Charles University

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Non-Linear Classification as a Tool for Predicting Tennis Matches

Author: **Bc. Jakub Hostačný**

Supervisor: **RNDr. Matúš Baniar**

Academic Year: **2017/2018**

Abstract

In this thesis, we examine the prediction accuracy and the betting performance of four machine learning algorithms applied to men tennis matches - penalized logistic regression, random forest, boosted trees, and artificial neural networks. To do so, we employ 40 310 ATP matches played during 1/2001-10/2016 and 342 input features. As for the prediction accuracy, our models outperform current state-of-art models for both non-grand-slam (69%) and grand slam matches (79%). Concerning the overall accuracy rate, all model specifications beat backing a better-ranked player, while the majority also surpasses backing a bookmaker's favourite. As far as the betting performance is concerned, we develop six profitable betting strategies for betting on favourites applied to non-grand-slam with ROI ranging from 0.8% to 6.5%. Also, we identify ten profitable betting strategies for betting on favourites applied to grand slam matches with ROI fluctuating between 0.7% and 9.3%. We beat both benchmark rules - backing a better-ranked player as well as backing a bookmaker's favourite. Neural networks and random forest are the most optimal models regarding the total profitability, while boosted trees yield the highest ROI. Besides, we show that bet size based on the half-sized Kelly criterion outstrips constant bet size for betting on favourites.

JEL Classification	C01,C38, C45, C51, C52, C53, C55
Keywords	neural networks, logistic regression, random forest, boosted trees, tennis forecasting, tennis betting, tennis modeling

Author's e-mail	`jakubhstn@gmail.com`
Supervisor's e-mail	`matus.baniar@gmail.com`