NCAA BASKETBALL FOCUS

Evan Zimmerman
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22904
ewz9kg@virginia.edu

Chris Barfield
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22904
cdb8da@virginia.edu

Kai Helli
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22904
dpd3zd@virginia.edu

December 8, 2022

ABSTRACT

We would like to help the UVA basketball team improve by investigating the position of each player on the team. Through machine learning techniques, we analyzed the performance of players, and whether their contribution on the court was better suited for a position different from their current position. From here, we want to make sure Coach Tony Bennett sets his lineups to maximize the effectiveness of players at their position.

1 Introduction

So long are the glory days of UVA basketball. The Virginia community has faced two years of disappointment and sadness, and we have set out to put a stop to it. With the help of machine learning, we can evaluate each player's performance to provide better insights that will hopefully improve the overall performance of the team.

For this project, we found two main datasets related to college basketball. The first dataset focuses on various statistics of college basketball players in each year. It contains data from 2009 to 2022 and includes a total of 65039 entries for 65 features. Therefore, it is a very comprehensive source to use machine learning in order to achieve a good accuracy and generalization. The record contains basic information such as the player's name, number of games and the conference in which he played, but also important statistics such as the percentage of minutes played or the offensive rating. The dataset already has a fairly high usability, with only some data missing that was not collected until after the first few years.

In addition, we considered a college basketball dataset that contains mainly team statistics. This dataset contains data from 2013 to 2021 and provides 24 features in 2455 rows. Its primary purpose is to evaluate a team's performance in a given year. Since the first dataset lists the team and the year in which a player played, we can use this dataset to match a player's performance to the team's overall performance. Using the Kaggle's statistics of each feature, we can also see that this dataset does not contain any missing data, which is part of the reason why it has high usability.

We were able to find several experiments and research that attempted to apply machine learning techniques to basketball datasets, either in NBA or college basketball. Our group investigated two experiments which used k-means to classify NBA players position based on their grouped clusters [1] [2]. Our goal is to apply a similar k-means algorithm to a college basketball dataset, and dive a bit deeper to build tangible insight on the UVA basketball team.

2 Approach #1: Natural Grouping

2.1 Method

Before we could apply any sort of algorithm to find natural groupings within our player data, we first needed to preprocess the data. When analyzing our dataset, we found that entries prior to 2013 contained large amounts of missing data, so we decided to remove all entries before that year. We believed that it was very important that our model have as much access to a player's seasonal data as possible. However, not all of the features in the dataset were relevant to the insights we wanted from our groupings. We believed the categorical features *player name, team, conference, player id, number, and recruit rank* had no impact on evaluating a player's performance on the court, so we dropped those features from our dataset.

A pipeline was then created to standardize the preprocessing for the remaining numerical and categorical features. For the numerical pipeline, missing values were imputed using the median, and StandardScaler was used to standardize values. The categorical data was prepared using OneHotEncoder.

We believed that the k-means algorithm would be best for constructing natural groupings amongst our data. Given a certain number of clusters, the k-means algorithm groups the data into that number of clusters based on how similar each data point is to the center in each cluster. The k-means algorithm is an unsupervised learning algorithm that is effective at finding groups not officially labeled in the data. In the context of our dataset, this could inform us which players are performing similarly to each other and should possibly be playing the same position.

2.2 Experiments

To discover the best number of clusters for our dataset, we ran k-means on a range of (2,20) clusters. We then plotted inertia vs the number of clusters to determine the number of clusters that fit our data the best. Inertia essentially measures how well a dataset was clustered by k-means, and the ideal number of clusters can be found by looking at the "elbow" of the inertia plot that is shown in figure 1. The "elbow" of the inertia plot for our trials was at 5 clusters.

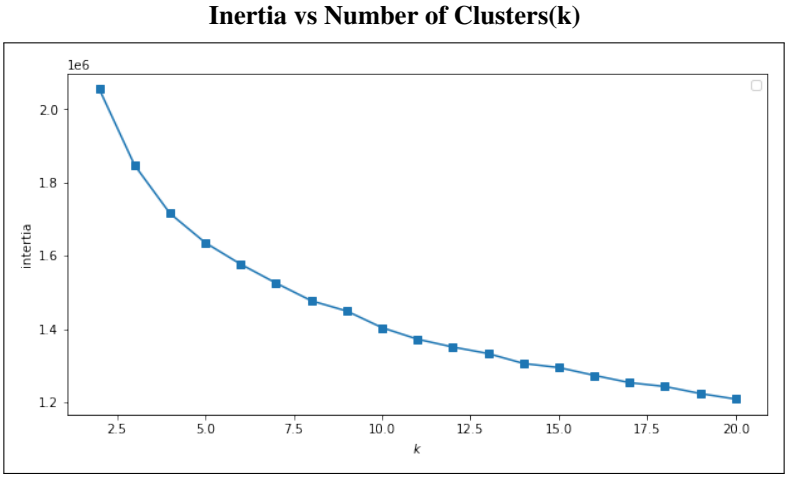


Figure 1: This graph visualizes the relationship between the number of clusters and the inertia for our k-means algorithm using that number of clusters

2.3 Results

The purpose of applying the k-means algorithm was to find the significance of groupings among player performance. To understand what each of the five clusters represented from our dataset, we performed a quantitative analysis of data within each cluster. We visualized the differences in data for the same features across each cluster by generating box plots for each feature in each cluster.

Boxplots of features per Cluster

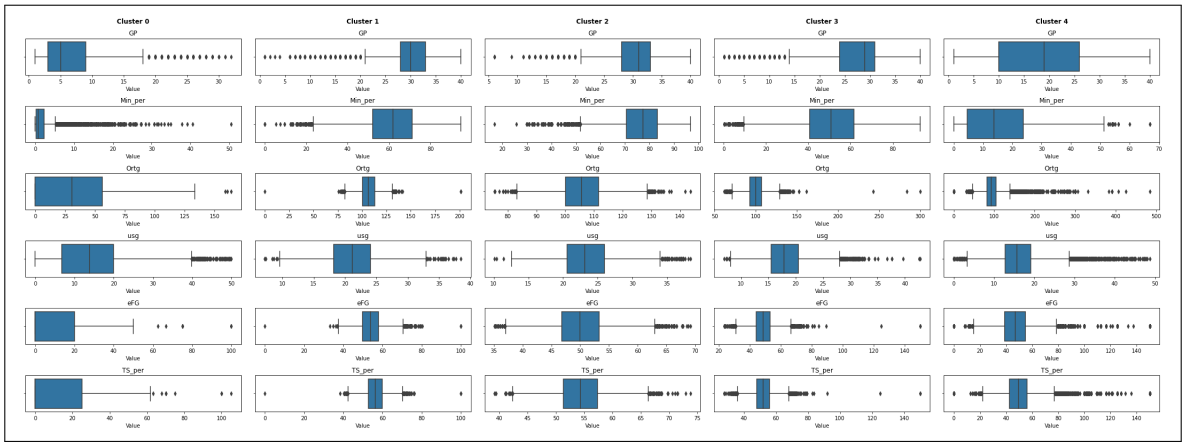


Figure 2: This figure contains a sample of the many boxplots that were created to compare the data for each feature across clusters. Each column represents a cluster, and each row represents a feature. **Not all boxplots are shown

One of the most glaring discrepancies that jumped out to us between cluster 0 and the other clusters was the games played. There was a median of five games played over the course of a season for players of cluster 0, leading us to conclude that the cluster mostly consisted of bench players. All of the box score data (points, assists, blocks, steals etc.) for this cluster had a median of approximately zero, reinforcing our conclusion.

At the other end of the spectrum, clusters 1 and 2 contained players whose median games played was a full season and whose box score stats indicated that they were significant contributors when they played. One of the major differences between these two clusters was the height distribution, players in cluster 1 were overall much taller than players in cluster 2 (5 inch difference). This difference, in addition to players in cluster 1 having greater blocks and rebounds per game, but less assists, led us to conclude cluster 1 contained starting forwards while cluster 2 contained starting guards.

Although the players in cluster 3 also had a median games played of approximately a full season, their median minutes percentage was lower than that of clusters 1 and 2. Overall, their box score stats were also lower than clusters 1 and 2, but not insignificant. Based on these findings we believed that these players were impactful substitutes. We found that the players in cluster 3 also had significantly less two point attempts, while still having respectable efficiency ratings, further confirming our hypothesis.

The players in cluster 4 seemed to be something of an anomaly at first. Although they were playing enough games and minutes to be considered part of the rotation, the distribution of their box score data was much closer to zero in all categories. It seemed as if despite being a part of a team's rotation, player's in cluster 4 were non-factors on the stat sheet. This led us to conclude that these players were essentially the role players of the team. It is common for every college basketball team to have a few players who get significant playing time because of their defense or how well they know their team's system. Their impact is strong despite not appearing so in the statistics. We concluded that players in cluster 4 represented this group.

All in all, the k-means algorithm gave us excellent insight on the makeup of our dataset, however it did not provide the actionable information we were hoping for that could actually help the UVA basketball team.

3 Approach #2: Predicting a Players Position

3.1 Method

We wanted to generate results that may deliver tangible feedback for the current UVA basketball team, so we drafted a driving question: Are UVA players playing in the best position? We sought out to build a model that would predict a given player's position based on their season stats. Overall, the classifcation from this model came from the insurance that we were training on the best players at each position, and not bad ones. Our end goal was to verify that the UVA players were playing in the best possible position through a predictive model when given their stats.

To ensure we were testing on the good players, we needed to identify a quantitative value which would measure a players performance. One of the features in our dataset was box plus-minus (BPM), which is a general indicator on how much a player contributes to a teams success. BPM is calculated by taking a player's box stats (such as points,

rebounds, and assists) and normalizing them by their team's pace. This calculation is then compared to other players in the NCAA to produce the final value. [3]

However, to be sure that BPM is indeed a good measure of a player's contribution to a team's success, we wanted to find some statistical evidence to support it. To do this, we incorporate our dataset for team statistics for a given season, which includes the total number of games played as well as the number of games won. The ratio of these two statistics can then be used as a measure of a team's success in a given season. Based on this, we try to find a correlation between the average BPM of a team and its performance. However, our dataset contained outliers, for example, people who played for only a few minutes and scored during that time had an incredibly high BPM value. Therefore, we decided to only consider players who have played an average of at least 10 minutes per game and have appeared in at least 10 games in general. We then found that the BPM and the percentage of games won is highly correlated with a factor of 0.743. To better illustrate this relationship, a scatter plot including the regression line is shown in figure 3. We can also infer why the correlation is not higher: since there is a large level of uncertainty in sports in general, a team with statistically many good players may still be beaten by a statistically weaker team. In general, however, the high correlation validates the use of this statistic to filter out statistically not as good performing players.

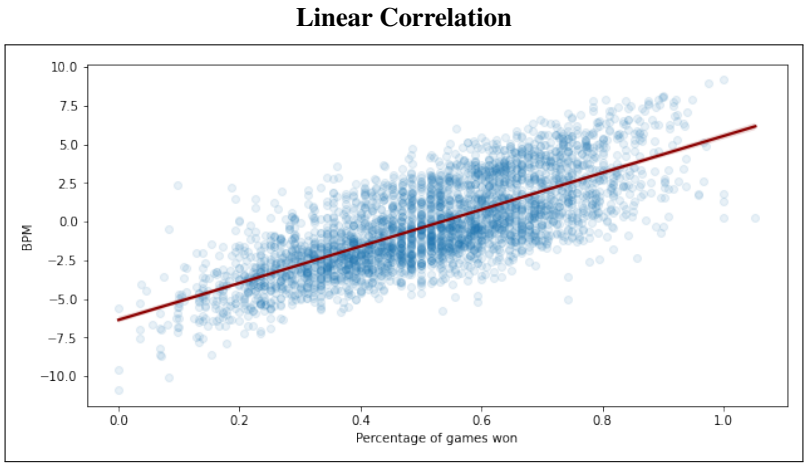


Figure 3: The scatter plot visualizing the correlation between the average BPM of a team and the percentage of games won in a certain year.

After filtering our dataset to include only the top quarter (75th percentile or higher) of players in the NCAA at each position in terms of BPM, we were left with about 10,000 rows and 5,000 individual players. Having the best players left us with one more task: to remove any bias from our model, we had to remove all data of current UVA players from our dataset and separated the most current stats from them into another dataframe. After running the remaining dataset through the pipeline we had set up in our first approach and splitting off another test dataset from it, we were then able to start training our models.

3.2 Experiments

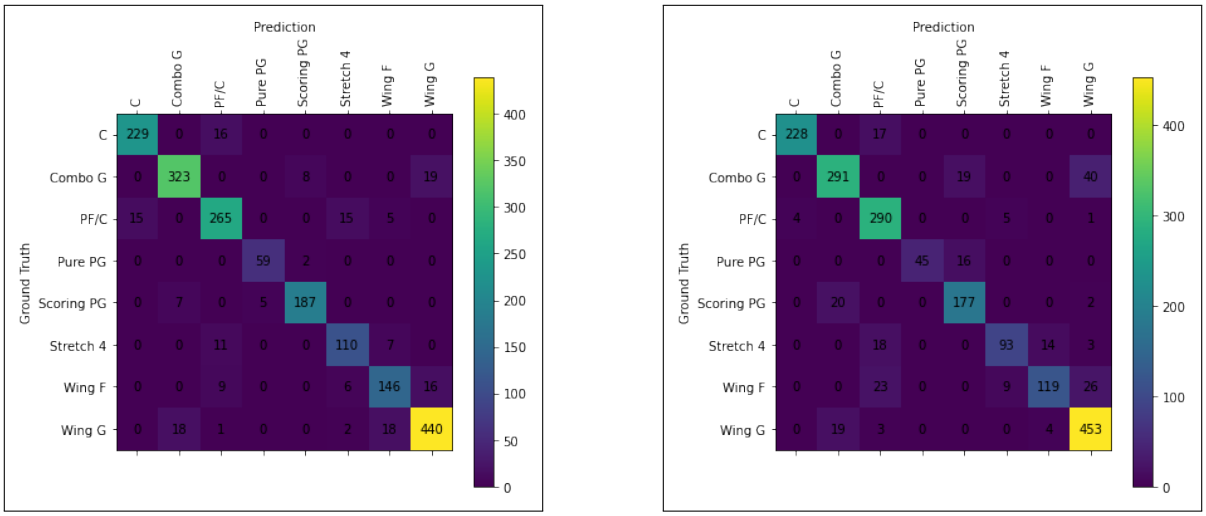
After evaluating the different models for classifying a player's position, we were left with two models that were especially suitable for our task: SVM and Random Forest. The first model we trained was the SVM classifier. In order to achieve the best results, we needed to properly determine the dimensionality of our data. For this, we trained a linear, polynomial, and Gaussian RBF SVM over several iterations with RandomSearchCV to optimize the different hyperparameters. It turned out that a polynomial SVM with a degree of 4 was the best fit for our data. With an overall accuracy of 91%, as can be seen from table 2, the model performed very well on the test set. The confusion matrix from this prediction can be seen in figure 4a.

	Precision	Recall	F1 Score
Polynomial SVM	91%	91%	91%
Random Forest	88%	87%	87%

Table 1: The evaluation metrics of our final models on the test set.

The second model we trained was a Random Forest Classifier. Although we could not achieve an accuracy as high as with polynomial SVM using RandomSearchCV, the resulting model still achieved a reasonably good accuracy of 87%. The confusion matrix of the prediction of our final model on the test set can be seen in figure 4b.

Confusion Matrices



(a) Polynomial SVM

(b) Random Forest

Figure 4: The confusion matrices of our final models predictions on the test set.

3.3 Results

When we finally applied the current UVA players to our model, we input all rows from UVA players during the 2021-2022 season. Because of this, no current first-year students were included in the results. Also, walk-on Tristan How did not appear in any games last year, so he did not have a row in the dataset either.

Our Random Forrest model correctly predicted four out of eight players tested, and determined the remaining four should be classified as a different position. In regards to our Polynomial SVM, out of the eight players returning from last year, our model correctly predicted six of the UVA players, suggesting that 2 players should be classified as different positions. Interestingly enough, across both the SVM and Random Forrest models, two players were determined to be playing out of position.

The models were both in agreement that Jayden Gardner and Ben Vander Plas were classified as the incorrect position. Ben was determined to play similar to a Wing Guard, while he is actually listed as a wing forward. Jayden Gardner was predicted to play PF/C, when he is actually classified as a wing forward.

Player Name	Actual Position	SVM Prediction	RF Prediction
Kihei Clark	Scoring PG	Scoring PG	Scoring PG
Jayden Gardner	Wing F	PF/C	PF/C
Francisco Caffaro	C	C	PF/C
Ben Vander Plas	Wing F	Wing G	Wing G
Armaan Franklin	Combo G	Combo G	Wing G
Kadin Shedrick	C	C	C
Chase Coleman	Wing G	Wing G	Wing G
Reece Beekman	Combo G	Combo G	Combo G

Table 2: The predictions of our models on current players of the UVA basketball team.

4 Conclusion

According to the dataset, Jayden Gardner and Ben Vaander Plas are both listed as Wing Forwards. However, when evaluating the film, it does not take a machine learning model to see that they have completely different play styles. Vander Plas excels at three point shooting (he is currently shooting 36.4 percent from behind the three point line) while

Jayden Gardner has yet to attempt a three pointer this season. [4] [5] Both players clearly have different strengths and could potentially reach greater heights if put into the positions suggested by our model. Realistically, it may be too drastic of a change to suggest that Coach Tony Bennett completely restructure the team's lineup based on the results of our model. Nevertheless, Coach Bennett could instead make changes to the plays/system so that Vaander Plas and Gardner have the opportunity to impact the game in a manner suited to their ideal position.

Our model could not only be applied to the UVA basketball team, but any college or even professional basketball team. As long as advanced player data is available at the seasonal level, our model could predict the role that is the best fit for a given player. Coaches could use our model to evaluate a player's skills to get a better understanding of where they can have the best impact on the team. Even if there are other factors keeping a coach from changing a player's position, coaches can still use our model to change schemes in order to give players the best opportunity to succeed. Coaches willing to embrace machine learning could aggressively use all players in their recommended position to optimize the talent on their roster according to our model.

One of the shortcomings of our model that has room for improvement is predicting performance after position change. It would be even more insightful if coaches could have access to projected player statistics following a switch to a different position. We could use very similarly performing players already playing the position in question to somehow project how our player would perform given a position switch. Although this process could be used to effectively find the best offensive role for a player, it falls short when taking defense into consideration. Our model currently gives no insight into how well a player would guard other players at their suggested position switch. This is a very important factor to consider when changing a player's position, because they could become a defensive liability. Ben Vaander Plas, for example, despite having the offensive skills of a guard, would likely struggle to defend other guards due to his lack of speed, but our current model does not reveal this. There are currently little to none credible statistics that capture how well a player can guard other players of different positions at the college level. To accurately predict this we would likely need to implement some sort of unsupervised learning or create our own measure of defensive effectiveness. Including this layer of information to our current results would give coaches a much better overall picture of the outcome of any position change.

5 Contribution

All of our work to date has been done in group sessions, including report writing, searching for datasets, coding, final video editing and recording. Therefore, the workload was evenly distributed. There was no part that any of the team members worked on individually.

Jupyter Notebook available at:

1. `https://colab.research.google.com/drive/13Pt0NA9LYqEcjweAyFnLQjOBhXn7jaPI`

Datasets available at:

1. `https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset`
2. `https://www.kaggle.com/datasets/adityak2003/college-basketball-players-20092021`

References

[1] Christophe Brown. Simple modeling of nba positions using the k-nearest neighbors machine learning algorithm, 2021. [Online]. URL: `https://towardsdatascience.com/simple-modeling-of-nba-positions-using-the-k-nearest-neighbors-machine-learning-algorithm-223b8addb08f`.

[2] Jeremy Lee. Players, positions, and probability in the nba using supervised machine learning to build an nba position classifier, 2021. [Online]. URL: `https://towardsdatascience.comd/players-positions-and-probability-in-the-nba-c54360309616`.

[3] Daniel Myers, Developer of Box Plus/Minus. About box plus/minus (bpm), 2020. [Online]. URL: `https://www.basketball-reference.com/about/bpm2.html`.

[4] ESPN. Season Stats 2022/23 - Ben Vander Plas, n.d. [Online]. URL: `https://www.espn.com/mens-college-basketball/player/stats/_/id/4279448/ben-vander-plas`.

[5] ESPN. Season Stats 2022/23 - Jayden Gardner, n.d. [Online]. URL: `https://www.espn.com/mens-college-basketball/player/stats/_/id/4396614/jayden-gardner`.