MatchMetric Analytics – Unveiling Performance Insights and Trends in Tennis

Sai Prasanth Kanamarlapudi
College of Engineering
George Mason University
Virginia, United States
skanama@gmu.edu

Harry J. Foxwell
College of Engineering
George Mason University
Virginia, United States
hfoxwell@gmu.edu

***Abstract*—The field of sports analytics has had tremendous growth in the last few years, enabling previously unheard-of levels of understanding of player performances and game dynamics. With thorough data analysis and investigation, MatchMetric Analytics, an in-depth investigation of tennis analytics, aims to reveal the complex layers of professional tennis. This study explores the amount of information gathered from multiple tournaments, including player statistics, match results, tournament features, and environmental elements. The study uses cutting-edge statistical methods, machine learning algorithms, and graphic depictions to identify underlying trends, patterns, and predictive models in tennis. Using a thorough examination of player characteristics, such as age, handedness, rankings, and past performances, the study clarifies the influence of many elements on match results and player performance. It also closely examines how match length and competitiveness are affected by tournament-level factors like surface type, draw sizes, and tournament classifications. The results shed light on the complex dynamics of tennis matches and provide insightful information about player strategies, strengths, and weaknesses. Additionally, this research fosters a greater understanding of the sport and its changing landscape by providing coaches, players, analysts, and aficionados with a basic resource. MatchMetric Analytics is proof of the ability of data-driven insights to clarify the subtleties of professional tennis and influence future tactical choices and player development in the game.**

***Index Terms*—Tennis Analytics, Sport Analytics, Statistical Analysis, Machine Learning, Predictive Models, Player Performance Metrics.**

I. Introduction

The individualized nature of tennis and the variety of elements that affect performance make it a special case for in-depth analysis. One interesting research question is how to comprehend the complex interactions that exist between player traits, match conditions, and match outcomes. This work aims to provide a greater understanding of tennis dynamics by disentangling the complex web of relationships between different elements and match outcomes.

The research holds relevance as it has the potential to yield significant insights into the game of tennis. Through a thorough examination of a large dataset that includes player statistics, match results, tournament characteristics, and environmental factors from several tournaments, this study aims to identify trends, identify patterns, and build predictive models. These insights give tennis enthusiasts a deeper understanding of the game in addition to being beneficial to coaches, players, and commentators.

MatchMetric Analytics essentially aims to use data-driven insights to interpret the nuances of professional tennis and inform player development and strategy decisions. This project is set out to answer 1) Can we predict the player's winning chance based on the player and match parameters? 2) How much

This study aims to propose statistical and machine learning methods to analyze the player characteristics and the match statistics. It aims to use various statistical and graphical methods to answer the identified problems along with a machine-learning model.

As a result, a machine learning model was successfully developed to predict the player's winning chance based on all player and match parameters which has an accuracy of 78.30%. Few graphical methods were employed to find that the matches are 62.81% favorable towards player rankings and it is statistically proved that players with the right hand will be advantageous to win a tennis match.

To conclude, with all the player's parameters and the match parameters inputted, the model developed can successfully predict the player winning chances by 80%, and players with better ranks tend to win matches more obviously than a player with lesser rank. Also, a right-handed player has a chance of 51.26% to win a match while a left-handed player has just 44.81% chances.

II. Literature Review

The Analysis and Forecasting of Tennis Matches by Using a High-Dimensional Dynamic Model [4] article proposes a high-dimensional dynamic model for tennis match results with time-varying player-specific abilities for different court surface types. The researchers also considered several factors which are player-specific explanatory variables and the grand slam tournament match configurations. The results are used to construct rankings of players for different court surface types. This article's aim is relevant to my first research question in the way that the article proposes a model to forecast the match outcomes based on player strengths and court surface types and my research aims to determine the probability of a player winning the match based on the player rankings.

An article, Effect of a Seeding System on Competitive Performance of Elite Players During Major Tennis Tournaments [1] discusses the seeded and non-seeded players considering match performance statistics. This article considered several physical performance parameters, tactical parameters, and match performance factors and found out the effect of seed rankings in men's grand slams. This article is relevant to my second research question as this considers physical parameters and other factors and founds that there is a significant decline in performance with age.

Another article, Automated processes in tennis: Do left-handed players benefit from the tactical preferences of their opponents? [2] discuss whether left-handed people benefit from the tactics of their opponents. The researchers performed two studies to test the bias in tactics. The first test included 108 right and left-handed players with varying skills and analyzed their performance. The second study included 54 professional tennis matches involving both right-handed and left-handed and analyzed the ball placement frequencies. The results showed that left-handed players had a significant advantage in tennis. This article is different from my third research question as my question considers the distribution of match winnings across gender, tourneys, surface, etc., which is completely different from this research article.

III. Strategy & Methods & Tools

This study starts with performing various data cleansing techniques like finding null values and then either imputing the null values with value 0 or dropping the value according to the data description. Then specific tournament names were changed according to the column naming convention and then trimmed all the white spaces in the entire dataset. Then the data types are properly formatted to fit the analysis and a new month column has been extracted from the existing tournament date column. This preprocessed dataset has been downloaded from R studio and then used in Python for analysis.

To understand the data, a bar plot has been plotted describing the count of matches by tournament surface and found that nearly 1200 matches were played on hard surface type, 800 matches on clay, and 300 on grass till September 11, 2023, from the beginning of 2023. Another line graph was designed to find the number of matches played by players according to their seed value. It was found that almost 130 matches were played by the player of seed 1, 100 by seed 2, and so on. Furthermore, a bar plot was plotted to see the distribution of matches according to draw size and observed that 96 were played among draw size of 4, 48 among 18, 1031 among 32, 306 among 64, and 888 among 128. Lastly, another bar plot has been created to analyze the age distribution of winners and found that most of the winners are among the 26-30 age group.

Creating an RDS database in AWS and connecting to it from MySQL workbench locally, a query has been executed to fetch the count of matches per tournament level and found that 508 Grand Slams, 545 Master 1000's, 96 Davis Cup, and 1220 other tourney-level matches have been conducted. Another query to find the average match duration was executed and found that the average match duration was 106.3 minutes.

A new data frame has been created by separating the winner and loser columns from the original dataset and then a row bind operation has been performed on these two data frames to create new data with an extra column of match result which will act as the target variable for the model and all other variables are feature variables. A logistic model has been built and an accuracy of 78.3

To find the match-winning % favorable according to player rankings, a new column has been created in Python by comparing the winner rank and loser rank with values Yes and No which is a result of comparison between the 2 columns. Then a division operation is performed by dividing the sum of better ranks counting as Y with the total number of matches multiplied by 100 which yielded a result of 62.81

To analyze which hand is more favorable to win a match, the value count of 'R' and 'L' has been performed in winner hand and loser hand columns and concatenated the two values to find the total value counts and the percentage of wins by hand is calculated resulting in 51.26

A variety of tools like RStudio, Python, Aws, SQL, and frameworks like pandas have been utilized in this analysis.

IV. Results

A. Exploratory Data Analysis - Univariate and Multivariate

Fig 1 shows the count of matches according to the Tournament surface. From January 2023 to September 11, 2023, nearly 1200 matches were played on the Hard surface, 850 matches on the Clay surface, and 350 matches on Grass surface.

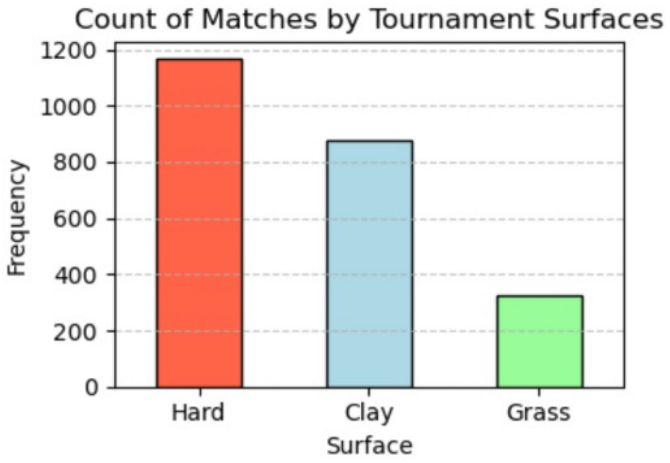


Fig. 1. Count of Matches by Tournament Surface

Fig 2 explores the no of matches won by the players corresponding to their seed values. 125 matches had been won by players with seed 1, 100 by seed 2, and so on. It is also observed that players with a seed beyond 10, had won not more than 25 matches, and players beyond 20 had won less than 10 matches.

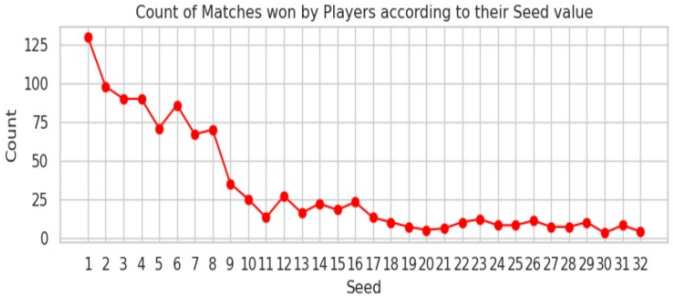


Fig. 2. Count of Matches Won by Players according to their Seed Value

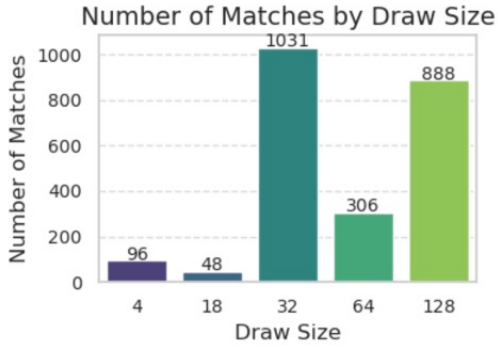


Fig. 3. Number of Matches by Draw Size

Fig 3 shows the distribution of matches by draw size. 96 matches were played with a draw size of 4, 48 matches with a draw size of 18, 1031 matches with a draw size of 32, 306 of 64, and 888 with a draw size of 128. The most common is a draw size of 32.

In Fig 4 bar plot depicts the distribution of age among winners. All the winning players were among the ages of 0 to 50 which is true in general as players would retire after 50. Most of the winners were players between the age of 26 -30 followed by 21-25. There were considerable players in the teenage also.

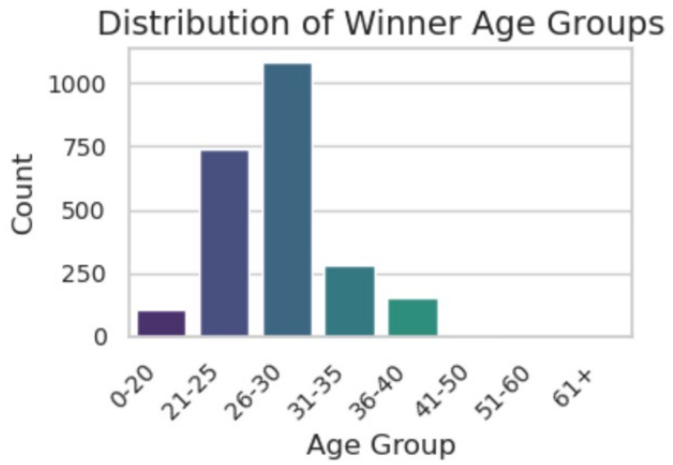
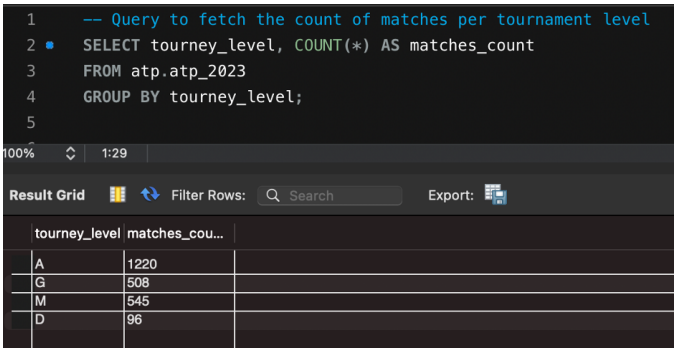


Fig. 4. Distribution of Winner Age Groups

Fig 5 shows pair wise plots among Player Age, Player Rank and Player Hand. Right-handed players were represented by blue dots, left-handed players by orange dots and green dots indicates players with unknown handedness. In all the plots right-handed players had a pure domination over all other players. Almost all the players were of rank below 1000.

Fig. 5. Pairwise Relationships Across Age, Rank and Hand

Fig 6 shows the distribution of matches across different tournaments. It can be seen that 508 Grand Slams were conducted, 545 Master's 1000 were played, 90 Davis Cup matches and 1220 all other tournament matches. This query is executed in SQL by connecting to the AWS RDS instances locally.



```sql
-- Query to fetch the count of matches per tournament level
SELECT tourney_level, COUNT(*) AS matches_count
FROM atp.atp_2023
GROUP BY tourney_level;
```

100% 1:29

Result Grid Filter Rows: Search Export:

tourney_level	matches_cou...
A	1220
G	508
M	545
D	96

Fig. 6. Distribution of Matches across Different Tournaments

Fig 7 shows the average duration of a tennis match in the tournaments. It is found that on average each match lasts for almost 107 minutes.

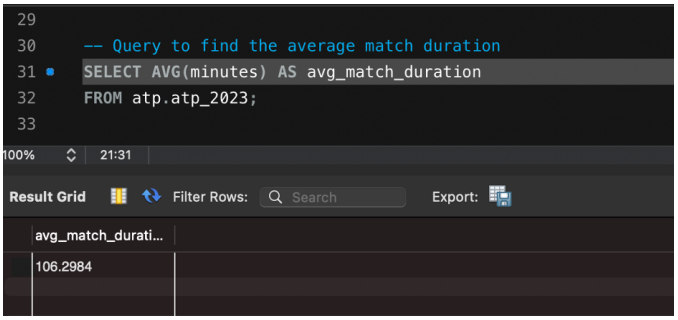


Fig. 7. Avg Duration of Match

Also, brief data analysis was performed using the AWS Glue Data Brew tool. It was used to understand more features of the data.

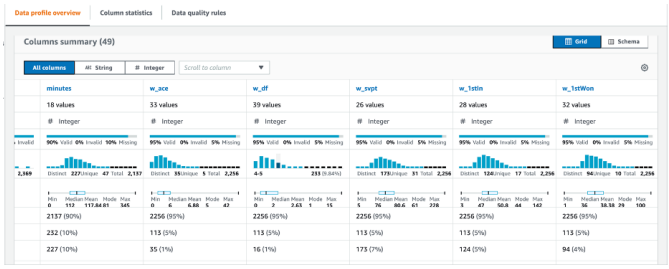


Fig. 8. Data Exploration using AWS Glue Data Brew

B. Research Question 1 Can we predict the player's winning chance based on the player and match parameters?

The original dataset has been modified according to the requirements. From the winner dataset the winner-specific columns and some common match-specific characteristics combined new data frame. Similarly, loser-specific columns and some common match-specific characteristics were combined into another data frame. A result column won or lost has been added to both the data frames and these 2 data frames were row bound to a new data frame. The resulting data frame is a cardinality of 4738 with degree 23. A 70 30 split was performed on the data and a logistic regression model was built on the training data and evaluated the accuracy by making predictions on the test data. The resulting model has an accuracy of 78.30%. Parameters like svpt, 1stWon, 2ndWon, bpSaved, bpFaced, player rank, and surface are statistically significant.

```
> accuracy
[1] 0.7830389
> # Summary of the model
> summary(model)

Call:
glm(formula = Result ~ ., family = binomial, data = train_data)

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.580e+00  5.932e-01   2.663 0.007747 **
seed         2.553e-03  5.168e-03   0.494 0.621372
entryAlt     1.987e+00  1.398e+00   1.421 0.155404
entryLL     -8.318e-02  3.076e-01  -0.270 0.786801
entryPR     -1.522e-01  4.795e-01  -0.317 0.750975
entryQ      -9.866e-02  1.789e-01  -0.551 0.581308
entrySE      9.453e-01  8.756e-01   1.080 0.280348
entryWC     -4.320e-01  2.319e-01  -1.863 0.062504 .
handR        1.041e-01  1.713e-01   0.607 0.543549
handU        4.670e-01  4.252e-01   1.098 0.272084
ht170        6.334e-01  4.519e-01   1.401 0.161080
```

Fig. 9. Logistic Model Summary and Accuracy

C. Research Question 2 How much % is the match favorable towards the player rankings?

To determine this, a comparison was made between the winner rank and loser rank columns, and a better rank column was created with value Y if the winner rank is less than the loser rank and N if the winner rank is greater than the loser rank. The comparison showed that in 62.81% matches players with a better rank had a positive result. So we can conclude that the matches were favorable to the player rank by 62.81%.

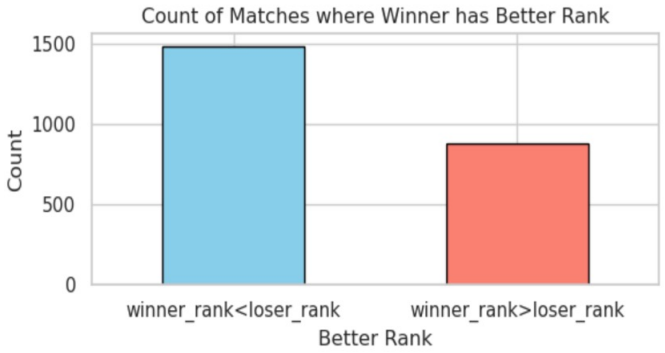


Fig. 10. Player Rank effect on Match Result

D. Research Question 3 Who is more advantageous to win a match? Right-handed or Left-handed?

The disparity in skill between right- and left-handed tennis players has long been a source of curiosity. Examining the results of matches reveals an intriguing pattern: right-handed players had a little greater winning percentage than left-handed players. Right-handed players win around 51.26% of the time, whereas left-handed players win about 44.48% of the time. These figures provide a compelling picture of the situation. This slight difference suggests that right-handed tennis players have a small advantage in the competitive game. However, because of the dynamic nature of the sport, different playing styles and techniques frequently test and contradict statistical

standards, making every game an exciting and unpredictable spectacle.

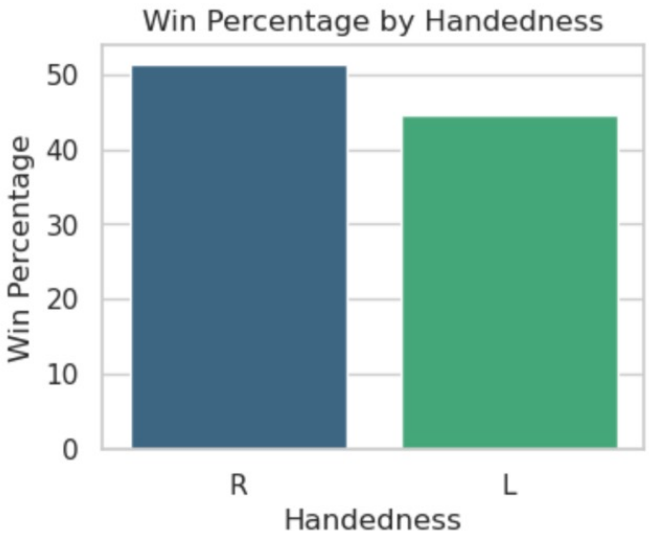


Fig. 11. Enter Caption

V. Limitations

All the results of this project are a result of the analysis conducted on just the matches that occurred in 2023 with male players. These results will vary when an extensive set of data is analyzed including all genders and all types of tournaments not just limited to singles. But remember that tennis is a dynamic sport, and all the analysis might go wrong depending on the players, match type, surface type, and even match day will have an impact. The player's performances might go high and low with time and results can change.

VI. Future Research

For more significant and strong results, an analysis of all the matches conducted is needed including all the genders. In this analysis, the climate is not considered but it will play a significant role on players performance. Hence in the future, climate should also be considered. Also, players' physical strength consideration and match-day player behavior should be considered for effective results as all these are dynamic. A deeper analysis of match scores, like set-wise scores, and tiebreakers analysis will help us understand more about the match. A long-term study on players' performance will help us to understand the players more accurately.

VII. Conclusion

To conclude this project provides a brief analysis of players' performance and match statistics with a model to predict the match-winning percentage with players and match specific parameters as inputs. With this analysis, right-handed players had better chances of winning the game, and the player with a better rank had a good probability of winning the game.

References

[1] J. Sackmann, "ATP Matches 2023," GitHub, https://github.com/JeffSackmann/tennisatp/blob/master/ atp-matches2023.csv (accessed Dec. 3, 2023).

[2] F. Loffing, N. Hagemann, and B. Strauss, "Automated processes in tennis: Do left-handed players benefit from the tactical preferences of their opponents?," Journal of Sports Sciences, vol. 28, no. 4, pp. 435–443, 2010. doi:10.1080/02640410903536459

[3] P. Gorgi, S. J. Koopman, and R. Lit, "The analysis and forecasting of tennis matches by using a high dimensional dynamic model," Journal of the Royal Statistical Society Series A: Statistics in Society, vol. 182, no. 4, pp. 1393–1409, 2019. doi:10.1111/rssa.12464

[4] Y. Cui et al., "Effect of a seeding system on competitive performance of elite players during major tennis tournaments," Frontiers, https://doi.org/10.3389/fpsyg.2020.01294 (accessed Dec. 3, 2023).