ISyE 6740 – Fall 2023
Project Proposal

Team Member Names
Declan Cunningham, David LLanso, Salman Mukhi

Project Title
Classifying NBA Players by Caliber

Problem Statement
At the beginning of an NBA season, it is difficult to guess which players will achieve certain accomplishments by the end of the season. Examples of player accomplishments include being one of the top starters for a team or being selected for the All-Star game or an All-NBA team. To gain insight into whether a player might achieve one of these feats, it is helpful to know their caliber based on how well they performed in a recent season. This project aims to empirically determine the calibers of NBA players in a recent season based on an unsupervised clustering analysis of player performances in many decades' worth of historical seasons. Our hypothesis is that we will be able to identify clearly separable clusters that categorize a player's caliber. Examples of cluster labels are the following in descending order of player caliber: superstar, all-star, starter, role-player, and a bench-player. We also believe training a classification model on the data labeled through unsupervised learning will result in a model that can accurately predict a player's caliber.

Data Source
The primary data source that we plan to use in this project is Basketball-Reference.com, a website that tracks and records professional basketball statistics from the NBA and WNBA leagues. Specifically, we retrieved a collection of 21 CSV files that contain ABA (past NBA competitor), BAA (NBA predecessor), and NBA league team and player statistics ranging from 1947 to the present. Many of the files contain unique player and season identifier variables, enabling us to easily combine datasets as needed. Additionally, there is a helpful glossary on Basketball-Reference.com that provides explanations of every statistic found in the datasets.

Methodology
At a high level, our goal is to first perform unsupervised learning on the data to label players with a caliber. We will then label the data points in each cluster and train several supervised learning classifiers. Lastly, we will compare the different supervised classifiers to determine which one produces the most accurate results with respect to the labels determined in k-means clustering.

Prior to getting started, we will need to explore all the data and perform standard preprocessing steps to ensure that we are working with the appropriate quantity and adequately cleansed/transformed subsets that we will need for our models. The underlying data contains records from each player per season they've participated

in so we will want to first aggregate the data in a meaningful way so that there is only 1 record per player. As the CSV files contain several variables, we also intend to perform variable selection. This will be performed using a combination of domain knowledge, such as by identifying which variables are most applicable to our problem statement based on their definitions, and also through the use of variable selection methods such as Principal Component Analysis (PCA). The latter will especially come in handy as we anticipate many of these predictors to be correlated (such as "Points Per Game" and "Field Goals Per Game").

Once we have our prepared data, the first part of our methodology is to perform a cluster analysis using player and game data variables (TBD) from several historical seasons. Data comprising one to ten (TBD) of the most recent seasons will be excluded. We will need to scale our selected variables to prevent any of them from having disproportionate influence over the others. To determine the optimal number of clusters, we will create an elbow plot by running *k*-means and plotting the sum of the squared distances to the nearest cluster center for several possible values of *k*. Then, using our selected *k*, we will run the *k*-means clustering algorithm against our scaled variables, append the resulting cluster labels to our data, and visualize the clusters in a scatter plot (if the number of selected variables/principal components allows).

Once we've determined the optimal number for k, we will seek to find a way to label these clusters based on the caliber of players within that cluster. For example, we may notice that the players in one of the clusters contains a higher volume of players that made the NBA all-star team, or that one of the clusters contains players that have a high volume of Rebounds Per Game. Based on the characteristics of these clusters, we will seek to give them a label or name based on our domain knowledge of basketball.

The second part of our methodology is to train a classifier on our data that is now labeled with the player-caliber cluster label determined during the unsupervised learning step. Some of the classifiers we plan to consider are Naïve Bayes, Support Vector Machine, Neural Network, and Gradient Boosting. We will first split the data into a training and testing set and train each classification model on the training set using k-fold cross validation. We will then evaluate each model's performance on the testing set. Additionally, it may be interesting to use this trained classifier to predict the caliber of players on a portion of young and current players from the most recent NBA seasons that we have intentionally excluded from the clustering analysis.

Evaluation and Final Results

Once we've performed k-means clustering on the data, we will evaluate how the number of clusters aligns with sensible player caliber divisions. This will be based not only on our domain knowledge of the NBA but also by seeing how NBA

accolades such as player all-star selections, MVP selections, and All-NBA team selections correspond with the different clusters. Regarding the trained classifiers, we also hope to evaluate how well each of the supervised learning classifiers performs in comparison to one another to determine which one produces the best classification rate.