SUMMARY

How does the classification performance compare across the 7 DR/FS methods?

Classification performance is best when dimensionality reduction has been done using LDA. It gives 100 %percent performance on test set. It also gives 100 % performance when cross validation is done. This shows the robustness of the technique. Amongst the feature selection techniques XGBoost feature selector has performed best with 15 features. The accuracy reached a value of 90% when 15 features were selected. However on the second classification dataset, no distinct performance improvement was seen across all the 7 DR/FS methods. All of them performed equally poor in classifying the minority class.

How does the regression performance compare across the 7 DR/FS methods?

The best performance has been achieved when features have been selected using random forest feature selection. The Mean Absolute Error achieved is 7.7 which is the lowest amongst all feature selection and dimensionality reduction techniques. Due to computational constraints t-SNE crashed while running hence has not been considered when concluding results.

What is the effect of increasing or decreasing the total number of your desired features on the classification and regression performance?

1. Classification: In dimensionality reduction methods(PCA), increasing the number of features required does not necessarily increase performance accuracy. Graphs of cross validation plotted support this notion as 10 principal components show a higher accuracy as compared to 15 principal components. With respect to feature selection techniques, it can be said that increasing number of features results in increase in performance. XGBoost feature selection gave best classification accuracy (90%) when 15 number of features were selected on dataset 1. However, dataset 2 proved too challenging for all algorithms and they failed to correctly classify the minority irrespective of increase or decrease in the number of features.

2. Regression: For regression, it can be said that increasing number of features results in higher performance irrespective of the nature of the methodology (DR or FS). All techniques used showed best performance when the number of features was 15.