NBA Player 5 year

Andrew Cramer, Marc Matvienko

# Problem Statement

What is the data science problem you are trying to solve? Why does the problem matter? What could the results of your predictive model be used for? Why would we want to be able to predict the thing you’re trying to predict?

Given statistics of a NBA player in a season, we predict whether or not the player remains in the league 5 years later. This could be valuable for a number of reasons including player acquisition, players on one’s team that are projected to fail soon can be sold at a high gain. Also…

### Data

Describe your dataset. This may also include insights from data exploration.

This dataset has about 500 entries each of which correspond to a single NBA Player from the 2014-15 Season, and 500 from the 2018-19 Season. Features included in the dataset are:

Player Name (Player)

Season (Season)

Season Type (Season Type)

Team (Team)

Games Played (Games)

Minutes played per game (Min)

Points scored per game (Pts)

Offensive Rebounds per game (OReb)

Defensive Rebounds per game (Dreb)

Rebounds per game (Reb)

Assists per game (Ast)

Steels per game (Stl)

Blocks per game (Blk)

Turn Overs per game (TO)

Personal Fouls per game (PF)

Field Goals Made per game (FGM)

Field Goals Attempted per game (FGA)

Field Goal % (FG%)

Three-Pointers Made per game (3PTM)

Three-Pointers Attempted per game (3PTA)

Three-Pointers % (3PT%)

Free Throws Made per game (FTM)

Free Throws Attempted per game (FTA)

Free Throws % (FT%)

We added the label *Survived* to each player in the 2014-15 dataset using the *Player* feature to identify if a player from the 2014-15 season was still playing in the 2018-19 season.

We also added a *Years in League* (YIL) feature to this by webscraping the data from Wikipedia. We were able to complete this using the Wikipedia API for about 300 of the entries, BeautifulSoup for another 150 and manually for the remaining entries. The scripts we made for that can be found in YIL\_Script1.py and YIL\_Script2.py respectively.

In our Feature Engineering Phase, we added 5 new features:

Field Goals per Total (FG/T)

Free Throws Per Total (FT/T)

Three-Pointers Per Total (3P/T)

Turn Overs and Personal Fouls per Minute (TOP/M)

Specialization (SPE)

The first three features represent the total number of points from each of the methods of point scoring with respect to the total number of points that player has scored (i.e. if I score one Three-Pointer (3pts), one Field Goal (2pts), and one Free throw(1pts), my 3P/T is 0.5, my FG/T is 0.333, and my FT/T is 0.167). The TOP/M feature represents a total number of Turn Overs and Personal Fouls a player commits per minute. Specialization represents how well balanced a player is across Offensive, Defensive and shooting attributes. Difference between Offensive and Shooting attributes lie in stats that don’t necessarily constitute good shooting from a player. For example, assists don’t increase point totals, and dunks can be Field Goals and yet have nothing to do with shooting. The feature is a measure of how well balanced a player is, players who are good at all aspects of the game will have a high score and those that specialize will have a low score, while those who are well balanced by being terrible in each will have an even lower score.

### Method

Describe your data science approach, any assumptions made, nuances, research done, feature engineering done, innovations in your procedure used, etc. Walk us through the process you used.

First we made sure there were no Null or NaN, which there were none of. However, if there were, some sort of imputation would have been appropriate, perhaps using the mean of that feature. Next we checked the head of the dataset to look for any features we thought were not appropriate to the task at hand. We ended up removing the season type, team, season year, and player name. These features had no inherent value that one could extract information from. The only thing of value is the team name, and that would be used to compare players to their teammates. Lastly we visualized the box plots of the data to see if perhaps some of the values were erroneous – any visible outliers could be attributed to the fact that the NBA has some extraordinary players.

# Challenges

Did you run in to any challenges? What worked well vs. what was more difficult than anticipated? Did you try anything without success?

aa

# Results

What were your results?

Aa

Overall accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SVM | KNN | RF | NB |
| Full Dataset | 63.3 | 55.6 | 53.4 | 70.9 |
| Correlated Dataset | 68.3 | 58.7 | 59.5 | 73.7 |
| Reduced Dataset | 67.1 | 60.1 | 60.5 | 67.1 |

Best accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SVM | KNN | RF | NB |
| Full Dataset | 65.7 | 55.2 | 59.7 | 70.9 |
| Correlated Dataset | 68.3 | 58.3 | 64.7 | 73.7 |
| Reduced Dataset | 64.9 | 60.5 | 63.7 | 67.1 |

Best F1 (macro)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SVM | KNN | RF | NB |
| Full Dataset | 0.69 | 0.60 | 0.60 | 0.71 |
| Correlated Dataset | 0.72 | 0.63 | 0.64 | 0.74 |
| Reduced Dataset | 0.70 | 0.64 | 0.65 | 0.67 |

# Next Steps

What could be done with this next? How could the results be used in the real world? What additional data analysis could be done on the data? Could additional data be added/supplemented to predict more things? Etc.

Adding more features such as *Age* and *Draft Round* that logically correlate to player longevity is an obvious step.

One problem with the current data is that it doesn’t have insight for nagging injuries a player might suffer from that temporarily drives his stats down. One solution to this might be to take the average stats over available years in the last three years, therein giving a more stable perspective on the player while not discounting players drafted in the three-year timespan. Another option would be to add an *Injured* flag or categorical feature.

Another course of action would be to get multiple years

Maybe also include the deviation from team statistics. How well does a player perform compared to all of his teammates that year. It is easier to have a lot of points if one is playing on a good team.