The battles of neighborhoods

Taewoo Lee

June 24, 2020

1. Introduction

Introduction where you discuss the business problem and who would be interested in this project. Data where you describe the data that will be used to solve the problem and the source of the data. Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why. Results section where you discuss the results. Discussion section where you discuss any observations you noted and any recommendations you can make based on the results. Therefore, it is advantageous for teams to accurately predict whether and how much a player will improve in the next season. This project aims to predict whether and how much a player will improve the next season based on these data.

2. Data Cleaning

Most player stats, position, age, and draft position data can be found in two Kaggle datasets here
and here. These two datasets, however, lack data for certain years. For example, the player stats
dataset ends in 2017, and the player draft dataset starts in 1978 and ends in 2015. To complement
these two datasets, I scraped basketball-reference.com for player season stats of 2018 and player
draft positions of 1965-1977 and 2016-2017 (players drafted in 2018 has yet to play in NBA).

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a
lot of missing values from earlier seasons, because of lack of record keeping. I decided to only
use data from 1980 season and after, because of later seasons have fewer missing values and
basketball was a lot different in the early years from today's game.

There are several problems with the datasets. First, players were identified by their names. However, there were different players with the same names, which cause

their data to
mix with each other's. Though it was possible to separate some of them based on the years,
teams, and positions they played, I decided that it was not worth the large effort to do so, because
such players only accounted for ~1% of the data. Therefore, players with duplicate names were
removed.
Second, multiple entries existed for players who changed teams mid-season. This cause
their seasonal data to represent multiple samples with incomplete data. I wrote script to extract
total season stats for these players, and discarded partial season rows.
Third, there were two short seasons in recent NBA history, during which less than the
normal 82 games were played. This has caused stats in those seasons to be artificially smaller
than other seasons. To correct that, I normalized cumulative features such as points, rebounds,
etc. as if 82 games were played.
After fixing these problems, I checked for outliers in the data. I found there were some
extreme outliers, mostly caused by some types of small sample size problem. For example, some
players had only played a few games or a few minutes the entire season, and had performed
extremely well or poor in those minutes. Therefore, seasons during which less than 20 games or
100 minutes were played were dropped from the dataset. Similarly, there were players who only
took one 3-point shot, but made it, therefore had 100% shot accuracy. I changed the shot
accuracies for players who shot less than 10 shots to missing values.
There were 4 features which had missing values. Games started were imputed from minutes played because starters usually play more minutes. Missing 3-point accuracies were
imputed with a very small value (0.05) because if a player rarely shoots 3s, it is probably because
he is not very good at it. Missing free throw accuracies were imputed using the mean of all
players. Missing draft positions, meaning undrafted, were imputed using position 61

(the
position after the last position in the draft, 60th).

2.3 Feature selection

After data cleaning, there were 13,378 samples and 49 features in the data. Upon examining the
meaning of each feature, it was clear that there was some redundancy in the features. For
example, there was a feature of the number of rebounds a player collected, and another feature of
the rate of rebounds he collected. These two features contained very similar information (a
player's ability to rebound), with the difference being that the former feature increased with
playing time, while the latter feature did not. Such total vs. rate relationship also existed between
other features. These features are problematic for two reasons: (1) A player's certain abilities
were duplicated in two features. (2) A player's playing time were duplicated in multiple features.