

A Comparative Study of Predictive Power in Football Match Outcomes

Gabriel Lindqvist*

June 2025

Abstract

This thesis investigates the use of machine learning methods to predict outcomes of football matches in the English Premier League. Specifically, the study compares the performance of multinomial logistic regression and random forest classifiers on a three-class prediction task, where the outcomes are either home win, draw or away win. The analysis is based on data that includes prior seasons starting from the 2016/2017 season up to the current 2024/2025 season. Predictive performance is evaluated using accuracy, class-specific F1 scores, and macro-averaged F1. To ensure robust results, each classifier was trained and tested across 10 independently stratified train/test splits. The results show that both classifiers performed similarly on the majority classes, but struggled to accurately classify draws. Further analysis of feature importance and predicted class probability distributions highlights challenges associated with class imbalance and limited separation in the feature space. The findings underline the importance of feature engineering for this particular prediction task.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: gali8113@student.su.se. Supervisor: Johannes Heiny, Ola Hössjer.