# Abstract

In this paper we evaluate the predictive power of classification tree based methods for rare event data, namely ship casualties. We use single classification trees, random forests and boosted ensembles (GentleBoost and LogitBoost) to model these casualties and compare their performance to that of a logit model. To deal with a small amount of ship casualty occurrences, we use asymmetric misclassification costs and other (asymmetric) techniques to predict these rare incidents

The data contains 41009 observations, 3198 (7.80%) of which concern a ship casualty, i.e. a serious incident involving severe damage to the ship which include but are not limited to fires, explosions, grounding and heavy water damage. Several independent variables are available for each observation including basic ship properties (type, size, age, etc.), the flag under which it sails, how often the ship was inspected and whether any deficiencies where found during inspection.

As is often the case in classification problems, there is asymmetry in the misclassification's effects. In the case of casualty data, misclassification of casualties, that is, failing to predict such events, is more severe than misclassification of non-casualties. Therefore, we use asymmetric measures to improve predictive accuracy for the casualties. The first asymmetric measure is undersampling the non-casualties to create a balanced sample on which the model can be based. This yields the best results for all tree-based methods. The second is the use of asymmetric misclassification costs, i.e. making it more costly to mispredict a casualty than a non-casualty. This also improves results greatly, especially recall values of predictions, albeit in return for lower precision. The third and fourth measures are simple sampling bias corrections.

All three tree methods outperform the logit model. The random forest using balanced sampling without any correction yields the best (cross-validated) results. Depending on the score value threshold, the recall of this method can be as high as 0.98, with decent values for precision and hitrate.


October 25, 2013