Energy Data Analysis - ABC Paper Company

Suresh Ooty

Table of Contents

This document is the *Final Project Report* submitted by Suresh Ooty during the Fall semester of 2016 for the course *IMSE 514 - Multi Variate Statistics*.

The choice of using RMarkdown for documentation was made to enable continous research on this subject

Acknowledgement

The researcher expresses his gratitude to *Prof.Yung-Wen Liu*

- for letting the researcher to choose a dataset that belongs to an organization to which the researcher works for
- for giving the freedom to explore a dataset that is seemingly humungous

- for being a motivation to atttempt something which could be challenging to complete during a course project and
- for extending the support beyond the limits of course

"I hope that I tried to deliver the intent of the course project using this dataset" - Suresh Ooty

How to read this document

- All the research work has been captured in separate html documents. To navigate, please browse through the "Navigation Banner" above in this page.

Project Background Summary

ABC paper Company produces variety of toilet papers from their factories across the globe. The company wants to know what could be the unknown patterns or variables that could be potential levers to improve their performance or save energy consumption.

Dataset Overview

An IT monitoring solution deployed across its global manufacturing facilities captures sensor data for every 15 minutes and streams meaningful detail to dashboards that are monitored by Plant Managers. The dataset used for this project is a subset of a large dataset, that is specific to single facility. This dataset contains site level atttributes for 2015, with ~30,000 records.

This is a subset of a large dataset that has millions records concerning to machine level attributes. The dataset taken for this study is for a single year with attributes at site level.

The chosen dataset has the following attributes

```
##  [1] "RecordID"       "DateTime_Out"   "TimeStamp"      "Production"
##  [5] "Hay_out_waste"  "ShutDownFactor" "LeadToFailure"  "CmpACon"
##  [9] "EleCon"         "NatGCon"        "SteCon"         "WatGCon"
## [13] "WatMCon"        "WatWGen"
```

No.of Observations & No. of Attributes

```
## [1] 32930    14
```

- RecordID - Unique Id for each observation, captured at every 15 minutes
- DateTime_Out - Date in mm/dd/yyyy format, derived value from 'TimeStamp' attribute
- TimeStamp - TimeStamp of each observation (eg., 2014-05-08 17:15:00)
- Production - Paper produced in Metric Tonnes (MT)
- Hay_out_waste - Waste produced in Metric Tonnes (MT), if this is '0' when 'Production' is nill means the plant is shut down
- ShutDownFactor- a calculated to determine based on *Hay_out_waste* and *Production*
- LeadToFailure - a calculated DUMMY variable, if it is TRUE means the plant was shut down in next 15, 30 min observations
- CmpACon - Compressed Air Consumed
- EleCon - Electricity Consumed
- NatGCon - Natural Gas Consumed

- SteCon - Steam Consumed
- WatGCon - Water (Ground) Consumed
- WatMCon - Water (Municipal) Consumed
- WatWGen - Water (Waste) Generated

The initial dataset had more variables that were calculated using one of the above varibles or found highly correlated (close 1), hence removed from the core dataset considered for this study. For eg., Compressed Air Generated, Electricity Purchased, Natural Gas Purchased, Steam Generated, Water Ground Purchased, Water Municipal Purchased, Waste Water Purchased, Waste Water Generated value and calculated costs on all these attributes.

The cost attributes were removed as the costs are always derived.

Tools

Given the wast scope and exploratory nature of the project, the reasearcher employed his liberty to use advanced Data Wrangling & Analysis tool such as Alteryx for quicker Data Transformation, R Markdown to bring power of *R* to documentation and the mighty *R* for Analysis on the R Studio platform.

Study

Variable Selection :

The variables *Production, Hay_out_waste, LeadToFailure, CmpAcon, EleCon, NatGCon, SteCon, WatGCon, WatMCon & WatWGen* were chosen for the correlation analysis to see if any of them are correlated with each other.

Outliers were removed by using scatter plots, residual plots and by reducing the sample sizes in a very random manner.

Application of Models :

Linear Regression Model on Full dataset Linear Regression Models on 3 slices of dataset based on production value Logistic Regression model to predict failures on a random train & test sample Neural network model to predict failures on a random train & test sample

Conclusion

Given the time constraint (Data collection from SME during holiday season!!), a very random and iterative methods of variable selection, different models to identify patterns and to predict were tried by the researcher. The observations were made & documented as professionally as possible so that this study could be continued.

Take aways from this study:

- Never start with a very large dataset - Make reasonable small datasets for analysis
- Understanding of dataset lead to make some reasonable judgements on creating DUMMY variables
- Applying Logistic & Neural Network did not come out effective enough to predict failures (at least with given dataset). This is the key focus on the future study
- A site level attribute set defines the function of the production plant, but it may prove out that it captures every noise, deviation & interactions caused by individual machines in the plant. So, to predict failure in such scenarios, a more complex system simulation model may be necessary to predict the failures.