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**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, January 2, 2018

Signature

Acknowledgments

The author is grateful especially to his family members for their unprecedented support during his studies. Besides, special thanks belong to the supervisor RNDr. Matúš Baniar for his willingness to help and insightful comments during last 18 months. Also, the author is grateful to one of his best friends Daniel Piaček for his valuable insights into machine learning algorithms and their practical implementation in R. Moreover, the author would like to express his gratitude to Mgr. Adriána Lelovská for her willigness to read the whole thesis in her free time and suggestions for further improvement. Lastly, but not least, special thanks belong to PhDr. Jiří Kukačka Ph.D. for consultation and helping out with insufficient computational capacity.

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`

Abstrakt

V tejto diplomovej práci skúmame predikčnú presnosť a výkon pri stávkovaní u štyroch strojovo učiacich sa algoritmov - penalizovaná logistická regresia, náhodny les, posilnené stromy a neurónové siete. Pri práci využívame 40 310 ATP zápasov hraných počas obdobia 1/2001-10/2016. Čo sa týka predikčnej presnosti, naše modely prekonávajú najlepšie modely súčasnosti pre predikovanie negrandslamových (69%) ako aj modely pre predikovanie grandslamových zápasov (79%). Všetky špecifikácie modelov sú presnejšie ako predikovanie na základe rebríčkového postavenia hráčov, zatiaľ čo väčšina špecifikácií je presnejších ako predikovanie na základe vypísaných kurzov stávkových kancelárií. Čo sa týka návratnosti pri stávkovaní, vytvorili sme šesť profitabilných stratégií pre stávkovanie na favoritov pre negrandslamové zápasy (návranosť investície v rozmedzí 0.8-6.5%). Taktiež sme identifkovali desať profitabilných stratégií pre stávkovanie na favoritov pre grandslamové zápasy (návranosť investície v rozmedzí 0.7-9.3%). Naše modely prinášajú vyššiu návratnosť ako stávkovanie na rebríčkového či kurzového favorita. Neurónové siete a náhodné stromy prinášajú najvyšší celkový zisk, zatiaľ čo posilnené stromy vykazujú najvyššiu percentuálnu návratnosť. Výsledky ďalej ukazujú, že veľkosť stávky závislá na Kellyho kritériu je optimálnejšia ako konštantná stávka pre stávkovanie na favoritov.

Klasifikace JEL	C01,C38, C45, C51, C52, C53, C55
Klíčová slova	neurónové siete, logistická regresia, náhodný lest, posilnené stromy, tenisové predpovede, stávkovanie na tenis, tenisové modelovanie

E-mail autora	`jakubhstn@gmail.com`
E-mail vedoucího práce	`matus.baniar@gmail.com`

Contents

List of Tables

List of Figures

Acronyms

ATP Association of Tennis Professionals

AUC area under curve

BT boosted trees

BTM McHale and Morton (2011)

CM combined model

F-NG betting exclusively on favourites for non-grand-slam matches only

F-G betting exclusively on favourites for grand slam matches only

FN False Negative Rate

FP False Positive Rate

G Grand Slam Matches

i.i.d independent and identically distributed

ITF International Tennis Federation

LR penalized logistic regression

LR1 Klaasen & Magnus (2003)

LR2 McHale & Morton (2011)

LR3 Konaka (2017)

LR4 Lisi & Zanella (2017)

LR5 Clarke & Dyte (2000)

LR6 Sipko & Knottenbelt (2015)

LR7 Koning (2011)

LR8 Ma et al. (2013)

NG Non-Grand-Slam Matches

NN artificial neural networks

NN1 Somboonphokkaphan et al. (2009)

NN2 Sipko & Knottenbelt (2015)

PCA Principal Components Analysis

PCs principal components

PR1 Boulier & Stekler (1999)

PR2 Gilsdorf & Sukhatme (2008)

PR3 Del Corral & Prieto-Rodriguez (2010)

RF random forest

ROC Receiver Operating Characteristic

ROI return on investment

TN True Negative Rate

TP True Positive Rate

U-NG betting exclusively on underdogs for non-grand-slam matches only

U-G betting exclusively on underdogs for grand slam matches only

Master's Thesis Proposal

Author	Bc. Jakub Hostačný
Supervisor	RNDr. Matúš Baniar
Proposed topic	Non-Linear Classification as a Tool for Predicting Tennis Matches

Motivation Undoubtedly, tennis is among the world's most popular sports. The popularity growth of the sport, paired with the expansion of the online sports betting market, has led to a large increase in tennis betting volume in recent years (e.g. Murray-Djokovic Wimbledon final in 2013 saw £48 million traded on Betfair). The potential profit, as well as academic interest, has fueled the search for accurate tennis match prediction algorithms. Considering the availability of an immense amount of diverse historical tennis data, an alternative approach to tennis prediction could be based on machine learning algorithms. The goal of this master thesis is to investigate the applicability of machine learning methods to the prediction of professional tennis matches. Moreover, by employing larger dataset with greater range of features along with the analysis of betting odds of various betting companies as an additional input, we aim to outperform current-state-of-art strategies.

Hypotheses

Hypothesis #1: The application of supervised machine learning algorithms (support vector machines, random trees, and artificial neural networks) is more appropriate approach to modeling of tennis match outcome than simple logistic regression.

Hypothesis #2: The prediction that uses larger set of features (e.g. head-to-head record, current form of players, age etc.) outperforms simple decision rule based only on the current ranking of the players.

Hypothesis #3: The profitable strategy for betting, taking into account the model prediction and betting odds, can be formulated (profit of 4% of total stake is set as a benchmark).

Hypothesis #4: The predictive ability of the model for women's tennis matches is inferior to men's tennis matches due to higher presence of unavailable or difficult to measure features.

Methodology The analysis using comprehensive dataset from tennisabstract.com with more than 50 000 observations covering time span 1998-2016. Subsequently, the data are split into training, validation, and test set. In addition, new matches (10/2016 - 4/2017) with the analysis of betting odds across several betting companies (Bet365, Bwin, Pinnacle, and Betfair) are used to further evaluate the profitability of suggested strategy. The critical part of the thesis is related to the features construction. As we deal with symmetric match feature representation, we experiment with two different constructions of features. Firstly, we take difference between the values of both match participants. As a second option, we use values for both players as two independent features. This approach preserve more information, allowing for more accurate prediction, although deals with some potential complications such as the prediction depending on the labelling of the players. We also consider historical averaging based on various time spans (e.g. percentage of first serve in last 15 matches, during the last year ...). Last but not least, we employ common-opponent approach proposed by Knottenbelt that uses only features calculated using the same set of opponent allowing for the fair comparison of the players. Also, time discounting is considered (more recent matches are of the higher importance). Moreover, feature scaling and regularization are applied. In order to test hypothesis 1 and 4 various evaluation metrics are taken into account (F1-score, ROC curve, learning curve (bias-variance trade-off, and confidence interval of accuracy). To evaluate hypothesis 2, we use lift chart to compare predictive accuracy of the model with baseline model (simple decision rule based on the current ranking of the players). Finally, in order to assess hypothesis 3, return on investment is calculated as total profit divided by total stake.

Expected Contribution Most current state-of-the-art approaches to tennis prediction take advantage of the hierarchical structure of the tennis match (match being composed of sets, which in turn are composed of games, which are composed of individual points) to define hierarchical expressions for the probability of a player winning the match. Although, several authors in the recent past (Sipko & Knottenbelt) employed logistic regression along with artificial machine learning algorithm to forecast the outcome of tennis matches, we aim to exploit much larger dataset (tens of thousands compared to few thousands) that in turns allows us to extend the range of features (hundreds compared to dozens). Furthermore, alternative approaches of ML are also considered (random trees and support vector machines). As an ad-

ditional input, betting odds of several betting companies (Pinnacle, Bwin, Bet365, and Betfair) are analyzed and bet is placed using the most optimal alternative. As a consequence, we believe that our model can outperforms others and profitable and sustainable strategy can be formulated. Also, separate model for women's tennis matches is formed and results of both versions of models are compared.

Outline

1. Introduction
2. Literature Review
3. Data Description
4. Methodology
5. Empirical Results
6. Conclusion
7. Suggestions for Further Research

Core bibliography

BARNETT, Tristan J., et al. Using Microsoft Excel to model a tennis match. In: Proceedings of the Sixth Australian Conference on Mathematics and Computers in Sport. Bond University: Queensland. 2002. p. 68.

BARNETT, Tristan; CLARKE, Stephen R. Combining player statistics to predict outcomes of tennis matches. IMA Journal of Management Mathematics, 2005, 16.2: 113-120.

BARNETT, T.; BROWN, A.; CLARKE, S. Developing a model that reflects outcomes of tennis matches. In: Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport, Coolangatta, Queensland. 2006. p. 3-5.

BARNETT, Tristan; POLLARD, Graham. How the tennis court surface affects player performance and injuries. J Med Sci Tennis 2007, 2007, 12.1: 34-37.

BICKEL, J. Eric. Some comparisons among quadratic, spherical, and logarithmic scoring rules. Decision Analysis, 2007, 4.2: 49-65.

BREIMAN, Leo. Bagging predictors. Machine learning, 1996, 24.2: 123-140.

CLARKE, Stephen R.; DYTE, David. Using official ratings to simulate major tennis tournaments. International transactions in operational research, 2000, 7.6: 585-594.

COATE, Douglas; ROBBINS, Donijo. The tournament careers of top-ranked men and women tennis professionals: Are the gentlemen more committed than the ladies?. Journal of labor research, 2001, 22.1: 185-193.

DEL CORRAL, Julio; PRIETO-RODRIGUEZ, Juan. Are differences in ranks good predictors for Grand Slam tennis matches?. International Journal of Forecasting, 2010, 26.3: 551-563.

DINGLE, Nicholas; KNOTTENBELT, William; SPANIAS, Demetris. On the (page) ranking of professional tennis players. In: European Workshop on Performance Engineering. Springer Berlin Heidelberg, 2012. p. 237-247.

FARRELLY, Daniel; NETTLE, Daniel. Marriage affects competitive performance in male tennis players. Journal of Evolutionary Psychology, 2007, 5.1: 141-148.

FERNANDEZ, Jaime; MENDEZ-VILLANUEVA, A.; PLUIM, B. M. Intensity of tennis match play. British journal of sports medicine, 2006, 40.5: 387-391.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. The elements of statistical learning. Springer, Berlin: Springer series in statistics, 2001.

GILSDORF, Keith F.; SUKHATME, Vasant A. Tournament incentives and match outcomes in women's professional tennis. Applied Economics, 2008, 40.18: 2405-2412.

GUYON, Isabelle; ELISSEEFF, Andre. An introduction to variable and feature selection. Journal of machine learning research, 2003, 3.Mar: 1157-1182.

HOLTZEN, David W. Handedness and professional tennis. International Journal of Neuroscience, 2000, 105.1-4: 101-119.

HORNIK, Kurt; STINCHCOMBE, Maxwell; WHITE, Halbert. Multilayer feedforward networks are universal approximators. Neural networks, 1989, 2.5: 359-366.

JAMES, Gareth, et al. An introduction to statistical learning. New York: springer, 2013.

KLAASSEN, Franc JGM; MAGNUS, Jan R. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. Journal of the American Statistical Association, 2001, 96.454: 500-509.

KLAASSEN, Franc JGM; MAGNUS, Jan R. On the probability of winning a tennis match. Medicine and Science in Tennis. 2003, 2003(8), 10-11. ISSN 1567-2352.

KLAASSEN, Franc JGM; MAGNUS, Jan R. Forecasting the winner of a tennis match. European Journal of Operational Research, 2003, 148.2: 257-267.

KNOTTENBELT, William J.; SPANIAS, Demetris; MADURSKA, Agnieszka M. A common-opponent stochastic model for predicting the outcome of professional tennis matches. Computers and Mathematics with Applications, 2012, 64.12: 3820-3827.

KOVACS, M. S. Applied physiology of tennis performance. British journal of sports medicine, 2006, 40.5: 381-386.

KUPER, Gerard H., et al. Using tennis rankings to predict performance in upcoming tournaments. University of Groningen, Research Institute SOM (Systems, Organisations and Management), 2014.

MADURSKA, Agnieszka M. A set-by-set analysis method for predicting the outcome of professional singles tennis matches. Imperial College London, Department of Computing, Tech. Rep, 2012.

O'DONOGHUE, Peter G. The most important points in grand slam singles tennis. Research quarterly for exercise and sport, 2001, 72.2: 125-131.

O'MALLEY, A. James, et al. Probability formulas and statistical analysis in tennis. Journal of Quantitative Analysis in Sports, 2008, 4.2: 15.

MCHALE, Ian; MORTON, Alex. A Bradley-Terry type model for forecasting tennis match results. International Journal of Forecasting, 2011, 27.2: 619-630.

NEWTON, Paul K.; KELLER, Joseph B. Probability of winning at tennis I. Theory and data. Studies in applied Mathematics, 2005, 114.3: 241-269.

PASERMAN, Daniele. Gender differences in performance in competitive environments: evidence from professional tennis players. Vol, 2007.

RADICCHI, Filippo. Who is the best player ever? A complex network analysis of the history of professional tennis. PloS one, 2011, 6.2: e17249.

SCHEIBEHENNE, Benjamin; BRÂ¨ODER, Arndt. Predicting Wimbledon 2005 tennis results by mere player name recognition. International Journal of Forecasting, 2007, 23.3: 415-426.

SIPKO, Michal; KNOTTENBELT, William. Machine Learning for the Prediction of Professional Tennis Matches. 2015.

SOMBOONPHOKKAPHAN, Amornchai; PHIMOLTARES, Suphakant; LURSINSAP, Chidchanok. Tennis winner prediction based on time-series history with neural modeling. In: Proceedings of the International Multi Conference of Engineers and Computer Scientists. 2009.

SPANIAS, A. Demetris; KNOTTENBELT, B. William. Tennis Player Ranking using Quantitative Models.

SPANIAS, Demetris. Professional Tennis: Quantitative Models and Ranking Algorithms. 2014. PhD Thesis. Imperial College London.

WATERHOUSE, J.; REILLY, T.; EDWARDS, B. The stress of travel. Journal of sports sciences, 2004, 22.10: 946-966.

WEINBERG, Robert S.; RICHARDSON, P. A.; JACKSON, Allen. Effect of situation criticality on tennis performance of males and females. International Journal of Sport Psychology, 1981.

YOUNG, Janet A.; PEARCE, Alan J. Attributes of champion female tennis players and challenges faced by aspirants. Medicine and science in tennis, 2009, 14.2: 16-19.

Author

Supervisor

Chapter 1

Introduction

Undeniably, tennis is among the world's most popular sports. Every year, The Association of Tennis Professionals (hereafter ATP) organizes about 60 tournaments worldwide, with more than 2600 matches played. Nevertheless, this portion is still just a tiny fraction - tens of thousands more are played on both challenger and International Tennis Federation (hereafter ITF) circuit. The popularity growth of the sport, paired with the expansion of the online sports betting market has led to a substantial increase in tennis betting volume in recent years (e.g., Murray-Djokovic Wimbledon final in 2013 saw £48 million traded on Betfair). The potential profit, as well as academic interest, has fueled the search for accurate tennis match prediction algorithms.

Although tennis can be played either against a single opponent or between two teams of two players, this thesis explores tennis singles exclusively. Professional singles tennis matches are an attractive proposition to model mathematically due to the three main reasons. Firstly, each match is played only between two players, as opposed to the multitude of participants involved in a team-based sport such as ice-hockey or football. Therefore, the complexity of predicting the outcome is greatly simplified. Secondly, a plethora of match statistics is publicly available leading to virtually unlimited modeling options. Thirdly, there are only two possible match outcomes.

Albeit an effective forecasting model may have various applications such as the analysis of the psychology of betting markets (Dixon & Pope, 2004; Graham & Scott, 2008), construction of ranking systems (Macmillan & Smith, 2007),

estimating the probability that a player with a particular ranking advances to a specific round (Kupper et al., 2014), or as an aid in the design of tournaments, in this thesis we examine how and if the effective prediction model can be applied to obtain positive ROI on online tennis betting market.

Tennis represents an attractive option from betting perspective. An advantage of tennis compared to team sports is that team sports have a large number of supporters with permanent allegiance. This sentiment distorts bookmaker odds offered in betting markets. Also, with only two possible outcomes, the overall margin of betting agencies is lower compared to multiple outcome sports such as horse racing. Last but not least, the level of transaction costs is low and comparable to other types of financial markets (Forrest & McHale, 2005).

With the introduction of live online betting, financial markets related to tennis have proliferated allowing traders to speculate on numerous outcomes, e.g., the likely winner of a match, the score of different sets, length of the match, or the expected number of aces. Nonetheless, taking the insufficient historical availability of the betting odds for various subtle match characteristics, this thesis examines only betting on the winner of the match. Bets can be placed both before the match and in-game. Various authors discuss in-game betting for tennis (Barnett et al., 2002; Klaassen & Magnus, 2003; Barnett & Clarke, 2005; Easton & Uylangco, 2010; Huang et al., 2011). However, due to the computational intensity of presented machine learning algorithms along with insufficient data on a point-by-point basis for the examined period, this thesis focuses solely on pre-match betting.

This thesis aims to explore the applicability of machine learning methods - regularized logistic regression, random forests, boosted trees, and artificial neural networks (hereafter NN) models) for the prediction of ATP men's tennis matches. Our models utilize a large dataset consisting of 40 310 ATP tennis matches played during 2001-2016. Moreover, we employ a broad set of input features that includes all relevant features found across literature related to the tennis modeling and extend them with new features. By searching through different betting strategies, we not only aim to formulate a profitable betting strategy, but we aim to show that machine learning algorithms presented here will outperform both simple decision rules based on the ranking of the players or betting odds and the models used for tennis betting so far.

Contribution

Thus, the contributions of this work are fourfold. First, we aim to investigate the applicability of previously untapped machine learning methods concerning tennis forecasting - random forests and boosted trees. Secondly, while the majority of tennis models employ only a limited set of explanatory variables (i.e., usually well below ten), we not only re-employ all significant features found in all papers written on the topic, but we also introduce a set of new features. Thirdly, following the notion 'here is a need to test the relative performance of heuristics, experts, and complex forecasting methods more systematically over the years rather than in a few arbitrary championships' (Goldstein & Gigerenzer, 2009), this thesis covers whole ATP seasons during the period $2001 - 2016$, as opposed to only a few seasons or few specific tournaments majority of other papers use. As a result, presented results are more robust. Finally, our thesis examines whether a constant bet size or a variable bet size based on the half-sized Kelly criterion is preferred for tennis betting.

Hypotheses

In this thesis, we aim to examine the following four hypotheses. Firstly, we predict that presented machine learning techniques combined with selected betting strategies can generate a positive return on investment if applied to online pre-match tennis betting market. Secondly, we assume that more complex machine learning algorithms such as random forest, boosted trees and NN will provide an improvement in predictive power to logistic regression and naive decision rules such as predicting the winner of the match based on ranking or bookmarker odds. Thirdly, we presume that determining bet size for each match depending on Kelly betting criterion presented by Kelly (1956) and refined by Dixon & Coles (1997) will lead to higher betting return compared to placing the constant bet size. Finally, we expect our models to perform better for grand slam tournaments, both regarding predictive power and betting return.

Structure

This thesis has the following structure. Chapter 2 Literature Review presents the brief overview of the approaches applied for tennis modeling along with their predictive power and betting performance. This section also discusses features that have been used so far to predict the outcome of the tennis match. Chapter 3 Methodology presents machine learning techniques we employ for modeling in this thesis, namely logistic regression, random forests, boosted trees, and NN. Furthermore, it introduces evaluation metrics along with betting strategies. Chapter 4 Data Description and Manipulation discusses data sources and data manipulation, presents feature engineering and feature introduction, summarize feature pre-processing operations, and discusses model tuning procedure. Chapter 5 Results and Discussion presents and discusses the empirical results. It also compares the prediction accuracy and betting return across all models and betting strategies investigated in this thesis as well as with other papers previously dealing with tennis forecasting and pre-match betting. Chapter 6 Conclusion summarizes key findings and assesses our hypotheses. Chapter 7 Further Extensions discusses limitations of this work and proposes extensions for further research. Finally, Chapter 8 Bibliography outlines the bibliography.

Chapter 2

Literature Review

Concerning tennis modeling and forecasting, various researchers have used a myriad of models and approaches with a different set of features. Two main issues are incessantly at the forefront - what model is the most optimal for tennis prediction and what features are most relevant in formulating such a prediction. Recently, various researchers have also investigated the efficiency of online pre-match tennis betting market and tested whether they can achieve sustainable monetary gains in this market.

2.1 Models for Tennis Forecasting

Concerning tennis modeling and forecasting, several different approaches have been proposed: **hierarchical Markov models** (Barnett & Clarke, 2002; Newton & Keller, 2005; Barnett & Clarke, 2005; Knottenbel et al., 2012; Spanias & Knottenbelt, 2012; Madurska, 2012), **logistic regression** (Clarke & Dyte, 2000; Klaassen & Magnus, 2003; Koning, 2011; McHale & Morton, 2011; Ma et al., 2013; Sipko & Knottenbelt, 2015; Konaka, 2017; Lisi & Zanella, 2017), **probit regression** (Boulier & Stekler, 1999; Gilsdorf & Sukhatme, 2008; Del Coral & Prieto-Rodriguez, 2010), **artificial neural networks** (Somboonphokkaphan et al., 2009; Sipko & Knottenbelt, 2015), and alternative models such as **paired comparison-based models** (McHale & Morton, 2011) or **recognition heuristics models** (Serwe & Frings, 2006; Scheibehenne et al., 2007).

2.1.1 Hierarchical Markov Models

Hierarchical Markov models exploit hierarchical structure of a tennis match - match consists of sets, which in turn contain games, which comprise individual points. Under this algorithm[1], the probability of winning a match in tennis is derived hierarchically starting from the probability of each player winning a point on his serve. Basic models apply only the probability of winning a point on serve for both players without taking opponent's ability on return into account (Barnett & Clarke, 2002; Newton & Keller, 2005). Previous matches of both players and average serve winning probabilities across all players participating on ATP circuit are employed to obtain estimates of serve winning probability for both players. Barnett & Clarke (2005) extend the model by adjusting the probability of winning a point on serve by opponent's return ability. Low-level point model introduced by Spanias & Knottenbelt (2012) further refines the model by combining more granular metrics to estimate the serve winning probabilities for both players such as server's historical ability to hit aces with receiver's historical vulnerability to aces.

As players historically face different opponents, Knottenbelt et al. (2012) argue that averaging across all opponents results in bias. Therefore, their Common-Opponent model further enhances existing hierarchical Markov models and considers only matches played against common opponents for averaging.

All models above rely on assumption of individual points during tennis match being independently and identically distributed (hereafter i.i.d) and thus ignore point-by-point and set-by-set dynamics of tennis match. Several authors have investigated the plausibility of i.i.d assumption. By examining almost 90 000 points played at Wimbledon 1992-1995, with a dynamic binary panel data model employed, Klaasen & Magnus (2001) found a robust evidence against the assumption of points in tennis match being independently and identically distributed, with the stronger violation for weaker players. According to their research, winning a previous point has a positive impact on winning the next point. In line with these findings, Jackson & Mosurski (1997) conclude that there is a psychological momentum in tennis. Also, Malueg & Yates (2010) confirm that the winner of the first set exerts higher effort in the second set

[1]Mathematical derivation for hierarchical Markov models with specific reference to tennis is widely discussed (Riddle, 1988; Barnett et al., 2006; O'Malley, 2008; Walker et al., 2011).

compared to the player who lost the first set.

In an attempt to build a more realistic model, Set-by-Set model presented by Madurska (2012) relaxes the assumption of invariant point-winning probabilities on the set-by-set level. Allowing for set-by-set dynamics results in dramatic improvement of predictive performance (see Table 2.1).

Nevertheless, despite its simplicity and applicability for in-game betting, hierarchical Markov models have their limitations. Sipko (2015) argues that the representation of the quality of players by only single value (percentage of points won on serve) fails to capture more subtle factors, such as the location of the tournament, weather, accumulated fatigue over the tournament and tennis season, or player's ability to react to various playing. Also, as the tennis rules have often been changing recently (e.g., ten-point tiebreak instead of the third set or shift to non-advantage games for some matches), the refinement of mathematical formulas is usually needed. Hence, more complex methods such as logistic/probit regression or artificial neural networks are gaining increasing popularity.

2.1.2 Logistic Regression

Logistic regression represents a traditional approach to binary classification problem and has been a popular candidate for tennis modeling. Majority of models based on logistic regression have employed a sole explanatory variable - either ATP official rankings or ATP points (Clarke & Dyte, 2000; Klaasen & Magnus, 2003; McHale & Morton, 2011; Konaka, 2017). Other authors extend the set of explanatory variables - ranging from six to 21 (Koning, 2011; Ma et al., 2014; Sipko & Knottenbelt, 2015; Lisi & Zanella, 2017). In those more complex models, in addition to some transformations of either ATP rankings or ATP points, player's physical characteristics (age and height), previous matches statistics (first and second serve, return, break points, total points won), match characteristics (surface, tournament level, tournament round), and the dummy indicating home advantage are included.

2.1.3 Probit Regression

Probit regression is a commonly used alternative to logistic regression for binary classification challenges. Similar to logistic regression, some authors employ only a sole explanatory variable - ATP rankings (Boulier & Stekler, 1999), while others opt for more comprehensive models, with 20 to 27 explanatory variables (Gilsdorf & Sukhatme, 2008; Del Coral & Prieto-Rodriguez, 2010). Besides additional variables mentioned for logistic regression models above, these models also introduce head-to-head balance, handedness, or prize money.

The main limitation of both logistic and probit regression is the fact that all features have to pre-specified. Due to complex nature of tennis, one is unlikely to capture all essential interaction terms and non-linearities in the data.

2.1.4 Artificial Neural Networks

Artificial neural networks do not require a precise specification of all input features, as opposed to both logistic and probit regression. Somboonphokkaphan et al. (2009) apply Multi-Layer Perceptron with back-propagation learning algorithm with three layers - input layer, hidden layer, and the output layer. Authors formulate three different models with three, 15, and 27 input nodes. Features include both previous matches match statistics (first and second serve, return, break points, and total points won) and dummies indicating match surface. As opposed to the majority of papers that employ differences, separate input features for both players are used. Sipko & Knottenbelt (2015) employ single layer NN with tanh activation function for hidden layer and logistic activation function for the output layer. Instead of separate input features for both players, authors use differences of statistics between match participants. Researchers claim that this approach reduces the variance of the model, prevents overfitting as well as it avoids the problem with the labeling of the players[2]. As opposed to all models above, neither ATP rankings nor ATP points are used in the model. Similar to Somboonphokkaphan et al. (2009), previous matches statistics are employed, with authors presenting new ones such as overall serve advantage or completeness. Also, head-to-head balance, fatigue,

[2]The model should predict the same probability of winning a match for a specific player, regardless of he is labeled as Player 1 or Player 2. This is not necessarily the case if separate features for both players are included in the model.

and the dummy indicating that a player plays the first match after a long injury are included. Motivated by Knottenbelt et al. (2012) only statistics against common opponents are utilized. Furthermore, surface and time-weighting of previous matches are applied, accounting for the fact the players perform differently across surfaces and more recent events are more important for assessing current quality of the players.

2.1.5 Other Modeling Approaches

Some authors also use alternative approaches. For instance, McHale & Morton (2011) employ Bradley-Terry model as a popular approach for handling data on paired comparison. The abilities of professional tennis players are inferred from a likelihood of games won and lost between player and opponent, with an exponential decay function to weigh more recent matches more heavily. Previous matches statistics, surface, ATP rankings, and the easiness of prior wins are utilized as explanatory variables. A different approach based on recognition heuristics is proposed by Serwe & Frings (2006) and Scheibehenne et al. (2007). Under this methodology, an alternative ranking system is constructed based on surveying both amateur tennis player and laymen. Players are ranked depending on how frequently their names were recognized. Subsequently, the winner of the match is predicted based on the recognition rankings.

2.1.6 Prediction Accuracy of Previous Models

The prediction accuracy of presented models ranges from 64% (Knottenbelt et al., 2012) to 77% (Lisi and Zanella, 2017). Nevertheless, one should be cautious about making the comparison between models due to the following reasons. Firstly, the best-performing models assess their prediction accuracy only on grand slam matches (Barnett et al., 2006; Serwe & Frings, 2006; Scheibehenne et al., 2007; Somboonphokkaphan et al., 2009; Madurska, 2012; Lisi & Zanella, 2017). Underdogs are much less likely to win a match at grand slam tournaments as they need three sets to win the match instead of two for lower-profile tournaments. Also, better players are more motivated due to higher prize money and ATP points awarded at grand slam tournaments. Secondly, the predictive performance of some models is difficult to compare with others as

only Brier score, or log loss is provided (Boulier & Stekler, 1999; Del Coral & Prieto-Rodriguez, 2011; Sipko & Knottenbelt, 2015).

Kovalchik (2016) validates the prediction accuracy of selected models on all 2014 ATP matches played, yielding range 59-67%. She also finds that models perform best for matches where TOP 30 players are involved, for grand slam matches, and for matches played on hard court.

Table 2.1 presents the overview of predictive performance for all tennis models we were able to extract information from. Table 2.1 also shows the prediction accuracy validation by Kovalchik (2016).

Table 2.1: Overview of Predictive Performance

Paper	Model Type	Accuracy - Claimed	Accuracy - Kovalchik (2016)
Barnett et al. (2006)	Markov model	72%	67%
Knottenbel et. al (2012)	Markov model	64%	63%
Spanias & Knottenbelt (2012)	Markov model	67%	64%
Madurska (2012)	Markov model	70%	NA
McHale & Morton (2011)	logistic regression	65%	NA
Lisi & Zanella (2017)	logistic regression	77%	NA
Sipko & Knottenbelt (2015)	logistic regression	0.613 logistic loss	NA
Del Coral & Prieto-Rodriguez (2010)	probit regression	0.158 Brier score	67%
Boulier & Stekler (1999)	probit regression	0.173 Brier score	59%
Somboonphokkaphan et al. (2009)	neural networks	74%	NA
Sipko & Knottenbelt (2015)	neural networks	0.611 logistic loss	NA
Scheibehenne et al. (2007)	recognition heuristics	70%	NA
Serwe & Frings (2006)	recognition heuristics	70%	NA
McHale & Morton (2011)	Bradley-Terry model	67%	65%

Note: Models with NA are not assessed by Kovalchik (2016).
Source: Author's own elaboration

2.2 Features Used for Tennis Modelling

A plethora of variables has been implemented across models. The majority of variables can be categorized as follows.

1. **Player's past performance**
   - ATP rankings, ATP points, previous matches statistics (e.g., serve, return, break point performance, tiebreak performance, total points won), head-to-head record
2. **Player's physical attributes and the current physical state**
   - age, height, handedness, fatigue, injury
3. **Match characteristics**
   - surface, tournament level, tournament round, prize money, match quality, home advantage

Besides presented variables, additional ones such as the effect of marriage, social support, or temperature have been investigated. Only a limited emphasis has been put on mental characteristics and the current mental state of the players so far.

2.2.1 Player's Past Performance

Player's past performance represents the most significant feature category across all models. The **official ATP ranking** is compiled to reflect player's performance in previous tournaments while using use a rolling 12 months window. It is designed to serve as a proxy for real player's ability. ATP rankings enter models in various forms. While the majority of models employ the difference in rankings between match participants (Boulier & Stekler, 1999; Del Coral & Prieto-Rodriguez, 2010; McHale & Morton, 2011), others opt for different transformations. For instance, Klaasen & Magnus (2003) utilize difference of transformed ATP rankings with transformation reflecting expected tournament round reached. Klaassen & Magnus (2003) and Koning (2011) include the sum of ranking as a measure of the absolute quality of the match. Lisi & Zanella

(2017) assume some homogeneity among the players within some intervals of ATP ranking, and thus five rank intervals are used instead.

Despite ATP rankings being widely used in models, some authors question the official tennis ranking system as a proxy for real player's ability. Instead, various alternatives are proposed and compared to the official ATP rankings in their predictive power. For example, Irons et al. (2014) show that the ranking based on the number of games won instead of the number of matches won results in improved predictive power. Bedford & Clarke (2000) use exponential smoothing method based on the margin of victory expressed in sets and games won. Despite its simplicity compared to the ATP ranking, this alternative algorithm performs similarly. Several authors also suggest surface-specific ranking (McHale & Morton, 2011; Irons et al., 2014). Glickman (1999) presents alternative ranking based on dynamic paired comparison similar to ELO ranking system used in chess.

Despite their similarity with ATP rankings, some authors opt for **ATP points** instead. Points depend upon progression within tournaments as well as tournament level (i.e., grand slam tournaments award more points compared to ATP 250). While some models use difference in points (Clarke & Dyte, 2000; Gilsdorf & Sukhatme, 2008), others select the ratio of ATP points instead (Konaka, 2017).

The evidence across prediction models emphasize the importance of the official ATP rankings or ATP points as predictors, these two variables themselves do not include all the necessary information for precise forecasting (McHale & Morton, 2011).

Albeit the previous matches statistics are statistically significant across all models, the evidence on the most important one is ambiguous. While Reid et al. (2010) show by examining the relationship between rankings and 14 statistics describing the match performance of TOP 100 ATP players during 2007 the prevalence of second serve points won and second serve return points won, Ma et al. (2013) favour first serve statistics. Sipko & Knottenbelt (2015) reveal that combining raw match statistics into more complex ones such as overall serve advantage can provide further improvement. Besides the match statistics above, total points won along with break point and tiebreak performance are often used.

As for other measures of player's past performance, Del Coral & Prieto-Rodriguez (2010) show that previous results on the same tournament are correlated with the probability of winning a match. Gilsdorf & Sukhatme (2008) and Sipko & Knottenbelt (2015) include head-to-head record and find its effect significant.

2.2.2 Player's Physical Attributes and Current Physical State

The most commonly used features related to physical attributes and current physical state are age, height, and handedness. The older the player gets, the more experienced he becomes. On the other hand, as repeated bouts of matches characterize professional tennis, often with only a little time to rest, the older players are at disadvantage due to their lower ability to recover. Del Coral & Prieto (2010) conclude that the probability of a higher-ranked player win decreases as he plays against the younger player. Schulz & Curnow (1988) argue that top tennis player reach their peak performance at age 24, while Lisi & Zanella (2017) show that the players achieve their peak rank at around age 27. Both age and square of age are used in models with both being significant (Gilsdorf & Sukhatme, 2008; Del Coral & Rodriguez, 2010).

Height affects the playing style as taller players have the advantage over shorter players for serve speed (Cross & Polland, 2009). On the other hand, taller players usually have worse coordination abilities. Ma et al. (2013) conclude that the most optimal height lies within interval 181-185 cm. Nevertheless, the effect of height on probability of winning a match is negligible across specifications.

Due to the significantly smaller proportion of left-handed players on ATP circuit, the effect of reduced familiarity with their playing style on the success of left-handed players is examined. By analysing the period 1968-1999, Holtzen (2000) shows that left-handed players were significantly over-presented among TOP players (World Number One and TOP10) and grand slam finalists, including champions, with the rate of left-handedness ranged from two to five times higher than expected in these highly successful players. Del Coral & Prieto-Rodriguez (2010) found some evidence of left-handers having higher probability of winning a match.

A standard explanation for the under-performance of a player in a match is the accumulated fatigue from previous matches. By utilising 20 320 Grandslams matches from 1992 till 2011, Goosens et al. (2015) conclude that there is an impact of the relative effort invested in winning a match on the probability of winning the next match. For men, a set difference of two was found to decrease the winning probability. Sipko & Knottenbelt (2015) include the number of games a player played in the past three days as a proxy for fatigue.

Sipko & Knottenbelt (2015) argue that a player's form is affected by recent injuries. Thus, they use player's withdrawal from the tournament as an approximation of injury. Subsequently, a dummy variable indicating that player plays his first match after an injury is included in the model. The effect of injury is found significant, although less important than fatigue.

2.2.3 Match Characteristics

As far as match characteristics are concerned, surface, tournament level, tournament round, home advantage, and prize money are discussed most frequently.

Tennis is played on four main surfaces - clay, hard, grass, and carpet; with clay and hard court accounting for the majority of matches. Surfaces differ in their characteristics such as speed and bounce of the ball, or physiological and physical demands (Fernandez et al., 2006). As a result, players perform differently across surfaces as some playing styles are more effective on a particular surface. For example, current No.1 player Rafael Nadal won 10 out of his 14 grand slam titles on clay. Tennis models incorporate information about the surface in various ways. For example, Gilsdorf & Sukhatme (2008) account for the familiarity with the surface by including the difference in career wins on the particular surface. Lisi & Zanella (2017) include dummy that indicates whether the match is played on player's favourite surface. Somboonphokkaphan et al. (2009) include set of dummies for each surface. McHale & Morton (2011) and Sipko & Knottenbelt (2015) utilize a more comprehensive approach and apply surface-weighting for assessing the player's past performance. The effect of the surface is found to be significant across specifications.

As awarded ATP points, prize money, and prestige differ through tournament levels and tournament rounds, the players are likely to exert higher effort

at high profile tournaments and in later rounds. While Gilsdorf & Sukhatme (2008) include dummy indicating whether the tournament is grand slam or master series, Del Coral & Prieto-Rodriguez (2010) and Ma et al. (2013) use separate dummies for each grand slam. Some of the dummies in these three studies are statistically significant. As for tournament rounds, Gilsdorf & Sukhatme (2008) include the number of remaining matches after the current match while Del Coral & Prieto-Rodriguez (2010) create a separate dummy for each tournament round. After controlling for tournament level, these variables are insignificant.

Home advantage suggests that player will perform above their expected performance level providing that the tournament is held in their own country. Several explanations for home advantage have been proposed - increased crowd support, lowered fatigue due to traveling, or familiarity with the home venue. Koning (2011) and Lisi & Zanella (2017) found its effect significant, being increasingly substantial with the absolute quality of the match, as measured by the sum of the rankings of both players.

Similar to tournament level and tournament round, prize money should exhibit a positive correlation with motivation and exerted effort. Sunde (2003) shows that prize money exhibits a highly significant and positive effect on effort. Substantially higher prizes to be won in finals also result in higher effort exerted by players, compared to semi-finals. In line with these findings, Gilsdorf & Sukhatme (2008) confirm that increase in prize money differential have a positive, statistically significant impact on the stronger player's probability of winning the match.

Although aforementioned variables represent most common features used in models, others are also considered. For example, Farrelly & Nettle (2007) show that married players suffer a significant drop in ranking points between the year before and the year following the marriage. Rees & Hardy (2004) examine the effect of social support on player's performance. Smith et al. (2017) explore the impact of temperature on playing style. By analysing Australian Open matches, authors show that increased temperature reduces net approaches and leads to more aces, indicating that temperature can have a substantial effect on specific playing styles (e.g., serve-volley strategy), but an only negligible impact on others (e.g., defensive play from baseline).

So far only a little effort has been exerted to examine the effect of player's intrinsic mental characteristics and current mental state of the player on the probability of winning a tennis match. Even though Chitnis & Vaidya (2014) use the number of tiebreaks won, the number of matches won after losing the first set, and the number of matches lost after winning the first as proxies for player's mental toughness, these variables were employed only in a non-parametric approach called Data Envelopment Analysis. Nevertheless, this technique was only used to assess the relative performance of an individual player and provide an alternative to the official ATP rankings.

Table 2.2 presents an overview of features used in tennis modeling.

Table 2.2: Overview of Features Used

Model	LR1	LR2	LR3	LR4	LR5	LR6	LR7	LR8	PR1	PR2	PR3	NN1	NN2	BTM
# variables	1	1	1	6	1	20	8	21	1	27	20	27	20	4
ATP Rankings	*	*		*			*	*	*		*			*
ATP Points			*	*	*					*				
Age or Height				*				*		*	*			
Handedness								*			*			
Head-to-Head						*				*			*	
Tournament Level							*	*		*				
Tournament Round										*	*			
Surface		*		*		*						*	*	*
Home Advantage				*			*							
Match Stats [1]				*		*		*				*	*	*
Betting Odds				*										
Injury						*							*	
Fatigue						*							*	
Other										* [2]				* [3]

Source: Author's own elaboration

Notes:
LR - logistic regression, PR - probit regression, NN - neural networks, BTM - Bradley-Terry model

To find out in which paper each model can be found, see List of Acronyms.

[1] Includes statistics from previous matches such as % first/second serve in, % first/second serve points won, % return point won, % break point converted / saved, or % total points/games/sets won.
[2] Career wins and prize money
[3] Career wins

2.3 Sport Betting Markets and Betting Return for Tennis

2.3.1 Efficiency of Sport Betting Markets

Various authors have examined the efficiency of sport betting markets. While some have questioned efficiency of sport betting markets across variety of sports (Pope & Peel, 1989; Gandar et al., 1998; Sauer, 1998; Cain et al., 2003; Rosenbloom & Notz, 2006; Graham & Stott, 2008; Spann & Skiera, 2009; Koning, 2012; Kopriva, 2015), others have elaborated explicitly on tennis betting market (Forest & McHale, 2005; Forest & McHale, 2007; Easton & Uylangco, 2010; Lahvička, 2014; Abinzano et al., 2016). Besides, few authors have tried to exploit these inefficiencies in search for profitable betting strategies on pre-match tennis betting market (Forest & McHale, 2005; McHale & Morton, 2011; Knottenbelt et al., 2012; Sipko & Knottenbelt, 2015; Lisi & Zanella, 2017).

According to the efficient market hypothesis, the bookmaker odds should reflect all available information relevant to the match outcome. While some authors advocate bookmaker odds as an efficient reflection of match outcome (Gandar et al., 1998; Pope & Peel, 1989), others question the efficiency of sport betting markets.

For instance, Sauer (1998) and Cain et al. (2003) suggest that betting on favourites tend to have higher expected return than betting on longshots. There are three main reasons for this phenomenon. Firstly, bettors are risk-lovers and betting agencies exploit this by lowering odds for longshots. Secondly, bettors overestimate winning probability for longshots. Finally, betting agencies hedge themselves against information asymmetry and thus bookmakers may lower odds on outsiders as an insurance against bet made by private information holders (Forrest & McHale, 2005).

Other researchers analyze bettor's behaviour and cognitive biases associated with it. By examining the world's largest betting exchange Betfair, Kopriva (2015) argues that bettors tend to overweight small and underweight vast differences in probabilities. As a result, this behaviour leads to biased bookmarker odds. Subsequently, inefficient odds can be exploited. Dixon & Pope (2004)

imply that differences in bookmaker odds often generate an arbitrage opportunity.

Some authors test market efficiency with specific reference to soccer betting market. Koning (2012) reveals odds on soccer betting market are not entirely informationally efficient. Graham & Stott (2008) confirm the existence of systematic biases and claim that these biases cannot be explained merely by the omitted variable or excluding extraneous information. The systematic deviations for games between strong and weak teams are found.

In line with findings as mentioned earlier, Rosenbloom & Notz (2006) conclude the superior accuracy of the real-money market for non-sports events compared to sports events.

As far as the market efficiency in tennis betting market specifically is concerned, the empirical evidence is ambiguous, however with a majority of empirical evidence implying the presence of biases.

Easton & Uylangco (2010) show that there is a high level of efficiency in the tennis betting market and demonstrate that betting odds are a good predictor of tennis match outcome. Furthermore, Klaasen & Magnus (2003) reveal that their hierarchical Markov point-by-point model yields probabilities that exhibit extremely high correlation with probabilities implied by bookmaker odds.

On the contrary, in line with findings of Sauer (1998) and Cain et al. (2003), several papers have found an evidence of positive favourite-longshot bias[3]. Lahvička (2014) and Abinzano et al. (2016) support the existence of positive favourite-longshot bias with the bias being most pronounced in matches between lower-ranked players, in high-profile tournaments, and in later-round matches. Forest & McHale (2007) also confirm the existence of positive favourite-longshot bias.

[3]Favourite-longshot bias is an observed phenomenon where on average, bettors tend to overvalue "long shots" and undervalue favourites - the bets on low probability outcomes have lower expected return than bets on high probability outcomes.

2.3.2 Betting Performance of Previous Models

To our best knowledge, the profitability potential of seven different models has been assessed against pre-match tennis betting market so far. Naive decision rule presented by Forest & McHale (2005) places a bet on favourite of each match (i.e., the player with lower bookmaker odds). By betting on grand slam matches during 2001-2004, this simple heuristics was able to generate 2,1% ROI. These results further support the existence of positive favourite-longshot bias. The advantage of this approach is that it places a bet on every single grand slam match.

Other betting models are based on predictive tennis models. These models compare model-implied probabilities with bookmaker odds-implied probabilities to identify matches to bet on. As a result, only a fraction of matches is identified to bet on.

Overall, all presented models exhibit a positive ROI ranging from 3.8 to 16.3%. Nevertheless, one should be cautious about comparing these models based only on their ROI due to the following reasons. Firstly, some betting models are evaluated only on grand slam matches (McHale & Morton, 2011; Lisi & Zanella, 2017), while others consider betting on all ATP matches (Knottenbelt et al., 2012; Sipko & Knottenbelt, 2015). As grand slam tournament account for only about 20% matches played, later models naturally lead to much more placed bets.[4] Secondly, betting rules are not uniform across all models. Even though all models bet only on matches where model-implied probability exceeds bookmaker odds-implied one, some models bet only on predicted winner (Knottenbelt et al., 2012; Sipko & Knottenbelt, 2015), while others also bet on underdogs providing that minimum threshold difference[5] in implied probabilities is suggested (McHale & Morton, 2012; Lisi & Zanella, 2017). The latter approach naturally leads to more bets placed. Lastly but not least, bet size differs across models. While some betting models place a constant bet size (Knottenbelt et al., 2012; McHale & Morton, 2012), others determine the proportion of bankroll to bet for each match based on Kelly criterion[6] (Sipko & Knottenbelt, 2012; Lisi & Zanella, 2017). Under this criterion, the bet size is

[4] In terms of absolute profitability, it is better to exhibit 4% ROI on 1000 bets than 10% ROI on 200.

[5] Ranging from 5 to 10%

[6] For more details, see Kelly (1956) or Baker & McHale (2013)

higher for betting on favourites and when the difference between model-implied and bookmarker odds-implied probability is more substantial.

Table 2.3 shows ROI for various betting models along with information about approaches to bet sizes, the proportion of ATP matches to bet on, and whether only bets on the favourite of the match are considered.

Table 2.3: Overview of Betting Returns and Strategies

Paper	Model Type	ROI	Proportion of ATP Matches	Bet Size	Favourite / Underdog
Knottenbelt et. al (2012)	Markov model	3.8%	40.1%	constant	only favourite
Lisi & Zanella (2017)	logistic regression	16.3%	2.4%	Kelly criterion	both
Sipko & Knottenbelt (2015)	logistic regression	4.2%	50.4%	Kelly criterion	only favourite
Sipko & Knottenbelt (2015)	neural networks	4.4%	50.4%	Kelly criterion	only favourite
McHale & Morton (2011)	Bradley-Terry model	10%	8%	constant	both
Forrest & McHale (2005)	naive decision rule	2.1%	19.3%	constant	only favourite

Source: Author's own elaboration

Chapter 3

Methodology

This chapter presents algorithms used for modeling, evaluation metrics employed to assess their performance, and betting strategies tested to determine whether monetary gains can be generated on online pre-match tennis betting market.

3.1 Models

In this paper, we apply four methodologies to build our models for prediction - logistic regression, random forest, boosted trees, and artificial neural networks.

3.1.1 Logistic Regression

The logistic regression is one of the most commonly used classification algorithms due to its simplicity and ability to make inferential assertions about model features.

Logistic regression models the log odds of the event as a linear function:

$$\log\Big(\frac{p(x)}{1-p(x)}\Big) = \beta_0 + \beta_1 x_1 + \ldots + \beta_P x_P, \tag{3.1}$$

where:

$p(x) = Pr(Y = 1|X)$
P - number of features

In binary classification setting, we can infer event probability as:

$$p(X) = \frac{1}{1 + e^{[-(\beta_0+\beta_1 x_1+...+\beta_P x_P)]}},$$

where $p(X) = Pr(Y = 1|X)$.

This non-linear function is a sigmoidal function of the model features and maps real-valued inputs between $-\infty$ and ∞ to values within the interval (0,1), allowing for output to be interpreted as the probability approximation. Similar to linear regression, even though the equation for p is nonlinear, this model also produces linear class boundaries, unless the predictors used in the model are non-linear transformations of the data, such as quadratic, cubic or higher polynomial functions.

Although other methods[1] are used to fit the model (see Equation 3.1), maximum likelihood based on the conditional likelihood of Y given X is most widely used. This method seeks parameter estimates of $\beta_0, \beta_1 \ldots, \beta_P$ such that the predicted probability for observations corresponds as closely as possible to their true values (i.e., for observations with Y=1, predicted probability should be as close to 1 as possible, while for those with Y=0, predicted probability should approach 0). Mathematically, parameters $\beta_0, \beta_1 \ldots, \beta_P$ are selected to maximize the likelihood function:

$$l(\beta_0, \beta_1 \ldots, \beta_P) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_{i'}=0} (1 - p(x_{i'}))$$

Penalized Logistic Regression

As we include 342 features in our model, we opt for a penalized logistic regression with built-in feature selection mechanism and shrinkage of less relevant predictors. Both lasso and ridge penalty terms are commonly used. While lasso implements built-in feature selection, the ridge is more suitable for cases

[1] Alternative methods such as non-linear least squares, gradient descent, conjugate gradient, BFLS, or L-BFGS could also be considered.

when features exhibit high collinearity. As both of these characteristics are beneficial for our modeling, we employ elastic-net penalty term introduced by Friedman et al. (2001) that combines both lasso and ridge penalty terms in one equation:

$$min_{\beta_0,\beta}\frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda\big[(1-\alpha)||\beta||^2/2 + \alpha||\beta||\big],$$

where:
N - number of observations
w_i - weight of the observation i
$l(y_i, \beta_0 + \beta^T x_i)$ - negative log-likelihood contribution of observation i
α - elastic-net penalty that bridges the gap between lasso (α=1) and ridge (α=0)
λ - overall strength of the penalty

The algorithm is implemented by using *glmnet* method in *R package caret*.

An efficient logistic regression model would require a comprehensive inspection to parametrize the model in a way that would account for non-linear effects through including quadratic, cubic or higher polynomial functions of original features as well as interaction terms. Thus, prescribing an exact functional form for the predictors is difficult (Kuhn & Johnson, 2013). An alternative solution could be using cubic splines (Harrel, 2015) or generalized additive models (James et al., 2013). Nevertheless, as we employ a penalized logistic regression model with a limited set of interaction and quadratic terms, non-linear methods presented below are expected to lead to a better performance.

3.1.2 Tree-based Methods, Random Forest, and Boosted Trees

Tree-based methods stratify the feature space into a set of rectangulars, and then fit a simple model (like a constant) for each one. For classification problems, we predict the response for each new observation by the most frequently occurring class in the region where it belongs.

These algorithms aim to find a segmentation of feature space that mini-

mizes the selected error rate [2]. As it is computationally infeasible to consider all possible partition of feature space, the recursive binary partition is often considered. Friedman et al. (2001) present this procedure for classification as follows:

1. Starting with all of the data, we consider a splitting variable j and split point s and define the pair of half-planes as:

$$R_1(j,s) = \{X|X_j \leq s\} \text{ and } R_2(j,s) = \{X|X_j > s\}$$

2. Then we seek the splitting variable j and split point s that solve:

$$min_{j,s}\Big(min_{c_1} \sum_{x_i \in R_1(j,s)} I(y_i \neq c_1) + min_{c_2} \sum_{x_i \in R_2(j,s)} I(y_i \neq c_2)\Big),$$

   where c_n takes value 0 or 1, depending on which class of Y prevails in the region n.

3. Having found the best fit, we segregate data into the two resulting regions and repeat the splitting process on each of the two regions.

4. Then, this process is repeated on all of the resulting regions until some stopping rule is reached[3].

Nevertheless, as this approach is likely to overfit the data and lead to poor test set performance (Friedman et al., 2001; James et al., 2013), cost-complexity pruning is performed. This procedure, as presented in Friedman et al. (2001), finds, for each α, the subtree $T_\alpha \subseteq T_0$ that minimizes the following cost complexity criterion:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|,$$

where:
N_m - number of observations assigned to terminal node m represented by region R_m

[2] While both Gini index and cross-entropy are preferred as a criterion for making the binary splits, only for the demonstration purposes, the minimization presented below employs misclassification error rate to keep the notation more straightforward. For more details, see Friedman et al. (2001).

[3] For example, once some minimum terminal node size is reached.

$Q_m(T)$ - node impurity measure[4]
α - tuning parameter that governs trade-off between goodness of fit to the data and tree size. Large values of α result in smaller trees.

Tree-based methods are simple and easy to interpret. Moreover, these methods are non-parametric and require few statistical assumptions. Nevertheless, one major problem of tree-based methods is their high variance. Thus, the prediction accuracy of simple tree-based methods is often inferior to other machine learning algorithms. To alleviate high variance of a single tree, we employ random forest and boosted trees instead.

Random Forest

Random forest builds an extensive collection of de-correlated trees, and averages them afterwards. Each of the B de-correlated trees of random forest is obtained as presented in Friedman et al. (2001):

1. We draw a bootstrap sample Z of size N from the training data.
2. We grow a random-forest tree to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size is reached.
   (a) We select m out of p variables at random.
   (b) We pick the best variable/split among m selected variables.
   (c) We split the node into two daughter nodes.

Subsequently, we can obtain class prediction as:

$$\hat{C}^B_{rf}(x) = \text{Majority vote } \{\hat{C}_b(x)\}_{b=1}^B,$$

where:
$\hat{C}_b(x)$ - class prediction for observation x made by b_{th} random-forest tree
B - number of trees in random forest

The algorithm is implemented by using *rf* method in *R package caret*.

[4] Any of misclassification error rate, Gini index, and cross-entropy can be used, but typically misclassification error rate is selected (Friedman et al., 2001).

Boosted Trees

Boosted trees combine an output of many 'weak' trees to produce a powerful 'committee'. A weak tree is one whose error rate is only slightly better than random guessing. The purpose of boosting is to sequentially apply the weak classification algorithm to repeatedly modified versions of the data, thereby producing a sequence of weak classifiers $S_m(x)$, $m = 1, 2, \ldots, M$. The data modifications at each boosting step consists of applying weights $h_1, h_2, \ldots, h_N$ to each of the training observations (y_i, x_i), $i = 1, 2, \ldots, N$. Initially, all of the weights are set to $h_i = 1/N$. At step m, those observations that were misclassified by the classifier S_{m-1} induced at the previous step have their weights increased, whereas the weights are decreased for correctly classified observations. The predictions from all of the weak classifiers are then combined through a weighted majority vote to produce the final prediction. Formally, following Friedman et al. (2001), the algorithm is implemented as follows:

1. We initialize the observation weights $h_i = 1/N, \quad i = 1, 2, \ldots, N$
2. For m=1 to M:
   (a) We fit a classifier $S_m(x)$ to the training data using weights h_i
   (b) We compute
   $$err_m = \frac{\sum_{i=1}^{N} h_i I(y_i \neq S_m(x_i))}{\sum_{i=1}^{N} h_i}$$
   (c) We compute $\alpha_m = log((1 - err_m)/err_m)$
   (d) We set $h_i := h_i e^{\alpha_m I(y_i \neq S_m(x_i))}, \quad i = 1, 2, \ldots, N$

   Given weak classifiers $S_m(x)$ produce a prediction taking either -1 or 1, we can obtain final class prediction as:

   $$S(x) = sign\Big[\sum_{m=1}^{M} \alpha_m S_m(x)\Big],$$

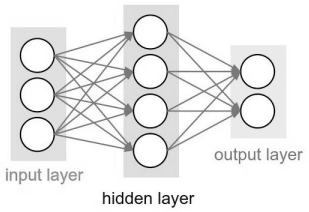
   where $\alpha_1, \alpha_2, \ldots, \alpha_M$ are computed by the boosting algorithm, and weight contribution of each respective $S_m(x)$.

The algorithm is implemented by using *ada* method in *R package caret*.

3.1.3 Artificial Neural Networks

Artificial neural networks are inspired by the learning processes that occur in the biological system. In these systems, neurons receive information from another neuron, process it, and pass it to the next neuron. Although the term neural network encompasses a large class of models and learning methods, in this thesis, we consider only the most widely used neural net - the single hidden layer back-propagation network. As Hornik et al. (1989) show that a single hidden layer with a finite number of neurons can approximate any continuous function, provided that a sufficient number of hidden layers is used. The applied neural network is a two-stage classification model. Figure 3.1 presents the example of single hidden layer network for binary classification problem.

Figure 3.1: Example of a single-layer neural network with four hidden and two output neurons



Source: http://cs231n.github.io/neural-networks-1/

In this model, the outcome is modeled by an intermediary set of unobserved variables called hidden neurons. As presented in Friedman et al. (2001), hidden neurons H_m are created from linear combination of some or all the original features, and then the target Y_k is modeled as some function of linear combinations of H_m:

$$\begin{aligned} H_m &= g\Big(\alpha_{0m} + \alpha_m^T X\Big), \quad m = 1, \ldots, M, \\ T_k &= \beta_{0k} + \beta_k^T H, \quad k = 1, \ldots, K, \\ f_k(X) &= g_k(T), \quad k = 1, \ldots, K, \end{aligned}$$

where $H = (H_1, H_2, \ldots, H_M)$, and $T = (T_1, T_2, \ldots, T_K)$.

The sigmoid function $\sigma(u) = \frac{1}{1+e^{-u}}$ is typically selected as an activation function $\sigma(u)$. The output function $g_k(T)$ performs a final transformation of

the vector of outputs T. For classification problems, softmax function is used:

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}$$

This transformation produces positive estimates that sum to one. The neural network has the following unknown parameters, often called weights, that have to be estimated:

$$\begin{aligned} \{\alpha_{0m}, \alpha_m; m = 1, 2, \ldots, M\} \qquad & M(p+1) \text{ weights} \\ \{\beta_{0k}, \beta_k; k = 1, 2, \ldots, K\} \qquad & K(m+1) \text{ weights.} \end{aligned}$$

For classification, either squared error:

$$R(\Delta) = \sum_{k=1}^{K} \sum_{N}^{i=1} (y_{ik} - f_k(x_i))^2$$

or cross-entropy can be used as a measure of fit:

$$R(\Delta) = \sum_{k=1}^{K} \sum_{N}^{i=1} y_{ik} \log f_k(x_i)$$

The parameters are usually initialized to random values, and then specialized algorithms for solving are used. Gradient descent called back-propagation is a general approach to minimize $R(\Delta)$. In the back propagation algorithm, each hidden neuron passes and receives information only to and from neurons that share a connection. The gradient uses the chain rule for differentiation[5].

Neural networks have tendency to overfit. Two most widely used methods to moderate overfitting are early stopping rule and weight decay. An early stopping rule terminates the optimization procedure when some estimate of error rate starts to increase. A more rigorous method is to use weight decay, a penalization method similar to ridge regression.

[5]For more technical details on back propagation algorithm, see Rumelhart et al. (1985) or Friedman et al. (2001).

Neural Networks Using Model Averaging

Instead of employing a single neural network, we follow approach presented in Ripley (2007). Thus, the same neural network model is fit using different random number seeds. All the resulting models are used for prediction. The model scores are first averaged, then translated to predicted classes. As we employ 342 features and neural networks are not robust to predictor noise nor perform automatic feature selection, this procedure is preferred to fitting a single neural network.

The algorithm is implemented by using *avNNet* method in *R package caret*.

Table 3.1 presents an overview of all algorithms we use in this thesis and summarizes some of their characteristics.

Table 3.1: Overview of Applied Models and Their Characteristics

Model	**Pre-Processing Required**	**Automatic Feature Selection**	**Robust to Predictor Noise**	**Computational Time**
Penalized Logistic Regression	C, S, NZV	YES	NO	FAST
Boosted Trees	-	YES	YES	SLOW
Random Forests	-	PARTIALLY	YES	SLOW
Neural Networks	C, S, NZV, CF	NO	NO	SLOW

Note: C-centering, S-scaling, NZV-near-zero variance filter, CF-correlation filter

Source: Kuhn & Johnson (2013)

3.2 Evaluation Metrics and Benchmarks Used

This section presents all evaluation metrics employed to evaluate and compare the performance of all presented machine learning techniques with each other as well as with models presented in the literature so far. Also, we present simple rules that serve as benchmarks for both prediction accuracy and betting return.

Confusion matrix is a conventional method to describe the performance of a classification algorithm. It is a simple cross-tabulation of the predicted and observed classes for the data. While diagonal cells denote cases where the classes are correctly predicted, off-diagonal elements depict the number of errors for each class. Table 3.2 shows the confusion matrix for the two-class problem. The table cells illustrate the number of the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Table 3.2: Confusion Matrix for Two-Class Problem

Predicted	**Observed**	
	Event	Nonevent
Event	TP	FP
Nonevent	FN	TN

Overall accuracy rate can be inferred from confusion matrix as follows:

$$\text{Overall accuracy rate} = \frac{TP + TN}{TP + FP + TN + FN}$$

Receiver operating characteristic (ROC) is a graphical plot that depicts the true-positive rate against the false-positive rate at various thresholds.

The **area under curve** (hereafter AUC) estimate corresponds to the probability that the classification algorithm would assign a higher score to a randomly selected positive example than to a randomly selected negative example. It represents the area under the ROC curve with values of $[0, 1]$. The value of 0.5 corresponds to a coin flip, while value of 1 represents a perfect classifier.

In this thesis, we employ both overall accuracy rate and AUC to evaluate and compare all model specifications.

Return on investment (ROI) represents a non-accuracy-based evaluation

metric that assesses the betting performance of presented machine learning techniques. In the context of tennis betting, ROI is defined as:

$$\frac{BR_{\text{end}} - BR_{\text{initial}}}{\sum_{i=1}^{n} s_i},$$

where:
n - number of matches we bet on
s_i - bet size placed on i_{th} match
BR_{initial} - initial betting bankroll
BR_{end} - betting bankroll at the end of betting period

We consider the following two simple rules as benchmarks for both prediction accuracy and betting return:

1. **Backing a better-ranked player**
   We favour the player with the lower ranking.

2. **Backing a bookmaker's favourite**
   We opt for the player with lower bookmaker odds.

3.3 Betting Strategies

The fundamental problem for bettors is to find positive expectation bets. Also, the bettor needs to know how to manage his money, i.e., how much to bet for any selected match.

3.3.1 Decision Criteria for Matches to Bet on

To identify positive expectation bets, our model-implied probability of a player winning a match has to exceed bookmaker odds-implied probability. Betting odds represent the return that bettor receives if he correctly predicts the outcome. For example, if bettor accurately predicts the winner of the match at 1.75, he will obtain 1.75 times his initial bet. If the bettor mispredicts the outcome, he will lose the full amount of his initial bet. In theory, odds should reflect the bookmaker's estimate of true probabilities of both players winning the match. However, as the betting agency operates with margin, the odds are lowered[6] in practice. We can infer bookmaker odds-implied probability of a player winning a match as follows[7]:

$$p(Xwins) = \frac{1}{X},$$

where X is a pre-match bookmaker odds for a player winning a match.

[6] If a bettor placed a bet on both possible match outcomes, he would obtain less money than the sum of his bets

[7] Following approach presented by Dixon & Coles (1997) and Sipko & Knottenbelt (2015), we could alternatively use a scaled version of bookmaker odds-implied probability:

$$p(Xwins) = \frac{Y}{X+Y},$$

where:
Y - pre-match bookmaker odds for opponent winning the match
X - odds for the player for whom we want to calculate the odds-implied probability.
Nonetheless, employing an inverse of bookmaker odds instead leads to more prudent and realistic betting rules (i.e., if real bookmaker's estimate of match-winning probability were 0.5, he would not list 2.00, but rather 1.9 to operate with margin. As a result, bookmaker odds-implied probability would be 1/1.9=0.526. Consequently, we consider betting only if the model-implied probability exceeds 52.6%, not just 50% (1/2.00).

Motivated by Lisi & Zanella (2017), we bet on positive expectation matches if and only if the following criteria are met:

1. The model and the bookmaker odds must agree on the favourite/underdog player
2. Expected gain has to exceed some safety threshold
3. Bookmaker odds belong to a moderately sized interval of values

   We bet on favourites only if both model-implied probability and odds-implied probability are below the selected upper boundary. Similarly, we bet on underdogs only if both model-implied probability and odds-implied probability are above the selected lower boundary.

While majority of previously tested betting strategies consider betting exclusively on match favourite (Forrest & McHale, 2005; Knottenbelt et al., 2012; Madurska, 2012; Sipko & Knottenbelt, 2015), some place a bet on any player once the match with positive expected gain is identified (McHale & Morton, 2011; Lisi & Zanella, 2017). In this thesis, as betting on favourites and betting on underdogs are mutually exclusive, we evaluate and compare the profitability of these two approaches separately.

The following indicator function summarizes our decision rule for betting on a match favourite:

$$I_f(bet) = \begin{cases} 1 & \text{if } \frac{p_{M_f}}{p_{B_f}} - 1 \geq r \wedge p_{M_f} > \frac{1}{2} \wedge p_{B_f} > \frac{1}{2} \wedge p_{M_f} \leq b_u \wedge p_{B_f} \leq b_u \\ 0 & \text{else} \end{cases}$$

where:
p_{M_f} - model-implied probability of a favourite winning the match
p_{B_f} - bookmaker odds-implied probability of a favourite winning the match
r - safety threshold
b_u - upper boundary

And for an underdog as follows:

$$I_u(bet) = \begin{cases} 1 & \text{if } \frac{p_{M_u}}{p_{B_u}} - 1 \geq r \wedge p_{M_u} < \frac{1}{2} \wedge p_{B_u} < \frac{1}{2} \wedge p_{M_u} \geq b_l \wedge p_{B_u} \geq b_l \\ 0 & \text{else} \end{cases}$$

where:
p_{M_u} - model-implied probability of an underdog winning the match
p_{B_u} - bookmaker odds-implied probability of an underdog winning the match
r - safety threshold
b_l - lower boundary

The higher r we select, the stricter betting regime we follow. Nevertheless, with higher r, fewer bets are placed. To select the optimal safety threshold, we test six different levels of safety threshold - 1, 1.05, 1.1, 1.15, 1.20, and 1.25. The same logic applies for boundary levels. For upper boundary, we evaluate six different choices - 100%, 95%, 90%, 85%, 80%, and 75%, while for lower boundary, we examine six different values - 15%, 20%, 25%, 30%, 35%, and 40%.

As we aim to compare betting return between grand slam and non-grand-slam matches, the following four betting strategies are examined:

1. Betting exclusively on favourites for grand slam matches only
2. Betting exclusively on favourites for non-grand-slam matches only
3. Betting exclusively on underdogs for grand slam matches only
4. Betting exclusively on underdogs for non-grand-slam matches only

3.3.2 Bet Size

So far, the majority of authors have opted for constant bet size for each match (Forrest & McHale, 2005; McHale & Morton, 2011; Knottenbelt et al., 2012; Madurska, 2012). Nonetheless, several recent papers propose dynamic bet size based on Kelly criterion (Sipko & Knottenbelt, 2015; Lisi & Zanella, 2017).

Kelly criterion is designed to maximize the long-run wealth of the investor by maximizing the period by the period expected utility of wealth with a loga-

rithmic utility function (Macleand et al., 2011). For a binary outcome, bankroll proportion s placed for the match should maximize the following function:

$$f(s) = p\log(1+s) + (1-p)\log(1-s),$$

where:
p - probability of winning a bet
s - bankroll proportion to bet

Therefore, the bankroll proportion s that maximizes the expected wealth is calculated as:

$$s = \frac{Bp-1}{B-1},$$

where:
p - probability of winning a bet
B - bookmaker odds

Nevertheless, various authors argue that while Kelly criterion is asymptotically optimal, it is less appealing for short-run as it can lead to considerable losses a small percent of the time (Grant & Johnstone, 2010; Macleand et al., 2011; Baker & McHale, 2013). Therefore, the size of the bet is often shrunk, especially in the presence of parameter uncertainty, with the half-size to be widely used choice (Thorp, 1969; Baker & McHale, 2013).

In this thesis, both constant bet size and variable bet size based on half-sized Kelly criterion are tested.

As we consider 12 different model specifications [8], betting on both favourites and underdogs, betting on both grand slam and non-grand-slam matches, two different approaches to determine the bet size, six different levels of safety thresholds, and six boundary values, we evaluate 3456 different betting settings.

[8] We employ four machine learning algorithms and three set of features. For details on set of features applied see Subsection 4.2.1 Feature Engineering.

Chapter 4

Data Description and Manipulation

4.1 Data Sources and Their Manipulation

4.1.1 Overview of Data Sources

Our dataset combines data from three online sources. The core data source originates from www.tennisabstract.com and comprises information about match characteristics, player's physical characteristics, match outcome, and match statistics for both players since 1970. As data are not downloadable in standard forms such as Excel or CSV, we perform web scrapping through Wolfram Mathematica and Bash. Table 4.1 provides an overview of all information gathered from www.tennisabstract.com. The second online data source www.tennis-data.co.uk provides bookmaker odds for all ATP matches since 2001. Odds listed by 11 betting agencies[1] are included. Finally, we gather ATP points awarded for each win based on the tournament level and tournament round from www.atpworldtour.com.

[1]The comprehensive list of all betting agencies from which odds are collected can be found at http://tennis-data.co.uk/notes.txt. For the analysis, we use the best odds available defined as the maximum of all available odds at the time of the match. In reality, as one bets some time before the match, one is likely to obtain different bookmaker odds as bookmakers continually adjust the odds based on the total amount of bets placed on both players.

Table 4.1: Overview of Information Gathered from www.tennisabstract.com

Match Characteristics	**Player's Physical Attributes**	**Match Outcome**	**Match Statistics**
Tournament	Age	Score	Duration of Match
Tournament Level	Height	Winner of the Match	# Winners
Round	Country of Origin		# Unforced Errors
Surface	Playing Hand		# Double Faults
Date	Backhand Style		# Aces
ATP Official Rankings			First Serve Statistics
# Sets Needed to Win			Second Serve Statistics
			Break Points Statistics

Source: Author's own elaboration

4.1.2 Data Pre-Handling Flow

To obtain final data for analysis, the following intermediate steps are performed:

1. **Merging Data Sources**
   Firstly, we merge our data sources. Initially, we add ATP points that players are awarded after winning a particular match by merging an auxiliary table inferred from www.atpworldtour.com with our core data from www.tennisabstract.com. Subsequently, we join our newly obtained dataset with bookmaker odds. Due to the availability of bookmaker odds only since 2001, we proceed only with ATP matches since 2001. Both joins are performed in statistical software R by employing dplyr package for data manipulation. After performing both joins, we obtain Raw Data #1. At this point, our dataset comprises 40 897 ATP matches played during 2001-2016.

2. **PostgreSQL Calculations**[2]
   After obtaining Raw Data #1, we employ them in PostgreSQL to calculate historical match statistics through a different set of opponents, surfaces, tournament levels, and time spans. As a result, Raw Data # 2 is obtained.

[2] A few thousand lines of code are written to handle this intermediary step. SQL code is available on request.

3. **Data Cleansing**
   At this step, we perform the following data cleansing operations with Raw Data #2:

   (a) **Omission of variables with more than 50% missing values**
   Two variables (height, seed number) exhibit more than 50% missing values. As height is not widely shown to be significant across previous papers and seed number is extremely positively correlated with the official ATP rankings, we delete these two variables.

   (b) **Omission of matches with unlikely high margin and sure bets exhibiting unrealistic profitability**
   189 matches exhibit sure bets with profitability higher than 15% or betting agency's margin above 15% which is an obvious error.

   (c) **Omission of matches when the official ATP rankings is missing**
   398 matches do not include official ATP ranking for at least one player. As we cannot infer whether it is caused by an error or player is unranked, we rather delete these less than 1% observations.

   (d) **Omission of matches with obvious-error percentages**
   30 matches exhibit negative or above 100 percent for some of the percentage variables. We replace this values with NA and subsequently use K-nearest neighbours imputation (see next data cleansing step).

   (e) **Imputing missing values**
   209 out of 342 features have at least one value missing. In total, 1,3% values are not present in the dataset. As a widely used approach, K-nearest neighbour-based imputation is applied. This technique is carried out by finding the k closest samples (Euclidian distance). It imputes mode value for binary/dummy variables and mean for continuous ones.

After performing all previous data pre-handling steps, we obtain our dataset that contains 40 310 ATP matches and 342 features. As the last step, this dataset is split into training and test set:

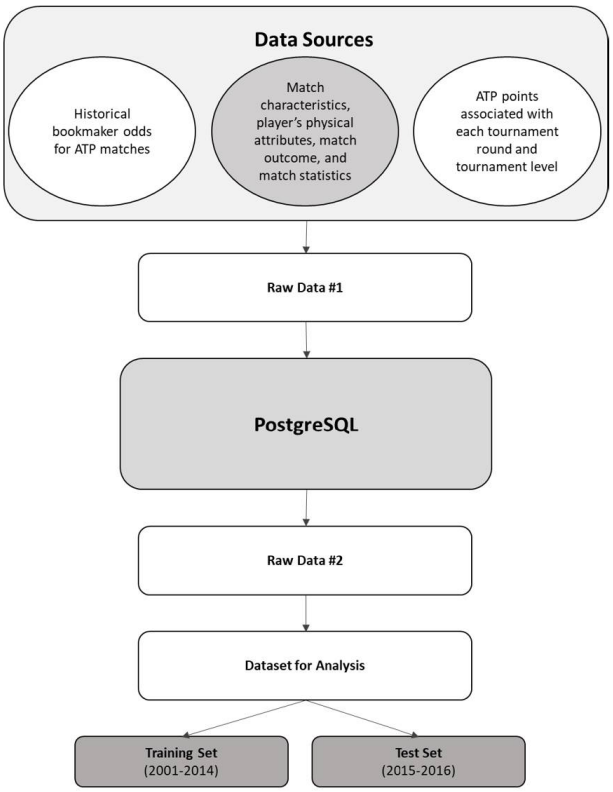
1. **Train Set (1/2001-12/2014)**
   The train set is utilized to train a model and obtain parameters estimates

that minimize selected loss criterion for training data. It is used to build the model and select tuning parameters. We select 35 628 ATP matches played during 2001-2014 as our test set.

2. **Test Set (1/2015-10/2016)**
   The test set is used to quantify the predictive performance of the model. It is utilized on previously unseen data. The most recent data (2015-2016) are used as the test set as they reflect the current competitiveness of betting market more closely than previous years. In total, 4 682 ATP matches are used in the test set.

Figure 4.1 depicts a schematic of whole data pre-handling flow.

Figure 4.1: A Schematic of Data Pre-Handling Flow



Source: Author's own elaboration

4.2 Feature Engineering and Feature Introduction

4.2.1 Feature Engineering

Dealing with Symmetric Match Representation

The tennis match has symmetric form. For illustration, winning a point by one player implies losing a point for his opponent. Also, better statistics on the serve of one player indicate worse return statistics of his opponents during the match. Thus, the efficient model has to consider characteristics of both players participating in a match. So far, there is no consensus on how information for both players should enter the model.

Various authors employ differences between two values. For instance, Sipko & Knottenbelt (2015) argue that keeping both values may cause the model to give different weights to those two initial inputs. Thus, under keeping both values separately, the predicted outcome could differ if players were re-labeled. Also, halving the number of features reduces the variance of the model and alleviates overfitting. Nevertheless, using differencing alone leads to a loss of information. For example, rank difference 50 cannot be perceived as the same across different rankings of better-ranked player. In other words, if player 1 is World Number One and player 2 is 51, the player 1 is arguably more likely to win than in the case of the rank of player 1 to be 51 and rank of player 2 to be 101. Taking simple differences fails to capture this well-known and intuitive fact. Alternatively, some studies such as neural network-based tennis modeling by Somboonphokkaphan et al. (2009), utilize information for both players and include two separate features in the model.

While inspired by recent approaches, we want to address their limitations. Therefore, we employ features in the following form. Firstly, we model outcome as the probability of a rank favourite winning match instead of match-winning probability of a randomly selected player. Thus, rank favourite always enters a model as Player 1, while the underdog is labeled as Player 2. This prevents labeling problem. Secondly, we experiment with the following three approaches for how to incorporate information for both players into the model:

1. Including original values for both players

2. Including favourite's original value along with difference between favourite's and underdog's value

3. Including favourite's original value along with the ratio of favourite's and underdog's value

By examining the prediction accuracy of all three features specifications across various set of features and models[3], we opt for including favourite's original value along with the difference between favourite's and underdog's value. This approach offers not only a slight improvement over other options regarding accuracy, but the variance[4] of the model is much lower.

Set of Features Used

In this thesis, we employ all four machine learning algorithms with three different set of features:

1. **Baseline Model**
   As a starting point, we include only nine features related to the official ATP rankings and bookmaker odds. Besides, we incorporate the dummy indicating that three sets are needed to win the match. As for rank-related features, we include dummy expressing that the match does not have a favourite based on the rank group, favourite's rank, and rank difference between players. For bookmaker odds-related features, we use external info, external info squared and the set of three dummies pinpointing the extent to which a rank favourite is also a bookmaker's favourite.

2. **Limited Model**
   To our best knowledge, only Lisi & Zanella (2017) include information about bookmaker odds among explanatory variables. To warrant the meaningful comparison with previous research, one of our specifications do not incorporate any knowledge about bookmaker odds into the model. As a result, limited models include 337 features.

[3]We use all original features as well as set of principal components with various thresholds (0.9,0.95) across penalized logistic regression, random forests, and flexible discriminant analysis.

[4]The variance and prediction accuracy of all model specification are assessed through 10-fold cross-validation.

3. **Full Model**
   The full model extends the limited model by including six bookmaker odds-related features. In total, we utilize 342 features in full models. While we expect this setting exhibiting higher prediction accuracy, the predicted probabilities will likely be highly correlated with bookmaker odds-implied ones and thus our betting rules may detect fewer matches to bet on.

The correlation between all model-implied and bookmaker odds-implied probabilities for all sets of features can be found in Appendix (Tables A.1-A.3).

Feature Selection

While presented features reduction techniques (see Subsection 4.3.2 Feature Reduction) alleviate overfitting, we still might deal with redundant features that may negatively affect model performance. Nevertheless, both penalized logistic regression and boosted trees have built-in feature selection procedures. Also, as presented in Kuhn & Johnson (2013), random forest performs well in the presence of irrelevant features. Also, weight decay in neural networks also controls for unnecessary features. The susceptibility of neural networks is further diminished by applying model averaging.

As our presented feature reduction and selection methods are based on simple heuristics, one could also opt for more elaborate alternatives such as wrapper methods[5]. Nevertheless, as these methods are much more computationally intensive in our setting[6], we do not utilize any of them. Furthermore, various simulations show that built-in feature selection procedures make algorithms reasonably susceptible to inclusion of redundant features (Kuhn & Johnson, 2013).

[5]For example, recursive feature elimination, genetic algorithms, or simulated annealing could be used. For more details, see Kuhn & Johnson (2013).

[6]Even employing 10-fold cross-validation for model tuning leads to the 2-day runtime for both support vector machines and neural networks. An additional optimization for feature selection would lead to the infeasible runtime for both models.

4.2.2 Features Introduction

In this thesis, we utilize 342 input features that can be categorized into 12 different groups. Tables 4.2 summarizes the number of features within each category.

Table 4.2: Number of Features Across Feature Categories

Category	**# Features**
Rank-related features	15
Bookmaker odds-related features	5
Player's physical attributes and fatigue	10
Player's mental toughness and consistency	22
Player's motivation	33
Player's historical in-match statistics	31
Match characteristics	14
Historical number of matches played and winning percentages	143
Common-opponents statistics	13
Surface-specific features	34
H2H-related features	12
Miscellaneous features	10

Source: Author's own elaboration

Rank-Related Features

As the majority of previous papers related to tennis modeling include ATP official rankings in some form, we include a set of rank-related features as well. Current rank of match favourite is a natural candidate and is included among put-aside variables. In addition, we include variables indicating career/last one-year best rank for both favourite and underdog, dummies indicating whether favourite or underdog is a former TOP10 or TOP11-30 player, dummies expressing whether the match favourite is currently ranked in TOP10/TOP11-30, dummy pinpointing that match is played between players from the same rank group[7], and dummies indicating whether favourite or underdog is a seeded player.

[7]We categorize players based on their current rank into five different rank groups - G1 for player from TOP10, G2 for players ranked 11-30, G3 for players ranked 31-100, G4 for 101-300, and G5 for players with rank higher than 300.

Bookmaker Odds-Related Features

Inspired by Lisi & Zanella (2017), we include information about bookmaker odds to account for the *external information* that is not captured by other features in the model. Variable *external info* equals 0 if the rank favourite is the same as bookmaker's favourite, while it takes rank favourite's odds if this player is bookmaker's underdog. Quadratic term for *external info* is also included. Besides, we include dummies indicating that bookmaker odds for rank favourite are below a certain threshold - 1.8, 1,5, and 1.25.

Player's Physical Attributes and Fatigue

As for physical attributes, we employ age, age squared and dummies representing whether a favourite or underdog is left-handed. Furthermore, following Sipko & Knottenbelt (2015), we employ proxies for fatigue - time spent on the current tournament before current match and number of matches played during last three months.

Player's Mental Toughness and Consistency

To our best knowledge, only Chitnis & Vaidya (2014) address the player's mental performance and consistency in their model. In this thesis, we significantly extend the set of features related to mental aspects of the tennis match. As proxies for player's mental toughness, we add the percentage of matches in which a player won the first set and eventually lost the match, the percentage of matches in which a player lost the first set and eventually won the match, and the percentage of matches won in which final decisive set was played. To account for consistency, we employ the percentage of matches in which a player lost while was a bookmaker's favourite and the percentage of matches in which a player won while was a bookmaker's underdog. All features here are calculated both during last three months and last one year.

Player's Motivation

To our best knowledge, only Sunde (2003) and Gilsdorf & Sukhatme (2008) address the effect of motivation on exerted effort and thus the probability of winning a match. In this thesis, we employ the number of ATP points the player would gain by winning a corresponding match, dummies indicating that favourite/underdog plays in his home country, and dummies for each level of match absolute quality [8] as proxies for motivation. Besides, we also include several interaction terms. Motivated by Lisi & Zanella (2017), we incorporate interaction term between home country dummy and absolute match quality quartiles for both favourite and underdog. Also, we employ interaction of ratio of favourite and underdog rank with the set of match characteristics dummies (tournament level, tournament round) to control for the fact that a better player's motivation may vary across different match characteristics.

Player's Historical In-Match Statistics

As for historical in-match statistics, we employ historical averages during whole career and last one year. As serve statistics are a common choice in previous papers, we include three serve statistics - the percentage of first serve points won, the percentage of serving points in which an ace was hit by a serving player, and percentage of serving points in which a double fault was made by a serving player. Besides serve statistics, to proxy for a performance during crucial points of the match, we include four in-match statistics related to break points performance - the number of break points created per match, percentage of break point chances converted, percentage of break points saved, and the ratio of later two variables as a measure of overall break point performance.

Match Characteristics

We apply set of dummies related to match characteristics. Separate dummies for each tournament level and tournament round are employed in our models to control for the fact that the favourite's performance is likely to vary across these factors. Moreover, dummies indicating whether two or three sets are

[8] We divide all matches into four groups based on the sum of favourite and underdog rank. The lower the sum, the higher absolute quality match is played.

needed to win a match are included as underdog is less likely to win three sets against a more-skilled player compared to only two sets.

Historical Number of Matches Played and Winning Percentages

Historical winning percentage in matches across different match conditions may be a strong indication of a player's performance once he faces the same condition in the future. Also, we utilize the number of matches across different factors as well. The number of matches across different categories is included due to the two main reasons. Firstly, the number of matches played at specific conditions (e.g., in the specific round, in specific tournament level, against the specific rank group) is used as a proxy for player's experience with particular match conditions. Secondly, using historical percentages alone might be misleading as some players have played only a few matches under the particular conditions and thus the sample to calculate historical averages is not sufficient[9]. Historical performance and the number of matches are calculated across different opponent's rank group, surface, tournament level, tournament round. Several timeframes are used for computation - last three months (as a proxy for current form), last one and two years (as a proxy for medium-term form), and whole career as a proxy for player's true ability. This feature category represents by far the most extensive set of features used in this thesis.

Common-Opponents Statistics

The simple averaging of player performance across all past matches is biased if two players have had different average opponents. Therefore, as suggested in Knottenbelt et al. (2012), Madurska (2012) and Sipko & Knottenbelt (2015), we alleviate this issue by using the set of features that consider only matches played against common opponents. The number of matches played against common opponents along with winning percentage in those matches during the whole career, last two years, and last one year are employed. Besides, average opponent rank during last one year is included for both favourite and underdog in our models.

[9]If we compare the winning percentage of a player that won 19 out of 20 matches with winning percentage of a player that won his single match under particular condition, the second player would be favoured if we include winning percentages only.

Surface-Specific Features

Tennis is played on four main surfaces - clay, hard, grass, and carpet; with clay and hard court accounting for the majority of matches. Surfaces differ in their characteristics such as speed and bounce of the ball, or physiological and physical demands (Fernandez et al., 2006). As a result, players perform differently across surfaces as some playing styles are more effective on the particular surface. We account for surface effect by including the number of matches previously played on each surface along with winning percentages during whole career, last two years, and last one year. Furthermore, separate dummies for each surface are included among input features. Also, dummies indicating that a match is played on favourite's/underdog's most favourite/least favourite surface[10] are utilized in our models.

H2H-Related Features

As player are susceptible to different playing styles, previous results against the current opponent might be a strong indicator of how the player will perform in the next match against him. To account for head-to-head balance, the number of matches, sets, games, and tiebreaks played in all previous encounters along with their winning percentages are used in our models.

Miscellaneous Features

Besides all features above, we experiment with few more. We utilize dummies representing that favourite/underdog had to win qualification matches to get to the main draw, dummies indicating that favourite/underdog obtained a wild card, and dummies pinpointing that favourite/underdog is a lucky loser[11]. Also, we include dummies indicating that favourite/underdog is close to his best form[12]. Finally, we utilize dummies expressing that favourite's/underdog's last

[10] To determine the most/least favourite surface, we calculate historical winning percentage on each surface at the time of the match. The surface with the highest value is selected as favourite, while the one with the lowest is perceived as the least favourite.

[11] A player lost in the final qualification match, but as some players accepted to the main draw cancel their participation in the tournament, the 'loser' eventually gets to the main draw despite losing his final qualification match

[12] If player's current ranking is at most 120% of his career-best rank, a player is expected to play close to his best.

tournament was successful[13].

Descriptive Statistics

Here we present descriptive statistics for rank and bookmaker odds. We also show the proportion of matches played across different categories.

Table 4.3: Descriptive Statistics for Ranking and Boookmaker Odds - Train Set

	Min	Q1	Median	Mean	Q3	Max
Ranking - favourite	1.00	12.00	30.00	38.66	55.00	1078.00
Ranking - underdog	2.0	51.0	81.0	108.6	122.0	2159.0
Bookmaker odds - favourite	1.010	1.270	1.480	1.632	1.781	16.500
Bookmaker odds - underdog	1.090	2.200	2.900	4.322	4.350	121.000

Note: Q1 - 1st quartile, Q3 - 3rd quartile

Source: Author's own elaboration in R

Table 4.4: Descriptive Statistics for Ranking and Boookmaker Odds - Test Set

	Min	Q1	Median	Mean	Q3	Max
Ranking - favourite	1.00	13.00	28.00	37.1	53.00	837.00
Ranking - underdog	2.0	48.0	77.0	108	120	1809.0
Bookmaker odds - favourite	1.001	1.240	1.450	1.623	1.760	11.170
Bookmaker odds - underdog	1.080	2.240	3.050	4.573	4.740	81.000

Note: Q1 - 1st quartile, Q3 - 3rd quartile

Source: Author's own elaboration in R

[13]Last tournament is considered to be successful if the player's rank after the last tournament improved compared to the rank he started the last tournament with

Table 4.5: Proportion of Matches Played

	Train Set	**Test Set**	**Category**
# Matches - Total	35 628	4 682	
Grand Slam	19.3%	18.5%	Tournament Level
Non-Grand-Slam	80.7%	81.5%	Tournament Level
Clay	33.65%	33.79%	Surface
Hard	51.41%	51.39%	Surface
Grass	11.50%	11.62%	Surface
Carpet	3.44%	3.20%	Surface
Favourite - TOP10	21.48%	21.21%	Favourite's Rank
Favourite - TOP30	51.25%	53.25%	Favourite's Rank

Source: Author's own elaboration in R

4.3 Feature Pre-Processing

In this section, we present all feature pre-processing operations.

4.3.1 Data Transformations

Following James et al. (2013) and Kuhn & Johnson (2013), we perform both centering and scaling for all non-binary variables[14]. As a result of centering, the feature has a zero mean, while by scaling data are coerced to have a standard deviation of one. These two common data transformations are used to improve the numerical stability of various calculations (Kuhn & Johnson, 2013).

Formally, the transformed variable is obtained as:

$$X_{new} = \frac{X_{original} - \bar{X}_{original}}{\sigma_{X_{original}}},$$

where:
X_{new} - new variable obtained after standardization
$X_{original}$ - an original variable
$\bar{X}$ - the mean of an original variable
$\sigma_{X_{original}}$ - the standard deviation of an original variable.

4.3.2 Feature Reduction

In addition to learning general patterns in the data, the models also learn unique noise of each sample. Using too many features (especially irrelevant ones) leads to complex models that can overemphasize patterns in the data that are not reproducible. This phenomenon is widely known as overfitting. Such models exhibit great accuracy on training data, but poor performance on previously unseen data. To mitigate overfitting, we reduce the number of features used in models through removing predictors with near-zero variance, removing highly-correlated features and principal components.

[14]While the majority of our variables is continuous, some exhibit discrete values, but over large set of values (e.g., the number of matches played ranking from 0 to 1 000).

Removing Predictors with Near-Zero Variance

Employing variables with sparse values and unbalanced distributions can invalidate our modeling procedures. Thus, inspired by Kuhn & Johnson (2013), we filter out near-zero variance predictors. This procedure removes features with the frequency of unique values being severely disproportionate or with only one unique values.

Removing Highly-Correlated Features

Using highly correlated predictors can result in numerical errors, highly unstable models and degrade predictive performance (Kuhn & Johnson, 2013). Also, including redundant features increases the complexity of the models and thus their computational time[15].

Inspired by Kuhn & Johnson (2013), we apply the following heuristic algorithm for removing features that ensures that all pairwise correlations are below a selected threshold:

1. Calculating correlation matrix of all features
2. Determining the two predictors with the largest pairwise correlation (call them X and Y)
3. Determining average correlation of X and Y with all other features
4. Removing either X or Y, depending on which one exhibits larger average correlation with other predictors
5. Repeating Steps 2-4 until no pairwise correlation is above a selected threshold

While this method only investigates pairwise correlation, it can have a substantially positive effect on the performance of the models (Kuhn & Johnson, 2013).

[15] As we employ computationally intensive models such as support vector machines and neural networks, this issue is crucial in our modeling .

Constructing Principal Components

Principal component analysis (hereafter PCA) is a commonly used unsupervised technique for feature reduction. This approach seeks to find linear combinations of features, known as principal components (hereafter PCs), which capture the most variance. The first PC is constructed as the linear combination of the features that captures the most variability of all possible linear combinations. Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs. Formally:

$$PC_k = (s_{j1} * Feature \quad 1) + (s_{j2} * Feature \quad 2) + \ldots + (s_{jP} * Feature \quad P),$$

where:
P - number of features
$s_{j1}, s_{j2}, \ldots s_{jP}$ - component weights that imply which features are most important to each PC

We construct PCs until 90% of the original variance is captured by our set of PCs.

Following Kuhn & Johnson (2013), data transformation and feature reduction procedures are performed in the following order:

1. Near-zero variance filter
2. Correlation filter
3. Centering
4. Scaling
5. Principal components construction

Table 4.6 summarizes data transformation and feature reduction procedures that we perform for both binary and continuous variables.

Table 4.6: Overview of Data Transformation and Feature Reduction Procedures

Non-Binary Variables	**Binary Variables**
Near-zero variance filter	Near-zero variance filter
Centering	Correlation filter
Scaling	
Principal components construction	

Source: Author's own elaboration

For put-aside variables (favourite's rank, rank difference, and five bookmaker-odds related features), we perform all procedures, except near-zero variance filter and correlation filter as we aim to keep all of them in our models. As a result of all aforementioned procedures, we decrease the number of features from 342 (55 binary, seven put-aside, and 280 non-binary) to 122 (43 binary, 72 PCs, and seven put-aside) for the full model and 117 (43 binary, 72 PCs, and two put-aside) for the limited model.

4.4 Model Tuning

Many models have essential parameters which cannot be directly estimated from the data. All machine learning techniques discussed in this text have at least one tuning parameter. Since many of these parameters control the complexity of the model, poor choices for the values can result in overfitting.

There are different approaches [16] to searching for the optimal tuning parameters. A general approach is to define a set of candidate values. In this thesis, we employ a pre-defined set of tuning parameters accessible in *R package caret* for all four machine learning algorithms.

Table 4.7 presents all tuning parameters that are optimized in our models.

Table 4.7: Overview of Optimized Tuning Parameters

Penalized Logistic Regression	**Random Forest**	**Boosted Trees**	**Neural Networks**
α - relative weight of lasso and ridge penalty	mtry - # randomly selected predictors	iter - # trees	W - weight decay
λ - penalty term		maxdepth - max tree depth	H - # hidden neurons
		nu - learning rate	

Source: Author's own elaboration

Once we select a set of candidate values, we must obtain reliable estimates of model performance for each combination. For this purpose, resampling techniques are often used.

4.4.1 K-Fold Cross-Validation

While alternative resampling techniques such as bootstrapping can be used, we employ K-fold cross-validation in this thesis. K-fold cross-validation uses part of available data as a training set and the rest of the data as a test set.

[16] As an alternative to a pre-defined set of candidates, one could opt for the random search. Nevertheless, due to its excessive computational time, this method is not utilized in this thesis.

Firstly, we split the data into K roughly equally-sized parts. Subsequently, we fit the model using K-1 parts, with the remaining k_{th} part used to determine the prediction error of the fitted model. We repeatedly estimate the model for $k = 1, 2, \ldots, K$ and combine the K estimates of prediction error. The performance on the K hold-out samples is then aggregated into a performance profile which is then used to determine the final tuning parameters. Although there is no formal rule for the selection of k, we opt for 10-fold cross-validation as this choice has desirable properties both regarding computational efficiency and bias-variance trade-off (Kuhn & Johnson, 2013).

In addition to obtaining optimal tuning parameters, 10-fold cross-validation allows us to produce appropriate estimates of model performance using the test set. As several researchers show that validation using a single test can be a poor choice (Martin & Hirschberg 1996; Hawkins et al. 2003; Molinaro et al., 2005), by employing 10-fold cross-validation, we can obtain more robust estimates of the model's predictive performance.

Once the model performance has been quantified across sets of tuning parameters, we opt for one with the best mean prediction accuracy[17]. Subsequently, we refit the model with the entire training set by using the final tuning parameters.

Table 4.8 presents values of all tuning parameters for all full models.

Table 4.8: Selected Values of Tuning Parameters - Full Model

LR	**RF**	**BT**	**NN**
$\alpha = 0.55$	mtry=62	iter=100	H=3
$\lambda = 0.027$		maxdepth=1	W=0.0001
		nu=0.1	

Note: LR - logistic regression, RF - random forest, BT - boosted trees, NN - neural network

Source: Author's own elaboration

[17]. Alternatively, one could favour less complex models providing that they still yield acceptable performance. The conventional approach is to use "one standard error rule" that chooses the simplest model whose performance is still within one standard error of the numerically best model. Nevertheless, as the interpretability of our model is not the primary concern, we select the one with the best performance regardless its interpretability.

Table 4.9 presents values of all tuning parameters for all limited models.

Table 4.9: Selected Values of Tuning Parameters - Limited Model

LR	**RF**	**BT**	**NN**
$\alpha = 0.55$	mtry=59	iter=150	H=3
$\lambda = 0.002$		maxdepth=3	W=0.1
		nu=0.1	

Note: LR - logistic regression, RF - random forest, BT - boosted trees, NN - neural network

Source: Author's own elaboration

Finally, Table 4.10 outlines values of all tuning parameters for all baseline models.

Table 4.10: Selected Values of Tuning Parameters - Baseline Model

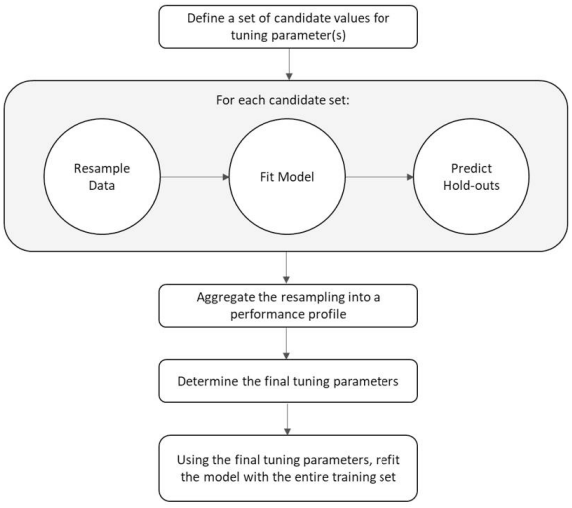
LR	**RF**	**BT**	**NN**
$\alpha = 0.55$	mtry=2	iter=150	H=1
$\lambda = 0.003$		maxdepth=2	W=0.1
		nu=0.1	

Note: LR - logistic regression, RF - random forest, BT - boosted trees, NN - neural network

Source: Author's own elaboration

Figure 4.2 summarizes a schematic of the whole parameter tuning process.

Figure 4.2: A Schematic of the Parameter Tuning Process



Source: Kuhn & Johnson (2013)

Chapter 5

Results and Discussion

In this chapter, we assess and compare the predictive performance, the betting performance, and the variable importance.

5.1 Predictive Performance

We use two evaluation metrics to assess and compare the predictive performance - the overall accuracy rate and AUC.

5.1.1 Overall Accuracy Rate

As for the comparison with both benchmarking rules, all model specifications for both train and test data outperform backing a better-ranked player (the difference ranges from 2% to 6%). As for backing a bookmaker's favourite, only all baseline and full model specifications (except the full model for neural networks) perform better, indicating that without any knowledge about betting odds, our models are not able to beat bookmakers' models.

Except for the full model for neural networks, all models perform better on the test data. There are two main explanations. Firstly, both better-ranked players (67.3% compared to 65.5%) and bookmaker's favourites (70.7% compared to 69.7%) won more regularly during 2015-2016. As can be seen from variable importance (see subsection Variable Importance), all our models are

strongly influenced by bookmaker odds and rank-related features. Secondly, the application of 10-fold cross-validation prevents overfitting on the train data[1].

As for the best-performing model, the full model for neural networks [2] along with the baseline model for random forest yield the best overall accuracy rate. For the test data, the full model for boosted trees and the baseline model for random forest perform the best. Nevertheless, differences across models are relatively small - for the train data, the difference between the best-performing and the worst-performing model is 3%[3], while for the test data only 1.6%.

For all models, benchmarking rules and data sets, the overall accuracy for grand slam matches is significantly higher than the overall accuracy for non-grand-slam matches (the difference ranges from 6.9% to 10.1%).

One can note that both the baseline and the full set of features outperform the limited set of features indicating that the overall accuracy is negatively affected by omitting bookmaker odds-related features. As the baseline models consistently bet limited models, we can infer that the inclusion of our more than 300 variables cannot sufficiently compensate for missing information about the bookmaker odds.

To compare with previous papers, the overall prediction accuracy exceeds the current state-of-art models for non-grand-slam matches (67%) evaluated by Kovalchik (2016). Also, we surpass current state-of-art models that employ grand slam matches exclusively (Somboonphokkaphan et al., 2009 - 74%; Lisi & Zanella, 2017 - 77%).

[1] To obtain the overall accuracy rate for the train set, we firstly find the best model by using 10-fold cross-validation, and then apply the selected model on the full train set. For more details, see Section Model Tuning in Chapter Methodology.

[2] As can be seen from the comparison of train and test overall accuracy rate, the full model for neural networks is the only model specification performing worse on the test data. As this algorithms does not have a built-in feature selection mechanism (as opposed to penalized logistic regression and boosted trees), it is likely that we slighly overfit despite using 10-fold cross-validation.

[3] If we omit the full model for neural networks, the difference is only 1.9%.

Table 5.1 outlines the overall accuracy rate across all model specifications for both train and test data. Also, it provides the overall accuracy rate for our two benchmarking rules.

Table 5.1: Overall Accuracy Rate - Overview

	Train Set			**Test Set**		
Model	All	NG	G	All	NG	G
LR - full	69.9%	68.4%	76.1%	70.8%	68.9%	79.0%
RF - full	69.9%	68.5%	76.2%	70.4%	68.7%	77.9%
BT - full	69.9%	68.3%	76.3%	70.9%	69.0%	79.0%
NN - full	71.5%	70.1%	77.4%	70.4%	68.5%	78.7%
LR - limited	69.1%	67.7%	75.1%	69.9%	68.1%	78.1%
RF - limited	69.0%	67.7%	74.6%	69.3%	67.7%	76.2%
BT - limited	68.5%	67.1%	74.4%	69.7%	68.0%	77.1%
NN - limited	69.1%	67.7%	74.8%	69.7%	67.9%	78.0%
LR - baseline	69.9%	68.4%	76.1%	70.8%	68.9%	78.9%
RF - baseline	70.4%	68.9%	76.8%	70.9%	69.1%	78.8%
BT - baseline	69.9%	68.4%	76.1%	70.8%	69.0%	78.8%
NN - baseline	69.9%	68.4%	76.3%	70.8%	68.9%	78.9%
Better-ranked player	65.5%	64.0%	71.7%	67.3%	65.4%	75.6%
Bookmaker's favourite	69.7%	68.2%	76.1%	70.7%	68.8%	79.0%

Note: All - all matches, NG - non-grand-slam matches only, G - grand slam matches only

Source: Author's own elaboration

For details on the baseline, the limited, and the full set of features see Subsection 4.2.1 Feature Engineering.

5.1.2 AUC

As for AUC, the inference made for the overall accuracy rate holds with three exceptions. Firstly, the full model for boosted trees is no longer among the best-performing models on the test data. Baseline models for logistic regression and boosted trees along with the full model for neural networks exhibit the highest AUC. Secondly, the limited model for neural networks outperforms some full and baseline models, indicating that the limited models are not inferior in all cases. Finally, despite one of the best overall accuracy rates on the test data, the baseline model for random forest shows the second lowest AUC.

Table 5.2 presents AUC across all model specifications for both train and test data.

Table 5.2: AUC - Overview

	Train Set			Test Set		
Model	All	NG	G	All	NG	G
LR - full	0.716	0.698	0.777	0.715	0.698	0.773
RF - full	0.715	0.697	0.775	0.703	0.686	0.755
BT - full	0.714	0.697	0.774	0.711	0.695	0.759
NN - full	0.750	0.735	0.805	0.716	0.698	0.773
LR - limited	0.710	0.694	0.764	0.710	0.694	0.756
RF - limited	0.709	0.693	0.759	0.683	0.670	0.726
BT - limited	0.706	0.690	0.758	0.700	0.685	0.752
NN - limited	0.724	0.707	0.779	0.700	0.686	0.748
LR - baseline	0.715	0.697	0.771	0.716	0.698	0.773
RF - baseline	0.733	0.716	0.806	0.695	0.682	0.752
BT - baseline	0.717	0.699	0.776	0.716	0.697	0.770
NN - baseline	0.716	0.697	0.771	0.714	0.695	0.768

Note: All - all matches, NG - non-grand-slam matches only, G - grand slam matches only

Source: Author's own elaboration

5.2 Betting Performance

5.2.1 Betting Benchmarks

To assess the profitability of our two selected betting benchmarks, we consider both constant and variable bet size [4] . We employ a constant bet size set to 100, while variable bet size is determined as follows:

$$bet \quad size = \frac{\bar{b}_f * 100}{b_s},$$

where:
$\bar{b}_f$ - average bookmaker odds for a better-ranked player
b_s - bookmaker odds for the player we want to bet on

The average bookmaker odds for a better ranked player are 1.62 for the test set and 1.63 for the train set.

The results show that both benchmark rules lead to a loss if applied to non-grand-slam matches exclusively. The loss for the train set ranges from -0.98% to -2.96%, while it fluctuates from -1.18% to -1.93% for the test set. Although for the train set the loss is more pronounced for betting on a better-ranked player, the opposite is true for the test set.

As for betting on grand slam matches, betting on a bookmaker's favourite is profitable for both train (ROI 1.25-1.4%) and test set (ROI 1.87-2.05%). The betting on a better-ranked player for grand slam matches is shown to be profitable under both bet size for the test set (ROI 1.87-2.2%) and under variable bet size for the train data (ROI 0.61%). The evidence of the positive ROI for betting on favourites applied to grand slam matches favours the existence of favourite-longshot bias and support the results of Forrest & McHale (2005).

As for the comparison of the constant bet size with the variable bet size, the variable bet size leads to a higher profit if applied to both train and test set. The evidence on the magnitude of the loss for non-grand-slam matches varies.

[4] As we cannot use Kelly criterion without the model-implied probability, we employ variable bet size as an alternative.

Tables 5.3-5.4 present the performance of both betting benchmarks on the train set.

Table 5.3: Betting Benchmarks on Train Set - Total Profit

	Constant = 100		**Variable = 163/odds**	
Model	NG	G	NG	G
Better-ranked player	-85065	-2498	-69593	4854
Bookmaker's favourite	-31918	8567	-31612	11787

Source: Author's own elaboration

Table 5.4: Betting Benchmarks on Train Set - ROI

	Constant = 100		**Variable = 163/odds**	
Model	NG	G	NG	G
Better-ranked player	-2.96%	-0.36%	-2.27%	0.61%
Bookmaker's favourite	-1.11%	1.25%	-0.98%	1.40%

Source: Author's own elaboration

Tables 5.5-5.6 depict the performance of both betting benchmarks on the test set.

Table 5.5: Betting Benchmarks on Test Set - Total Profit

	Constant = 100		**Variable = 162/odds**	
Model	NG	G	NG	G
Better-ranked player	-4485	1618	-4839	2286
Bookmaker's favourite	-7343	1774	-7472	2034

Source: Author's own elaboration

Table 5.6: Betting Benchmarks on Test Set - ROI

Model	Constant = 100		Variable = 162/odds	
	NG	G	NG	G
Better-ranked player	-1.18%	1.87%	-1.18%	2.20%
Bookmaker's favourite	-1.93%	2.05%	-1.73%	1.87%

Source: Author's own elaboration

5.2.2 Searching for the Most Optimal Betting Strategy

To find the most optimal strategy, we firstly find the most optimal betting setting for each model specification (12 model specifications in total) across all four main betting strategies (F-NG: betting of favourites for non-grand-slam matches only; F-G: betting of favourites for grand slam matches only; U-NG: betting of underdogs for non-grand-slam matches only and U-G: betting of underdogs for grand slam matches only) on the train set. As a result, we seek the optimal solution for 48 different combinations. For each of them, we select the most profitable strategy that meets the following two criteria:

1. **The number of matches to bet on is at least 1000 for non-grand-slam matches and 250 for grand slam matches**
   As we want to identify consistent strategies, we omit strategies with only a small fraction of matches eligible for betting.

2. **ROI of selected betting setting is at least 2%**
   We disregard all settings with profitability below betting benchmarks.

The search is performed through six different safety threshold levels, six lower/upper boundary values, and two bet sizes - constant bet size and the best sized based on the half-sized Kelly criterion. Therefore, for each of 48 combinations, 72 different betting settings are tested.

We manage to find at least one profitable betting settings with both ROI and the number of matches to bet on exceeding presented thresholds for 35 out of 48 combinations. The full and the limited model for neural networks exhibit the best performance followed by the limited model for penalized logistic regression and the full model for the random forest.

Table 5.7 presents the rankings of all models for each of the four main betting strategies applied on the train set. The ranking is based on the total profitability.

Table 5.7: Ranking of Betting Strategies - Train Set

	F-NG	**F-G**	**U-NG**	**U-G**
LR - full	-	10	10	-
RF - full	5	8	5	4
BT - full	-	11	-	-
NN - full	1	1	1	1
LR - limited	3	4	4	6
RF - limited	6	-	8	3
BT - limited	4	5	3	-
NN - limited	2	2	2	2
LR - baseline	-	7	9	7
RF - baseline	-	3	-	-
BT - baseline	7	8	7	4
NN - baseline	-	6	6	-

Source: Author's own elaboration

Subsequently, we evaluate all 35 scenarios on the test data. 17 out of 35 scenarios exhibit the positive return on the test data. While six out of seven betting scenarios for betting on favourites applied to non-grand-slam matches and 11 out of 11 betting scenarios for betting on favourites applied to grand slam are profitable on the test set, none of the betting strategies for betting on underdogs exhibits a positive return once employed on the test data. The full models for neural networks and random forest show the best performance regarding the total profitability.

Table 5.8 presents the rankings of all models for each of the four main betting strategies applied on the test set. The ranking is based on the total profitability.

Table 5.8: Ranking of Betting Strategies - Test Set

	F-NG	**F-G**	**U-NG**	**U-G**
LR - full	-	8	-	-
RF - full	1	2	-	-
BT - full	-	7	-	-
NN - full	2	1	-	-
LR - limited	3	11	-	-
RF - limited	5	-	-	-
BT - limited	6	9	-	-
NN - limited	-	3	-	-
LR - baseline	-	4	-	-
RF - baseline	-	6	-	-
BT - baseline	4	10	-	-
NN - baseline	-	5	-	-

Source: Author's own elaboration

In majority of cases, ROI is higher for betting on grand slam matches. Although we do not successfully replicate any of the betting strategies for betting on underdogs, the identification of several profitable strategies for betting on underdogs on the train set indicates that other approaches to betting strategy selection should be examined. We assume that with more comprehensive approach to betting strategy selection, one could identify consistently profitable betting strategies for betting on underdogs as well [5].

Table 5.9 outlines the median ROI for each of the four main betting strategies applied to both test and train set.

Table 5.9: Median ROI of Betting Strategies

	F-NG	**F-G**	**U-NG**	**U-G**
Train Set	6.8%	5.5%	3.7%	6.4%
Test Set	1.6%	2.9%	-	-

Source: Author's own elaboration

The analysis of betting on favourites applied to non-grand-slam matches favours the full model for random forest and the full model for neural networks

[5] Our in-sample optimization for test set reveals various profitable betting settings for betting on underdogs, both for non-grand-slam and grand slam matches.

as two most profitable options in absolute terms [6]. Regarding ROI, the baseline model for boosted trees is preferred. Overall, ROI fluctuates between 0.8% and 6.5%. As for the proportion of matches to bet on, models range from 5% to 46%.

Table 5.10 outlines the betting strategies for betting on favourites applied to non-grand-slam matches.

Table 5.10: Betting on Favourites for Non-Grand-Slam Matches - Overview

Model	**Bet Size**	**Safety Threshold**	**Upper Boundary**	**Profit**	**ROI**	**# Bets**	**% Matches to Bet on**
RF - full	Kelly	1.05	90%	3000	2.4%	606	16%
NN - full	Kelly	1.00	100%	2000	0.8%	1737	46%
LR - limited	Kelly	1.00	100%	1800	2.3%	820	22%
BT - baseline	Kelly	1.10	100%	1200	6.5%	630	17%
RF - limited	Kelly	1.00	95%	1000	0.8%	843	22%
BT - limited	Kelly	1.15	100%	490	0.9%	200	5%
Average		1.05	98%	1582	2.3%	806	21%

Source: Author's own elaboration

The analysis of betting on favourites applied to grand slam matches favours the full model for random forest and the full model for neural networks as two most profitable options in absolute terms. Regarding ROI, the full model for boosted trees is preferred. Overall, ROI fluctuates between 0.7% and 9.3%. As for the proportion of matches to bet on, models range from 13% to 44%. The betting on favourites for grand slam matches leads to both the higher ROI (4% compared to 2.3%) and the higher proportion of matches to bet on (30% compared to 21%) as compared to betting on favourites for non-grand-slam matches. Therefore, our results strongly favour betting on grand slam matches once the betting on favourites is considered.

[6] To calculate the total profitability (column profit) and the bet size based on the half-sized Kelly criterion, we assume the initial bankroll of 2000 USD. For the simplicity and the meaningful comparison with the previous models, the bet size is determined based on the initial balance and does not evolve dynamically with the bankroll balance.

Table 5.11 shows the betting strategies for betting on favourites applied to grand slam matches.

Table 5.11: Betting on Favourites for Grand Slam Matches - Overview

Model	Bet Size	Safety Threshold	Upper Boundary	Profit	ROI	# Bets	% Matches to Bet on
NN - full	Kelly	1.00	100%	4900	7.4%	344	40%
RF - full	Kelly	1.05	100%	2700	6.1%	130	15%
NN - limited	Kelly	1.00	95%	2000	2.5%	322	37%
LR - baseline	Kelly	1.00	100%	1800	2.9%	383	44%
NN - baseline	Kelly	1.00	100%	1700	3.0%	373	43%
RF - baseline	Kelly	1.25	100%	1500	1.7%	350	40%
BT - full	constant	1.00	100%	1400	9.3%	153	18%
LR - full	Kelly	1.00	100%	740	5.6%	179	21%
BT - limited	Kelly	1.05	100%	470	1.8%	117	13%
BT - baseline	Kelly	1.00	100%	450	1.2%	320	37%
LR - limited	Kelly	1.00	95%	280	0.7%	226	26%
Average		1.03	99%	1631	4%	263	30%

Source: Author's own elaboration

One can notice that the half-sized Kelly criterion puts only soft restrictions on both safety threshold (1.05 for non-grand-slam matches and 1.03 for grand slam matches on average) and the upper boundary (98% for non-grand-slam matches and 99% for grand slam matches on average). This finding is not surprising as unlike the constant bet size, the bet size based on the half-sized Kelly criterion is adjusted automatically based on the size of the encountered betting edge and bookmaker odds.

Although some of the previous papers show higher ROI (Lisi & Zanella, 2017 - 16.3% for grand slam matches) or the higher proportion of matches to bet on (Sipko & Knottenbelt, 2015 - 50.4% for all matches), one should be cautious about making the comparison with their betting performance. From the previous papers, it is not clear whether the presented betting strategies were selected on the train set and subsequently evaluated on the test set or authors performed in-sample optimization only. Therefore, their performance might be artificially high. As we initially find the optimal betting settings on the train data and eventually assess their performance on the test set, our betting strategies should be reasonably robust, and we should obtain a realistic estimate of their performance on previously unseen matches.

5.2.3 Determining the Optimal Bet Size

As for the optimal bet size, the bet size determined by the half-sized Kelly criterion is consistently preferred for betting on favourites. Only one out of 18 identified betting scenarios for betting on favourites prefers the constant bet size. On the contrary, the bet size determined by the half-sized Kelly criterion is preferred in only seven out of 17 selected scenarios for betting on underdogs.

Table 5.12 summarizes the proportion of optimal betting strategies with bet size based on the Kelly criterion.

Table 5.12: Proportion of Optimal Betting Strategies with Bet Size Based on the Kelly Criterion

	F-NG	**F-G**	**U-NG**	**U-G**
Train Set	7/7	10/11	3/10	4/7
Test Set	6/6	10/11	-	-

Source: Author's own elaboration

The analysis of the identified betting scenarios shows that the betting that utilizes the half-sized Kelly criterion leads to a significantly larger average bet size for betting on favourites - from 1.46 to 1.91 times larger compared to a constant size. As for betting on underdogs, bet sizes are roughly the same.

Table 5.13: Median Kelly Bet Size Across Models

	F-NG	**F-G**	**U-NG**	**U-G**
Train Set	177	174	101	102
Test Set	146	191	-	-

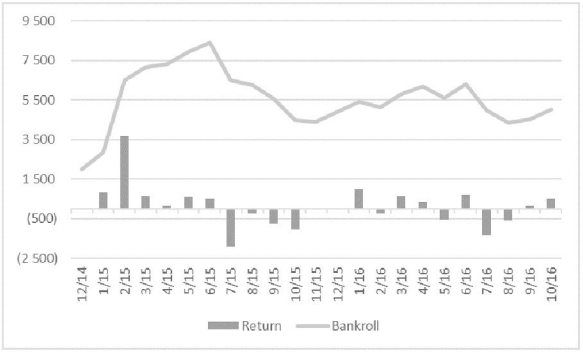
Source: Author's own elaboration

5.2.4 Payoff Profile of the Best Models

Here we present the payoff profile for three best-performing models for both betting on favourite applied to grand slam matches and betting on favourites applied to non-grand-slam matches. Besides, we compare ROI and % of matches to bet on between train and test set for all best-performing strategies.

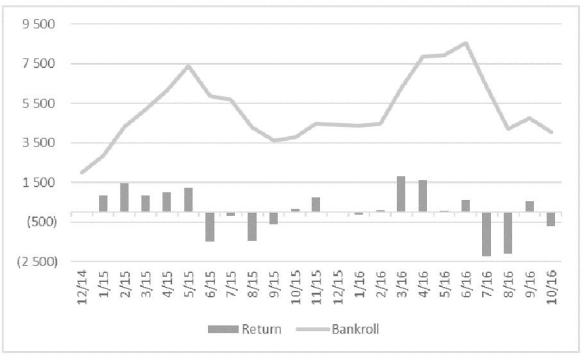
Figures 5.1-5.3 show the payoff profile for three best-performing strategies for betting on favourites applied to non-grand-slam matches. Full models for random forest and neural networks exhibit the highest absolute profitability, while the baseline model for boosted trees exhibits that highest ROI.

Figure 5.1: Random Forest for Non-Grand-Slam Matches - Payoff Profile



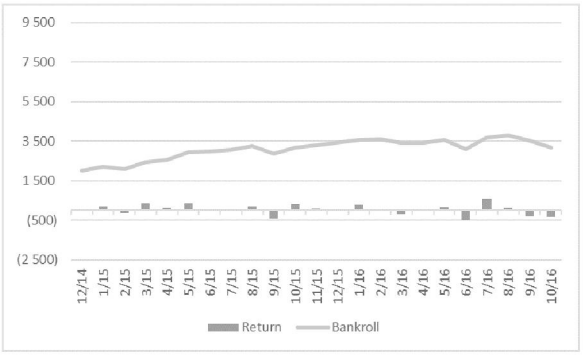
Source: Author's own elaboration

Figure 5.2: Neural Networks for Non-Grand-Slam Matches - Payoff Profile



Source: Author's own elaboration

Figure 5.3: Boosted Trees for Non-Grand-Slam Matches - Payoff Profile



Source: Author's own elaboration

Table 5.14 outlines the comparison of ROI and % of matches to bet on between train and test set for betting on favourites applied to non-grand-slam matches.

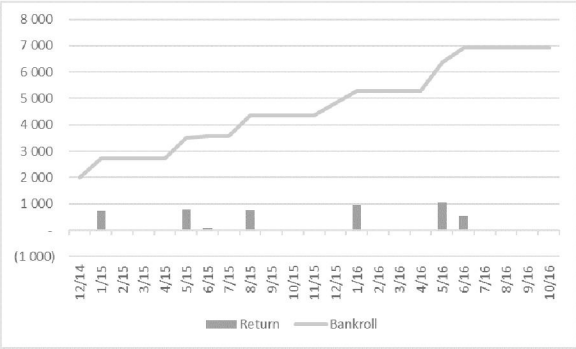
Table 5.14: Best Models for Non-Grand-Slam Matches - Overview

	Train Set		Test Set	
Model	ROI	% Matches	ROI	% Matches
RF - full	2.8%	17.0%	2.4%	15.9%
NN - full	14.0%	49.0%	0.8%	45.5%
BT - baseline	3.9%	7.0%	6.5%	16.5%

Source: Author's own elaboration

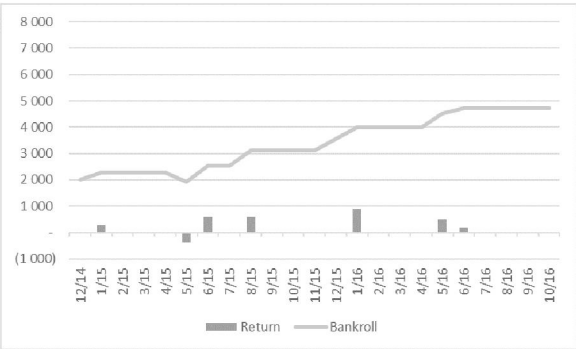
Figures 5.4-5.6 show the payoff profile for three best-performing strategies for betting on favourites applied to grand slam matches. Full models for random forest and neural networks exhibit the highest absolute profitability, while the full model for boosted trees exhibits that highest ROI.

Figure 5.4: Neural Networks for Grand Slam Matches - Payoff Profile



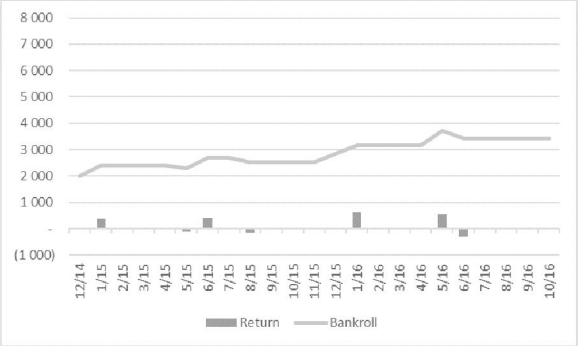
Source: Author's own elaboration

Figure 5.5: Random Forest for Grand Slam Matches - Payoff Profile



Source: Author's own elaboration

Figure 5.6: Boosted Trees for Grand Slam Matches - Payoff Profile



Source: Author's own elaboration

Table 5.15 outlines the comparison of ROI and % of matches to bet on between train and test set for betting on favourites applied to grand slam matches.

Table 5.15: Best Models for Grand Slam Matches - Overview

	Train Set		**Test Set**	
Model	ROI	% Matches	ROI	% Matches
NN - full	14%	46.7%	7.4%	39.7%
RF - full	5.3%	16.0%	6.1%	15.0%
BT - full	2.7%	22.1%	9.3%	17.6%

Source: Author's own elaboration

5.3 Variable Importance

As we apply principal components to all non-binary features (280 out of 342), we cannot compare variable importance across original features. Nonetheless, we can still infer which binary and put-aside features are the most relevant and how their importance compares to PCAs. Here we discuss the most relevant binary and put-aside features across all three set of features - the baseline, the limited, and the full models. One can also examine how they compare to PCAs.

As expected, the full model is driven by bookmaker-odds related features with the external info and the dummy indicating that rank favourite is also a huge favourite based on bookmaker odds (odds below 1.25) exhibiting the highest importance. Besides, favourite's rank, rank difference, dummy indicating that the match does not have a favourite based on a rank group, and dummies pinpointing that the favourite is a former TOP10/TOP11-30 as representatives of rank-related features are also significant. Two of our miscellaneous and experimental variables show a high importance as well - dummy indicating that the favourite's last tournament was successful and the dummy expressing that an underdog had to qualify to the main draw.

As for the model without bookmaker odds-related features (limited model), additional important binary variables are revealed - three sets are needed to win the match, the favourite is seeded, favourite/underdog plays home, underdog got a wild card, underdog's last tournament was successful, the match is played on underdog's least favourite surface, and the match is an R16 match.

Finally, the variable importance for baseline models confirms that bookmaker odds-related variables are more important than rank-related ones. Either external info or the dummy indicating that rank favourite is also a huge favourite based on bookmaker odds is shown to be the most important variable across all four machine learning algorithms. Favourite rank, as the most relevant rank-related features, is ranked in the third place at best.

Tables 5.16-5.18 outline the scaled variable importance[7] for the full, the limited, and the baseline set of features across all machine learning algorithms. TOP10 most important variables are selected for each model unless there are less than ten variables with the importance above 1.

Table 5.16: Variable Importance - Full Models

	LR	**RF**	**BT**	**NN**
External info	33	92	84	66
External info squared	3	100	84	48
Favourite is a former TOP11-30 player				29
Favourite is a former TOP10 player			51	
Favourite rank			72	
Favourite's last tournament was successful				29
Match without favourite - rank group			50	
PC1	1	35	100	
PC11		30		
PC13		30		
PC20		29		
PC3	2	29	77	
PC30		30		
PC55				23
Rank difference			66	
Rank favourite is a big favourite - odds	26			32
Rank favourite is a favourite - odds	13	37		
Rank favourite is a huge favourite - odds	100	65	71	100
Tournament level is A250				22
Underdog is a qualificant				34

Note: PCX - x_{th} principal component

Source: Author's own elaboration

[7]The variable with the highest importance takes value 100, while others are scaled back based on their relative importance to the most important variable.

Table 5.17: Variable Importance - Limited Models

	LR	**RF**	**BT**	**NN**
3 sets to win	62			48
Favourite is a former TOP10 player			51	
Favourite is seeded			49	
Favourite is a TOP10 player			44	
Favourite plays home	68			66
Favourite rank		26	72	
Match without favourite - rank group			49	
PC1		100	100	
PC12	26	35	31	
PC13		28		
PC16	28	29	37	
PC2		27		
PC21				58
PC3	38		77	
PC44				49
PC55	32	26		
PC6		27		
PC75				46
PC9		79		
R16 match				81
Rank difference	26	28	66	
Underdog got a wild card	38			56
Underdog is a qualificant	100			100
Underdog plays home	60			
Underdog's last tournament was successful				66
Underdog's least favourite surface				48

Note: PCX - x_{th} principal component

Source: Author's own elaboration

Table 5.18: Variable Importance - Baseline Models

	LR	**RF**	**BT**	**NN**
External info	12	100	100	100
External info squared	13	92	100	68
3 sets to win	7			29
Favourite rank	1	36	81	25
Match without favourite - rank group		6	46	9
Rank difference	3	30	72	50
Rank favourite is a big favourite - odds	37	24	44	32
Rank favourite is a favourite - odds		24	23	10
Rank favourite is a huge favourite - odds	100	70	80	82

Note: PCX - x_{th} principal component

Source: Author's own elaboration

Chapter 6

Conclusion

In this thesis, by employing 40 310 ATP matches played during 1/2001-10/2016 and 342 input features, we examine the prediction accuracy and betting return of four machine learning algorithms applied to men tennis matches - penalized logistic regression, random forest, boosted trees, and artificial neural networks. Three sets of features are investigated across all four machine learning algorithms - the baseline sets that includes only five bookmaker odds-related features along with three rank-related and one additional feature, the limited set that contains all but bookmaker odds-related features, and the full set that utilizes all input features. Therefore, we evaluate 12 different model specifications.

The contributions of this work are fourfold. First, we investigate the applicability of previously untapped machine learning methods concerning tennis forecasting - random forests and boosted trees. Secondly, while the majority of tennis models employ only a limited set of explanatory variables (i.e., usually well below ten), we re-employ all significant features found in all papers written on the topic and add a set of new features resulting in 342 features in total. Thirdly, following the notion 'here is a need to test the relative performance of heuristics, experts, and complex forecasting methods more systematically over the years rather than in a few arbitrary championships' (Goldstein & Gigerenzer, 2009), this thesis covers whole ATP seasons during the period 2001 - 2016, as opposed to only a few seasons or few specific tournaments majority of other papers use. As a result, presented results are more robust. Finally, our thesis examines whether a constant bet size or a variable bet size based on the

half-sized Kelly criterion is preferred for tennis betting.

The analysis provides the following answers to our four hypotheses. Firstly, we show that a profitable strategy for online tennis betting market can be formulated. We develop six profitable betting strategies for betting on favourites applied to both 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.

Secondly, as far as the prediction accuracy is concerned, all model specifications beat backing a better-ranked player, while the majority also surpasses backing a bookmaker's favourite. While some of the more complex model specifications outstrip the penalized logistic regression in both overall accuracy rate and AUC, the improvement is negligible. Overall, our models outperform current state-of-art models for both non-grand-slam (69%) and grand slam matches (79%).

Thirdly, we conclude that the bet size based on the half-sized Kelly criterion leads to a higher betting return if applied to betting on favourites for both non-grand-slam and grand slam matches. 17 out of 18 identified betting strategies for betting on favourites prefer the bet size based on the half-sized Kelly criterion to the constant bet size. On the contrary, as for the betting on underdogs, the bet size based on the half-sized Kelly criterion is preferred only in seven out of 17 cases.

Finally, we examine that grand slam matches outperform non-grand-slam matches both in terms of prediction accuracy and betting performance. As for the prediction accuracy, the overall accuracy rate for grand slam matches exceeds the overall accuracy rate for non-grand-slam matches by 6.9% to 10.1%. The significant difference is also encountered by AUC - from 0.062 to 0.090. As far as the betting performance is concerned, the betting on favourites for grand slam matches surpasses the betting on favourites for non-grand-slam matches regarding the total profitability (1631 compared to 1582 on average), ROI (4% compared to 2.3% on average), and the proportion of matches to bet on (30%

compared to 21% on average).

With further optimization of tuning parameters of more complex models (e.g., the number of trees or the number of features considered at each split for random forest; learning rate or the number of iterations for boosted trees; the number of hidden neurons or weight decay for neural networks), more comprehensive feature selection procedures such as simulated annealing, the extension of the set of input features, or by employing alternative approaches to the betting strategy selection, the improvement of both prediction accuracy and betting return might be achievable. Chapter Further Extensions presents some of the areas for potential improvement.

Chapter 7

Further Extensions

Despite our best effort, we believe that one can further enhance this work by the following:

Firstly, presented algorithms and betting strategies could be tested on women matches taking pronounced differences between men and women's tennis into account. For instance, Magnus & Klaasen (1999) show that men play fewer points per game and more games per set as the dominance on serve is greater for men. Also, Del Coral & Prieto-Rodriguez (2010) reveal that tennis skills are much more surfaced-biased for men than for women. Besides, De Paola & Scoppa (2017) reveal that women losing the first set are much more likely to play poorly in the second set, compared to men. Early modeling attempts to capture women's tennis dynamics suggest that an efficient model for women's tennis can lead to better predictive power as well as higher betting return, compared to men's tennis (Madurska, 2012).

Secondly, although we significantly extend the set of explanatory variables, several important aspects of tennis match are still not captured sufficiently (e.g., player's susceptibility to particular playing style, fear against specific opponents, form at the day of the match, or current mental state of the players). Furthermore, more extensive experimentation with interaction terms could also lead to an improved model performance.

Thirdly, one may employ presented machine learning techniques with different dependent variables, such as the number of games/sets played or the duration of the match. Although betting performance of these models is vir-

tually impossible to assess due to historical unavailability of bookmaker odds other than pre-match odds for winning the match, the data on the number of games/sets played along with the duration of the match are easily available and thus at least the predictive power of these alternative models can be inferred.

Also, even though we introduce previously untapped machine learning algorithms - random forest and boosted trees, alternative techniques such as k-nearest neighbours, stochastic gradient boosting, flexible discriminant analysis, support vector machines, or C 5.0 can be used. Furthermore, even with presented algorithms, further optimization can be performed. Due to the computational limitations, we deal with a limited set of hyperparameters of the model to tune. Enlarging set of tuning parameters might lead to an improvement, especially for more complex approaches such as neural networks and boosted trees.

Moreover, with the improved availability of point-by-point data, the potential of presented machine learning techniques for in-game betting can also be investigated.

Last but not least, although we identify 17 profitable betting strategies for betting on underdogs on the train set, none of them yield a positive return on the test data. Therefore, as far as the betting on underdogs is concerned, the alternative approaches for betting strategy selection could be investigated.

Chapter 8

Bibliography

ABINZANO, Isabel; MUGA, Luis; SANTAMARIA, Rafael. Game, set and match: the favourite-long shot bias in tennis betting exchanges. Applied Economics Letters, 2016, 23.8: 605-608.

BARNETT, Tristan J., et al. Using Microsoft Excel to model a tennis match. In: 6th Conference on Mathematics and Computers in Sport. Queensland, Australia: Bond University, 2002. p. 63-68.

BARNETT, Tristan J., et al. Mathematical modelling in hierarchical games with specific reference to tennis. 2006. PhD Thesis. Swinburne University of Technology.

BARNETT, Tristan; POLLARD, Graham. How the tennis court surface affects player performance and injuries. Medicine and Science in Tennis, 2007, 12.1: 34-37.

BARNETT, Tristan; CLARKE, Stephen R. Combining player statistics to predict outcomes of tennis matches. IMA Journal of Management Mathematics, 2005, 16.2: 113-120.

BARNETT, T.; BROWN, A.; CLARKE, S. Developing a model that reflects outcomes of tennis matches. In: Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport, Coolangatta, Queensland. 2006. p. 178-188.

BEDFORD, Anthony B., et al. A comparison of the ATP ratings with a smoothing method for match prediction. In: Fifth Australian Confer-

ence on Mathematics and Computers in Sport. University of Technology Sydney: Sydney, Australia. 2000. p. 43-51.

BOULIER, Bryan L.; STEKLER, Herman O. Are sports seedings good predictors?: an evaluation. International Journal of Forecasting, 1999, 15.1: 83-91.

BOULIER, Bryan L.; STEKLER, Herman O. Predicting the outcomes of National Football League games. International Journal of Forecasting, 2003, 19.2: 257-270.

BRADLEY, Ralph Allan; TERRY, Milton E. Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 1952, 39.3/4: 324-345.

CAIN, Michael; LAW, David; PEEL, David. The Favourite - Longshot Bias, Bookmaker Margins and Insider Trading in a Variety of Betting Markets. Bulletin of Economic Research, 2003, 55.3: 263-273.

CHITNIS, Asmita; VAIDYA, Omkarprasad. Performance assessment of tennis players: Application of DEA. Procedia-Social and Behavioral Sciences, 2014, 133: 74-83.

CLARKE, Stephen R.; DYTE, David. Using official ratings to simulate major tennis tournaments. International transactions in operational research, 2000, 7.6: 585-594.

CROSS, Rod; POLLARD, Graham. Grand Slam men's singles tennis 1991-2009 Serve speeds and other related data. COACHING & SPORT SCIENCE REVIEW, 2009.

DEL CORRAL, Julio; PRIETO-RODRIGUEZ, Juan. Are differences in ranks good predictors for Grand Slam tennis matches?. International Journal of Forecasting, 2010, 26.3: 551-563.

DE PAOLA, Maria; SCOPPA, Vincenzo. Gender differences in reaction to psychological pressure: evidence from tennis players. European Journal of Work and Organizational Psychology, 2017, 26.3: 444-456.

DIXON, Mark J.; COLES, Stuart G. Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1997, 46.2: 265-280.

DIXON, Mark J.; POPE, Peter F. The value of statistical forecasts in the UK association football betting market. International journal of forecasting, 2004, 20.4: 697-711.

EASTON, Stephen; UYLANGCO, Katherine. Forecasting outcomes in tennis matches using within-match betting markets. International Journal of Forecasting, 2010, 26.3: 564-575.

ELLIOTT, Bruce; SAVIANO, Nick. Serves and returns. World-class tennis technique, 2001, 207-222.

FARRELL, Michael James. The measurement of productive efficiency. Journal of the Royal Statistical Society. Series A (General), 1957, 120.3: 253-290.

FARRELLY, Daniel; NETTLE, Daniel. Marriage affects competitive performance in male tennis players. Journal of Evolutionary Psychology, 2007, 5.1: 141-148.

FERNANDEZ, Jaime; MENDEZ-VILLANUEVA, A.; PLUIM, B. M. Intensity of tennis match play. British journal of sports medicine, 2006, 40.5: 387-391.

FORREST, David; MCHALE, Ian. 8 Longshot bias: insights from the betting market on men's professional tennis. Information Efficiency in Financial and Betting Markets, 2005, 215.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. The elements of statistical learning. New York: Springer series in statistics, 2001.

GANDAR, John M., et al. Informed traders and price variations in the betting market for professional basketball games. The Journal of Finance, 1998, 53.1: 385-401.

GESCHEIT, Danielle T., et al. Effects of consecutive days of match play on technical performance in tennis. Journal of sports sciences, 2017, 35.20: 1988-1994.

GILSDORF, Keith F.; SUKHATME, Vasant A. Testing Rosen's sequential elimination tournament model: Incentives and player performance in professional tennis. Journal of Sports Economics, 2008, 9.3: 287-303.

GLICKMAN, Mark E. Parameter estimation in large dynamic paired comparison experiments. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1999, 48.3: 377-394.

GOLDSTEIN, Daniel G.; GIGERENZER, Gerd. Fast and frugal forecasting. International Journal of Forecasting, 2009, 25.4: 760-772.

GOOSSENS, R. Dries, et al. Winning in straight sets helps in Grand Slam tennis. International Journal of Performance Analysis in Sport, 2015, 15.3: 1007-1021.

GRAHAM, I.; STOTT, H. Predicting bookmaker odds and efficiency for UK football. Applied Economics, 2008, 40.1: 99-109.

HAIGH, John. Taking chances: winning with probability. OUP Oxford, 1999.

HARRELL JR, Frank E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, 2015.

HAWKINS, Douglas M.; BASAK, Subhash C.; MILLS, Denise. Assessing model fit by cross-validation. Journal of chemical information and computer sciences, 2003, 43.2: 579-586.

HERZOG, Stefan M.; HERTWIG, Ralph. The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. Judgment and Decision Making, 2011, 6.1: 58.

HOLDER, Roger L.; NEVILL, Alan M. Modelling performance at international tennis and golf tournaments: is there a home advantage?. Journal of the Royal Statistical Society: Series D (The Statistician), 1997, 46.4: 551-559.

HOLTZEN, David W. Handedness and professional tennis. International Journal of Neuroscience, 2000, 105.1-4: 101-119.

HORNIK, Kurt; STINCHCOMBE, Maxwell; WHITE, Halbert. Multilayer feedforward networks are universal approximators. Neural networks, 1989, 2.5: 359-366.

HUANG, Xinzhuo; KNOTTENBELT, William; BRADLEY, Jeremy. Inferring tennis match progress from in-play betting odds. Final year project), Imperial College London, South Kensington Campus, London, SW7 2AZ, 2011.

IRONS, David J.; BUCKLEY, Stephen; PAULDEN, Tim. Developing an improved tennis ranking system. Journal of Quantitative Analysis in Sports, 2014, 10.2: 109-118.

JACKSON, David; MOSURSKI, Krzysztof. Heavy defeats in tennis: Psychological momentum or random effect?. Chance, 1997, 10.2: 27-34.

JAMES, Gareth, et al. An introduction to statistical learning. New York: springer, 2013.

KANAZAWA, Satoshi. Why productivity fades with age: The crime-genius connection. Journal of Research in Personality, 2003, 37.4: 257-272.

KELLY JR, John L. A new interpretation of information rate. In: The Kelly Capital Growth Investment Criterion: Theory and Practice. 1956. p. 25-34.

KLAASSEN, Franc JGM; MAGNUS, Jan R. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. Journal of the American Statistical Association, 2001, 96.454: 500-509.

KLAASSEN, Franc JGM; MAGNUS, Jan R. Forecasting the winner of a tennis match. European Journal of Operational Research, 2003, 148.2: 257-267.

KNIGHT, Gareth; O'DONOGHUE, Peter. The probability of winning break points in Grand Slam men's singles tennis. European Journal of Sport Science, 2012, 12.6: 462-468.

KNOTTENBELT, William J.; SPANIAS, Demetris; MADURSKA, Agnieszka M. A common-opponent stochastic model for predicting the outcome of professional tennis matches. Computers & Mathematics with Applications, 2012, 64.12: 3820-3827.

KONAKA, Eiji. Match results prediction ability of official ATP singles ranking. arXiv preprint arXiv:1705.05831, 2017.

KONING, Ruud H. Home advantage in professional tennis. Journal of Sports Sciences, 2011, 29.1: 19-27.

KONING, Ruud H. Regression tests and the efficiency of fixed odds betting markets. International Journal of Sport Finance, 2012, 7.3: 262.

KOPRIVA, Frantisek. Constant Bet Size? Don't Bet on It! Testing Expected Utility Theory on Betfair Data. 2015.

KOVALCHIK, Stephanie Ann. Searching for the GOAT of tennis win prediction. Journal of Quantitative Analysis in Sports, 2016, 12.3: 127-138.

KUHN, Max; JOHNSON, Kjell. Applied predictive modeling. New York: Springer, 2013.

KUPER, Gerard H., et al. Using tennis rankings to predict performance in upcoming tournaments. University of Groningen, Research Institute SOM (Systems, Organisations and Management), 2014.

LAHVIČKA, Jiří. What causes the favourite-longshot bias? Further evidence from tennis. Applied Economics Letters, 2014, 21.2: 90-92.

LISI, Francesco. Tennis betting: Can statistics beat bookmakers?. Electronic Journal of Applied Statistical Analysis, 2017, 10.3: 790-808.

LEITNER, Christoph; ZEILEIS, Achim; HORNIK, Kurt. Is Federer Stronger in a Tournament Without Nadal? An Evaluation of Odds and Seedings for Wimbledon 2009. Austrian Journal of Statistics, 2016, 38.4: 277-286.

LOFFING, Florian; HAGEMANN, Norbert; STRAUSS, Bernd. The serve in professional men's tennis: Effects of players' handedness. International Journal of Performance Analysis in Sport, 2009, 9.2: 255-274.

LOFFING, Florian; HAGEMANN, Norbert; STRAUSS, Bernd. Left-handedness in professional and amateur tennis. PLoS One, 2012, 7.11: e49325.

LYOCSA, Stefan; VYROST, Tomas. To bet or not to bet: a reality check for tennis betting market efficiency. Applied Economics, 2017, 1-22.

MA, Shang-Min, et al. Winning matches in Grand Slam men's singles: An analysis of player performance-related variables from 1991 to 2008. Journal of sports sciences, 2013, 31.11: 1147-1155.

MADURSKA, Agnieszka M. A set-by-set analysis method for predicting the outcome of professional singles tennis matches. Imperial College London, Department of Computing, Tech. Rep., 2012.

MAGNUS, Jan R.; KLAASSEN, Franc JGM. The effect of new balls in tennis: four years at Wimbledon. Journal of the Royal Statistical Society: Series D (The Statistician), 1999, 48.2: 239-246.

MALUEG, David A.; YATES, Andrew J. Testing contest theory: evidence from best-of-three tennis matches. The Review of Economics and Statistics, 2010, 92.3: 689-692.

MARTIN, J. Kent; HIRSCHBERG, Daniel S. Small sample statistics for classification error rates II: Confidence intervals and significance tests. 1996.

MCHALE, Ian; MORTON, Alex. A Bradley-Terry type model for forecasting tennis match results. International Journal of Forecasting, 2011, 27.2: 619-630.

MOLINARO, Annette M.; SIMON, Richard; PFEIFFER, Ruth M. Prediction error estimation: a comparison of resampling methods. Bioinformatics, 2005, 21.15: 3301-3307.

NEVILL, Alan M., et al. Identifying home advantage in international tennis and golf tournaments. Journal of Sports Sciences, 1997, 15.4: 437-443.

NEWTON, Paul K.; KELLER, Joseph B. Probability of winning at tennis I. Theory and data. Studies in applied Mathematics, 2005, 114.3: 241-269.

O'DONOGHUE, Peter G. The most important points in grand slam singles tennis. Research quarterly for exercise and sport, 2001, 72.2: 125-131.

O'DONOGHUE, Peter; INGRAM, Billy. A notational analysis of elite tennis strategy. Journal of sports sciences, 2001, 19.2: 107-115.

O'MALLEY, A. James. Probability formulas and statistical analysis in tennis. Journal of Quantitative Analysis in Sports, 2008, 4.2.

PASERMAN, M. Daniele. Gender differences in performance in competitive environments: evidence from professional tennis players. 2007.

POPE, Peter F.; PEEL, David A. Information, prices and efficiency in a fixed-odds betting market. Economica, 1989, 323-341.

RADICCHI, Filippo. Who is the best player ever? A complex network analysis of the history of professional tennis. PloS one, 2011, 6.2: e17249.

REES, Tim; HARDY, Lew. Matching social support with stressors: Effects on factors underlying performance in tennis. Psychology of sport and exercise, 2004, 5.3: 319-337.

REID, Machar; MCMURTRIE, Darren; CRESPO, Miguel. Title: The relationship between match statistics and top 100 ranking in professional men's tennis. International Journal of Performance Analysis in Sport, 2010, 10.2: 131-138.

RIDDLE, Lawrence H. Probability models for tennis scoring systems. Applied Statistics, 1988, 63-75.

RIPLEY, Brian D. Pattern recognition and neural networks. Cambridge university press, 2007.

ROSENBLOOM, Earl S.; NOTZ, William. Statistical tests of real-money versus play-money prediction markets. Electronic Markets, 2006, 16.1: 63-69.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

SAUER, Raymond D. The economics of wagering markets. Journal of economic Literature, 1998, 36.4: 2021-2064.

SCHEIBEHENNE, Benjamin; BRODER, Arndt. Predicting Wimbledon 2005 tennis results by mere player name recognition. International Journal of Forecasting, 2007, 23.3: 415-426.

SCHULZ, Richard; CURNOW, Christine. Peak performance and age among superathletes: track and field, swimming, baseball, tennis, and golf. Journal of Gerontology, 1988, 43.5: P113-P120.

SERWE, Sascha; FRINGS, Christian. Who will win Wimbledon? The recognition heuristic in predicting sports events. Journal of Behavioral Decision Making, 2006, 19.4: 321-332.

SIPKO, Michal; KNOTTENBELT, William. Machine Learning for the Prediction of Professional Tennis Matches. 2015.

SLATTON, Thomas Grant. A comparison of dropout and weight decay for regularizing deep neural networks. 2014. PhD Thesis.

SMITH, Matthew T., et al. Heat stress incident prevalence and tennis matchplay performance at the Australian Open. Journal of Science and Medicine in Sport, 2017.

SOMBOONPHOKKAPHAN, Amornchai; PHIMOLTARES, Suphakant; LURSINSAP, Chidchanok. Tennis winner prediction based on time-series history with neural modeling. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. 2009.

SPANN, Martin; SKIERA, Bernd. Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. Journal of Forecasting, 2009, 28.1: 55-72.

SPANIAS, Demetris; KNOTTENBELT, William J. Predicting the outcomes of tennis matches using a low-level point model. IMA Journal of Management Mathematics, 2012, 24.3: 311-320.

SUNDE, Uwe. Potential, prizes and performance: Testing tournament theory with professional tennis data. 2003.

SUNDE, Uwe. Heterogeneity and performance in tournaments: a test for incentive effects using professional tennis data. Applied Economics, 2009, 41.25: 3199-3208.

SZYMANSKI, Stefan. The economic design of sporting contests. Journal of economic literature, 2003, 41.4: 1137-1187.

WALKER, Mark; WOODERS, John; AMIR, Rabah. Equilibrium play in matches: Binary Markov games. Games and Economic Behavior, 2011, 71.2: 487-502.

Appendix A

Appendix

Correlation of Predictions

Table A.1: Correlation of Predictions - Full Model

	LR	**RF**	**BT**	**NN**	**Implied Prob**
LR	1.00	0.90	0.98	0.94	0.97
RF	0.90	1.00	0.89	0.88	0.90
BT	0.98	0.89	1.00	0.92	0.95
NN	0.94	0.88	0.92	1.00	0.93
Implied Prob	0.97	0.90	0.95	0.93	1.00

Table A.2: Correlation of Predictions - Limited Model

	LR	**RF**	**BT**	**NN**	**Implied Prob**
LR	1.00	0.84	0.92	0.93	0.90
RF	0.84	1.00	0.86	0.82	0.80
BT	0.92	0.86	1.00	0.88	0.86
NN	0.93	0.82	0.89	1.00	0.85
Implied Prob	0.90	0.80	0.86	0.85	1.00

Table A.3: Correlation of Predictions - Baseline Model

	LR	**RF**	**BT**	**NN**	**Implied Prob**
LR	1.00	0.81	0.99	0.99	0.97
RF	0.81	1.00	0.82	0.81	0.79
BT	0.99	0.82	1.00	0.99	0.97
NN	0.99	0.81	0.99	1.00	0.97
Implied Prob	0.97	0.79	0.97	0.97	1.00