Background

The sport of tennis has a rich history dating back to at least 1873, evolving over the years to become the dynamic game it is today. Advancements in equipment, strategy, rules, and playing surfaces have significantly shaped the modern version of tennis. In 2023, the rules were revised to allow coaching during matches. This introduces a valuable opportunity for in-match analysis and strategic adjustments for players and their coaching teams. With the increasing emphasis on data-driven decision-making in sports in general, there is great potential for leveraging advanced analytics to enhance player performance, optimize game strategies, and gain a competitive edge on the court.

Motivation

My analysis aims to identify variables that contribute to winning a tennis match. I will center my analysis around the following questions:

- **Which variables impact whether a tennis player wins a match?**
  This question forms the core of the analysis, seeking to identify the most influential factors that determine match outcomes. By examining a range of variables, including match statistics, player demographics, and tournament characteristics, we can pinpoint predictors of success on the tennis court.
- **Do player age and height influence winning a match?**
  Exploring the relationship between player age, height, and match outcomes can provide valuable insights into the role of physical attributes in tennis performance. Understanding how age and height may impact players' abilities to compete effectively can inform training programs and talent development strategies.
- **Do top players face fewer break points?**
  Examining the frequency of break points faced by top-ranked players compared to their lower-ranked counterparts can offer insights into the strategies employed by top players to maintain their competitive edge during a tennis match.
- **Which countries produce top tennis players?**
  Investigating the distribution of top tennis players across different countries can provide insights into the global landscape of tennis talent development. Identifying countries with a strong track record of producing elite players can inform recruitment efforts and talent scouting.
- **Do the variables that impact whether a tennis player wins a match depend on the surface type?**
  Exploring how surface type influences the significance of various match variables can enhance our understanding of the nuances of tennis strategy and gameplay. By analyzing match data across different surfaces (grass, clay, hard court, carpet), we can uncover surface-specific patterns and trends in player performance and match outcomes.

Description of the Data

The tennis data I will be using was collected and shared by Jeff Sackmann via his Github repository. Match data was saved in separate CSV files for the years 1968 to present. I am focusing on ATP tour level singles matches from the past twenty years (2003-2023). This includes data from 61,932 matches. The dataset includes 49 variables. These variables include information about the tournament (surface type, size of draw, level, etc.), demographic

information about the winner and the loser (handedness, height, age, country, etc.), and match statistics about the winner and the loser (aces, number of break points faced, number of double faults, etc.).

The data is relatively clean, but there are a few variables that I plan to recode. This includes winner and loser seed and winner and loser entry. There are a lot of NA values for these variables, because most players will enter a draw unseeded (thus having an NA value), and most players are part of the main draw without having to play a qualifying tournament prior, receiving a wild card entry, or having another special circumstance regarding their entry into the tournament. I will create a binary variable for seeding that will indicate whether the player is seeded or unseeded. I may also create binary variables for the entry variables, or I may exclude them entirely.

Because the data for the winner and loser are contained in one row for each match ('winner_age' and 'loser_age'), I plan on reformatting the data so that match data is on two lines, consolidating 'winner_age' and 'loser_age' to 'age' and coding a response variable 'outcome' for 'winner' or 'loser'. This will be essential for performing analyses.

In addition to the match dataset, I have access to a CSV of player information and a CSV of ranking information. I don't anticipate needing these CSVs because ranking and demographic information are already included in the match dataset.

Proposed Analysis

The focus of my proposed analysis of tennis matches is one of classification: predicting whether a player will win or lose a tennis match. This comprehensive analysis will employ various statistical techniques including logistic and linear regression analysis, linear discriminant analysis (LDA), Quantitative Discriminant Analysis (QDA), random forest, chi-square tests, and visual representations of the data to enhance our understanding of match outcomes.

Through logistic regression analysis, I will investigate the influence of various variables on match outcomes. Logistic regression is well-suited for this task due to its ability to handle categorical response variables and provide interpretable coefficients for each predictor variable. I will explore whether demographic factors such as height, age, or country of origin impact a player's likelihood of winning a match, as well as the significance of in-match performance variables such as first serves made and break points faced. In addition to logistic regression, I will explore other classification models such as Linear Discriminant Analysis (LDA), Quantitative Discriminant Analysis (QDA), and random forest. By comparing the performance of these models, I can determine which approach provides the most accurate predictions of match outcomes. This comparative analysis will allow me to identify the most effective modeling technique for the dataset overall.

Next, I can use linear regression analysis to examine the relationship between player age and height and match outcomes. Similarly, I could also conduct correlation analyses to assess the strength/direction of the relationship between player age or height and match outcomes. This analysis will provide insights into the impact of physical attributes on player performance and match outcomes.

To investigate whether top players face fewer break points, I can compare the mean number of break points faced by top-ranked players and lower-ranked players using either an independent samples t-test (if the data is normally distributed) or a Mann-Whitney U test (if the data is non-normally distributed). This analysis will help determine whether top players really do

face fewer break points on average compared to lower-ranked players. I will also need to decide what ranking I will consider a “top player”.

I will use a chi-square test of independence to examine the relationship between player nationality and player ranking. This test will help determine whether there is a significant association between the country of origin of players and their ranking. Additionally, I can visualize the distribution of top-ranked players across different countries using a heat map.

To address the question regarding surface type, I can use a multi-[way ANOVA test to assess the interaction effect between surface type and other predictor variables on match outcomes. This analysis will help determine whether the impact of certain variables on match outcomes varies across different surface types, providing insights into the influence of surface conditions on player performance.

By leveraging these various analytical techniques, I will be able to provide insight into what variables are important in influencing a tennis match outcome.

References

Lorge, Barry Steven , Bruce, Morys George Lyndhurst and Aberdare, 4th Baron. "tennis". Encyclopedia Britannica, 6 Mar. 2024, https://www.britannica.com/sports/tennis. Accessed 9 March 2024.

Sackmann, Jeff, “tennis_atp”, (2024), GitHub Repository, https://github.com/JeffSackmann/tennis_atp/tree/master.

Appendix – Variable Descriptions

The raw data: https://github.com/JeffSackmann/tennis_atp

After combining match data CSVs from 2003 to 2023, there are 61,932 observations with 49 variables. The variables are described as followed:

- tourney_id - a unique identifier for each tournament, such as 2020-888. The exact formats are borrowed from several different sources, so while the first four characters are always the year, the rest of the ID doesn't follow a predictable structure.
- tourney_name
- surface – Type of court surface: carpet, clay, grass, or hard.
- draw_size - number of players in the draw, often rounded up to the nearest power of 2. (For instance, a tournament with 28 players may be shown as 32.)
- tourney_level - 'G' = Grand Slams, 'M' = Masters 1000s, 'A' = other tour-level events, 'C' = Challengers, 'S' = Satellites/ITFs, 'F' = Tour finals and other season-ending events, and 'D' = Davis Cup
- tourney_date - eight digits, YYYYMMDD, usually the Monday of the tournament week.
- match_num- a match-specific identifier. Often starting from 1, sometimes counting down from 300, and sometimes arbitrary.
- winner_id - the player_id used in this repo for the winner of the match
- winner_seed
- winner_entry - 'WC' = wild card, 'Q' = qualifier, 'LL' = lucky loser, 'PR' = protected ranking, 'ITF' = ITF entry, and there are a few others that are occasionally used.
- winner_name

- winner_hand- R = right, L = left, U = unknown. For ambidextrous players, this is their serving hand.
- winner_ht- height in centimeters, where available
- winner_ioc- three-character country code
- winner_age- age, in years, as of the tourney_date
- loser_id
- loser_seed
- loser_entry
- loser_name
- loser_hand
- loser_ht
- loser_ioc
- loser_age
- score
- best_of- '3' or '5', indicating the the number of sets for this match
- round
- minutes- match length, where available
- w_ace- winner's number of aces
- w_df- winner's number of doubles faults
- w_svpt- winner's number of serve points
- w_1stIn- winner's number of first serves made
- w_1stWon- winner's number of first-serve points won
- w_2ndWon- winner's number of second-serve points won
- w_SvGms- winner's number of serve games
- w_bpSaved- winner's number of break points saved
- w_bpFaced- winner's number of break points faced
- l_ace
- l_df
- l_svpt
- l_1stIn
- l_1stWon
- l_2ndWon
- l_SvGms
- l_bpSaved
- l_bpFaced
- winner_rank- winner's ATP or WTA rank, as of the tourney_date, or the most recent ranking date before the tourney_date
- winner_rank_points- number of ranking points, where available
- loser_rank
- loser_rank_points