Camden Miller
CSS 777 – Big Data

Introduction

Research Question: Can you predict whether any given NBA player is a frontcourt or backcourt player based on their per game statistics?

Objectives and Background Information: In this project we are looking at historical NBA statistics to determine the extent of relationship between a player's stats and their position. Specifically, I am curious if a player's basketball position can limit potential statistical output and if whether stats serve as a reliable identifier for position. This question is even more relevant today in a game that is fueled by social media and sport betting, two developments that put a large emphasis on statistical output over pure winning and positional excellence. This project could help prove that certain positions receive unfair criticism for poor statistical output when their own position limits this output. To examine this question, I focused on NBA historical data from the NBA Stats (1947-present) dataset on Kaggle. This source is a collection of datasets which includes 3 informational files, 3 team files, and 7 player files for National Basketball Association (1950-present), Basketball Association of America (1947-1949) and American Basketball Association (1968-1976) overall historical stats. For my project I am going to focus specifically on the Player Per Game data set which includes 30+ per game statistics such as ppg, rpg, apg, fg%, etc. for 31,870 player seasons (# of rows in table). By labeling PGs and SGs as backcourt players and SFs, PFs, and Cs as frontcourt players this data set is perfect for addressing the question at hand.

Limitations: Possible limitations that are addressed with data cleaning and feature selection during methodology are:

- Game evolution corresponding with positional evolution over time
- Positional anomalies like Magic Johnson, Lebron James (as a PG), and Nikola Jokic
- Records from players with limited minutes

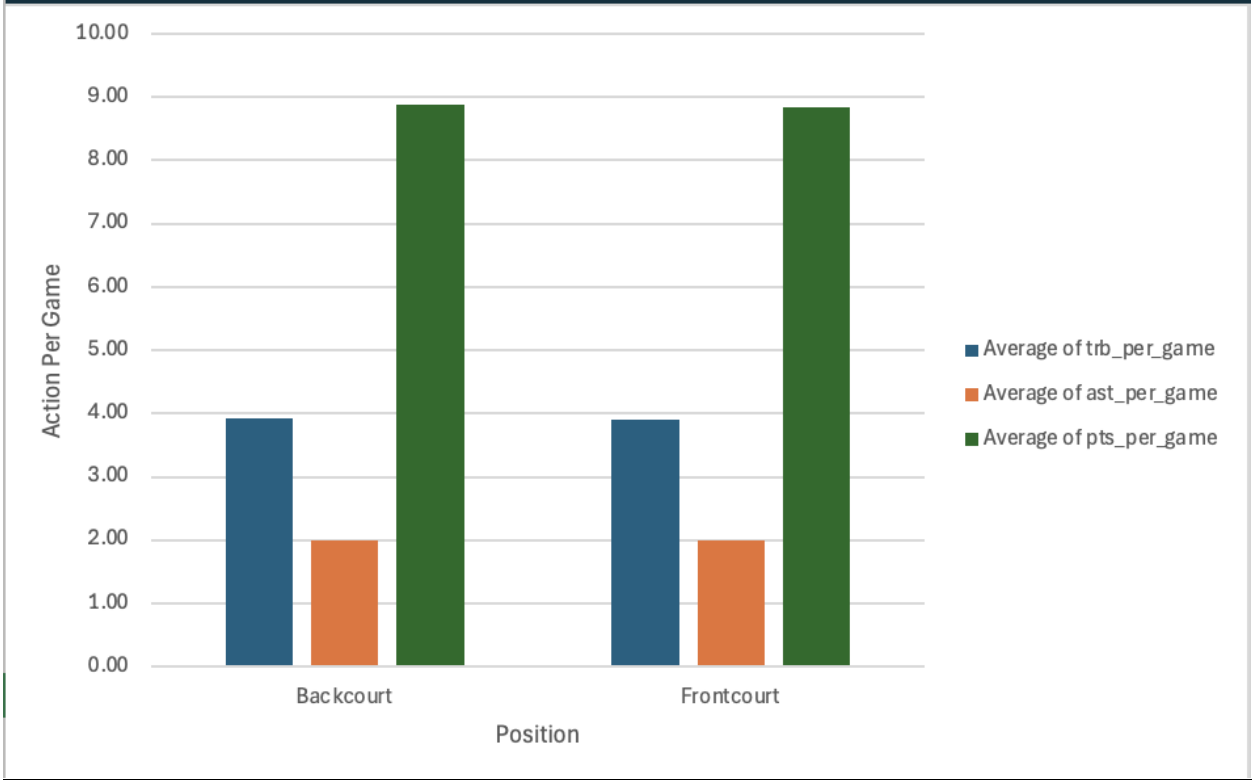
Methodology

Data Preprocessing: One the of the first steps to tackling the research question was data cleaning. One important distinction to note is the eventual use of two datasets that will be addressed in discussion below. One dataset compares backcourt and frontcourt players, while the second dataset compares only Point Guards and Centers. Otherwise, they both include the following preprocessing to follow. To address game evolution, I focused only on player seasons from 1980 and on. This should encompass most of the modern game and modern positional attributes. I also only focused on one positional players as a limited number of records included multiple positions (Ex: Pg-Sg) and would make it difficult for backcourt and frontcourt distinction. I also included players who averaged more than 10 minutes played per game to make sure we are focusing on records where players actually produced enough statistical output in general.

Exploratory Data Analysis: The tables and graphs below show the average points, rebounds, and assists per game for each positional group. The first graph focuses on comparing frontcourt and backcourt players and immediately highlights a potential issue of statistical similarities between frontcourt and backcourt players. Due to foreseeing difficulties in labeling with classification algorithms, I developed a second data set that only compares centers and point guards under all the same data cleaning constraints. As shown below, centers (the stereotypical frontcourt position) average more rebounds while point guards (the stereotypical backcourt position) average more points and assists. This points to a statistical separation across all 18 features trained in the model.

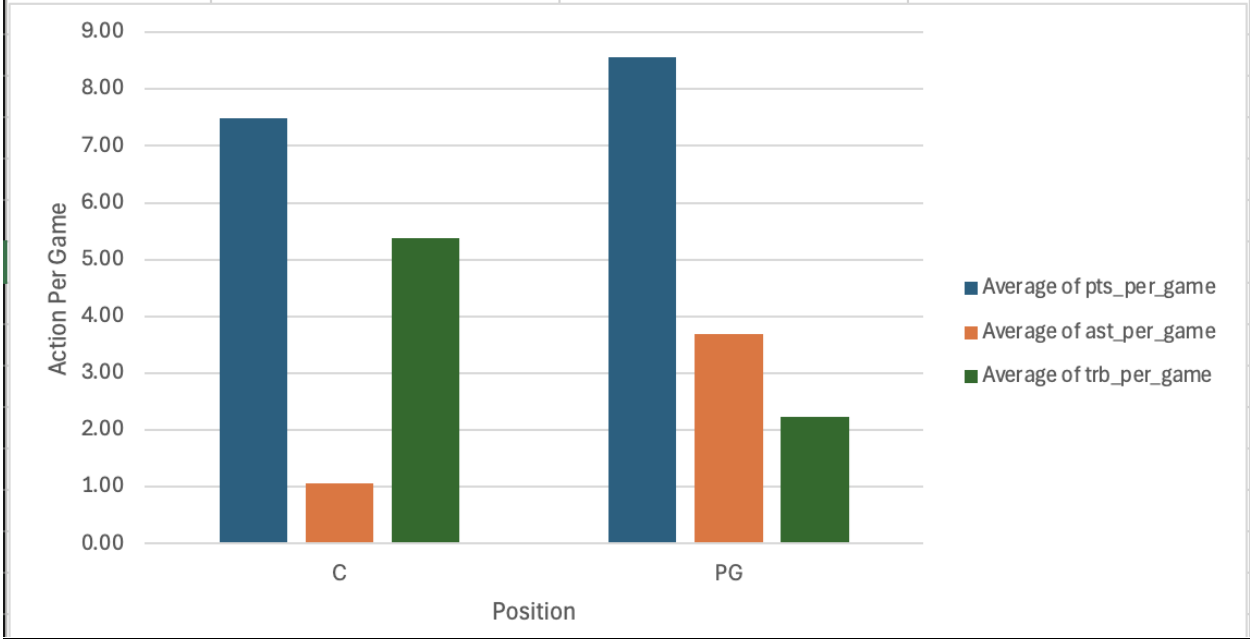
Backcourt vs. Frontcourt

Row Labels	Average of trb_per_game	Average of ast_per_game	Average of pts_per_game
Backcourt	3.92	2.00	8.88
Frontcourt	3.90	1.99	8.84
Grand Total	**3.91**	**1.99**	**8.86**



Point Guards vs. Centers

Position	Average of pts_per_game	Average of ast_per_game	Average of trb_per_game
C	7.48	1.06	5.37
PG	8.56	3.69	2.24
Average	**8.01**	**2.35**	**3.83**



Feature Selection: I chose the following 18 numerical features due to their potential relevance in distinguishing positional attributes for NBA players. This includes metrics such as field goal % (centers are historically poor), assists per game (point guards and shooting guards are strong passers), and rebounds per game (frontcourt players can dominate this statistic). I didn't include 10+ categorical variables such as team, name, division, career year, etc. as they aren't relevant to statistical output and would confuse my classification models.

Features:

fg_per_game	fga_per_game	fg_percent	x3pa_per_game	x3p_percent	x2pa_per_game	x2p_percent	
fta_per_game	ft_percent	orb_per_game	drb_per_game	trb_per_game	ast_per_game	stl_per_game	blk_per_game
tov_per_game	pf_per_game	pts_per_game					

Model Selection: Due to the binary nature of my research question and interest in comparing backcourt and frontcourt statistics I chose to utilize classification machine learning models to evaluate whether player statistics can accurately be used to classify a player's position. If classification performs well, this provides introductory evidence to the statistical limitations of a player's position. The two types of classification I chose were Logistic Regression and Support Vector Machine. Logistic Regression is used for binary classification tasks and utilizes a sigmoid function to map predicted values to 0 or 1

(backcourt or frontcourt / PG vs. C). SVM is a classification algorithm that finds the optimal plan that separates the binary data classes.

Training and Performance Metrics: Starting with my original plan, I completed data cleaning in excel and mapped player positions to binary frontcourt and backcourt metrics. Specifically, point guards and shooting guards are backcourt players and small forwards, power forwards, and centers are frontcourt players. After extracting features and labels I split my data into an 80/20 split training and test data set, checking the distribution of backcourt and frontcourt records to rule out the need for stratification. I first trained the data on a logistic regression model with 100 iterations and regParam = 0.000001, evaluating its performance on the test data set using precision, recall, F1 score and Confusion Matrix metrics. I then trained the data on a SVM model with 100 iterations and regParam = 0.00001 with the same performance metrics listed above. After seeing poor model performance as outlined below and touched on above during EDA, I completed the same model training, but with only point guard and center NBA records. The process was identical but found success with a logistic regression model with 100 iterations and regParam = 0.001 and a SVM model with 100 iterations and regParam = 0.0001.

Results

Model Performances:

Logistic Regression Results:		
Precision:	0.0013	1 = Backcourt
Recall:	0.75	0 = Frontcourt
F1 Score:	0.0027	
Total Time(s):	10.18	
Confusion Matrix	Positive (1)	Negative (1)
Positive (1)	TP: 3442	FP: 2257
Negative (0)	FN: 1	TN: 3

Logistic Regression Results:		
Precision:	0.9685	1 = Point Guard
Recall:	0.9685	0 = Center
F1 Score:	0.9685	
Total Time(s):	9.87	
Confusion Matrix	Positive (1)	Negative (1)
Positive (1)	TP: 1145	FP: 35
Negative (0)	FN: 35	TN: 1076

Support Vector Machine Results:		
Precision:	0.013	1 = Backcourt
Recall:	0.6	0 = Frontcourt
F1 Score:	0.0026	
Total Time(s):	6.21	
Confusion Matrix	Positive (1)	Negative (1)
Positive (1)	TP: 3334	FP: 2305
Negative (0)	FN: 2	TN: 3

Support Vector Machine Results:		
Precision:	0.9658	1 = Point Guard
Recall:	0.9684	0 = Center
F1 Score:	0.9671	
Total Time(s):	5.44	
Confusion Matrix	Positive (1)	Negative (1)
Positive (1)	TP: 1145	FP: 38
Negative (0)	FN: 35	TN: 1073

*The tables on the left are for the first dataset that separates records by backcourt vs frontcourt players while the tables on the right are for the second dataset that separates records by point guards and centers.

Discussion

Looking at the results above it is evident that I ran into problems addressing my initial research question. As exhibited by the EDA and poor model performance, there was limited statistical separation between frontcourt and backcourt players. Specifically, the logistic regression model had a precision of less than 1% with a F1 score of 0.27%. The model classified all but 4 records as backcourt players. Performing similarly, the SVM model had a precision and F1 score less than 1% and classified most records as backcourt players. This led to modifications in my methodology and need to readdress my question. Because separation of statistics appeared to be the issue, I decided to instead compare only Point Guards and Centers, the two positions that stereotypically represent the backcourt and frontcourt positions respectively. Without closely related positions like Small Forwards and Shooting Guards, I suspected the models would perform better at accurately classifying a player's positions by stats. Looking first at the logistic regression model, I saw immediate success with the changes to the data. This model had a precision, recall, and F1 score of 0.9685 and only misclassified 70 records. Similarly, the SVM model also performed well with precision, recall, and F1 score all above 0.96 and only 73 records misclassified. Therefore, comparing all 4 models across both datasets, my logistic regression model on the second data version saw the greatest success in accurately classifying a player's position based on stats. I also wanted to note that the SVM model ran almost twice as fast (5s vs. 10s) as the logistic regression model, suggesting a possible advantage of SVM with bigger data.

Conclusion

Overall, while I wasn't directly able to answer my research question, I was able to build two models that successfully classified point guards and centers as their correct position based on 18 relevant NBA statistics. This builds on the question posed in my introduction, supporting the notion that a player's statistical output is limited by their position on the court. For example, centers are grabbing the majority of rebounds and can struggle from the free throw line while point guards are accumulating assists and steals.

While I did ultimately find success, I also encountered numerous limitations with my data and models. My initial plan of classifying frontcourt and backcourt players failed due to this division not separating class statistics enough. As shown with my EDA, many major statistics are almost identical suggesting both classes share similar distributions. The variability of coaching and playing style could also limit the suspected separation, as certain teams/players are more offensively and defensively minded.

In the future, the use of clustering algorithms could be effective in further delving into my research data and discover what clusters exist within the data. For example, it may be possible to cluster players as offensively or defensively minded or as oversized or undersized.