Following on from logistic regression, decision tree models were another avenue that was explored. Firstly, owing to the tabular nature of the data, they were determined as a particularly amenable approach. Decision trees have demonstrated their efficacy in this domain, even beating the likes of more advanced models such as neural networks. They are also suitable in this case because they can cope with the presence of superfluous and unhelpful features, which are a concern due to the substantial number of features present (95). There are also mechanisms such as weighting that can be used to address the class imbalance that is present in our data. These considerations led to the choice of the following 3 decision tree models: random forests, XGBoost and AdaBoost. These are all decision tree ensemble models that should lead to more robust outcomes.

Random forests operate on the premise of creating a forest of multiple decision trees that are trained on randomly selected data subsets and feature subsets at each splitting rule.

XGBoost operate by fitting successive decision trees that correct the residuals of the previously fitted trees.

AdaBoost operates by fitting successive decision trees that focus on the previous misclassified or hard-to-classify examples.

Training was completed with 5-fold cross validation that focused on optimising ROC, following from the same motivations of logistic regression. Hyperparameters were set and tuned with the aid of this cross validation. Notably, this involved structural and regularisation hyperparameters in order to combat overfitting. For example, with the numbers of trees/rounds, tree maximum depth and depth regularisation.

The ROC plots of the trained models, based on the probability threshold for positive predictions, are depicted below. Thus, the desired model can be chosen along the ROC curve based on the trade-off between sensitivity and specificity. The AUC results from these ROC curves are depicted below. Also, the confusion matrices of the models, at the optimal point on the ROC curve, are also depicted below.

From the results, we can see that the random forest model performed the best with an AUC of \_, which outperforms the prior logistic regression model and positions it as the optimal solution to the classification task that was developed. This is in line with expectations as decision trees introduce feature interactions and non-linearity. This is evident as feature splitting rules along branches are conditioned on the ancestor feature splitting rules. Thus, decision trees are able to capture greater complexity compared to linear models that assign coefficients to individual features and therefore don’t account for their interactions.

Another point of interest is that the application of decision tree models was better across the board on the balanced datasets compared to their application on the original unbalanced datasets with weighting. This highlights the efficacy of the up-sampling methods that were utilised to address the class imbalance.