NBA Shot Prediction and Analysis Team Project - Final Version
Section C, Team C49
Team Members: Cara You, Juan Brando, Pritisha Punukollu, Julie Liu, Zhicheng Zhang
Date: 10/10/2019

Business Understanding

In the NBA, basketball teams are looking into increasing their chances of winning each game by attempting the most efficient shots. Attempting the shots with highest expected points could result in higher scoring average and therefore a higher winning percentage. In a multimillion dollar league, having every edge counts towards a goal of winning a championship.

Understanding a player's most efficient shot area could help a coaching staff identify the team's strengths and weaknesses, build a roster with the correct balance or even prepare for an upcoming match by scouting the opposing team. By using data from every shot attempted, we could determine the most effective ranges for different types of players. By understanding these conditions, it can be possible to try to assemble the most effective team by balancing players that are better from short, mid and long range. It could also be determined if a player is better as a spot-on shooter or as a dribbler or even how much does a player's effectiveness decreases when a defender is guarding him compared to when he is open.

In this project, we will focus on determining where each player is most efficient, helping to increase each teams expected point average by modelling the probability of making a shot. Then, by combining this result with clustering analysis, we are going to make recommendations for players and coaches to consider shots where they excel and become more efficient. Even by just changing from a couple of low efficiency shots into more efficient ones can be the difference between losing or winning a game.

Data Understanding and Cleaning

A couple of years ago, tracking cameras were installed in NBA arenas which allowed to record information of every player in the court for every second of play. This permitted to generate new information which was previously not available. Every record in the database is a shot that was attempted in every game during the 2014-2015 NBA season. Information was scraped from NBA's REST API (https://www.kaggle.com/dansbecker/nba-shot-logs). There are 21 variables and 128 thousand records in the dataset. A detailed Dictionary is available in the appendix. Some of the most important features include: the period of the shot, the time remaining in the quarter, the time remaining in the shot clock, times the player dribbled, how long did the player held the ball and how far was the shot from the rim.

Information regarding the player taking the shot and the defensive player was also included to complement the original database which was obtained from Basketball Reference (https://www.basketball-reference.com/). Player database contains information like position, age, height, weight, salary and nationality. Databases are being joined two times, one to include information for the closest defender and another to include the information from the shot taker.

Data Cleaning: Treating NAs and spurious value

For SHOT_CLOCK, we found that there were 5567 records with no shot clock, which occurs when a quarter is ending and there is no shot clock; we also found that there were 3857 shots with maximum value 24 sec, which means players make shots the same second when the clock starts. We viewed this 24 as offensive rebounds. For LOCATION, we transfer location A, which means away from home, to 0, and H, which means at home, to 1. For TOUCH_TIME, there are 312 shots with negative seconds and 3046 shots with 0 seconds. We decided to keep the shots with 0 or positive TOUCH_TIME and created a new subset called shot1. For GAME_CLOCK, we changed 'mm:ss' format to numerical second for logistic regression.

Outlier Analysis and Feature Engineering

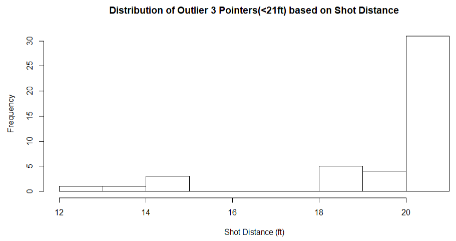
After taking care of the NULL values, we proceeded to conduct sanity checks and outlier treatment as follows:

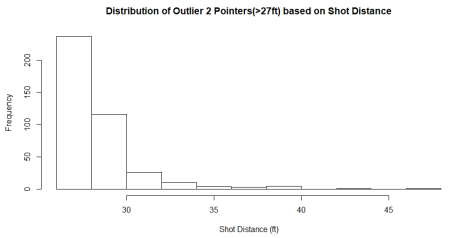
Cross Verification of Points per Shot and Successful Shots: There are four columns in question: SHOT_RESULT, FGM, PTS_TYPE, PTS. FGM is the binary version of SHOT_RESULT (categorical) and hence we drop the latter as the values tally. PTS=PTS_TYPE*FGM (all values tally) and thus poses as redundant but we retain it for ease of use for creation of future columns.

Closest Defender Distance:Most defenders (97%) defend within 10ft of the shooter. We believe that we could introduce one profile of the shots made based on how far away the closest defender was as opposed to looking for outliers: "*open_shots*". We applied a threshold of 6ft to CLOSE_DEF_DIST beyond which shots were classified as open shots i.e. *open_shots=1.*

Column based on Dribble: We've also added a column to categorize shots based on whether they were made directly or post dribbling that we think will be useful for later classification: "*spot_shots*" (i.e. spot_shots=0 if there was no preceding dribbling).

Outliers based on Point Value of Shots versus Shot Distance & Column for Corner 3 Shots: We then proceeded to sense check the value of the shot taken based on where it was taken from. The NBA has defined distances based on which shots can be classified as 2/3 pointers. Despite making slight allowance we found that some point values of the shots given their distances were simply absurd. These can be observed in the following charts.



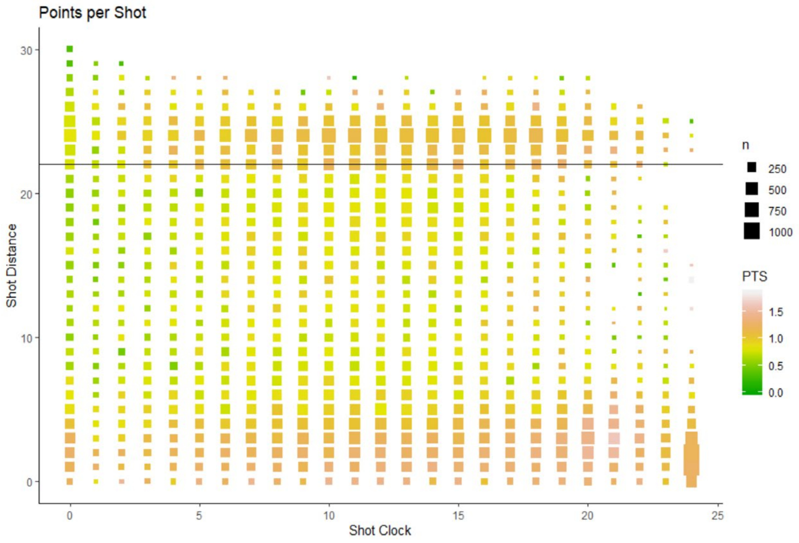


Based on both the charts we have chosen to drop the following rows:

- 2 pointers over with a shot distance > 27 feet (0.3% of data)
- 3 pointers over with a shot distance < 21 feet (0.03% of data)

The shot value and distance were then used to determine whether the shot was a corner 3 shot or not and stored in "*corner_shots*". Corner 3 are defined as 3 point shots with a distance to the rim of less than 24 feet.

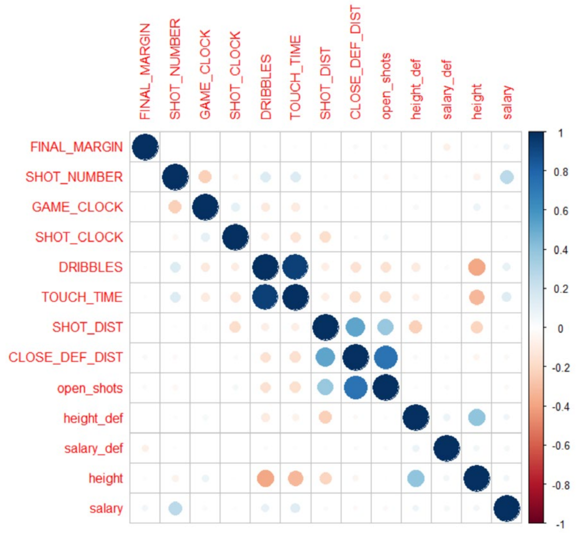
Data Exploration



The graph above shows several trends about the distribution of shots. It shows how teams have focused on being more efficient and are trying to maximize the expected points per shot. Regarding the shot distance, there are two main zones where there is a higher concentration of shots, around the three-point line (the black one) and near the rim. These are also the areas where there is a higher average points per shot. The volume of shots in the mid-range area is small, which are indeed the least efficient shots. In terms of the shot clock, as it winds down, there is a lower volume of shots and they are less efficient. The most efficient shots appear to be near the rim when the shot clock is just starting, however these could be a product of offensive rebounds where the shot clock restarts. As a matter of fact, there is a low volume of shots at the start of the shot clock as the play develops and then the volume increases. However, the efficiency starts to decrease when there are less than 5-7 seconds and probably the play intended didn't develop and the attacking team had to settle for less efficient options.

Next, we are going to plot a correlation matrix among features to further explore the data:

From the matrix, there are some obvious insights such as: the height of the player is positively correlated with the height of defenders;there are high positive correlation among close defence distance, open shot, and shot distance, as it is how open shot is defined;there are also some insightful discoveries:

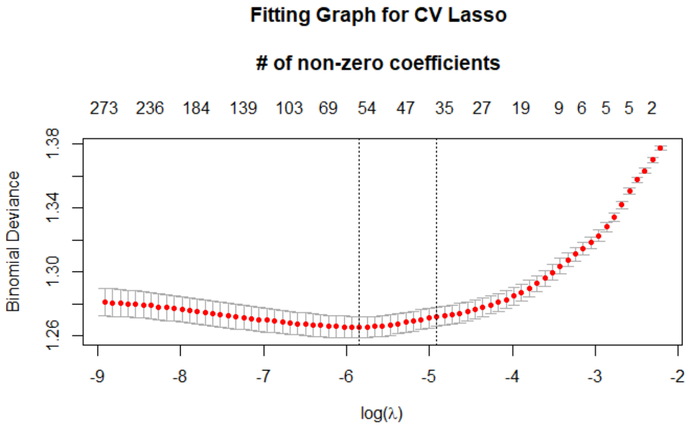


Game clock is negatively correlated with shot number, and shot distance is negatively correlated with shot clock, because the less time left, the more pressure to make a shot from a longer distance; close defender distance and open shot are negatively correlated with dribbles and touch time, because players with close defenders will choose to pass the ball; height of defender is negatively correlated with the shot distance and dribble, because centers who make shot in shorter distances are usually defended by tall defenders; height is negatively correlated with dribbles, touch time, and shot distance, because taller plays either dunk, or make the shot from a shorter distance; salary is positively correlated with shot number, shot time, and dribbles, because players with a higher salary will participate more in the game.

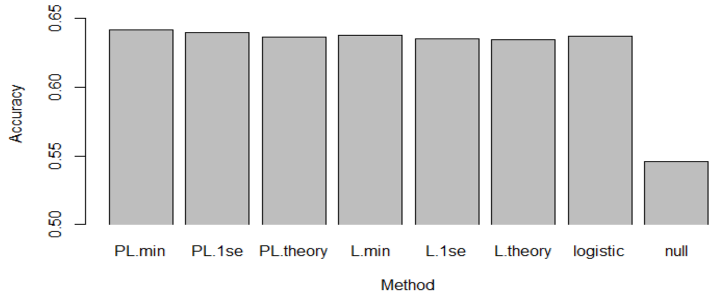
Modeling: Logistic Regression

As we mentioned before, we intend to evaluate each NBA player’s shooting performance and determine under which circumstances, each player is most efficient in making successful shots. In order to do this, we start by modelling the probability of making a successful shot, given shooting data and player characteristics. Specifically, FGM (1: successful shot, 0: unsuccessful shot) will be our dependent variable. Independent variables include SHOT_CLOCK, TOUCH_TIME, HEIGHT etc.. We are going to estimate different classification models for this task.

We first build a simple logistic regression model and compare it with the null model. Based on this, we use the Lasso method to select variables: Lasso with theory λ, Lasso with min λ, Lasso with 1se λ. Left is the Lasso path:



We then build the Post Lasso models with three different λ choices and apply the 10-fold cross validation for all the models using out of sample accuracy.



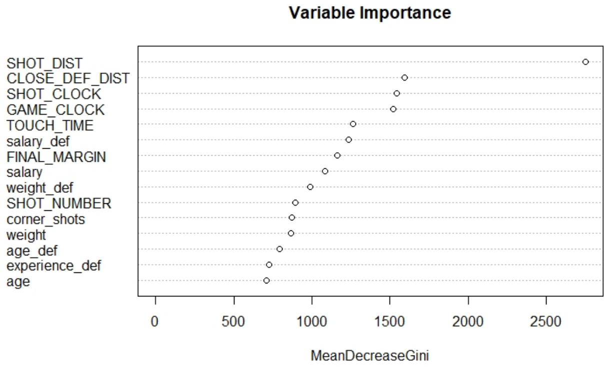
As is shown in the graph, PL.min model (49 variables selected) performs the best with an out of sample accuracy 0.64.

We also want to compare PL.min with other machine learning methods.

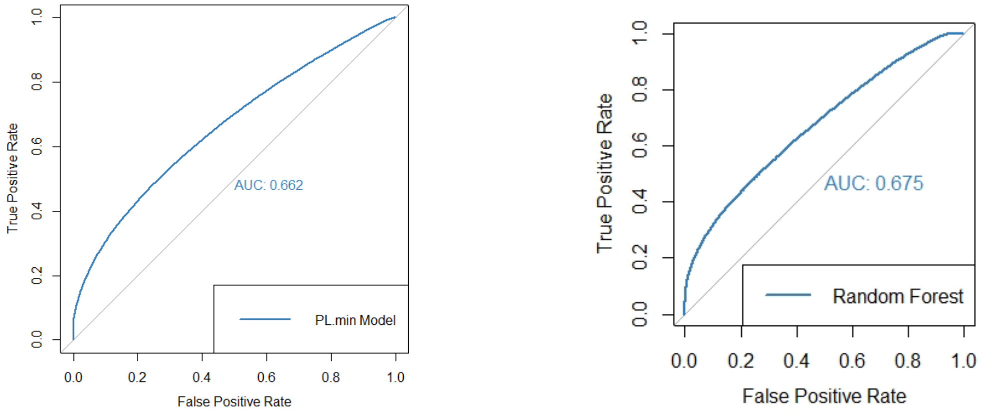
Modeling: Random Forest

A second machine learning algorithm that we are going to evaluate is Random Forest. The advantage of this model compared to others is that it avoids overfitting by computing an out of bag error while only trying a subset of the available variables as the error can be reduced by adding more trees to the model. Nevertheless the model is not as interpretable as other alternatives and it also has a considerable amount of variables to be tuned.

After splitting the data into training and test samples the first step is to determine the optimum number of trees needed to estimate the target variable. This is shown in the first graph. As the out of bag error stabilizes pretty fast it is not worth to have a very big number of trees. Instead we are settling with 200 trees and proceeding to tune the other parameters. In order to optimize the other parameters we proceed with a hyperparameter tuning in which we want to find the model which produces the least amount of error. We have to determine the number of variables to be selected for each tree, the size of the node and the sample of the tree to estimate. This process is detailed in the accompanying code. We finally settle for a model with 5 variables, a node size of 9 and a sample fraction of 0.55. The most important variables are detailed next:



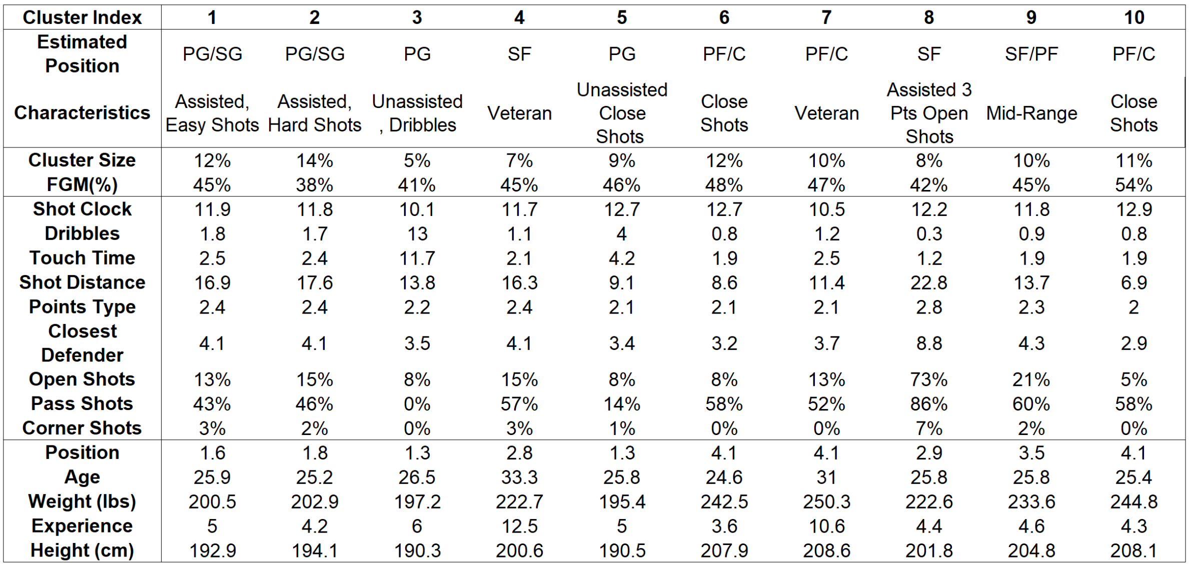
The most important variables are as expected, where the most important one to determine whether a shot was made or not is the distance to the basket. The next variables are how close is the closest defender, the time remaining in the possession and the game clock. It’s interesting that salary from the player taking the shot and the salary from the defender are very important. This could mean that teams are valuing correctly the better players and rewarding them with high salaries. Finally, we compare the performance of random forest model with that of the PL.min model, it turns out that random forest performs slightly better with an AUC of 0.675 so this is the model that is going to be predicted on the deployment section.



Clustering

In order to further understand different segments of the data and make suggestions based on that, we did unsupervised clustering. The PCA outcome was not very interpretable, therefore, we ran AIC BIC and HDIC for K means. The smallest cluster size recommended was 17. But we ran 10 clusters for easier future interpretation and suggestion making purpose.

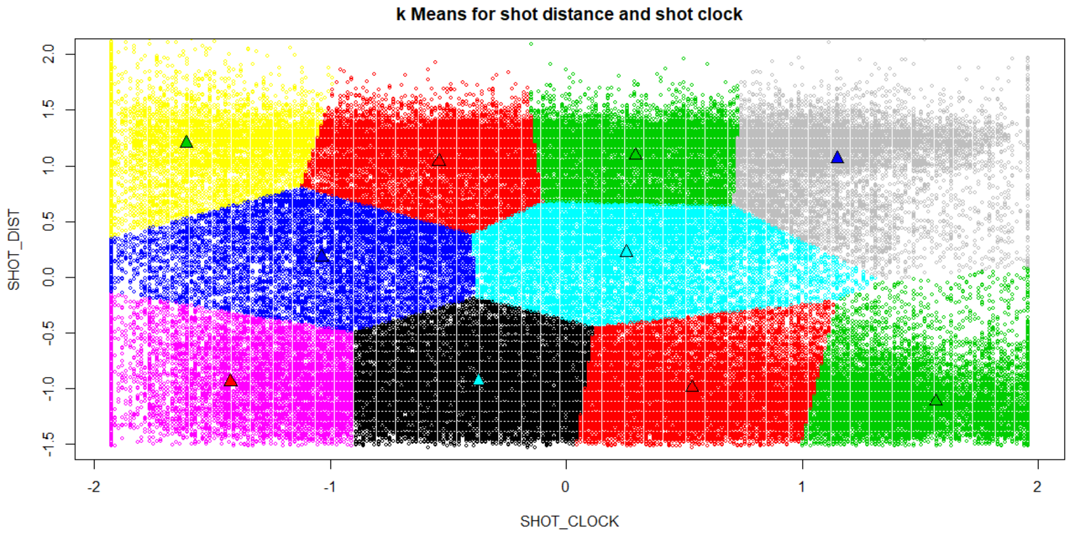
We chose 15 features that represented the player characteristics and shot performances of each cluster and distinguish one cluster from another the best. We then built a table with the cluster size, the field goal made (FGM) percent that we calculated, the features, and the estimated position and characteristics of each clusters based on the feature values.



Cluster Index	1	2	3	4	5	6	7	8	9	10
Estimated Position	PG/SG	PG/SG	PG	SF	PG	PF/C	PF/C	SF	SF/PF	PF/C
Characteristics	Assisted, Easy Shots	Assisted, Hard Shots	Unassisted , Dribbles	Veteran	Unassisted Close Shots	Close Shots	Veteran	Assisted 3 Pts Open Shots	Mid-Range	Close Shots
Cluster Size	12%	14%	5%	7%	9%	12%	10%	8%	10%	11%
FGM(%)	45%	38%	41%	45%	46%	48%	47%	42%	45%	54%
Shot Clock	11.9	11.8	10.1	11.7	12.7	12.7	10.5	12.2	11.8	12.9
Dribbles	1.8	1.7	13	1.1	4	0.8	1.2	0.3	0.9	0.8
Touch Time	2.5	2.4	11.7	2.1	4.2	1.9	2.5	1.2	1.9	1.9
Shot Distance	16.9	17.6	13.8	16.3	9.1	8.6	11.4	22.8	13.7	6.9
Points Type	2.4	2.4	2.2	2.4	2.1	2.1	2.1	2.8	2.3	2
Closest Defender	4.1	4.1	3.5	4.1	3.4	3.2	3.7	8.8	4.3	2.9
Open Shots	13%	15%	8%	15%	8%	8%	13%	73%	21%	5%
Pass Shots	43%	46%	0%	57%	14%	58%	52%	86%	60%	58%
Corner Shots	3%	2%	0%	3%	1%	0%	0%	7%	2%	0%
Position	1.6	1.8	1.3	2.8	1.3	4.1	4.1	2.9	3.5	4.1
Age	25.9	25.2	26.5	33.3	25.8	24.6	31	25.8	25.8	25.4
Weight (lbs)	200.5	202.9	197.2	222.7	195.4	242.5	250.3	222.6	233.6	244.8
Experience	5	4.2	6	12.5	5	3.6	10.6	4.4	4.6	4.3
Height (cm)	192.9	194.1	190.3	200.6	190.5	207.9	208.6	201.8	204.8	208.1

In the above image, cluster 1 2 and 8 are highly assisted based on the high percentage of pass shots while cluster 3 and 5 are in the opposite; cluster 1 and 2 are similar across all features except FGM, cluster 1 has a higher FGM and therefore makes easier shots; cluster 3 has the most dribbles; cluster 4 and 7 have the most experienced players; cluster 5, 6 and 10 have the lowest percentage of open shots, therefore, they make closer shots, while cluster 8 makes open shots; cluster 8 has the longest shot distance with a mean of 22.8 feet, the highest percentage of open shots and pass shots, and a shot type towards 3, which makes mostly 3 point shots; cluster 9 has mid range feature values.

Then we plotted the 10 clusters by shot distance and shot clock. Different clusters have different level of shot distance and shot clock.



cluster color	yellow	top red	top green	grey	dark blue	light blue	pink	black	bottom red	bottom green
shot distance	long	long	long	long	middle	middle	short	short	short	short
shot clock	short	middle	middle	long	short middle	middle long	short	middle	middle	long

Business Insights and Deployment

We calculated the average points expected (APE) for each shot by multiplying the probability of making the shot (FGM) by the point type (2 or 3), and aggregated them on a player level.

$$APE_j = \frac{\sum_1^i \hat{y_i} * \text{shot type}}{\text{i}}, \text{for shot i by player } j \text{ in cluster } k$$

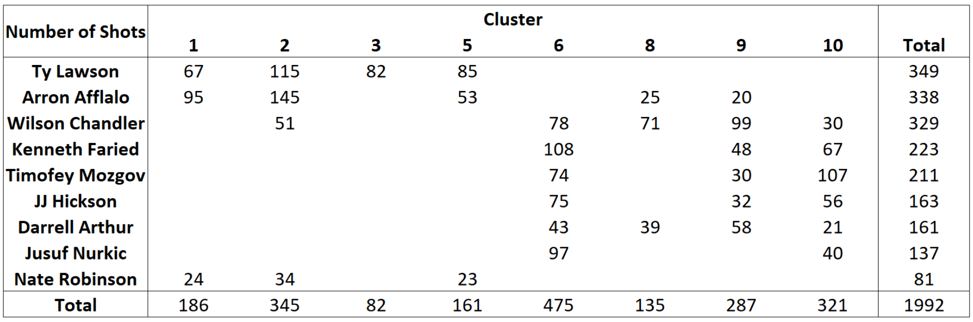
$$DAPE_j = y_j - APE_j, \text{for shot i by player } j \text{ in cluster } k$$

Then we calculated the deviance from the APE (DAPE) by deducing the true points made by the APE. We wanted to compare the DAPE for each cluster with the number of shots each player made in each cluster. Combining this with the insight for each cluster that we gained from clustering section, we gave a recommendation on which cluster each player should focus on attempting more shots in order to increase the personal APE, and eventually, increase the APE of the whole team. In this case, we chose the Denver Nuggets as our targeted team as an example and analyzed the players who made the most shot over the season.

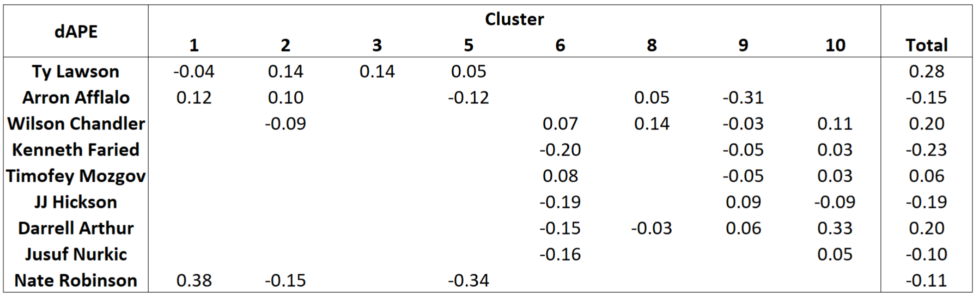
For Ty Lawson, instead of making more cluster 1 shots, which is assisted shots, he should make more cluster 3 shots and keep on making lots of cluster 2 shots, which are unassisted or dribbling shots. For Arron Afflafo, instead of making more cluster 5 or 9 shots, which are dribbling and unassisted close shots, he should make more cluster 1 shot and keep on making lots of cluster 2 shot, which are both assisted shots. Also, from the total point difference column, we see that Ty Lawson was out-performing his teammates, while Arron was under-performing compared to the grand total point difference of the team, which contradicted with his relatively high shot number compared to his teammates.

Also from the two tables, we can see that the Nuggets lack cluster 4 and 7 players. We suggest them to hire players who attempt a high amount of shots from cluster 4, for instance, veterans to balance the skillset and position within the team.

Looking ahead, we can also externalize the previous process to other team in season 2014 to 2015. Moreover, in the future, once we get access to the updated data, we can make suggestions for not only what cluster should the player focus on making the shot, but also the player hiring process for the following season for any team.



Number of Shots	Cluster								
	1	2	3	5	6	8	9	10	Total
Ty Lawson	67	115	82	85					349
Arron Afflalo	95	145		53		25	20		338
Wilson Chandler		51			78	71	99	30	329
Kenneth Faried					108		48	67	223
Timofey Mozgov					74		30	107	211
JJ Hickson					75		32	56	163
Darrell Arthur					43	39	58	21	161
Jusuf Nurkic					97			40	137
Nate Robinson	24	34		23					81
Total	186	345	82	161	475	135	287	321	1992



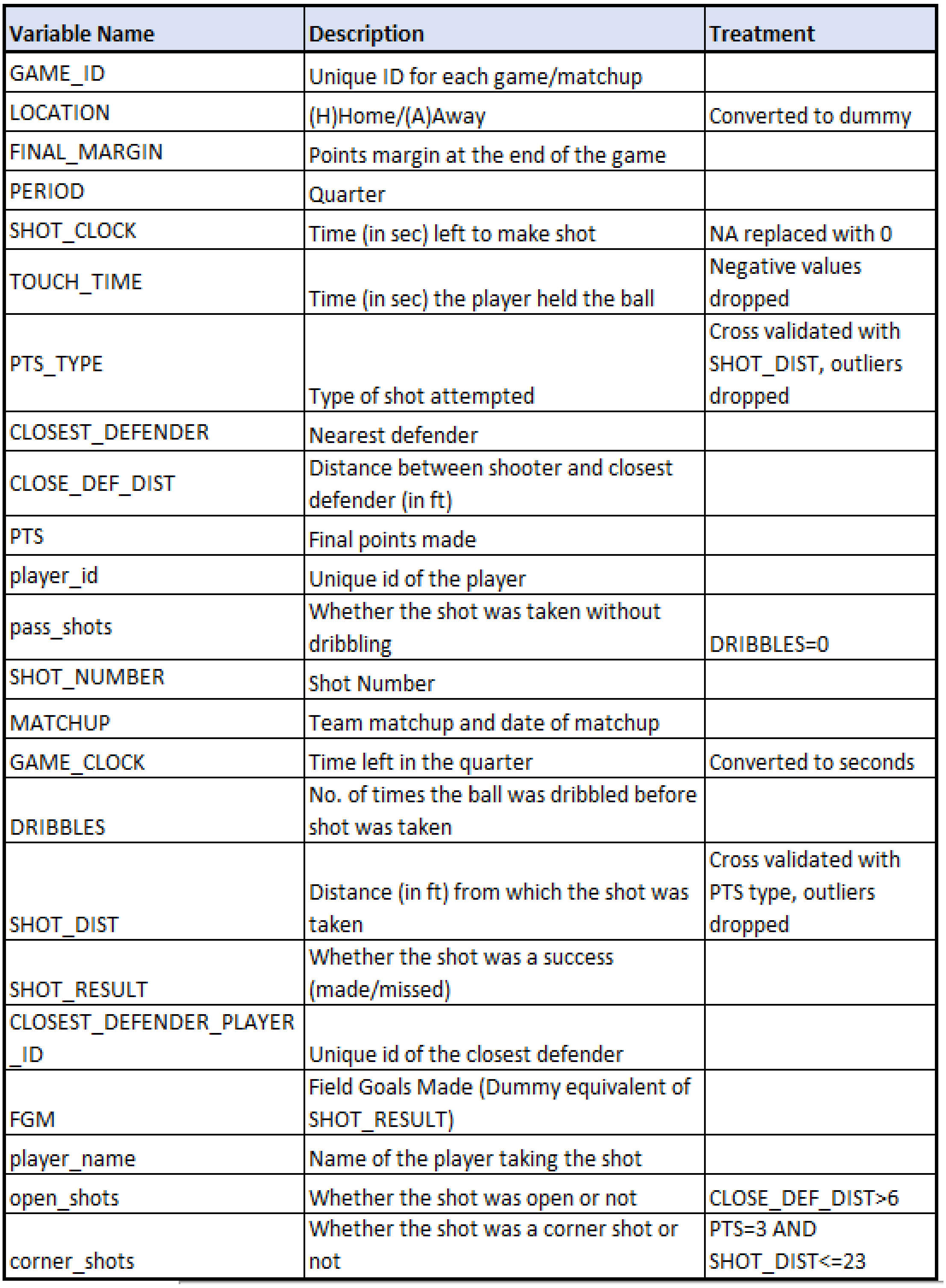
dAPE	Cluster								
	1	2	3	5	6	8	9	10	Total
Ty Lawson	-0.04	0.14	0.14	0.05					0.28
Arron Afflalo	0.12	0.10		-0.12		0.05	-0.31		-0.15
Wilson Chandler		-0.09			0.07	0.14	-0.03	0.11	0.20
Kenneth Faried					-0.20		-0.05	0.03	-0.23
Timofey Mozgov					0.08		-0.05	0.03	0.06
JJ Hickson					-0.19		0.09	-0.09	-0.19
Darrell Arthur					-0.15	-0.03	0.06	0.33	0.20
Jusuf Nurkic					-0.16			0.05	-0.10
Nate Robinson	0.38	-0.15		-0.34					-0.11

Appendix

Contribution table:

	Cara You	Juan Brando	Zhicheng Zhang	Julie Liu	Pritisha Punukollu
Business Understanding		Business Understanding		Business Understanding	
Data Understanding		Data Understanding	Data Understanding		Data Dictionary
data cleaning	Treating NAs		Spurious value treatment		Outlier Analysis and Feature Engineering
Data Exploration				Data Exploration	Problem Breakdown
Modeling		random forest	logistic modeling		
Clustering	Clustering				
Business Insights and Deployment	Business Insights			Business Deployment	

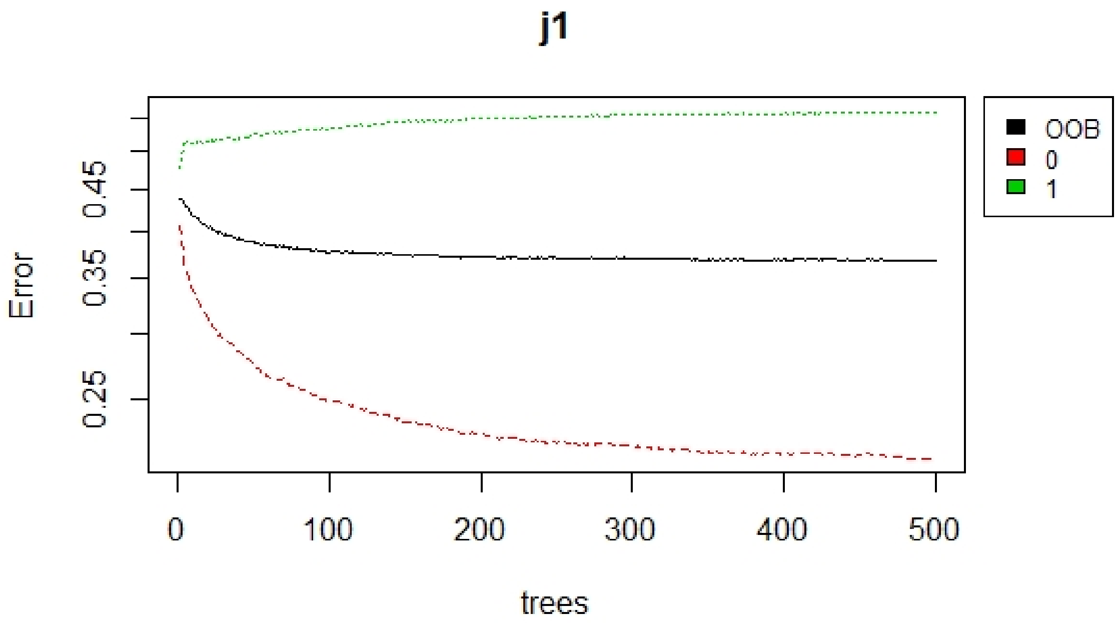
Data Dictionary



Variable Name	Description	Treatment
GAME_ID	Unique ID for each game/matchup	
LOCATION	(H)Home/(A)Away	Converted to dummy
FINAL_MARGIN	Points margin at the end of the game	
PERIOD	Quarter	
SHOT_CLOCK	Time (in sec) left to make shot	NA replaced with 0
TOUCH_TIME	Time (in sec) the player held the ball	Negative values dropped
PTS_TYPE	Type of shot attempted	Cross validated with SHOT_DIST, outliers dropped
CLOSEST_DEFENDER	Nearest defender	
CLOSE_DEF_DIST	Distance between shooter and closest defender (in ft)	
PTS	Final points made	
player_id	Unique id of the player	
pass_shots	Whether the shot was taken without dribbling	DRIBBLES=0
SHOT_NUMBER	Shot Number	
MATCHUP	Team matchup and date of matchup	
GAME_CLOCK	Time left in the quarter	Converted to seconds
DRIBBLES	No. of times the ball was dribbled before shot was taken	
SHOT_DIST	Distance (in ft) from which the shot was taken	Cross validated with PTS type, outliers dropped
SHOT_RESULT	Whether the shot was a success (made/missed)	
CLOSEST_DEFENDER_PLAYER_ID	Unique id of the closest defender	
FGM	Field Goals Made (Dummy equivalent of SHOT_RESULT)	
player_name	Name of the player taking the shot	
open_shots	Whether the shot was open or not	CLOSE_DEF_DIST>6
corner_shots	Whether the shot was a corner shot or not	PTS=3 AND SHOT_DIST<=23

Additional Graphs

Random Forest Estimation



Deployement

Additional analysis are included in an excel file which is also being submitted.