**MoM AML Week 1**

**Summary of preprocessing so far**

We have in total now four datasets that we can run our models on:

1. Reduced Dataset with all the features with zero variance taken out
2. Dataset (1) + Normalised the dataset by rescaling every row in the matrix and then turning each row of the matrix into length one - used Normalizer()
3. Dataset (1) + turned the dataset into more gaussian like with zero mean and standard deviation of one – used Powertransform()
4. Dataset (1) + rescaling each feature between 0 and 1 – used MinMaxScaler()

Additional comments:

1. Alberto and Ali tried the pipeline and found that it causes a black box effect and will need to be implemented at the end rather than at the beginning.

**Next stage – Feature Selection/Extraction**

Ideas regarding Feature selection:

1. PCA
2. Random Forest
3. Naïve Bayes
4. Mutual Information (Entropy)
5. Extra Trees
6. Chi-square

As discussed I will also be attempting Stack Auto Encoder(SAE) and when I get it working I will make sure to upload to GitHub.

**Deadlines**

Alberto – 1st draft of preprocessing report for **21st November**

Dave – 1st attempt with a model and any delegation **24th November**

Adam – Attempted SAE & upload to GitHub **24th November**

Andy – Have 1st draft of our team plan report by **28th November**

All of us – have a draft of report for your stage by **12th December**