DSC 423 - Winter 2021

Sports/Video Game Group: El Dorado

Ken Thomas Wei Tong Su (James Su) Zhong Xie
Chunnan Liu Devin Carroll

The final dataset was cleaned with the response variable of basketball FG% in mind. It was prepared from a combination of two sets of data containing professional basketball player statistics. The first dataset (International Basketball Stats) consisted of 53949 observations and 34 variables, the second dataset (Seasons Stats) contained 24691 observations and 53 variables. The two sets of data were joined using a combination of filtering and joining techniques used to match naming conventions across columns and to also ensure the most pertinent data was made available in the final dataset. The final dataset consists of 9171 observations with 59 variables – the variables being a mix of both continuous and categorical data. Between the two datasets there were a number of NA values. The challenge was in maintaining data fidelity while choosing which features or observations to drop.

Our first step in cleaning the data was to remove unrelated observations like non-NBA leagues in the dataset “InternationBasketballStats”.

```
>cleaned_InternationalBasketballStats <- filter(InternationalBasketballStats, Stage == "Regular_Season")
```

We chose Stage == “Regular_Season” as a condition because only NBA games have the stage as Regular_Season. Then we stored data into a new dataset called cleaned_InternationalBasketballStats by using Filter function from dplyr. We also can use this function to remove other redundant observers later.

Our next step was to create a new column true_draft which is the overall draft pick that player was selected in. The true_draft is calculated by (draft_round-1)*the number of teams + draft_pick. Then we split the dataset into 2003 and earlier when there were 29 teams, and 2004 and later when there were 30 teams.

```
>pre2004 <- filter(cleaned_InternationalBasketballStats, Year <= 2003)
>pre2004$true_draft <- (pre2004$draft_round-1)*29 + pre2004$draft_pick
>post2004 <- filter(cleaned_InternationalBasketballStats, Year > 2003)
>post2004$true_draft <- (post2004$draft_round-1)*30 + post2004$draft_pick
```

The data before 2004 stored into pre2004, and the others stored into post2004. After that we recombine them into cleaned_InternationBasketballStats by using bind_rows function from dplyr.

```
>cleaned_InternationalBasketballStats <- bind_rows(pre2004, post2004)
```

Formatting the “Year’ variable to the correct form is essential as the InternationalBasketballStats has a different format compared to our NBA only dataset. To match the formatting, we split the season by the delimiter “-“ into starting year and the ending year. Once we completed the split, the variable was then converted into integers.

```
cleaned_InternationalBasketballStats <-
separate(cleaned_InternationalBasketballStats,Season,into=c('Start','Year'),sep="-")
cleaned_InternationalBasketballStats$Year <-as.integer(cleaned_InternationalBasketballStats$Year)
```

The international dataset was reduced to contain only the essential features that were going to be used in the final dataset. This was done to prevent the need to drop unused

columns after the merging of data took place because many of the continuous numeric variables in the International Basketball dataset were already present in the NBA only dataset. The only information we needed from the International Basketball data were related to the draft and the player's physical measurements. The following line of code was used to select for features that would be used in the final dataset.

```
>cleaned_InternationalBasketballStats <- select(cleaned_InternationalBasketballStats, Year, Player, height_cm, weight_kg, draft_round, draft_pick, true_draft)
```

After selecting all the features we need from the international dataset, we merged it with our NBA season stats dataset with a left join operation. We specified the join operation based on the Player and Year variable because they are the primary key to the season stats and the foreign key to the international dataset.

```
full_NBA <- left_join(Seasons_Stats, cleaned_InternationalBasketballStats, by=c("Player","Year"), all.y=TRUE)
```

In this step of the cleaning process, we wanted to remove a few variables which are not useful such as: x3Par, blan1 and etc. To drop these columns we assigned them a null value. After taking these steps, the variables were pared down to what we need and anticipate on analyzing.

The final data cleaning step we took was transforming the Position variable into variables that were more usable. Players have often played multiple positions and our merged dataset started with 24 levels at the Position variable. We felt this number of levels would create too many dummy variables if used in the regression model, making it less parsimonious. Instead, we decided the best way to simplify this variable was to adapt it into 5 binary variables, one for each of the traditional positions, if a player was considered to have played that position, the value would be 1, otherwise 0. This method of identifying players' positions would reduce the number of variables in the model as well as individually capture the value-added of playing at each position. We created these 5 variables by initializing the value for all players at all 5 positions to 0. Then splitting the dataframe into 10 sub-dataframes, once again using filter(). Each sub-dataframe represented a different permutation of positions: Centers, Centers/Power Forwards, Forwards, Guards, Center/Small Forward, Point Guards, Shooting Guards, Small Forwards, Shooting Guard/Small Forwards, and Point Guard/Small Forwards. Once each player was in the appropriate group, we assigned them a value of 1 in the corresponding columns and then rebound the sub-dataframes using dpylr's 'bind_rows' function.

The final filtering step we took was to restrict the data set to only include players after the 2001 season. This decision was made because we felt it more accurately captured the current NBA landscape. A number of rule changes were instituted that season which drastically affected how defenses and offenses operated in the NBA and as such we feel this restriction creates a more precise dataset. This restriction has the positive knock-on effect of removing a large number of our observations that were missing significant amounts of values.

After taking these munging steps, we are left with 9171 observations where each observation captures a single NBA player's season long statistical output for that particular year. We dropped most of the features in the International Basketball data as they were already present in our main dataset. The features we did keep added draft information and physical

attributes of each player to our final dataset. We also dropped 4 features from the main data set, 1950-2017 NBA Season Stats, because they did not contain relevant information. We created a new variable to accurately present the overall draft selection each player was selected at. We also created 5 binary variables to replace the large variety of levels present in the Position feature. We also restricted our date to start in the 2002 season to create a more complete and consistent dataset.

Appendix - Variable Glossary:

Year/Player Name - Define each observation as the given player's statistical output in the season indicated by the year. The year is defined as the year the Finals took place in a given season. For example the 1999-2000, with the Finals occuring in 2000 is 2000.
Age - Is the player's age in years as of February 1st of the given year
G - Games, the number of games a player appeared in as either a starter or off the bench.
GS - Games Started, the number of games a player appeared as a starter.
MP - Minutes Played, the number of minutes played by the player across all games.
PER - Player Efficiency Rating, an advanced stat developed by John Hollinger. Meant to present a player's positive contributions subtracted by their negative contributions on a per minute basis.
VORP- Value Over Replacement, an estimate of the relative contributions of a player compared to a "replacement" player defined as (-2.0), over 100 possessions on an average team over 82 games.
ORB. - Offensive Rebound Percentage, estimate of Offensive Rebounds grabbed as percent of Offensive Rebounds opportunities while playing.
ORB - Offensive Rebound Total, total number of offensive rebounds a player accumulated DRB.
DRB. - Defensive Rebound Percentage, an estimate of Defensive Rebounds collected as a percent of the Defensive Rebound opportunities while playing.
DRB - Defensive Rebound Total, total number of defensive rebounds a player accrued
TRB. - Total Rebound Percentage, an estimate of total rebounds collected as a percent of the total rebounding opportunities while playing. Combines offensive and defensive rebounding.
TRB - Total Rebound Total, total number of rebounds by a player. Combines offensive and defensive rebounding.
AST. - Assist Percentage, an estimate of the percentage of teammate's field goals assisted by the player while playing
AST - Assist Total, total number of assists by a player in a given season
STL. - Steal Percentage, an estimate of the percentage of opposing possessions that end with that player making a steal.
STL - Steals Total, total number of steals by a player in a given season.
BLK. - Block Percentage, an estimate of the number of opponent's 2-point field goal attempts that a player blocks while on the floor.
BLK - Blocks Total, total number of blocks by a player in a given season.
TOV. - Turnover Percentage, estimate of the number of turnovers a player commits over 100 possessions
TOV - Turnover Total, total number of turnovers a player commits in a season.
USG. - Usage Percentage, estimate of a percentage of a team's possessions that end with that player taking a shot.
OWS - Offensive Win Share, estimate of the number of win's a player contributes on the offensive end of the court over the season
DWS - Defensive Win Share, estimate of the number of win's a player contributes on the defensive end of the court over the season.

WS - Win Shares, estimate of the number of win's a player contributes on both ends of the court over the season
WS/48 - Win Shares per 48 Minute, number of win shares a player contributes scaled to 48 minutes of play
OBPM - Offensive Box Plus Minus,
DBPM - Defensive Box Plus Minus,
BPM - Box Plus Minus, an estimate of the point difference per 100 possessions that a player contributed above a league-average player, translated to an average team
FG - Field Goals, number of total field goals a player successfully shot over the season
FGA - Field Goal Attempts, number of attempted field goals over the season
FG. - Field Goal Percentage, the percentage of shots taken a player hit
3P - 3 Point Field Goals, total number of 3-point shots made over the season
3PA -3 Point Field Goal Attempts, total number of 3-point shots attempted over the season
3P. - 3 Point Field Goal Percentage, the percent of attempted 3 Point shots a player successfully hit
3P - 2 Point Field Goals, total number of 2-point shots made over the season
3PA - 2 Point Field Goal Attempts, total number of 2-point shots attempted over the season
3P. - 2 Point Field Goal Percentage, the percent of attempted 2-Point shots a player successfully hit
FTr - Free Throw Rate, number of Free Throws attempted per game
FT - Free Throw Made, number of Free Throws a player hit in a given season
FTA - Free Throw Attempts, number of Free Throws a player attempted over a season
FT. - Free Throw Percentage, the percent of Free Throws a player successfully shot eFG - Effective Field Goal Percentage, Field Percentage adjusted for the value of the shots taken. Does not include Free Throws.
TS. - True Shooting Percentage, measure of overall shooting percentage weighted by the value of the shoot. Includes Free Throws.
PFouls - Personal Fouls, number of Personal Fouls committed by a player over a season
PTS - Total Points, total number of points a player scored over a season
Height_cm - Height in Centimeters, each player's height in centimeters
Weight_kg - Weight in kilograms, each player's weigh in kilograms
Draft_round - Draft Round, which round of the draft a player was drafted in
Draft_pick - Draft Pick, which selection within a round a player was selected in the draft
True_draft - True Draft Pick, the number of selections that occured before the player was selected in the draft when accounting for the draft round and draft pick.
PG - Binary variable representing if the player is considered a Point Guard, 1 if yes, 0 if no.
SG - Binary variable representing if the player is considered a Shooting Guard, 1 if yes, 0 if no.
SF - Binary variable representing if the player is considered a Small Forward, 1 if yes, 0 if no.
PF - Binary variable representing if the player is considered a Power Forward, 1 if yes, 0 if no.
C - Binary variable representing if the player is considered a Center, 1 if yes, 0 if no.