Group Project Final Report

Gray Team (Nick Carroll, Emmanuel Ruhamyankaka, Jiaxin Ying, Song Young Oh)

2022-12-04

Abstract

This project analyzed Tennis Abstract's men's ATP tour-level dataset to understand the ability to predict a player's rank and the factors that impact an upset with respect to men's tennis. The main findings are: 1) a player's attributes (i.e. height, age, etc.) alone are not sufficient to predict a player's rank, more analysis is required to understand how to quantify and analyze a player's skill and how it impacts his rank, and 2) tournament type, match surface, and age are the primary predictors of an upset. The model used to analyze an upset had a 95% accuracy.

Introduction

Within recent decades sports have become a major use case of data analytics, and tennis is no exception. Furthermore, in 2021, more than half of states began the process of legalizing sports betting, whereas, until recently, sports betting was only allowed in select regions. This has exploded an industry where there are large financial implications to the ability to predict a sports performance. For this analysis, the dataset of men's ATP tour-level matches was studied for the 2021 season. This dataset has been uploaded by Jeff Sackmann manager of Tennis Abstract[1]. The dataset contains a sample size of 2,733 tennis matches and includes 49 variables which cover both continuous and categorical variables. It consists of the basic information for each match. The data can be grouped into two parts: data that describes each match's attributes, and data that describes the results of the match, including the winner's and loser's attributes, and performance separately.

Specifically, the first part of the data includes the date of the match, along with a match-specific identifier for the tournament. There is also information describing the tournament level, which is broken down into five categories. Meanwhile, the second part of the data covers the player specific data, with a wide range of variables about the winner and loser, including their name, age, height, nationality, and handedness. The data also includes each player's ATP ranking information at the tournament date. Moreover, this portion of the data describes the performances of the winner and loser, such as each player's number of first serves made, and break points saved during the match.

With this data, the main goal of this analysis is to answer the following two research questions:

1) Can a tennis player's rank be predicted by his attributes, specifically: height, age, nationality, and handedness?
2) Which factors have the greatest impact on the likelihood of an upset? (Throughout this analysis, an upset is defined as when the winner held a rank of at least 10 ranks lower than the loser.)

Methods

Research Question 1

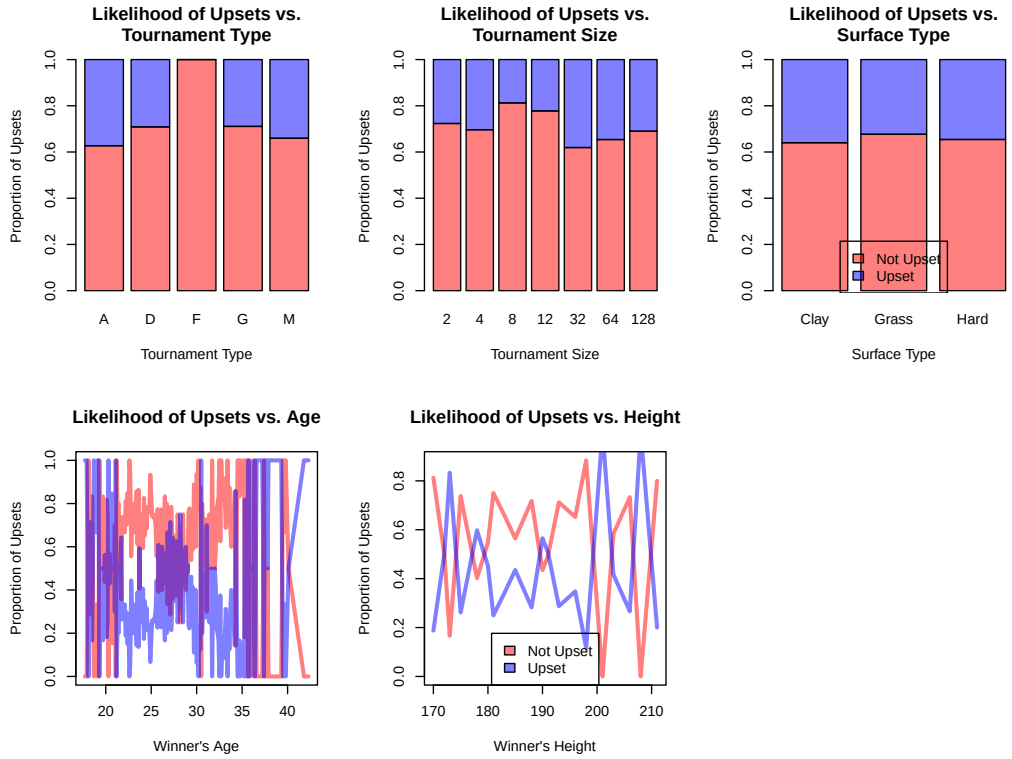
To analyze the first research question, a multiple linear regression model was used to predict the outcome of a tennis player's rank based upon his attributes. Although rank is not technically a continuous variable, it can be estimated as a continuous variable on this scale, because rank has at least 200 possible options in this dataset. Furthermore, while rank is the desired outcome, this analysis actually seeks to predict average rank over the course of the season. While rank changes over the season in accordance with performance, a player's attributes are (approximately) static over the season, so it makes more sense to predict average rank. In this scenario, average rank is actually a continuous variable because it can contain any degree of decimal values.

This dataset includes 49 variables; however many of the variables are not appropriate for the focus of this analysis. Therefore, all variables have been chosen a priori based upon domain knowledge of the drivers of a player's performance in sports. Since the first question's analysis focuses on predicting a tennis player's average rank, only factors which are player specific will be considered. Specifically, variables which are tournament-specific (i.e. tournament name) or match-specific (i.e. round or serve points) were included in the model. In professional sports, high performers usually have unique attributes compared with the rest of the normal population, including specific height ranges, age ranges, and regional and socio-economic backgrounds. Handedness is an attribute with particular importance in tennis, because it is common for one player to have a dominant swing known as a "forehand". Due to these expectations, the variables selected for the linear regression model to predict a player's average rank is height, age, nationality, and handedness.

Research Question 2

To analyze the second research question, a multiple logistic regression model was used to evaluate the probability of an upset occurring. As previously mentioned, an upset is defined as the winner's rank being at least 10 ranks lower than the loser's. Due to the binary nature of an upset and the multiple potential predictor variables, a multiple logistic regression is the appropriate model for this analysis.

Similarly, for the second research question, most of the variables in the dataset are not appropriate for this analysis; and again, predictor variables are chosen a priori based upon expectations from sports knowledge. Variables that are not appropriate for this analysis include match statistics (because this information would not be available prior to a match, and would be collinear with the outcome of the match) and tournament information that is not play-specific (i.e. tournament id and name). Conversely, based upon the exploratory data analysis, variables that are of particular interest are Tournament Type and Size, Surface Type, and the winner's attributes (i.e. height). The player's attributes (height, age, etc.) are chosen to understand how the player can impact the likelihood of an upset. The tournament/match information (surface, size, etc.) can provide inference information into how external factors impact the likelihood of an upset. Finally, tournament type can provide inference information into how a player's mentality can change in different match situations. Below plots show the likelihood of an upset against different predictors.



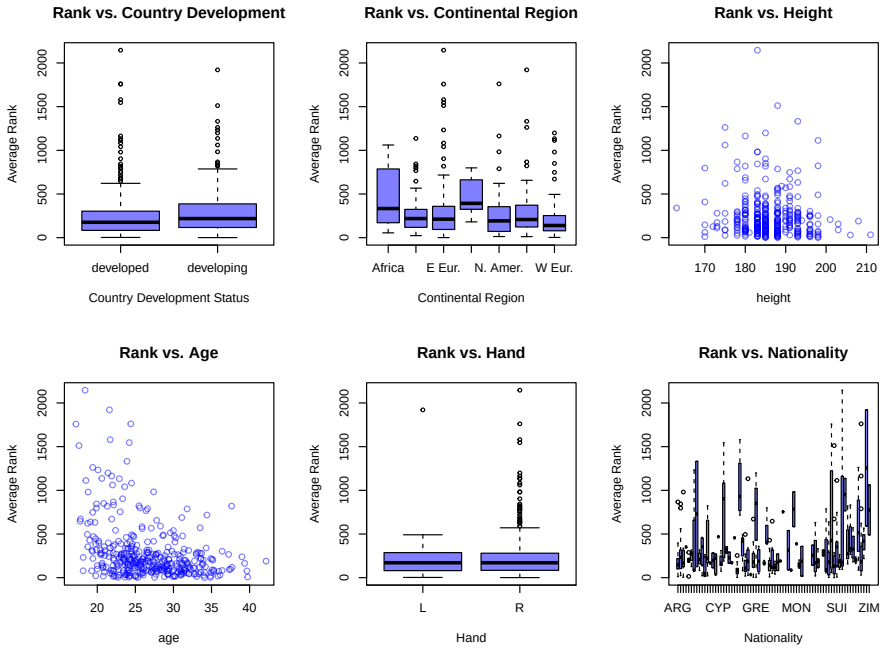
Finally, there are a total of 5,570 missing values in the original dataset, and 65 missing values exist in the subset used for selecting each player's statistics. This analysis assumes that the missing data is random, and safe to be excluded from the dataset. Furthermore, since the exploratory data analysis suggests weak correlations in the research questions, with substantial noise, data imputation methods would bias the analysis to suggest there are greater correlations than the data actually suggests.

To analyze the effectiveness of the model, the r-squared value was reviewed, and the predictions were compared to the actual values. After understanding the effectiveness of the model, the model's assumptions are assessed to ensure the model is appropriate. The model's validity is based upon its adherence to the linear regression assumptions: 1) that there is a linear relationship between the predictors and a player's rank, 2) that the players are independent of each other, 3) that the error in the model is approximately normal, and 4) that the residuals do not have a pattern (i.e. growing with their variables).

Results

Research Question 1

To explore the relationship between a player's rank and his natural characteristics (height, age, nationality, and handedness), scatter plots were reviewed showing the player's rank against each of his characteristics. The scatter plots did not show a clear correlation between a player's rank and characteristics.



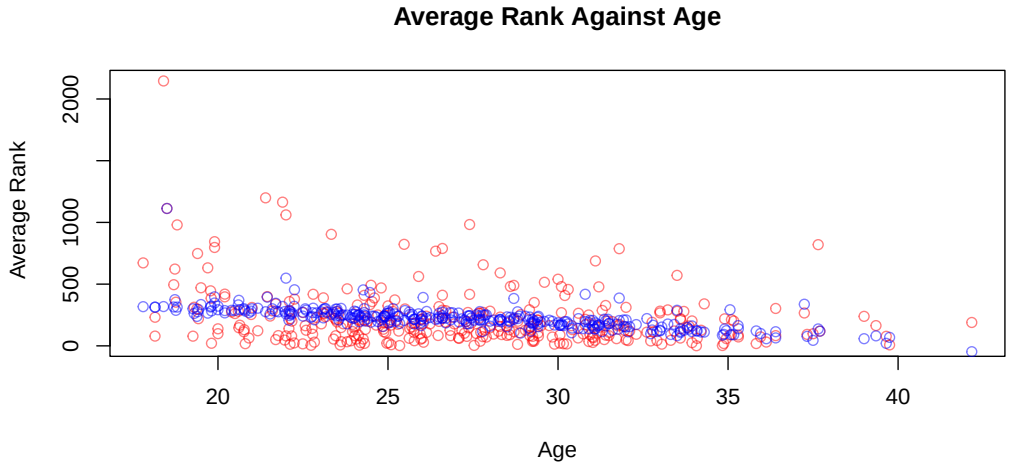
Additionally, while the exploratory plots didn't show a clear correlation between nationality and rank, there were some nationalities that clearly tend to have higher ranks than others. It is assumed that some nationalities tend to perform better than others because some countries have better access to certain sports than others, due to underlying socio-economic and cultural reasons. For example, countries with strong economies might invest more in athletic training, have better facilities/courts, and invest in access for sports where there is greater demand. To explore simpler relationships between nationality and rank, the 72 countries in the dataset were grouped into subsets. The first subset split the nationalities into "developed" and "developing" countries. A "developing" country has a lower GDP than developed countries and a less mature economy. Conversely, a "developed" country has a higher GDP and a more mature economy. For the purposes of this analysis, the "developed" countries are listed by following the country classification of UN[2] and all other countries are considered "developing". Additionally, the subset of developing and developed countries can be compared with the subset of countries by region[3]. However, with these subsets, exploratory analysis did not show a clear correlation between rank and grouped nationalities.

The coefficients of the linear regression model used to predict a player's rank, along with the statistical analysis, are shown in the table below. From the table below, handedness is not a statistically significant predictor variable, and region is only statistically significant if you are from Africa. According to the model, if you are from Africa, you are expected to have a rank over 200 higher than a player who has otherwise similar attributes. Height and Age are the attributes that are statistically significant. For every year older a player is, he is expected to reduce his average rank by approximately 11. Similarly, for every cm increase in height, a player is expected to be able to reduce his average rank by almost 4. Despite the fact that these two attributes are statistically significant, the R-squared value is less than 15%, meaning that the model can only account for less than 15% of the error in the data. When comparing the predicted average rank values to the actual average rank values, on average, the predictions are only within 277% of the actual values.

Below is a scatter plot of average rank against age, showing the predicted values overlaid with the actual values. The predicted values clearly do not accurately account for the vast amount of variability in the players' average ranks.

Table 1: Predicting Players' Average Ranks

	Estimate	Standard Error	Statistic	p-value
Height	-3.8556	1.9081	-2.0207	0.0442
Age	-11.3301	2.5613	-4.4237	<.0001
Hand: Right	——-	——-	——-	——-
Hand: Left	-37.9741	34.7338	-1.0933	0.2751
Region: North America	——-	——-	——-	——-
Region: Africa	211.1458	85.2622	2.4764	0.0138
Region: Asia	-3.7804	58.0921	-0.0651	0.9482
Region: Australia	16.7798	58.0895	0.2889	0.7729
Region: Europe	-28.9559	38.3149	-0.7557	0.4504
Region: Middle East	139.1449	115.1752	1.2081	0.2279
Region: South America	-3.2483	49.5387	-0.0656	0.9478

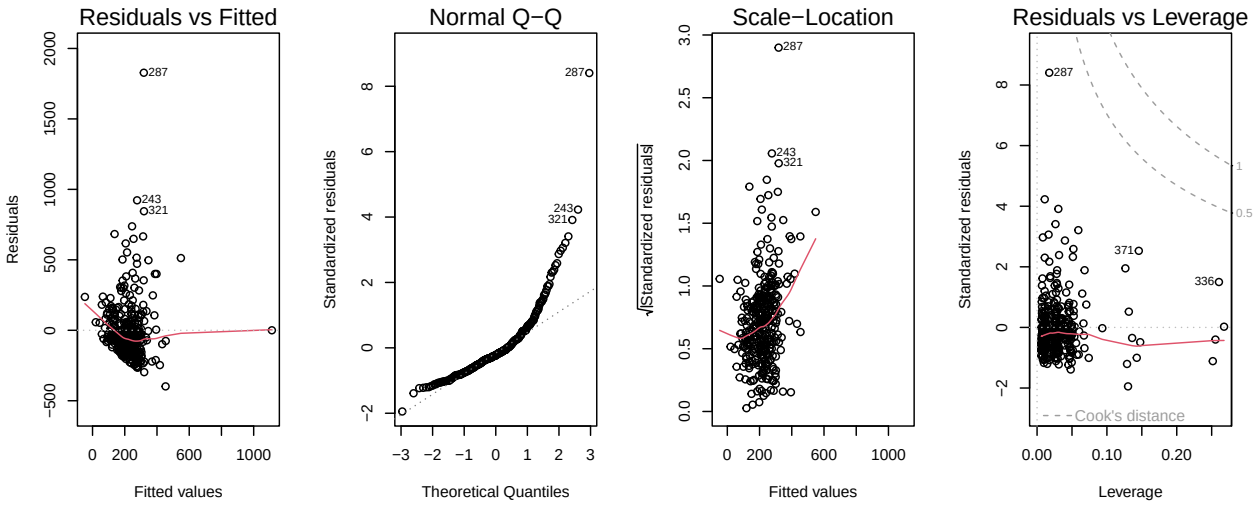


Since the model is not a good predictor of a player's average rank, it is particularly important to assess the model's validity and review the linear regression assumptions: 1) that there is a linear relationship between the predictors and a player's rank, 2) that the players are independent of each other, 3) that the error in the model is approximately normal, and 4) that the residuals do not have a pattern (i.e. growing with their variables).

First, as previously discussed, the scatter plots did not show a clear linear relationship between the predictors and a player's rank, which means this assumption may be violated. Second, while the players should largely be independent of each other, they are in competition with each other for lower ranks, and therefore this assumption may also be violated. Third, when reviewing the normal Q-Q plot below there does appear to be some deviation from normality. Fourth, while the residuals do not have a clear pattern, they do appear to be more positive than negative, suggesting their may be some deviation from this assumption. The plots below also show that while there are some outliers in the data, there are no high leverage points impacting the model.

Overall, it would seem that a linear regression model is not an excellent fit for the data; nevertheless, linear regression is generally robust to these assumptions and there is nothing suggesting that an alternative model would provide better performance. This would suggest that it is not possible to

reasonably predict a player's average rank based on their attributes; and therefore, learned skills are the primary driver of a player's success in tennis.



Finally, to assess multicollinearity in the model, the variance inflation factor was analyzed for each of the predictors in the model as shown in the table below. From the table, there are no multicollinearity concerns in the model.

Table 2: Checking Multicollinearity of Linear Regression Model

	VIF	Degrees of Freedom	Scaled VIF
Height	1.0569	1	1.0281
Age	1.0421	1	1.0208
Hand	1.0525	2	1.0129
Region	1.0826	6	1.0066

Research Question 2

For the second research question, similar exploratory analysis was reviewed using bar plots to review the factors. Specifically, bar plots were expected to show a relationship between the comparative number of upsets and the categorical variables, such as tournament type and surface types. The exploratory analysis did not establish a clear relationship between any of the categorical variables and the likelihood of an upset. There did appear to be more upsets in the older and younger players; however, this is also potentially because these age ranges are only more likely to compete at the professional level, if they are particularly skilled. Furthermore, there also appeared to be a potential relationship between upsets and height in the extremes, but, similarly, extreme heights also have small sample sizes.

The coefficients of the model used to infer the factors that lead to an upset, along with the statistical analysis, are shown in the table below. Based on the table, the model suggests that the most important factors that lead to an upset are the tournament type and the surface. If the tournament is a Master's type tournament, the odds of an upset are reduced by approximately 1,595 times. If the match is played on grass, as opposed to a hard surface, or a clay surface, the odds of an upset increases by approximately 30 times. The next most important factor that leads to an upset is the lower-ranked player's age. Every year the player's age increases, the odds of an upset increase by

approximately 25%. The model also shows a positive odds correlation with the winner's seed and rank, but this likely has to do with the fact that there is more potential for an upset with a lower ranked player. For example, a player with rank 9 can never be in position to have an upset.

Table 3: Predicting Probability of Upset

	OR	Log Odds Est.	S.E.	Statistic	P-value
Height (Underdog)	0.8661	-0.1437	0.1274	-1.1285	0.2591
Age (Underdog)	1.5244	0.4216	0.1896	2.2237	0.0262
Tournament Type: Standard	——-	——-	——-	——-	——-
Tournament Type: Grand Slam	0.0057	-5.1621	5.9712	-0.8645	0.3873
Tournament Type: Master's	0.0001	-8.8214	3.9712	-2.2214	0.0263
Draw Size	1.0653	0.0633	0.0495	1.2796	0.2007
Match Surface: Hard	——-	——-	——-	——-	——-
Match Surface: Clay	11.8366	2.4712	1.9963	1.2379	0.2158
Match Surface: Grass	0.5593	-0.5811	4.5279	-0.1283	0.8979
Tournament Date	1.0024	0.0024	0.0028	0.8496	0.3956
Seed (Underdog)	2.0563	0.7209	0.2648	2.7221	0.0065
Hand (Underdog): Right	——-	——-	——-	——-	——-
Hand (Underdog): Left	0.0799	-2.5274	2.2942	-1.1016	0.2706
Hand (Favorite): Left	——-	——-	——-	——-	——-
Hand (Favorite): Right	4.5549	1.5162	2.9639	0.5115	0.609
Height (Favorite)	1.1838	0.1687	0.1212	1.392	0.1639
Age (Favorite)	1.0749	0.0722	0.1318	0.5482	0.5836
Round: of 64	——-	——-	——-	——-	——-
Round: Quarter Finals	190.2806	5.2485	2.808	1.8691	0.0616
Round: of 16	51.4032	3.9397	2.7198	1.4485	0.1475
Round: of 32	29.4149	3.3815	4.1418	0.8164	0.4143
Round: Round Robin	0.7153	-0.335	7.8848	-0.0425	0.9661
Round: Semi-Finals	1903.9767	7.5517	3.4988	2.1584	0.0309
Rank (Underdog)	1.2323	0.2089	0.0807	2.5872	0.0097

To assess the accuracy of the model, a confusion matrix is shown below. The accuracy of the model is 95%, and it is clear from the confusion matrix that the model is very strong at predicting upsets. To confirm the validity of the accuracy, a cross validation analysis was run with five folds, and the average accuracy of the analysis was found to be 88%. Therefore, it is reasonable to suggest that the identified factors do indeed impact the likelihood of an upset.

Table 4: Confusion Matrix for Multiple Logistic Regression Model

	FALSE	TRUE
FALSE	97	1
TRUE	3	45

Conclusion

After analyzing the data of the men's ATP tour-level matches in 2021, this analysis has shown that the data is insufficient to predict a player's average rank for a season based upon that player's attributes, but that the data is sufficient to understand the factors that impact an upset. There is substantial variability in a player's rank that cannot be accounted for by the player's attributes alone. To better understand how to predict a player's rank, there are two key factors that are unaccounted for in this analysis. First, this analysis has been based upon a player's average rank for the season, but a season lasts several months and can have unexpected events, like injuries, which are not accounted for in this dataset. To better understand a player's rank, the rank must be analyzed over the course of a season and identify reasons other than a player's attributes that impact rank. Second, this analysis has only regressed a player's average rank based upon the player's attributes, and not upon the player's learned skills. It is reasonable to expect that a player's skills are more impactful than their height and age on their performance. Skills such as serve speed, lateral speed, and reaction time, among others, are all potential drivers of performance that are not captured in this dataset.

After analyzing the factors that impact the likelihood of an upset, it has been shown that tournament type, match surface, and age are all strong indicators of the likelihood of an upset. While the accuracy has been shown to be extremely high with the logistic model from this dataset, it is still possible that there are underlying confounding variables which impact the likelihood of an upset. While age can be an indicator of maturity, it would be reasonable to believe that there is a mental component to an upset that is not accurately captured in this dataset. Furthermore, tournament type and match surface are both external variables, and this analysis would suggest that a player has little ability to impact the likelihood of an upset on his own. This is hopefully false, as a player would want to believe he can train to prepare for an upset. Similar to research question 1, it would be beneficial to study how a player's skills or "recent" performance can impact the likelihood of an upset.

References

[1] Jeff Sackmann: Tennis Abstract. https://github.com/JeffSackmann/tennis_atp/blob/master/atp_matches_2021.csv. Accessed: 2022-10-01.

[2] UN: Country Classification. https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf. Accessed: 2022-11-01.

[3] Wikimedia: List of countries by regional classification. https://meta.wikimedia.org/wiki/List_of_countries_by_regional_classification. Accessed: 2022-11-01.