Classifying Modern NBA Players (& Teams)

Tejas Bala

Introduction

Initially this project was motivated from an idea to provide a Recommendation Engine for NBA teams to help them target Free Agent players that will be beneficial to them. The engine will be informed by similarity models for players and team. For the purpose of this course's projects I decided on just the similarity models.

Data Collection

Data was collected from Basketball-reference.com All data was collected for the following NBA Seasons: 2015-2016, 2016-2017, 2017-2018, 2018-2019 (collected upon the conclusion of the season).

Team data was taken from the Miscellaneous Team Stats table since it included high-level statistics instead of having to compute them myself. Here is a view of the Misc. Team Stats table:

Miscellaneous Stats Share & more ▾ Glossary

															Offense Four Factors				Defense Four Factors						
Rk	Team	Age	W	L	PW	PL	MOV	SOS	SRS	ORtg	DRtg	NRtg	Pace	FTr	3PAr	TS%	eFG%	TOV%	ORB%	FT/FGA	eFG%	TOV%	DRB%	FT/FGA	Arena	Attend.	Attend./G
1	Houston Rockets*	29.8	65	17	61	21	8.48	-0.27	8.21	114.7	106.1	+8.6	97.6	.298	.502	.590	.551	12.7	21.3	.233	.521	13.4	79.9	.171	Toyota Center	732,722	17,871
2	Toronto Raptors*	25.8	59	23	60	22	7.78	-0.49	7.29	113.8	105.9	+7.9	97.4	.250	.377	.575	.539	12.1	23.0	.198	.501	13.0	77.7	.212	Air Canada Centre	813,431	19,840

Player data was collected from Per 100 Possession Stats and Advanced Stats tables. Here they are below:

Player Per 100 Poss Share & more ▾ Hide non-qualifiers for rate stats Glossary Hide Partial Rows

Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	ORtg	DRtg
1	Alex Abrines	SG	24	OKC	75	8	1134	5.0	12.7	.395	3.7	9.7	.380	1.4	3.1	.443	1.7	2.0	.848	1.1	3.9	5.0	1.2	1.7	0.4	1.1	5.4	15.4	116	110
2	Quincy Acy	PF	27	BRK	70	8	1359	4.6	13.0	.356	3.6	10.4	.349	1.0	2.6	.384	1.8	2.1	.817	1.4	7.8	9.2	2.0	1.2	1.0	2.1	5.3	14.7	99	110

Advanced Share & more ▾ Hide non-qualifiers for rate stats Glossary Hide Partial Rows

Rk	Player	Pos	Age	Tm	G	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP
1	Alex Abrines	SG	24	OKC	75	1134	9.0	.567	.759	.158	2.5	8.9	5.6	3.4	1.7	0.6	7.4	12.7	1.3	1.0	2.2	.094	-0.5	-1.7	-2.2	-0.1
2	Quincy Acy	PF	27	BRK	70	1359	8.2	.525	.800	.164	3.1	17.1	10.0	6.0	1.2	1.6	13.3	14.4	-0.1	1.1	1.0	.036	-2.0	-0.2	-2.2	-0.1

Data Dictionary

Including some descriptions for the advanced statistics that were used in the analysis:

3PAr – (Three-point Attempt Rate) Percentage of shot attempts from 3P range

FTr – (Free-Throw Rate) Number of FT attempts per shot attempt
Pace – An estimate of Possessions per 48 minutes (length of average game)

WS – (Win Share) An estimate of the number of wins contributed by a player
OWS – Offensive Win Share
DWS – Defensive Win Share

Data Cleaning

In order to get my Player data in a form to perform analysis and apply models on I completed the following steps:

- Merged player tables together
- Added [Year] to player name (as I would have the same player over 4 seasons)
- Mapped Team Abbreviation to Team Full Name so that both were present in the table
- (Feature Engineering) Added a rough All-Star Caliber flag that was based on each player's Win Share

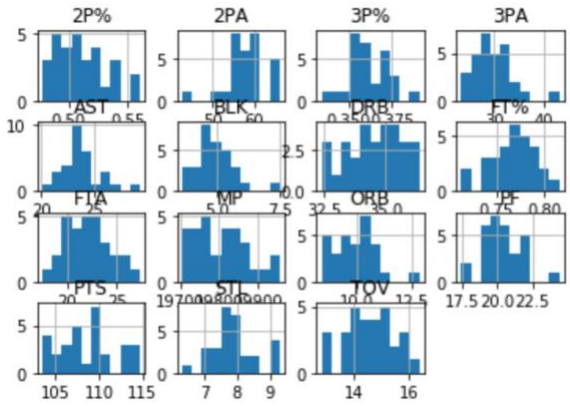
In order to get my Team data in a form to perform analysis and apply models on I completed the following steps:

- (Feature Engineering) Create Playoff flag for teams that made the playoffs that season, indicated in data by an asterisk after team name: [Team Name]*
- Remove the asterisk after team name for merging later on
- Added [Year] to team name (as I would have the same team over 4 seasons)
- Renaming columns appropriately that were misread from dual headers from the data source
- Mapped Team Abbreviation to Team Full Name so that both were present in the table

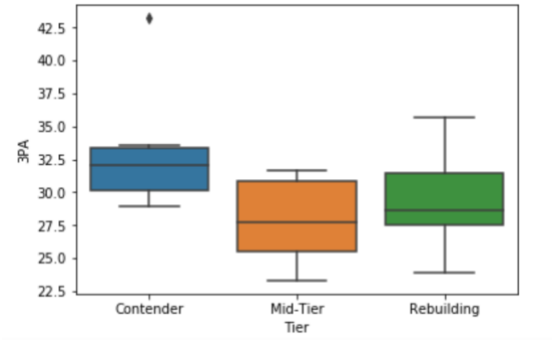
Data Exploration

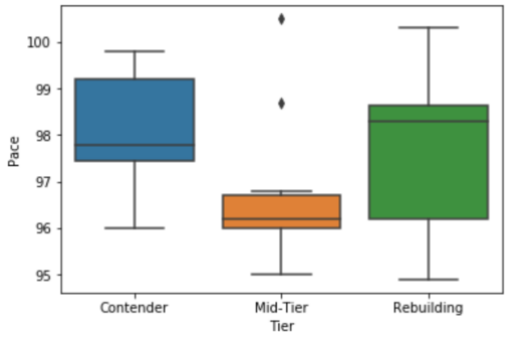
I completed some Data Exploration with my initial set of data pulled for Team which included 4 additional tables to the Miscellaneous Stats, they were: Per100 Team Stats, Per100 Opponent Stats, Team Shooting, Opponent Shooting.

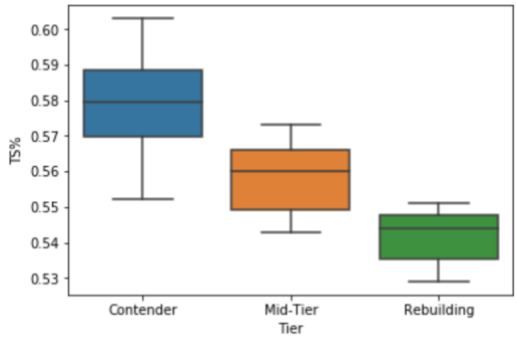
I plotted histograms of features to understand their distributions and identify any skews:



I also created a Tier classification based on Team Wins and observed separation of those classes by different features. Here are a few:







Lastly, I created a correlaton matrix and identified if any features were high collinear with other features and this influenced my variable choices in my models.

Player Clustering Analysis

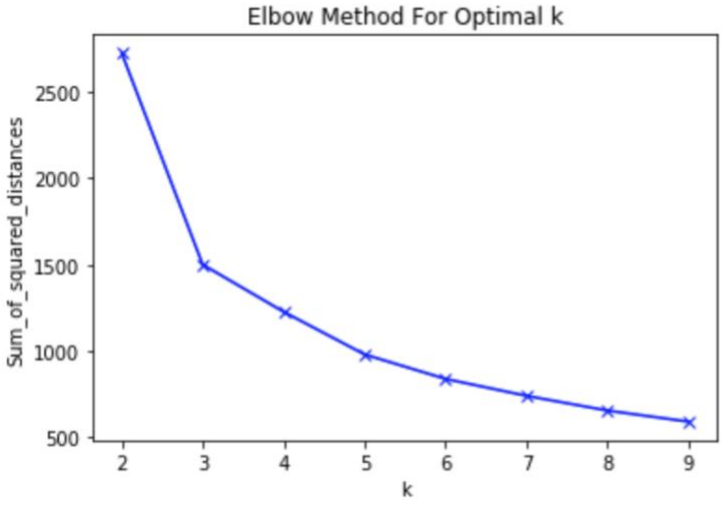
The intuition behind my player clustering approach was that we will be able to re-classify NBA players based on various statistics that will determine style of play and behavior on the court. With the clean data our first step is feature selection; we want to make sure we drop most features that are representative of a player's success on the court so that we can extract style of play. I also decided to keep mostly the percentage statistics and drop the raw statistics

to create clusters of players that are frequently used as well as not so frequently used. We are left with:

Pos_x	Tm_x	3P%	2P%	FT%	ORB	DRB	TRB	AST	STL	BLK	PF	ORtg	DRtg	Player_year	Year	TS%	3PAr	FTr	OWS	DWS	Abrv
SG	OKC	0.323	0.500	0.923	0.4	3.4	3.8	1.6	1.3	0.5	4.2	103.0	111	Alex Abrines\abrinal01/2019	2019	0.507	0.809	0.083	0.1	0.6	OKC

Once we have selected our features we can perform LDA to maximize the separation between the existing classes (traditional positions). I decided on 2 principal components for the purpose of plotting and visualization since that would be a big part of how we assess the clusters. Before performing clusters I did an analysis to see what would be the optimal number of clusters. This was the output:

```
For n_clusters = 2 The average silhouette_score is : 0.47980387092455745
For n_clusters = 3 The average silhouette_score is : 0.48652575276179294
For n_clusters = 4 The average silhouette_score is : 0.41871453865282027
For n_clusters = 5 The average silhouette_score is : 0.38314284114887065
For n_clusters = 6 The average silhouette_score is : 0.3496620672892624
For n_clusters = 7 The average silhouette_score is : 0.35747896389759176
For n_clusters = 8 The average silhouette_score is : 0.3396943606868568
For n_clusters = 9 The average silhouette_score is : 0.3440934944151359
```



There is a clear elbow at K = 3 and he silhouette score is very high with K < 5 but I wanted clusters that further broke down the 5 main positions that players are labelled as so I tried clustering with K = [6:9] and assessed the clusters. I decided on K = 8 as a good clustering output. You can see the player clusters here.

Team Clustering Analysis

The intuition behind my clustering approach was that if we wanted to assess team similarity for roster recommendations it would be very useful to consider the roster as well as a handful of other statistics to drive the clustering model.

Once we have the player cluster analysis complete we merge it with the full cleaned player dataset. This is done so that for the players with multiple rows (changed teams mid season) will be reflected in the data when we aggregate up to the team level. To aggregate up to the team level we iterate through the merged data by team and select the top 9 players based on minutes played. This is done because the players at the end of the roster don't play that much and are not reflective of the team's style of play. After filtering columns we are left with this output:

	Player_year	Team_full	Cluster
0	Jayson Tatum\tatumja01/2019	Boston Celtics/2019	Primary Distributing Ballhandler
1	Kyrie Irving\irvinky01/2019	Boston Celtics/2019	Pace & Space Forward

After creating dummy variables for the cluster labels and grouping by team we get the roster composition for each team. Last step is merging on team to add 3PAr, FTAr, Pace for clustering:

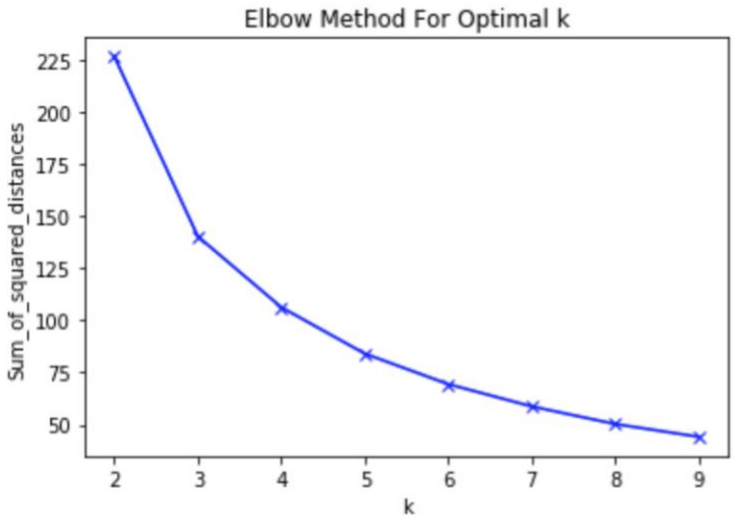
	3 & D Wing	Combo Guard	Defensive Forward	Offensive Big Man	Pace & Space Forward	Post-Heavy Big Man	Primary Distributing Ballhandler	Slashing Guard	Team	Team_year	3PAr	FTr	Pace
0	0	1	2	0	0	1	5	0	Atlanta Hawks/2019	Atlanta Hawks/2019	0.403	0.255	103.9
1	2	0	3	0	1	0	3	0	Boston Celtics/2019	Boston Celtics/2019	0.381	0.215	99.6

Before PCA and Clustering we perform Z-score standardization on our data so that each element has a equal influence on the model. Principal Component Analysis is performed for dimensionality reduction. I decided on 2 principal components for the purpose of plotting and visualization since that would be a big part of how we assess the clusters. Below is the output of the PCA:

```
Total Variance: 3.45150930044816
Percentage of Var represented by PC:
[0.54 0.46]
Cumulative Percentage of Var represented by PC:
[0.54 1.  ]
```

With the PCA complete we run analysis to determine the optimal number of clusters. Here we look at the Shadow Score and the Elbow plot:

```
For n_clusters = 2 The average silhouette_score is : 0.3346575844079631
For n_clusters = 3 The average silhouette_score is : 0.36466926385599585
For n_clusters = 4 The average silhouette_score is : 0.3819167344619265
For n_clusters = 5 The average silhouette_score is : 0.35975992356900105
For n_clusters = 6 The average silhouette_score is : 0.3604037607630322
For n_clusters = 7 The average silhouette_score is : 0.3477480127152779
For n_clusters = 8 The average silhouette_score is : 0.355096812200371
For n_clusters = 9 The average silhouette_score is : 0.3648776156013548
```



There is no clear elbow in the plot but the silhouette score peaks at K=4 and 6. I tried both 4 and 6 and assessed the clusters to decide that 6 lead to a good clustering of teams. You can see the team clusters here.

Conclusion

These methods were highly successful in re-classifying modern NBA players into 8 new groupings: Post-Heavy Big, Offensive Big, Defensive Forward, Pace & Space Forward, 3 & D Wing, Slashing Guard, Combo Guard and Primary Ball-Handler/Distributor. The team clustering was fairly insightful in terms of seeing which teams are built similarly and I think this can definitely help drive recommendations for roster decisions.

Not having a hard evaluation metric or criteria is a huge threat to validity in my opinion. Throughout the course of this project I discussed and iterated with domain experts many times to make sure that the output made sense.

If I were to do this project with this data over again, I would focus much more on trends. How a player progresses over his career, does he change clusters? How do the top teams in the league change year over year? Is the best team usually built in one specific way? I look forward to exploring these questions further. You can find all of my work on my GitHub.