Final Paper

STOR 320.02 Group 4

November 16, 2020

INTRODUCTION

Although baseball could once be considered “America’s pastime,” basketball has sharply risen in popularity since its invention in Massachusetts in 1891. Popularized by American servicemen abroad during World War II, international leagues such as the Euroleague and the Chinese Basketball (CBA) league have since flourished and produced many exceptional players. Basketball is a lucrative industry, proven by the NBA’s 2018-2019 revenue of $8.76 billion and the salaries of its highest players, ranging from $64.2 million to $134.9 million. With such talent and money at stake, it becomes incumbent on scouts, teams, and coaches to accurately assess players and make the wisest choices possible.

Just as trade routes for oil and resources between hemispheres have flourished in the last 50 years, so has player movement and trade between international leagues. While it is not uncommon to see a player switch from the NBA to CBA or Euroleague to NBA, performance of these players tends to vary greatly. The first part of our analysis looks specifically at players who change leagues, given that it is a well-accepted fact that competition, compensation, and playing level differs from league to league. We ask: can we predict a player’s change in performance as they move from one league to another? Essentially, we explore the influence of the league on player performance, based on prior players who have moved from League A to League B.

Naturally, green basketball players freshly graduated from high school or college would improve as years pass in their respective professional leagues. However, their performance necessarily cannot increase forever. According to Forbes, performance tends to peak in a player’s mid-20s, stays constant until his early 30s, then increasingly steeply declines. The legendary Kobe Bryant declined to an effective field goal percentage of 41% by the age of 35. However, we understand that age alone is likely not an accurate metric to understand the age of peak performance, as many other factors influence athlete health. Basketball experts and writers have often commented on the short careers of players over seven feet, which in theory could make them prone to injury and early career decline. We sought to analyze these factors and better answer our question by making a model predicting the age of peak performance for players of different builds, and hypothesized that tall, heavyset players may experience an early age of peak performance. To this end, we asked the question: can we predict a player’s age of peak performance based on their physical measurements, height and weight?

By exploring and analyzing these data, we can provide accurate information to those in the NBA and other leagues who want to recruit internationally. If we are able to predict a player’s future performance in their proposed league of destination, given their prior performance in another league, scouts and managers will be better able to assess their investments. Furthermore, managers could analyze their existing teams in order to better understand their team makeup based on their predicted age of peak performance. In the profitable industry surrounding basketball players, information is certainly equivalent to money and accurate performance analysis plays a vital role.

DATA

Our dataset consists of 44066 observations across 31 variables and includes personal details and statistics for every player in 49 basketball leagues from the year 1999 to 2020. The data was scraped from the website RealGm.com by Jacob Baruch and published on the database website Kaggle. RealGM.com is a well-known sports website that publishes news, raw data, and analyses for a variety of different sports including basketball.

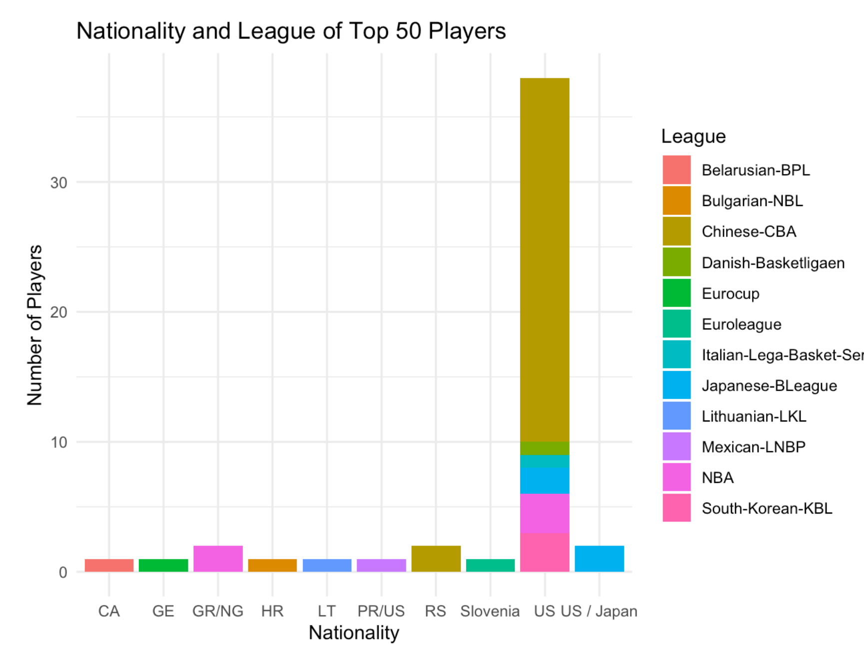
Since our original dataset was very large in size and scope, and we found ourselves asking many different and unrelated initial questions, we chose to narrow our focus by only focusing on several variables that we felt would yield an interesting analysis. These variables were “League,” “Player,” “Nationality,” “Season,” “height_cm,” and “weight.” The variables “height_cm,” and “weight_kg” were renamed to be “Height,” and “Weight,” respectively. These variables indicate a player’s recorded nationality, name, height, weight, and the league they played in for each season between the years 1999-2020.

To further simplify our analysis, as well as to yield results that would be more understandable and applicable to other people, we chose to combine several variables into a single metric of player performance. This variable, named “Performance,” was created by calculating two separate metrics: the player’s efficiency in scoring overall points and the player’s efficiency in making shots. First, the player’s total points (“PTS”) scored in each season was divided by the number of minutes (“MIN”) that they played in that season. Second, the sum of the variables “FTM,” “FGM,” and “TPM” was divided by the sum of the variables “FTA,” “FGA,” and “TPA” to produce the ratio of collective free throw, field goal, and three-point shots the player made to the same collective shots the player attempted. These two calculations were then multiplied together in order to relate them into one number. Only players who had played more than 5 games and more than 60 minutes in a season were included in this dataset in order to remove inexperienced players that might have abnormally high or low performance ratings.

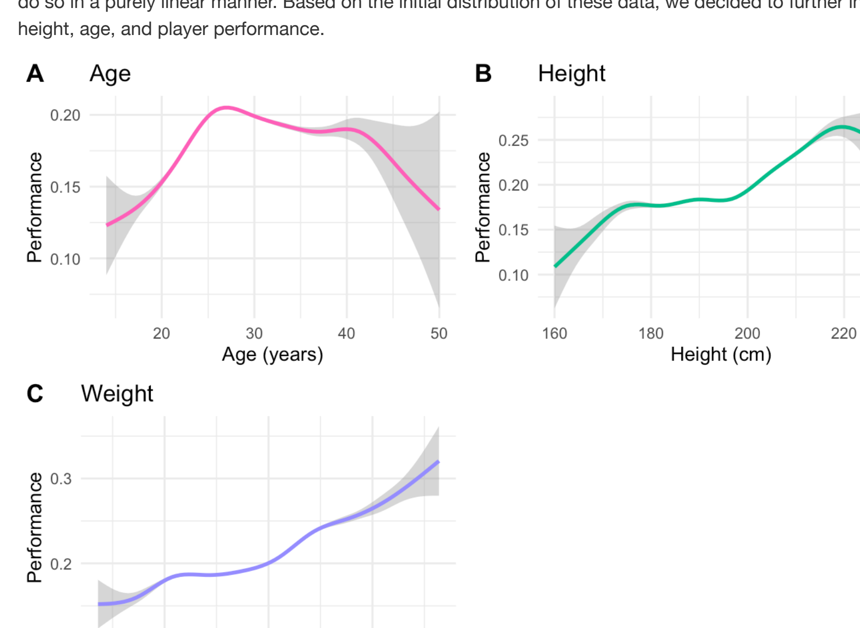
To show the player’s age at the time of a particular season, we created another variable named “Age.” This variable was calculated by subtracting the player’s birth year, represented by the “birth_year” variable, from the end year of the variable “Season.” The overall variables used in this analysis are represented in the table below:

League	Nationality	Player	Season	Height	Weight	Age	Performance
NBA	United States	James Harden	2019 - 2020	196	100	31	0.495447682356372
NBA	United States	Damian Lillard	2019 - 2020	191	88	30	0.426082068524196
NBA	United States	Devin Booker	2019 - 2020	198	93	24	0.418602824803912
NBA	Greece / Nigeria	Giannis Antetokounmpo	2019 - 2020	211	110	26	0.525637458628225
NBA	United States	Trae Young	2019 - 2020	188	82	22	0.434850207974117
NBA	Slovenia	Luka Doncic	2019 - 2020	201	99	21	0.42922582368876

Our initial questions looked at the relationship between a player’s nationality, league, and performance. Below is a graph of the top 50 players with the highest performance overall, showing that the majority of the best players are American but play in the Chinese Basketball Association (CBA) league. This result is surprising; we expected that the best players would be American and play in the NBA instead. Our subsequent analysis focused on understanding this unexpected result as well as characterizing the predictive value of the relationship between a player’s league and performance.



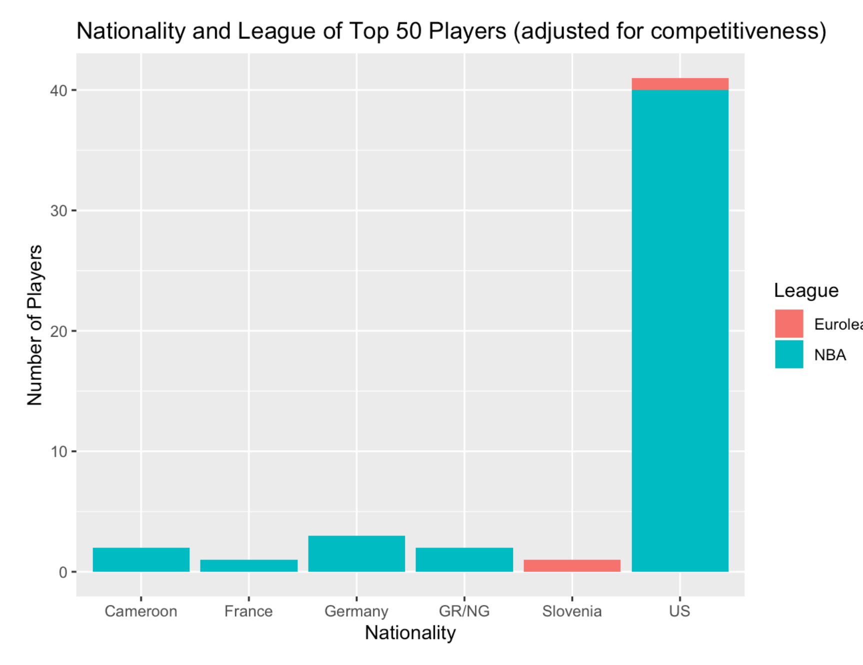
We also initially explored the relationship between age, height, weight, and performance. Based on prior knowledge, we believed that there would be a clear relationship between age and overall performance, given a predicted pattern of improvement when a player initially joined the NBA that would then decline after their age of peak performance. The graphs below show how overall performance clearly reaches an initial peak around 27 years of age, then declines. We also observed that performance tends to increase with height and weight, which would be expected, but none do so in a purely linear manner. Based on the initial distribution of these data, we decided to further investigate the relationship between weight, height, age, and player performance.



RESULTS

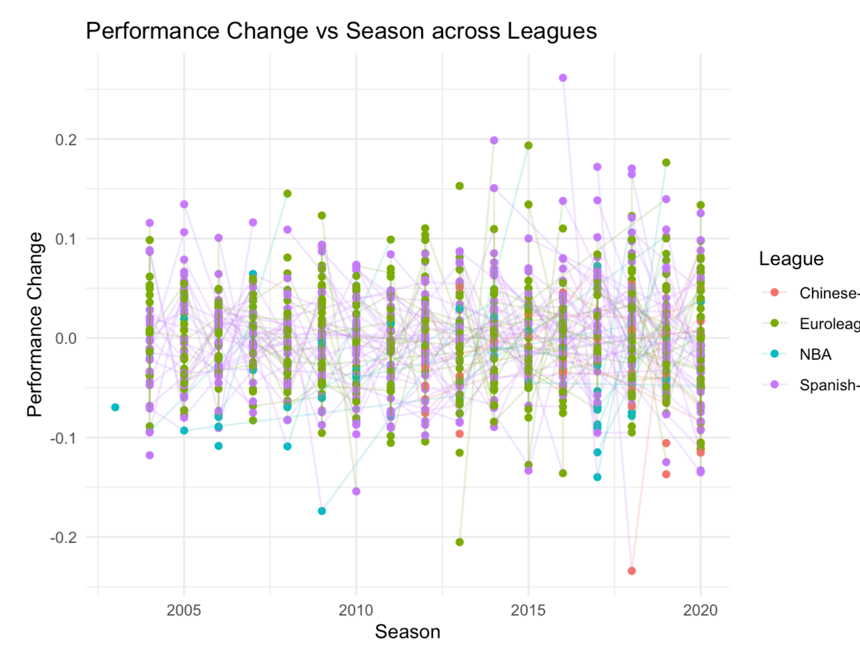
Question 1: Moving Leagues

In our initial exploration of the data, we were surprised to see that the majority of the top players belonged to the CBA league and not the NBA. Given this result, we reasoned that there must be another factor specific to individual leagues influencing performance. Since a player’s performance is likely higher when playing in a less competitive environment compared to a more competitive one, we determined that this factor is the overall “competitiveness” of each league. We decided that we wanted to determine this competitiveness factor to systematically weight calculated performance accordingly across different leagues. A 2017 ESPN article that ranked the top professional basketball leagues in the world was used to determine the correct ranking of the top 11 leagues. The top-ranked league, the NBA, was assigned a competitiveness factor of 1, and each subsequent league was assigned a competitive factor of 0.05 less than the prior league. Using this information, we could attempt to more accurately map a player’s change in performance if they moved from one league to another.



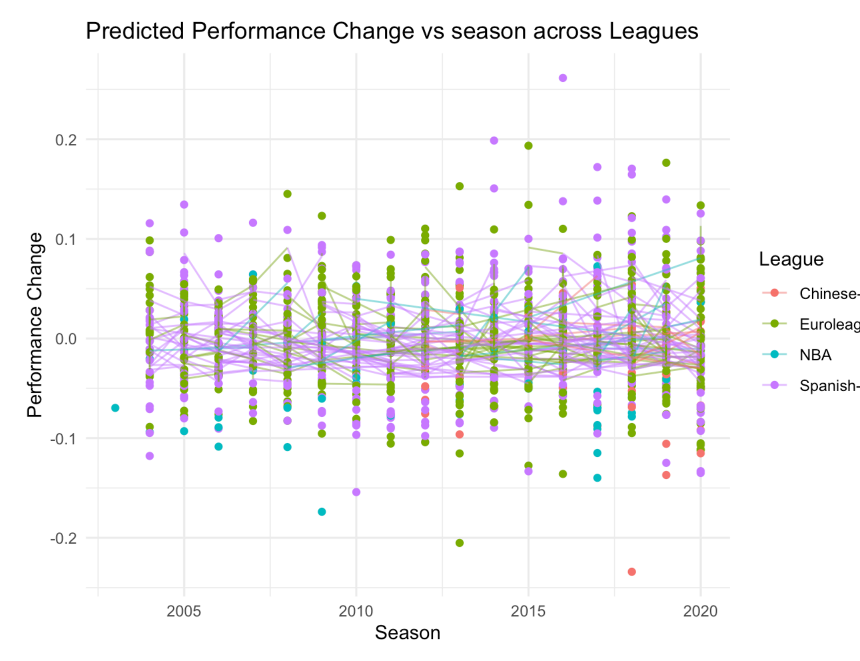
In order to address the first question, we first created a graphical representation of the change in performance across different leagues over time. A new dataset was created from the original that was ordered by Player, descending Season, and League. From this dataset, it seemed simple to take the difference between the Performance of a Player and the Performance for the following Season of the same Player and calculate Performance Change.

However, we soon realized that players who only occur once in the dataset were unduly influencing the Performance Change calculation for players above and below them alphabetically. By using the group_by() function, the Performance Change of individual players was changed to a single occurrence into NAs. Once removing the NAs and filtering out the instances when the League was the same across consecutive Seasons, the dataset became filtered down to 1092 observations from 9928 observations. It is important to note here that only the NBA, Euroleague, Chinese-CBA, and Spanish-ACB leagues were chosen for this analysis so as to facilitate our understanding of any observed patterns. These four leagues made up 21.5% of the entire dataset, as well as four of the most well-known leagues in the world. Therefore, these four were most pertinent to our analysis to generate a relevant, focused model showing the players’ movement between these leagues. The distribution of the raw data is displayed below, titled “Performance Change vs Season across Leagues’”. The point color is associated with League, the lines are grouped by Player, and the line color is associated with League.



While this plot may seem disordered, it does display trends about the change in performance that players display as they switch Leagues over time. The majority of observations where players are switching into the NBA, the turquoise-colored dots, have negative performance changes, which indicates that players are decreasing in performance once they enter the NBA. The average performance change for players entering the NBA was calculated to be -0.037, or -3.7%. Players entering into the Chinese-CBA league had similar results, with an average performance change of -2.5%. The Euroleague and Spanish-ACB leagues, on the other hand, produced average performance changes of .125% and .315%, respectively. These findings further intrigued us to create a predictive model for performance change.

To answer our first question, we used a polynomial model to fit the predicted performance change to actual performance change for players as they switch leagues. We chose a polynomial model since performance change over time is not linear, but instead oscillates up and down as players switch between multiple leagues during their career. First, we used a 10-fold cross-validation technique on the filtered dataset. After creating a function to calculate root mean squared error (RMSE) and a function to create a polynomial model for Performance Change based on Performance and Age, with given values of I and J, a double-loop was utilized to assign the calculated RMSE for each combination of I and J used in the model. The result of this process showed that the model with the lowest RMSE, and therefore the most desirable model, was the combination of I= 2 and J=1, the RMSE being 0.04697. A second plot is displayed below, titled “Season vs Predicted Performance Change”.



This plot is similar to the first one, but the lines now display the predicted performance change for each individual player as they switch from one league to another. Interestingly, the spread of performance change is noticeably narrower for the modelled lines compared to the actual performance changes depicted in the first plot. This indicates that numerous outliers exist in the dataset, as instances where performance changed by more than 0.05 in either direction are seemingly eliminated. Compared to the average actual performance change for players entering the NBA, the average predicted performance change was -.517%, which is more than 2% less. Additionally, players entering the Chinese-CBA had an average predicted performance change of only .042%, which is more than 2.5% different from the actual average performance change of -2.5%. The average predicted performance change for the Euroleague and Spanish-ACB leagues were essentially the same as the actual, as both were within .04 of the actual.

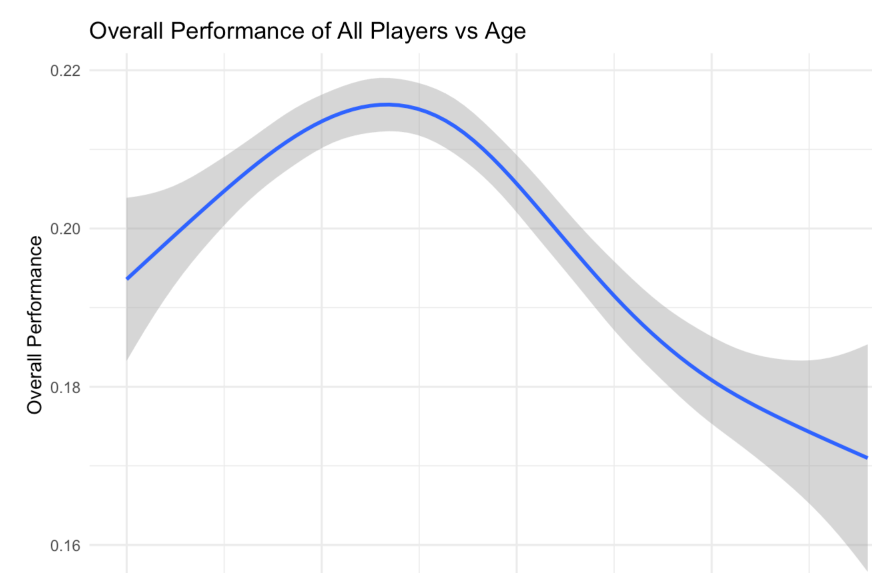
In order to gauge the significance of the model, the summary() function was utilized with the name of the model, sortedfinal_mod, used as the argument. The second-order polynomial fit for Performance was statistically significant at the 99% confidence level on both indexes, both with p-values less than 0.001. However, the Age variable was not statistically significant, with a p-value greater than 0.05. Yet the overall model produced an adjusted R-squared value of 0.2258, which means that 22.58% of the variations in performance change can be explained by the variations in performance and Age. While the Age variable is insignificant, the estimated coefficient of -.039 is interesting to observe, as it indicates a negative performance change as Age increases. Performance, as is expected, has an estimated coefficient of 0.826, which indicates a positive performance change as Performance increases.

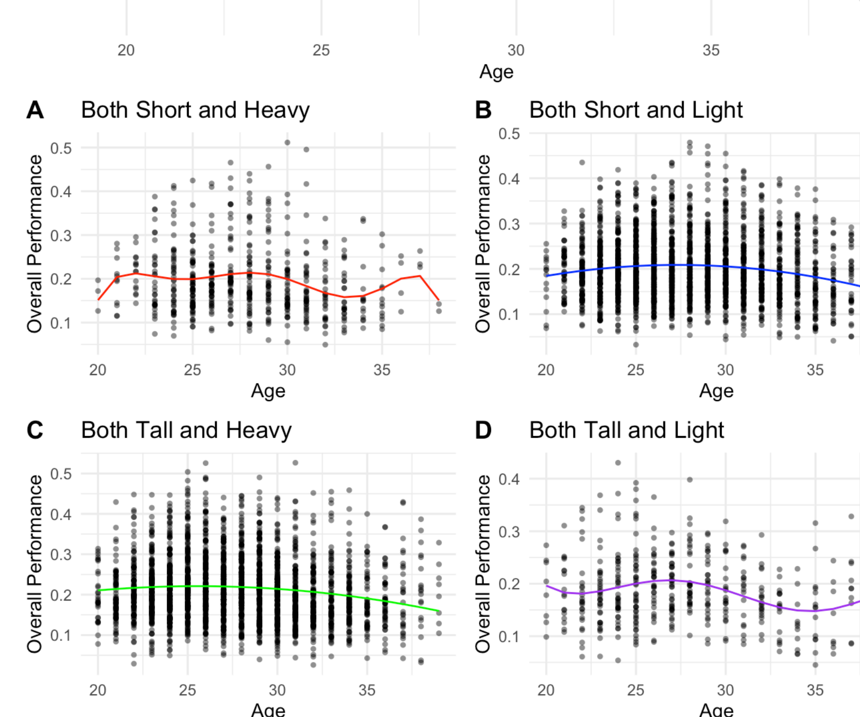
Question 2: Performance and Age

We noticed that there is a clear peak in performance corresponding with age, and that height and weight were positively correlated with performance, in our initial data exploration. We hypothesized that it would be possible to determine a player’s age of peak performance, and that this peak performance age would differ based on the player’s physical characteristics. In particular, we hypothesized that heavier and taller players would peak in performance at a younger age than players with lighter and shorter physical characteristics.

To answer these questions, we classified each player as being either above or below the calculated average weight and height of all the players, then separated the data into four groups based on each height-weight combination: “Tall and Heavy,” “Tall and Light”, “Short and Heavy,” and “Short and Light.” We excluded all players above the age of 40 to remove unnecessary outliers. After initial tests to eliminate a linear relationship were successful, we turned to the polynomial model to best fit our data. This was used because of the unique quality of polynomial functions to have distinct maxima and minima, of which maxima are most relevant to our model. We considered using sine and cosine functions, but were dissuaded by the fact that maxima and minima are distinctly repetitive, which could eliminate the possibility of modeling a second peak in an athlete’s career.

We developed four distinct polynomial models to predict the age of peak performance given a player’s height and weight classification. For each of the four models, we picked the model of best fit through the distinct combination of values that minimized our RMSE value. These combinations of I and J values differed between physical classifications, which gave us the preliminary impression that there would be a marked difference in the ultimate predictive polynomial models, which was evident in the graphs of the models. The combinations yielded RMSE values of: 0.07939997, 0.07067345, 0.07355429, and 0.06522199 for “Tall and Heavy,” “Tall and Light”, “Short and Heavy,” and “Short and Light” classifications, respectively. From our small RMSE values, we can conclude that our models serve as good predictors for our datasets. The shape of the polynomial of best fit for each combination evidently differs greatly, as well as the age of predicted peak performance. The residuals plot for each model shows that our models are good predictors, given our dataset.





Physical Stature Combination	Short and Heavy	Short and Light	Tall and Heavy	Tall and Light
Maximum Performance	0.27758226	0.2086880	0.2209266	0.2066194
Age	28	27	26	27

From this analysis, we were able to obtain our most significant results of our analysis focusing on physical metrics. Players classified as “Tall and Heavy” had predicted age of peak performance of 26 years old, had the greatest contrast with players classified as “Short and Heavy,” who had a predicted age of peak performance of 28 years old. We can also notice key differences between the shapes of the modeling graphs. The two somewhat paradoxical height-weight classifications, “Tall and Light” and “Short and Heavy,” are less smooth curves than their counterparts. Though they share the characteristic of a peak in performance around a player’s late 20s, Short and Heavy players are shown to have a 2nd peak in performance after 35 years, and Tall and Light players steadily increase in performance after 35.

Although our models have the nonlinear relationship between “overall performance” and “Age”, the polynomial regression is still considered linear regression since it is linear in the regression coefficients. We did a summary on the four models by using the summary() method, and the result is different from what we expected. In the summary, we noticed a low adjusted R-squared, which means little of the variation is explained by the independent variables “Age”, and “AgeSq”. In order to check if the “Age” variable is statistically significant, we also looked into the result of the F-test. With p-value < .01, we could reject the null that both coefficients are 0, which also means we are 99% confident that two variables are statistically significant. Here we see a conflict in our result, we have a low adjusted R-squared and a low p-value. The figure we plotted shows a clear relation between “overall performance” and “Age”, and we are getting confused that the “Age” variable can only explain little of the variation of the “overall performance”.

CONCLUSION

In Question 1 of this analysis, we first confirmed our prediction that a player’s overall performance can change as they switch between playing in different leagues. After analyzing the summary for our model, we were able to both tell that “Performance” is statistically significant for our prediction and predict a player’s change in performance to a high degree of confidence. We expected a player to decrease in performance as they entered the NBA, since that is the most competitive league in the world, and that is exactly what we found, with a negative relationship present in both the actual and predicted performance change variable. However, the (small, positive) predicted performance change for players entering the Chinese-CBA was not expected, as the actual data displayed a strong negative relationship. It is possible that outliers have caused significant skews in the predictions for the Chinese-CBA league in particular. For players entering the other two leagues, the Euroleague and Spanish-ACB, such a small performance change occurred in both the actual and predicted variables that the expected performance change for these leagues is 0.

In Question 2, we found that the age at which a player’s performance peaks is dependent upon the player’s physical characteristics of height and weight. Modeling for this question was effective at predicting a noticeable difference between age of peak performance, given a player’s height and weight class. The analysis showed differing fitted curves for each combination of physical characteristics, maintaining our intuitive separation of the dataset into four classes. Our analysis also supports our preliminary hypothesis that particularly heavy, tall players would have an earlier peak in their careers, possibly indicating that their careers are more limited as compared to their (relatively) shorter colleagues.

In the highly competitive and lucrative industry of professional basketball, our findings would be extremely important to recruiters and all those involved in the team. In reference to Question 1, a recruiter scouting a player from a different league would need to know how the player is likely to perform differently in a new league. For example, a player with a relatively low performance rating in a very competitive league might perform much better in a less competitive league, making them a good recruit. Similarly, a player with a very high performance rating might seem like a good recruit at first, but if they are playing in a less competitive league they likely will not perform as well when moved to a more competitive one. Since recruiters need to ensure that the players they recruit will be able to contribute to the team and make the team better-performing and more competitive overall, knowing how a player’s performance is likely to change between leagues is essential for a recruiter looking to recruit a player from a different league. For a recruiter just scouting players within the same league, however, it is essential to be able to predict how a player’s performance will change with time. Common knowledge dictates that athletic performance is not static and rather fluctuates with age and a player’s physical characteristics in a parabolic manner. Knowing at what age a player’s performance is likely to peak would enable a recruiter to make smarter long-term decisions by recruiting players who are more likely to improve in the future. Additionally, knowing the predicted peak performance age of each player on a team would enable a coach to assess the overall makeup of their team and determine the team’s long-term viability in highly competitive leagues such as the NBA. Since the choice to recruit a player is also a hefty financial investment, this knowledge is particularly important in the professional basketball industry in order for recruiters, coaches, and owners alike to make informed financial decisions about their team.

Though our dataset was certainly robust, we are nonetheless aware of limitations of our data and resulting analysis. We encountered an obstacle in its sheer size; there were too many leagues and variables to analyze in an effective manner. We restricted the size of our dataset many times over the course of this analysis to be able to produce focused, relevant results. Future analysis of this data could incorporate the excluded data, particularly the leagues that were excluded. Including more data would make our model more compelling. We also only used 8 variables in this analysis out of the 31 original variables, and it would be interesting to investigate these unused data in future analysis. The addition of health factors that might affect players in each physical class would also be interesting to analyze. This information would allow us to delve further into the analysis behind relating peak age of performance, determine if frequency of injury relates to physical class and peak age performance, and make the models we created to answer Question 2 more predictive.

We should note that it would be unethical to deny a player a contract or playing time in a game solely based on these predictions alone. Many of the best players in the NBA and other leagues have been “outliers” who play at a high level longer than their age or previous performance might imply, including famed player Kareem Abdul Jabbar, who retired at the age of 42, well past the predicted age of peak performance shown by our model. Additionally, a player’s league environment certainly has a strong impact on their performance in the league. For example, a player in the CBA might have significant negative predicted change in performance, but given more intense or higher quality coaching in another league the player could very well exceed expectations and perform at a high level. Therefore, we stress the importance of placing our analysis in context.