WARWICK
THE UNIVERSITY OF WARWICK

PREDICTIVE ANALYSIS IN TENNIS FOR SPORTS JOURNALISM

CS350 FINAL REPORT

Author: Danielle O'Sullivan
Student ID: 1402161
Course: Data Science G302
Year of Study: 3

Supervisor: Dr. Graham Cormode

ABSTRACT

With the growth of sports statistics, sports betting and sport engagement, the demand for real-time analytics reaches ever higher. Predicting the outcome (winners as well as losers) is a preoccupation of every sport, and the problem of forecasting the results of matches, has had fans fascinated and perplexed for almost as long as tennis has been a competitive sport. Conceptual ideas around statistics in the sports industry will be explored thoroughly within this paper as well as addressing the drawbacks surrounding predictive analysis industry wide. This document also demonstrates how tennis match results (recorded by ATP) can be used to train a predictive statistical model using the attributes provided in the data in decision trees. The intention is that the algorithm learns as each match progresses, taking in more information as it becomes available. Alongside a discussion on the results of the model training, any legal, social and ethical considerations within the (digital) sports journalism industry are reflected on before potential future work for the predictive tool is reviewed.

Keywords: Big Data, open source, dataverse, tennis, Markov, Bernoulli, decision tree, digital journalism.

CONTENTS

INTRODUCTION

Data analytics is coming to every sport in some capacity, though at varying speeds. Baseball, basketball, American football, and hockey have all made great strides in the last few years. As has the English Premier League (although with Association football, the "match analysis" at the end of the match largely focusses on *what* went wrong, rather than *why* it went wrong in the first place). Tennis seems to be struggling to keep up analytically, as most of the rich data that has been collected, is not being utilised in the most efficient way.

"A player's true strength and weaknesses only become evident during matchplay. Research show that players do not always do what they think they are doing in competition, and that coaches are less than 45% correct in their post-match assessment as to why a match is won or lost."[1]

There are some wider problems that the tennis analytics world faces however, for example:

- **Tournaments**: Tennis is a fragmented, decentralized landscape. Some of the main associations hold some sort of power over tournaments, but most events are very autonomous, and it's difficult to align a uniform data tracking system, especially when directors are most interested in maximizing their sponsorship deals for their players.
- **Players**: Individual sports are not as incentivized to gather analytics, as there's no team-constructing component to the game. There's no need to isolate individual contributions in tennis, since we can already tell just how much a player contributed to his or her success.
- **Sport**: Given the lack of parity in tennis as a whole, the public demand for analytics may not be too great, as even the most comprehensive data will be unlikely to turn the likes of John Isner or Jo-Wilfried Tsonga into a regular title contender.

So there's a huge gap to fill in freely accessible tennis data, which has the power of some of the richest data sources out there. This project has focussed on exploring prediction which can be used to decode the patterns within the raw data which would be exciting not only for the tennis players themselves, but their entire coaching team and, of course, the fans.

This project uses decision trees as a simple way of trying to predict the outcome of any given tennis match using set by set results as well as world ranking, court surface and match pressure (i.e. whether the match is a first round match or a final). This documentation showcases the main motivations behind this project; an in-depth discussion on the predictive algorithm; as well as legal, social & ethical considerations of potential use of the final tool; and any future work that is relevant to the project and industry.

[1] http://www.tennisanalytics.net/

BACKGROUND MATERIAL

Big Data is not just the phrase *du jour*, it's making an impact on modern day life in many diverse ways - some quite unexpected. Out of all the data that has been collected throughout history, around 90% of all of it was generated in the most recent decade.[2] Data just keeps on building: EMC predicts that by 2020, the practice of data generation will have reached staggering heights of 40 trillion gigabytes collected per year (which is roughly 1.7MB of new information per person for every second of every day).[3] This means that every two years, the "digital universe" will approximately double in size.

It makes sense when you think about search engines like Google and Yahoo. For every second on Google, over 40,000 search queries are performed - this accumulates to 3.5 billion search terms per day and 1.2 trillion searches in a year worldwide.[4] To further put this *dataverse* into perspective, social media collects hundreds thousands of gigabytes of data per minute, Facebook alone reporting its users send an average of 31.25 million messages and view 2.77 million videos every minute, according to statistics released by Facebook in March 2015.[5]

When all these astronomical figures are floating about, it can be hard to grasp what it means to truly harness data and to understand the power of data interpretation. The majority of the general public do not possess the mathematical awareness in order to convert the numbers to meaning in a short space of time, reflected in the fact that most retailers are missing out on managing to adjust their sales and marketing strategies to their fullest potentials - these companies could see their profit margins shoot up by as much as 60%.[6] Of course, more and more companies are becoming aware of this power of statistical analysis and around three quarters of all organisations worldwide had invested in some sort of Big Data plan for 2016 and beyond.[7]

With the ever increasing demand for statistical analysis, it makes sense that the sporting industry took a hold of the power of analytical science to improve performance. It all starts with data collection, and in today's world, there is more data being collected than being utilised. Just over a year ago, it was estimated that 0.5% of all data ever collected was actually analysed and used.[8] From geo-tagging to loyalty cards, almost every shop or company you come into contact with has collated some sort of data profile of you. This has sparked debates on whether it's a breach of privacy to store such large amounts of data for professional sports players, however, the fact that so much daily life has the potential to contribute to success is somewhat of a 'hidden weapon'.

---

[2] http://www.sintef.no/en/latest-news/big-data--for-better-or-worse/
[3] https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
[4] http://www.internetlivestats.com/google-search-statistics/
[5] http://www.cio.com/article/2915592/social-media/7-staggering-social-media-use-by-the-minute-stats.html#slide2
[6] http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/getting-big-impact-from-big-data
[7] http://www.gartner.com/newsroom/id/2848718
[8] https://www.technologyreview.com/s/514346/the-data-made-me-do-it/

If you were to look into the preparation that goes into some of the major leagues, you wouldn't be far off if you were to say that for every professional athlete, there is a statistical expert - this area has expanded so much because of the undeniable benefits. How can you become faster, stronger and better? The answer lies within the capability of data comprehension. Who benefits the most from this capability? The athletes, the teams, the coaches, the sponsors, and of course, the spectators who get to witness groundbreaking feats of athleticism which have been refined with the aid of statistics and drive.

Any athlete who doesn't have an analytics expert on hand is at a surprisingly high disadvantage to their competitors. For up and coming sports stars, however, this understanding can be difficult to grasp unless you have the right academic training, or more specifically, unless you have the money to buy the people who are in the know. Otherwise you are fighting an increasingly steep uphill battle.

For athletes, the relevant data that could be collected ranges from in-play statistics to training methods to nutrition. All this data is 'open' to the individual athlete, which means there's more consumable content for the fans to follow their favourite athletes. There is a huge amount of online resources dedicated to the research and analysis of any data collected within certain sports. In North America, Major League Baseball (MLB), the National Basketball Association (NBA) and the National Football League (NFL) are some of the best examples of data driven decision making.

In MLB, one of the more widely used approaches to select players for a team was analytical (i.e. who can run to get *on base* the fastest) in that the performance of the players was measured and proven. Much of the data used in many of the professional leagues is actually available to view online, but the details given at no cost don't necessarily give that much of an extra insight into player performance, since well-known betting companies will often buy the rights to an entire statistical website which provide the following information to bookmakers; so keeping this all out of the public eye is, of course, what makes betting sites successful:

- **A structured stream of events** (e.g. All European Football matches with names of the relevant players). Bookmakers consume this into their databases.
- **Prices for markets**. Often this is based on what major bookmakers are offering on a particular market, so their customers can see if their own prices are competitive.
- **And outcomes**. Which are the lifelines for bookmakers as it allows them to settle bets (and make money).

For much of the MLB, baseball enthusiasts and analysts have made use of the data given in order to construct their own views for matches on a play-by-play basis. For example, many intricate hitting, pitching and fielding metrics are used to determine whether a player is fit for the lineup. These metrics have been initially documented by the fans (many of whom are professional

analysts, and with their skills, they manage to get hired by MLB teams to provide their insights directly).[9] This is one of many motivations as to why the application of statistical analysis in sports is quite high-paying - those who possess this understanding are in very short supply, when the demand is extremely high. This is important to note because of the intensity at which the rate of the *dataverse* is expanding; the more people who can fill these roles, the better, and the more immersed the fans can be within live sports coverage.

Another example is more to do with larger, and more interactive, team sports - leagues such as the NFL and Barclays Premier League. It has been noted that the game performance analytics tend to be less sophisticated than baseball[10], however this is mainly due to the very nature of the sports. Both American Football and (European) Football, or Soccer, use highly integrated teamwork unlike baseball which is more individual-focussed. For baseball, the main *team* tactics that have to be decided are the physical lineup for batsmen and the positioning of certain players for fielding - although, then again, this naturally doesn't have the highest impact on overall performance as the interaction between players is minimal.

For footballers (both types), this is a much more complex field due to the sheer number of permutations of interactions the players can have between each other as well as the types of play each athlete can viably choose. A simple example could be: if a footballer were to be playing from a defensive into an offensive position, he has a number of different plays he could choose from:

1. **Pass to Player 2 on their left** - the ball will unlikely reach this player without being intercepted, or with too poor aim, the opposing team could win the ball. There is a low probability that this move is the 'best' in this oversimplified scenario.
2. **Pass to Player 3 on their right** - the ball will likely reach, but there is a strong threat of a tackle taking place, or at least an attempt to block a potential shot at goal, by Opponent 2.
3. **Pass behind to Player 4** - the ball will most likely reach this player with a low chance of interception and poor aim, however, this does not advance the ball *upfield* so Player 1 may decide this is not the right play for the situation.
4. **Shoot for goal** - this option again has a low probability of success, although depending on the player, if Player 1 has been trained as a *Striker*, in theory they will have more

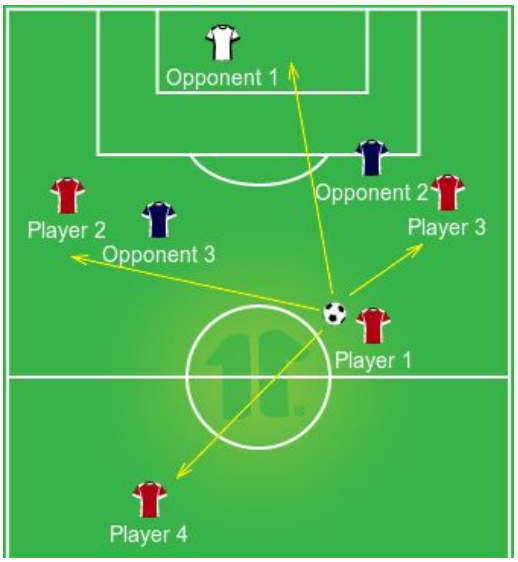

Figure 1: Simple tactical view for individual football players

[9] https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/iia-analytics-in-sports-106993.pdf
[10] https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/iia-analytics-in-sports-106993.pdf

chance of succeeding in scoring than someone who has been trained as a *Midfielder* or *Defender*. Again, many factors to take into account, even here.

Even taking into account the simplicity of this example, *Figure 1* gives a bit more depth into why football and team sports are much harder to model when looking for predictive capability because you have to notice that every individual will choose differently based on their opponents and personal position. Running through the 4 possible moves Player 1 could choose once more, it is apparent that all four options have a low probability of success (where success is taken to mean the play contributes to a 'better' opportunity to score a goal), so at this stage in the world of technological advancements, the kinds of software available are not quite up to speed in order to predict outcomes to a high technical accuracy.

On the other hand, regardless of publicly available data, industry wide gambling takes into account an extraordinary amount of factors to update their odds during live games. From player profiles that include injury likelihood, on-pitch and off-pitch behaviour as well as the more general (relative) average sprint speed and height and weight statistics, to more complicated probability distributions for entire teams to predict performance outcomes. When building a model, *bookmakers* have some of the sharpest minds in the industry working for them to create intricate algorithms to produce their betting odds, so naturally, it won't be easy to beat them - although it's not entirely impossible. Essentially, a betting model is used to assess potential in teams or individuals, but in the sense that a statistical model can show you insights that an enthusiast's gut feeling can't predict. As these models are data-driven, which at the very least means using final scores for testing the models, but ideally it means in-depth stats that you can breakdown and incorporate into an algorithm. The most important aspect of understanding betting models is using historical odds for which to test on; so a lot of research has to go into obtaining the data required, as well as in the format needed.

Extrapolating logically from this observation, the choice of sports to delve deeper into became more apparent: in order to keep the project challenging whilst also keeping in mind the time constraints, it would be sensible to avoid highly 'interactive' team sports such as football, basketball or hockey. This narrows it down to team sports with lower interactivity (e.g. baseball, cricket) and individual sports. The reason why such a sport as baseball was not chosen for this particular project was due to the vast number of resources and research that has already gone into the development of the sport; there was simply no real conceivable 'gap' to fill for this particular sport, as sophisticated as it is. Similar reasoning goes for cricket, in that most of the metrics collected are for the batsmen, and as the batter cannot *pass* the ball to a teammate, this makes it quite closely related to the problems of baseball. Searching through the many individual sports played to a professional level, there are a surprisingly few number that collect substantial and rich enough data.

Take Taekwon-Do as an example, this is one of the few sports that directly integrates wearable technology as point scoring systems, and yet, the available data out there for the public is rather

lacking[11], with very sparse entries and with next to no explanation of how the data was collected or how the point systems worked - which leaves too much space to move, how exactly should or *could* this data be interpreted? And is there any benefit to the sport by crunching these numbers? By asking these questions the sport that looked the most promising was tennis. The cornerstones was the amount of rich data that the official organising bodies had accumulated over the years; organisations such as the Association of Tennis Professionals (ATP), the Women's Tennis Association (WTA), the International Tennis Federation (ITF) and the Lawn Tennis Association (LTA) all conduct massive amounts of data collection during each of their respective tournaments.

ATP have a particularly interesting interactive comparison tool on their website *Head2Head*[12] which allows users to compare and contrast their favourite professional tennis players by showing a range of basic stats that allows the user to loosely make their own mind up on who is more likely to win, however it's not difficult to see that this really doesn't add much more in the decision making process, especially if the user already knows many of the listed statistics. The WTA version for player comparison, *Head to Head*[13], shows fewer details side by side as well as in an order that is not consistent with its male counterpart - which makes the scraping of such data from each association site that much more of a headache due to the nature of preprocessing. So it becomes clear that one aim is to explore the different avenues for open source data and to create a tool that is able to distribute *meaningful* content, which means collating data from all the different associations into a consistent and clean way.

Journalists in particular take notice that almost as vital as having access to enough data, "is being able to present the information to coaches, managers and club owners, as well as players, in a format they can work with" because "simply having more information does not always lead to better decision making."[14] Which is why the next layer of data driven sports science lies in excellent User Experience (UX) - if a user can understand the content intuitively, then the user interface (UI) has succeeded.

A recent open source project aims to tackle this consistency problem in noting down individual tennis match statistics manually. So-called *Match Charting*[15] was created by Jeff Sackmann in 2013 which records "shot-by-shot data for every point of a match, including the type of shot, direction of shot, depth of returns, types of errors, and more." The project, which has been running for over 3 years now, has charted almost 3,000 individual tennis matches which can be broken down into the following attributes:

- Match ID
- Round

[11] An example being this website: http://www.taekwondodata.com
[12] http://www.atpworldtour.com/en/players/fedex-head-2-head/novak-djokovic-vs-andy-murray/D643/MC10
[13] http://www.wtatennis.com/headtohead
[14] https://www.theguardian.com/sport/2015/jan/22/marginal-gains-the-rise-of-data-analytics-in-sport
[15] http://www.tennisabstract.com/charting/meta.html

- Player 1
- Player 2
- Player 1 Dominant Hand
- Player 2 Dominant Hand
- Gender
- Date
- Tournament
- Time
- Court
- Surface
- Umpire
- Best of
- Final Set Decided By Tiebreak?
- Charted by

Match charting is only the start of a bigger and more intrinsic change to the sporting world, the next step would be to enable real-time charting along with insights to enhance the general public's engagement in sport. It's no surprise that an individual can get more excited about entertainment if there's more to the eye than just the game. IBM, one of the largest data-centered companies in the world states that "there is an increase in demand for real-time insight in numerous organizations" so in maintaining a "current, real-time view of what is happening… [via] key performance indicators, [such as] highlighting a drop off in performance, or triggering an alert if something requires action. In all cases, the user knows something needs to be done now, rather than finding out after the event."[16]

In terms of real-time analytics, tennis lacks statistical engagement from its fans. One of the most comprehensive data collectives out there, "Hawk-Eye, which tracks the speed, trajectory, and spin rate of every shot, is not currently publicly available, and fans and reporters are usually subject to … largely useless data such as break point percentage, aces, and winner totals."[17]

The demand for publicly available data is at a breaking point, fans will surely get tired of seeing the same plain statistics; and journalists, who often have more insights than the general public, but are still obliged to translate data for consumption by a general public, are consistently supplied by official sources in physical paper copies which they have to then manually transcribe into their computers before they can even begin to make sense of what they've been given.

> " During the U.S. Open, tennis is played simultaneously on up to 19 courts, resulting in more than two weeks' worth of action. That action broadcasts around the world, but it's also captured by digital cameras and analyzed by computers in near real-time - to help players dispute bad calls and help fans see every point, game and set score, as well as the speed of a player's serve. "[18]

The future of Hawk-Eye (and other cutting-edge technology) data can only be brighter, as the introduction of 360-degree instant replay systems to the Grand Slam tournaments allowed

[16] http://www.ibmbigdatahub.com/blog/wimbledon-using-real-time-sports-statistics-fan-engagement
[17] https://www.forbes.com/sites/jimpagels/2015/03/03/why-is-tennis-so-far-behind-other-sports-in-data-analytics/#1e603a615e1d
[18] https://iq.intel.com/pro-tennis-serves-fans-riveting-real-time-data/

spectators viewing on court and at home to be involved that little bit deeper. This type of immersion can only stay on the top of its game if much of the official associations allow their data to be publicly available, as Nicole Jeter-West (former managing director of digital strategy for the U.S. Tennis Association) points out that "the tournament is only relevant in the eyes of tennis fans for two weeks [and] it's critical fans are given reliable and meaningful insight to make the most of their experience."[19]

In keeping with this view, it makes sense that an interactive tool is made for the fans so that they can continually seek out which players have the better odds of succeeding and reaching further into the competition. IBM enabled fans to interact with the tournament via social media at precisely the right times using the IBM Watson to compute predictive analytics about the tournament in order to "feed fans the right social media at the right time."[20] This is just one of many highly sophisticated approaches to keeping the data alive.

So where's the gap? Due to the "low profile" of the public statistics surrounding tennis. If it can be revolutionised in such a way that it becomes a highly exciting and sought-after topic of research, it could hugely impact the game itself. For example, freely available resources that could track how many types of strokes an individual played that resulted in a loss for that game could create flexible, targeted and personalised training - although the exact way this would happen remains to be seen.

The ATP & WTA low-level comparison tools have been used as a starting point to the project, in order to predict who will win out of a given pair of players, based on past performance and individual attributes (dominant hand for example). For the project to move forward, existing data collated from ATP was used to train a statistical model using set results to track time and progression through each individual match. The code looks at which time step it is at (i.e. which set is 'being played'), then looks at past performance from the previous sets (if any), and uses this information combined with initial attributes, such as world ranking of both players, to determine who is more likely to win overall. This would then be implemented into a 'Head to Head' web application allowing users to compare athletes. Eventually, the model used for prediction would be trained a further step by looking at how each point was played - which is where Match Charting would really take off.

METHODOLOGY & RESULTS

Data Collection

When searching for data, there were certain criteria each dataset had to comply with - the richness of the data had to be to a sufficient level, there had to be enough data to sufficiently train the resulting predictive model and ideally (although not required), it would be easier for the different versions (i.e. data from different years) of datasets to be standardised.

[19] https://iq.intel.com/pro-tennis-serves-fans-riveting-real-time-data/
[20] https://iq.intel.com/pro-tennis-serves-fans-riveting-real-time-data/

There were a few resources that were looked at and subsequently reviewed:

1. **UCI Machine Learning Repository**[21]: A dataset that contained 8 files containing match statistics for both women and men at the four major Grand Slams of the year 2013. Each file had 42 columns and a minimum of 76 rows. This dataset was eventually rejected because there was only the data from 2013 and no other years were available in the same format.
2. **ATP World Tour Official Website**[22]**:** There was an existing project sourced on Github [23] which used Python scripts to scrape data from the official website. The scripts had the ability to scrape data by player or by year. By player, the scraped data listed 99 headers (columns) for the resulting CSV files, and depending on which player was chosen, the number of columns depended on both career length and consistency in tournament performance (e.g. Roger Federer has consistently ranked highly in most of the tournaments he has participated in, as well as having a long career on top of that). By year, the scraped data listed 42 headers and a variable number of rows (depending on the year). In terms of rich data this resource comes out on top, however, due to the runtime of the scraping and the size of the resulting files, it was a long stretch to generate these files whilst also building the appropriate model (that was sufficiently trained to predict) that would use specific columns *(refer to 'Future Work' for more insights)*.
3. **Tennis-Data.co.uk**[24]**:** This site offered a 'quick fix' in terms of data collection. This contained historical records from ATP tennis tournaments between the years 2000-2017. The data was initially in an Excel spreadsheet format, but later converted into CSV file format for compatibility with the prediction program. It has a total of 42 columns and a varying number of rows dependent on the year.

Predictive Model

There were many different methods considered that could have been pursued in the building of the predictive algorithm, the main approaches researched were: Markov chains, Bernoulli distributions and Decision trees.

1. **Markov Chain**

   The definition of a Markov chain is, according to *Wolfram MathWorld*[25], "a collection of random variables $\{X_t\}$ (where the index t runs through 0, 1, ...) having the property that, given the present, the future is conditionally independent of the past. In other words,

$$P(X = j \mid X_0 = i_0,\ X_1 = i_1, ..., X_{t-1} = i_{t-1}) = P(X_t = j \mid X_{t-1} = i_{t-1})$$

   If a Markov sequence of random variates X_n take the discrete values $a_1, ..., a_N$, then

[21] https://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics
[22] http://www.atpworldtour.com/en/tournaments
[23] https://datahub.io/dataset/atp-wta-professional-tennis-tournament-data
[24] http://www.tennis-data.co.uk/alldata.php
[25] http://mathworld.wolfram.com/MarkovChain.html

$$P(x_n = a_{i_n} \mid x_{n-1} = a_{i_{n-1}}, ..., x_1 = a_{i_1}) = P(x_n = a_{i_n} \mid x_{n-1} = a_{i_{n-1}})$$

and the sequence x_n is called a Markov chain."

Tristan J. Barnett, author of *Using Microsoft Excel to Model a Tennis Match (2002)*[26], derived a simple model which is set up as follows:

"[We] have two players, A and B. Player A has a constant probability p of winning a point. We set up a Markov chain model of a game where the state of the game is the current game score in points (thus 40-30 is 3-2). With probability p the state changes from (a,b) to $(a+1,b)$, and with probability $1-p$ it changes from (a,b) to $(a,b+1)$. Thus if $P(a,b)$ is the probability that player A wins when the score is (a,b), we have:

$$P(a,b) = p \cdot P(a+1,\ b) + (1-p) \cdot P(a,\ b+1)"$$

Barnett then goes on to discuss more realistic aspects of the game, such as deuces and tiebreaks. As concluded in Agnieszka M. Madurska's Analysis Project[27], the final probabilities for a particular player winning a set are shown as follows:

"Assuming that Player A serves first in the set then the probabilities of winning a set from set score (a,b), $P_{set}(x,\ y)$ are:

$P_{set}(a,b) = p_A^{game} P_{set}(a+1,b) + (1-p_A^{game}) P_{set}(a,b+1)$, for even $(a+b)$

$P_{set}(a,b) = p_B^{game} P_{set}(a+1,b) + (1-p_B^{game}) P_{set}(a,b+1)$, for odd $(a+b)$

Where:
p_A^{game} is the probability of Player A winning a game from score (0,0) while serving.
p_B^{game} is the probability of Player B winning a game from score (0,0) while serving."

This approach goes into such a high level of detail, that it would be near impossible to create a whole new algorithm without infringing on past authors' work, and would leave little room for designing a front-end, which was what was initially planned for, with all the (foreseen) time that a new implementation would take.

2. Bernoulli Distribution

A Bernoulli distribution is defined as follows, according to *Wolfram Mathworld*:

[26] http://strategicgames.com.au/excel.pdf
[27] https://www.doc.ic.ac.uk/teaching/distinguished-projects/2012/a.madurska%20.pdf

"The Bernoulli distribution is a discrete distribution having two possible outcomes labelled by $n = 0$ and $n = 1$ in which $n = 1$ ('success') occurs with probability p and $n = 0$ ('failure') occurs with probability $q = 1 - p$, where $0 < p < 1$. It therefore has probability density function: $P(0) = 1 - p$ and $P(1) = p$, which can also be written: $P(n) = p^n(1-p)^{1-n}$."[28]

In the paper *Computers & Mathematics with Applications*[29], the authors discuss another approach to predicting the outcome of a tennis match using the O'Malley closed-form equations. O'Malley derives some equations for the probabilities of a certain player winning a match, set, game and tiebreaker assuming that winning any point in play is a Bernoulli random variable.

According to these probabilities, "a fundamental insight arising from O'Malley's equations is that the probability of winning a match is mainly dependent on the difference of the probabilities of players winning a point while serving", for example, let there be 3 individual players who each have a fixed probability of winning a point on their serve against another opponent. If Player 1 has a probability of 0.7, Player 2 had a probability of 0.2 and Player 3 has a probability of 0.6 - then Player 1 is much more likely to win the match against Player 2 than against Player 3 since the difference in probability between Player 1 and Player 2 is 0.5 which is larger than the difference in probability between Player 1 and Player 3 which is 0.1[30]. This is actually just a mathematical way of describing common sense when observing a match.

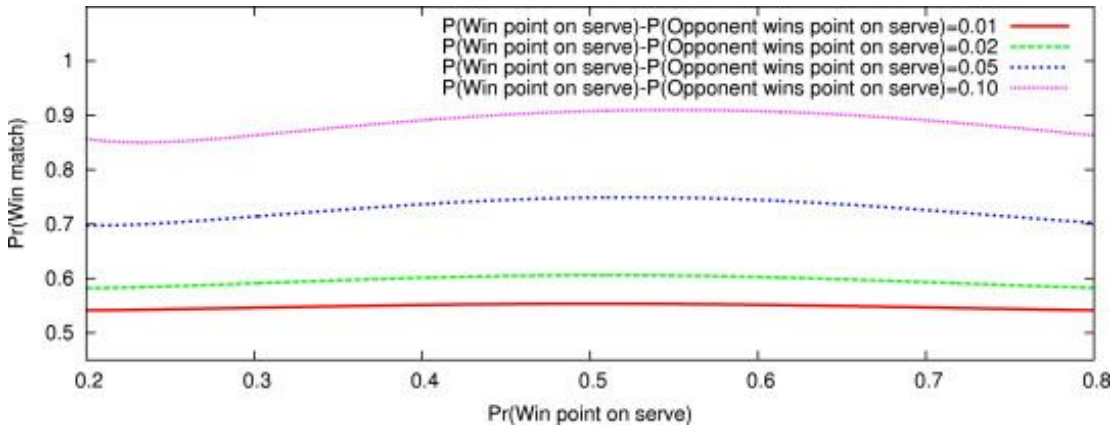

Figure 2: Probability of the better player winning a best-of-three-sets tennis match with fixed differences of 0.01, 0.02, 0.05 and 0.10 in the two players' probability of winning a point on serve.

[28] http://mathworld.wolfram.com/BernoulliDistribution.html

[29] http://ac.els-cdn.com/S0898122112002106/1-s2.0-S0898122112002106-main.pdf?_tid=c9e8f8fc-2c7e-11e7-99bc-00000aab0f26&acdnat=1493431052_8fe08b68b5bcbd93de3df8206d86e789

[30] Another way of looking at this is Player 1 has probabilities of winning a point of 0.7 when serving, (1-0.2)=0.8 whilst Player 2 is serving, and (1-0.6)=0.4 whilst Player 3 is serving.

Figure 2 depicts a graph[31] that demonstrates this observation using likely probabilities that would be encountered in a professional tennis tournament. Again, this approach was interesting and highly detailed; to the point where even the explanation can only be best described in the original form the authors used. The main reason why this method was not explored further was due to the time complexity the implementation would likely have taken. In terms of resources available, it may have taken several hours to train (as well as correctly build) these statistical models in Python, the language of choice.

3. Decision Tree

The final approach considered for implementation was the relatively simple approach of using decision trees. Decision trees are used commonly in the machine learning field for classification; informally, a decision tree is a "disjunction of conjunctions"[32], which means it is a set of rules which input is taken through and a particular value is output depending on which constraints were applied to the input.

More formally, a decision trees "classify instances by sorting them down the tree from the root to some leaf node, which provides classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute … This process is then repeated for the subtree rooted at the new node."[33]

Decision trees have many advantages, such as: being easy to interpret, not needing the data to be preprocessed to a great extent; the cost of using the tree to predict is logarithmic in the number of data points used to train the tree; and they are also able to handle both numerical and categorical data.[34] Some of the disadvantages include the risk of overfitting (i.e. where a predictive model 'learns by heart' the outcomes of the test data, rather than generalising to a wider population of instances) and the potential for bias if one or two attributes are heavily weighted (the split may become uneven due to 'unbalanced' data). The potential disadvantages were monitored through various means, such as 'pruning' the tree before it grew too large and complex (avoiding overfitting), and ensuring the data was balanced (there were no missing observations) before fitting it to the tree. *Figure 3*[35] shows a simple decision tree visualisation.

---

31 http://ac.els-cdn.com/S0898122112002106/1-s2.0-S0898122112002106-main.pdf?_tid=c9e8f8fc-2c7e-11e7-99bc-00000aab0f26&acdnat=1493431052_8fe08b68b5bcbd93de3df8206d86e789

32 http://www2.warwick.ac.uk/fac/sci/dcs/teaching/material/cs342/cs342-lec5.pdf

33 http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf

34 http://scikit-learn.org/stable/modules/tree.html

35 http://www.saedsayad.com/decision_tree.htm

Figure 3: A simple decision tree based on whether one should practice based on the weather.

Implementation

It was decided that there would be a series of decision trees that would be called depending on the number of sets played. This was to ensure the simplicity of the trees so as to avoid overfitting. The trees follow a simple set of rules, as shown in *Figure 4*, in order to arrive at an outcome.

The language predominantly used was Python because of the module `Scikit-Learn` which has an inbuilt `DecisionTreeClassifier` function, providing the training data is in the correct format. To begin with, however, the unprocessed data was converted into a `Pandas DataFrame` and run through a series of `if...else` statements. By comparing *Figure 4* to *Figure 5*, it becomes clear how much the complexity of the trees increase as more attributes are taken into account.

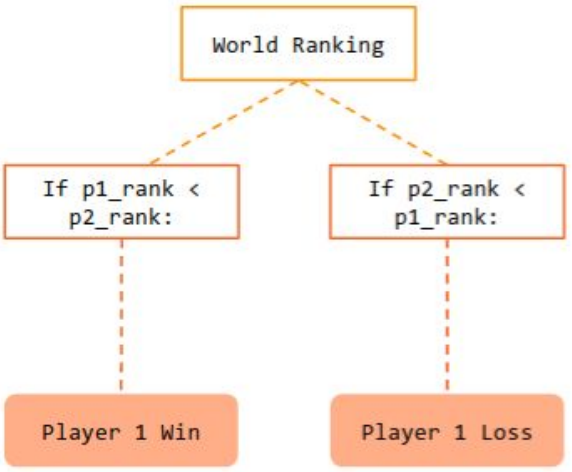

Figure 4: Decision tree to predict the outcome before any sets have been played

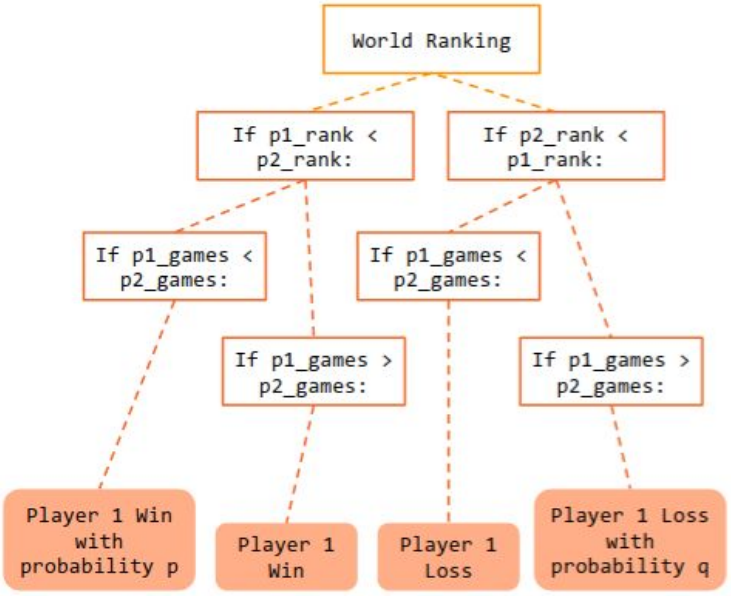

Figure 5: Decision tree to predict the outcome after 1 set has been played.

The probabilities shown in *Figure 5* are based on the Markov chain model discussed earlier. To explain in further detail, if the player with the higher ranking ends up losing the first set, it is not an automatic win for the lower ranked player since they are still ranked lower (and common sense tells us that they have less skill and/or experience).

Looking more closely at the data, *Figure 6*, through observation alone it is clear that world ranking is the best attribute split the data on, since most of the time, the player with the higher rank will come out on top. The headers taken into account for the first decision tree (i.e. before any sets had been played) were "Winner", "WRank" (the winner's world ranking), "Loser", "LRank" (the loser's world ranking). These columns were extracted, renamed and randomised. Randomisation was necessary since any good decision tree would definitely pick up on the pattern that the winners were all in one column - a clear sign of overfitting.

	Series	Court	Surface	Best of	Round	Winner	WRank	Loser	LRank	W1	...	W2	L2	W3	L3	W4	L4	W5	L5	Wsets	Lsets
0	International	Outdoor	Hard	3	1st Round	Dosedel S.	63	Ljubicic I.	77	6.0	...	6.0	2.0	NaN	NaN	NaN	NaN	NaN	NaN	2	0
1	International	Outdoor	Hard	3	1st Round	Enqvist T.	5	Clement A.	56	6.0	...	6.0	3.0	NaN	NaN	NaN	NaN	NaN	NaN	2	0
2	International	Outdoor	Hard	3	1st Round	Escude N.	40	Baccanello P.	655	6.0	...	7.0	5.0	6.0	3.0	NaN	NaN	NaN	NaN	2	1
3	International	Outdoor	Hard	3	1st Round	Federer R.	65	Knippschild J.	87	6.0	...	6.0	4.0	NaN	NaN	NaN	NaN	NaN	NaN	2	0
4	International	Outdoor	Hard	3	1st Round	Fromberg R.	81	Woodbridge T.	198	7.0	...	5.0	7.0	6.0	4.0	NaN	NaN	NaN	NaN	2	1
5	International	Outdoor	Hard	3	1st Round	Gambill J.M.	58	Arthurs W.	105	3.0	...	7.0	6.0	6.0	4.0	NaN	NaN	NaN	NaN	2	1
6	International	Outdoor	Hard	3	1st Round	Grosjean S.	26	Ilie A.	51	6.0	...	6.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	2	0
7	International	Outdoor	Hard	3	1st Round	Henman T.	11	Balcells J.	218	6.0	...	7.0	6.0	NaN	NaN	NaN	NaN	NaN	NaN	2	0
8	International	Outdoor	Hard	3	1st Round	Hewitt L.	24	Woodforde	129	6.0	...	2.0	6.0	6.0	1.0	NaN	NaN	NaN	NaN	2	1

Figure 6: DataFrame showing the attributes available for training.

The headers, "Court", "Surface", "Best of", "Round" and the set results (denoted by "W*n*" and "L*n*" for number of sets won by the overall winner in set *n*, and the number of sets won by the overall loser in set *n*, respectively) were used to build more complex trees for each match. Trivially, the "Best of" header enabled the algorithm to choose which trees to visit and which to bypass (since not every match reaches the full amount of sets played).

The code below shows the simplistic approach of using control flow statements to demonstrate the point of the decision tree classifier:

```
from __future__ import division
import numpy as np
import pandas as pd
from sklearn import tree
atp2000 = pd.read_csv("2000.csv")

rel_2000 = atp2000[['Series','Court','Surface','Best
of','Round','Winner','WRank','Loser','LRank','W1','L1','W2','L2','W3','L3','W4','L
4','W5','L5','Wsets','Lsets']]

winners = []
ranks = rel_2000[['Winner','WRank','Loser','LRank']]
rank_matrix = ranks.as_matrix()

for game in range(len(ranks)):
    if rank_matrix[game][1] < rank_matrix[game][3]:
        winners.append(rank_matrix[game][0])
    else:
        winners.append(rank_matrix[game][2])

print winners
```

Results

The `for` loop (above) looks at each row, evaluates which player has the higher ranking and appends the highest-ranking player out of the two into a list, which is then returned. The accuracy for the 2000 ATP results was given a prediction score of 72.0% (this was calculated by simply comparing the predicted winners list to the original "Winners" column); when tested on 2017 results, the accuracy score was 68.0%.

When fitted to the `DecisionTreeClassifier` the accuracy score stayed the same for the time step where no sets had been played. The comparison between the `DecisionTreeClassifier` and the second for loop in the algorithm was that the accuracy score for the for loop stayed roughly the same, at 73.5%, and the `DecisionTreeClassifier` gave a score of 75.2%[36].

[36] For the 2000 ATP results.

In training the classifier with more data, from the results from 2000 until 2016, the resulting accuracy score was 78.8% for the 2017 data (where the result had been omitted by randomisation of the columns).

LEGAL, SOCIAL & ETHICAL

Dealing with data has, for a long time, been surrounded with controversy. There are a few considerations that should be taken into account when collecting data, for example, an individual's privacy.

As has been discussed in my selection process for choosing a suitable sport for working with decision trees, the conditions under which this kind of data work is effective is determined by how the sport is structured. The conclusion that sports must follow a reasonably workable and appropriately simple structure, to suit machine learning, relates to its legal implications. Like the law, machine learning of this nature works best in environments which are controlled, without much in the way of abstraction or any sense of flexible decision-making[37]. To have a sport such as European Football (soccer) and then to apply the complexities of machine learning to influence or advise in ways where decisions relating to player selection, betting, or anything with human consequences, are made would be unethical. Using the way law works in our society as a comparison, the environment in which sports of this nature operate is not suitably structured to resonate with these techniques.

Like any project that deals with personal data, there are legalalities that have to be looked at. The *Data Protection Act* controls the use of personal data by any organisation. It states that "everyone responsible for using data has to follow strict rules called 'data protection principles'. They must make sure the information is:

- used fairly and lawfully
- used for limited, specifically stated purposes
- used in a way that is adequate, relevant and not excessive
- accurate
- kept for no longer than is absolutely necessary
- handled according to people's data protection rights
- kept safe and secure
- not transferred outside the European Economic Area without adequate protection"[38]

In keeping with the Data Protection Act, this tool must not be used to sell odds to unlawful or unauthorised betting organisations that aim to undermine the betting industry, it should be used for fair observation and fan engagement. The current and future predictions would only be stored as a historical reference, and as the data is not 'personal' (no specific details are kept with

[37] http://www.mlandthelaw.org/papers/burri.pdf
[38] https://www.gov.uk/data-protection/the-data-protection-act

reference to any individual other than name and world ranking), the data that is generated and stored from the use of this tool is fine to be kept as a historical record of the tool's accuracy.

PROJECT MANAGEMENT & FUTURE WORK

The foundations for the project were set loosely according to the Gantt chart, *Appendix 1*, which was a first attempt at including contingency time for any minor delays in completing the programming section of the project. There were a few areas that held me back in completing this project.

The ability to look back, readjust and refocus is difficult. I realised this when I found it incredibly difficult to narrow down my project research area. The journey from wanting to build and collect fitness data from my peers, to then finding some footing in the area of existing data. I looked into Olympic data and discussed heavily with my supervisor, Dr Graham Cormode, the potential this could bring, before coming to the conclusion that the dataset I had found was too sparse for the plans I had initially wanted to carry out.

Olympic data as a whole, of course, did not hold enough continuity so I wandered into the realm of individual sports within the Olympics, before finally settling on predictive analysis for tennis. The reason for this was based on a number of factors: the richness of the data, the amount of data available and the independence of the data (i.e. individual players). This allowed me to evolve my project into something that seemed more in reach - although this was another mistake. I again, set my ambitions too high for the programming side of the project, as I intended to create an entire user interface, complete with working backend, that would allow the user to experience the match on a whole new level. The program would learn from the data that would be scraped in real-time giving the user insights that would enhance their view on the current game.

Learning to cope with current personal mental health issues, which hindered my ability to focus on both research and development. Through this project, I have truly tested the limits of my welfare and what I have learnt from this experience is that I am able to recognise when I am overestimating myself.

This allowed me to continue on a basic level, where I created decision trees to predict the outcomes of the matches based on time series data (i.e. the tree would evolve as time went on, and as more of the match had been played). In future, more work would be done to integrate the Match Charting data, to enable a near real-time experience for predicting the outcome of the match.

Given more time, I would like to have also finished a user interface to showcase the interactivity, which would be marketed at data journalists as a tool to be used during the Grand Slams to engage spectators in how they feel about individual athletes. The rise of interactivity in digital journalism has brought a sense of statistical sophistication to the masses: journalists must be as

conscious to use data correctly for a multitude of reasons, as touched upon in the ethical reflection[39]. The rise of data journalism as a standard method means that capabilities in the newsroom are improving and, as such, developers and reporters must be able to work together[40]. Products such as this tool can be a culmination of these efforts and it's something we're seeing more and more of.

CONCLUSION

I believe this project has been a (partial) success, in that it was well conceived; regardless of the fact it was after much trial and tribulation, the idea behind the project made sense to fill a gap in the industry that has previously not been tackled in this way before. The challenges were plentiful, and overcoming them were difficult. But it produced an understanding of what is required for sports data analytics to truly take off, and how it can impact the entertainment and reporting value of tennis, and of sports as a whole.

If given more time, it would be entirely manageable to complete the future work stated above. The main aim and intention of gaining a deep theoretical knowledge of betting odds, predictive modelling and sports analysis has been achieved through thorough literature review and presentation. The project management aspect was what let me down, and if I were to complete a task of a similar nature, I would consider asking for more help, earlier on in order to face any troubles I had. I would also put more time into the programming section in order to give some backbone to the presentation.

It has been a learning experience first and foremost, and that is why I can still consider this project successful.

[39] http://www.interhacktives.com/2017/03/24/alberto-cairo-steal-best/
[40] http://www.bbc.co.uk/sport/olympics/36984887

APPENDICES

Appendix 1: *Initial Gantt Chart outlining project and time management*