Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2026:11**
http://www.math.su.se

A Comparative Study of Random Forest and Gradient Boosting for Imbalanced Multiclass Classification

Liam Cuclair*

May 2026

Abstract

In this thesis, multiclass classification under varying levels of class imbalance is investigated using two machine learning methods: random forest and gradient boosting. The study is conducted on two datasets with different characteristics, while varying class sizes, hyperparameter settings, and methods for handling class imbalance. Model performance is evaluated primarily using balanced accuracy, supplemented by the F1-score. The imbalance methods considered are oversampling, undersampling, and class weights. The results show that oversampling is generally the most effective method for handling class imbalance, especially when the imbalance is large. When the imbalance is more moderate, class weights also performs well. Undersampling generally performs worse than the other methods, although it can be competitive in datasets that are more prone to overfitting. For datasets that are more prone to underfitting, gradient boosting generally outperforms random forest across all levels of class imbalance, whereas the two methods perform similarly for datasets that are more prone to overfitting. The results also indicate that hyperparameter tuning is important for both methods. For random forest, larger maximum depths generally improve performance or have little negative effect, suggesting a relatively low risk of overfitting. For gradient boosting, however, performance is more sensitive to model complexity, and the best results are obtained at intermediate maximum depths, where for datasets prone to overfitting higher maximum depths offer very similar results to intermediate maximum depths.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: liam.cuclair@math.su.se. Supervisor: Martina Favero & Jan-Olov Persson.