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 an increasing number 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 they 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). Some 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.

Once we attained the URLs of each player, we could access the information on their webpage. The data we extracted was in the first table on the webpage, where each row represented a season of that player's career, and each column a season average statistic such as Points per Game. We stored the scraped data in a dataframe. To that dataframe we also appended extra information that might be useful later such as the Weight, Height, 'Year in League' and Draft Rank (if a player was un-drafted, we assigned a -1 label as a signpost for later wrangling steps) as columns. Finally, we determined the origin of the player (US or EU) and added it as a column also. We did 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 was not from an EU country, or a US state, we labelled him as 'Neither'. We did 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 exported the dataframe to a .csv file.

Wrangling the 'Player Data' Dataframe

We read in the dataframe that we exported to a .csv file in the previous section, replaced the -1 values as NaN and corrected the data types 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 un-drafed 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.

EDA and Inferential Statistics Results

These are described in detail in 'NBA Player Analytics Inferential Statistics' document, so I will provide only a summary of the findings.

We found out that since 1980, 3804 seasons were played by players originating in the US and 592 by European players. The ratio of EU players has increased since 1980, but most notably in the last decade. When comparing the distributions US and EU players in terms of Points, Assists, Rebounds, Steals, Blocks, Turnover and Fouls we found there to be very little difference in the distribution of data, the quartile values, the median and so on. Thus, we concluded, perhaps surprisingly, that US and EU players (in totality of all the seasons played in the NBA) are the same (from a statistical standpoint). At that point we had effectively, and uninterestingly, answered a key question of the project *'What is difference between US and EU centers playing in the NBA, and what are their strengths and weaknesses'.*

The next part of the EDA focused on feature analysis/selection for the Linear Regression models that were to follow in the machine learning part of the project. We found out that: Age, Draft Rank, Effective Field Goal Percentage and Minutes Played have some relationship with the Points per Game dependant variable (however all these columns showed a non-linear relationship). For other features we used a heatmap. However, after analysing those first 4 features ‘manually’ it was clear that a heatmap is not a sustainable (time-wise) way to choose features for the model since correlation coefficients cannot account for non-linear relationships. The conclusion then, in the interest of time mostly, was to proceed to the machine learning stage and return to EDA only if we found the accuracy of the model to be insufficient.

Hypotheses were formed and ready to be tested, however we found out that the data was not normally distributed (thus breaking the assumption of the Hypothesis Test).

Building the Predictive Models/Machine Learning

The findings concerning the Machine Learning part of the project are described in detail in the Python Notebook, however, like with EDA, I will provide a summary of the findings and conclusions here.

Firstly, we split the data into 2 groups: drafted and undrafted players so we could create models for both separately. Then we performed some rudimentary data wrangling to get rid of some outliers. After building a few models, it became evident that a lot of code was being rewritten so we decided to define our custom functions to reduce the number of lines of code. Some of these functions served the purpose of plotting residual plots, others were used to easily predict values from unseen data in a simple way.

We went through many iterations of building a model, consulting the residual plot, and making changes with the aim to reduce the mean squared error and achieve homoscedasticity. The Points per Game and Assists per Game models performed much better than their counterparts (the reasons for that I covered in the Jupyter Notebook) both for drafted and undrafted player data, however heteroscedasticity was still present even after transforming the features and attempting weighted least squares regression.

Conclusion

We found that, in the end, there is no statistical difference (as a whole) between NBA centers from the US or EU in the major statistical categories and that european players are making up an ever-increasing ratio of the NBA player base. The predictive models are able to fulfil the purpose of the project, which was to allow the user (fan or scout of an NBA team for example) to get predictions for the main performance statistics, however only the models for Points and Rebounds are sufficiently accurate to be used in a real-world scenario, however this depends on the required accuracy of the user. Perhaps by introducing more data and tinkering with the models heteroscendasticity could be resolved and the accuracy of the other models improved.