Play to Win—— Analysis and Prediction of NBA Game's Dataset

Dong Wenhui (G2001154H), He Hongshan (G2001279B), He Yuting (G2001441C)

Abstract

Machine learning is now widely used in a variety of fields. Algorithms can be used to analyze a basketball games' data to provide useful guidance on strategies to improve team performance and deployment. This paper focuses on the subject of "the era of three-pointers" and "four factors of basketball success", and proposes research questions based on these two subjects. The study uses regression and classification algorithms to construct a machine learning prediction model based on historical NBA game data sets and analyze factors that affect the 3-pointer and the prediction of game outcomes. The research aims to provide scientific advice and technical basis for basketball game tactics through machine learning and data analysis, and hopes to provide some ideas for applying algorithm science to sports.

1.Introduction

The NBA, known as the National Basketball Association organized in the United States, is the most popular basketball league around the world.It is the need for team cooperation, tactical arrangements and on-the-spot performance that all make basketball games so wonderful and exciting to watch. Besides watching players' performance on the court, statistical experts and fans are also curious about the factors that affect the winning rate of the game, and how to arrange a team's winning tactics based on these factors. Therefore, this project will try to use several machine learning algorithms to analyze past statistical data of NBA games, build machine learning models to predict the results of NBA games, so as to get insights for tactical arrangements.

2. Literature Review

Nowadays game analysts have seen information explosions when conducting the analytical job, therefore it's important to extract data and decide which elements are the most influential to the game. Besides, feature engineering is the essential process when building a machine learning model, since the model's output will largely depend on the input features. This paper seeks to look for the stats that have great influences on the game's results.

Instead of extracting features from a dataset randomly and aimlessly, this paper turned to classic basketball theories to look for support for the feature engineering.

2.1 The Three-Pointer Effect in the "Three-Point Era"

There are three kinds of field goals in one basketball match: One-Pointer from the free throws, Two-Pointer which is the shot within the three-point line, and Three-Pointer which is the shot beyond the three-point line. Since reducing the distance to the rim will increase the chance of scoring, NBA teams used to favor two-point jump shots much more than three-point attempts. While in 1985 a team only made 2.4 three point attempts per game, the trend has changed in the past 20 years. 2014-15 was the first season that NBA teams were more likely to make three-point attempts rather than two-point jump shots. Media claimed that the NBA has entered the "Three-Point Era" (Shea, 2019).

As nowadays every NBA team is shooting more and more three-pointers, people might wonder could the increase of three-point attempts really lead to a team's higher winning rate? And what other effects will be brought to the game's tactics when every team is significantly encouraging their players to shoot three-pointers?

2.2 The "Four Factors of Basketball Success" Theory

The " Four Factors of Basketball" was a theory put forward by data analyst and NBA assistant coach Dean Oliver. In this theory, shooting efficiency, turnovers, rebounds and free throws are the key factors to win a game. In order to evaluate team's performance in these four factors, Effective Field Goal Percentage (eFG%), Turnover Percentage (TOV%), Offensive and Defensive Rebound Percentage (ORB% and DRB%), and Free Throw Percentage (FTR%) were introduced, both can be used on offence and defense dimensions with different weights assigned. In Dean's equation to predict the team's game results, shooting consists of 40%, turnovers consists of 25%, rebounding consists of 20%, and free throws has the weight of 15% (Kotzias, 2018).

Since the prediction equation was manually created, people might wonder how these four factors can contribute to a team's win? And what if a machine learning model can be built based on real game data to offer a computer's solution to this model? Will the model have better performance than manually created equations?

Based on the literature review, the following two research questions were proposed:
RQ1*: What can more three-point attempts bring to the NBA game?*
RQ2: *How do the "Four Factors of Basketball Success" perform in the machine learning prediction model?*

3. Datasets Descriptions

Two datasets from *Kaggle.com* were introduced to conduct the machine learning models. Each dataset contains detailed information of every NBA game.

	index	GAME_ID	TEAM_ID	TEAM_ABBREVIATION	TEAM_CITY	PLAYER_ID	PLAYER_NAME	START_POSITION	COMMENT	MIN	...	OREB	DREB
0	0	21900895	1610612749	MIL	Milwaukee	202083	Wesley Matthews	F	NaN	27:08	...	4.0	4.0
1	1	21900895	1610612749	MIL	Milwaukee	203507	Giannis Antetokounmpo	F	NaN	34:55	...	2.0	18.0
2	2	21900895	1610612749	MIL	Milwaukee	201572	Brook Lopez	C	NaN	26:25	...	2.0	5.0
3	3	21900895	1610612749	MIL	Milwaukee	1628978	Donte DiVincenzo	G	NaN	27:35	...	1.0	6.0
4	4	21900895	1610612749	MIL	Milwaukee	202339	Eric Bledsoe	G	NaN	22:17	...	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
576777	576777	11200005	1610612743	DEN	Denver	202706	Jordan Hamilton	NaN	NaN	19	...	0.0	2.0
576778	576778	11200005	1610612743	DEN	Denver	202702	Kenneth Faried	NaN	NaN	23	...	1.0	0.0
576779	576779	11200005	1610612743	DEN	Denver	201585	Kosta Koufos	NaN	NaN	15	...	3.0	5.0
576780	576780	11200005	1610612743	DEN	Denver	202389	Timofey Mozgov	NaN	NaN	19	...	1.0	2.0
576781	576781	11200005	1610612743	DEN	Denver	201951	Ty Lawson	NaN	NaN	27	...	0.0	2.0

576782 rows × 29 columns

Table 1. 2004-2020 NBA season dataset

The first dataset captured general information for Regular Season and Playoff Season NBA games from October 2004 to March 2020,with 23194 games counted. It has 576781 rows of player stats in every game, and 22 features ranging from the date, player's name,home court, to game stats such as field goals, rebounds,etc. This covers almost every game in the 2000's, therefore is suitable for the trend analysis in RQ1. However, since this is the dataset with player's details counted, it cannot show key values on a team's aspect such as total 3-point attempts per game. Therefore data preprocessing is required before building the model.

	index	gmDate	gmTime	seasTyp	offLNm1	offFNm1	offLNm2	offFNm2	offLNm3	offFNm3	...	opptFIC40	opptOrtg	opptDrtg	opptEDiff	oppt
0	0	2012/10/30	19:00	Regular	Brothers	Tony	Smith	Michael	Workman	Haywoode	...	61.6667	105.6882	94.4447	11.2435	
1	1	2012/10/30	19:00	Regular	Brothers	Tony	Smith	Michael	Workman	Haywoode	...	56.0417	94.4447	105.6882	-11.2435	
2	2	2012/10/30	20:00	Regular	McCutchen	Monty	Wright	Sean	Fitzgerald	Kane	...	80.8333	126.3381	112.6515	13.6866	
3	3	2012/10/30	20:00	Regular	McCutchen	Monty	Wright	Sean	Fitzgerald	Kane	...	62.7083	112.6515	126.3381	-13.6866	
4	4	2012/10/30	22:30	Regular	Foster	Scott	Zielinski	Gary	Daken	Eric	...	58.6458	99.3678	108.1034	-8.7356	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
14753	14753	2018/4/11	10:30	Regular	Garretson	Ron	Mott	Rodney	Richardson	Derek	...	77.2917	113.0207	98.2788	14.7419	
14754	14754	2018/4/11	10:30	Regular	Cutler	Kevin	Kennedy	Bill	Lewis	Eric	...	67.9167	104.4268	95.2126	9.2142	
14755	14755	2018/4/11	10:30	Regular	Cutler	Kevin	Kennedy	Bill	Lewis	Eric	...	54.7718	95.2126	104.4268	-9.2142	
14756	14756	2018/4/11	10:30	Regular	Tiven	Josh	Orr	J.T.	Foster	Scott	...	65.8333	104.3633	90.2307	14.1326	
14757	14757	2018/4/11	10:30	Regular	Tiven	Josh	Orr	J.T.	Foster	Scott	...	34.8548	90.2307	104.3633	-14.1326	

14758 rows × 122 columns

Table 2. 2012-2018 season team box score dataset

The second dataset captured details of every game from 2012 to 2018, both Regular Season and Playoff Season. There are 14757 rows of game information, and 122 columns of game features. The stats are on a team's aspect, and more advanced metrics are calculated such as team's total turnovers, opponent team's offensive rebounds, etc. Compared to the first dataset, the second one is more suitable for RQ2 because it's more convenient to analyze the "Four Factors",which requires both on-court teams' advanced metrics.

4. Building Machine Learning Model to Evaluate the Three Pointers' Effects in Game Wins

4.1 Data processing

Import the two tables of games and game_details which contain detailed information about each game and each player from 2003-2019. Since the Research Question 1 focuses on effects that more 3-pointer attempts will bring, the variables associated with the 3-Pointer and the outcome in the original data are selected and a new table is merged for analysis.

GAME_ID	TEAM_ID	FGM	FGA	FG_PCT	FG3M	FG3A	FG3_PCT	PTS	REB	OREB	DREB	SEASON	HOME_TEAM_ID	VISITOR_TEAM_ID	WIN_TEAM_ID
21900895	1610612749	3.0	11.0	0.273	2.0	7.0	0.286	8.0	8.0	4.0	4.0	2019	1610612766	1610612749	1610612749
21900895	1610612749	17.0	28.0	0.607	1.0	4.0	0.250	41.0	20.0	2.0	18.0	2019	1610612766	1610612749	1610612749
21900895	1610612749	4.0	11.0	0.364	1.0	5.0	0.200	16.0	7.0	2.0	5.0	2019	1610612766	1610612749	1610612749
21900895	1610612749	1.0	5.0	0.200	0.0	3.0	0.000	2.0	7.0	1.0	6.0	2019	1610612766	1610612749	1610612749
21900895	1610612749	2.0	8.0	0.250	0.0	1.0	0.000	4.0	1.0	1.0	0.0	2019	1610612766	1610612749	1610612749

Table 3. Merged table for analysis

Since the initial data only include the player data for each team, it is necessary to calculate the team data by counting the sum of the data of the players. The

Groupby function in python makes it available to group, cluster, and find the sum of the various points of each team in each game. This method first filters and groups the data according to the dimensions of the GAMA_ID, SEASON, TEAM_ID, WIN_TEAM_ID, HOME_TEAM_ID, VISITOR_ TEAM_ID, and then extracts features of PTS, FG3M, FGM, FG3A, FGA, REB, DREB, DREB and DREB according to these dimensions, summarizes several fields to get the following tables.

GAME_ID	SEASON	TEAM_ID	WIN_TEAM_ID	HOME_TEAM_ID	VISITOR_TEAM_ID	PTS	FG3M	FGM	FG3A	FGA	REB	OREB	DREB
10300001	2003	1610612742	1610612762	1610612762	1610612742	85.0	2.0	34.0	8.0	76.0	38.0	12.0	26.0
10300001	2003	1610612762	1610612762	1610612762	1610612742	90.0	1.0	32.0	7.0	70.0	41.0	9.0	32.0
10300002	2003	1610612749	1610612763	1610612763	1610612749	94.0	2.0	32.0	13.0	75.0	43.0	11.0	32.0
10300002	2003	1610612763	1610612763	1610612763	1610612749	105.0	4.0	40.0	15.0	81.0	48.0	14.0	34.0

Table 4. Results after conducting GroupBy function

Data in data tables are executed in rows within the loop, and each row of data executes a conditional statement in the function. Next, feature vectorization is used by using the If Else Statement in Python. If the feature of WIN_TEAM_ID matches the feature of TEAM_ID, it will be assigned a value as 1, or else it will be considered as loss thus the value as 0 will be assigned in order to transfer strings into numerical values that algorithms can understand.

Algorithm 1 Feature Vectorization

```
def set_win_team_id(a, b, c):
    if c == 1:
        return a
    else:
        return b
def set_win_or_lose(a, b):
    if a == b:
        return 1
    else:
        return 0
def set_home_or_visitor(a, b):
    if a == b:
        return 'HOME'
    else:
        return 'VISITOR'
    def set_fg3a_level(a):
    if a <= 0.25:
        return 1
    elif a <= 0.35:
        return 2
    else:
        return 3
```

For the next step, calculate the 3-Point attempts rates (FG3A_ratio) of each team in each game. Relating to the literature review which shows that average NBA teams shot lower than 20% of 3-pointers from their total attempts per game in 2005, and then increased to over 30% in 2016 (Shea, 2019), the 3-Point attempts rates (FG3A_ratio)were splitted into the following three intervals (FG3A_level): 0-25%, 25-35%, and above 35%. Additionally,since in the basketball game,if a team missed a shot, the opponent team might get a defensive rebound and add up its chance for shooting in the next round, it will be interesting to consider adding the opponent's defensive rebound to evaluate if more 3-pointer attempts could affect the opponent team's stats. To calculate the opponent's long rebounds (OPPONENT_DREB) in each game and other features, the equations are shown below:

FG3A_ratio = FG3A (3-Point Field Goal Attempts) / FGA (Field Goal Attempts)
OPPONENT_DREB = sum_DREB - DREB (Defensive Rebounds)
FG3_PTS_ratio = FG3M (3-Point Field Goals Made) / PTS (Points)
FG3M_PCT (3-Point Field Goal Percentage) = FG3M (3-Point Field Goals Made) / FG3A(3-Point Field Goal Attempts)

4.2 FG3A & OPPONENT_DREB

By using the seaborn package in python and drawing the plot of a heatmap, it is available to tell the correlations between two features.

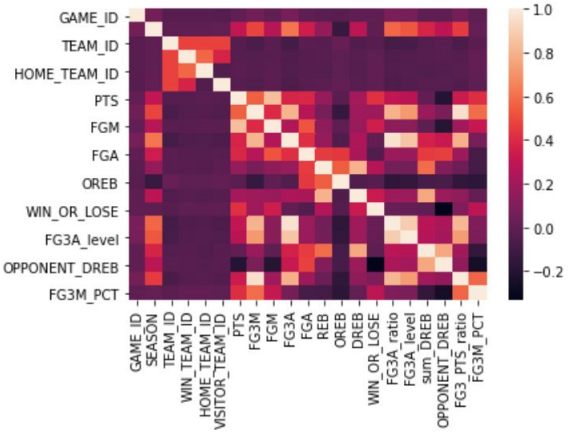

Figure 1. Heatmap of the correlations between two features

In the heatmap drawn above, the lighter color of each small matrix represents a higher correlation between two features. From the heatmap it is available to tell that the opponent's defensive rebound has correlation with the 3-pointer attempts level.

To further develop this finding, we calculated the relationship between 3-Point attempts (FG3A) in each interval and opponent Defensive rebounds (DREB) from 2003 to 2019, as shown in the figure below. It can be seen from the figure that the 3-Point attempts (FG3A) basically increased over the years. As the three-point attempts have increased over the years, opposing rebounding has essentially increased.

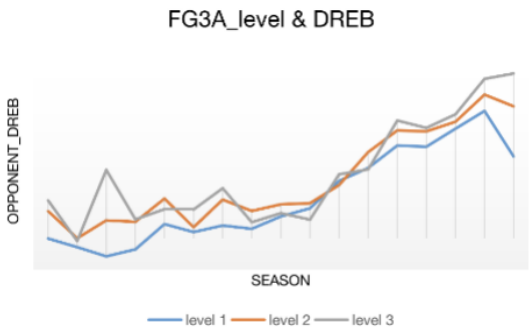

Figure 2. Relationship between FG3A and DREB

Then the relationship between the number of three-point attempts taken at home and away and the number of rebounds an opponent has obtained has been evaluated. Figures show that the number of home rebounds is slightly higher than the number of away rebounds, indicating that home and away games have a certain influence on the number of rebounds.

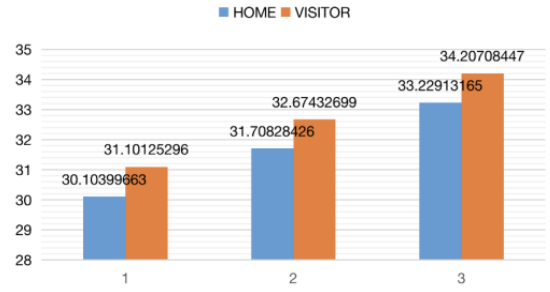

Figure 3. Home and visitor

4.3 FG3A & WIN_RTO

The 3-Point attempts rate also showed an upward trend from the previous season. The figure shows that in the third interval, where the 3-Point attempts rate is greater than 35%, there is a significant increase, while the first interval, where the 3-Point attempts rate is less than 25%, is significantly reduced.

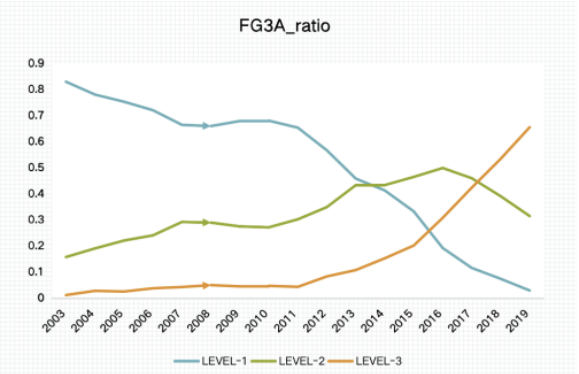

Figure 4. Trend of FG3A

However, in contrast, the game winning percentage corresponding to the 3-Point attempts rate has fluctuated almost horizontally with the changing trend of the season, and there has been no obvious increase or decrease.

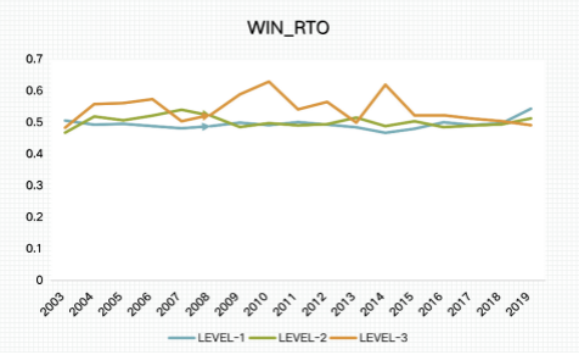

Figure 5. Winning percentage corresponding to FG3A

4.4 Use GBDT to examine factors affect wins from perspective of Three-point

Accordingly, we turn our attention to exploring factors that influence winning from the perspective of the 3-Point attempts. Combining the above exploration, we use the GBDT model to examine the rankings of the FG3_PTS_ratio, FG3A_ratio, FG3M_PCT, FG3A and OPPONENT_DREB at the impact on winning percentage. GBDT (Gradient Boosting Decision Tree) is an iterative weak learner algorithm. The algorithm consists of multiple weak learners and the results of all trees are added together to generate a final response.

Algorithm 2 Gradient Boosting Decision Tree

```
model_data = games_3fg[['FG3_PTS_ratio',
'FG3A_ratio', 'FG3M_PCT', 'FG3A',
'OPPONENT_DREB']]
model_target =  games_3fg[['WIN_OR_LOSE']]
```

```
X_train,X_test,Y_train,Y_test =
train_test_split(model_data,model_target,test_size =
0.2, random_state=3)
clf = GradientBoostingClassifier(learning_rate=0.1
      , n_estimators=300
      , max_depth=5
      , subsample=0.5
      , min_samples_split=5
      , min_samples_leaf=1
      , init=None
      , random_state=1
      , max_features=None
      , verbose=0
      , max_leaf_nodes=None
      , warm_start=False)
clf.fit(X_train,Y_train)
y_trainP =clf.predict(X_train)
y_testP = clf.predict(X_test)
```

Algorithm 3 Classification report

```
from sklearn.metrics import classification_report
print(classification_report(Y_test,y_testP))
```

	precision	recall	f1-score	support
0	0.72	0.70	0.71	4656
1	0.71	0.72	0.72	4583
accuracy			0.71	9239
macro avg	0.71	0.71	0.71	9239
weighted avg	0.71	0.71	0.71	9239

```
feature_importances = clf.feature_importances_
print(feature_importances)
```

```
[0.21543465 0.14755908 0.25204254 0.11661939 0.26834434]
```

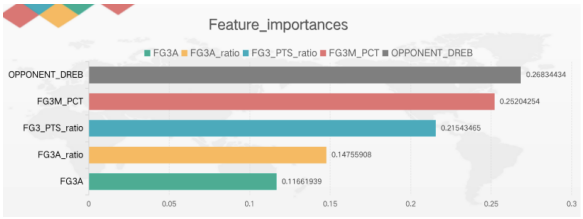

Figure 6. Feature importance

According to the results, an accuracy rate of 71.4% has been gained from the model, therefore it is reasonable to come up with several valuable insights from it. As shown in the figure above, OPPONENT_DREB, FG3M_PCT and FG3_PTS_ratio have the greatest impact on the winners. This is consistent with the relationship between the three-point shot (FG3A) and the opponent's rebounds (OPPONENT_DREB). It is by overall consideration of the opponent's rebounds, 3-point mades and three-point attempt ration that improves the model performance.

Based on the 3-pointer analysis results above, historical data shows that while more three-pointers are shot, the opponent's defensive rebounds increase, and this trend is strengthening. Judging from the game results, the opponent's rebounds play an important role in the victory. The more rebounds an opponent receives, the greater the opponent's winning percentage. From the tactical aspect, the explanation could be: while more 3-pointer attempts are made, if the accuracy of every shot doesn't increase accordingly, it will be easier for the opponent to grab the defensive rebound, push more fast breaks, increase the game's pace and possessions. Thus more 3-pointer attempts is not equal to a higher chance of winning. For the three-point tastics it is important to improve shooting percentage by assigning more elite 3-point shooters while assigning defensive players that are good at grabbing offensive rebounds and defending fast breaks at the same time.

5. Building Machine Learning Model to Predict Game's Results Based on Oliver's Four Factors

In this section, we will discuss four factors that have a significant impact on the winning rate of basketball games, which has been stated in the literature review. We will use seven different models to predict the accuracy of "Four Factors of Basketball Success" from Dean Oliver in influencing the winning rate of the game, so as to prove the importance of these four measures, which subsequently to further improve the team's strategy in the later stage.

5.1 Data definitions and preprocessing:

The second dataset of NBA Enhanced Box Score and Standings Data from 2012 to 2018 has been used to conduct the model, and the equations of measuring the "Four Factors" are listed as follows:

eFG%: Effective Field Goal Percentage, the formula is (FG + 0.5 * 3P) / FGA. This statistic will adjust based on the fact that a 3-point field goal is worth one more point than a 2-point field goal. We calculated the team's and opponent's Effective Field Goal Percentage as given in the dataset.

TOV%: Turnover Percentage (available since the 1977-78 season in the NBA), the formula is 100 *

TOV / (FGA + 0.44 * FTA + TOV). This is an estimate of turnovers per 100 plays. We calculated the team's and opponent's Turnover Percentage (TOV%) as not given in the dataset.

OREB%: Offensive Rebound Percentage, DREB%: Defensive Rebound Percentage. The formulas are teamOREB%= teamORB / (teamORB + opptDRB), teamDREB%= teamDRB / (opptORB + teamDRB). We calculated the team's Offensive and Defensive Rebound Percentage (OREB% & DREB%) as not given in the dataset.

FTR%: Free Throw Percentage, the formulas are: teamFTR%= teamFTA/teamFGA, opptFTR%= opptFTA / opptFGA. We calculated the team's and opponent's Free Throw Rate (teamFTR% & opptFTR%) as not given in the dataset.

Algorithm 4 "Four Factors" values measuring

```
data['teamTOV%']=data['teamTO']/ (data['teamFGA'] + 0.44*data['teamFTA']+data['teamTO'])

data['opptTOV%']= data['opptTO'] / (data['opptFGA'] + 0.44*data ['opptFTA']+data['opptTO'])

data['teamOREB%']=data['teamORB']/ (data['teamORB']+data['opptDRB'])

data['teamDREB%']=data['teamDRB']/ (data['opptORB']+data['teamDRB'])

data['teamFTR%']=data['teamFTA']/ data['teamFGA']

data['opptFTR%']=data['opptFTA'] / data['opptFGA']
```

5.2 Models:

In order to construct the machine learning model better and evaluate the performance of the final trained model, before using machine learning algorithm, we usually need to divide the data set into training set and test set. When the training set and the test set are allocated, if the data of the test set is smaller, the estimation of the generalization error of the model will be more inaccurate. Therefore, in this section, based on the size of the whole dataset, we set the partition ratio of training set data and test set data as 7:3.

Classifiers can help us better understand and compare algorithms. We built seven models: Logistic Regression Model, Support Vector Classifier Model, Linear Support Vector Classifier Model, BaggingSVC Ensemble classification Model, Random Forest Classifier Model, KNeighbors Classifier Model and XGBoost Model.

The following table shows the different accuracy and training time obtained by the seven models:

Model	Accuracy	Training Time (s)
Logistic Regression Model	93.1%	0.036
Support Vector Classifier Model	94.5%	4.71
Linear Support Vector Classifier Model	94.4%	0.051
BaggingSVC Ensemble classification Model	94.5%	3.35
Random Forest Classifier Model	89.7%	2.58
KNeighbors Classifier Model	88.7%	0.017
XGBoost Model	93.0%	0. 362

Table 5. Comparisons between different models

In order to compare the difference of accuracy and training time more simply and intuitively, we use the charts below:

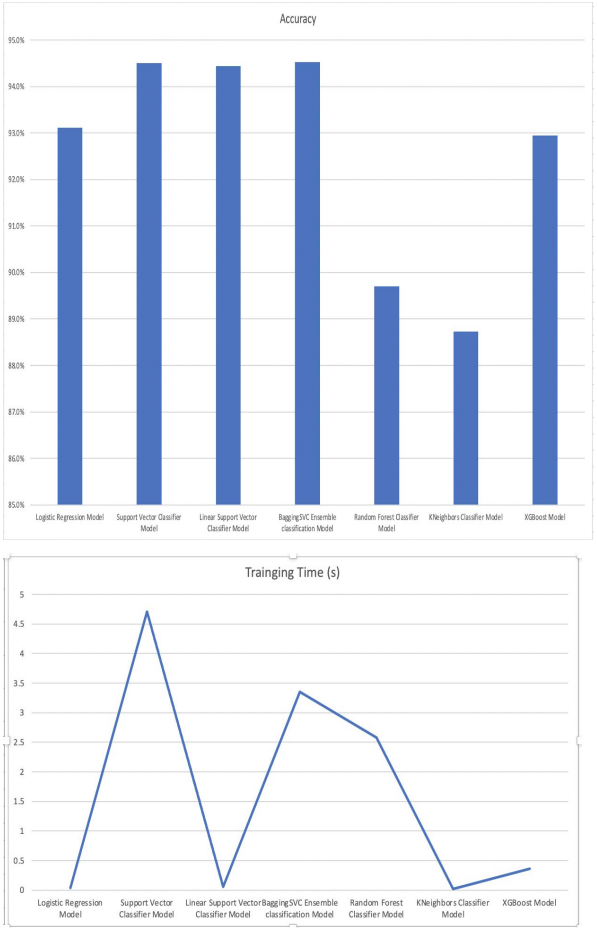

Figure 7. Accuracy and Training Time of different models

The results show that the support vector classifier model and the BaggingSVC Ensemble classification Model have the highest accuracy (94.5%). Correspondingly, their training time is longer. However, the random forest classifier model has poor comprehensive performance in training samples. The support vector classifier model helps us to select the few training samples which are most effective for the prediction task, which not only saves the data memory required for model learning, but also improves the prediction performance of the model. The BaggingSVC Ensemble classification model can also effectively avoid over fitting

Algorithm 5 Support Vector Classifier Model

```
from sklearn.svm import SVC
from sklearn import svm
clf = SVC(gamma='scale',probability=True)
%time clf.fit(X_train, y_train)
clf.score(X_test,y_test)
```

Algorithm 6 BaggingSVC Ensemble classification model

```
from sklearn.svm import SVC
from sklearn.ensemble import BaggingClassifier
advclf=BaggingClassifier
(base_estimator=SVC(gamma='scale'),
n_estimators=10, random_state=0)
%time advclf.fit(X_train, y_train)
advclf.score(X_test,y_test)
```

On the one hand, through the detection of the above models, it is easy to see that four factors (Shooting, Turnovers, Rebounding and Free Throws) have a great influence on the winning rate of a game. Therefore, data analysis based on these four factors can help a team to formulate and adjust its tactics both before and during the games.

On the other hand, although Oliver coined the term of "four factors", the process is actually eight factors. Teams that score more don't necessarily win, they have to stop their opponents from scoring at the same time. But rebounding is an exception. In an NBA game, finding a specific balance in scoring, rebounding, controlling the ball and attacking the basket is a necessary condition for winning. Although these four factors can not guarantee the victory of a game, at least, it can let us know how far a team lags behind other teams in these four core areas, and then further adjustment can be made. It can even help the team to develop the players' directional ability through these four specific indicators.

6. Discussion and Conclusion

6.1 Conclusion

Based on the analysis shown above, it is powerful to use machine learning methods to test classic basketball theories, getting insights from the pass-game data, and reflect on team's tactics to improve performance on court. In this paper, the theory of 'Three-Pointer Era" and "Four Factors of Basketball Success" have been reviewed. Based on these two theories, NBA datasets were applied for features analyzing and engineering, and then machine learning prediction models were built to get more valuable insights that can reveal which factors in the game can have high influence on teams' winning performance.

In conclusion of the first analysis part relating to Research Question1, the results revealed that while the trend of more three-pointer attempts does exist in NBA games after 2003 and it seems unstoppable, the increase of such attempts doesn't directly lead a team to win. Instead, more attempts with lower quality (less accuracy) might lead to the increase of opponent's rebounds, which is also an important feature affecting the game's result. It is the combination of 3-point attempts ratio per game, 3-point attempts made and the opponent team's defensive rebounds that have higher effects on the game rather than 3-point attempts alone. Team's tactics can be adjusted considering how to make more 3-point attempts with higher quality while effectively controlling the opponent's rebounds performance at the same time.

For the second analysis relating to Research Question 2, it has been revealed that the "Four Factors" concerning a team's shooting, rebounding, turnovering and free throwing performance do have significant importance on the game's result. Therefore NBA teams should positively reflect on these four features, and develop players' abilities accordingly. Besides, since the final prediction model has a high accuracy rate, it is possible to use it for game prediction and resource for tactical arrangements. For instance, a NBA team can input their average seasonal performance values into the model, and the output will be the game's results for the rest season, so as to provide valuable insights for the coach.

6.2 Limitations and Reflections

Due to the timing and technical capacity limitations, several improvements can be made in the future research. The first limitation is that, to conduct two model buildings by using two datasets from different time ranges seems to be lack of consistency and not so convincing.To solve this problem, more data preprocessing could be done or a better dataset could be found to improve the model performance.

The limitations of usage of the GBDT model used in RQ1 is that in the case of fewer features, there is still room for optimization in adjusting parameters. Methods such as calculating feature average, variance and getMaxWindow could be considered to enrich features. Besides, more features should be taken into consideration in order to have a more general insight of how the "3-Point Era" can change the games.

For the second model-building process relating to RQ2, limitations existed in the model comparison part. By simply calculating the testing accuracy and training time of each model is not enough, the errors and outliers should also be considered.

For the future research, a more advanced and convincing model that can not only effectively combine the two main research insights, but also reveals how a team can respond to every metric emerged from every possession could be done.

The code files and datasets are available at:
https://github.com/paulheyt30/Group-Pythonsonic-Project

Acknowledgement

The research team wants to take this chance to thank Dr.Chen, Dr.Zhao and Dr.Li for providing valuable guidance and feedback throughout the whole research process.

References:

Kotzias, K. (2018, March 15). The Four Factors of Basketball as a Measure of Success. Retrieved November19, 2020, from https://statathlon.com/four-factors-basketball-success/

Basketball: A Guide to Stats. (n.d.). Retrieved November 19, 2020, from https://www.hudl.com/support/v3/breakdown-stats-and-reports/stat-reports/basketball-a-guide-to-stats

Glossary. (n.d.). Retrieved November 19, 2020, from https://www.basketball-reference.com/about/glossary.html

Khan,E. (2013,October 18). Advanced NBA Stats for Dummies: How to Understand the New Hoops Math. Retrieved November 19,2020, from https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math

Shea, S. (2019).The 3-Point Revolution, Retrieved November 19,2020, from https://shottracker.com/articles/the-3-point-revolution

Fayad,A. (2020, July 10). Building My First Machine Learning Model | NBA Prediction Algorithm. Retrieved November 19, 2020, from https://towardsdatascience.com/building-my-first-machine-learning-model-nba-prediction-algorithm-dee5c5bc4cc1

Lauga, N. (February 2020). Dataset with all NBA games from 2004 season to Feb 2020. Retrieved November 16, 2020, from https://www.kaggle.com/nathanlauga/nba-games?select=games.csv

Rossotti, P. (2018). NBA Enhanced Box Score and Standings (2012 - 2018). Retrieved November 16,2020, from https://www.kaggle.com/pablote/nba-enhanced-stats