Pro Tennis Gameplay Analysis

Hiroka Tamura | ID: 912801779 **Andy Wu | ID: 912991265** **Ryan Gosiaco | ID: 912819444**

Bailey Wang | ID: 914955801 **Kevin Chu | ID: 913077890**

Department of Statistics

University of California, at Davis

Abstract

This paper works to explore the disparity in professional tennis players' career performance based on their historic game-play statistics. The goal is to identify the key features that distinguishes an outstanding player; We do this by evaluating different dimensions of match data and player characteristics.

1 Introduction

Tennis is the fourth most popular sport in the world. The sport has been around for a multitude of years and is estimated to have over one billion followers. As a result, tennis has a rich history and it has been made popular by players with highly diverse backgrounds and playing styles. Tennis is also unique because unlike a vast majority of sports such as soccer and basketball, tennis is an individual sport.

Because of the sport's popularity, there are many datasets available that captures several dimensions. Using a combination of these datasets, we will answer a variety of questions ranging from general trends we discover across the entire tour to more specific aspects of the upper echelon of players. We are interested in answering the following questions: How important is rank difference when two players compete? How can we define the separation between a top 10 and top 100 player? Can we quantify mental stability that affects a player's gameplay? What are the significant features that allow us to predict match outcomes?

Furthermore, in tennis there is a term called “the Big Four”. This term is used to refer to four male players that have dominated the modern era of tennis. These four players are Roger Federer, Rafael Nadal, Novak Djokovic, and Andy Murray. We wish to extend our questions to compare the big four to the rest of the players.

Terminology *Ace*: Serve where the ball lands inside the serve box and is untouched by the receiver. *Break Point*: The point which, if the receiver takes, wins the set. *Double Fault*: Failing to serve the ball twice resulting in losing a point. *Forced Error*: Error caused by opponent’s good play. *Serve*: The starting stroke for each point. *Tiebreak*: Special round played when the game count is 6-6 to decide the winner of the set, to reach at least seven points with a difference of two points over the opponent. *Unforced Error*: Error in serving or returning a shot that cannot be attributed to any factor other than poor judgment or execution by the player.

2 Data Acquisition

Unfortunately, there are no publicly available API’s from the ATP website. However, we were able to find two datasets to work with. The first is a set of publicly crowdsourced datasets from Jeff Sackmann’s github. The datasets are called the Match Charting Project and it is a compilation of shot by shot data for each match recorded. It contains a lot of information which includes types of shots, directions of shots, depth of returns, and more. The second dataset is a web scraping of ATP rankings throughout many years. This dataset contains information about every ranked player on the tour such as age and player profiles. This is done with a public web scraping script from Kevin Lin. We updated the script to obtain more recent data.

Jeff Sackmann’s webscrape consists separating all of the information into different CSV files. These files provide more in depth details to analyze but make it harder to understand the overall picture. We used Kevin Lin’s python code to webscrape current statistics to fill in the gap from when the dataset was last updated.

3 Data Pre-processing

Since the Match Charting Project datasets don’t contain information about player ranks, we first worked to merge the ranking information from rankings dataset into the Match Charting Project datasets. Doing this allows for us to conduct analysis about differences between players of different ranks. To do this, we have to split up the ‘match_id’ column which contains information about ‘date’, ‘tournament’, ‘stage in tournament’, ‘player 1’, and ‘player 2’. Doing this also allows for us to be able to separate the statistics of each player during a match up. The Match Charting Project datasets also don't contain information about which player won the match so we also had to engineer that feature based on the data.

4 Methods

The main method used to gather the data was web scraping. Luckily, the ATP website is well formatted and the bulk of the dataset up until 2017 included Python2 code that was used to gather it. We then converted it to Python3.

We perused Jupyter Notebook with python for data wrangling and graphic result display. Some notable packages used are *plotnine, plotly, pandas*, and *numpy*. As for machine learning methods, we implemented *xgboost, sklearn*, and *keras* for Gradient Boosting and Neural Network methods.

5 Results: Data Exploration

5.1 Evolution of Tennis

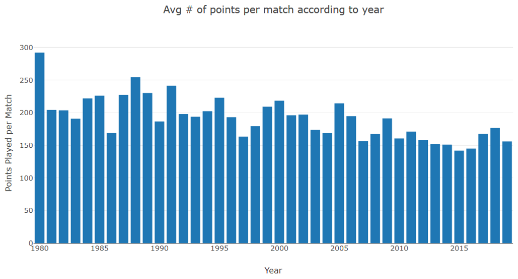

Figure 1: Average number of Points per Match by Year

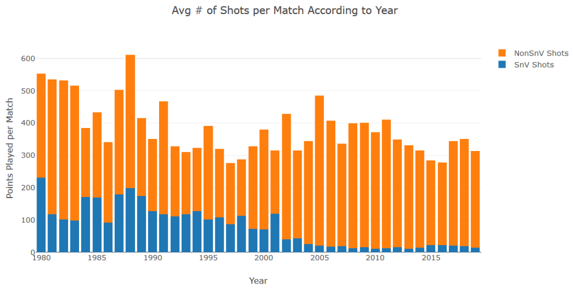

Figure 2: Average number of Shots per Match by Year

According to fig. 1, since the beginning of the sport, the average number of points played per match has slowly decreased. There are two possible explanations. The first explanation is that the level of competition has decreased. The second explanation is that players have gotten better at closing out matches when they have a lead. From general intuition, the second explanation is more plausible.

In fig. 2, there isn't a clear trend that shows the evolution of the sport. From this current plot however, it is clear that the sport has drastically changed since its inception. In tennis a commonly known strategy is to serve and volley. This tactic was widely used in the years prior to 2000. In the most recent two decades however, this strategy has been widely phased out making it seem to be more of a gimmick than a playstyle.

5.2 Average Match Stats by Rank

With fig. 3-6, we can see the differences between players of different ranking categories. Each plot compares players of four different calibers: top 10 players, top 50 players, top 100 players, and players outside of the top 100. The metrics compared are player/opponent points won, winners, forced errors, and unforced errors. From intuition, one would assume that the largest difference in these statistics would be between top ten players and players outside of the top 100. Almost every plot indicates a small difference between the different rank categories but in a sport such as tennis, every point matters because it only takes a minimum of five points to win a game.

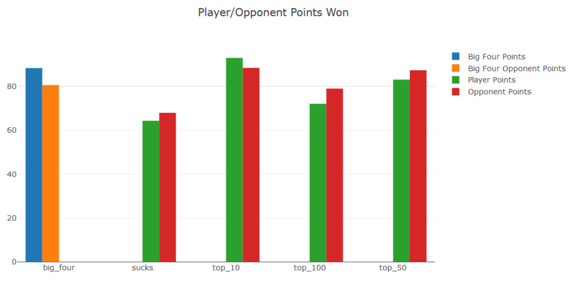

Figure 3: Player/Opponent Points Won

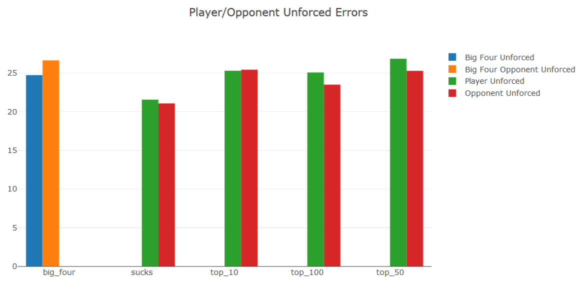

Figure 4: Player/Opponent Unforced Errors

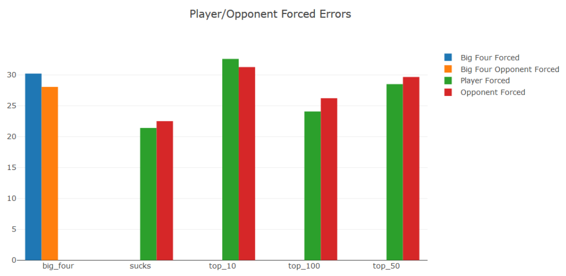

Figure 5: Player/Opponent Forced Errors

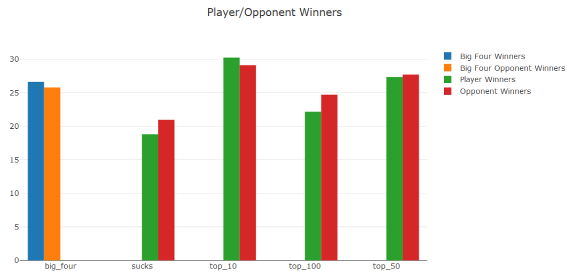

Figure 6: Player/Opponent Winners

5.3 Longevity of Top Players

In order to better explain the fig. 7, it must be noted that there are two major types of tournaments that hold extreme prestige. These two types of tournaments are called ‘Grand Slams’, and ‘ATP Masters’ and they are respectively worth 2,000 and 1,000 points. There are four grand slams and seven atp masters tournaments per year totalling up to 15,000 points. There are also smaller tournaments that award up to 500 points but their prestige pales in comparison to grand slams and ATP masters tournaments.

We can further investigate the longevity of the top four players by examining how much their ranking fluctuates over the years as in fig.8; Though they may be alternating amongst each other, for nearly an entire decade they maintain their spots in the global top pool. Despite their quick rise to fame, their skillset is proven not to be just a phase.

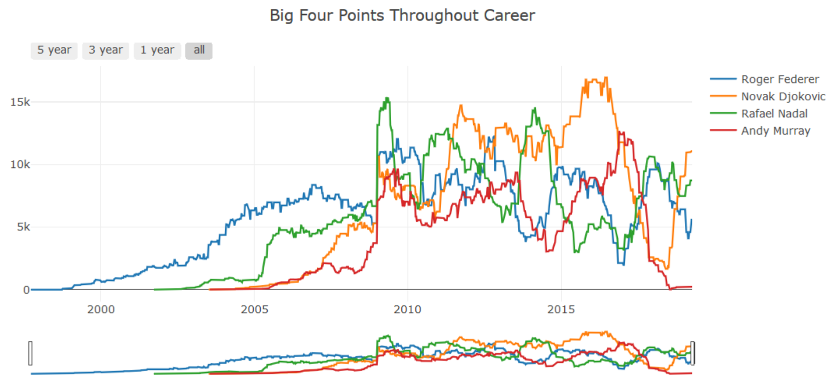

Figure 7: Big Four Points Throughout Career (Overview)

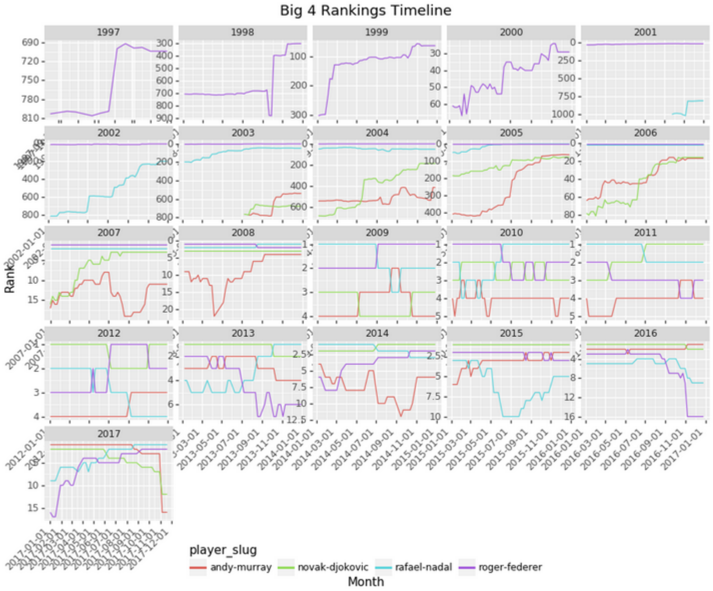

Figure 8: Big Four Ranking Throughout Career (By Year)

5.4 Comparing the Big Four

Radar graphs of each pairwise match statistics for the big 4 players in fig. 9 intends to visualize the aspects in gameplay that differ by the pairing of players and features that are significant (or not significant) in deducing the outcome of the

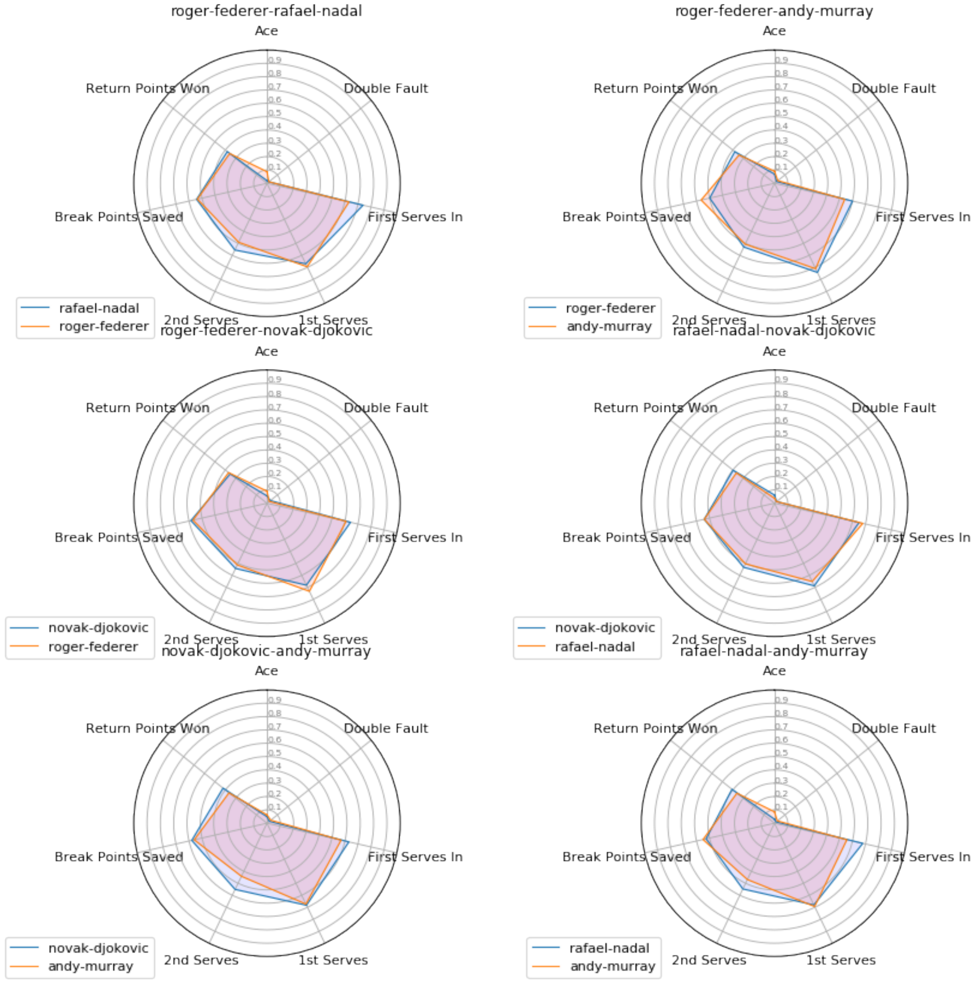

Figure 9: Piece-wise Match Statistics Averages by Big Four

match. The blue indicates the player that has won more matches between the pair. One thing to note, as mentioned above, is the fact that rankings do not necessarily the indicator for what the outcome will be for a particular match.

For example, in the first radar graph you will see that Rafael Nadal wins more matches against Roger Federer whilst Federer holds the higher rank for significant number of years. Further investigation shows that Nadal actually wins 23 out of the 34 matches that they play together. Since ranks are determined by the cumulation of points earned by their entire career, player-to-player playstyles are not taken into account. Perhaps rankings are heavily biased on luck for who they are matched to play against. Overall, you will see that the blue lines of the winning players have a higher rate of “First Serves In” and “2nd Serves” which indicate the proportion of points won from their second serves. These features may be a significant part to include in the model, and influences us to believe that deeper analysis on service performance will be beneficial in understanding what makes a good player.

Observe the contrast between the graph below of Djokovic vs. Anderson and those above in fig. 10. Kevin Anderson is currently ranked at 8 in the world with the highest historic career title at 5. Although that alone is an impressive profile, there is a drastic difference in the piece-wise radar charts of the stats among the big 4. Kevin Anderson is known for his extraordinary service performance - a distinct jump on the Ace tick and high rates in first serve percentages. He has had more than 6000 aces in his career while others of similar rankings range between 3000 and 5000. However, what really sets the difference between the big 4 and Anderson is the break point and returns performance. We hypothesize that strong serves are one of the main factors that determines winners and high ranking players (which is also true for Anderson), however the lack of well roundedness in Anderson's gameplay stats suggest that his performance as a receiver is insufficient to be as competitive.

Not only can we infer differences in shots, but the distinct stats for break points saved in particular can provide insight in the mental ability in a player as well. Break Point Saved refers to the percentage of break points that the player has retained; in other words, how well the player won the point that is at a very high stake. Losing the break point would mean to lose your serving game, which breaks the flow for the set negatively. Every competitive player is expected to save their serving game. Mental strength is also a highly discussed aspect in tennis since players are highly aware of the moments that may change the course of the match. Often gameplay stability wavers at these points, and to be able to overcome these nerves is a key skill to keep up the score. With a significantly lower break point save rate and a drastically different graph structure altogether, Anderson comes short in playing consistently aside from his serves, which ultimately sets him apart from being on top.

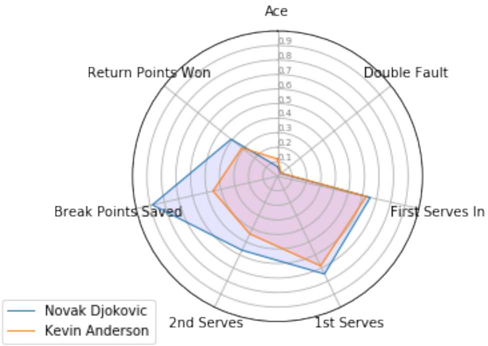

Figure 10: Big Four player vs. Non-Big Four Match Statistics Averages

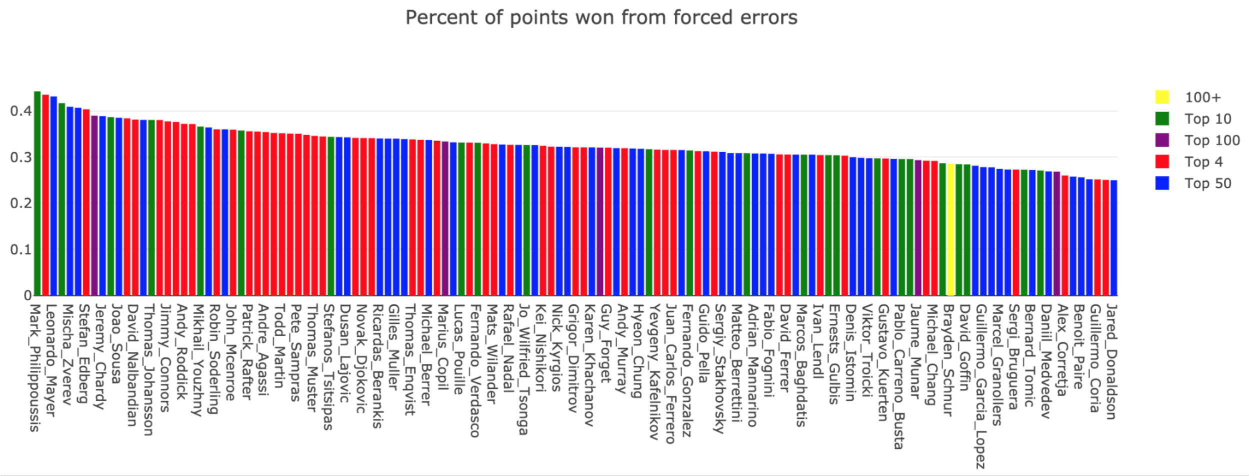

Figure 11: Percent of Points Won by Forced Error

5.5 Forced Error Effects

In tennis, a forced error is an error where a player loses a point as a result of their opponent's previous shot that resulted in no other option but a forced error. Some examples of forced errors are drop shots, where the player must sprint to the net to attempt to return the ball, and sudden changes in pace of the ball. In these cases, the player loses the point due to the shot from the opponent and thus the point is considered a forced error. In contrast, an unforced error is when a player loses a point due to their poor shot, such as a ball in the net.

In general, a player who wins more points on forced errors should be a better player, because a higher rate of points won on forced errors is a direct indicator of points won and thus games won. We explored this in the Match Charting Project point-by-point dataset and compared it to the peak ranks of each player.

Fig. 11 shows the percentage of all points won by a player by virtue of a forced error. The color of each bar indicates what the player's highest career rank is/was, binned into categories. We separate the top 4 from the top 10 because a player in the top 4 ranks receives the best seeding in tournaments as they are mostly brackets made up of 4 quadrants where the top player in each quadrant is one of the top 4 ranked players.

In fig. 11 , we see that Mark Phillippoussis has the highest career percentage of points won by forced error. Additionally, his career highest rank is in the top 10. This indicates that forced error rate does correlate to overall player performance but also that there are other factors that contribute. Additionally, we can see that there is a higher density of top 4 players on the left of the plot, indicating that the best players in the world are also the best players at forcing their opponent to commit an error.

5.6 Match Length Effect on Service

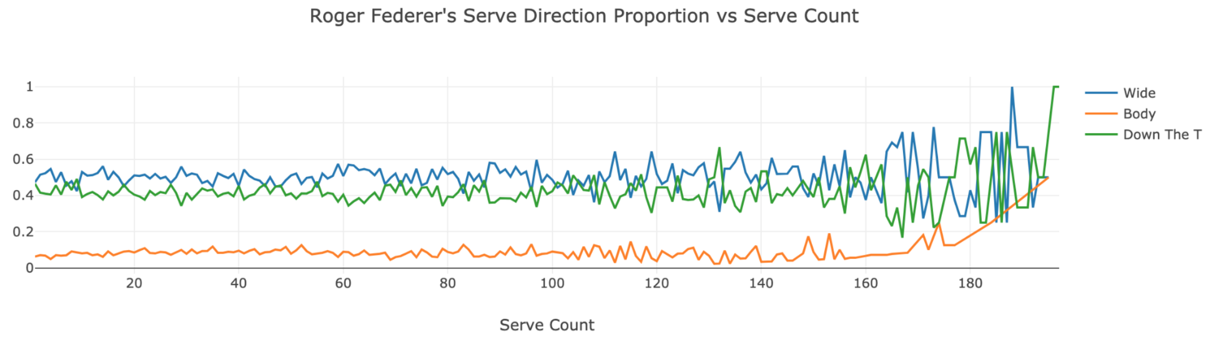

Figure 12: Roger Federer Serve Direction vs. Serve Count

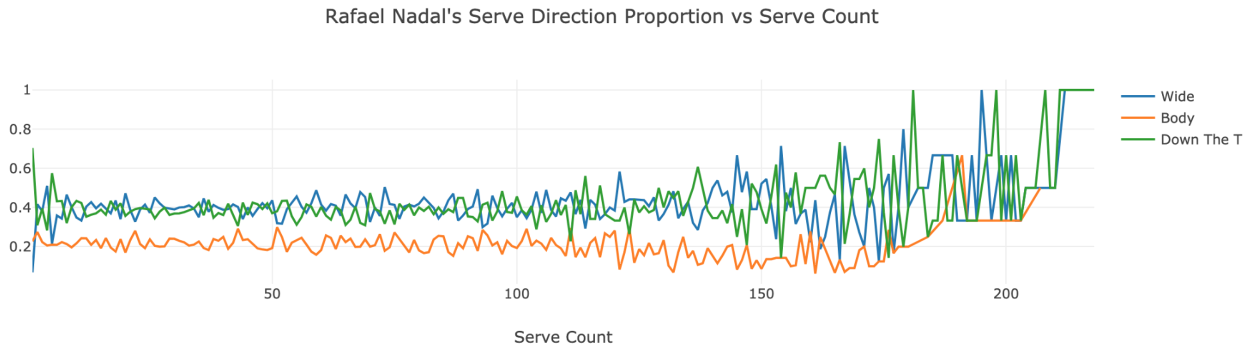

Figure 13: Rafael Nadal Serve Direction vs. Serve Count

The Tennis Match Charting Project tracks serve directions in three directions. A serve made wide is a serve made toward the outside of the court, typically also on the outside of the opponent. A serve made toward the body is a serve that is hit directly at the body of the opponent. A serve down the T is a serve made down the middle of the court.

We wanted to examine whether the length of a match had an impact on how the top players directed their serves. Perhaps as the length of the match gets longer, a player may want to take more risks and serve out wide, which while more likely to result in an ace, also is more likely to land out of bounds. Or perhaps, as the length of the match gets longer, a player may want to serve the ball down the T, a safer shot than a wide shot, in order to stay in the game, especially if they are behind.

To answer these questions, we used the table of all points charted in the Match Charting Project database, aggregating on the serve number for each point and the proportion of each serve direction.

In fig. 12, we see serve statistics for Roger Federer. In general, he takes more wide serves than serves down the T for about the first 100 serves, but around 100 serves into the match, which generally equates to about the third or fourth set, he starts experimenting more and serving more shots down the T, although still primarily utilizing the wide serve. This shift in serve strategy could mean many things, but considering that Roger Federer is a top 4 all-time player, it is likely that he changes up his serve game toward the end game to finish off his opponent and stave off the chance for a comeback.

In fig. 13, we see the serve statistics for Rafael Nadal. Nadal is the player in our exploration who mixes up his serve directions the most. Interestingly, Nadal by strong majority opens up his serve game with a shot down the T, utilizing the shot on 70% of his first serves. Additionally, Nadal tends to have longer matches than either Federer or Djokovic. However, it appears that for Nadal at least, the length of the match does not change where he serves the ball as he tends to use a mix of wide serves and serves down the T.

As the match goes on, we can find in both figures that the oscillation for each serve type fluctuates; This suggests that players' service grows unstable with time likely due to fatigue.

6 Results: Machine Learning Based Methods

In addition to the graphical methods we explored earlier, we decided to consider and try to apply various popular machine learning techniques such as gradient boosted trees and neural networks. Since our data is structured, we wanted to try and train a model that could predict the rank of a player based on the total match stats.

Gradient boosted trees (GBT) can be applied to supervised learning problems and work by essentially creating an ensemble of trees which are created one at a time. Each new tree helps to correct any errors made by the previous tree and because of this, with each new tree, the more the model captures. However, GBTs can take a long time to train due to the fact that the trees are generated sequentially and they are inherently prone to overfitting. These downsides can be minimized by tuning the parameters such as number of trees, depth of trees, learning rate, etc.

These parameters can make a big difference in the overall model performance too and because of this, it is critical to find the best set of parameters to maximize the final model performance. A grid search is time intensive but it does not require supervision so other tasks can be completed as the grid search runs over the parameter space. This implementation we used is sklearn's GridSearchCV which not only performs a grid search but does a cross validated grid search over the parameter grid. The cross validation combined with the grid search results in a very effective way to generate the “best” model from the parameter space because the cross validation verifies the effectiveness of the model over the whole dataset.

Since the grid search is computationally and time intensive, it is wise to try and “hone in” the model parameters as best as you can beforehand instead of blindly running a grid search. We found this out the hard way because we decided to blindly run the grid search on the data and it resulted in very poor results. We thought it was due to the variables we were predicting on, so we limited that but it still had poor results. Since our dataset extends as far back as 1970s we figured that the scope of the dataset was skewing the results because various aspects of the game like playstyles, sports science/medicine and other things have changed over time. Taking this into consideration, we only looked at the last 5 years of data but even then the accuracy was not very good. The accuracy of the model was around 12% which was very low, and that was using the “best” model created from the grid search.

We then realized that a multiclass classification problem is very difficult because the dataset is very imbalanced. Since each rank is not equally represented in the dataset and also because we looked only at the rank of player 1 and the total match stat instead of the stats after each set.

Therefore we decided to turn it into a binary classification problem where we wanted to predict whether or not a player was rank 1 based on the total match stats. This small change ended up simplifying the problem significantly and resulted in fairly good accuracy (86%). This shows how a class imbalanced dataset can result in poor performance when trying to do multiclass classification.

Furthermore, we then went to explore neural networks through the use of Keras with a Tensorflow backend. We constructed a basic neural network with 9 hidden nodes, 3 hidden layers, a batch size of 32 and learning rate of 0.1 using a elu activation function. This model trained very quickly and resulted in a similar performance to the prior models. Along with a neural network, we also ran logistic regression and random forest classification models as a baseline for accuracy/performance since those two models are relatively simple and easy to run because there are only one or two parameters that can be changed.

The ROC (fig. 14)and PR (fig. 15) curves tell us about the overall model performance and from those graphs we can see that for some reason, both GBTs and neural networks have a ROC curve that resembles a random classifier. This leads us to believe that there might be something else going on within the dataset or something with the way I created the testing/training set.

Looking closer at the proportions of each rank for player 1, we found out that the dataset is very imbalanced and for obvious reasons, does not contain an entry for each rank. Therefore, creating any sort of model to try and determine a player's rank based off of match statistics is very difficult because there is not enough representative training data and there does not seem to be one stat that is significantly more important than the rest.

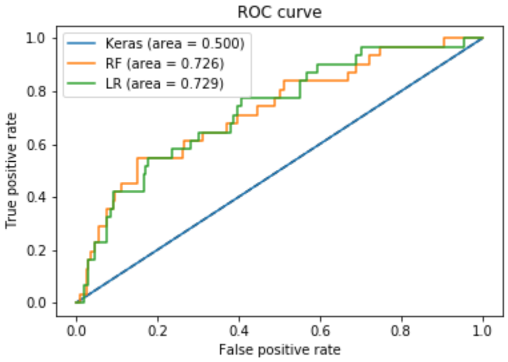

Figure 14: ROC Curve

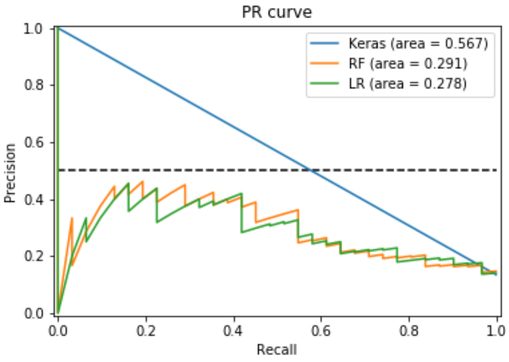

Figure 15: PR Curve

7 Conclusion

Overall, we were able to analyze the data and gain insight as to how to apply this in the future to a more in-depth dataset. The main drawback to the dataset we were using is the fact that it does not contain every match that has happened because each match has to be recorded by a volunteer. Another problem we had to solve was incorporating the ranks of each player at the time of each match into the main match charting dataset. We solved this by web scraping the ranks utilizing code from a few years ago that we found on GitHub and then merging the two datasets together.

Our first step in analyzing the dataset was to consider key points of areas of interest to focus on. We began by looking at the distribution of the age of each player based on their rank. We found that a majority of players are around ages 20 and 21 which makes sense because many professional players make their debut at that age. Furthermore, advances in sports medicine and technology have most likely contributed to elongating a player's "prime years". After that we focused on points and shots per match and how it changed over the years. This helped us later on in understanding current player styles/trends and how they evolved over the years.

From there, we took a closer look at the Big Four. The Big Four is comprised of Roger Federer, Novak Djokovic, Rafael Nadal, and Andy Murray. These four players have dominated the modern era of tennis and each have their own unique style. By taking a closer look at the points over their careers so far, we can see that these four players have held a majority of the points over the past decade. Furthermore, when we look at their ranks over the years, we can see that they maintain their spots at the top. This cements their greatness and that their skills/performance are not just a phase. The next idea we had was to create a radar graph of each pairwise match statistic for the Big Four. We intended this to visualize the aspects in gameplay that differ by the pairing of players. By these graphs, we were able to discover what is the difference between the Big Four and other players. This led us to look further into points won from forced errors and the effect of match length on serve direction. By graphing each serve direction of the serve count, we discovered trends in player's style and noticed how unique each graph was for each player.

In regards to common machine learning related techniques, applying them to our dataset was difficult. Our main focus was on trying to determine the rank of a player based on the overall match statistics which is a multiclass classification problem. This becomes difficult to model properly because the dataset is very imbalanced. This means that each rank is not represented equally frequency-wise which makes sense. This issue can be partially solved by converting the task from multiclass classification to one vs all classification which is essentially a binary problem. As mentioned earlier, doing this changed our model's performance from around 13% to 86% which is a significant increase. This means that there is some overall match statistic that defines a rank 1 player but because of how volatile the ranks could be, a rank 1 player could be rank 2 the next match. This leads to difficulty in classifying a player by their rank because their rank is not always representative of their actual skill level.

Ultimately, the best technique for analyzing this specific dataset was doing graphical/visual analysis. It was much more straightforward to understand and explain to other people.

Supplementary

7.1 Age

As seen in fig. 16, the distribution of age in Tennis is centered around 20 and 21. However, despite the majority of Tennis players around ages 20 and 21, they start to peak when they reach their later years of playing tennis. As in fig. 7, the Big Four players continues to achieve results, despite being in their 30's.

Nearly decade ago, the majority of tennis players would retire around 27. This most likely results from improvement fitness and nutrition which helps players retain longevity. This also includes players taking recovery more seriously retaining their own ice packs, oxygen tent, and physiotherapist. However, the most important factor in which these players continue to play is the monetary gain. These factors have held more players to play the game longer contributing to the increase in the average of player ages.

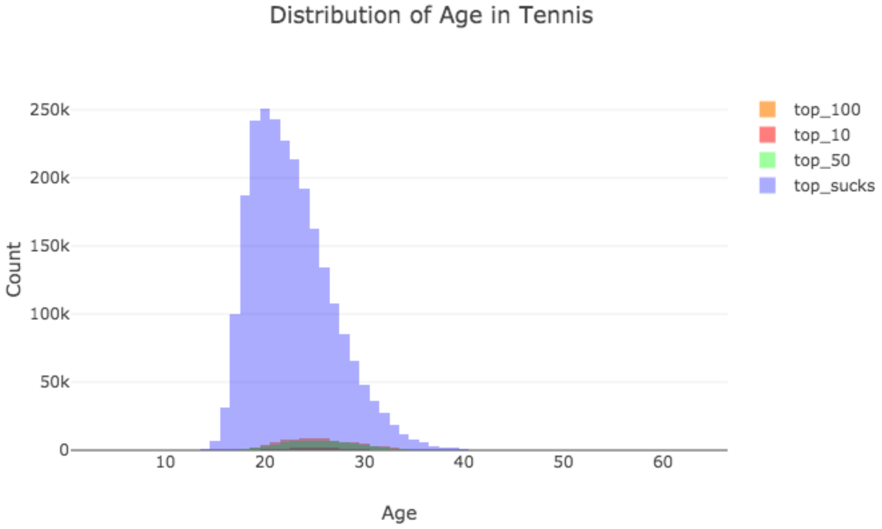

Figure 16: Age Distribution