Machine Learning-Based Player Performance Analysis for Association of Tennis Professionals Tour

Youchen Qing *

UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai, 200240, China

* Corresponding Author Email: rex29qyc@sjtu.edu.cn

Abstract. This study explores the use of data analysis in professional tennis, using the dataset of the world-renowned Association of Tennis Professionals (ATP) Tour, which is the top-class men's professional tennis tour. Utilizing the match data of all the players, mainly Novak Djokovic, the research applies various models and methods of data preprocessing and selection to predict match outcomes. The main objective is to provide data-supported insight to professional tennis players and coaches to help analyze and further improve their tennis. The analysis and model take factors like player performance, court types, and tournament importance into consideration. This study shows the potential of data analysis in sports where such resources are often limited to well-funded teams. Not only the effect of different models in forecasting tennis match outcomes is evaluated, but also the influence of non-critical tournaments and external factors like injuries and policies on the accuracy of predictions are explored. The research findings offer insights into the evolving role of data analysis in enhancing sports strategies and performance, indicating a promising future for data-driven decision-making in professional tennis for different levels of players.

Keywords: ATP Tour; machine learning; data analysis.

1. Introduction

In today's society, data is considered to be the new oil driving the booming of every industry in the world, and sports is no exception [1]. The work of data analysts has become increasingly important in professional sports teams, and they use detailed data to examine the factors that effect on-field performance and game results to help professional athletes and coaches improve their performance [2, 3]. However, such data analysis is often only available for sports teams with deep pockets, such as football clubs or basketball clubs, and for sports, such as professional tennis, where the players themselves have to afford all the funding, the support of data will be out of reach for many athletes.

Previous work proposes a measure based on eigenvector centrality for predicting the winner in tennis matches. The resulting ratings are then used as a covariate in a simple logit model. The proposed approach largely and consistently outperforms all the alternative models considered in terms of prediction accuracy [4]. Another study proposes a statistical approach for predicting the match outcomes of Grand Slam tournaments, in addition to applying exploratory data analysis (EDA) to explore variables related to match results. The proposed approach introduces new variables via the Glicko rating model, a Bayesian method commonly used in professional chess [5].

In this article, the author will use data from the world's top tennis tour Association of Tennis Professionals (ATP) Tour to predict the results of tennis matches. In order to make the results more meaningful, the match data of Novak Djokovic will be used as the main analysis object. In the course of the experiment, different models will be used to incorporate different features for analysis and comparison.

2. Method

2.1. Data Selection

In terms of data selection, the experimental data came from the data set of ATP Tour matches on Kaggle [6]. The dataset contains detailed information on all ATP matches played since 1968. Competition statistics since 1991 are available. In the experiment, the author extracted all matches

associated with Novak Djokovic's player_id for training and testing. Since Djokovic officially became an ATP professional player in 2004, the data has been complete (compared to those before 1991), and the technologies of the game and tournament settings have not changed much till today. In addition, Novak Djokovic, as a top player, has a wide enough sample of matches, while average players are eligible for fewer tournaments and can last a shorter number of rounds in the same tournament. In addition, compared to two great world-known tennis players, Rafael Nadal, who has suffered many injuries and is better at and signed up more for clay courts, and Roger Federer, who retired in 2021, Novak Djokovic plays almost all year round and evenly attends each tournament. This will avoid some bias in the collected data, such as the uneven distribution of the type of the courts and the bias brought by aging.

2.2. Data Preprocess

Because in the raw data set, column names have feature prefixes that distinguish winners from losers, these strongly indicative column names can overfit the model used later in the experiment, making the model meaningless. Therefore, prefixes such as "novak_" and "oppenent_" will be used instead of "w_", "winner_", "l_", “and loser_" and so on to pave the way for prediction. To make a comparison to the model trained by all players’ match data, another dataset is created by replace the "w_", "winner_", "l_", "loser_" with "player_1_" and "player_2_", and then randomly exchange the information of player_1 and player_2, otherwise all the player_1 will be the winner. The ‘result’ in this dataset, including all players, is 1 if the features of the players are not exchanged and 0 if the features of the players are exchanged. At the same time, the classification features such as tournament name, players’ dominant hand, the round of the match, and the type of the court are one-hot encoded to facilitate the training of the model. In this special case of making a prediction for a professional tennis player, to fit in with their individual demand and situation is one of the priorities. Therefore, finding out the games that the players pay a lot of effort to is important since players from different levels have different tour plans, which leads them to sign up for different tournaments some of the tournaments are just for warming up and they will spare some efforts. Novak Djokovic, for example, has been the world's top athlete for most of his career, so the competitions he focuses on and puts all his efforts into are usually high-value competitions, such as the four Grand Slams and nine Masters Tournaments. However, he often doesn’t perform as well as he plays in national team competitions such as Davis Cup or tournaments with less than 1,000 points. In this way, the warm-up and not serious participation won’t be something the athletes want to study. Thus, for the selection of unimportant matches, taking the maximum intersection of all the tournament names that the top four players in the world (Andy Murray, Novak Djokovic, Rafael Nadal, and Roger Federer) played during this period can prove that these matches were relatively important to some extent. This operation will also benefit the further comparison of the model’s performance on different players.

2.3. Models

In terms of model selection, this work chooses the basic logistic regression, support vector machine (SVM), LightGBM, XGBoost, random forest and deep neural network (DNN) in common one-dimensional data classification problems.

Here, the data set is divided by time, taking the effective features and results of Djokovic's selected matches from 2004 to 2017 as the training set, and from 2018 to 2022 as the test set, to predict the outcome of the test set. There may be some confusion as to why the test set and the training set are divided this way. Why not practice all the games before each game to predict that game? Why not use a sliding window to select some features? The main reason for this choice is the unique nature of the tennis ATP Tour. First of all, there is a huge difference between the court types of tennis matches, and even the same hard-court matches will vary greatly because of the speed of the court bounce, and at the same time, a player may have to participate in a variety of different court types of matches in a short period of time. Therefore, using the previous year's data as a prediction does not affect the prediction of a match in a field that has not been played that year. Second, the rule of most of the

tournaments is that the player will be eliminated after lose one match, which means that the level of the opponent tends to grow from weak to strong as a player goes farther in a tournament and after ending his trip in one tournament and starts another tournament the player's opponents will also begin to enter another cycle, so collecting the data from matches happened close to the one which is going to be predicted will cause a bias.

In the approach here to predict tennis match outcomes, this work employs a set of six different predictive models. The models are chosen for their unique strength in dealing with this type of data.

Logistic Regression turns the classification problem into a regression problem. It is a foundational model that can be clearly interpreted and is able to estimate win probabilities based on a variety of predictors, such as the players' past performances and different conditions given in this case.

XGBoost is a machine learning model based on Gradient-Boosting Decision Tree (GBDT) with optimized loss function and regularization terms. It provides more accurate results when handling complex datasets. In this moderate-sized tennis matches dataset, it will be efficient [7].

Random Forest is an ensemble of decision trees. It can effectively handle both numerical and categorical data and avoid overfitting. In addition, it can show the importance of the features that will be inspiring in this study for tennis matches.

SVM is a model that is effective in doing a classification job, especially in high-dimensional spaces, which is just the case here. The ATP tennis match datasets include many features where non-linearity is sure to exist and the SVM model can handle the situation by applying different kernel functions to divide them without losing robustness [8].

LightGBM can process large datasets with high efficiency and less memory. It also performs well when dealing with categorical data, which is a frequent and essential part of the dataset [9]. The dataset here is not extremely large and the computation is not extremely costly. Thus, LightGBM may not give the top performance here but will be more valuable when bigger datasets are given.

Deep Neural Networks is good at dealing with complex and non-linear relationships in the dataset. DNNs excel at automatically learning feature interactions, which is an essential aspect in tennis, where factors such as player style, match arrangement, and tournament information significantly influence match outcomes [10].

To better leverage the strengths of each of these models, the hyperparameters have all been fine-tuned with grid searching to optimize the model and avoid overfitting. In addition, cross-validation is also applied to ensure fair comparison and robustness.

3. Experiment and Result

3.1. Research Targets

In this paper, the following experiments mainly aim to answer four questions. Firstly, for the player Novak Djokovic, whether the model trained by all players' data or by his own individual data works better? Secondly, whether the data without unimportant tournaments will get a more accurate prediction? Thirdly, will the accuracy of prediction of the results of the matches decrease from the year of 2018 to the year of 2022? In this process, whether the events out of court, like injuries or policies about pandemics, affect the accuracy of the prediction? Fourthly, how well will a model trained on one player's match data predict the match result of another player's match? What factor will affect the accuracy of this operation?

Hypothesis for the First Question: When predicting the match outcomes of Novak Djokovic, a model trained by his individual match data will be more accurate compared to a model trained on data from all players on the prediction task.

Hypothesis for the Second Question: Excluding data from unimportant tournaments and external factors in the predictive model won't affect the accuracy of predictions of match outcomes.

Hypothesis for the Third Question: The accuracy of predicting match outcomes will decrease from the year 2018 to 2022, potentially influenced by external factors such as injuries and pandemic-related policies.

Hypothesis for the Fourth Question: A model trained on one player's match data will not perform well in predicting the match result of another player's match. The less similar the two players are, the less accurate the prediction will be.

3.2. Model Comparison

From Fig 1, the accuracy of the prediction made by the model trained by all the players' match data from 2004-2017 is only around 0.5, which is just the performance of a random model. This suggests that all the players' data is not indicative of Novak Djokovic's matches.

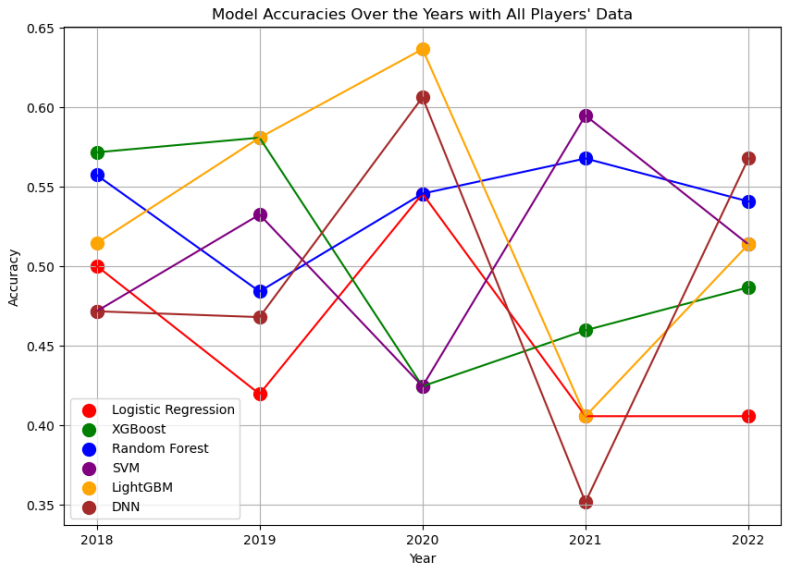

Fig 1. Accuracy of various models on all players' data (Figure Credits: Original).

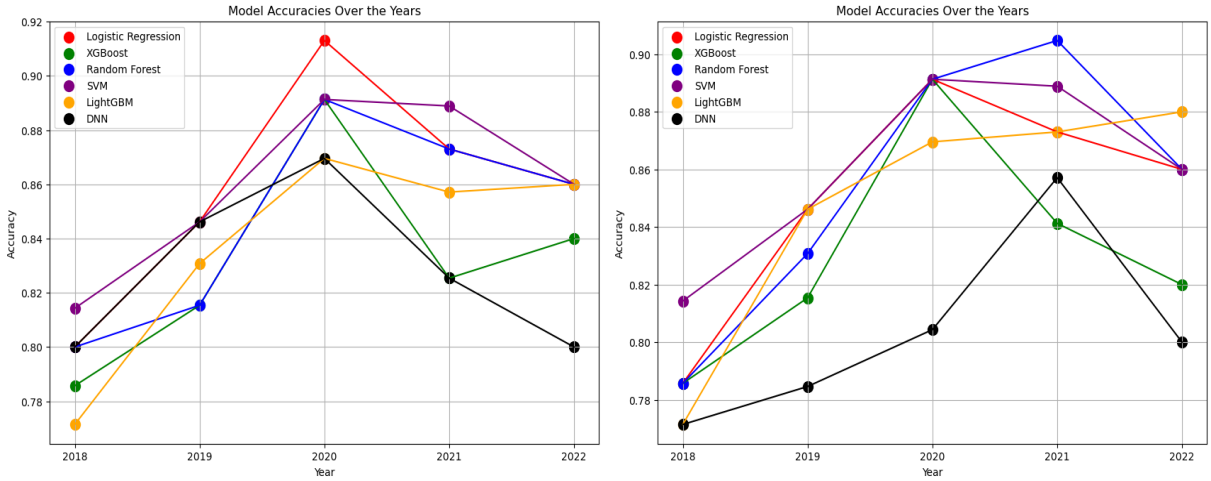

Fig 2. Trained with all the matches (left) and trained without unimportant matches (right) accuracy of various models over different years (Figure Credits: Original).

In the experiment, Fig 2 is used here to show the prediction results of six models that include non-important tournament data and do not include non-important tournament data. It is not difficult to find that the model trained without data from these unimportant tournaments has a close accuracy to the unprocessed one in predicting Djokovic's match, especially in the case of the SVM model, where the prediction accuracies over all five years are all the same. In these two-line plots, the accuracy of different models is also different. SVM is a relatively more efficient and accurate model, so SVM will be used as the model for prediction in the following experiments. This is paralleled with the previous hypothesis that using only the data of important games will be an effective way.

From Fig 3, the accuracy of model predictions rose between 2018 and 2020 and then declined. This result is different from the hypothesis that the accuracy should decrease with time.

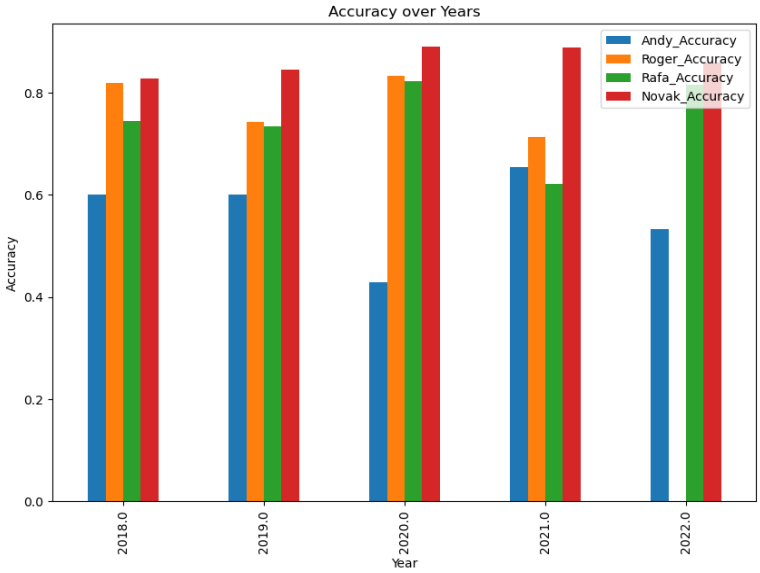

Fig 3. Accuracy from 2018 to 2020 (Figure Credits: Original).

One intriguing question in the experiment is whether the model developed from Djokovic's match data can also be applied to other players at a similar level. Therefore, this author used the model trained on Djokovic's 2004-2017 match data to predict the outcome of the 2018-2022 match data of the other three players in the so-called Big Four of this era: Nadal, Murray, and Federer. As the table above shows, the model's accuracy in predicting the outcome of Murray's match is relatively low, ranging from 0.4-0.6. For the two more powerful players, Nadal and Federer, the accuracy of the prediction is even higher, except for 2021, only about 0.65, and the accuracy of the other years is 0.8. Therefore, the model is meaningful for other athletes at the same level to some extent.

4. Discussion

From the result of the experiment, it can be found that the model trained by the match data from the player works better for himself. Indeed, this is reasonable considering that in the data set, including all the players, their mean rank is only 59.5, while Novak Djokovic is always ranked at the top of all the players. Thus, it’s normal that their data are not indicative of Novak’s matches.

For the experiment focused on the second research question, the results were similar to those assumed before the experiment, and filtering out the important matches slightly improved the model's performance. Statistically speaking, most of the games left after the screening are indeed higher-level games (500 points and above) and these players play consistently rather than just a few times over the years. Therefore, as the results show, the outcomes of these games remain predictable because the players are more involved in such games, and the outcome is not affected by attitude or other off-field factors.

In the third research question, the results did not agree with the previous assumptions, and the accuracy of the prediction did not monotonically decrease over time, but first increased and then decreased. While training the model, each match feature does not include information on injuries or off-court factors, but this is the situation that happened around 2018 when Djokovic was coming back from a two-year slump. After removing the data of Djokovic's 2017 matches due to injury and condition problems (41 matches), it was found that the prediction results using the SVM model did not change. This shows that the most basic SVM used is not affected by poor race data over a period of time.

The model trained using Djokovic's match data to predict the results of other top players is not entirely consistent with the experimental hypothesis. Djokovic is right-handed and uses two-handed backhand shots. The SVM model trained only has an accuracy of nearly 0.5 when predicting the match performance of Andy Murray, also a right-handed player of the same age as Djokovic, who

uses two-handed backhand shots. The performance is almost poor with the random model. For Spanish star Nadal, who uses his left hand and backhand, and Swiss star Federer, who uses his right hand and one-handed backhand, the accuracy of this SVM model can exceed 0.8 in most years. From this point of view, the portability of the model should be measured by the rank or points of the players rather than the apparent similarities or habits, which is a topic that calls for further exploration.

The analysis still has a huge space for improvement. Since the available dataset is limited, if more detailed and individual data are provided, the prediction may be more accurate. For example, the experiment carried out in this article cannot use the features like first-serve points won or second-serve points won, which are highly related to the result of the game and will make the prediction meaningless. If data like the serve speed or forehand and backhand return speed are provided, the analysis will be more instructive for the players to adjust their strategy to prepare for the upcoming matches. Also, if the data on the cost of the players to take each tournament is provided, the model will also help them to fully utilize their limited budget by giving them a competition plan after comparing the estimated income and points from each tournament with their cost.

5. Conclusion

The research demonstrates the effectiveness of data analysis in predicting tennis match outcomes, particularly in the context of the ATP Tour. Key findings include the higher accuracy of predictions when using the player's individual dataset, and the operation of excluding non-important tournaments is harmless. The study also reveals that models trained on Novak Djokovic's match data can also predict outcomes for other top players with satisfactory accuracy, while it doesn't work well for players from a lower level. These results underscore the potential of data analytics in professional tennis, not just for predicting the results of the matches but also for strategy optimization of the players in the future. The study paves the way for more detailed and personalized analytics in sports, tailored to individual athletes' needs and the specificities of tennis.

References

[1] Wang Jin, Yaqiong Yang, Tian Wang, R. Simon Sherratt, and Jingyu Zhang. Big data service architecture: a survey. Journal of Internet Technology, 2020, 21(2): 393-405.

[2] Zhu Pan, and Feng Sun. Sports athletes' performance prediction model based on machine learning algorithm. In International Conference on Applications and Techniques in Cyber Intelligence, 2020, 7: 498-505.

[3] Oytun Musa, Cevdet Tinazci, Boran Sekeroglu, Caner Acikada, and Hasan Ulas Yavuz. Performance prediction and evaluation in female handball players using machine learning models. IEEE Access, 2020, 8: 116321-116335.

[4] Arcagni Alberto, Vincenzo Candila, and Rosanna Grassi. A new model for predicting the winner in tennis based on the eigenvector centrality. Annals of Operations Research, 2023, 325 (1): 615-632.

[5] Yue Jack C., Elizabeth P. Chou, Ming-Hui Hsieh, and Li-Chen Hsiao. A study of forecasting tennis matches via the Glicko model. Plos one, 2022, 17(4): e0266838.

[6] Kaggle ATP matches, URL: https://www.kaggle.com/datasets/sijovm/atpdata, Last Accessed 2023/11/20.

[7] Chen Tianqi, and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016: 785-794.

[8] Noble William. What is a support vector machine? Nature biotechnology, 2006, 24(2): 1565-1567.

[9] Ke Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017, 30: 1-9.

[10] Sharma, Poonam, and Akansha Singh. Era of deep neural networks: A review. In 2017 8th International Conference on Computing, Communication and Networking Technologies, 2017: 1-5.