

Mathematical Statistics Stockholm University Bachelor Thesis **2022:4** http://www.math.su.se

Gradient Boosting vs Random Forests: Predicting the Overnight Return of the OMXS30-index

Filip Bergkvist*

June 2022

Abstract

Statistical learning is an important tool for statistical analysis in many areas of finance. In the current economy, transactions are made all across the world making different stock markets increasingly integrated. This thesis will compare the prediction accuracy of two different machine learning techniques; gradient boosting and random forest. Using only a few predictor variables consisting of stock indices, currencies and commodities the different models will try to predict whether the overnight return of the OMXS30 index is negative or not. In this binary classification problem, the gradient boosting model achieved a slightly better prediction accuracy than the random forest model and both of them outperformed the Zero rule classifier. The variable importance when predicting the response showed for both models that the Hang Seng Index, Nikkei 225 Index and Nasdaq Composite Index had the most relative influence out of the predictor variables. Even if gradient boosting outperformed random forest in prediction accuracy, the difference in computational cost between the models must be taken into account when evaluating overall model performance. The GBM-model, even with a restricted parameter space, was very time consuming to train compared to the random forest model. Which model to use in a classification task depends on how sensitive the data is. If it is of high importance to get correct classifications it could justify training a model that requires more computational cost.

^{*}Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: filipbergkvist@hotmail.com. Supervisor: Ola Hössjer and Nils Engler.