Hello everyone, I am we are ST4 Black Swan – Group A, and today we’ll walk you through our Feature Analysis workflow.

Our study focuses on three key tasks:

Correlation Matrix – to identify and remove redundant features;

Systematic Bias Impact – to measure how feature distributions shift under simulated uncertainties;

Feature Importance – to understand which features are most helpful for our final classification task.

Here’s a quick overview of our project timeline: from Mon to Fri, we have worked on a serie of tasks, from ... And on Wednesday we initially estimated the feature importance by using the results of BDT model, but it didn’t work, so we changed to the NN model to finish our estimation.

The first task that we did is the calculation of correlation matrix to identify and remove the redundant features. Before the calculation, it is necessary to consider the different natures of features, the primary ones, which are measured directly from the experiments, and the derived ones, which are calculated from the primary ones. Then we separate the features according to the number of jets, 0, 1 and 2 or more. And finally we calculated the correlation matrix to identify the features that are highly correlated, that means they don’t add the information. For example, here we have the corre

The second task that we did is to check the impact of the systematic bias on the features to measure how feature distributions shift under simulated uncertainties. When applying bias, we considered two version of biased data affected by each biased parameter as alpha +1sigma and alpha -1sigma, where ... Then we calculated the chisquare errors of each feature under each bias. The picture below is part of the chisquare errors. After that we add all the chisquare errors for each parameter to get the score of each feature. Here we have the total scores of each feature, and the higher ones mean that the correspond feature are more sensitive to the bias, which may not be used later.

The last task that we did is to measure the importance of each feature to understand which features are most helpful to predict the outcome. We have used the results of NN and our propose are to ... Here is the ranking of the feature importance. But in practice we combine the feature importance with the total score of the features, to make a balance between the predictive power and stability. As the right graph shows, the overall less important features are the ones lower and to the right, so the upper left features will be taken into consideration.

For further works, we can apply the SHAP method to quantify each feature’s contribution to individual predictions, and also the PCA method to reduce the dimensionality of our model. Then I pass the