Tennis match forecasting; how AI can beat book makers on their own platforms

Patrick Schall* and Vahid Toomani†

(Dated: March 5, 2024)

* patrick.schall@gmail.com

† vahidtoomani2002@gmail.com

CONTENTS

I. INTRODUCTION

In 1943, Walter Pitts and Warren McCulloch laid the foundation for artificial intelligence (AI) and its subfield, machine learning, with their first mathematical model of a neural network.

Seven years later, Alan Turing posed the fundamental question: "Can machines think?" He developed the Turing test, originally known as the imitation test, which assesses whether a machine exhibits intelligent behavior. In his seminal work "Computing Machinery and Intelligence" (1950), Turing didn't settle for a simple definition of machines or thinking; instead, he asked a more profound question: "Can machines do what we, as thinking entities, can do?"

Over the past 70 years, fueled by ever-growing computational power, increased storage capacity, and an exponentially expanding pool of data, machine learning (ML) and deep learning algorithms have revolutionized the field of data science. ML models can predict whether a customer will cancel their contract with a telecommunication provider, determine the authenticity of an unknown painting, verify the origin of wine analyzed in a laboratory, and even classify tumors as benign or malignant. Recent breakthroughs include autonomous driving cars, defeating the world's best Go player (a game more complex than chess and reliant on intricate pattern recognition), and predicting the 3D structure of proteins from their amino acid sequences (AlphaFold 2).

One significant advantage of machine learning models lies in their ability to forecast the future with a high degree of certainty. Today, ML models are employed to predict weather patterns, estimate fuel consumption for new cars, and optimize maintenance intervals for machinery. A practical application of ML is in the prediction of sports betting outcomes. By analyzing historical data, player performance, and other relevant factors, ML models can provide valuable insights for informed betting decisions.

In the following work, we aim to design a machine learning model, employing

standard ML algorithms like Decision Tree, Random Forest, Support Vector Machine (SVM), and more advanced deep learning models like Convolutional Neural Network (CNN), capable of predicting the outcomes of tennis matches played on the ATP tour. We aim to better understand which features influence the outcome of a tennis game. Additionally, we intend to develop a betting strategy that minimizes investment losses and maximizes return on investment by placing bets on the odds provided by two bookmakers: Bet365 and Pinnacle Sports. In general, we aim to develop a betting tool which:

1. Predicts tennis matches with high accuracy.
2. Yields a net plus when betting money on tennis matches on Bet 365 and Pinnacle. with their respective odds.

This work, carried out by Vahid Toomani with a scientific background in math and physics, and Patrick Schall with a scientific background in molecular biology, will not only assist gamblers in maximizing their ROI and aid bookmakers in improving their odds, but it will also help tennis players and their coaches better understand the main features influencing the outcome of a tennis game.

A. International tournaments and book makers

ATP, Pinnacle sports, Bet 365

B. Elo system

Elo rate

C. objectives

Goals

- What are the main objectives to be achieved? Describe in a few lines.
- For each member of the group, specify the level of expertise around the problem addressed?
- Have you contacted business experts to refine the problem and the underlying models? If yes, detail the contribution of these interactions.
- (Are you aware of a similar project within your company, or in your entourage? What is its progress? How has it helped you in the realization of your project? How does your project contribute to improving it?).

II. EXPLORATORY DATA ANALYSIS (EDA)

A. Data fetch

Our initial data set consist of ATP matches from 2000 until 2018 fetched from Kaggle. This Dataset is provided by Eduard Thomas to Kaggle.com. Since our initial data frame was fairly outdated and were missing a lot of odds for the years 2000 till 2003 we fetched new data from tennis-data and deleted all the data with missing data for the odds. This resulted in a data frame consisting of data from 2004 until 2024.

B. Data enrichment & cleaning

Which parts of data dropped or added?

C. Data statistics and visualisations

Our dataset consists of $48,824$ tennis matches involving $1,283$ unique male players from January 5, 2004, until February 4, 2024 (Figure 1). From 2004 to 2023,

we observe an average of approximately $2,500$ games per year. Notably, in 2020, only about $1,450$ games were played due to Covid-19 restrictions. Furthermore, the data belonging to year 2009 is missing. That is because the bet odds of Pinnacle sports is not recorded in our data source, hence the related data is droped all together in cleaning procces. One of the most important metrics for assessing the

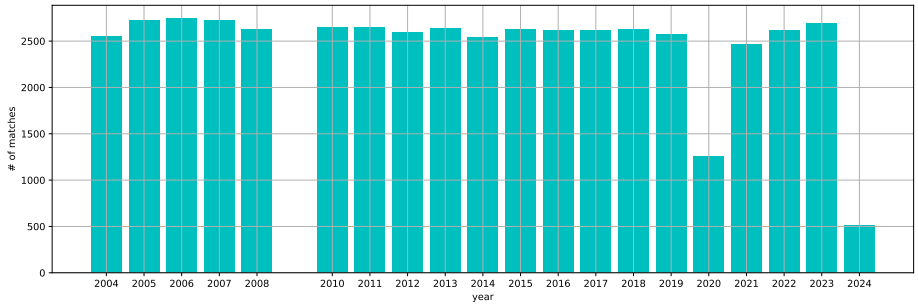

FIG. 1. Number of matches held on each year

performance of individual players in tennis is the Elo rating. The distribution of Elo ratings on the ATP tours reveals that the median rating is 1640 (Figure 2). Interestingly, only one player (N. Djokovic) has an Elo rate exceeding 2000 (Figure 3).

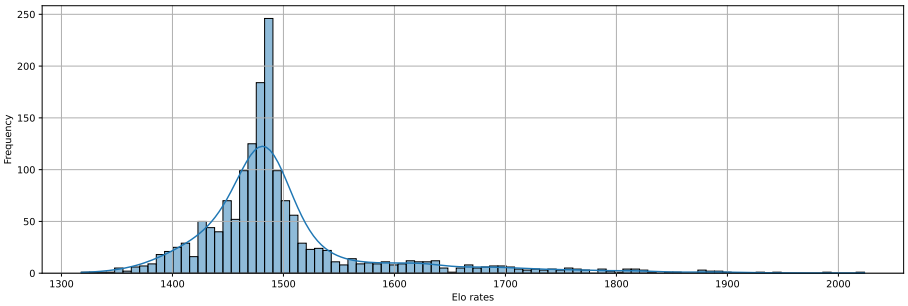

FIG. 2. Distribution of Elo rates

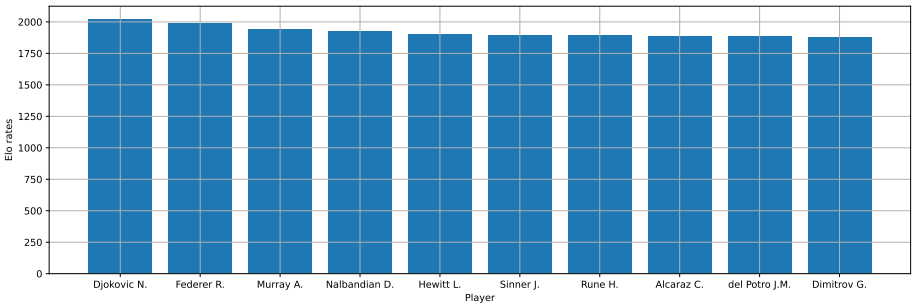

FIG. 3. Top 10 players with highest Elo rates

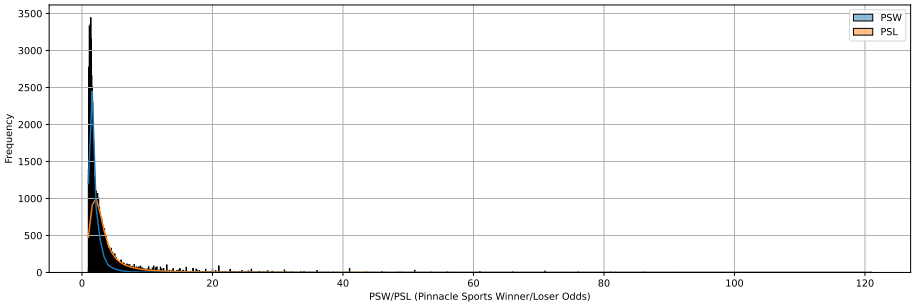

FIG. 4. Distribution of Pinnacle Sports Winner/Loser Odds

D. Strategies

Strategies

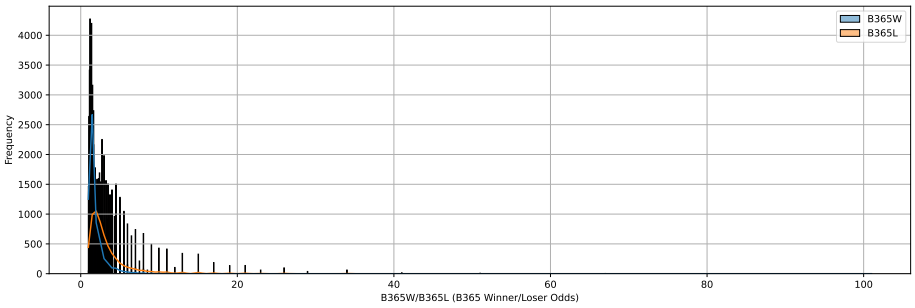

FIG. 5. Distribution of B365 Winner/Loser Odds

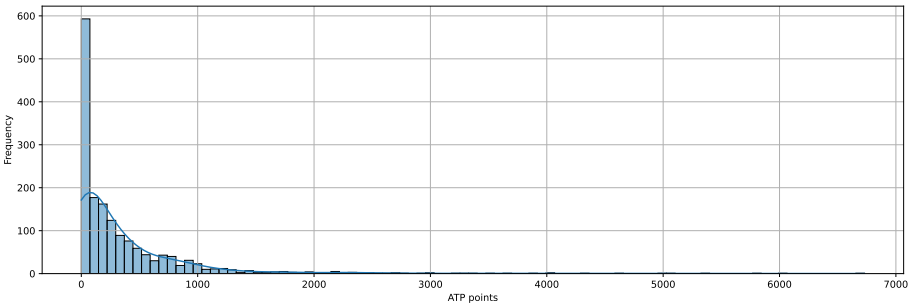

FIG. 6. Distribution of atp points

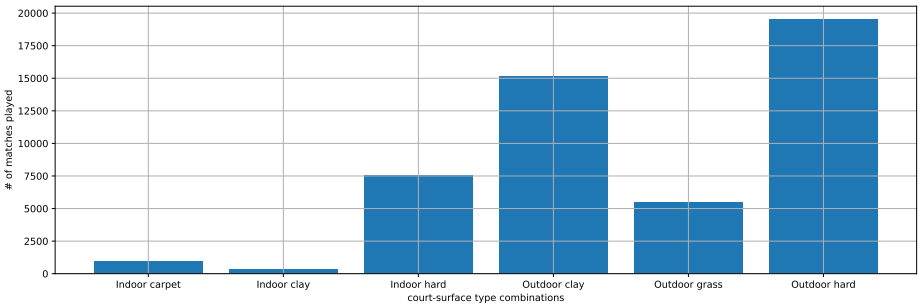

FIG. 7. Total number of matches held on each court/surface type

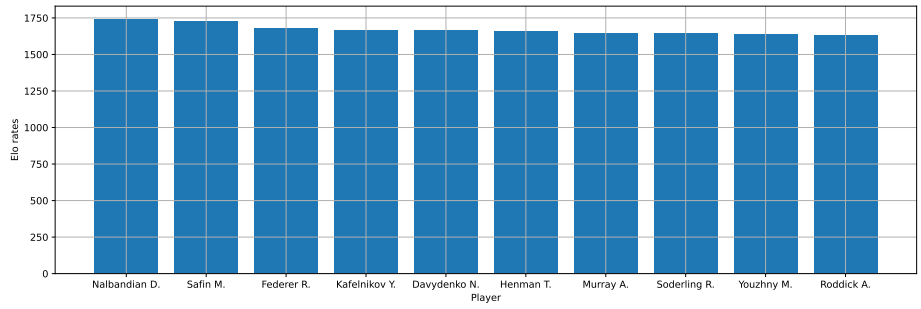

FIG. 8. Top 10 players in indoor courts and on carpet surfaces

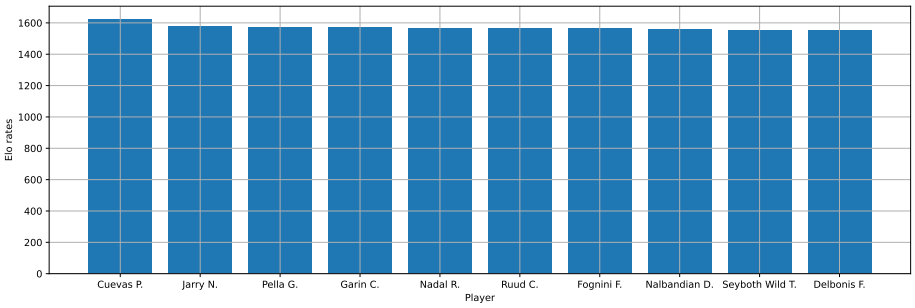

FIG. 9. Top 10 players in indoor courts and on clay surfaces

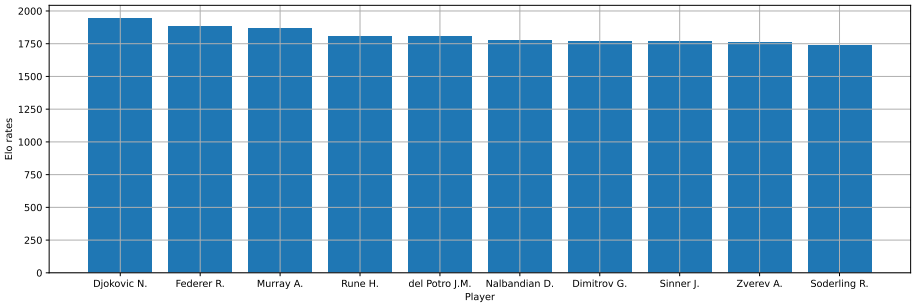

FIG. 10. Top 10 players in indoor courts and on hard surfaces

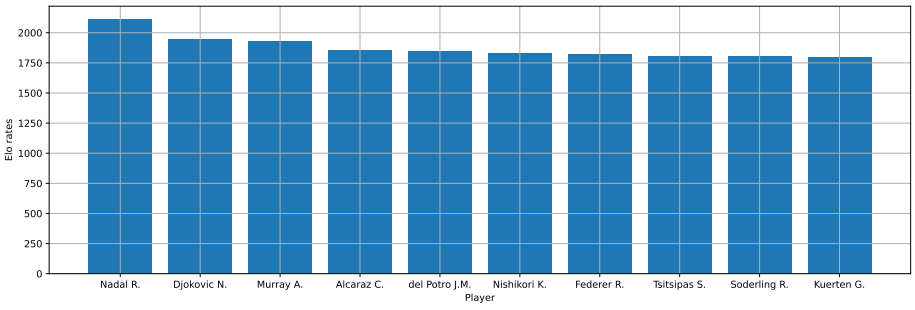

FIG. 11. Top 10 players in outdoor courts and on clay surfaces

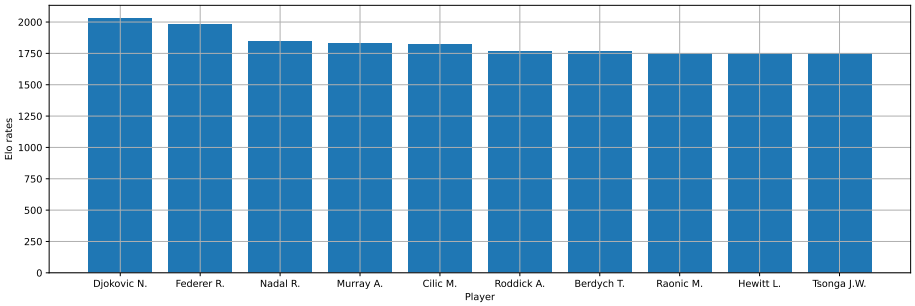

FIG. 12. Top 10 players in outdoor courts and on grass surfaces

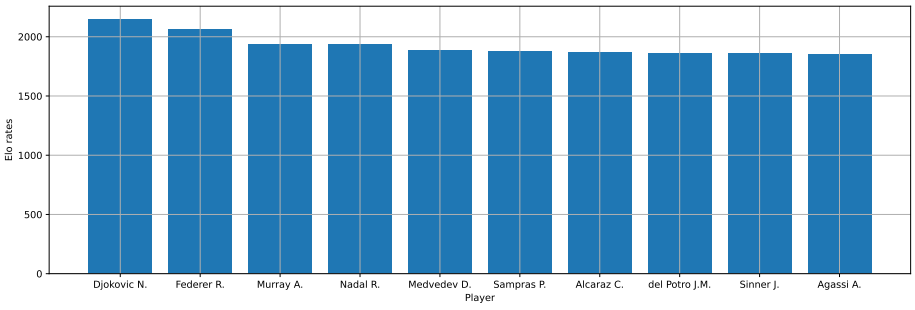

FIG. 13. Top 10 players in outdoor courts and on hard surfaces

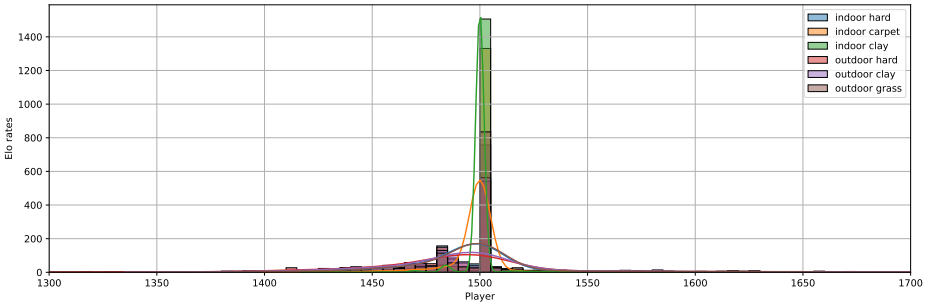

FIG. 14. This graph shows the distribution of Elo rates on all court/surface type combinations. As it is observed, Elo rates are highly concentrated about 1500 in indoor carpet and clay field types in comparison with other court/surface types. This means that Elo rates in these field types carry less information about the skills of the players, as they have more or less the same rates. Accordingly, it doesn't sound reasonable to rely on the Elo rate for prediction in these field types.