ENSAI 20 ANS École nationale de la statistique et de l'analyse de l'information

energiency

Big Data Project

Machine Failure Prediction

Submitted by

Moustapha OUSMANE BAWA GAOH
Shashank SHARMA
Shishir DUBEY

Master of Science in Big Data Students

Supervised by

Aymeric HERVIEU
R&D Director, Energiency

Date: 30/03/2016

Contents

I. Introduction

1. The Enterprise

Energiency is a French technology startup located at Rennes, and comprises a team of 10 engineers engaged in the development of cloud based predictive energy analytics solutions for European manufacturing companies. The company has developed big data analytics software to enable energy management in manufacturing industries.

2. Problem Description

The objective of the project is to predict the failure of machines to enable energy management in manufacturing industries. This helps in the planning of industrial tasks to improve energy management. The project has been executed by statistical analysis of large amounts of machine data collected from French industries over the past two years. Results obtained from statistical analysis will help the industries in reducing undesirable machine stoppages during manufacturing. Data including energy consumption, production, and maintenance planning are considered for the analysis, and the probability of machine failure for the following days is analyzed.

II. Available Data

The data used for the project was collected at regular intervals from machines operating within industrial plants. The data includes the variables of production quantity, electricity consumption, maintenance planning and production weight, spanning a period of three years, from 2013 to 2016. Time interval frequency of the data ranges from 10 minutes to 1 hour.

1. Data Preparation

Before proceeding with the analysis, the available data was sanitized by using the following steps:

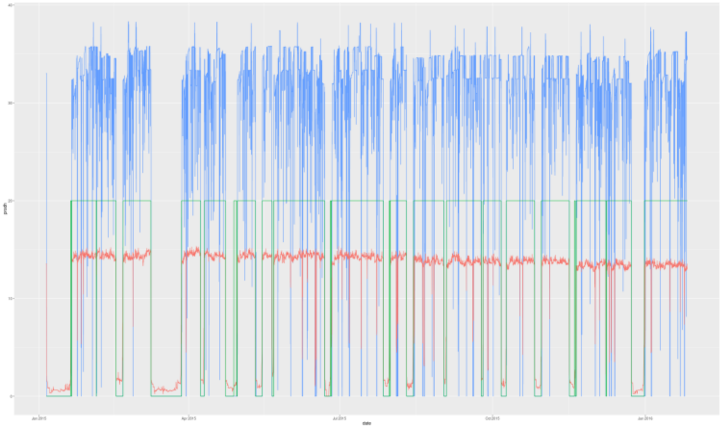
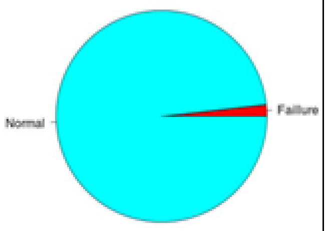
a. Re-sampling of the data into one hour time intervals.

b. Merging of the data together and removing rows with null values, in order to create a target variable. (The first plot in the Appendix presents the data before removing the null values and the non-relevant variables)

c. Creating the target variable of machine failure prediction with binary levels. Failure implies no planned machine stop and production quantity lower than 20.

d. Transforming the data to consider only the first instance as machine failure and the remaining instances as normal machine operation.

e. Exploring the data to seek the proportion of failure:

First Failure Proportion	Pie Chart of Target
	

f. Creating the following new variables:

Variable Name	Variable Type	Description
Worktime	Continuous	Counter for normal machine operation.
DeltaElec	Continuous	Variation of electricity consumption for a time unit
DeltaProd	Continuous	Variation of production quantity for a time unit
Month	Factor	Month of machine operation
Weekday	Factor	Day of the week of machine operation
Log(Elec+1)	Continuous	--
Sqrt(Elec)	Continuous	--
Elec^2	Continuous	--
Elec^3	Continuous	--
PolyElec5	Continuous	--
Log(Prod+1)	Continuous	--
Sqrt(Prod)	Continuous	--
Prod^2	Continuous	--
Prod^3	Continuous	--
PolyProd5	Continuous	--

g. Transforming the data to create variables for 24 hours before a given time, to obtain a final dataset of 723 variables and 1 target variable (which considers the first failure instance).

h. Removing the current time variables and considering only the variables with observations for the 24 preceding hours.

III. Methods & Tools Used

The following excerpts obtained from research publications accurately capture the essence of the statistical analysis undertaken in the project:

"Failure prediction is about assessing the risk of failure for some time in the future", from the research paper titled 'Predicting Failure with Hidden Markov Models' by Felix Salfner.

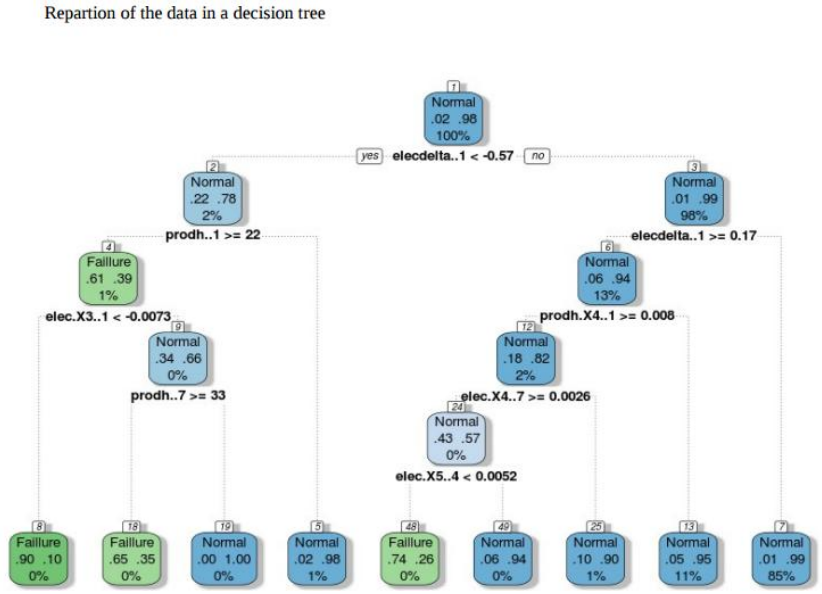
"The problem is to characterize as one of detecting rare events from a time series of noisy and non parametrically distributed attributes" from the research paper titled 'Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application' by Joseph F Murray, Gordon F Hughes, and Kenneth Kreutz-Delgado.

In the project, the classification probability of machine failure was the desired output, and accordingly different statistical models for classifier prediction were considered for two industrial machines (PM1 and PM2). Each statistical model was applied on the partitioned training and test data (70% training : 30% testing) as well as on time slices of the data. A similar approach was applied on re-sampled data.

IV. Results Obtained

Before applying any predictive model, a decision tree was first applied to observe how the data was split. Some variables were observed to be highly discriminant.

Decision Tree



Different statistical models were considered to predict machine failure. ROC curve was used as the criterion to compare the different models. This is an easy, simple, and understandable cmparison criteria. However, in the present case, we firstly considered a model's ability to predict machine failure (by checking the sensitivity and AUC). A complete list of the most significant variables for each model has been included in the Appendix section of this report.

1. Generalized Linear Model (GLM)

The following results were obtained by fitting the GLM model onto the data:

GLM Results	Randomly Split Data	Re-sampled Randomly Split Data	Time Sliced Data	Re-sampled Time Sliced Data
Optimal Prediction Threshold	0.01	0.01	0.01	0.01
Good Prediction Rate	96.64%	74.53%	95.64%	83.88%
AUC	59.44%	59.27%	58.26%	61.66%
Sensitivity	20.80%	43.40%	19.23%	38.50%
Specificity	98.10%	75.10%	97.28%	84.90%

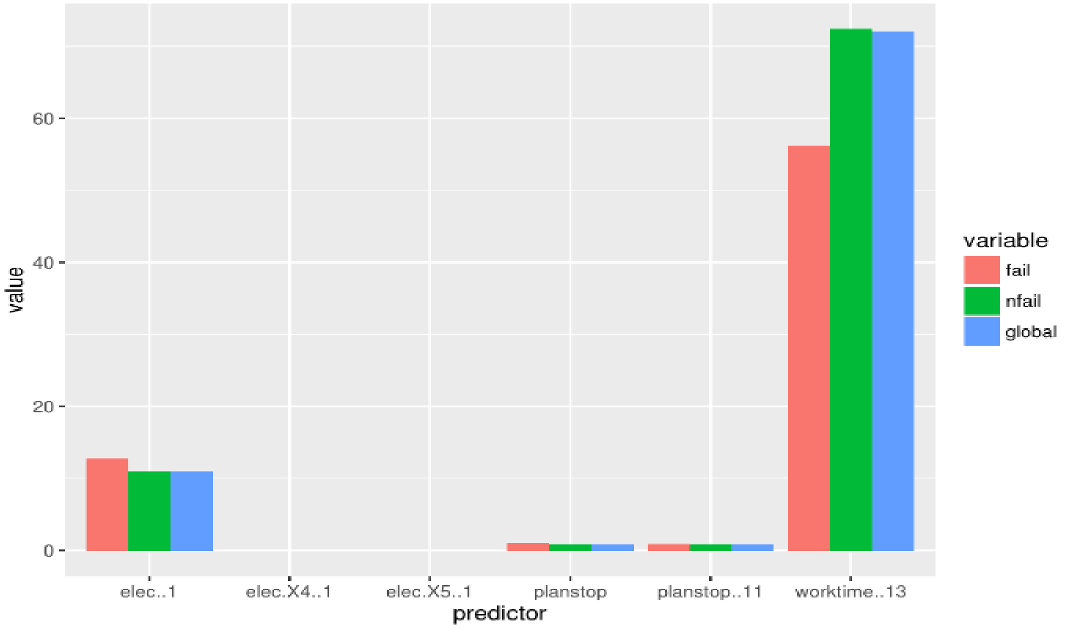
Comparison of the prediction accuracy of the randomly split dataset (with and without re-sampling) reveals that the AUC remains almost constant, while the sensitivity almost doubles when we use re-sampled data. A similar trend is observed for the time sliced data. Hence, we can infer that use of re-sampled data is better in predicting machine failure.

Considering the variables of importance in the case of both randomly split and time sliced data, we obtain the following most significant variables in our dataset:

Most Significant Variables	
Re-Sampled Randomly Split Data	**Re-Sampled Time Sliced Data**
Planstop	wtime..11
elec..x5.1	wtime..12
wday..12	ffail..8
elec..1	state..8
wday..11	ffail..11

We observe that in the case of GLM, there is no overlap among the 5 most significant variables in the two best model types.

Variation of Variables According to the State of the Machine (Re-Sampled Time Sliced Data)



Here, we observe that in the case of time sliced data, the electricity consumption an hour preceding an instance of machine failure is higher than that observed during normal machine operation. 'worktime..13' is also observed as a significant variable for GLM in Re-Sampled Time Sliced Data. The worktime 13 hours preceding an instance of machine failure is lower than that observed during normal machine operation.

2. Support Vector Machine (SVM) Model

The following results were obtained by fitting the SVM model onto the dataset:

SVM Results	Randomly Split Data	Re-sampled Randomly Split Data	Time Sliced Data	Re-sampled Time Sliced Data
Optimal Prediction Threshold	0.03	0.17	0.02	0.25
Good Prediction Rate	92.96%	86.14%	72.03%	66.88%
AUC	64.04%	64.26%	57.49%	64.27%
Sensitivity	33.96%	41.50%	42.30%	61.53%
Specificity	94.11%	87.01%	72.67%	67%

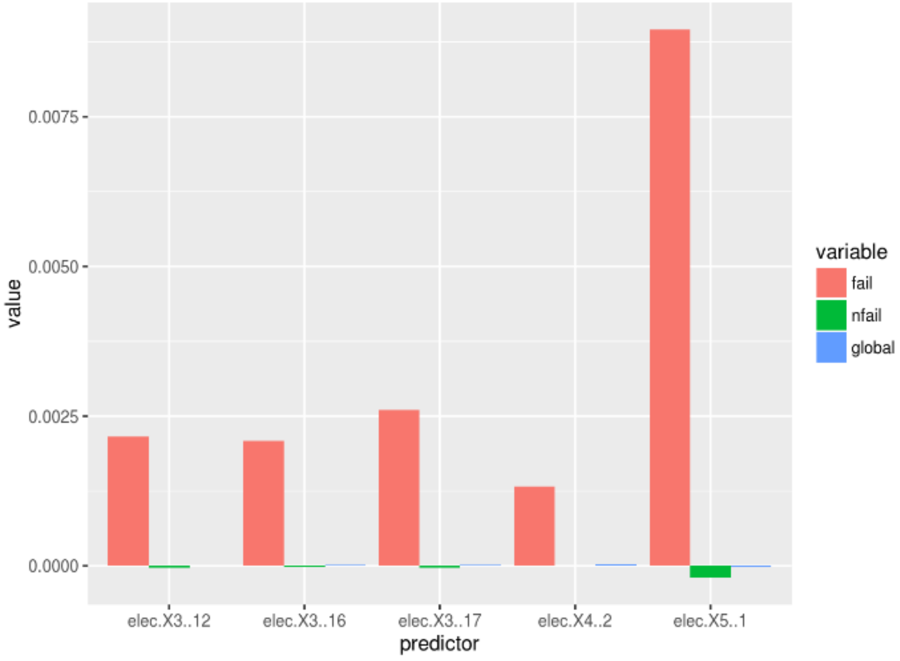
Now, we compare the prediction accuracy of the randomly split dataset (with and without re-sampling). We observe that the AUC remains almost constant, while the sensitivity slightly increases when we use re-sampled data. For the time sliced data, the AUC shows a slight increase with re-sampling, while the sensitivity shows a significant increase. Hence, we can infer that re-sampled data leads is better for predicting machine failure.

Considering the variable importance in the case of both randomly split and time sliced data, we obtain the following most significant variables in our dataset:

Most Significant Variables	
Re-Sampled Randomly Split Data	**Re-Sampled Time Sliced Data**
prodh.delta..24	elec.x3..16
elec.x5..1	elec.x3..12
elec.x4..1	elec.x3..17
elec.vect.p3..12	elec.x4..2
elec.vect.p2..12	elec.x5..1

We observe that the 5 most significant variables for both the models include transformed electricity consumption variables at different preceding instances of time, especially in just 1 hour and 12 hours before.

Variation of Variables According to the State of the Machine (Re-Sampled Time Sliced Data)



It clearly appears that all the top 5 most significant variables are transformation versions of the electricity consumption variable, and their values are much higher in the case of machine failure.

3. Treebag Model

The following results were obtained by fitting the Treebag model onto the data:

Treebag Results	Randomly Split Data	Re-sampled Randomly Split Data	Time Sliced Data	Re-sampled Time Sliced Data
Optimal Prediction Threshold	0.05	0.20	0.28	0.20
Good Prediction Rate	89.68%	75.00%	97.82%	73.00%
AUC	70.69%	72.46%	59.37%	71.16%
Sensitivity	50.90%	69.81%	19.23%	69.23%
Specificity	90.40%	75.10%	99.50%	73.08%

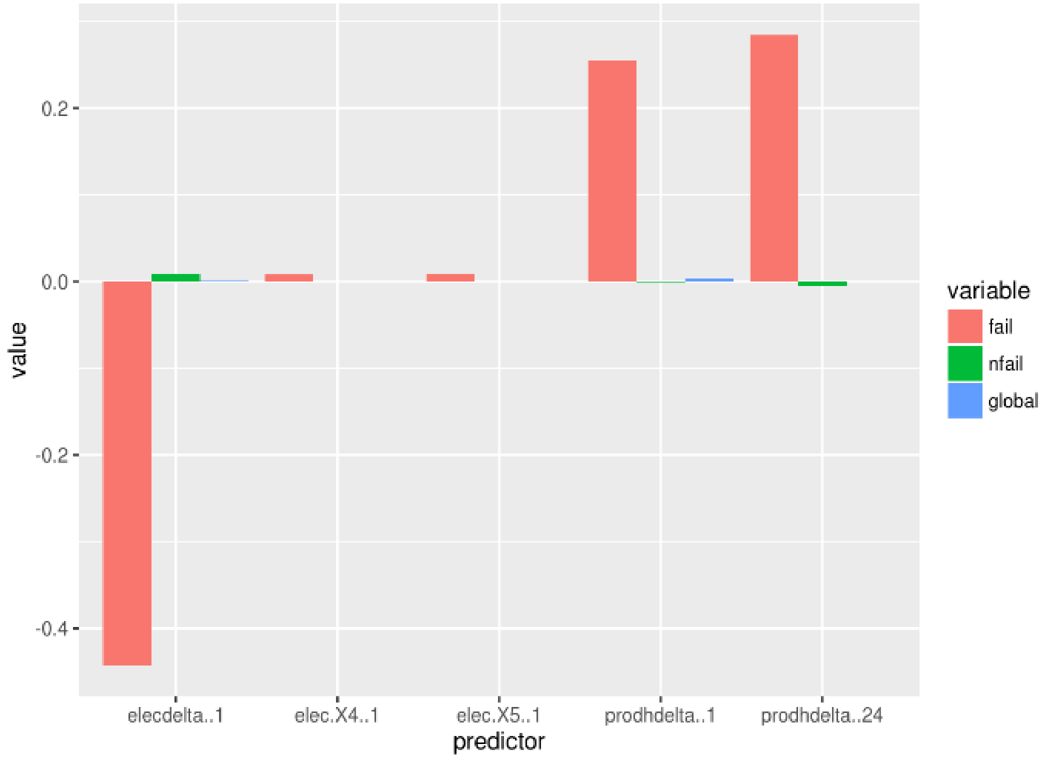
Upon comparison of the prediction accuracy of the randomly split dataset, with and without re-sampling, we observe that the AUC increases slightly, while the sensitivity shows a significant increase with re-sampling. For the time sliced data, the AUC increases by 20%, while the sensitivity increases by almost three times. Hence, we can infer that re-sampled data is better for predicting machine failure.

Considering the variable importance in the case of both randomly split and time sliced data, we obtain the following most significant variables in our dataset:

Most Significant Variables	
Re-Sampled Randomly Split Data	**Re-Sampled Time Sliced Data**
elecdelta..1	elecdelta..1
prodhdelta..24	prodhdelta..19
elec.x.5..1	elecdelta..5
elec.x.4..1	state..5
prodhelta..1	ffail..5

Here, we first observe that the variation of electricity an hour before is the first one in both the models. The variation of the production also appears at different hours each time.

Variation of Variables According to the State of the Machine (Randomly Split Data)



Here, we observe that the variation of electricity is higher one hour preceding the instance of machine failure. On the contrary, the value of production is lower than the normal 1 hour and 24 hours before. This pattern is the same that was revealed by the decision tree.

4. Stochastic Gradient Boosting Machine (GBM)

The following results were obtained by fitting GBM model onto the data:

GBM Results	Randomly Split Data	Re-sampled Randomly Split Data	Time Sliced Data	Re-sampled Time Sliced Data
Optimal Prediction Threshold	0.01	0.01	0.01	0.01
Good Prediction Rate	97.47%	86.14%	97.98%	87.99%
AUC	64.04%	70.74%	59.45%	69.40%
Sensitivity	28.30%	54.70%	19.20%	50.00%
Specificity	98.82%	86.80%	99.70%	88.88%

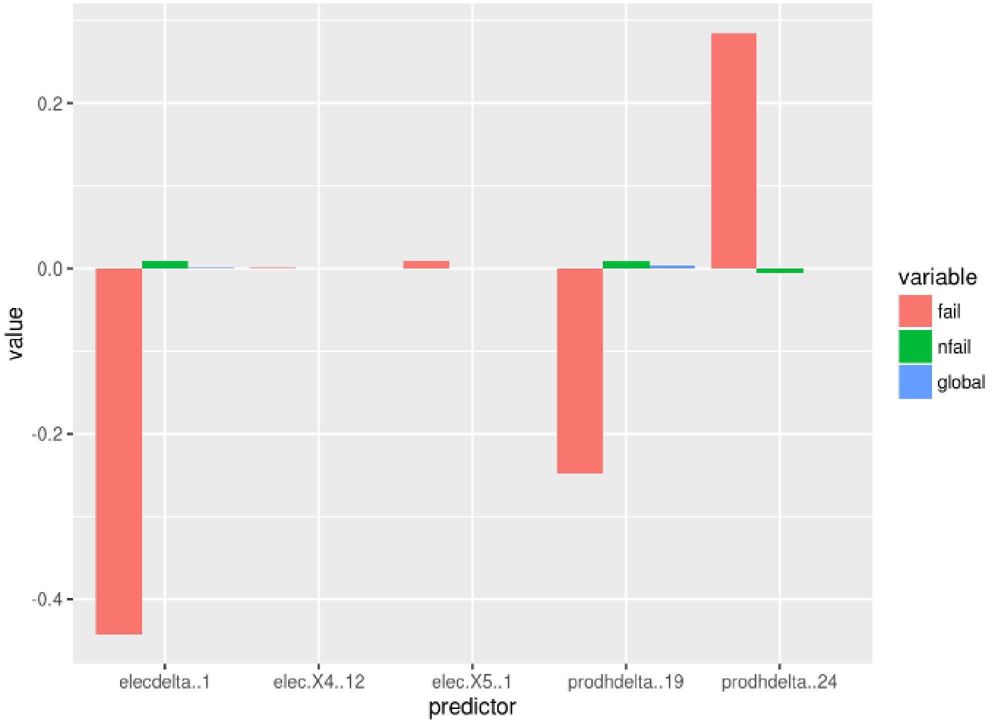
Now, we compare the prediction accuracy of the randomly split dataset, with and without re-sampling. We observe that the AUC shows an approximately 15% increase, while the sensitivity almost doubles when we use re-sampled data. For the time sliced data, the AUC increases by 10%, while the sensitivity increases by about 30% with re-sampling of the data. Hence, we can infer that re-sampled data is better for predicting machine failure, as was the case in earlier models.

Considering the variable importance in the case of both randomly split and time sliced data, we obtain the following most significant variables in our dataset:

Most Significant Variables	
Re-Sampled Randomly Split Data	**Re-Sampled Time Sliced Data**
elecdelta..1	elecdelta..1
prodhdelta..24	elecdelta..5
elec.x5..1	prodhdelta..19
prodhdelta..19	ffail..5
elec.x4..12	elec.x5..1

Here, we observe that the variables elec.delta..1, elec.x5..1, and prodhdelta..19 are significant for both the models.

Variation of Variables According to the State of the Machine (Randomly Split Data)



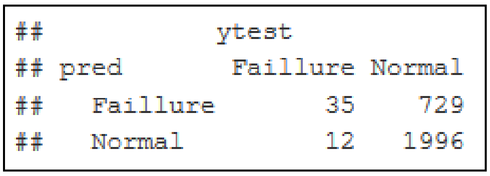
We observe that the variation in electricity 1 hour before is higher than that at the instance of machine failure. The variation in hourly production 19 hours before the instance of machine failure is higher and that 24 hours before is lower than that at the instance of machine failure.

5. Model Comparison

Best Model with Randomly Split Data

Models	GLM	SVM	Treebag	GBM
AUC	59.27%	64.26%	72.46%	70.74%
Sensitivity	43.40%	41.50%	69.81%	54.70%

Here, we observe that Treebag is the best model based on both the AUC and Sensitivity creteria. Its confusion matrix is as follows:

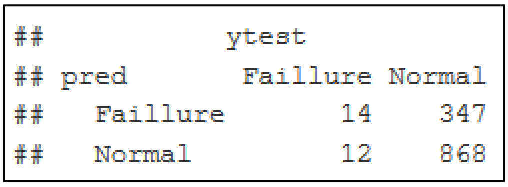


```
##            ytest
## pred        Faillure Normal
##    Failllure      35    729
##    Normal         12   1996
```

Best Model with Time Sliced Data

Models	GLM	SVM	Treebag	GBM
AUC	61.66%	64.27%	71.16%	69.40%
Sensitivity	38.50%	61.53%	69.23%	50.00%

Here also, Treebag is the best model based on both the AUC and the Sensitivity creteria. Is confusion matrix is as follows:

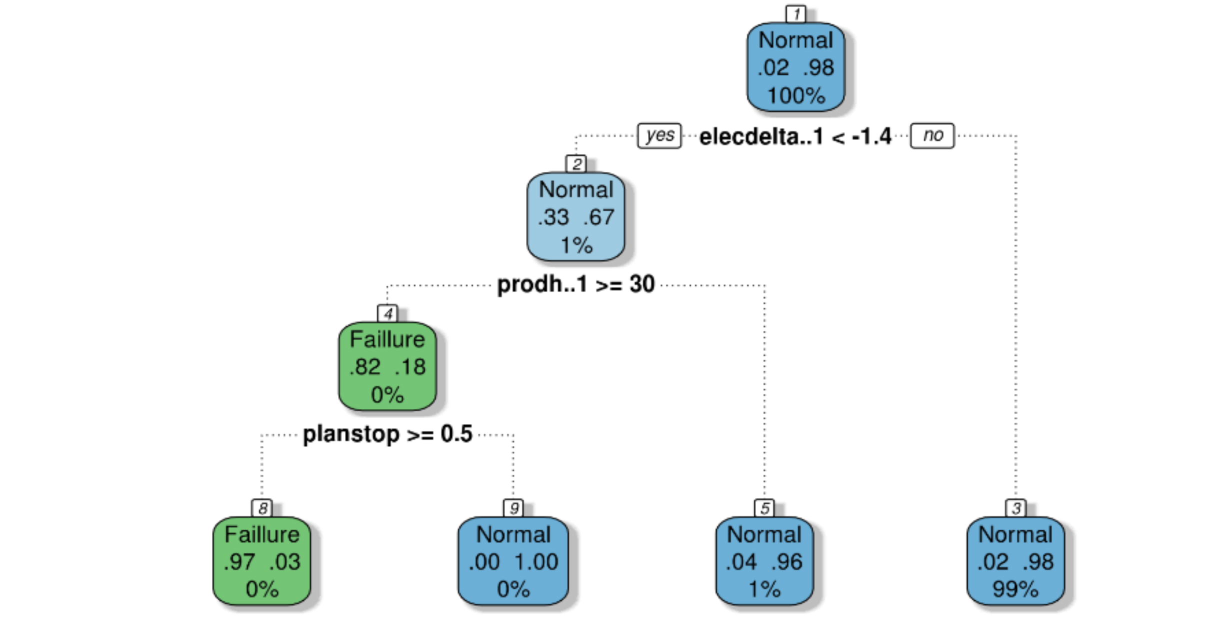


```
##            ytest
## pred        Faillure Normal
##    Faillure       14    347
##    Normal         12    868
```

6. Machine 2

The aforementioned statistical analysis procedure was also applied to the datasets obtained from Machine 2. It was observed that the most discriminant variable was the variation of electricity. Model fitting led to the conclusion that Treebag model fitted to Smote data was the best model for prediction machine failure. Moreover, the Treebag model provided us with a higher accuracy of machine failure prediction in the case of Machine 2, when compared with Machine 1.

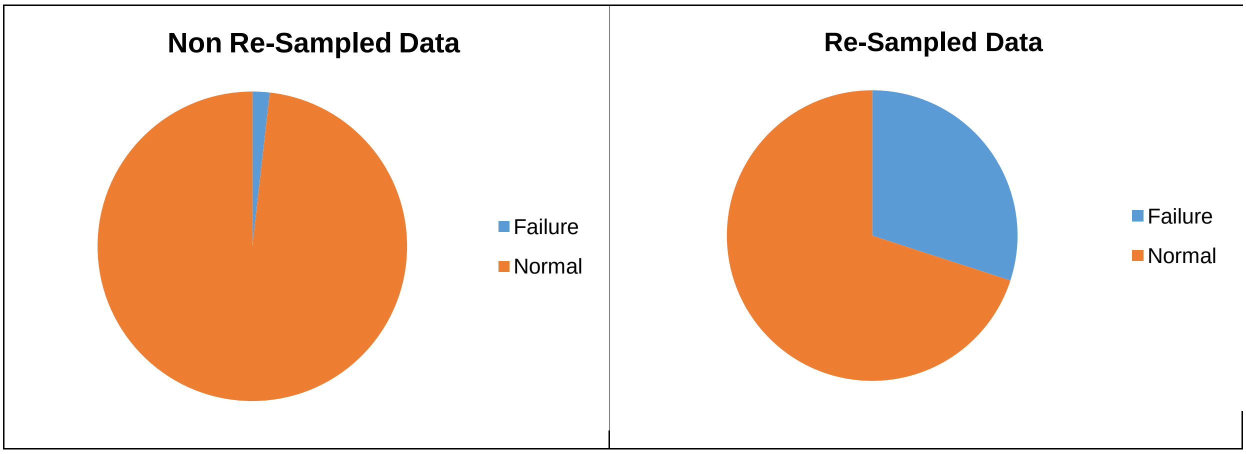
Decision Tree for Machine 2



V. Conclusion

In the case of machine failure, electricity consumption and production quantity are higher than that during normal machine operation, and this variable or its transformation is discrimanant in all the models considered.

The method of super resampling of data was used, and this enabled in better prediction of the results. The proportion of the failure and normal state for each of the datasets used has been displayed in the figure below.



Considering the value of sensitivity obtained as the criteria for model selection, we can conclude that the Treebag model is the best statistical model for predicting machine failure (to address our problem statement). However, we need to keep in mind that globally the model predicts at a rate of 75% and has an Area Under the Curve of 72.46%.

All different models were tuned on a cross validation of 2. The result could be improved by tuning the parameters of the models. Limitations posed by the limited number of variables could be improved by the availability of additional variables (or additional data).

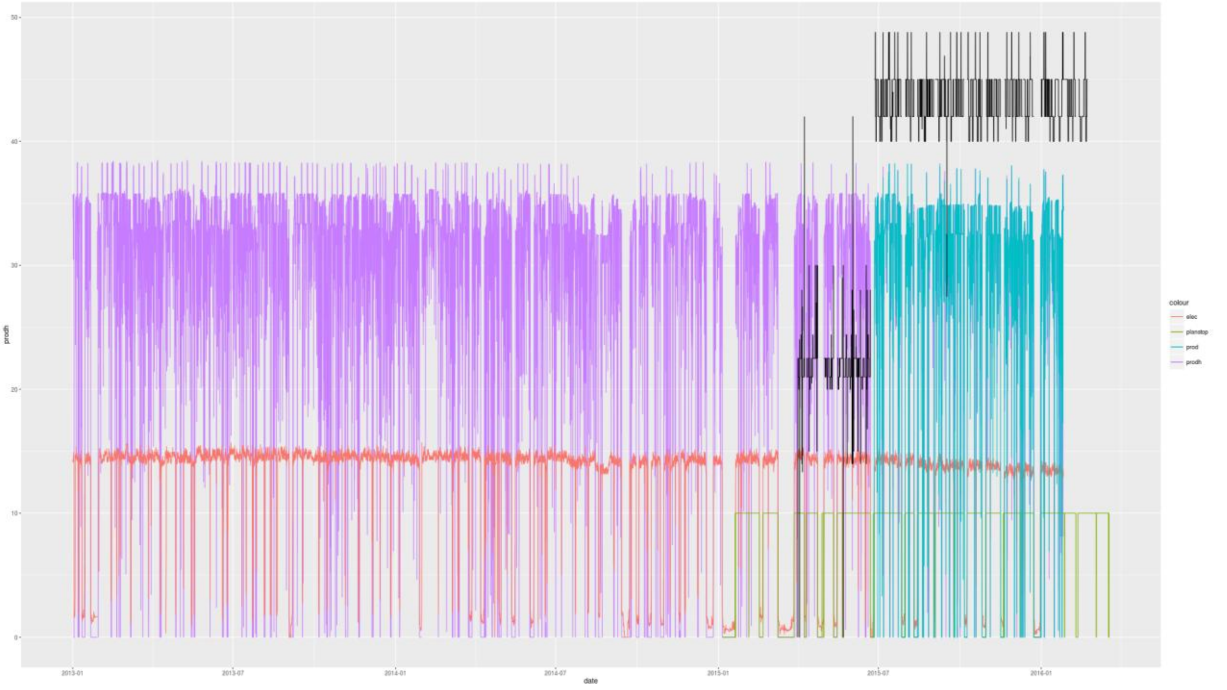
Further, certain additional statistical models (such as Hidden Markov Model) were also applied to the data. However, a successful implementation could not be achieved.

VI. References

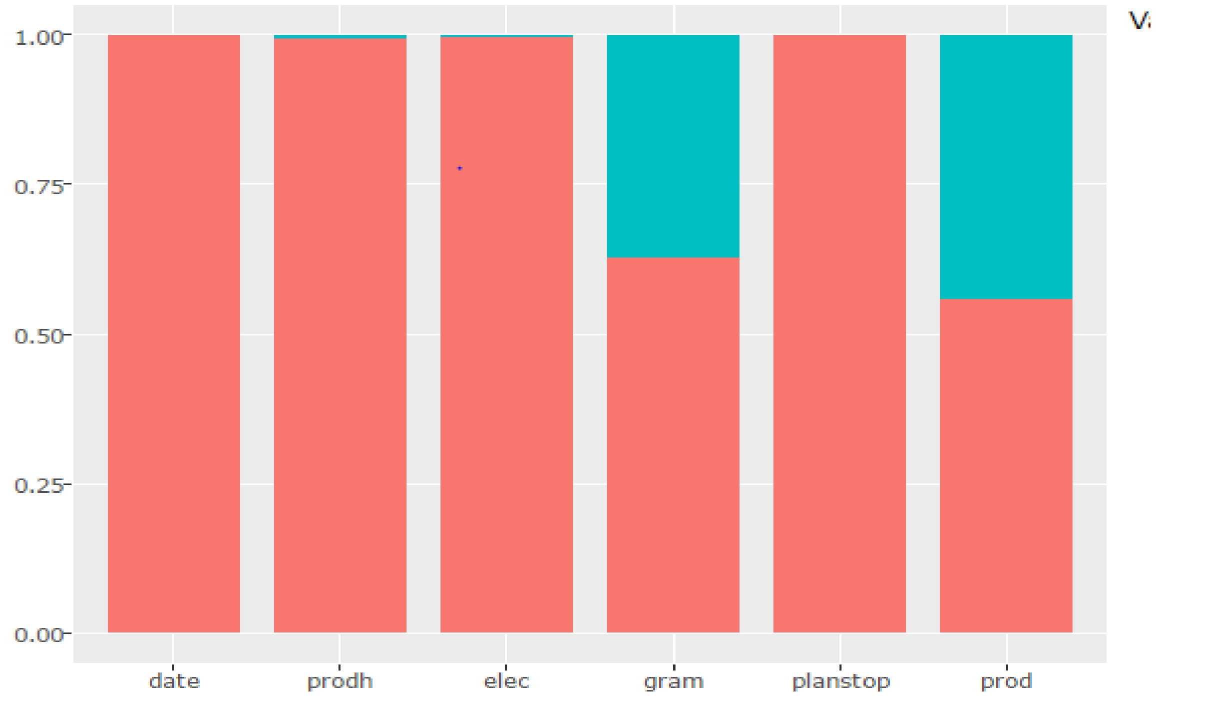
- 'Learning to Predict Rare Events in Categorical Time-Series Data', by Gary M. Weiss, Fordham University, USA.
- 'Time Series Analysis and Its Applications: With R Examples'
- 'SMOTE' - Supersampling Rare Events in R, (http://amunategui.github.io/smote/)
- 'Modèle Markov Caché', (https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_Markov_cach%C3%A9)
- 'Hidden Markov Models', (http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter10.html)
- 'Preliminaries: Mixtures and Markov Chains', (https://ayorho.files.wordpress.com/2011/05/chapter1.pdf)
- 'Neural Networks for Time Series Prediction', (https://www.cs.cmu.edu/afs/cs/academic/class/15782-f06/slides/timeseries.pdf)
- 'Hidden Markov Model: Identifying Changing Market Conditions', (https://inovancetech.com/hmm-tutorial-1.html)
- 'Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application', (http://jmlr.csail.mit.edu/papers/volume6/murray05a/murray05a.pdf)
- 'R Caret Package', (http://topepo.github.io/caret)

VII. Appendix

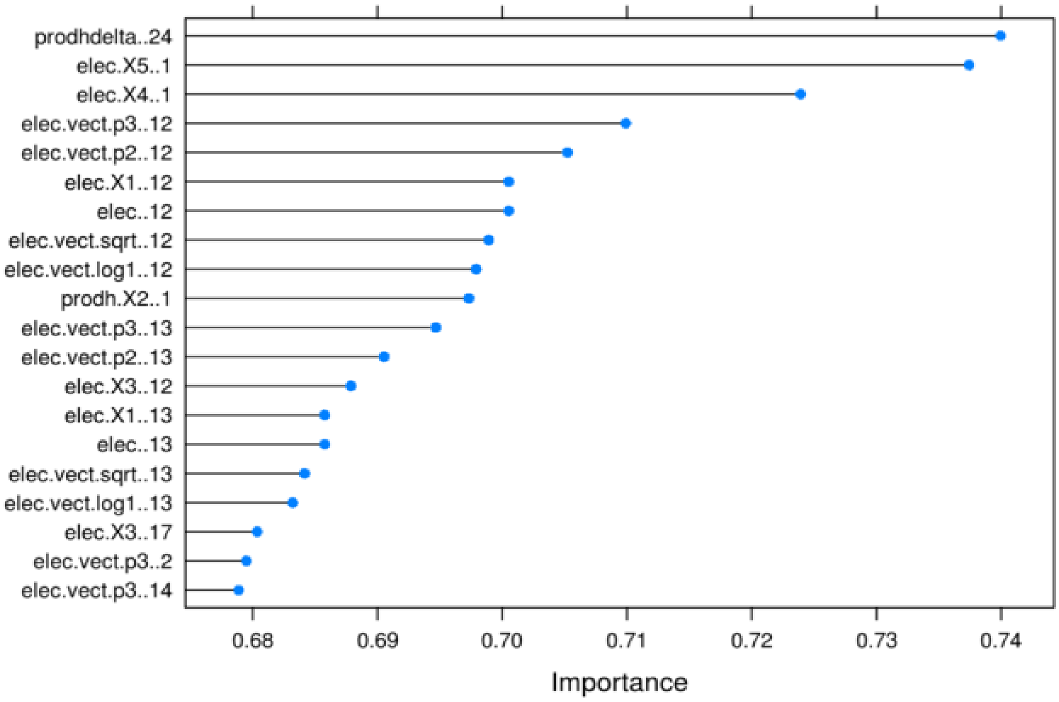
Plot of Sensor Values and Planning Stops



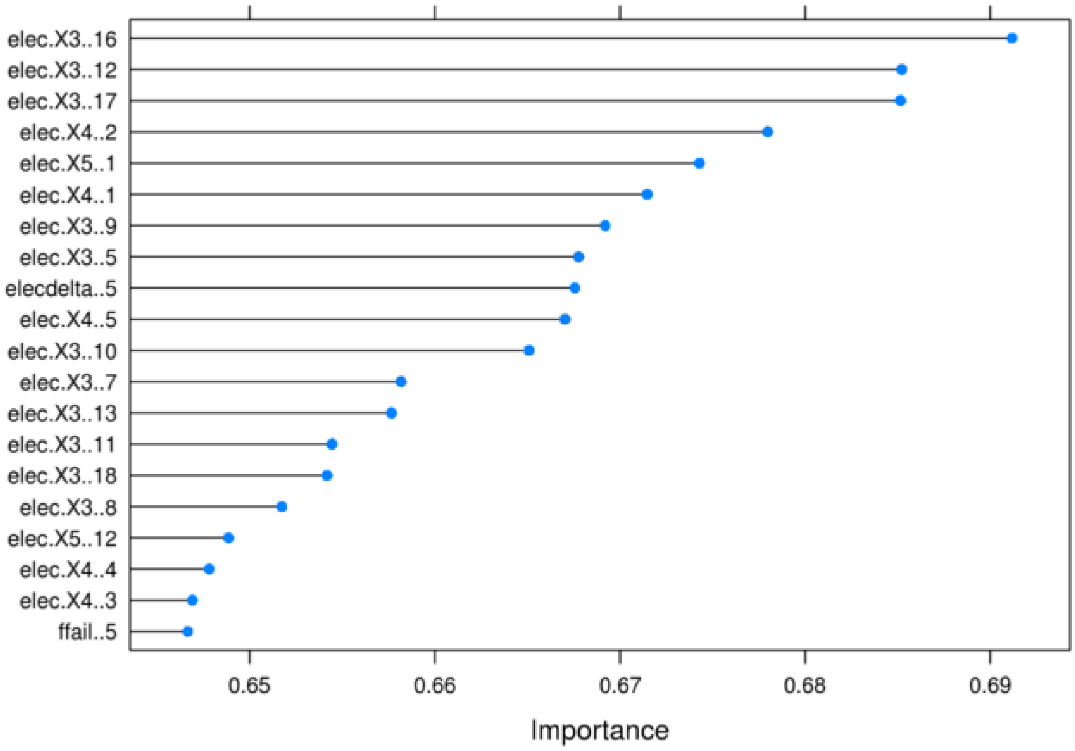
Plot of the Proportion of Missing Values



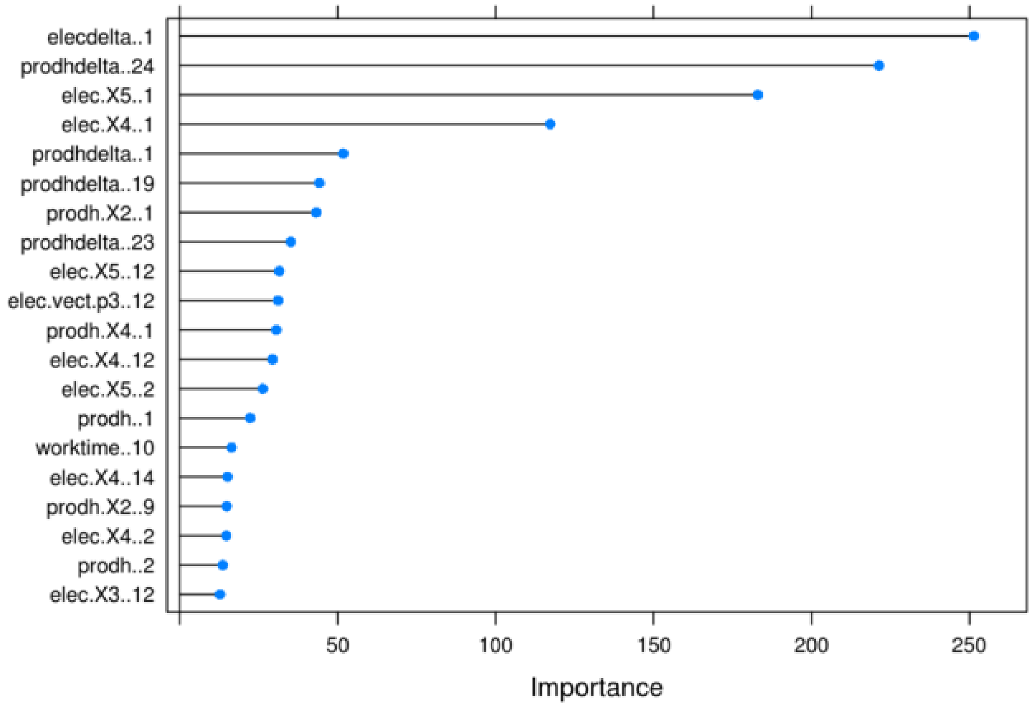
Variables of Importance for Bayesian Model with Re-Sampled Randomly Split Data



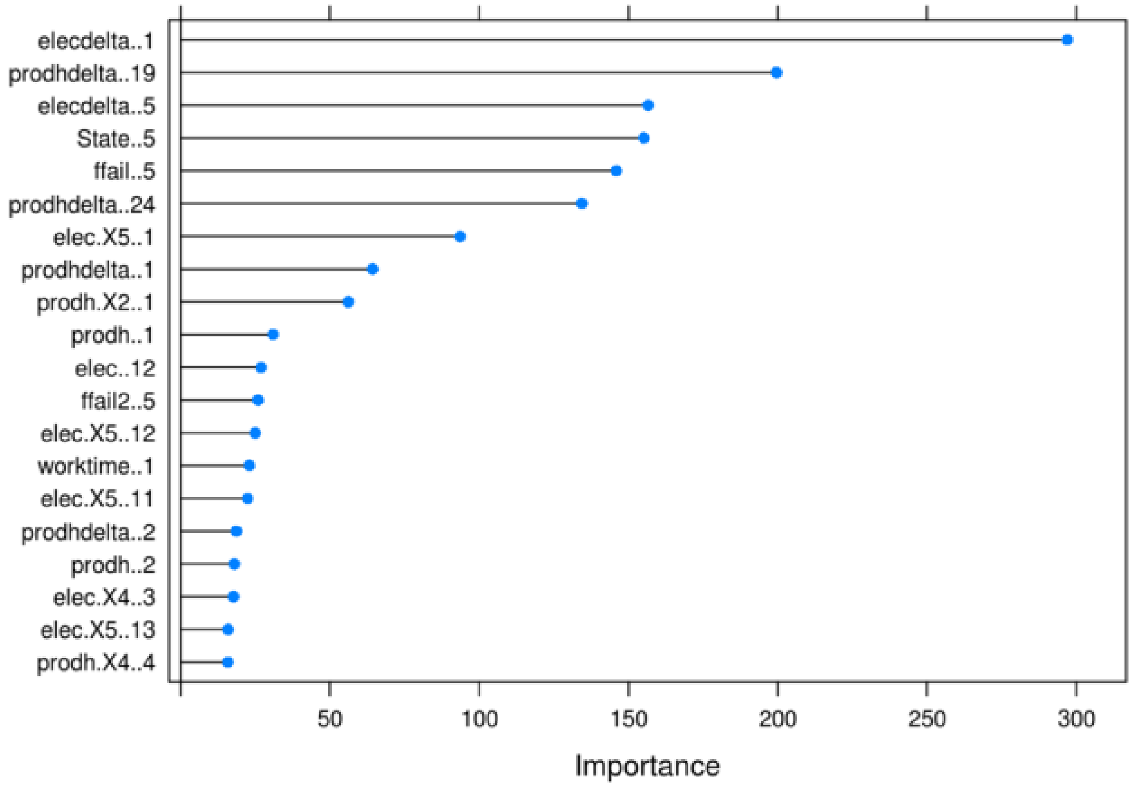
Variables of Importance for Bayesian Model with Re-Sampled Time Sliced Data



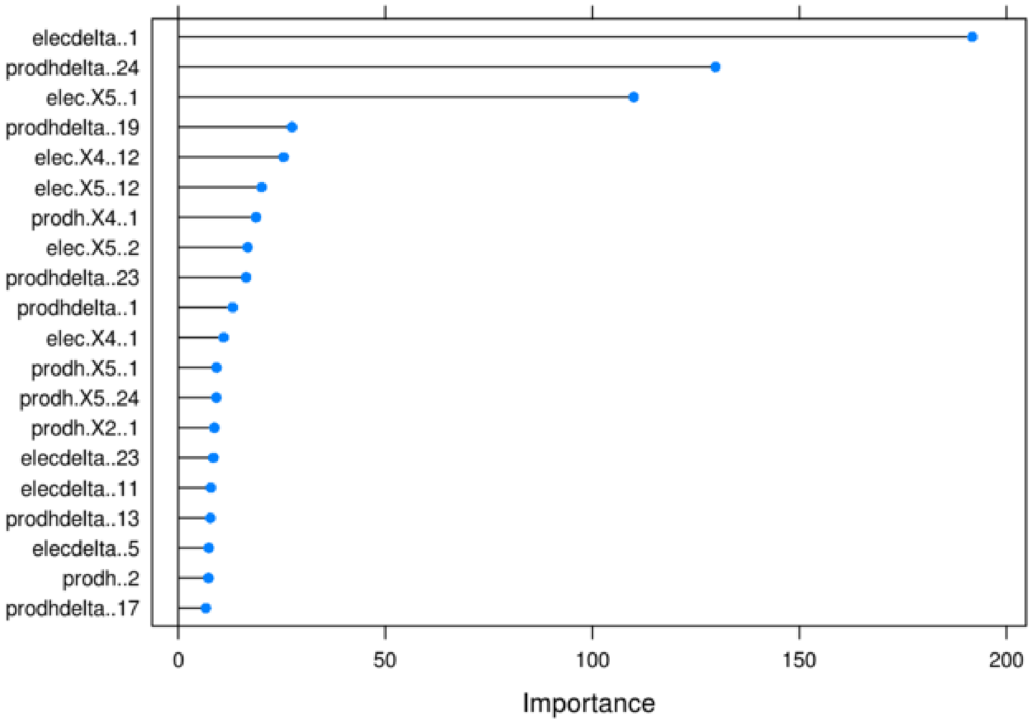
Variables of Importance for Treebag Model with Re-Sampled Randomly Split Data



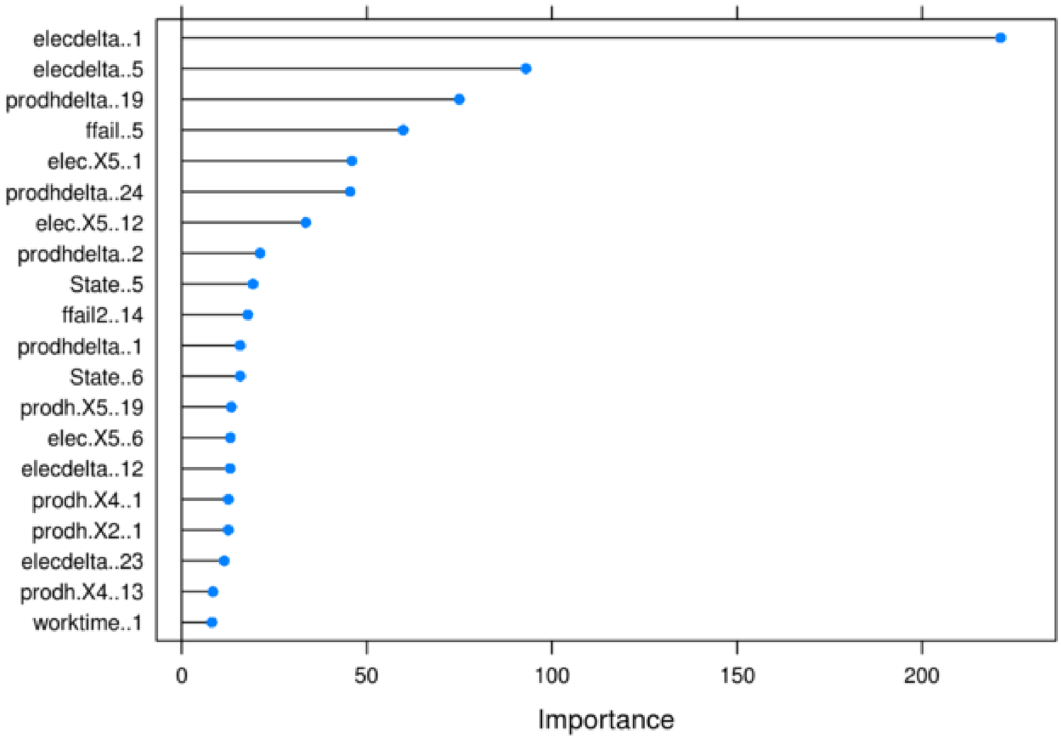
Variables of Importance for Treebag Model with Re-Sampled Time Sliced Data



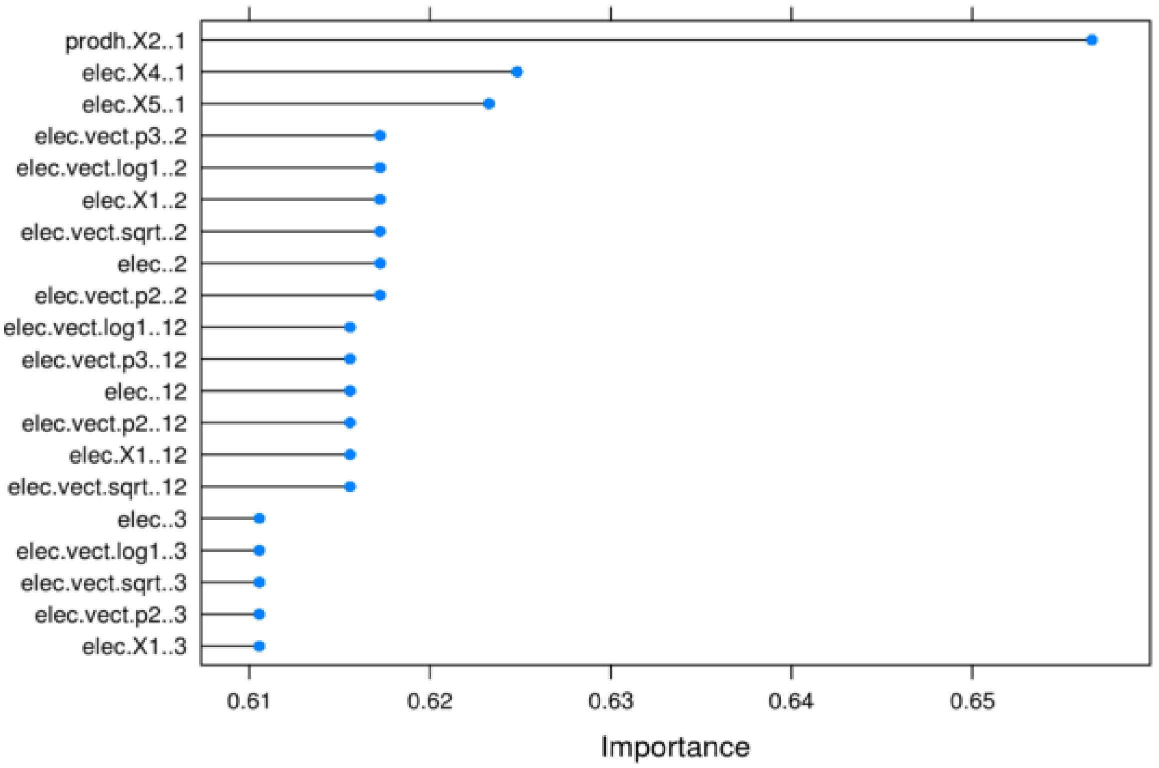
Variables of Importance for Stochastic GBM Model with Re-Sampled Randomly Split Data



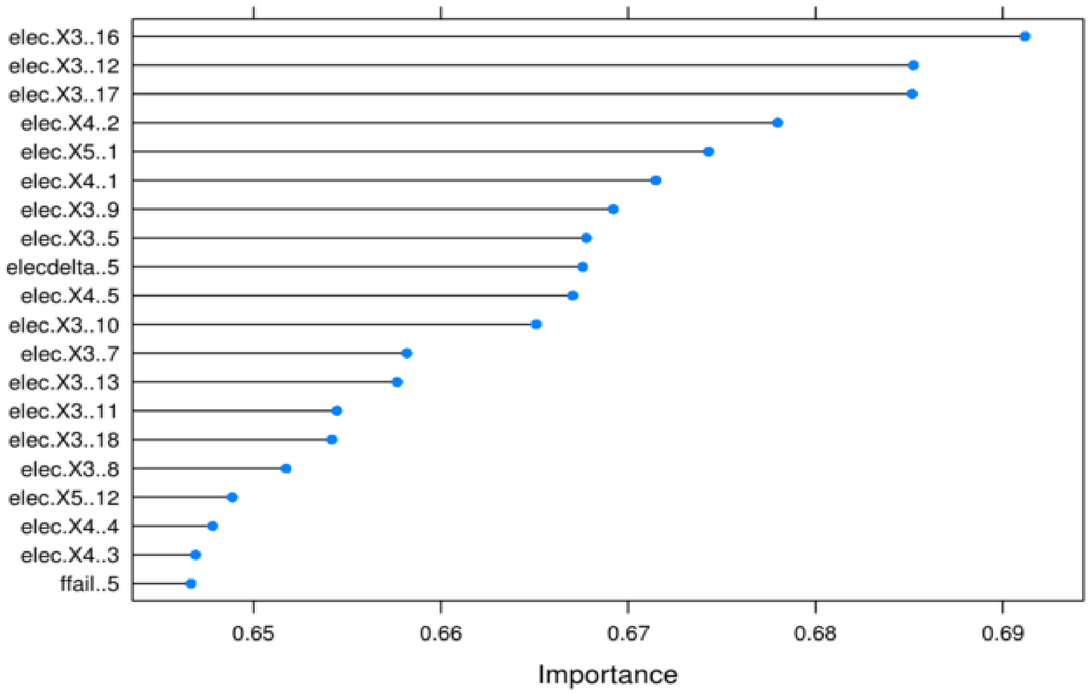
Variables of Importance for Stochastic GBM Model with Re-Sampled Time Sliced Data



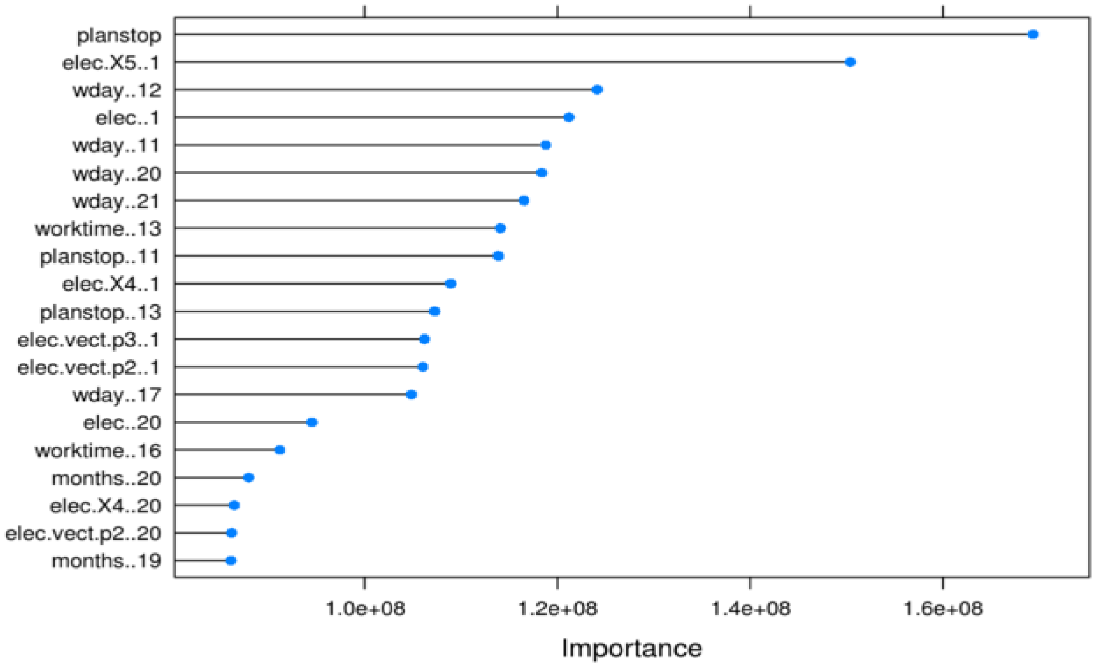
Variables of Importance for SVM Model with Re-Sampled Randomly Split Data



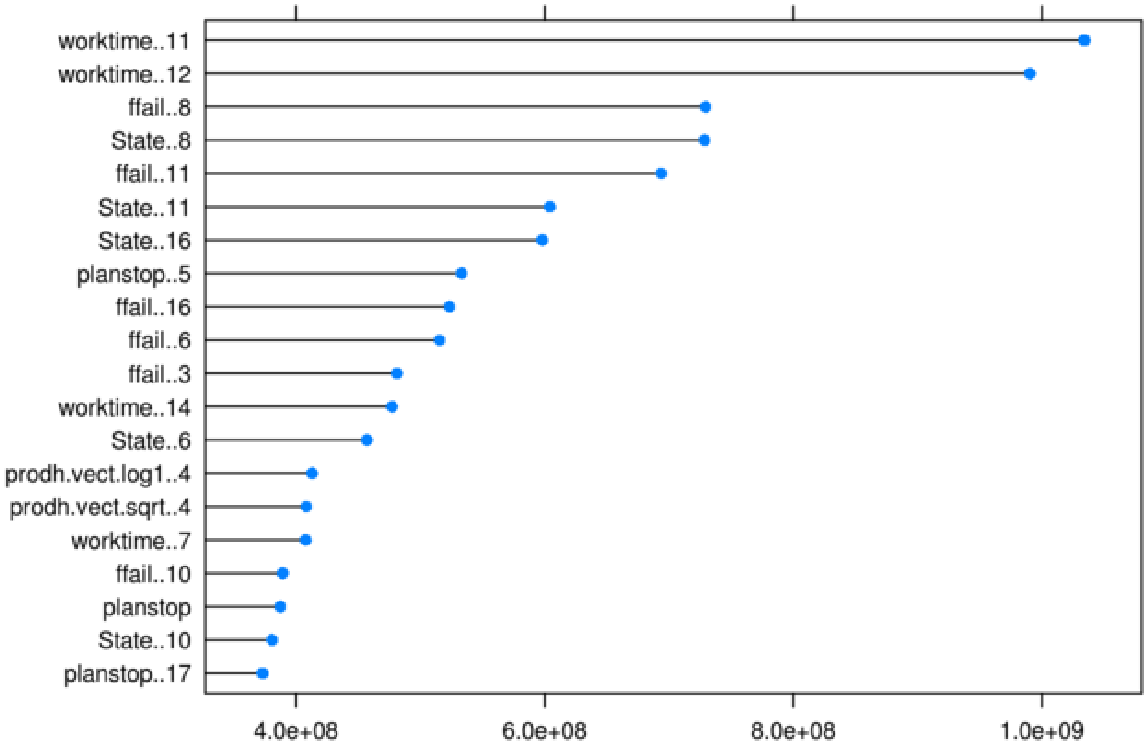
Variables of Importance for SVM Model with Re-Sampled Time Sliced Data



Variables of Importance for GLM with Re-Sampled Randomly Split Data



Variables of Importance for GLM with Re-Sampled Time Sliced Data


worktime..11
worktime..12
ffail..8
State..8
ffail..11
State..11
State..16
planstop..5
ffail..16
ffail..6
ffail..3
worktime..14
State..6
prodh.vect.log1..4
prodh.vect.sqrt..4
worktime..7
ffail..10
planstop
State..10
planstop..17
4.0e+08
6.0e+08
8.0e+08
1.0e+09