
Random forests were tuned over a grid of four parameter variables: mtry, ntree, max depth and nBins. Mtry is the number of variables randomly sampled to be tested for the best split at each split point in the tree, while ntree is the total number of trees grown. Max depth is the maximum depth to grow the trees, shallower trees (lower numbers) lead to simpler models and can control over fitting. Deeper trees will allow for more interactions between variables to be found. To find the optimal nBins, H20 builds a histogram to assess the splitting points, a histogram of this many bins is built and the best point selected as the split. As with max depth less bins (lower numbers) leads to a simpler model and can control over fitting.

For stochastic gradient boosting (GBM), max depth and nBins are as per random forest. The idea behind a GBM is for each successive tree is built from the error of the previous tree. ntrees controls how many of these iterations occur and the learning rate controls how fast the model learns the error. Setting of these two parameters can be viewed as a compromise between fast learning (low iterations, high learning rate) and good generalisation (high iterations and low learning rate). The sample rate used by the GBM subsamples the data without replacement at each boosting round and each model is trained on this subsample.

For logistic regression (LR) and the super learner (SL) two parameters were used: alpha and lambda. There are two forms of regularisation used in logistic regression, the L1 and L2 norms. Alpha sets the form of regularisation used, where alpha of 0 is L1 regularisation and an alpha of 1 is L2 regularisation. Whereas with an alpha of 0.5 the model uses both forms of regularisation each weighted by half. Lambda is the overall weight on the regularisation penalty. Higher lambda values will shrink coefficients in the regression towards zero to provide more regularised (simpler) models. The SL used two lambda values per alpha: the best preforming lambda, and the lambda value from the most regularised model such its cross-validated error is within one standard deviation of the minimum cross-validation error.