 1 | 2 |  |  |
| **4.5** |  |  |  |  | 1 |
| **5.0+** |  | 1 |  |  |  |

**Air Compressor Variable Speed**

**A close up of a logo

Description automatically generated**

| **# of potential outliers** | | Relative Error | |
| --- | --- | --- | --- |
| **20%** | **100%+** |
| **Absolute Error (St Deviations)** | **3.0** | 8 | 0 |
| **3.5** | 5 | 0 |
| **5.0+** | 1 | 1 |

**Product B Processing Facility**

**A screenshot of a cell phone

Description automatically generated**

| **# of potential outliers** | | Relative Error | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **20%** | **40%** | **60%** | **80%** | **100%+** |
| **Absolute Error (St Deviations)** | **3.0** | 7 | 6 | 2 | 1 | 4 |
| **3.5** | 2 | 2 | 3 | 1 | 4 |
| **4.0** |  | 1 |  | 2 | 2 |
| **4.5** |  |  | 2 | 2 | 1 |
| **5.0+** |  |  | 2 | 2 | 5 |

**Product M Processing Facility**

**A screenshot of a cell phone

Description automatically generated**

| **# of potential outliers** | | | Relative Error | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **20%** | | **40%** | | **60%** | **80%** | **100%+** | |
| **Absolute Error (St Deviations)** | **3.0** | | 9 | | 2 | | 2 | 1 | 5 | |
| **3.5** | |  | | 1 | | 1 | 1 | 2 | |
| **4.0** | |  | | 4 | |  |  | 1 | |
| **4.5** | |  | | 2 | | 2 |  | 1 | |
| **5.0+** | |  | | 1 | |  |  |  | |
|  | |  | |  | |  | | | |  | |  |

**Product B Further Processing Facility**

**A screenshot of a cell phone

Description automatically generated**

| **# of potential outliers** | | Relative Error | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **20%** | **40%** | **60%** | **80%** | **100%+** |
| **Absolute Error (St Deviations)** | **3.0** | 6 | 5 | 1 |  |  |
| **3.5** |  | 6 | 4 |  |  |
| **4.0** |  | 3 | 2 |  |  |
| **4.5** |  |  | 1 |  |  |
| **5.0+** |  |  | 2 | 5 |  |

**Main Refrigeration Compressor**

**A screenshot of a cell phone

Description automatically generated**

| **# of potential outliers** | | Relative Error | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **20%** | **40%** | **60%** | **80%** | **100%+** |
| **Absolute Error (St Deviations)** | **3.0** |  | 3 | 1 | 3 | 4 |
| **3.5** |  | 1 |  | 1 | 5 |
| **4.0** |  |  | 2 |  | 4 |
| **4.5** |  |  | 1 | 1 | 6 |
| **5.0+** |  |  |  |  | 10 |

## Appendix D: Python Scripts for Data Preparation and Model Diagnostics

HTML of Python script for data preparation



HTML of Python script for diagnostics.

