University of California, Berkeley

Predicting NBA Performance Given NCAA Statistics of Basketball Players

Marta Carrizo Vaque
SID: 3035384912

Luke Liu
SID: 3034958044

Data 100 Final Project

May 2020

Abstract

This project studies the NBA performance of basketball players based on their college statistics and physical features. In particular, we examine the factors that impact the PER (Player Efficiency Rating), a metric of performance, as well as the position of a player. We model PER using ridge regression with minimal success due to a lack of data and PER being a biased metric of performance. On the other hand, our majority vote model that predicts positions using logistic regression, PCA and random forests yields stronger results, but is very dependent on physical features.

1 Introduction

The goal of this project is to understand how we can best predict the professional career of male basketball players, given their NCAA statistics, biological features and educational background. Based on these given data, we will examine the following two questions rigorously:

1. How good will a player be in NBA?
2. Which position will the player play in NBA?

2 Description of Data

We are provided with five different datasets that contain information regarding basketball players and basketball teams. As the questions we wish to address are related to individual players, we use the following two datasets that contain such information.

2.1 2012-18_playerBoxScore dataset

The granularity of this dataset is a player's performance in a particular game that he has played in. It includes games from season 2012-13 to 2017-18. Each row contains general information about the game, the player's team and its opponent. However, the most interesting part is the statistics of the player's performance. They are very comprehensive and do not have any missing values. The only missing values of this dataset correspond to the first and last names of third officers when a game only had two officers, which are not relevant to our study.

2.2 college dataset

This dataset contains one row for each player. It has statistics about the player's performance during his entire career, including his performance in NBA and NCAA. While the column descriptions are missing, we were able to find their information on the basketball reference sites [1, 2], which are then inputted into our Jupyter Notebook. Unlike the previous dataset, this one is very messy and contains many missing values. Consequently, we performed data cleaning in the following sequence.

1. Drop columns with no information (e.g. "Unnamed", "URL").
2. Encode the birth date for players in a way that the day, month and year are in different columns. This allows us to work with the year of birth easily as a way to measure how statistics have changed over time.
3. A column encoded all US universities a player had attended. It was uninformative, so it has been changed to a binary feature that determines whether a player has attended one of the 10 top universities according to the site [3].
4. Height is encoded in feet instead of feet-inches to allow for linear scale comparison.
5. To be consistent with the previous dataset and allow for an easier merge, the player's name is divided into first and last name.
6. Positions are one-hot encoded. In addition, since some players play in multiple positions, their main position is encoded as a new feature.

7. We found that a significant number of players are missing data from NCAA (from players who did not attend a US college or did not attend college at all). Since the goal of this project is to understand the performance of NBA players based on their NCAA performance, we proceed to drop those rows. Also, some missing values could be interpreted from the rest of the statistics (e.g. "efgpct") or be given a reasonable value that addressed the reason they were missing.
8. Incorrect column data types are changed (e.g. Birthday, weight and NCAA_games are now encoded as integers rather than floats).

There are some missing values left that will be addressed individually when answering the questions.

3 Description of Methods

We address each of the questions we wish to answer separately.

3.1 Player's performance

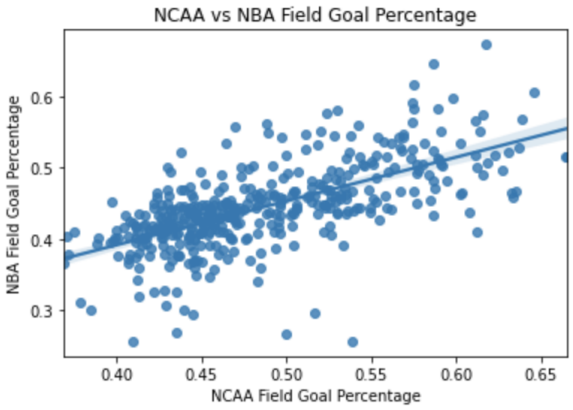
In this section, our goal is to predict how good a player will be in NBA based primarily on their statistics in NCAA. There are many factors that speak to the overall performance of a basketball player, such as field goal percentage, 3 point percentage, number of assists, turnovers, and many more. Luckily, there is a comprehensive statistic that NBA uses to boil down all of a player's contributions to one number, known as the Player Efficiency Rating (PER) [4], which we aim to predict. However, we keep in mind that this rating is nowhere near perfect, and may differ for players in different positions.

Since the PER is not included in any of our datasets, we had to manually calculate it ourselves given the available information. Using the playerBoxScore dataset, we first found the cumulative statistics for each player based on all the games they played between 2012 and 2018. Then, we calculated the PER based on the formula below and normalized it by the total minutes played the player has played.

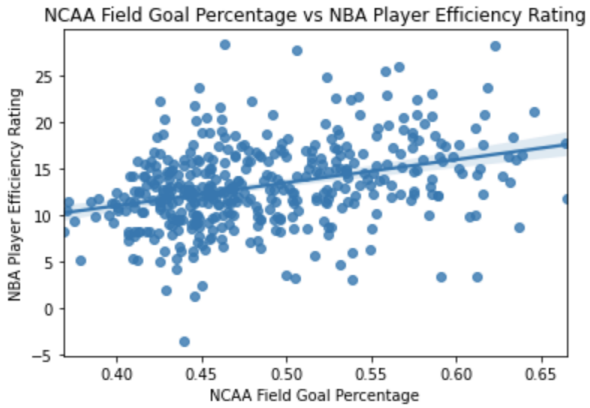
$$\begin{aligned} PER = & \frac{1}{Minutes}\left(FGM \times 85.910 + (Steals - TO) \times 53.897 + 3PTM \times 51.757 + FTM \times 46.845\right. \\ & + (Blocks + OffensiveReb - FGMiss) \times 39.190 + Assists \times 34.677 \\ & \left. + DefensiveReb \times 14.707 - Foul \times 17.174 - FTMiss \times 20.091\right) \end{aligned} \tag{1}$$

There are certain extreme outliers (PER score above 100 or below 0) caused by players having played very few minutes. Consequently, we removed those players that have played for less than an hour, and then merged the PER score with the college dataset. To enable us to build a stronger model and get more accurate predictions, we only considered players that are active for at least one year, and those that have three point shot statistics in their NCAA data.

With our final dataset ready, we did a train test split and conducted exploratory data analysis (EDA) on the training data. First, we are interested in finding if a direct correlation exists between two corresponding variables in the dataset. For instance, we found that there is a fairly strong positive correlation between the field goal percentage of a player in NCAA and NBA. We also want to see if this kind of correlation carries over to PER. While not as strong, there is a positive correlation between the field goal percentage in NCAA and the PER. The graphs are shown in figure 1.



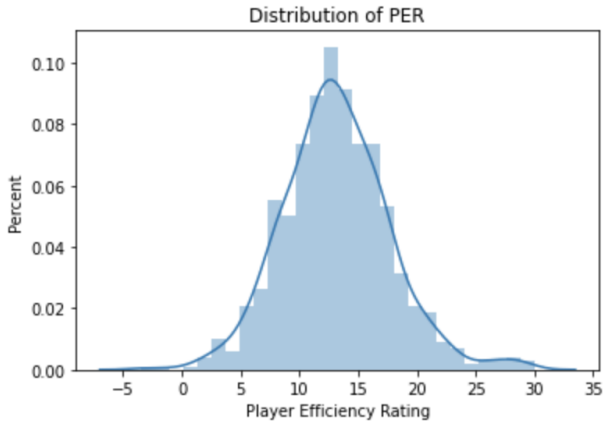
(a) FGP vs FGP



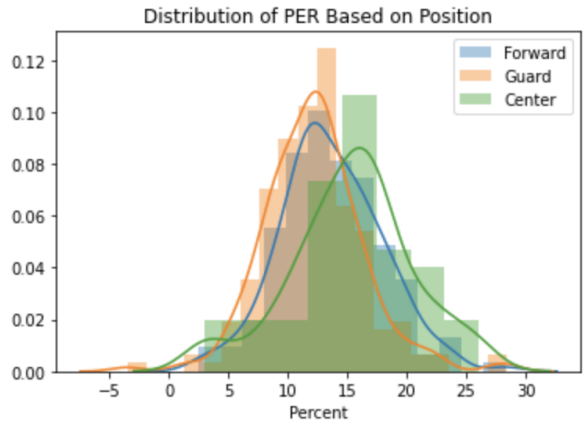
(b) FGP vs PER

Figure 1: Correlation between NCAA and NBA Data

We asserted earlier that PER may depend on the position of the players. The histogram in figure 2 shows that players in the center position have an overall higher PER. This confirms our assumption, and makes sense since centers are valued for their ability to protect their own goal while scoring with high efficiency.



(a) Unclassified



(b) Classified by Position

Figure 2: Histogram of Player Efficiency Rating (PER)

In order to find the best features, we extracted all the columns containing the NCAA statistics, background and biological data, and plotted all their correlations against PER on a bar graph in figure 3. Unfortunately, none of the features have a particularly strong correlation with PER. Surprisingly, whether players were in top basketball colleges does not correlate with their PER. The field goal attempted per game in NCAA also minimally correlated with PER.

Consequently, we excluded these two features from our model. Using the SKLearn Pipeline, we built a Ridge Regression model that predicts PER based on the rest of the features. We chose to use Ridge Regression rather than LASSO Regression because we do not have a lot of features in the first place, and most of the features we selected all roughly have a similar magnitude of correlation with PER.

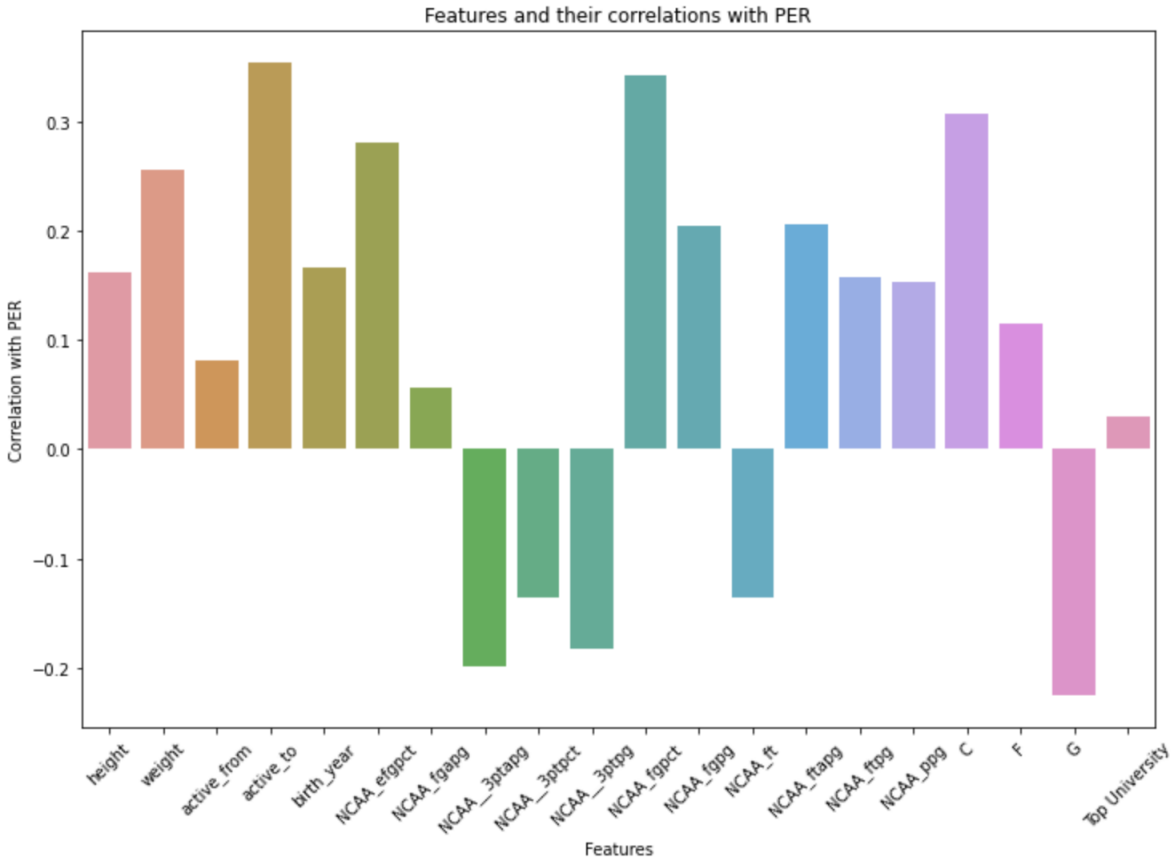


Figure 3: Bar chart showing correlations between selected features and PER

We standardized all features except the positions (c, f, g) because they were one-hot encoded from the categorical variable.

In order to find the best model, we iterated through 30 equally spaced data points between 0 and 15 to search for the optimal alpha hyperparameter with the lowest cross-validation error. The results will be discussed in the next section.

3.2 Player's position

The aim of this part is to create a model that predicts the NBA position of a player given their biological and college data. We first determined if biological information is useful when predicting positions. As shown in figure 4, both height and weight seem to be very good indicators for positions.

We were also interested in whether the month of birth is a good predictor for positions. Figure 5 indicates that it does not seem to be one. However, there seems to be a surprisingly low number of players born in April. This could be an interesting side revelation. As all children born in the same year are enrolled in the same school year regardless of their month of birth, they may be exposed to different concepts at different ages. This affects education, especially in the early years. We wondered if a similar effect could be seen in basketball players. We used hypothesis testing to see if there is a significantly lower number of players born in April with a p-value cutoff of 5%. We found a p-value of 1.5% and we rejected that the probability of being born in April is the same as any other month.

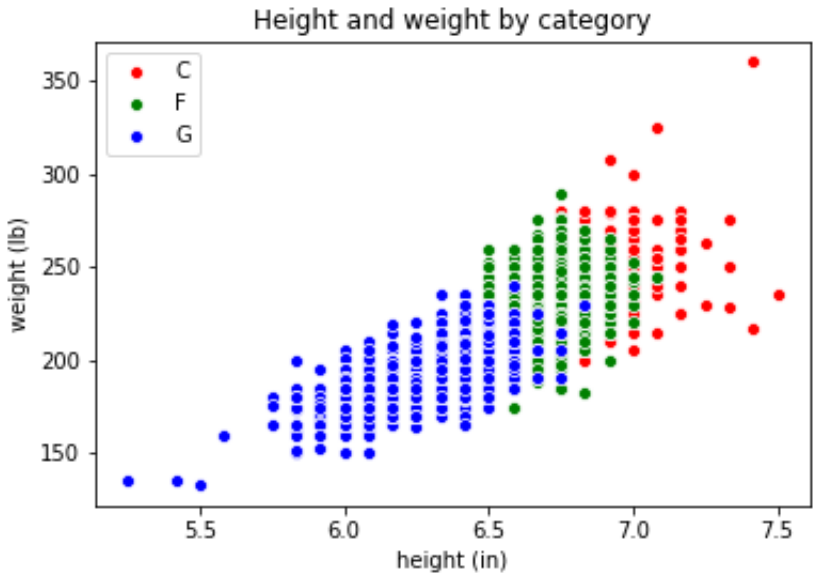
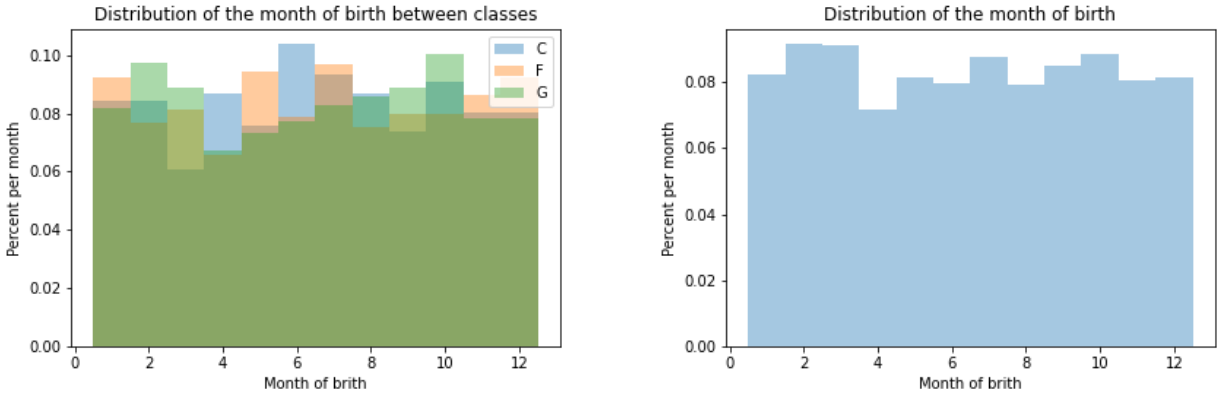


Figure 4: Height and weight of NBA players. Color encodes the position they play at. It can be seen that height and weight are good features to predict position.

However, this result is very surprising and further research should be done before claiming that there are fewer players born in April.



(a) By main position

(b) All players

Figure 5: Number of players born on each month

Finally, we used NCAA statistics as features since we understand that the differences between the offensive and defensive performance in these statistics are key to asses the position of a player. We also added the birth year as a measure of how statistics may have changed over the years.

We first used logistic regression as our classification technique. By doing so, we assumed that the boundaries between the classes are linear. We used cross-validation to determine the best type of norm penalty, as well as the constant associated with such penalty.

Using the results from the regression, we determined that there is a high degree of collinearity between some of the features. Since PCA is a dimensionality reduction technique, it can be useful to deal with this issue. Then, PCA was performed before logistic regression and cross-validation determined the best norm penalty, the penalty parameter and the number of principal components.

Using PCA, we observed that even if the boundaries of classes look somewhat linear, sometimes they have a peculiar shape. Because of that, we decided to apply random forests to the data.

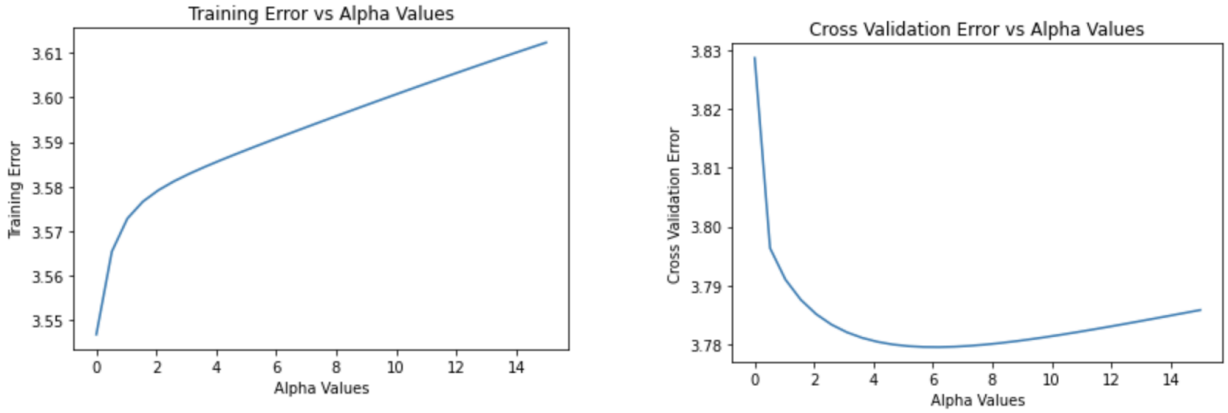
Finally, since the improvement between methods is very small, we decided to use a majority vote method, where each of the three described methods predicts an outcome and then a majority vote is taken. In the event that all methods predict a different outcome (this is possible since there are three positions), the technique that has the better cross-validation error will determine the output. We will find that this is random forests.

4 Summary of Results

We summarize the results obtained by applying the methods described before for each of our questions.

4.1 Player's performance

After multiple trials of feature engineering, our final Ridge Regression model reported a training error of 3.59, a cross-validation error of 3.78, and a testing error of 3.63. All of these values are calculated with Root Mean Squared Error (RMSE). To obtain this result, we ran our model on different hyperparameter values between 0 and 15, and found that an alpha of 6.2 provided the lowest cross-validation error, as shown in figure 6.



(a) Training Error

(b) Cross-Validation Error

Figure 6: Training and CV Error against different hyperparameter values in Ridge Regression

To gauge the strength of our model, we compared its performance to the baseline model using a Dummy Regressor that always predicts the mean of PER. The baseline model reported an error of 4.49, and we conclude our model has improved from the baseline. We also attempted to use LASSO Regression, and found that the best hyperparameter is 0, which simply turns the model into a simple linear regression model. The data confirms our assumption that Ridge Regression is the better model to use in this case.

Finally, we graphed the residual plot based on our Ridge Regression model shown in figure 7. It shows no pattern and a similar vertical spread throughout, as desired.

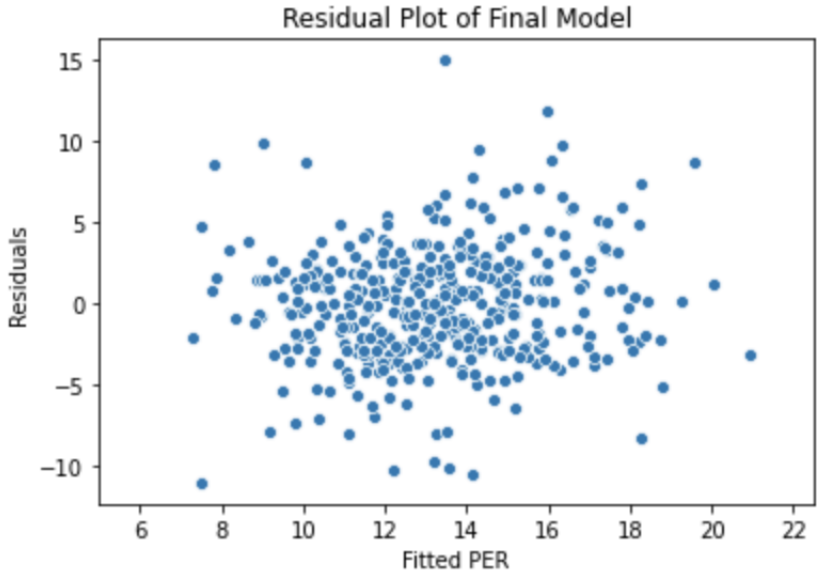


Figure 7: Residual Plot of Ridge Regression Model

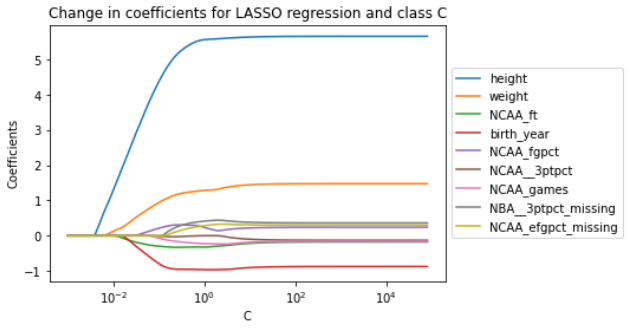
4.2 Player's position

The simple logistic regression model reported a cross-validation error of approximately 88.06%. To obtain this result, we used a $\mathcal{L}^1$ norm. But the most interesting thing about this norm is that it acts as a form of feature selection. Therefore, if we change the penalization term $C = 1/\lambda$ from smaller values to larger ones, we find that the coefficients of features gradually move away from 0. The first features that appear are the ones that are more relevant when predicting the output. As we have performed logistic regression with three categories, each category has its own coefficients. The visualizations can be found in figure 8.

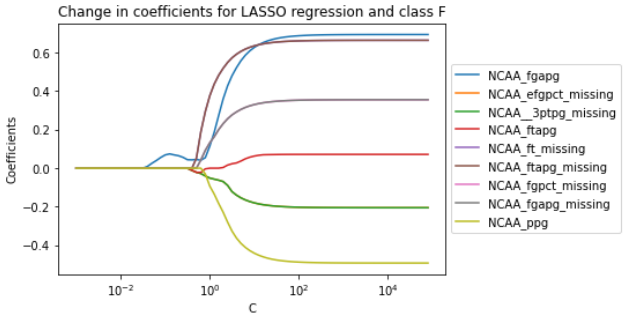
We observed that height and weight are indeed very good predictors for all positions. Additionally, we observed that for the center position, there are several features that overlap completely. This may be due to the fact that the data in these features may have similar values (e.g. the missing value features). This motivated us to perform dimensionality reduction with PCA.

Our logistic regression using PCA performed slightly worse than simple logistic regression, with an accuracy of 88.02%. However, the small difference in accuracy between methods could simply be due to the chance of the cross-validation procedure alone. We also plotted the data using the first two principal components to understand why logistic regression does not perform as well. The graph in figure 9 shows that even if decision boundaries are somewhat linear, they still have certain structures.

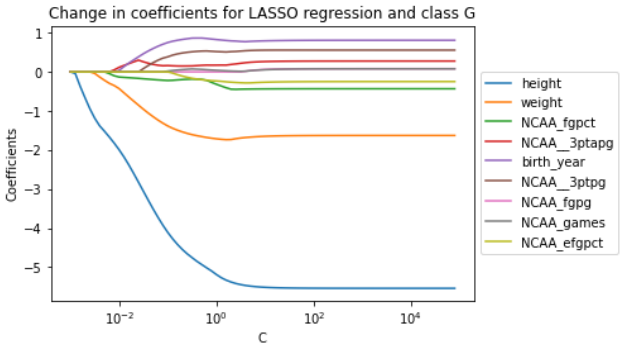
Consequently, we utilized random forests as a technique that does not have linear decision boundaries. Their performance is slightly better, with a cross-validation error of 88.16%. Finally, we estimated the error of our majority vote model and obtained a cross-validation error of 88.06%.



(a) Centers



(b) Forwards



(c) Guards

Figure 8: Weight of logistic regression parameters with respect to the $\mathcal{L}^1$ regularization parameter $C = 1/\lambda$ for different positions.

Given those results, it is clear that the random forest model performed the best. However, it is known that this technique tends to overfit. Furthermore, its difference in accuracy with the majority vote model is very small and the latter seems more robust, since it is built using three different models with similar cross-validation error.

Finally, we selected the majority vote model to predict a player's position. We found its accuracy using the test set to be 88.65%. We also studied its performance per class, as shown in table 1. It tells us that the greatest misclassification occurs when centers are thought to be forwards. If we recall figure 9, this seems reasonable since there is a blur in the boundary between them. Then, it is likely that the algorithm predicts the most common point on the boundary, which is the forwards position.

	Predicted C	**Predicted F**	**Predicted G**
Actual C	59	24	0
Actual F	6	193	14
Actual G	0	15	209

Table 1: Comparison between predicted and real positions by clas on the test set.

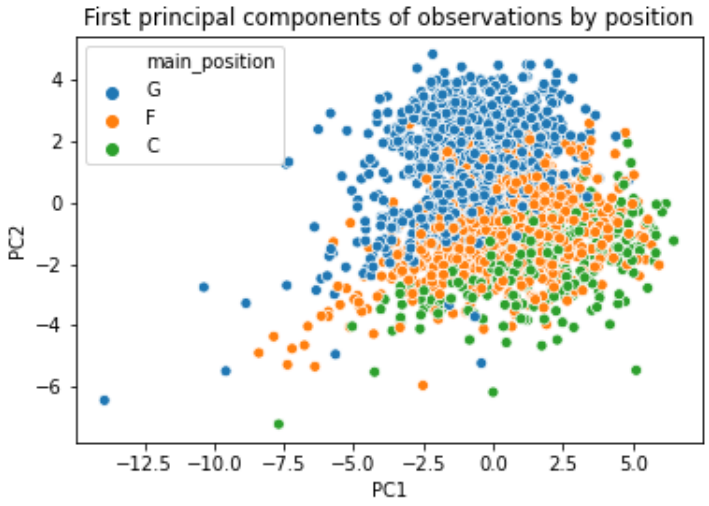


Figure 9: Two first principal components for the features used to predict positions and position for each observation.

5 Discussion

5.1 Player's performance

The strength of our Ridge Regression model is not as strong as we have hoped. Even after extensive feature engineering, we find a minimal correlation between features in the college data and PER. We are surprised that the college(s) that a player attended did not correlate with the PER at all. This could be due to several reasons. First, PER may be a biased metric of performance. While it is a standard comprehensive rating used by NBA, it largely measures offensive performance and does not paint an accurate picture of a player's defensive performance. Furthermore, since we only possessed player box data from 2012 to 2018, we were not able to calculate a player's career PER which leads to further bias. Particularly, players that only played for a few years between 2012 to 2018, or those with minimal data are more affected. To improve this model, we need to redefine our metric of performance, gather more data and seek out more relevant features in college data that correlate with this metric.

5.2 Player's position

When predicting positions, we found that, surprisingly, two of the most important features are the height and weight of the players. We used a majority vote model that uses logistic regression, logistic regression combined with PCA and random forests and we have obtained an accuracy of 88.65% on the test set.

This model could be improved by taking into account class imbalance (as there are fewer centers than any other class). It could also be useful to find other features that might have a greater impact on the player's position. Finally, it would be interesting to study if misclassified players are more likely to play multiple positions and if they have been misclassified to their secondary position.

5.3 Seven Questions

- **What were two or three of the most interesting features you came across for your particular question?** When we predicted the position of a player, height and weight were some of the more important features. It was also interesting to see in figure 8 that the missing values also played a very important role in determining positions. When we were predicting the PER, it was interesting to see that player positions affected PER quite significantly.
- **Describe one feature you thought would be useful, but turned out to be ineffective.** We thought that whether a player has attended a top basketball college in the US would be important in predicting the PER. As a result, we took great care in encoding the universities during feature engineering. However, it turns out that there is no correlation between these two variables. This could be due to several factors; for example, the top colleges can change over time, and the performance of college does not necessarily represent the performance of individual players in that college.
- **What challenges did you find with your data? Where did you get stuck?** We struggled when we tried to clean the list of colleges a player attended. The main issue was that they were encoded as strings, and for players that had attended more than one college, they were separated by commas. However, some colleges also have commas in their names, such as University of California, Los Angeles. We struggled to find a way to distinguish compound university names from different university names. We finally decided to split college strings by commas and then assume that every segment that did not contain words University, College or Institute were part of the previous split.
- **What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?** We assumed that PER is the best indicator of a basketball player's performance. While this statistic is comprehensive in including a player's achievements, it is difficult to boil down all the achievements to one number. As mentioned earlier, it largely measures offensive performance and might not paint an accurate picture of a player's defensive performance. The PER calculated is only based on the data from 2012 to 2018, and thus does not necessarily represent a player's career PER, and might not be an accurate response variable.
- **What ethical dilemmas did you face with this data?** While there are no significant ethical concerns with respect to the use of data in this study since all data used are publicly available, we acknowledge that there may be privacy issues that basketball players may not want their data to be used in certain ways.
- **What additional data, if available, would strengthen your analysis, or allow you to test some other hypotheses?** Having a player's career PER would strengthen our analysis significantly, as opposed to the PER only from 2012 to 2018. We are also interested in player's salaries, which could be an interesting variable to predict given NBA and/or NCAA data.
- **What ethical concerns might you encounter in studying this problem? How might you address those concerns?** After all, each and every basketball player has unique experiences and circumstances. It may not be fair to predict a player's performance simply based on his college statistics. We also observed that positions are very dependent on physical features, and basketball associations might be tempted to assign new players to a position based on that. We strongly urge teams to consider potential players holistically rather than a model with potential bias.

References

[1] *Basketball Reference*, 2000 (accessed May 7, 2020). https://www.basketball-reference.com/.

[2] *SRCBB: College Basketball Stats and History*, 2000 (accessed May 7, 2020). https://www.sports-reference.com/cbb/.

[3] NCAA, *Top Men's Basketball Schools*, 2002 (accessed May 7, 2020). https://www.ncsasports.org/best-colleges/best-basketball-colleges.

[4] S. B. Dime, *How to Calculate Player Efficiency Rating (PER)*, 2020 (accessed May 8, 2020). https://www.sportsbettingdime.com/guides/how-to/calculate-per/.