Predicting ATP Match Winners using Demographics

Sports betting has transformed the way viewers have consumed sports, with many fans worrying about player statistics and who will win certain matches more than enjoying the sport itself. With fans betting on who will win certain matches and certain tournaments, predictions and probabilities being calculated have rapidly increased and have called for new models to be made. The ATP, the men's professional tour has also seen this uptick in sports betting, with many fans betting on who wins certain tennis tournaments, especially the bigger ones. However, tennis is an individual sport and there are many factors affecting each player in each match. With so many moving parts in each match, accurately predicting the winner of each match is difficult yet crucial for these sports bettors. To try and figure out the demographics that matter the most and can help determine the winner of each match, I created visuals and models. These visuals and models help predict which factors should be tested for and used in a prediction of the winner. All of these findings help answer the question of who will win each tennis match?

To answer this question, this paper will describe three crucial aspects of the research and training I did. First, I will go over the dataset that I used to assist in creating the visuals and models I made. Understanding the underlying data is crucial in understanding what the models are built off of and how we can interpret them. Second, I will go over the visuals that I created to see what demographics interact with each other and how they ultimately impact the ability to predict which player will win specific matches. Finally, I will go over the models I built and trained in order to predict the winner of tennis matches. I used three different types of models, each with different accuracies and aspects that matter. Ultimately, this paper will examine the relationship between player demographics and their ability to predict who wins certain tennis matches on the ATP circuit.

As a brief introduction to the findings of my models and the answer to this paper, my models are slightly accurate in predicting the winner of tennis matches. Using factors like height,

country, hand, and age, the model was slightly accurate in predicting the winner. However, there is still room for improvement in terms of accuracy, meaning that while there is a slight indication that these factors predict the winner, there are still many unaccounted for factors that help predict the winner more accurately. As expected, height had an impact on these models and the chance of correctly predicting the winner. More surprisingly, the country also had a large impact on these models. However, there is still some inaccuracy, showing that these models can only be used as slightly accurate predictors. The rest of this paper will describe how we came to these conclusions, why they are not perfect indicators and the most accurate predictors, and how to interpret these findings for use in real world cases.

Dataset

In order to answer the question of who will win a certain tennis match, analyzing a good dataset is crucial. The best dataset would provide not only the results of as many tennis matches as possible, but it would provide the demographics and statistics of those matches as well. The dataset I used perfectly fits this description, with the result of every ATP match in 2023 along with statistics like aces, double faults, and points won, as well as demographics like the players heights, ages, and hands. The dataset includes every tournament for the 2023 season, every match played at those tournaments, and the accompanying results and statistics necessary. This dataset allows for the visuals to show important relationships for the winning players as well as matches throughout the season. It also allows for the model to use sufficient and accurate data to train and test for, theoretically allowing for accurate predictions. While the predictions were not perfectly accurate, as I will describe later, the type of data allows for clean and organized analysis of the data set. Overally, the dataset was the perfect match for the ability to answer the question of who will win certain ATP matches.

Exploratory Data Analysis

The exploratory data analysis, or my visualizations, allow me to see high level trends and fundamental relationships within the data. These visuals and the ability to see these relationships allow me to create hypotheses for what variable should be trained with the model, which is crucial for figuring out the most accurate model. I created five visuals to see these trends and relationships. For each visual I will go over why I decided for this visual and what the visual means in context.

For the first visual, I created a barplot comparing average height of tennis players for each country. When determining the best athletes in their respective sports, height is a main characteristic that many look for. Height is especially crucial in tennis, although for different reasons than other sports. In tennis, height is needed for the serve, groundstrokes, and for overall athleticism. The taller the player, the easier it is to serve harder, the easier it is to hit higher balls, and the easier it is to run with longer strides. It is then important to see the countries that produce the tallest players, ultimately meaning that these countries are more athletic and have a higher chance of winning. I created a barplot to analyze this relationship because barplots allow me to quickly see which bar is highest and which country has that bar. In terms of what this visual means, we can clearly see that Poland, Kazakhstan, Croatia and Greece are among the countries with the tallest players. This means that when looking at matches and trying to figure out which player has a higher chance of winning, players from these countries should have a higher chance of winning. While it is not a direct relationship, height and country are important factors to consider, especially for models that factor in country and height when comparing two players.

For the second visual, I created a scatterplot of winner's heights and their respective number of aces. In tennis, the serve is the one shot players can control. This means it is very important to get off to a good start and have a chance of winning. Building on the previous

visual, I wanted to see how heights relate to the chances of having more aces and therefore having a higher chance of winning. I used a scatter plot in order to see the trend in the data and see whether it is positive or negative. Looking at the second visual, it is clear that there is a positive correlation between height and number of aces. This means that taller players have a higher chance of hitting more aces, giving them a higher chance of winning. When putting this in context, we can use this variable and help the model predict who is the winner based on height. The model will know that when comparing two players, height is something to consider when factoring serves and chances of winning. However, as I will show later, more height does not necessarily mean more wins, which the model will need to train for.

For the third visual, I created a barplot for the number of top 100 ATP players per country. Knowing what country players come from is important, as each country has its own infrastructure and facilities that allow for better tennis training and tournaments. Therefore, by looking at which countries have the most top 100 ATP players, we can determine which countries have the best tennis facilities and training infrastructure. Then, we can use this knowledge for the models to determine which countries have a higher chance of winning when comparing two countries and their players. I created a barplot to see this visual because a bar plot allows me to quickly see and compare which countries have the most top players. This type of visual also allows me to quickly see which country has the most versus the least. Looking at the visual, it is clear that France, USA, and Spain have the most top 100 players. This means that they have some of the best tennis infrastructure to train their players. When using countries to predict winners, the model will be able to use country as a predictor, as there is a clear relationship between top players, wins, and country.

For the fourth visual, I created a line graph of height and number of top 100 ATP players. Once again building upon the previous visuals, seeing the relationship between height and their rank is crucial to see whether height indicates a higher chance of winning. In theory, taller players are more athletic and should win more. I used a line graph for this visual because it

makes it clear which height has a peak and has the highest number of top players. Looking at the visual, it is clear that being within 186 to 190 cm is the best height, as there are more top 100 ATP players at that range than any other range. These results are different from what would happen in theory. There is a clear sweet spot for the best height, as being too short is not the best and being too tall is not the best either. The 186 to 190 cm range has the most, presumably because it is a combination of both good serving, as shown in the other visual, and overall athleticism. The model will be able to compare player heights and predict based on the better height.

For the final visual, I created a scatter plot of winner height versus total serve points won. Finding the relationship between height and serve points won is crucial to understand weather height is the perfect indicator of success and winning. This visual helped answer whether height has a direct relationship with winning, and the taller a player is, the better. While the other visuals indicated that, this visual provides a direct insight into whether it related to serve points rather than a specific shot. I used a scatter plot because it shows the direction in which the data is trending. Looking at the fifth and final visual, there is no clear correlation or trend in the data. While this may be discouraging, it is very important for understanding why the model is not perfectly accurate. The scatterplot being noisy shows that a player's height is not a perfect indicator of winning, which means the model cannot be perfectly accurate since the indicator itself is not accurate. This means that the model must be combined with other factors to help make height more accurate, which I will describe later. Ultimately, this visual shows that there are other factors needed to predict winners, and that height can only be used in conjunction with these other factors.

Models

Building upon the EDA, I created 3 models to determine the accuracy in which these variables can be used to predict the winner. In terms of my results, I found that all the models were more accurate than a coin flip, but were not perfectly accurate. This means they can only be used as an indicator rather than a perfect prediction that bettors could use. I will now detail the three models, which factors impacted the models the most, and the accuracy of those three models. This ultimately will lead to my interpretation of these findings.

For the first model, I built a logistic regression model that used age, height, hand, country, and form as predictors for a player's chance to win against another player. Form was used as a predictor by calculating the number of wins a player had in their recent five matches. I built a logistic regression model as a baseline, as it is simple, quick, and has a probabilistic output, meaning it is not simply binary. Logistic regression can take into account numeric inputs and compare two players to each other. In terms of what I did to build the model, I cleaned the data by filtering out unnecessary rows and empty values, created categories for each variable that is used as a predictor, created two players that the model will compare to each other, calculated the difference between the variables for each player, and then created a training and testing split. These steps were done initially so that each model I created will be able to use these variables. The logistic regression model then trained and evaluated the data and the accuracy, ultimately resulting in an accuracy of 61.5%. This accuracy means that the model was more accurate than a 50/50 coin flip, but not accurate enough to be perfect and correct all the time. Logistic regression allowed for the two players to be compared to each other in a quantitative sense, especially with form being one of the variables. I then created a graph of which variable impacted the model the most, and the result was that form was the most important factor. This makes sense, as form is the most numerical category and has real world logic as well. Logistic regression is more accurate and influenced by numerical factors, which form is. In real world contexts, players with good momentum and wins recently have a higher chance of winning, so the model makes sense. In terms of false positives, I created a confusion

matrix to see how many times this occurred. The result was that there were around 310 perfect predictions and 190 false positives. This means that the model is slightly accurate, but cannot be completely accurate since there are enough times that it was a win but was not. Ultimately, the logistic regression model shows that these demographics and variables can only be used as an indicator and not a perfect predictor.

For the second model, I built a random forest model that used the same inputs and predictors as the logistic regression model. The random forest model was based on the same data and variables, and ultimately gave an accuracy of 59.5%. This essentially means the same thing as logistic regression, in the sense that it is more accurate than a coin flip but cannot be used as a perfect prediction for things like sports betting. Random forest is better than logistic regression for my type of dataset because it is best with structured data and non linear data, which my variables are. It also handles categorical and numerical variables better, so my country's variables have more of an impact as well. However, because countries and heights are not the best predictors, as the visuals showed, the random forest model that took these into account was not as accurate either. This is proven in the graph I created for the most important variable that affects the random forest model, as country was the most impactful variable. Because it is categorical, random forest was better at using it to predict the winner. But because this variable itself is not always the most accurate, the model is not either. This idea shows up again in the confusion matrix, with 200 false positives and 300 correct predictions. This means that the random forest model in theory should be more accurate, but in reality was influenced by variables that are not perfectly predictive. Ultimately, the random forest model shows that these variables and demographics can only be used as indicators and not perfect predictors of the winner of ATP matches.

Finally, the third model I built was a decision tree model based on the same variables to predict the winner of ATP matches between two players. This model gave an accuracy of 60.8%, which is more accurate than the random forest but less accurate than the logistic

regression. This result makes sense, as decision tree models are similar to random forest models in that they take into account categorical and numerical variables and they can handle logical splits that are in my data. This means that when higher ranked players win, the model can pick that up and apply it to the test data. However, as mentioned before, because countries and other categorical variables are not perfect predictors, this model is less accurate than the logistic regression model. The decision tree model is more accurate than a coin flip and can be used as an indicator, but not as a perfect predictor. In terms of the most influential variable, form followed by country are the most important. This makes sense, as this model took into account form, making it more accurate than random forest, but also took into account country, making it less accurate than logistic regression. In terms of the confusion matrix, it was similar to the others, with 170 false positives and 330 correct predictions. This means it is slightly accurate and can be trusted as an indicator of who could win matches, but should not be depended upon. In conclusion, the three models take into account different ways of training the data, but all result in around 60% accuracy, allowing us to use these models as indicators that are slightly accurate.

Interpretations and Conclusion

In conclusion, I found that age, height, country, and form are slight predictors of the winner of a match. With the visuals showing a clear relationship between each of the variables tested and the models being around 60% accurate, these variables can be used to slightly predict the winner of a tennis match. Essentially, when determining who would win between two players, these factors paint part of the picture, but not the whole picture. The three models all took into account different factors, but ended up having around the same accuracy.

These conclusions are important in the context of tennis, as it shows how complex the sport is and how difficult betting on it is. Tennis being an individual sport makes it unique in that

every match has multiple factors affecting the result, and that the player only has a few things in his control. This means it is difficult to use these statistics and demographics to predict the winner of certain matches. This is evident, with many millions of dollars being spent on models to train on tennis data so that tournaments can create advanced analytics and sports betting apps can be transparent to users. With that being said, my findings throughout the visuals and models do prove that these demographics can be used as indicators that are slightly accurate. While they may not give people perfect strike rates on bets, they can give fans an idea of who could win and how demographics impact their chances of winning. Overall, these findings show how complex tennis is and how demographics and form can be used to give an idea of who could win a certain tennis match.