RESULTS AND VISUALISATION

1) Dimensionality Reduction

Figure 1 shows the cumulative variance captured by the number of the principal components in this analysis. ***Figure 2*** shows the comparison between the reconstructed dataset using principal components (black) and the real observations (red) with 1; 75; and 150 principal components. The higher number of principal components perform better to reconstruct the real dataset (***Figure 2***) as the variances are captured more (***Figure 1***). However, the increasing number of principal components will increase the computing requirements and not necessarily lead to better classification performance. Thus, this essay will inspect the performance of the classification algorithms using three distinct value of principal components, and those are 1; 75; and 150.

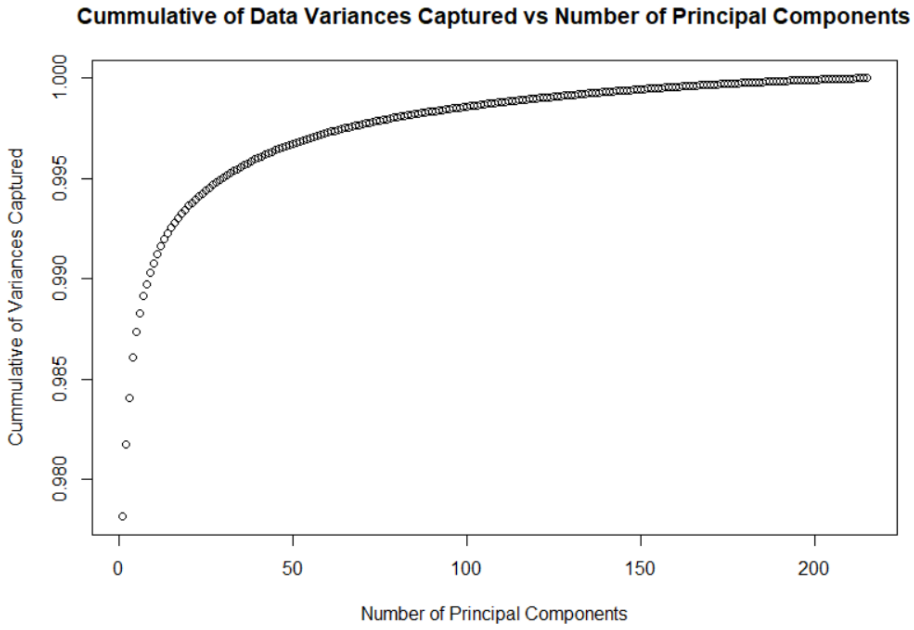
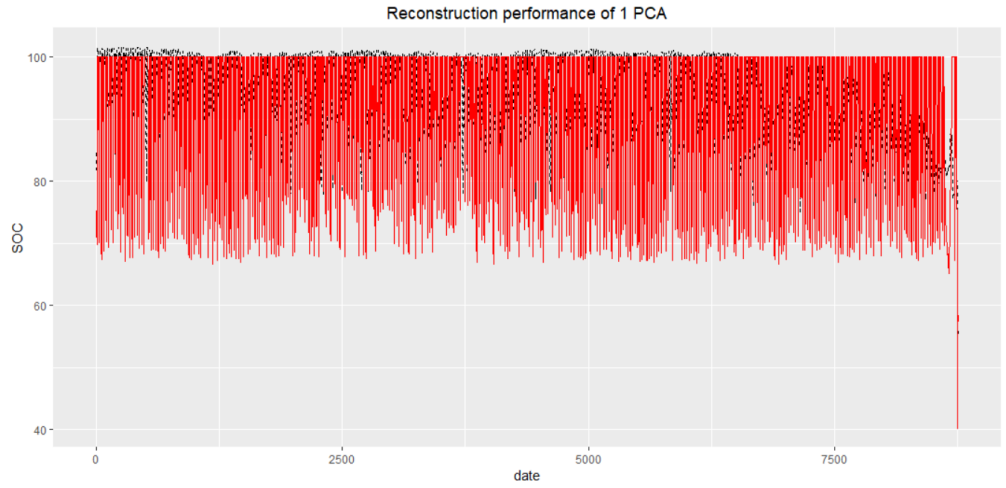
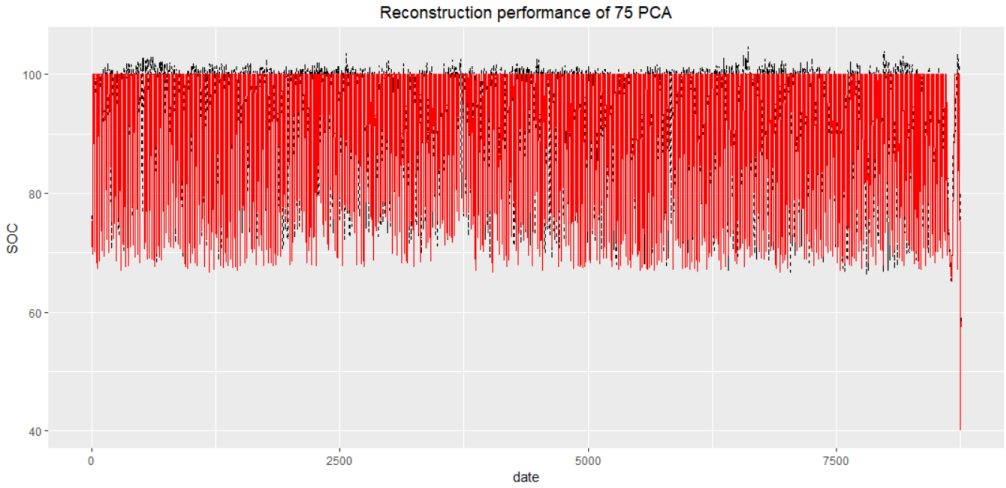


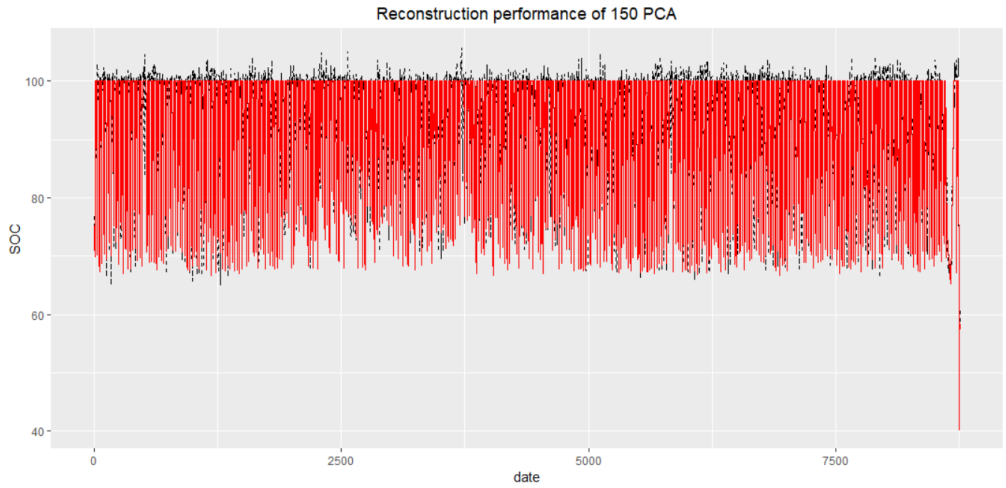
Figure 1. *Data variance captured by the principal components*



(a)



(b)



(c)

Figure 2. *Comparison of reconstructed data using PCA (black) and the real observations (red) with 1* ***(a)****; 75* ***(b)*** *and 150 (****c****) principal components*

2) Result

Three measures are used to evaluate the performance of machine learning algorithms, as follows,

- Accuracy calculates the portion of failed and passed batteries that the model predicted correctly.

$$\text{Accuracy} = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}}$$

- True positive rate or precision calculates the portion of failed batteries that the model predicted correctly.

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

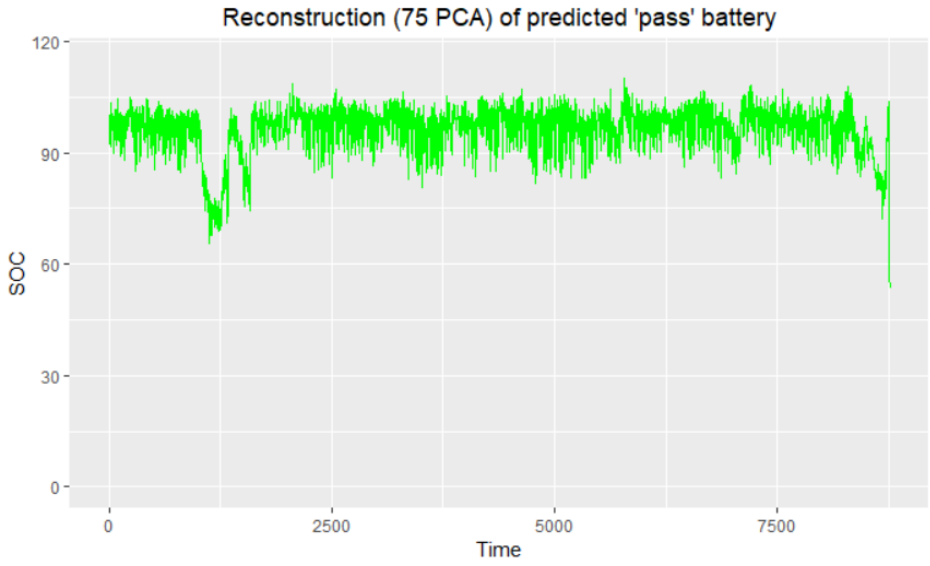
- False positive rate measures the portion of predicted failed batteries that do not actually fail.

$$\text{False positive rate (FPR)} = \frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

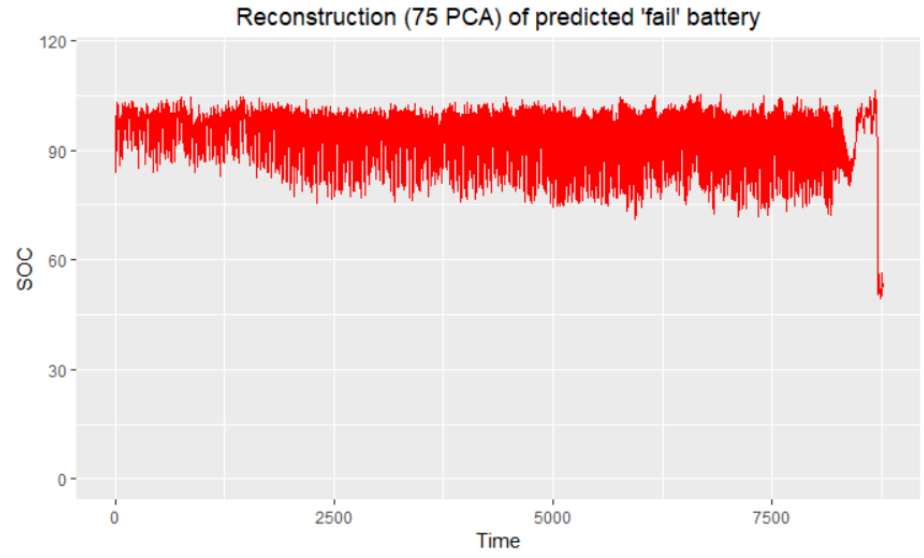
With high true positive rate, it means that the model can predict the failed batteries accurately. Whereas, low false positive rate means the model learns how to separate the fail and pass batteries well.

The result of the RF, SVM and GB classifiers in this analysis shows that SVM with the non-linear kernel is the best classifier in terms of the accuracy to predict which batteries are 'pass' or 'fail' on the battery failure test conducted by BBOXX with 86.1% accuracy (***Table 1***). However, the result poses 100% FPR which means all the actual 'pass' battery is predicted as 'fail' battery, which is not preferable because it would make unnecessary maintenance cost due to the assignment of 'fail' battery to the actually 'pass' battery.

Figure 3 shows the difference of a random predicted 'pass' and 'fail' battery on the test set from SVM algorithms and shows that the 'pass' battery visually differs from the 'fail' battery by seldom reach SOC below 90%. This linearity property might be the reason why the linear kernel is performing better (in terms of FPR) to predict the pass battery compared to the non-linear kernel.



(a)



(b)

Figure 3. *Comparison between a random SVM classification for (a) pass battery and (b) fail battery*

Table 1. *Result of Random Forest, Gradient Boosting and Support Vector Machine (SVM) on test set*

Classification Algorithms		Number of Principal Components		
		1	**75**	**150**
Random Forest	**Accuracy**	0.744	0.791	0.767
	Precision	0.775	0.786	0.767
	FPR	0.9	0.9	1
Gradient Boosting	**Accuracy**	0.814	0.814	0.814
	Precision	0.814	0.829	0.829
	FPR	1	0.875	0.875
SVM - Linear Kernel	**Accuracy**	0.861	0.767	0.767
	Precision	0.861	0.865	0.865
	FPR	1	0.833	0.833
SVM - Non Linear Kernel (Gaussian, Sigmoid, and Polynomials)	**Accuracy**	0.861	0.861	0.861
	Precision	0.861	0.861	0.861
	FPR	1	1	1