Predicting New Contracts in NBA Free Agency

Calvin Szeto

May 2, 2014

1 Introduction

1.1 Motivation

The goal of this project is to find meaningful relationships between player performance data and the salaries they earn. These relationships provide opportunities for basketball fans and professionals to enhance their understanding of the NBA player market. For a fan, the visualizations provide insight into whether their favorite players are truly overpaid or underpaid. For a professional, the models allow them to predict the market for certain players in certain years, and they can adjust their contract negotiations accordingly.

1.2 Problem

Prediction of salaries is a non-trivial problem. The amount a player is paid depends on a large number of features, not only performance and skill data, but also personality, athleticism, and the current market. Further, performance data such as box scores and even advanced statistics are still not completely representative of a player's true skillset. Nevertheless, there is enough meaningful substance in available data to cluster similar players together, and from these clusters perhaps find a reasonable estimate of the market for a certain type of player.

1.3 Data and Algorithms

To satisfy this goal, multiple algorithms are applied to scraped data and compared side-by-side. These algorithms include k-Means using both the Lloyd-Forgy method and Hartigan-Wong method, Spectral Clustering based on a k-Nearest-Neighbors graph, and k-Means with Principal Component Analysis. These clustering algorithms allow us to cluster players meaningfully, and to examine salary statistics for individual clusters.

The data used includes regular box score data and advanced statistics downloaded from Basketball-Reference [2], salary data scraped from Basketball-Reference [1], and free agency lists scraped from Wikipedia [3]. These datasets are joined

to relate a player's performance in a given season with their salary, and free agency data is used to identify persons of interest.

2 Implementation

2.1 k-Means

First, the forementioned box score data was clustered using the k-Means algorithm with the intention of clustering players by skillset and quality of performance.

To begin, the version of k-means commonly known as the Lloyd-Forgy method was implemented in R and applied to per-game datasets. The algorithm chooses an initial set of centroids at random, although each feature of the centers are kept within the bounds of the maximum and minimum values of the corresponding feature in the actual dataset, so as not to create unusually random centroids. An interesting problem initially encountered was the occurence of empty clusters early in the execution of the algorithm; this was easily resolved by moving a random point into each of the empty clusters. The algorithm converges quickly afterwards. A perhaps better implementation would be to choose the point with the highest error and move that into the empty cluster instead. This approach would be more data-driven and help separate clusters which are less dense.

The algorithm was run with several different values of k. The squared-sum-of-errors of some of these attempts are plotted here.

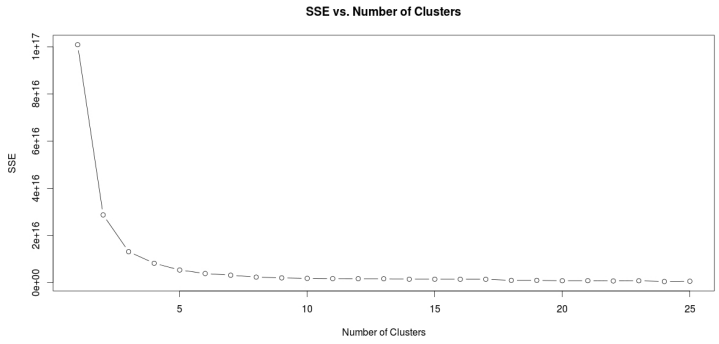

Figure 1: Finding an optimal number of clusters using an SSE graph.

As expected the SSE descreases sharply within the first few clusters, and slowly afterwards. In addition to examining the errors, however, we also browse the clusters manually in hopes of finding a clustering which divides players sensibly among positions, skillsets, and quality. From such manual investigation, we find that 10-15 clusters is the sweetspot for meaningful groups.

1102	Troy Murphy	3
26	Amar'e Stoudemire	4
215	Chris Paul	4
377	Dwyane Wade	4
753	LeBron James	4
25	Alonzo Mourning	5
29	Amir Johnson	5
30	Andray Blatche	5
32	Andray Blatche	5
46	Antawn Jamison	5
68	Antonio McDyess	5
88	Ben Wallace	5
89	Ben Wallace	5
109	Brandon Bass	5
112	Brandon Bass	5
144	Carl Landry	5
211	Chris Kaman	5
306	DeAndre Jordan	5
408	Elton Brand	5
446	Glen Davis	5
500	Jamario Moon	5
562	Jason Smith	5
572	JaVale McGee	5
619	Jermaine O'Neal	5
647	Jonas Jerebko	5
651	Jordan Hill	5

Figure 2: A sample of players ordered by cluster.

With the number of clusters decided and the algorithm converging smoothly, we run an instance of k-means on the dataset and merge it back with the salary data. We now have a dataset of player per-game statistics separated by season, joined with the salaries and cluster assignments.

	row.names	Player	season	cluster	PTS	TRB	AST	BLK	STL	FG	FT	X3P	TOV	amount
1	1	Aaron Brooks	12-13	8	7.1	1.5	2.2	0.2	0.6	2.7	0.8	0.9	1.3	3.250000
2	3	Aaron Gray	11-12	3	3.9	5.7	0.6	0.3	0.4	1.7	0.5	0.0	1.0	2.500000
3	5	Aaron Gray	12-13	7	2.8	3.2	0.8	0.1	0.2	1.1	0.5	0.0	0.9	2.575000
4	7	Aaron Williams	06-07	7	2.0	2.2	0.2	0.4	0.2	0.8	0.5	0.0	0.5	1.750000
5	8	Acie Law	10-11	10	4.2	1.2	1.6	0.0	0.6	1.6	0.9	0.1	0.8	0.656030
6	9	Adonal Foyle	07-08	7	1.9	2.5	0.2	0.5	0.2	0.9	0.2	0.0	0.4	1.219590
7	11	Adrian Griffin	06-07	7	2.5	2.0	1.1	0.1	0.6	1.1	0.3	0.0	0.7	1.475000
8	12	Alan Anderson	12-13	11	10.7	2.3	1.6	0.1	0.7	3.6	1.9	1.5	1.2	0.885120
9	14	Alexander Johnson	07-08	10	4.2	2.2	0.3	0.2	0.3	1.4	1.3	0.0	1.0	0.687456
10	16	Alexey Shved	12-13	11	8.6	2.3	3.7	0.4	0.7	3.1	1.4	1.1	1.9	2.934610
11	18	Al Harrington	10-11	11	10.5	4.5	1.4	0.1	0.5	3.8	1.2	1.6	1.5	5.765000
12	19	Allen Iverson	09-10	1	13.8	2.8	4.0	0.1	0.7	4.9	3.7	0.3	2.3	1.029794
13	21	Alonzo Gee	10-11	3	5.9	3.3	0.7	0.3	0.7	2.2	1.3	0.3	1.0	0.125538
14	23	Alonzo Gee	12-13	11	10.3	3.9	1.6	0.4	1.3	3.7	2.0	0.8	1.6	3.500000
15	25	Alonzo Mourning	06-07	5	8.6	4.5	0.2	2.3	0.2	3.1	2.4	0.0	1.7	2.500000
16	26	Amar'e Stoudemire	10-11	4	25.3	8.2	2.6	1.9	0.9	9.5	6.1	0.1	3.2	16.486611
17	27	Amir Johnson	07-08	3	3.6	3.8	0.5	1.3	0.4	1.5	0.6	0.0	0.6	3.666666
18	29	Amir Johnson	10-11	5	9.6	6.4	1.1	1.2	0.7	3.9	1.8	0.0	1.0	5.000000
19	30	Andray Blatche	07-08	5	7.5	5.2	1.1	1.4	0.6	3.1	1.3	0.0	1.4	2.470000
20	32	Andray Blatche	12-13	5	10.3	5.1	1.0	0.7	1.0	4.2	1.8	0.0	1.5	1.146337

Figure 3: A sample of the clustered dataset with a salary feature.

Finally, we are particularly interested in free agents, so we trim the current dataset to only include player seasons where the player is fresh off a new contract from free agency. This allows us to examine the type of pay players from different clusters receive on average, and to also observe players who were particularly overpaid or underpaid in the past compared to the norm.

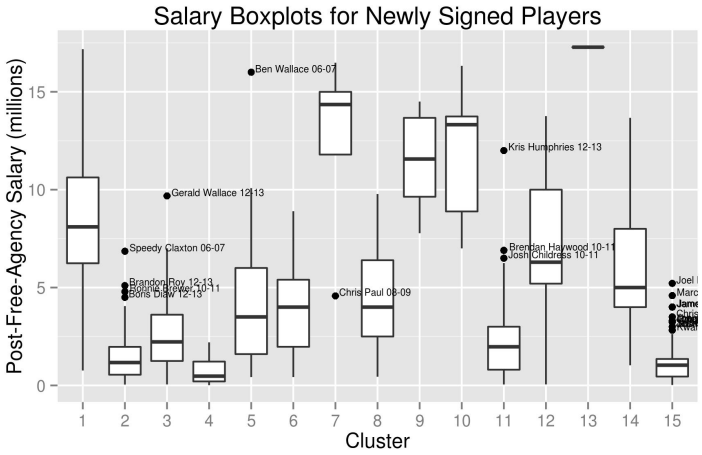

Figure 4: k-Means using the Lloyd-Forgy method.

In addition to the Lloyd-Forgy algorithm, another popular version of k-means is the Hartigan-Wong method. This method gives us similar results:

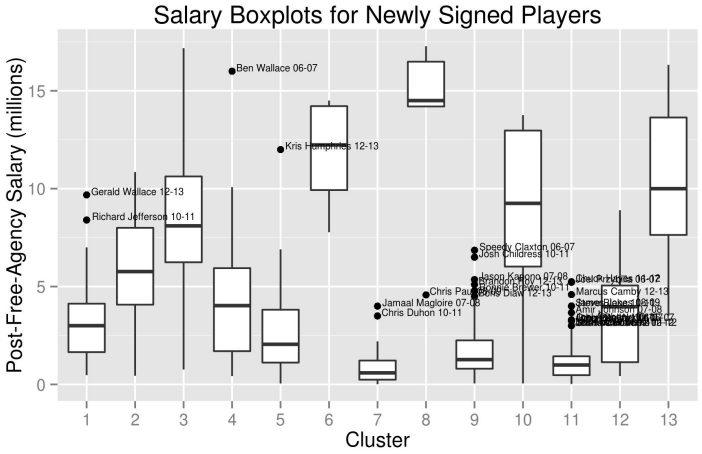

Figure 5: k-Means using the Hartigan-Wong method.

A major difference of Hartigan-Wong is the performance - compared to our implementation of the Lloyd-Forgy algorithm, the Hartigan-Wong implementation in the R libraries runs considerably faster. Note that every run of k-Means gives similar but different results - since we intitialize our centroids randomly, the algorithm will converge on different centroids on every execution.

2.2 k-Nearest-Neighbors

Another common clustering algoritm is a modification of the common classification algorithm, k-Nearest-Neighbors. This clustering algorithm has a few advantages - most importantly, it benefits from data which is not necessarily meaningful using Euclidean distance, as k-Means does.

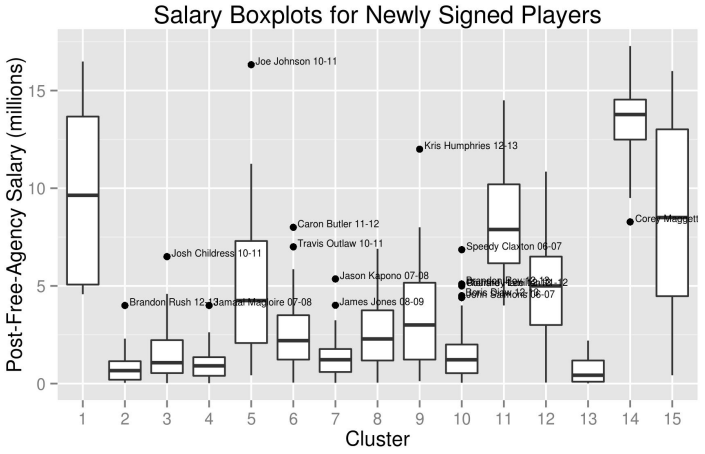

Figure 6: Spectral Clustering using a k-Nearest-Neighbors graph.

Here, we use Spectral Clustering based on a k-Nearest-Neighbors graph for similarity. Spectral Clustering calculates the eigenvectors of the data before clustering in order to reduce the dimensions - see the Principal Component Analysis subsection for more.

2.3 Principal Component Analysis

An issue which is common among machine learning situations is high-dimensional data. Most data is naturally multidimensional - for example, the simple box score data already has at least 9 major features, as well as other possibly important features such as age, games played, etc. Multidimensional data is difficult to reason with intuitively, and can introduce a lot of noise. One common method to reduce the number of dimensions is Principal Component Analysis, which decomposes the data matrix into eigenvectors and sorts them based on eigenvalues. The resulting vectors happen to be sorted by the amount of variance which they account for, with the highest eigenvectors tending to account for a large percentage of the total variance.

We use this method to our advantage by running PCA on the players dataset and running k-Means as before using the first two major eigenvectors as the input. The results are very similar:

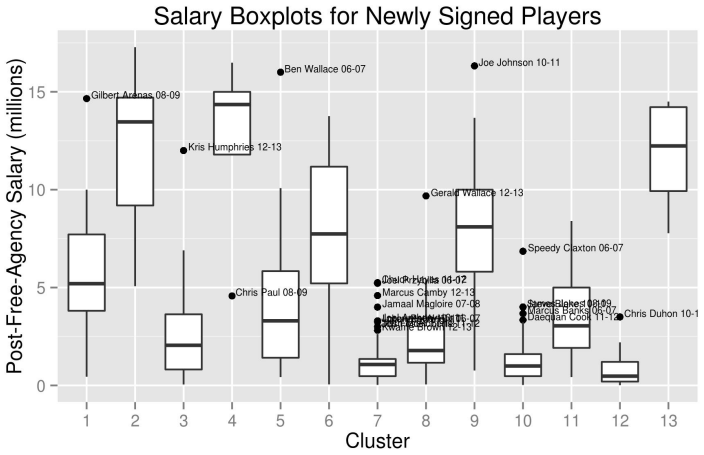

Figure 7: k-Means using the Hartigan-Wong method and PCA.

3 Analysis

From the results above, we can draw a few conclusions, as well as predict salaries for upcoming free agents.

First, from observing the PCA eigenvectors (not shown), we find that the features which provide the most variance in salary are indeed minutes played and points. This simply confirms the intuition that players are paid largely based on skill, and skilled players tend to play more minutes and score more points. This also hints at one of the weaknesses of our analysis: since box score data is heavily offense-oriented, defensive players will not be represented well in our clusterings.

Indeed, if we observe the k-Means boxplots, some of the high outliers are defensive players. However, in general, the outliers in the data are players who are/were generally agreed as "overpaid". For example, in Figure 7, Joe Johnson in 2010-11 had a relatively disappointing season despite being paid one of the highest salaries in the NBA. Similarly, Gilbert Arenas in 2008-09 was a famously crushing season for Gilbert and the Wizards, even though he was cashing in tens of millions of dollars.

Similarly, we can observe a few players who are low outliers, meaning they were relatively underpaid despite performing in an exceptional cluster. In Figure 4 and Figure 7, we see Chris Paul making under 5 million a year, despite playing at the level of players who make a median of just under 15 million.

Finally, we evaluate some values for upcoming free agents in 2014:

Player	Current Team	Cluster	Median Salary of Cluster (millions)
Avery Bradley	Boston Celtics	1	6.427
Luol Deng	Chicago Bulls	1	6.427
Aaron Brooks	Denver Nuggets	11	3
Jordan Hamilton	Houston Rockets	6	1.352
Lance Stephenson	Indiana Pacers	9	7.147
Eric Bledsoe	Phoenix Suns	5	13.32
Kevin Seraphin	Washington Wizards	12	1.137
Trevor Ariza	Washington Wizards	1	6.427

How accurate are these results? We will find out this offseason.

4 Bibliography

References

[1] Basketball-Reference. *NBA & ABA Player Directory*. URL: `http://www.basketball-reference.com/players/`.

[2] Basketball-Reference. *NBA Stats*. URL: `http://www.basketball-reference.com/leagues/NBA_2014_per_game.html`.

[3] Wikipedia. *List of NBA Season Transactions*. URL: `http://en.wikipedia.org/wiki/List_of_2012-13_NBA_season_transactions#Signed_from_free_agency`.