Matematisk statistik
Stockholms universitet
Masteruppsats **2019:11**
http://www.math.su.se/matstat

Hiding in the Trees

A case study of filtering approaches to Insurance fraud classification using tree based Gradient Boosting

Oliver Murquist*

September 2019

Sammanfattning

With easy to access complex machine learning methods that are powerful right out of the box, the risk of misuse and overuse increases. Are simpler methods like Gradient Boosted trees and Random Forests intuitive and flexible enough to compete? Previous studies made on the subject indicate that slow learning models produce the best performance in terms of prediction and robustness. The results of this study were partly contradictory to that. A gradient boosting model configured to be a fast learner was found objectively better than the other models fitted, however most of that success could be tied to model complexity rather than learning rate. Additionally both gradient boosted trees and random forests displayed great difficulties handling imbalanced data. Overall the analysis concluded that a considerable portion of performance could be attributed to tailoring the model configuration to the outcome of an initial analysis as well as to the specific application environment.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post:oliver.murquist@gmail.com. Handledare: Chun-Biu Li.