Capstone Project 1 Milestone Report

Introduction, Problem and Client

The NBA is widely regarded as the best basketball league in the world, boasting the highest level of play from the best players in the world. In the last decade or so, there has been an influx of 'european big men' (centers originating from EU countries) into the NBA. Great passing ability and sound fundaments of the game are aspects commonly attributed to EU centers. This has made drafting players harder as more and more are coming from oversees.

The NBA Draft is an annual event during which teams from around the NBA draft young players into their teams. Simply put those players can either be from a college or high school in the US, or international players. Not all players in the NBA were drafted, however the large majority enter the league through the draft. Naturally then, the teams that participate in the draft want to ensure that their player selection is optimal for their team. NBA scouts facilitate in this process and use various methods to evaluate talent. To make the jobs of NBA coaches and scouts easier, we will create models that will predict seasonal averages for the most common NBA statistics (Points per game, Assists per game etc). We will also compare US and EU centers and find the benefits/downsides of each category (we chose to compare only these 2 groups, as no other geographical group is large enough to make reliable statistical inferences possible). Equipped with this information, NBA teams (including its staff) will make better decisions regarding which players to draft, and what to expect from them as their progress through their careers.

Initial Data Acquisition and Wrangling

We had to scrape data from Basketball Reference (*www.basketball-reference.com*) as the website didn't have an API. To make the comparison as unbiased as possible, we scraped NBA center data from the 1980 season onwards (3pt shot introduced in 1980). Brief NBA player overview data as well as the URL of each player could be found on the website corresponding to the first letter of the surname (so the URL containing data about LeBron James could be found at the URL *www.basketball-reference.com/players/j).* We cycled over each letter (excluding X, as there was no X URL) and scraped the URLs of all NBA centers. Since the websites also included the start of the player's career and their position, we were able to filter out pre 1980 seasons and select only the players at the center (C) position.

Now that we have the URLs of each player, we can access the information on their webpage. The data we are looking to extract is the first table on the webpage, where each row is a season of that player's career, and each column is a season average statistic such as Points per Game. This is exactly the information that we scrape and store in a dataframe. To that dataframe we also append extra information found on the page that might be useful later such as the Weight, Height, 'Year in League' and Draft Rank (if a player is un-drafted, we assign a -1 label as a signpost for later wrangling steps) as columns. Finally, we determine the origin of the player (US or EU) and add it as a column also. We do this by scraping the birth place of the player and running it by a list of EU countries and US states to determine which group to place the player into. If the player is not from an EU country, or a US state, we label it him as 'Neither'. We do this for all players and append all of the dataframes to one 'master' dataframe where each row is a season played by an NBA center from the US or EU, and each column is a feature.

We then export the dataframe to a csv file so that we don’t have to re-scrape every time we run the program.

Wrangling the ‘Player Data’ Dataframe

We read in the dataframe that we exported to a .csv file in the previous section. Then we replace the -1 values as NaN and correct the datatypes of the columns in the dataframe. There were also multiple rows with the ‘Neither’ label in the ‘US or EU’ column. Some of these were really EU entries, but misclassified due to differences in spelling. We considered players originating in Canada as part of the US category as fundamentally we want to infer differences in the upbringing of the players, and not the geographical regions (also there was a substantial amount of data from Canadian born players).

Having fixed the misclassified rows, we removed players who were not from the US or EU. This was justifiable, as the number of such rows was very small in comparison to the whole dataframe. There were instances where almost entire rows of data were null values. These corresponded to cases where players did not play during that season (injury or playing overseas are a few examples for the reasons why). Again, the number of null rows was not substantially large, and by the same logic as applied above, we removed these rows.

Finally, we dropped a few low information columns (3 Point %, Free Throw %), reset the index of the dataframe, and converted the ‘%’ columns from ratios to percentages. It’s important to note that we did not deal with undrafed player data yet (rows where ‘Draft Rank’ column values are NaN) as this group, unlike the ones before, makes up a sizeable amount of data (around 10%) and represents a legitimate group of players. However, since we don’t know if ‘Draft Rank’ will be a feature in our predictive models, we can’t deal with these values yet. Furthermore, it doesn’t make sense to infer draft ranks for un-drafted players. As before, we exported the dataframe, ready for EDA.

Other Potential Data Sets

Other potential data sets include any other information that could have been scraped from basketball-reference, however in most cases this would represent the same information but transformed slightly (for example instead of reading in a table of ‘per game’ data, we read in ‘per 36 minutes’ data). The other option would have been to include advanced statistics, however most of these can be derived from the information we already have in our dataframe and as such would not provide much use in terms of predicting power.

Initial Findings