**Final Project Document**

**Sebastian Buxman**

The goal of the project was to gain better insight and understanding of attributes that create success for professional tennis players based on statistics. The data collected included professional tennis players statistics from the years 2010 to 2023. Based on my personal experience play college tennis, the three questions for the project include: firstly, how important the serve is based on the playing surface because on different surfaces the ball typically bounces differently (Lindsey, 2004). Secondly, to see whether height makes a difference on how well a player served, this can be gauged by the number of aces a player hit in a match. Lastly, the discrepancy between breakpoints saved between the losers and the winners- a break point saved is when a player saves a game point against their opponent when they are serving. Finally, my goal is to predict whether a player is in the top 100 in the world based on these attributes. This can be done by predicting accuracy value on the test set, and to see which of these features have the greatest importance by using random forests to predict feature importance.

I hypothesized that there would be a strong correlation between height and the number of aces a player hit. Furthermore, there will be a strong correlation between the number of aces hit and the surface. In addition, a player’s age will be a strong factor in determining success. Lastly, these attributes will be able to predict ranking with a high degree of accuracy.

The data used in this project was found on a GitHub repository (Sackmann, 2023). The data was collected by separating each match during a specific year on a separate row. The data was then grouped based on the player name because the algorithms are predicting ranking based on certain player attributes.

Within the professional tennis realm, tennis tournaments are divided into three categories of tournaments: Futures tournaments, Challenger tournaments, and ATP tournaments. Futures tournaments are at the lowest level with the lowest prize money and points for the athletes. Challengers tournaments are considered medium level and the ATP tournaments as the highest level. A complete view of professional tennis was needed to make a meaningful prediction, so a combination of all the data from 2010 to 2023 from the three types of tournaments was used. The next step was grouping the players’ data based on their name and summing the statistics from all their matches which resulted in total matches won, total matches lost and the players’ statistics. A player’s latest ranking was appended into the combined data frame – after 13 full seasons, these statistics for each player resulted in their latest updated ranking.

First, a correlation matrix is plotted to see whether any correlations stood out. After doing this, it was surprising to see that there weren’t any correlations off the bat. Using my knowledge of tennis and how the game is played, I wanted to see if there were any correlations between aces and serves. I plotted the number of aces vs the winner height and used regression lines to show the correlation between surfaces a little more clearly:

A graph of colored dots

Description automatically generated

Winner Height vs Number of Aces:

As one can see, there is a clear correlation between the number of aces the winner of a given match hits and their height. Players who are taller can swing downward with a much higher velocity with less chance of hitting the ball in the net. It is also evident that the surface plays a role in how many aces are served, so plotting a box plot to see the differences in the number of aces based on the surface. Below shows an individual that the surface of a court plays a role in the number of aces:

A graph of different sizes and colors

Description automatically generated

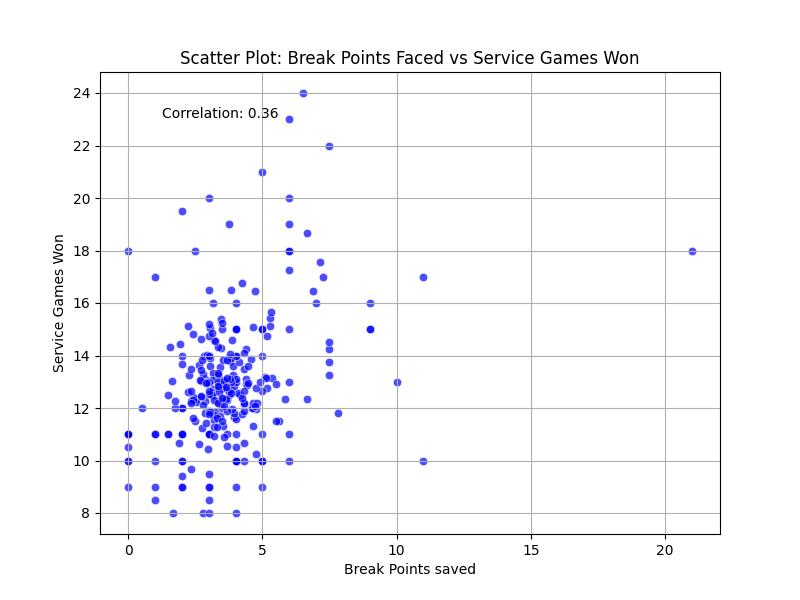
The graph shows there are much fewer aces hit on clay courts than any other court. When the ball bounces on clay it loses a lot of speed, the clay glues to the ball, slowing the ball down and allows the ball to bounce higher. Looking at grass and carpet there is nothing to slow the ball down – in fact, the grass is often slightly moist so as you can imagine the ball skids through more just like if you kick a soccer ball on moist ground.

A graph of a number of points

Description automatically generatedNow, aces are clearly affected by height and surface but going the other direction and looking at double faults (when a player misses both their serves and loses the point) do these two features play as big of a role?

There is not as much of a correlation when looking at double faults. Contrary to my initial speculations, I thought there may be a slight correlation between height and double faults, using the logic that the shorter a player is the harder it is to make the ball over the net; , at this level, most players have a decent enough serve to make a relatively low number of double faults. Although, it was interesting to see the slight correlation between surface, this may be due to players knowing the carpet and grass courts play faster, therefore increasing their chances of winning the point by an ace or a forced error on the opponent’s end and just going for more than they would on the hard or clay surfaces.

If we look further into the importance of serving in professional tennis, we can identify how important saving a break point is for a winner. I hypothesized the winners will face less break points and therefore save less break points because looking at the previous plotted data generally the winners generally have a better serve.



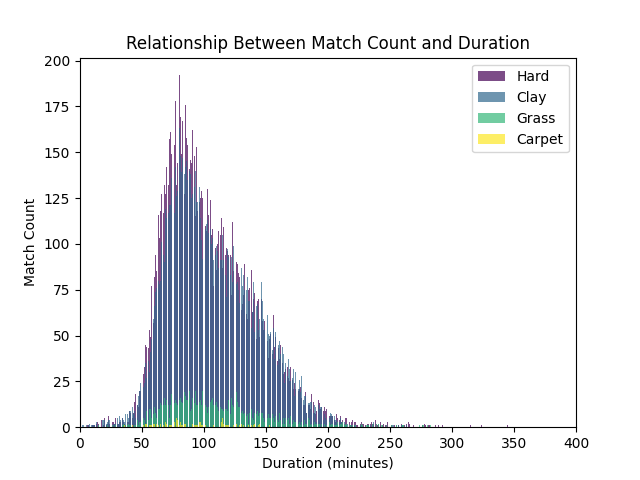
There is little correlation between break points and games won. This may be because winners served better in that specific match. If we examine this further and plot the mean difference in break points saved between the winners and loser based on the round that was being played:

A graph of blue bars

Description automatically generated

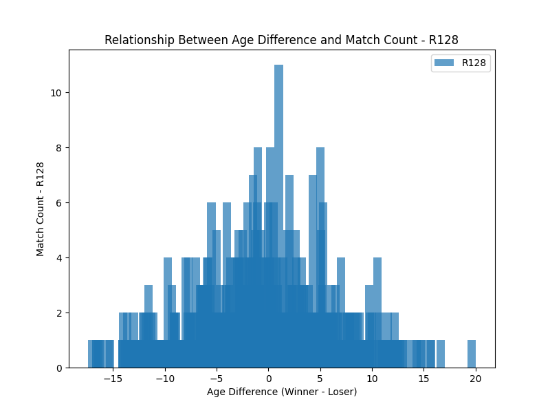
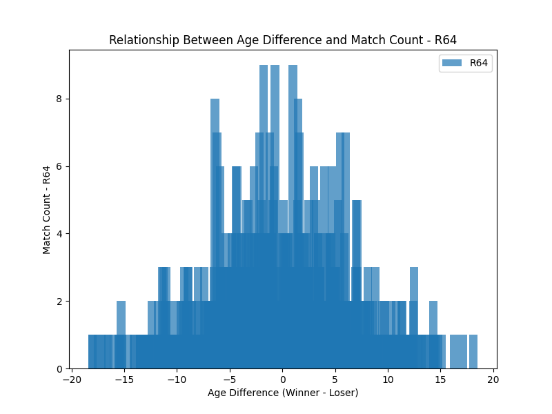
Here, one can clearly see that the winners are able to save more break points than the losers – it is evident that in the first rounds of tournaments (R128) there are more break points saved, this may be due to the fact that the highest ranked players usually play much lower ranked opponents in the first round of tournaments and the difference in their level is more apparent. The break points saved in the final of tournaments is also higher than the semi and quarter finals by almost 0.30.

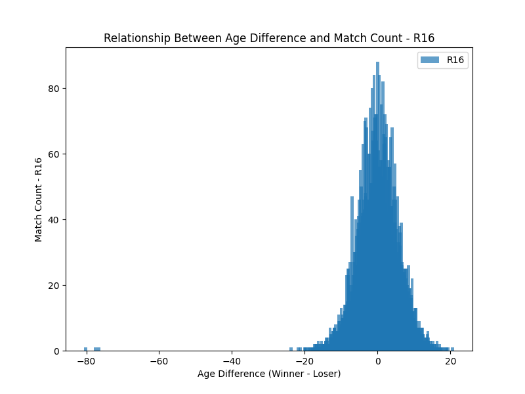
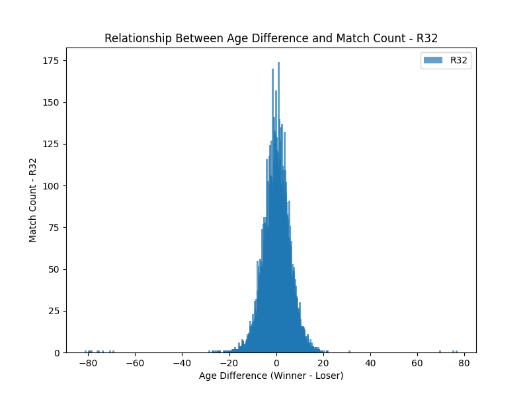
Subsequently, coming into this project I had the assumption that age and match duration would be important features when looking at the winners in this professional tennis dataset. Firstly examining match duration based on surface and round:

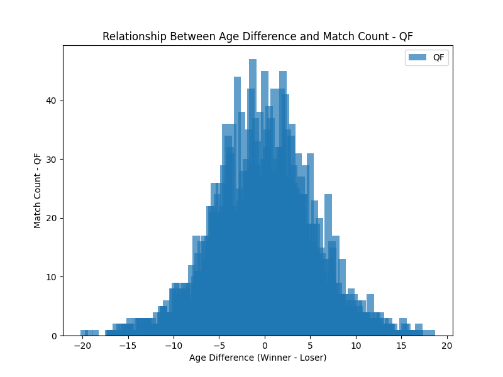
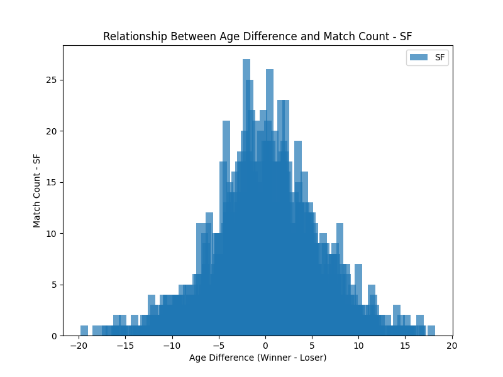


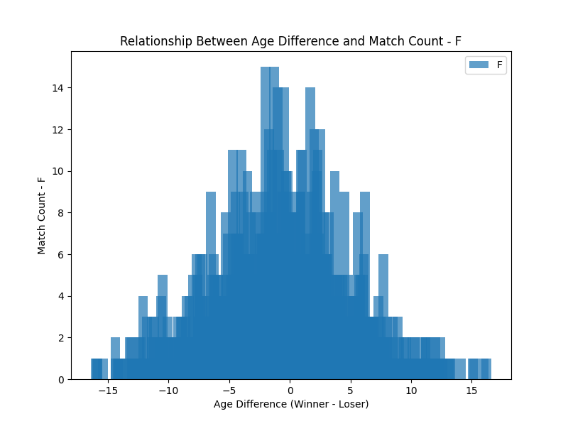
It is clear that hard and clay courts matches are much longer than the matches played on grass and carpet – however, it is important to note that the amount of tournaments played on grass and carpet are much less than the tournaments played on clay and hard courts (CNN, 2010). At the same time, these correlate to the amount of aces hit on the different surfaces and comes back to the fact that the ball bounces slower off of these surfaces.

There were no blaring differences between the differences in ages between winners and loser, in fact most of the time most of the data was grouped around 0 meaning the two players were around the same age, included rounds are R128, R64, R32, R16, QF, SF, F:









It is interesting to see how the data is spread out in R128, R64 and in the final when in the rounds in between them, the ages are closer to 0. This tells us that age is not an extremely important feature like my hypothesis stated.

Moving on to predicting a player’s ranking – I used a random forest to plot the feature importance of each feature from the combined data frame I created. Winner\_id and winner name were removed because the name and ranking had a direct correlation. Furthermore, name is not an attribute or trait that a player uses during a match and using this as a feature would result in inaccurate results.

A white screen with red text

Description automatically generated with medium confidence

Total\_wins is by far the most effective when it comes to feature importance of a random forest – the other features are very low in comparison. This makes sense because the more wins, the higher the ranking. However, when we analyze the fact that there are multiple levels when it comes to tennis, if a player is winning many matches on the future’s circuit, then it does not necessarily translate to the challenger’s circuit or the ATP in terms of results. Judging from the result of feature importance this does not seem to be the case. I could not get the regressor to show a higher importance than I got with 500 estimators at a max\_depth of 3. Using more estimators or depth resulted in the data being overfit and the predictor performing extremely poorly.

Initially, I didn’t want to drop the 0 values because it reduces the size of the database vastly; however, an interesting pattern occurred when I did:

A graph with a bar chart

Description automatically generated with medium confidence

Total\_matches had much less of an effect, whether the loser was able to save break points became the most important value by a big margin.

For the prediction part of this project, three different machine learning algorithms were used. SVM regressor, Random Forest, and a neural network because the dataset is relatively small so computational efficiency will not be a major factor. The problem being faced is a regression one because a quantity is being predicted. Furthermore, dimensionality reduction is not needed because of the size of the final dataset (12587 rows).

After scaling the data using StandardScaler() and splitting the data into a training and testing set:

A computer code with text

Description automatically generated

These were my final scores after using the RandomForestRegressor(), SVMRegressor and MLPRegressor: (on the 0s in the dataset)







After applying AdaBoost to the Neural Network, the following accuracy score was reached:

Boosting did not help much; in fact it decreased the accuracy of the model.

Seeing as the dataset is relatively small, I thought training a decision tree regressor might be beneficial, but it ended up being very similar to the Random Forest Accuracy:



Plotting the predicted values vs, the actual values the following result occurred:

A graph with blue dots and a red line

Description automatically generated

The data seems to follow the correct value reasonably well. However, there are some glaring errors in the predictions especially as the rankings get higher where there is less data to predict on in the test set. Dropping the 0 values yields this result:

A graph with blue dots and a red line

Description automatically generated

The data follows the perfect prediction line reasonably well when predicting on the top 500 tennis players, these players have more matches and therefore more data to use for the prediction, furthermore, to make a decent living from tennis a player needs to be in the top 300 in the world. Therefore this model performed reasonably well on the dataset when predicting the top 300 players. (Hadlich, n.d.)

But the overall accuracy values were much less accurate:







In conclusion, it is evident that there are certain correlations between attributes such as service, break points saved and double faults. However, these correlations are not necessarily good indicators of a professional player’s overall success when trying to predict their ranking. An idea for future work would be to gather more data on tennis players such as information on forehand and backhand winners as well as volleys, this would give a better overall insight into the game and therefore give us more data to work with when trying to predict ranking. I hypothesized that there was a relationship between surface, height and serves and in this prediction, I was correct, but with the same token, after examining these specific attributes and applying different machine learning algorithms I can deduce that they do not play a major factor when trying to predict ranking. Overall, this project highlights how nuanced the sport of tennis is – there is evidently no set of steps that can guarantee success, the only way to be successful is to work hard and to enhance what one can control. It is evident that being able to win points on one’s serve while losing one’s own serve as little as possible can lead to success but won’t guarantee a high ranking. The l\_bpSaved is a feature that stood out where how many break points a player was able to save directly affected ranking, if the player did not serve well and as a result lost many break points, they lost the match. Breaking serve as often as possible while holding one’s own serve seems as the best route to success.

# References

CNN. (2010, March 30). *The six seasons of the tennis year* . Retrieved from CNN: http://www.cnn.com/2010/SPORT/tennis/03/30/tennis.six.grand.slams/index.html

Hadlich, G. (n.d.). *How Much Do Tennis Players Make? – The Ugly Truth*. Retrieved from My Tennis HQ: https://mytennishq.com/how-much-do-tennis-players-make-the-ugly-truth/

Lindsey, C. (2004, April). *Follow the Bouncing Ball*. Retrieved from Tennis Industry Magazine: http://www.tennisindustrymag.com/articles/2004/04/follow\_the\_bouncing\_ball.html

Sackmann, J. (2023, 12 1). *tennis\_atp*. Retrieved from Github: https://github.com/JeffSackmann/tennis\_atp/tree/master