# CHAPTER 1: THE GAME OF TENNIS + WINNER PREDICTIONS

1. **Rules of the game**

The scoring system in tennis is hierarchical: matches are composed of *points*, *games* and *sets*. The first player to win 4 points and at least 2 more than his opponent wins the game. Points are counted using the sequence 0, 15, 30, 40. If the score reaches 40-40, this is called a *deuce* : at this point, each player needs to win two consecutive points to win the game. Scoring a point after the deuce gives a player the *advantage* : scoring the next point will win this player the game. If the next point is lost, the score returns to a deuce.  
In order to win a set, a player must win at least 6 games and at least two more than his opponent. If the game score reaches 6-6, players play a tie breaker game, which continues until one player has scored at least 7 points and at least 2 more than his opponent. \\  
Matches are scored using a best-out-of-three or best-out-of-five set system, meaning that the first player to win either two or three sets wins. Which scoring system is used depends on the tournament. \\

At the beginning of every point, the ball is put in play by an overhead shot called the *serve*. The player who puts the ball into play is called the *server*, and the opposing player is called the *receiver*. If the server hits the *first serve* into the net or out of bounds, he is awarded a *second serve*. The server alternates at every game. The server has a big advantage and is on average more likely to win the game: this is illustrated by the all-time statistics of the Association of Tennis Professionals (ATP)[[1]](#footnote-1). The player having won the highest percentage of service games over his career is John Isner at 91.8 %, whereas the leader in terms of return games is Guillermo Coria at only 35.3%.

The fact that “breaking” is so important comes from the fact that winning a set on return is a small upset given the average serve game win probz

**ATP Tour**

The data we will use in our analysis comes exclusively from ATP tour matches. ATP is the largest men’s professional tennis organization, and the ATP tour is the highest tier matching up the highest ranking men in the most prestigious tournaments. Four categories of tournaments exist, granting different amounts of prize money and ranking points. These categories are:

**1.ATP Tour tournaments**, including:

- ATP Tour Finals

-ATP Cup  
-ATP Tour Masters 1000  
-ATP 500  
-ATP 250

**2.ATP Challenger Tour tournaments**

**3. Grand Slam tournaments** ( US Open, Australian Open, Roland Garros, Wimbledon)

**4. Davis Cup.**

ATP tour finals

All ATP matches are played in a best-of three tiebreaker sets format. (depuis quand) The qualification and seeding of players for the tournaments is dependent upon the ATP rankings, which are established based on matches won and lost during ATP events.

PARAGRAPH: ATP points determine who qualifies for big events, lower players can decide which tourneys they play in, gaps in prize money (cite prize money study).

Who can benefit from predicting winners in tennis? (Stake holders)

=Players and coaches: compare to other sports where AI/ predictive modeling has helped   
-Media: entertainment value  
-Bettors + Bookmakers: need to refine their model as much as possible, using model can help bettors make educated guesses

**TENNIS INDUSTRY STAKEHOLDERS**

Collecting data on tennis matches and gathering insights enabling educated predictions is something that can benefit different actors in and around the tennis industry. \\

The most obvious group are the players themselves and their coaching staff. Analyzing data can help identify patterns in their own play and their opponent’s play, in order to fix leaks or find areas to exploit. Multivariate analysis can prove useful to gauge the influence of factors like surface or certain characteristics of the opponent on the player’s performance. Being able to aggregate the data and run analytics over multiple matches of a player can uncover patterns that might not be as apprehensible through human analysis. \\  
Tennis has been considered to be lagging behind in the use of analytics when compared to other sports. In the first issue of ESPN’s *The Magazine: Analytics[[2]](#footnote-2)* from 2012, sports were ranked based on the state of their analytics, and tennis was ranked second to last, outranking only boxing. Speakers at the MIT Sloan Sports Analytics Conference specialized in tennis analytics, like Craig O’Shanessy[[3]](#footnote-3) and Jeff Sackmann[[4]](#footnote-4), have attributed this to the lack of budget dedicated by players to analytics in tennis. Contrary to team sports like basketball or football, players cannot take advantage of their team’s investment in advanced analytics. However, as data collection and analysis technologies advance and our world as a whole becomes more aware of the power of data, it seems that the sport is beginning to catch up. In 2015, the Women’s Tennis Association (WTA) announced a partnership with the “market leader in enterprise application software” SAP[[5]](#footnote-5), resulting in the creation of the “SAP Tennis Analytics for coaches” software. In certain WTA tournaments, coaches are provided with tablets collecting real-time data from cameras and the umpire’s scorecard, that they can use to analyze the play and coach their player accordingly. Top ranked ATP players like Novak Djokovic and Rafael Nadal have added data analysts to their coaching teams, and companies like IBM, Tennis Analytics and Brain Game Tennis have developed platforms to help professional players analyze their games. \\  
  
Another category of stakeholder are the actors in the tennis media. Data analysis can be used to engage fans and make commentary more interesting. \\

Finally, tennis analytics are key for actors on the betting market. On the one side, informed bettors try to analyze data in order to make educated predictions that they will subsequently bet money on. The ones who have the most to gain from high quality predictive models, however, are the bookmakers. In 2015, the tennis betting market was estimated at 5 billion dollars and growing[[6]](#footnote-6). In 2012, Andy Murray’s Wimbledon final beat the record for biggest betting market on Betfair, the world’s largest online betting exchange, with more than £58 million of bets placed. The sport of tennis lends itself well to sports betting, given the great amount of tournaments played over a year that give many betting opportunities all year round. \\

**2012 SLOAN SPORTS ANALYTICS CONFERENCE- TENNIS ANALYTICS**

**AVAILABLE DATA**

“Tennis data” is an umbrella term that can be used to

Match data, umpire scorecards (photo), something else, Hawk-Eye

This section begins by detailing a few theoretical concepts necessary to understand tennis win prediction studies as well as the methods that will be used in this study.

1. **MARKOV CHAINS**

Markov chains are a type of stochastic model that can be used to represent events or systems. The model is made up of different *states* that make up the *state space* and are linked to one another, hence being called a chain. Different states are related to one another because the transition from one state to the other happen with a certain probability. The model must also respect the Markov property of “memorylessness”, meaning that the probability of transitioning to one state solely depends on the present state and not on previous movements. \\

Markov chains are named after Russian mathematician Andrey Markov and are widely used to describe real-world phenomena involving random processes. Due to the hierarchical scoring structure of tennis, they lend themselves well to the modeling of tennis matches. Every score line is a state, and the probability of transitioning to another state corresponds to the probability of either player winning the next point.\\

Diagram, schematic

Description automatically generated

The above graphic is a representation of a set as a Markov Chain. Player A has a point-winning probability p. (moet wss zelfgemaakte grafiek zijn, en p-1 moet eigenlijk 1-p zijn)

1. **Regression analysis**

Regression analysis is a statistical modeling technique widely used for prediction and forecasting. To build a regression model, one must define a *dependent variable* (or *outcome* *variable*) that is the variable of interest that one wants to be able to predict, and one or multiple *independent variables* (or *input variables*) linked in some way to the dependent variable. Regression can be viewed as a task of function approximation in that the model infers the relationship that links the dependent variable to the independent variables by “learning” from the observed relationships from a random sample and aggregating the information into a mathematical relation.   
  
One of the most basic forms of regression that is often used to introduce the concept is linear regression.

**B1) Linear regression**  
  
The word “linear” in linear regression stems from the fact that this type of model establishes a linear relationship between the dependent variable and the independent variables. The following is called a multiple regression model because it has multiple independent variables.

Y= β0 + β1X1 +β2X2 + …. +βKXK + u

Y is the dependent variable; X1, X2, XK are the independent variables; β0,β1,β2,βK are the parameters and u is the error term. This equation is established for every observation in a sample: Y and Xi are observed, whereas the βi and u are unobserved. The function approximation happens by estimating the parameters βi . A common technique for this estimation is the *ordinary least squares method* (or OLS), consisting of determing the estimators that minimize where denotes the estimator of This expression is the sum of the squares of the differences between the estimated and the observed values of the dependent variable: finding the estimators of the parameters that minimize this expression is a way of verifying the quality of our model. The regression function is in fact the conditional probability of Y given the dependent variables : .

Linear regression models are used when the output variable Y is quantitative. However, when trying to predict the outcome of a tennis match, the desired output variable is categorical in nature: win or lose, which can be represented by a 1 or a 0. In this case, a *logistic regression* is a more adequate model.

B2) Logistic regression

Predicting the winner of a tennis match is a regression problem with a binary dependent variable, which can also be called a classification problem. This means that the only possible outcomes of our model are Y=1 or Y=0. Moreover, the expected value of Y is equivalent to the probability that Y =1:

Thus it makes sense to use a function which returns values between 0 and 1, representing a probability. The *probit* and *logit* regression presented hereafter both do this but using a different function: the probit regression uses the standard normal cumulative probability distribution function, whereas the logit regression uses the logistic cumulative probability distribution function.

Probit model:

Logit model: where

Graphics: Normal cdf, logistic cdf

These two models give very similar results. Historically, the logit regression was preferred because the values of the logistic cumulative probability distribution function could be computed faster than the values of the standard normal cumulative probability distribution function, but this does not make a difference nowadays.

Include econometry hypothesis testing for regression? Multicollinearity, Heteroskedasticity?

1. **Neural networks (**[**https://www.nature.com/articles/nature14539#Sec1**](https://www.nature.com/articles/nature14539#Sec1)**, elements stat learn)**

Neural networks are a type of statistical model that have been associated with the advent of artificial intelligence. Over the last 10 years, neural networks combined with large databases and rapidly evolving computing power have substantially improved the performance of computers in many domains, including speech and image recognition. These improvements are directly accessible to anyone owning a smartphone, which has helped contributing to the “hype” around neural networks.

Just like for regression models, building a neural network is a task of function approximation. The network must infer the relationship between the inputs it is given and the desired output by learning from the data it is “fed”.

There are many different types of neural networks, but the overarching principles can be illustrated by looking at the simplest form of neural network called a *multi-layer perceptron*. This neural network can be represented as a diagram made up of *nodes* or *neurons* arranged in layers: an input layer, one or multiple hidden layers, and an output layer. Every neuron is connected to all neurons in the previous and following layers. Neurons are, in fact, computing units: they take a certain value as inputs from the previous layer (or from the user in the case of the input layer), apply a function to a linear combination of these inputs to compute an output value that then becomes an input for the neurons of the following layer.

The linear combinations in every neuron and the function that are applied depend on many unknown parameters called *weights*. The “learning” or training process of the model consists of the adjustment of these parameters based on the training data. Training data is data for which the desired output is known: when trying to predict the winner of tennis matches, this means matches for which the winner is known. In this case, one data point corresponds to one tennis match: the input values can be any characteristics determined by the user, like measures of the quality of the players or of the circumstances of the match ( average serve win percentage, percentage of matches won in the past, surface played on…), and the desired output values are either a 0 for “match lost” or a 1 for “match won” ( or the estimated probabilities of these events). When presenting a match to the untrained network, it will compute an output based on the input values and the weights that were fixed at that moment. This is called a *forward pass*. The output values computed are then compared to the observed output, and a measure of the error in the model’s prediction is made. This measure is then utilized to compute the adjustments of the weights to be made. This is called a *backward pass*. The combination and repetition of forward and backward pass make up the *backpropagation algorithm*, which is in fact the mechanism by which a neural network can improve its predictions: the more data it is fed, the better it is able to model the relationship between input and output.



Fig: Structure of a multi-layer perceptron.

Sources: Paula Gobbi Econs463, Elements of statistical learning, Stockn Watson: Introduction to Econometrics

**MONOGRAPHIE**

Researchers have examined the prediction of the outcome of tennis matches for many years. Different approaches have been taken and developed over time. As it was done by Kovalchik (2015), these approaches can be split up three main categories: point-based, regression-based and paired comparison models.

1. **Point-based extrapolation**

One of the earliest documented methods to determine the winner of a tennis match is based on the estimation of the point-winning probability for each player. The problem of determining the winner of a match is simplified to determining the winner of the next point. By using a stochastic model called a Markov chain, the probability of winning a point is linked to the probability of winning a match, and a formula for the match-winning probability taking the point-winning probability as a variable is created.

Hsi and Burych (1971) and Carter and Crews (1974) first use this approach to predict player’s probability of winning games, sets and matches, as well as the expected duration of a match. The authors determine the probability of winning a point as being approximately the average of the probability of winning a point when serving and when the opponent is serving. This probability is assumed to be constant, and the authors examine how variations in this probability affect the game, set and match-winning probabilities as well as their durations. Pollard (1983) later also develops this approach. Klaassen and Magnus (2003), Newton and Keller (2005) and Barnett and Clarke (2005) take an extra step by supplying the model with empirical data from past matches, estimating the observed point-winning probabilities and using them as inputs to predict the outcome of future matches . Additionally, Barnett and Clarke refine the point-winning probabilities by making them specific to the opponent and to the surface the match is played on instead of determining a single overall point-winning probability per player. Spanias and Knottenbelt (2012) bring further improvement by considering points themselves as Markov chains, thereby observing dynamics at an even lower level, like the likeliness of the serving player to win the point by ace or to win on his second serve. In a subsequent article, Knottenbelt, Spanias and Madurska (2012) try to make win probability estimates for every player more accurate by looking at common opponents. Instead of comparing one player’s overall probability to win at serve and the opponent’s probability of winning when returning, they only include data resulting from previous matches against opponents that both players have faced before to mitigate bias that could arise due to a difference in the level of competition that both players have faced. They also evaluate their model by computing its return on the betting market over a certain sample of games.

All these studies rest on the assumption that points in tennis can be considered random variables that are independent and identically distributed (iid), which is a necessary assumption to consider the probability of winning a point constant. Klaassen and Magnus (2001) demonstrate that this assumption specifically does not always hold but that it is a good enough approximation of reality to be employed. Although this assumption is hereby verified and many improvements have been proposed, extrapolating the match-winning probability from the point-winning probability might be an oversimplification, and information relative to the specific circumstances of a given match might be lost. (Deze zin is bof bof)

1. **Regression-based models**

As seen in the previous section, the prediction of the winner of a tennis match can be presented as a regression problem with a binary dependent variable representing win/lose. Multiple studies have been made using logit and/or probit models. For these type of models, the challenge often lies in the choice of the independent variables: from the available data on tennis matches, researchers have tried to select and combine the features that they deem relevant for the prediction of the winner of the match and that prove to lead to the most accurate predictions.

Boulier and Stekler (1999) use a probit model where difference in ranking ( calculated from results of past Grand Slam events) is used as the single independent variable to predict the winner of a match. Clarke and Dyte (2000) and Klaassen and Magnus (2003) used a similar model using a logistic regression and the official ATP rankings. Both studies showed that rankings were significant for the prediction of the winner of a match. Del Corral and Prieto-Rodriguez (2010) further explored the question by creating a probit model including difference in ranking and 19 other variables pertaining to player’s past performance, physical characteristics and match characteristics. They evaluated predictions when removing past performance and physical characteristics variables, and found that difference in rank to be the most relevant for the prediction of the winner for every selection of variables. They found that the effect was the same for men and women, but that it was generally more important when looking at higher ranked players. Gilsdorf and Sukhatme (2007) used a probit regression with multiple independent variables including age, head-to-head record, form difference in the current year and over the entire career, as well as potential financial gain from the tournament. They found that an increase in potential financial gain had a positive effect on the favorite’s probability of winning a match. They also found the performance variables to be positively correlated with the chance of winning. Many of the regression-based models from different authors use similar independent variables, or different variables with similar general goals like giving a measure of past performance or match characteristics. Some authors like Lisi and Zanella (2017) try to incorporate match odds from bookmakers into their model to take advantage of the fact that bookmakers might have more information than the one included in the available data, and including the bookmaker odds can be a way to account for eventual unknown information. In a more recent study, Srivastava (2018) uses a logistic regression and additionally attempts to rank the type of independent variable based on the insight they provide. The author confirms ranking or equivalent point indicators to be the most important in the prediction of a winner, followed by match characteristics like surface or tournament type. Physiological variables like age, height and weight are not significant. No other player performance statistics other than ranking are used.

Official rankings are a crucial explanatory variable in most regressions used to predict match results. This reflects the fact that these rankings are established based on an aggregation of past match results, and therefore hold relevant historical information. A challenging next step lies in finding variables to significantly improve the predictive power of a model already incorporating player rankings. Scheibehenne and Broder(2007) point to the fact that bookmaker predictions can be up to 10% more successful than predictions based on rankings. McHale and Morton (2011) find that they can significantly outperform a logistic regression based on rankings by creating their own measure of strength of a player that takes into account the quality of their past opponents and the margin with which past matches were won or lost. They also give recent form more importance relative to older matches and establish their strength measure specific to surface. These measures aim to regain the information that is lost by the ATP rankings in that they award points to players for winning matches solely based on the tournament and the stage of the tournament. For the same reason, Masuda (2012) creates a dynamical ranking of players that considers the fact that player’s quality levels fluctuate over time, and a win should be proportionally more important in rating a player the stronger the opponent was.

**Paired comparison**

1. **Machine learning and Neural Networks**

Boulier and Stekler (1999), Klaassen and Magnus (2003), Gilsdorf and Sukhatme (2008), Del Corral and Prieto-Rodriguez (2010), Sipko (2015),

Title could be +- “how much better than the rank can we predict using machine learning”? Zoals emiel zei, beginnen met enkel rank lijkt logisch omdat al een soort van sleutel is om te bepalen wie beter zou zijn, en een mens kan al gewoon voorspellen wie gaat winnen door te zeggen “de hoogst gerankte”, als je dat niet kan verbeteren is je model eigenlijk nutteloos. Praten over theoretical maximum in verband met de betting market, bookmakers probably have the best models right now because their business model rests on it and there is a ton of money involved

From klaasen2003: “See Boulier and Stekler (1999), Clarke and Dyte (2000),

and Lebovic and Sigelman (2001)on the forecasting accuracy of

rankings and related issues”

look at other studies making modified elo ranks/ network ranks and try to create a new rank feature based on their study

**REFERENCES**

Scheibehenne, B.,&Broder, A. (2007). *Predicting Wimbledon 2005 tennis results by mere player name recognition.* International Journal of Forecasting, 23, 415–426.

**II. DATA**

**Database Presentation**

1. **Source**

The database we will use for our analysis was published by Jeff Sackmann, an author and software developer with a passion for tennis analytics. This database is one of the most extensive and regularly updated databases available for free on the internet.   
It is available at <https://github.com/JeffSackmann/tennis_atp> and regroups singles match records from the ATP tour from 1968 until today. Jeff Sackmann has published much more tennis related data, including match records for the WTA, ATP doubles, ATP Challenger Tour qualifiers, point-by-point match charts.. Our analysis will limit itself to the ATP singles match database as it is the most complete and holds the highest amount of matches. Furthermore, only matches from 2009 to 2019 will be used. This is to make sure that the player pool is not too large (i.e. all players have or could have played against each other), and that we are considering a comparable generation of tennis players. This is to prevent a bias in the data due to the evolution of the game of tennis. It is often said for football players, for instance, that players from older generations and their statistics (mainly number of goals scored) cannot be compared to today’s players, simply because the game has evolved and goals were scored much more often in the past as defensive tactics were much less evolved. A recent illustration of this fact has come from the EURO 2021, where Cristiano Ronaldo became the all-time top scorer at European Championship finals by scoring his 10th goal for Portugal. This was the fifth EURO Ronaldo played and scored in since his first one in 2000. The player he was tied with until then, Michel Platini, scored all 9 goals for France in one single EURO campaign in 1984.

1. **Features**

|  |  |  |
| --- | --- | --- |
| CATEGORY | FEATURE NAME | DESCRIPTION |
| General tournament and match information | tourney\_id | Unique identifier for tournament |
| tourney\_name | Name of the tournament |
| surface | Surface type of tournament |
| draw\_size | Number of enrolled players |
| tourney\_level | ATP level of tournament |
| tourney\_date | Date of tournament |
| match\_num | Match-specific identifier |
| best\_of | Number of sets played (BO3/BO5) |
| round | stage of tournament |
| Player-specific information | winner\_id;loser\_id | Unique identifier of player |
| winner\_seed;loser\_seed | player seed for tournament |
| winner\_entry,loser\_entry | way of entry to the tournament |
| winner\_name;loser\_name | name |
| winner\_rank; loser\_rank | ATP rank at date of tournament |
| winner\_rank\_points; loser\_rank\_points | ATP rank points at date of the tournament |
| winner\_hand;loser\_hand | left or right handedness |
| winner\_ht; loser\_ht | height |
| winner\_ioc; loser\_ioc | 3 letter code for country of origin |
| winner\_age; loser\_age | age |
| Post-match statistics | score | final score |
| minutes | minutes played |
| w\_ace; l\_ace | number of aces |
| w\_df; l\_df | number of double faults |
| w\_svpt;l\_svpt | number of serve points |
| w\_1stIN; l\_1stIN | number of first serves made |
| w\_1stWon;l\_1stWon | number of first-serve points won |
| w\_2ndWon;l\_2ndWon | number of second-serve points won |
| w\_SvGms;l\_SvGms | number of serve games |
| w\_bpFaced;l\_bpFaced | number of break points saved |

Every match example has 49 features that can be divided into three overarching categories: general information pertaining to the tournament or the specific match, information about the players, and post-match statistics for each player.

**STATISTICS**

The database contains 32425 distinct match records. Matches are distributed close to evenly across all eleven years, with a slight decrease in the number of matches the more recent the year is : from 3085 matches in 2009 down to 2781 matches in 2019.   
The database contains matches from 96 different tournaments played over the years. Most matches are from the Davis Cup (3118), followed by the four Grand Slam tournaments (1397 each) and the Indian Wells and Miami Masters (1045). The tournaments with the lowest number of matches are the ATP Next Gen Finals (16), a competition for under-21 players, and the Zhuhai and Cordoba ATP 250 tournaments (27 each). The amount of matches vary as some tournaments have been discontinued over the years, whereas others were created after 2009.  
18426 matches were played on hardcourt, 10174 on clay, 3532 on grass and 95 on carpet ( all matches played on carpet come from the Davis Cup, as the use of carpet courts in ATP tour matches was halted in 2009).  
The players with the most matches are Novak Djokovic, Rafael Nadal and Roger Federer with respectively 839, 772 and 747 matches between 2009 and 2019. We have 100 matches or more for 179 players in our database, 30 matches or more for 310 players, 10 matches or more for 496 players, and 5 matches or more for 697 players, which means there are 682 players in the database on which we have less than 5 matches.

1. **Feature selection**

A subset of features relevant to the prediction of tennis matches can be selected, given that many of the features from the original database were included for descriptive purposes and most likely do not hold any predictive power towards the winner of the match, like player name or country of origin.  
We select the following features:

**- rank / ranking points**: Regression-based studies are unanimous about ranking being highly relevant for the prediction of match winners.

-**surface**: the review of the literature has shown that there can be a significant skill gap for one player depending on the surface he plays on, especially for matches played on clay.

-**player height and age**: Although some studies found that demographic factors like age, height and weight had no significant influence on their predictions, relevant features could be constructed from these measures. In general, athletes reach peak performance at a certain age, when an optimal balance between experience and physical fitness is achieved, and performance begins declining as they age further and their bodies lose strength and speed. Height could also be advantageous, for example for players relying on a hard serve that can be hit from a better angle.

-**handedness**: Whether a player is left- or right-handed will determine from which side he hits his forehand and backhand. This might have a tactical repercussion for the opponent.

-**score**: as pointed out by McHale and Morton (2011) , one of the inaccuracies of rankings is that they consider any win to be the same, whether one player dominated the other or the match was very closely fought. The information on how dominant a player has been in the past can be extracted from the exact score lines.

We also include all post-match statistics in this first selection, as they give a more accurate description of the performance of both players during each match.

**C)Data Preparation**

For any task requiring the use of a database, especially when using third-party data, it is important to start by checking the data for errors, missing values or outliers. If these remain in the data when starting the analysis and trying to create a model, they might create a bias in the results.

We begin by removing matches based on score lines. Some of the matches in the database ara walkovers (wins by default) due to the withdrawal of one player. These give no indication of the skill level of any player. Other matches have incomplete score lines because the matches finished as a retirement of one player during the match, leading the other player to win by forfeit. As these matches might not accurately represent the difference in skill between the two players, we decide to delete all 1229 observations of unfinished matches and default wins. We also remove one erroneous entry with a score of 0-0.

Other studies indicate the fact that rankings are a central feature in the prediction of match winners in tennis. Both rankings are missing in 97 matches of the database, and one of the rankings is missing in 596 more matches. All these matches are deleted.

There are 5 players whose ages are missing, resulting in 11 matches with missing age values.

There are 5000 matches where heights are missing for one player and 1405 matches where both heights are missing.

# CLEANSING: ERRORS

Before any data science task, it is important to make sure that the available data is void of errors. Especially when using large datasets provided by a third party, there can be errors in the collection of the data that the end user has no control over. Therefore, a careful cleansing is of the utmost importance. If they are not removed, errors in the form of missing values or outliers might bias the results when the data is fitted to a statistical model and invalidate the analysis.

## BASED ON SCORE

In order to cleanse our dataset, we begin by using the “score” feature. Some of the matches are labeled walkovers (wins by defaut) due to the withdrawal of one player before the match, some have a partial scoreline followed by the word RET signifying the retirement of a player during the game. In the case of a walkover, the fact that one player won is entirely independent of the difference in skill between both players; in case of abandonment or retirement, one of the players might have been impaired by an injury or other circumstances, and the fact that the match was ended prematurely means that we cannot compare the match statistics to those of other full matches. For these reasons, we decided to delete all these matches from our database. We remove 1031 matches where one of the players retired, 192 walkovers, 9 matches where players abandoned, 1 erroneous entry with a score of 0-0 and 1 where score is a date, for a total of 1234 matches.

## BASED ON RANK

The literature shows that rankings and ranking difference are the most important predictor of the match winner. Therefore, we decide to delete all matches where one of the player’s ranking is missing. We delete 95 matches with both rankings missing and 281+ 313 where either player’s ranking is missing for a total of 689 matches.

## BASED ON SURFACE

Only 150 matches in the database have a missing “surface” feature, with 145 matches being from the 2014 Davis Cup draw and 5 more from the 2017 Davis Cup draw. We obtain results and surface from the Wikipedia page and fill them in manually, in order to have the surface feature for all of our data.

## BASED ON SERVICE POINTS

The minimum amount of service games for either player in a best-of-three tennis match is 24, corresponding to a match score of 6-0 6-0, where the winning player wins every score on his own serve. We delete 6 matches where either player has less than 24 service points.

# CLEANSING: MISSING VALUES

Of the remaining matches, 1766 have no match stats. These matches are left in the databases but we mark the empty cells as being non available to make sure they are not interpreted as a 0 by our statistical analysis program, which could lead to wrong calculations when we take the historical

Deleted a few more matches with abandons, 7-8 with impossible scores (either player less than 24 serve points, which is the minimum for a 6-0 6-0 score

**D)Feature creation**

In order for a model to be used for classification, there must be a binominal variable that is to be predicted, in this case representing that a match is won or lost. In Jeff Sackmann’s database, the features have already been assigned specifically to the winner and loser of the match. We generate a new feature, “Player1\_wins”, that randomly assigns a 1 or 0 to every match. If the number assigned is 1, the winner of that match becomes Player 1, and the loser becomes Player 2: if the assigned number is 0, the opposite happens. All features specific to the match winner and loser are renamed accordingly. “Player1\_wins” is the feature that our model will need to predict, also called the *label*.

Wp\_surface, wp\_hand, differences in wp\_surface / wp\_hand, servewinpct, returnwinpct, win percentage difference vs common opponents, head-to head win percentage? Possible to have features where some players have no values bcos not enough dat/ no data?

**E)Feature Scaling**

**III. MODELING**

**A)Model assessment method**

**MODEL BASELINE AND THEORETICAL MAXIMUM- MODELING METHODOLOGY**

The central measure for the assessment of the performance of a classification model is accuracy. Accuracy is calculated as the percentage of correct predictions made by the model on the data it is fed. However, accuracy alone is not a reliable measure to rate the usefulness of a predictive model. For instance, when using a predictive model to classify data points into two categories with one of the categories being much more prevalent than the other, absolute accuracy can be misleading: if 98% of the data points belong to class 1, and 2% to class 2, a model that would predict “class 1” every time would attain 98% accuracy. However, it is clear that such a model is useless as it brings no added value to human prediction and has not “learned” significantly from the data, as the model amounts to a simple rule. Conversely, when trying to predict the winner of a tennis match, a model attaining 98% accuracy would be an astonishingly useful model, smashing all other tennis models ever created, and it would be as profitable on the betting market as it is impossible to create.

This illustrates the fact that accuracy is only a good indicator of performance when it is *relative accuracy*. In order to know whether a certain prediction accuracy is satisfactory, it must be compared to a certain baseline. One way to tell whether a model is valuable is to compare its accuracy to the accuracy of a human (average human?) on the same predictive task. We choose this approach by supposing that any human could predict tennis matches simply by supposing that the player with the better ranking is the one that is going to win. The baseline to which the models of this study will be compared to is therefore the accuracy of the rule-based prediction :

IF RANK(Player 1) < RANK(Player 2), Player1\_wins = 1. Else Player1\_wins =0.

Using this rule to predict all the winners of the matches in our cleansed dataset yields an accuracy of 67,29%.

**THEORETICAL MAXIMUM**

**STEP-BY-STEP APPROACH**

The starting point in the creation of the predictive model is a simple model using rank points as only feature. Rank points are preferred to rankings, as they conserve the same (though inverted) ordinal relationship between one another, but rank points offer an additional differentiation as the difference in rank points between two players ranked one after the other can be variable. Rank alone gives a constant measure of the skill gap between two players, which is less representative of reality, whereas the difference in rank points between two pairs of players having the same difference in rankings can be variable.

Other features are added one by one in order to isolate their effect on accuracy. Based on the results, features are kept or discarded in the process. Intermediate model accuracies are listed in the table in B).

**B) Results**

**-**After making our own point based model we can try to include serve win probz as well as derived match win probz using the formula’s into our model and evaluate results (when making a feature prob need to include best out of how many sets match )

-Could we try to make a “mental” feature with break points saved, maybe win percentage in crucial points, # of double faults (proxy for focus but might be reflection of aggressive serving strategy as well)  
Could be: difference in proportion of points won overall and break points saved vs break points faced, proxy for how well player performs under pressure

-look at other studies making modified elo ranks/ network ranks and try to create a new rank feature based on their study:

ST

**ELO NOTES**

We scrape Jeff Sackmann’s website and obtain

**IV. DEPLOYMENT**

-Betting market: the advantage of point-based model, be it less accurate for classification, is that it provides stronger rule for betting as we can decide whether or not to bet given the computed odds and the offered odds. With classification we could bet every match but don’t know the odds the model gives, how could we circumvent this? Clustering? Some test to see if class assigned to a match varies when we change certain things (sensivity analysis)?

# Bibliography

Barnett, T., & Clarke, S. (2005). Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, *16*, 113–120. <https://doi.org/10.1093/imaman/dpi001>

Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors?: An evaluation. *International Journal of Forecasting*, *15*(1), 83–91. <https://doi.org/10.1016/S0169-2070(98)00067-3>

Carter, W. H., & Crews, S. L. (1974). An Analysis of the Game of Tennis. *The American Statistician*, *28*(4), 130–134. JSTOR. <https://doi.org/10.2307/2683337>

Clarke, S., & Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research, 7*, 585-594.

del Corral, J., & Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *Sports Forecasting*, *26*(3), 551–563. <https://doi.org/10.1016/j.ijforecast.2009.12.006>

Gilsdorf, K. F., & Sukhatme, V. A. (2008). Testing Rosen’s Sequential Elimination Tournament Model: Incentives and Player Performance in Professional Tennis. *Journal of Sports Economics*, *9*(3), 287–303. <https://doi.org/10.1177/1527002507306790>

Hsi, B., & Burych, D. (1971). Games of Two Players. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* *20*(1), 86-92. doi:10.2307/2346634

Klaassen, F., & Magnus, J. (2001). Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model. *Journal of the American Statistical Association*, *96*(454), 500–509. JSTOR.

Klaassen, F., & Magnus, J. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, *148*, 257–267. <https://doi.org/10.1016/S0377-2217(02)00682-3>

Knottenbelt, W., Spanias, D., & Madurska, A. (2012). A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications*, *64*, 3820–3827. <https://doi.org/10.1016/j.camwa.2012.03.005>

Kovalchik, Stephanie Ann. "Searching for the GOAT of tennis win prediction" *Journal of Quantitative Analysis in Sports*, vol. 12, no. 3, 2016, pp. 127-138. <https://doi.org/10.1515/jqas-2015-0059>

Lisi, F., & Zanella, G. (2013). Tennis betting: Can statistics beat bookmakers? *Electronic Journal of Applied Statistical Analysis*, *00*, 1–35. <https://doi.org/10.1285/i20705948v10n3p790>

Newton, P., & Keller, J. (2005). Probability of Winning at Tennis I. Theory and Data. *Studies in Applied Mathematics*, *114*, 241–269. <https://doi.org/10.1111/j.0022-2526.2005.01547.x>

Pollard, G. (2008). An analysis of classical and tie-breaker tennis. *Australian Journal of Statistics*, *25*, 496–505. <https://doi.org/10.1111/j.1467-842X.1983.tb01222.x>

Scheibehenne, B.,&Broder, A. (2007). *Predicting Wimbledon 2005 tennis results by mere player name recognition.* International Journal of Forecasting, 23, 415–426.

Spanias, D., & Knottenbelt, W. J. (2013). Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*, *24*(3), 311–320. <https://doi.org/10.1093/imaman/dps010>

Srivastava, S. (2019). *Predicting Success Probability in Professional Tennis Tournaments Using a Logistic Regression Model* (pp. 59–65). <https://doi.org/10.1007/978-981-13-1208-3_6>

1. https://www.atptour.com/en/stats/leaderboard [↑](#footnote-ref-1)
2. https://www.sbnation.com/2012/3/4/2843710/sloan-sports-analytics-conference-day-2 [↑](#footnote-ref-2)
3. https://www.youtube.com/watch?v=QPyVJbKjmaw [↑](#footnote-ref-3)
4. https://www.sloansportsconference.com/people/jeff-sackmann?mode=grid [↑](#footnote-ref-4)
5. https://news.sap.com/2015/08/wta-sap-team-give-tennis-analytical-edge-real-time-performance-data/ [↑](#footnote-ref-5)
6. https://www.bloomberg.com/news/articles/2015-12-07/tennis-gamblers-demand-for-real-time-betting-nets-itf-data-deal [↑](#footnote-ref-6)