ARE NBA STATISTICS PREDICTIVE OF PLAYING POSITION?

Camden Miller
CS 777

Introduction

- Research Question: Can you predict whether any given NBA player is a frontcourt or backcourt player based on their per game statistics?
- Can a player's basketball position limit potential statistical output and are a reliable identifier for position?
  - Relevance in today's game with social media and betting
- Data Overview
  - NBA Historical Player Per Game data set
  - 30+ per game statistics such as ppg, rpg, apg, fg%, etc. for 31,870 player seasons
  - Ability to label positions as frontcourt or backcourt
- Limitations
  - Game evolution → Positional Evolution
  - Positional anomalies
  - Records from players with limited minutes

Methodology

- Data Preprocessing
  - Two dataset versions: backcourt vs. frontcourt / point guards vs. centers
  - Player seasons 1980+
  - Exclusion of limited multi-positional players
  - Over 10 minutes played per game on average
- Feature Selection
  - 18 numerical features
  - Excluded categorical variables

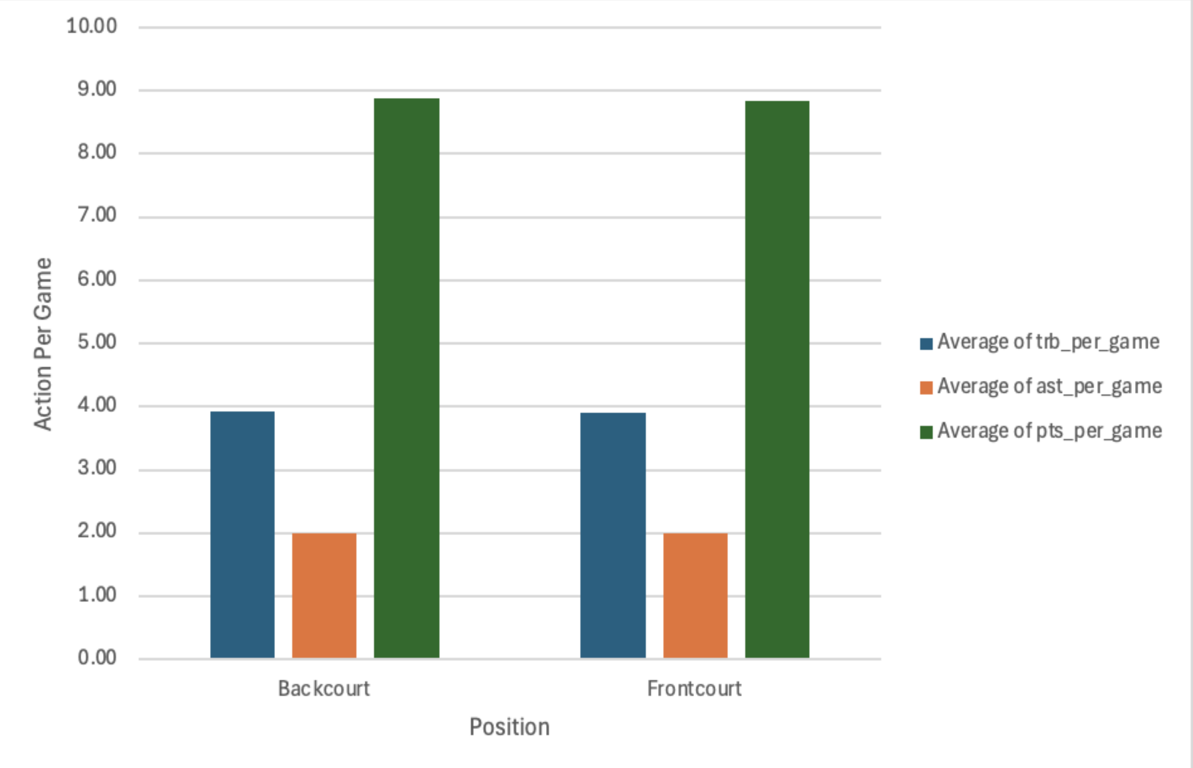
Features:

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					

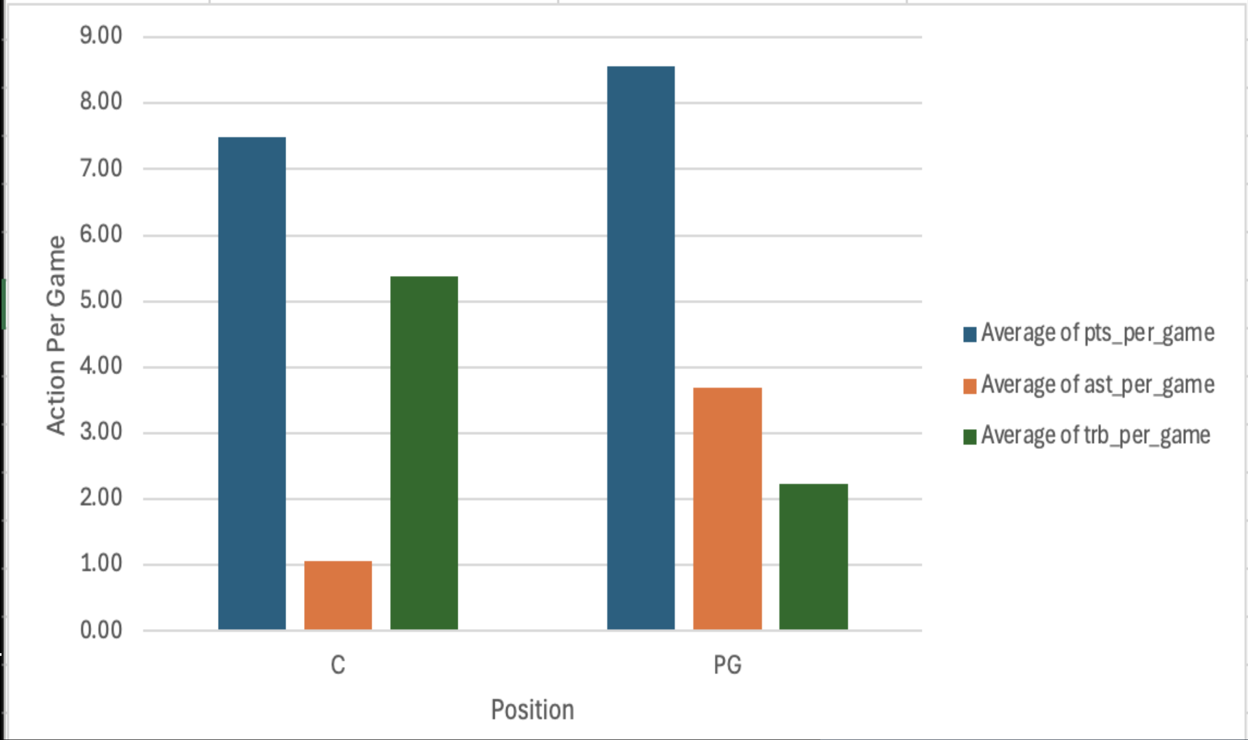
Methodology

- Exploratory Data Analysis

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**



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**



Methodology

- Model Selection
  - Binary nature of research question → Classification
  - Logistic Regression: Sigmoid function to map predicted values to classes
  - Support Vector Machine: Optimal plane to separate classes
- Training and Performance Metrics
  - 80/20 Training and Test Data Split
    - No need for stratification
  - Backcourt vs. Frontcourt Log Reg: 100 iterations and regParam = 0.000001
  - Backcourt vs. Frontcourt SVM: 100 iterations and regParam = 0.00001
  - Point Guards vs. Centers Log Reg: 100 iterations and regParam = 0.0001
  - Point Guards vs. Centers SVM: 100 iterations and regParam = 0.0001
  - Metrics: Precision, Recall, F1 Score, Confusion Matrix

Results

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

Discussion + Conclusion

- First model training trial
- Potential issues based on results
- Changes to data labeling and filtering
- Second model training trial
- Model success and performance
- Comparison of model runtime
- Limitations
  - Frontcourt vs. Backcourt
  - Coaching and Playing Styles
- Next steps with clustering