Capstone Project

Making Sense of the NBA Draft

Background

- Each year, NBA teams will draft players to their teams from College or International teams

- Players are scouted prior to the draft and their statistics, strengths and weaknesses are evaluated. However, it seems that a lot of this relies on scouts simply using the “Eye Test” to gauge potential

- Selecting the right player in the draft can translate to hundreds of millions of dollars of revenue for a team in the best cases, and similarly if a team makes the wrong decision they can lose out on massive amounts of revenue from ticket sales, merchandise and sponsorship

- The only problem with this process is that there is no concrete way for teams to be sure of the future development of their selections, as most of them are aged 18-20

Problem Statement

Which College statistics are most predictive of a player’s success in the NBA?

Assumptions/Limitations

- Data taken from the last 10 NBA drafts, going back any further would create problems with relevancy as the nature of the game changes

- International players were excluded, as there were no statistics about their international experience. Even If it was available, it wouldn’t be comparable due to the variance in quality of competition

- Rows with Null values were dropped, as each player is very different I felt that assigning median or regressed values to them would negatively affect the models (i.e. Combine stats)

Approach

- Gather data from basketballreference.com and sports-reference.com via exporting CSV files
- EDA/cleaning
- Create binary variables for whether a player exceeded the average in **Player Efficiency Rating**, **Win Shares per 48 minutes** and **WS/48 by the round they were drafted**
- Eliminate collinearity
- Use various classification models to predict the above variables to high accuracy
- Assess which statistics were most prominent in each model

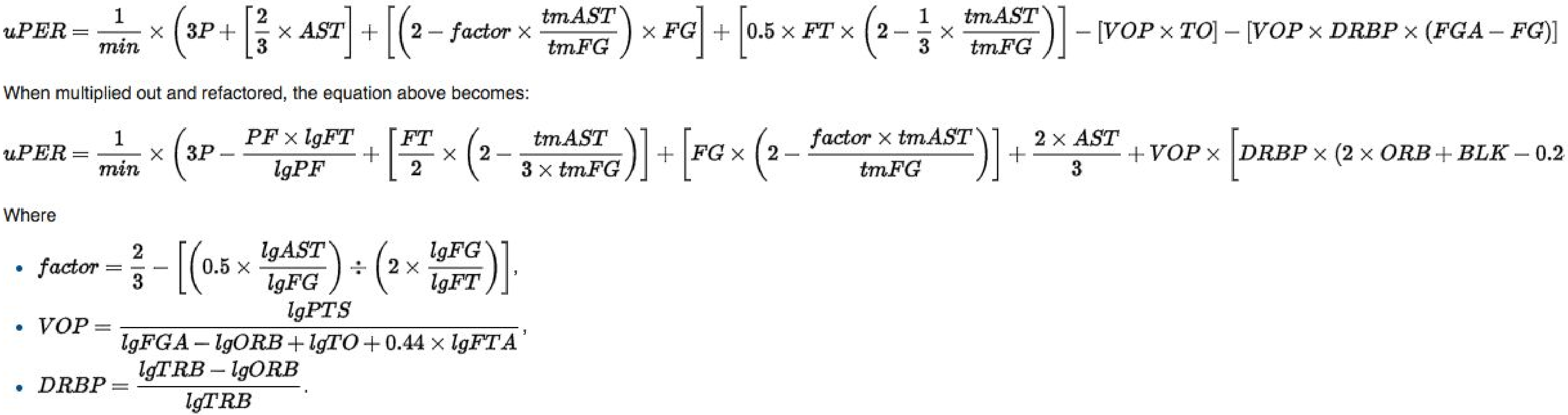
Dataset Information

- Information about every player drafted since 2008. This is 600 rows - I chose not to use drafts prior to that due to the shift in the nature of the sport

- Each row contains demographic information on the player, their College statistics (where applicable), their NBA statistics (where applicable) and their draft combine measurements (where applicable) - 61 columns in total

- After cleaning, the model dataset contained 223 rows

Player Efficiency Rating (PER)



$$uPER = \frac{1}{min} \times \Bigl(3P + \left[\frac{2}{3} \times AST\right] + \left[\left(2 - factor \times \frac{tmAST}{tmFG}\right) \times FG\right] + \left[0.5 \times FT \times \left(2 - \frac{1}{3} \times \frac{tmAST}{tmFG}\right)\right] - [VOP \times TO] - [VOP \times DRBP \times (FGA - FG)]$$

When multiplied out and refactored, the equation above becomes:

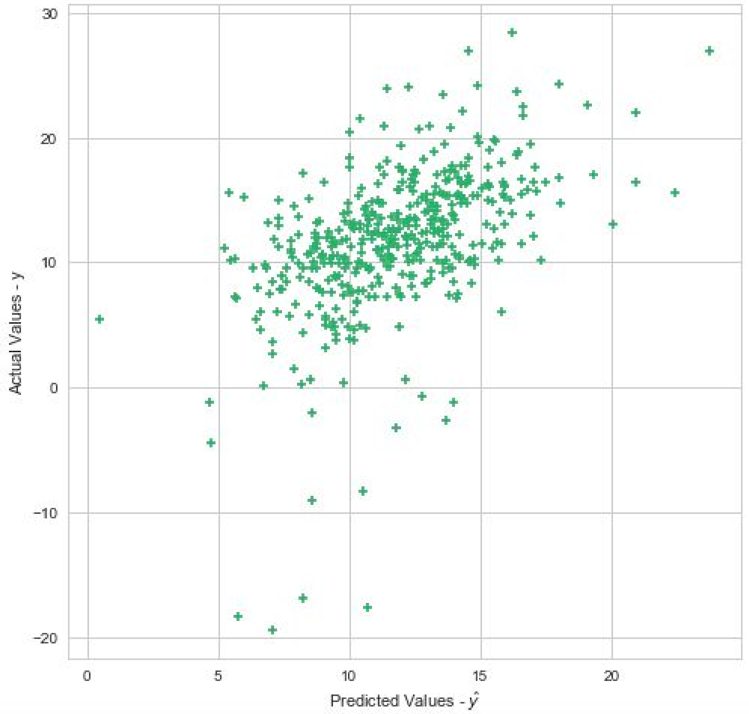
$$uPER = \frac{1}{min} \times \Bigl(3P - \frac{PF \times lgFT}{lgPF} + \left[\frac{FT}{2} \times \left(2 - \frac{tmAST}{3 \times tmFG}\right)\right] + \left[FG \times \left(2 - \frac{factor \times tmAST}{tmFG}\right)\right] + \frac{2 \times AST}{3} + VOP \times \Bigl[DRBP \times (2 \times ORB + BLK - 0.2$$

Where

- $factor = \frac{2}{3} - \left[\left(0.5 \times \frac{lgAST}{lgFG}\right) \div \left(2 \times \frac{lgFG}{lgFT}\right)\right],$
- $VOP = \frac{lgPTS}{lgFGA - lgORB + lgTO + 0.44 \times lgFTA},$
- $DRBP = \frac{lgTRB - lgORB}{lgTRB}.$

Credit: John Hollinger, 2002

Regression for Predicting NBA PER



```
score:  0.253725191317
```

- As shown, regression is not a great model for predicting NBA PER with given data

- Classification models should provide much better results

PER Model Performances - Baseline 65%

Model to predict whether a drafted player will have an above or below average NBA PER based on their college statistics.

Classification Model	Test Set Accuracy	Most Important Variables
Decision Tree (w/GS)	72%	**FG%**, PER/C, Hand Width
Random Forest (w/GS)	74%	**FG%**, BPG, Standing Reach
Logistic Regression (w/GS)	70%	PER/C, Standing Reach, **FG%**
SVM (w/GS)	**78%**	**FG%,** RPG, PER/C

Win Shares per 48 Minutes (WS/48)

- Metric to determine how many wins a player themself is directly responsible for
- Calculated by totalling a player's offensive and defensive impact and adding them together, then adjusting for 1 NBA game (48 minutes) so that it is standardised

WS/48 Model Performances - Baseline 58.29%

Model to predict whether a drafted player will have above or below average NBA WS/48 based on their college statistics.

Classification Model	Test Set Accuracy	Most Important Variables
Decision Tree (w/GS)	68.88%	FG%, **RPG**, Height
Random Forest (w/GS)	**72.98%**	**RPG**, Standing Reach, TOPG
Logistic Regression (w/GS)	71.11%	**RPG**, MPG, WS/C
SVM (w/GS)	69.46%	FG%, FTA, 3PA

WS/48 With Respect to Round Drafted

- As WS/48 performed slightly better against the baseline, I used this metric against the average WS/48 of players drafted in the 1st and 2nd Round to determine whether a player should be picked in that round.

- The rationale for this was that there are many ‘diamond in the rough’ players who slip through to the 2nd round, but go on to become superstars.

WS/48/Round Model Performances - BL 54.26%

Model to predict whether a drafted player will have above or below average NBA WS/48 for their drafted round based on their college statistics.

Classification Model	**Test Set Accuracy**	**Most Important Variables**
Decision Tree (w/GS)	63.45%	FG%, APG, TOPG
Random Forest (w/GS)	**64.28%**	APG, TOPG, 3PA
Logistic Regression (w/GS)	62.22%	FG%, WS/C, MPG
SVM (w/GS)	60%	FG%, 3PA, FTA

Findings

- Each NBA performance metric has its strengths and weaknesses, as outlined in the model performances
- Career performance is determined by so many factors (mindset, lifestyle, team culture, dedication) that an average of the model accuracy of roughly 10% over the baseline is very encouraging
- The statistic that featured most prominently across all 3 models was **FG%**, which indicates that a player's ability to score efficiently is the best indicator of success in the NBA.
- Other prominent statistics included Rebounding, Turnovers and Standing Reach

The Future

- Predicting the upcoming Draft
- Creating my own metric for performance (months)