Research into factors affecting the performance of tennis players in ATP

You Peng

2022-02-23

Introduction

Tennis is a popular sport around the world, and one of the worldwide top-tier tennis tours for men is ATP tour organized by the Association of Tennis Professionals. Those professional tennis players who maintained a high ranking in ATP seem to have something in common. There are already research studying the Association between body height and serve speed in elite tennis players. For example, body height of the men explained 27% of the variance of fastest serve in a match (Vaverka & Cernosek, 2013). Serving speed may not directly contribute to a tennis player's ranking in tournaments, and there could be more potential factors affecting the performance of tennis players. Therefore, we raise our research question as how could some physical factors and experience of a tennis player be used to explain the variation observed in rankings of 639 players participated in 2017 ATP tournaments. Moreover, given the physical characteristics and experience of a player, how can we predict his ranking in a range of 2074 tennis players from ATP? By exploring the relationship between tennis players' physical characteristics, experience, and their performance, we could help coaches to better recognize if someone could be a good tennis player and inspire tennis players what they can improve on in order to making progress. The goal of this research is to find a model that is not overly complicated, but also having reasonable properties required to make good predictions.

Methods

The dataset we used is the ATP World Tour tennis data. "This dataset contains tennis data from the ATP World Tour website. The data contains ATP tournaments, rankings and player's overview. The latest available data is for 2017." (ATP World Tour, 2018) The dataset can be downloaded from the url: 'https://datahub.io/sports-data/atp-world-tour-tennis-data/datapackage.json' or website 'https://datahub.io/sports-data/atp-world-tour-tennis-data#r', we only require the 9th and 10th csv file from this json file. We provided two methods to read in the dataset. The first one gives the url path provided by the json file to read.csv(), which will takes a long time to download from the website. The second one requires the downloaded csv and zip file stored in the data directory, and then apply read_csv() directly on the path 'data/rankings_1973-2017_csv.csv' and 'player_overviews_unindexed_csv.csv' after unzipping the zip file "player_overviews_unindexed_csv.zip" in data directory.

The outcome variable and predictors were contained in two different datasets we loaded. The first dataset contains the ranking of tennis players every week from 1973 to 2017, which is a lot of data. And each player's ranking varies several times in one year since the ranking is updated per race in that year. In order to perform the most up-to-date analysis as well as to obtain a continuous outcome variable, I averaged the ranking of each tennis player over the whole year 2017, so that each tennis player has a corresponding average of his ranking in 2017. Similarly, I averaged the age and number of tournaments played in 2017 as well. This process is done by using filter() from dplyr to leave only those observations about 2017 in the dataset, and aggregate() to calculate the mean of rank, age, and tournaments played to be new variables in the dataset. Then we used merge() from dplyr by player_id to merge two dataset into one, and only kept the following variables after applying select() on the joint dataset:

"weight_kg": body weight in kg of the tennis player in 2017. "height_cm": body height in cm of the tennis player in 2017. "handedness": whether the tennis player is left-handed or right-handed. "average_age": the average age of the tennis player during the 2017 tournament period. For example, some players were 24 years old during the first few tournaments held in 2017 and grew to 25 years old later in the 2017. In those cases, their average age would become decimals. "average_tourneys_played": the average number of tourneys played during the 2017 period. "backhand": whether the player is using one-hand backhand or two-hand backhand. "average_rank": the average rank a tennis player achieved during the whole 2017 tournament period. This is a continuous variable since the rank is averaged over the whole period. This is our response variable.

The joint dataset now only contains predictors and outcome variables that I need, and I dropped all rows that contained null values or contained 0 height or weight since these observations were missing the predictor values we need. Note that I also dropped one observation because that tennis player had a height of 3cm. Another tennis player had a weight of 675kg, which was modified as 67.5 after the cleaning.

Then we conducted exploratory data analysis on this cleaned dataset. The exploratory data analysis provides a general idea about the distribution of the data we have as well as a note of anything odd such as skews or outliers. In the exploratory data analysis. We first took a look at the dimension of our dataset, as well as a summary of mean and standard deviation for variables in the dataset. Moreover, we generated boxplots and histograms for each of the numeric variables using boxplot() and hist(). For categorical variables, we produced bar plots for them by using geom_bar() in the ggplot2. In order to visualize the correlation between predictors and response variable, scatter plots between each of the predictors and response were produced by plot() functions. Boxplots for average ranks after grouped by categorical variables were also generated by plot() functions in order to visualize the effect of these categorical variables on the response. Moreover, a scatter plot for average ranking vs. height after grouping by handedness and backhand, as well as a smoothing linear line, is generated by geom_point() and geom_smooth() from ggplot2. A mean-deviation plot on handedness and average ranking and a mean-deviation plot on backhand and average ranking were also generated using stat_summary() in ggplot2. Lastly, we provided an interactive scatter plot showing ranking, height, and age of players from different countries using plot_ly from plotly library.

We continued our research by checking the violations of 4 assumptions needed for fitting a multiple linear regression model. We checked these four assumptions by generating the QQ plot using qqnorm() and qqline(), and generating residual plots by using resid(). If any assumption is violated, we will apply powerTransform() from car package, which is an automated power transformation, to transform response variable, then check if the transformed model satisfies the assumptions. We have to be careful not making the transformation too complicated.

Then we move to the stage of model comparison. We built another GAM model where we put a cubic regression spline on height using gam() from mgcv package. Then we compared the adjusted R square of these two models to see which one is better. We would prefer a model that has fairly large adjusted R2 with appropriate number of predictors. A plot for linear and non-linear associations between rankings and heights of athletes is also generated by using ggplot2.

We also trained two machine learning models to predict a player's rank in a range of 2074 tennis players from ATP. First of all, we split the data into training and testing (70-30%). By using randomForest(), random forest was the first model trained because usually it performs better than the single regression tree model and also better than bagging without feature selection. This is because that random forest randomly picked features used to predict in each bagging, so that it decrease the dependence between trees and decreased the variance of the prediction. Then we trained an xgboost model using train() in caret package and set "method" to be "xgbTree", as well as performing a grid search for tuning the number of trees and the maxium depth of the tree. We also performed a 10-fold cross-validation and determine the variable importance. We provided a plot of variable importance for both model, and calculated their test MSE. Then we fit the one with smaller test MSE on the full dataset and provide the RMSE of it.

Results

After cleaning and wrangling, there are 639 observations and 8 variables in the dataset, and the response variable is the average rank a tennis player achieved during the whole 2017 tournament period. Among all these 639 players, the last place has a rank of 2074, this will be discussed in the limitation. The summary statistics for five numerical variables are given below, we can see that the mean height for all players is 185.3, which is much higher than the average height of men between 20 to 39 in US, which is 176.1 in that case (Fryer, 2018). This may suggest that a higher height could take advantage in tennis.

Table 1: Summary statistics for numerical variables in the dataset

Variable	mean (s.d.) in dataset
`average rank`	555.877 (490.213)
`weight`	79.154 (6.783)
`height`	185.252 (6.631)
`average tourneys played`	17.624 (8.455)
`average age`	25.991 (4.522)

For numerical variables in dataset, their histograms indicate that the response variable is heavily right skewed, and the distribution of "average tourneys played" has a heavy tail on the left. These observations suggest that we may need to transform some variables later. And for binary variables in the dataset, there are only 80 left-hand players where majority being right-hand players, and 104 one-handed backhand players where majority being two-handed backhand players. This indicates that data are very imbalanced in both variables. Details of histograms can be found on the EDA page on the website: https://fredooooooo.github.io/JSC370-FinalProject/.

Among all scatter plots between response and predictors, we can see a clear linear pattern between average rank and average tourneys played. Since the slope is negative, this suggest that more tourneys a player has played, the better rank he is expected to have. The correlations in other plots are not very obvious to see, while the right-handed players may perform a little bit better than left-handed players since the median of them is a little bit larger. The details of other plots can be found on the EDA page on the website: https://fredooooooo.github.io/JSC370-FinalProject/

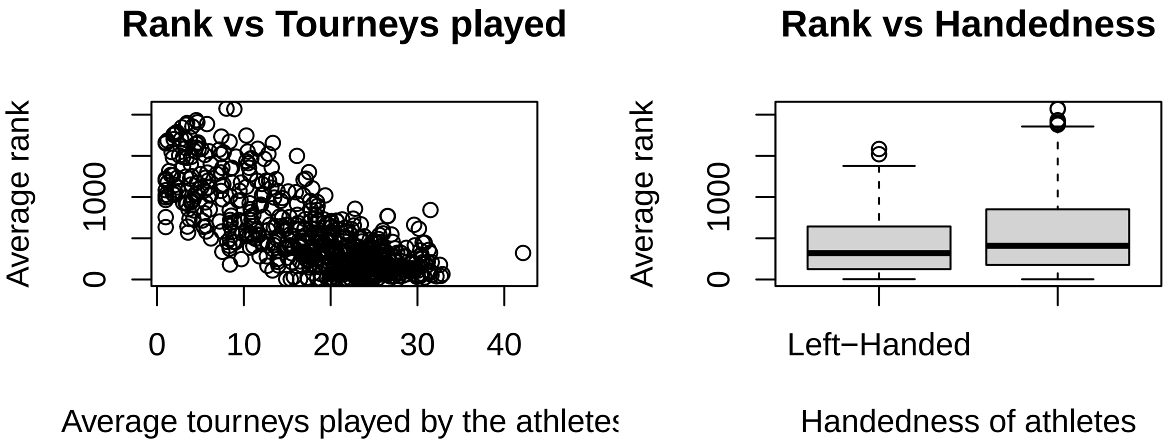

Figure 1: plots showing correlations between each predictor and the response variable

Boxplots for other numerical variables are also shown below. It's clear that there are some outliers in weight and height variables. Moreover, height and average tourneys played variables are left skewed since the first quantile deviates from median more. And the average rank is right skewed, which is consistent with the conclusion from histogram.

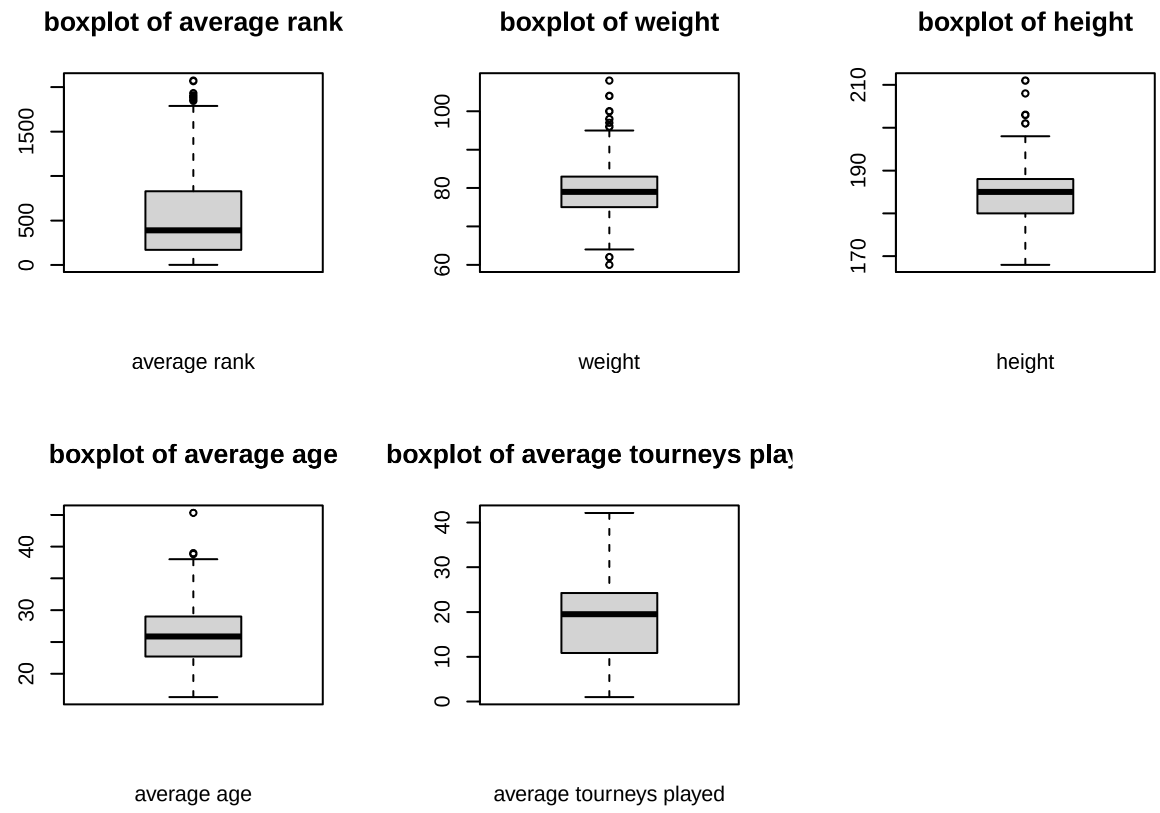

Figure 2: box plots for numerical variables

When we comparing average height and average rank in each category of handedness, we can see negative linear relations in both category, while the slope seems to be more negative in the right-handed group. This trend suggest that higher players are expected to have better rank, and a right handed player is more affected by height. This conclusion lines up with the hypothesis we had from the summary statistics. While when we comparing average height and average rank in each category of backhand, we see a interesting pattern that a negative linear relations lies in two-handed backhand category, while the slope become a slight positive one in the one-handed backhand group. This trend may suggest that the conclusion we had before saying that higher players are expected to have better rank is not suitable on one-handed backhand players. However, it's also possible that the slight positive trend is due to a lack of observations, since there are only 104 players in this category. Details of these two scatter plots can be found on the EDA page on the website: https://fredooooooo.github.io/JSC370-FinalProject/.

From the following Mean-std error plot, we can see that the mean rank of left-handed players is better, while the standard deviation is larger in right-handed players group. The means in two backhand categories are almost the same, while the standard deviation is larger in one-handed players group. This is interesting since one-handed players are the minority in the dataset, however, they varied a lot more than the two-handed backhand players.

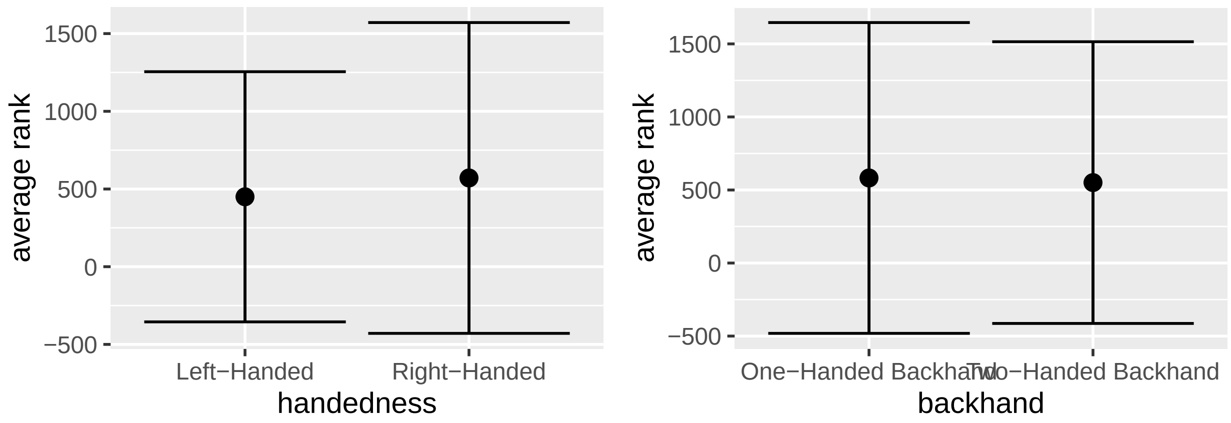

Figure 3: max-min plot for rank in different groups

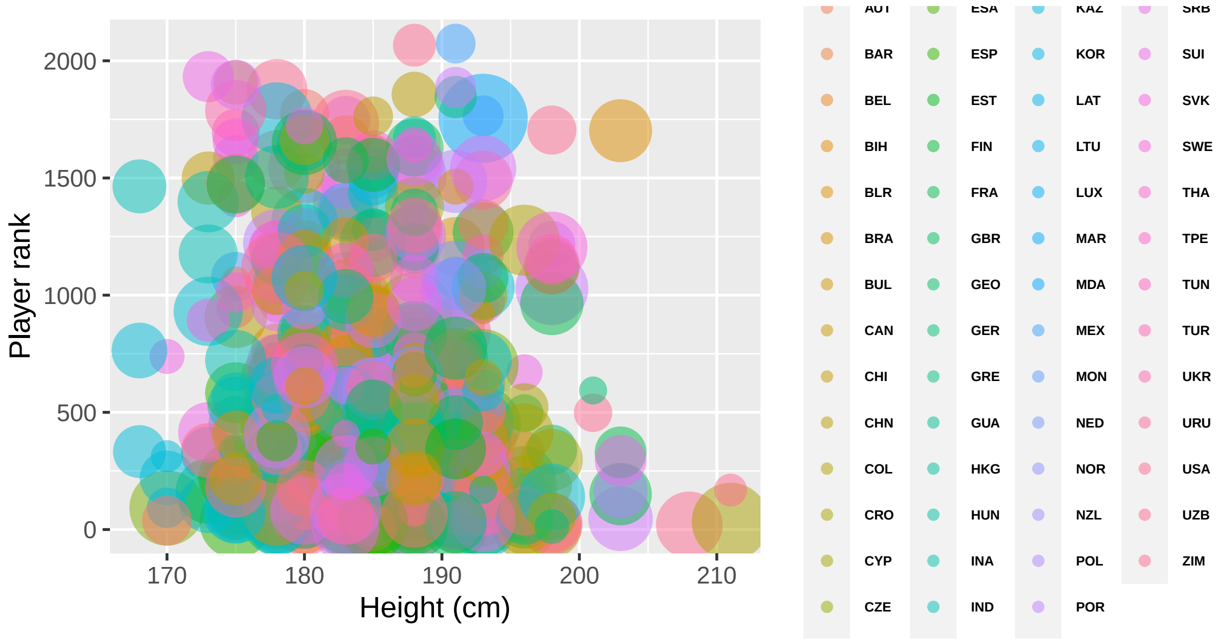

Figure 4: Scatterplot for height, age, and rank

To explore three numerical variables simultaneously, the above plot shows the relationship between athletes height and rank in different countries, with the size of the circle indicating the age of that athletes. For example, an athlete from MAR is significantly older than others. And we can also see that lots of older players had a better ranking in the bottom, this may suggest that older players with more experience would be more likely to obtain a better ranking. An interactive version of this plot is provided on the website: https://fredooooooo.github.io/JSC370-FinalProject/.

We first built a linear regression model taking account into all predictors. From the following QQ plot, we can see large deviations at the right end, which means that the normality assumption is not satisfied currently and we need to apply some transformations to our variables.

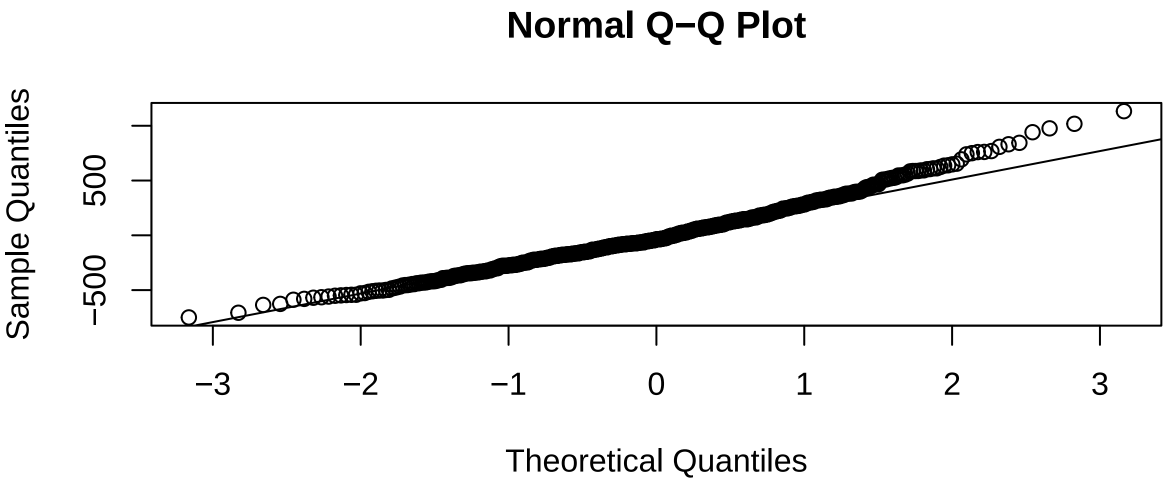

Figure 5: QQ plot for initial linear model

The automated transformation result in applying a power of 0.5 to average rank. Then we refit the model and here is the summary of the model. From this summary, we can see that the square root of rank is expected to be 85.83 if given all other predictors as 0. This interpretation will not be so useful since it's impossible for predictors such as height and weight to be 0. We can also see that the square root of rank is expected to decrease 0.96 given a one-unit increase in the average tourneys played while holding other variables in the model constant. Moreover, the square root of rank is expected to decrease 0.51 given a one-unit increase in the average age while holding other variables in the model constant, and the square root of rank is expected to decrease 0.17 given a one-unit increase in the average height while holding other variables in the model constant. From the p values of variables, we can see that only average tourneys played, age, and height are significant linear predictors in the model. And these three predictors all had a negative relationship with ranking meaning that a higher height and more experiences correspond to better ranking. We also obtained the R-squared to be 0.6436, which means that the linear model explains 64.36% of variance in Player's ranking. The adjusted R^2 is 0.6402, which is not bad. Moreover, the QQ plot and residual plots for the model after transformation is checked to see that the QQ plot looks better and the residual plots looks randomly scattered around 0.

	Estimate	Pr(>|t|)
intercept	85.8283558	0.0000000
weight(kg)	-0.0407366	0.4627676
height(cm)	-0.1652308	0.0033295
average tourneys played	-0.9631705	0.0000000
average age	-0.5108429	0.0000000
two-handed backhand	-0.7625122	0.2644566
right-handed	0.0662256	0.9298064

The following plot presents a linear regression line and gam regression line for average height and square root of rank. These two lines are very close which means that the relationship between height and square root of rank could be explained by linear regression well enough, and the gam regression don't need a non-linear line to explain this relationship. Moreover, we built the gam model where we put a cubic regression spline on average height. The adjusted R square for this gam model is 0.64, which is the same to the linear model.

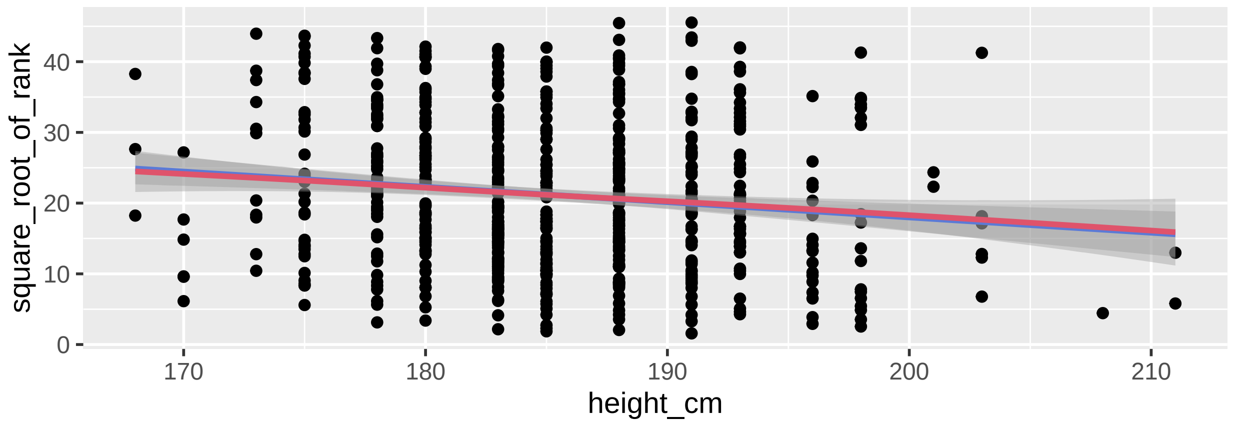

Figure 6: Scatterplot of height(cm) vs. square root of average rank

Another way to predict the response is using random forest. From the summary of the random forest model, we find that averaging across all 500 trees provides an OOB MSE being 81956.94. And their variable importance plot is shown below as well. We can see that the average tourneys played is the most important variable in predicting the ranking of the player. Average age and weight are in the second and third place, however, their importance is much less than the average tourneys played. The random forest model explained 66.17% of variance in train dataset. Lastly, we compute the test MSE of the random forest method to be 81997.48.

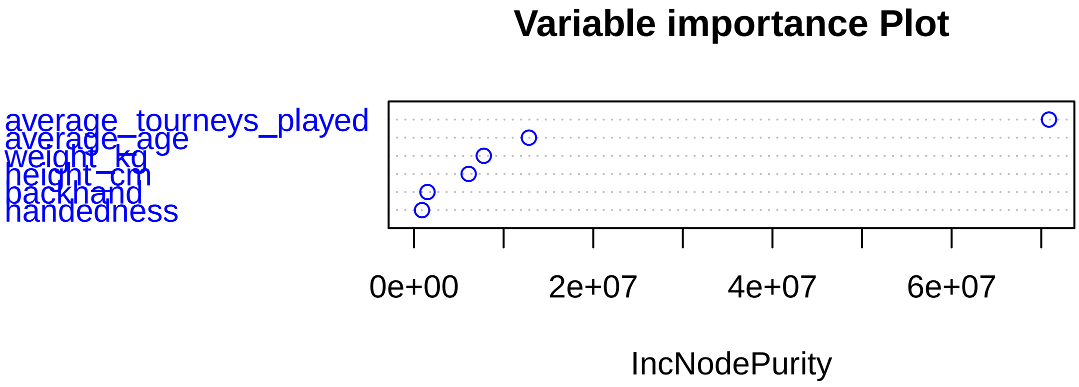

Figure 7: variable importance plot for random forest model

For the xgboost model, average tourneys played once again became the most important variable according to the following variable importance plot. Age is sill the second most important, and height became the third important one. And the test MSE is calculated as 79707.52. This is smaller than the random forest. This is because that xgboost tuned the number of trees and depth of trees well, and provide a better prediction through boosting. From the summary of bestTune, we can see that the number of trees is set to 50, learning rate is set to 0.2, and max_depth is set to 1.

Therefore, we fit the xgboost model with the whole dataset, the summary provides the RMSE being 278.0541. And the final values used for the model were number of trees being 150, max depth being 1, and learning

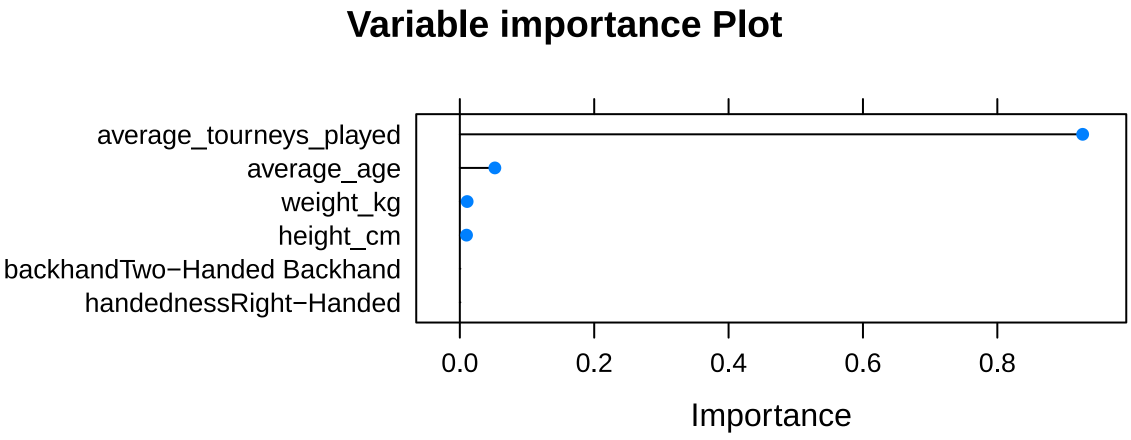

Figure 8: variable importance plot for Xgboost

rate being 0.1.

Conclusion

The linear model is telling that athletes are expected to have better rank if he played more tourneys while holding other variables the same, or he is older holding other variables the same, or he is higher holding other variables the same. Note that this model aligns with the study mentioned in introduction since height of the athletes could explain the variation of the serve speed which further explains the variation of ranking. The gam model has the same adjusted R square, which means that the linear model could fit the data well enough. And the gam model also suggest that a higher tennis athlete is expected to have a better ranking.

Given the physical characteristics and experience of a player, we can apply the Xgboost model trained with this dataset to predict his rank in a range of 2074 players from ATP. The Xgboost model achieved an RMSE of 278.0541, and shows that the average number of tourneys played by this player is the most important predictor used in predicting his rank.

The linear model has some limitations since the QQ plot still have some deviations at right end. The use case of our Xgboost model might also be limited to predict a rank among 2074 players, since the model is trained in a condition that 2074 is the last place, and a rank 5 in 10 players should be interpreted differently with rank 5 in 1000 players. Lastly, the research is conducted on the 2017 dataset since this is the latest dataset we found, the result concluded from this dataset may not be the case 10 years later.

References

ATP World Tour. (2018, October). ATP World Tour tennis data. Retrieved from DataHub: https://datahub.io/sports-data/atp-world-tour-tennis-data#r

Fryer CD, et al. (2018). Mean body weight, height, waist circumference, and body mass index among adults: United States, 1999–2000 through 2015–2016.

Peng, Y. (2020). Research into the performance of tennis players [Unpublished paper]. University of Toronto.

Vaverka, F., & Cernosek, M. (2013, March 12). Association between body height and serve speed in elite tennis players. Retrieved from Scholar Portal Journals:
https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/14763141/v12i0001/30_abbhassi etp.xml