Stockholm
University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2025:3**
http://www.math.su.se

A Comparative Performance Analysis of Random Forest and XGBoost in Regression for Various Noise Levels and Data Patterns

Fikrat Ibragimov*

February 2025

Abstract

This thesis compares two powerful tree based machine learning algorithms, Random Forest and XGBoost, with a focus on their ability to generalize under varying levels of noise. Simulated datasets were designed to include four regression functions. These regression functions corresponds to four distinct geometric shapes, referred to as Sphere-Truncated Cone, Cone, Rhombus-Truncated Pyramid, and Ridge, defined for five predictor variables and two noise levels, categorized as "high" and "low." These shapes serve as a challenge for both algorithms, evaluating their performance under structured and noisy conditions. The theoretical component of this study delves into the mathematical foundations of Random Forest and XG Boost, providing a step-by-step analysis of their mechanisms, including tree construction, optimization processes, and ensemble strategies. Using Mean Squared Error (MSE) and Mean Squared Error of Prediction (MSEP) for test data as evaluation metrics, this work systematically assesses the algorithms' performance across the simulated datasets. The findings demonstrate that Random Forest excels in handling regression functions with discontinuities, while Gradient Boosting performs better for regression functions with continuous and smooth patterns.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: Fikrat-ibragimov@live.se. Supervisor: Ola Gerton Henrik Hössjer och Johannes Heiny.