INTA 6450-4803 Homework 2

Instructions

Please submit all analyses using `R Markdown` .html format unless otherwise noted. Follow the instructions specified below in each section. All answers are to be submitted to Canvas by the due date and time specified on the syllabus and Canvas. All analysis is to be completed individually, but you may utilize the class Slack channel to ask questions and collaborate on finding answers (i.e. directing students to Stack Exchange and the like).

I

Show, using the homework 2 dataset provided on Canvas, that Fearon and Laitin's (2003) model of civil war onset is overfit.

- For reference in the data, Fearon and Laitin's (2003) model is given by `as.factor(warstds) warhist+ln_gdpen+lpopns+lmtnest+ncontig+oil+nwstate +inst3+pol4+ef+relfrac`
- Split the data into a training and test set. Let the training dataset be from 1945-1980, and the test set from 1981-2000.
- Report the number of onsets correctly predicted by Fearon and Laitin's model in the test dataset.
- Set the random seed to 38745, use 10-fold cross validation to estimate the following Elastic Net models.
- Use an Elastic Net model with the same features that Fearon and Laitin used in their model. Report the number of onsets correctly predicted in the test data. Provide the non-zero features returned by this model.
- Use an Elastic Net with the full set of features in the dataset. Report the number of onsets correctly predicted. What are the non-zero features in the final model? Provide these features, and and the features from the previous model implementation in a nicely formatted table.
- Explain your results.

II

Now, using the same dataset compare the predictive results obtained by Fearon and Laitin's original model to a random forest model.

- Set the random seed value to 38745
- Grow a forest of 1000 trees, using 10-fold CV. Use all features in the data.
- Report the number of onsets correctly predicted by the random forest model.
- Compare your results to the Elastic Net and explain.

III

Evaluate the predictive accuracy of each model using confusion matrices. Report the precision, recall, and F1 scores for each model. Which model is most accurate according to each measure?

Use ROC Curves to estimate the fit of each of the three models. Which model is the most accurate?

Since the dataset is class-imbalanced, use the the Area under the Precision-Recall Curve to estimate the fit of each model. Which model is the most accurate?

What does this model fitting tell us about how we estimate the fit of our models?

IV

Run one final model, boosted decision trees using a `gbm` model on the dataset. Use 10-fold CV, and set the seed value to 38745. Report it's out-of-sample predictive power using the metrics in III.

Now, using `caretEnsemble` stack these models with a single-layer feed-forward neural network as a super-learner over the ensemble. Report the accuracy of this stacked ensemble compared to the results of each model separately.

Why do you think the ensemble or a single model within that ensemble is the most accurate model? Justify your answer.