# Predicting New Contracts in NBA Free Agency

**Calvin Szeto**

## Introduction

### 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.

### 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.

### 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. In addition, Multiple Linear Regression will allow us to train a model to directly predict future salaries.

The data used includes regular box score data and advanced statistics downloaded from Basketball-Reference, salary data scraped from Basketball-Reference, and free agency lists scraped from Wikipedia. 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.

## Implementation

### $k$-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.

TODO: Insert plot of SSE

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.

TODO: Insert image of a cluster that makes sense.

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.

TODO: Insert image of some rows of clustered dataset

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.

TODO: Insert boxplots here

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

TODO: Insert boxplots here

### $k$-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.

### Principal Component Analysis

### Linear Regression

### Results

## Analysis

## Bibliography