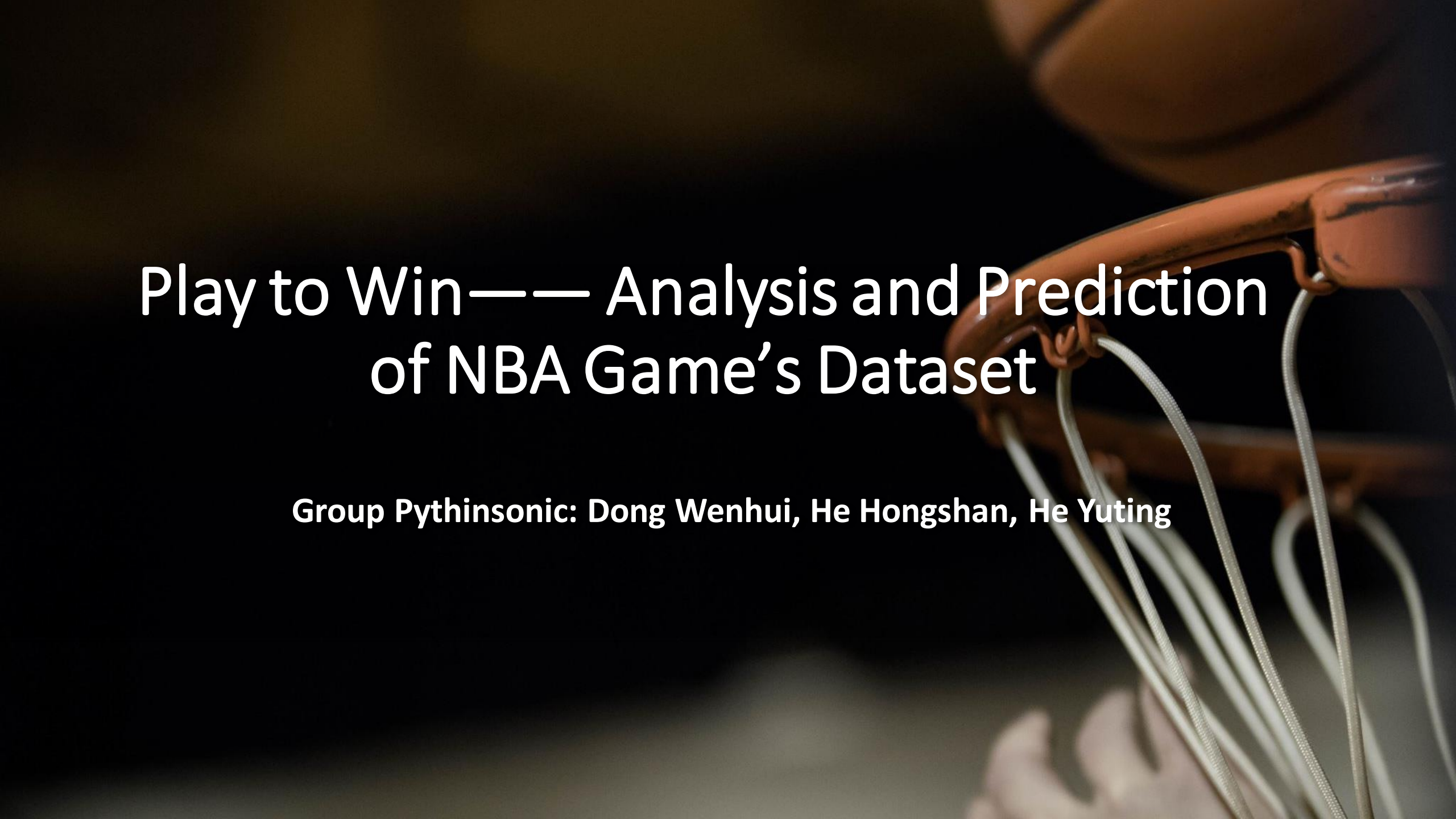
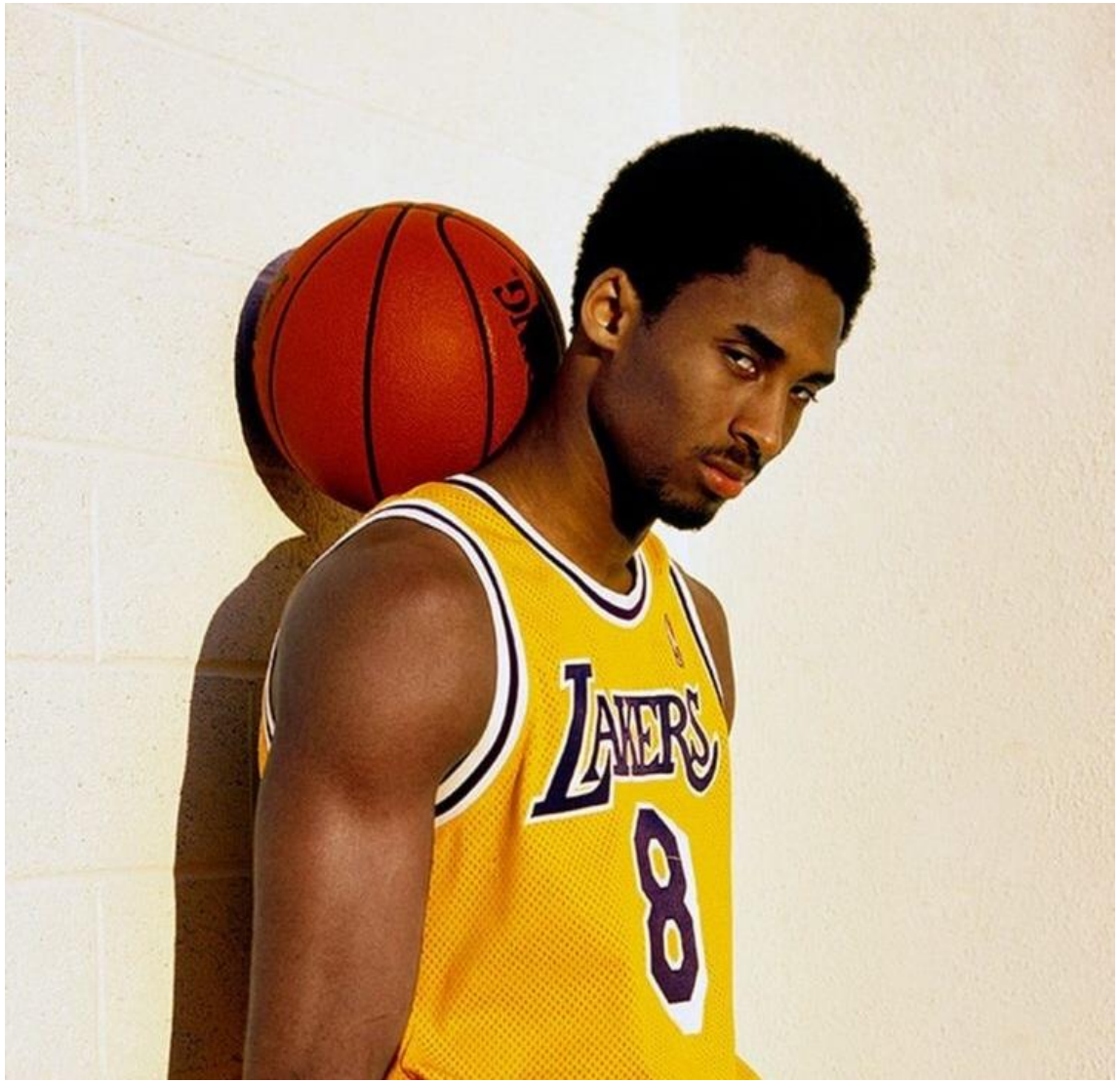
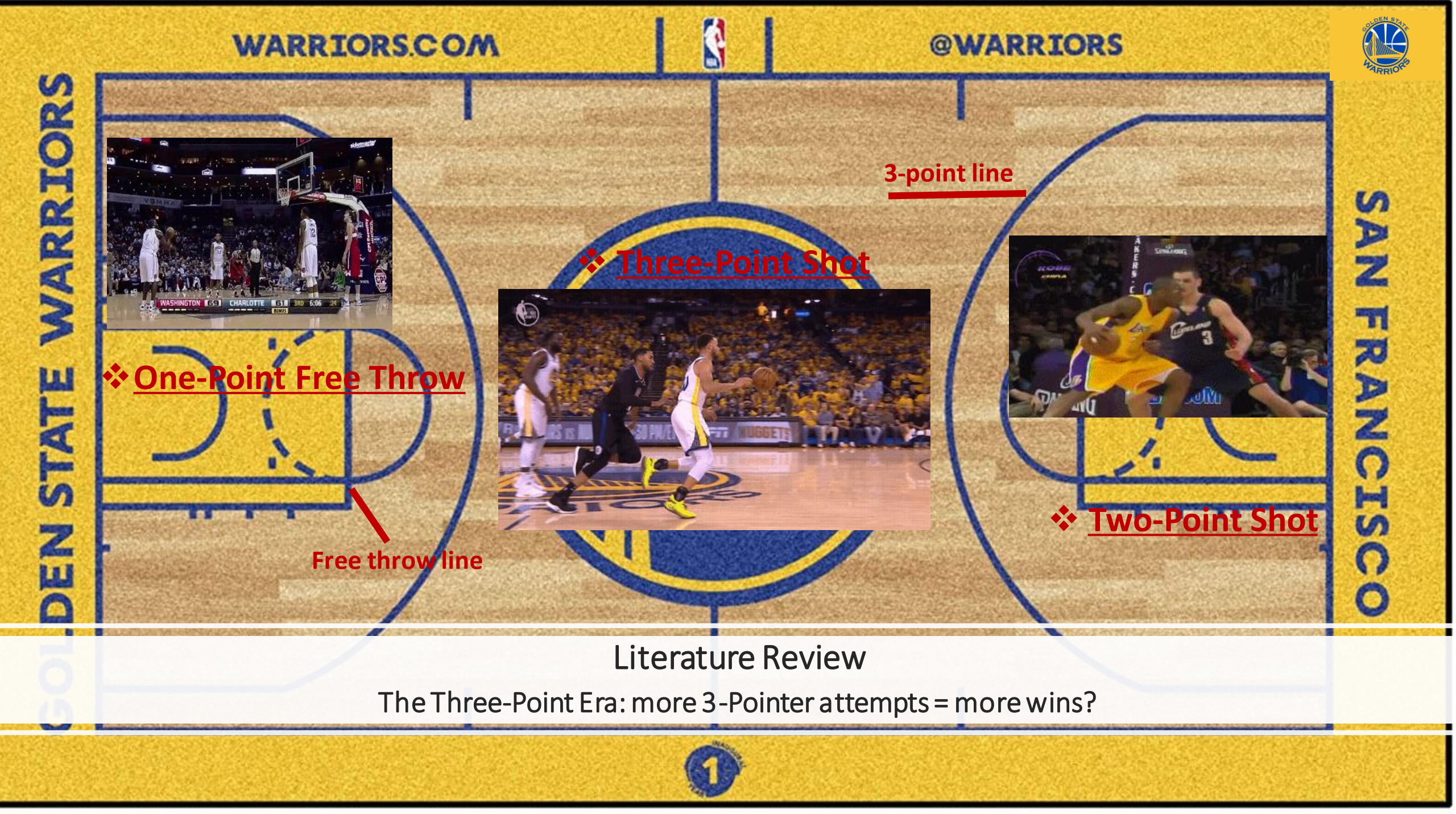


Play to Win—— Analysis and Prediction of NBA Game's Dataset

Group Pythinsonic: Dong Wenhui, He Hongshan, He Yuting

- Literature Review
- Data Description
- Model1: Evaluate the Three Pointers' Effects in Game Wins
- Model2: Predict Game's Results Based on Oliver's Four Factors
- Conclusions & Limitations

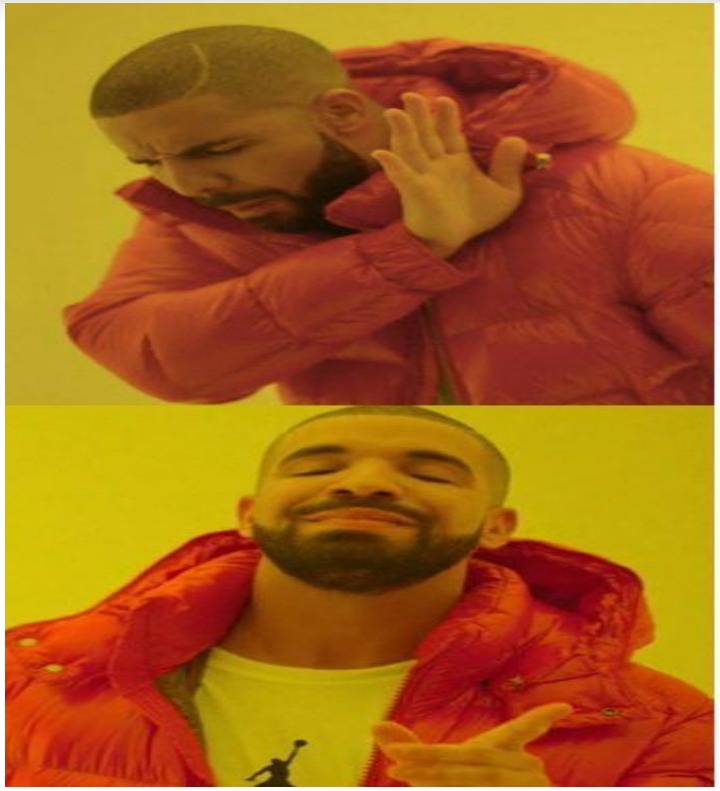




Literature Review

The Three-Point Era: more 3-Pointer attempts = more wins?

Literature Review 2
The "Four Factors"of Basketball Success



Dean Oliver's Model:

- Shooting (40%)
- Turnovers (25%)
- Rebounding (20%)
- Free Throws (15%)

Machine Learning Model:

Projected Wins=w1*teamEFG%+w2*opptEFG%
+w3*opptTOV% +w4*teamOREB%
+w5*teamDREB%+w6*teamFTR%
+w7*opptFTR%

Research Questions

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?

Data Descriptions

	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

Dataset1: 2004-2020 NBA season dataset

- Only contains each player's details,need data preprocessing to sum up team's stats
- Time range is long, suitable to analyze the trend
- Used for RQ1

	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	Dalen	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

Dataset2:2012-2018 season team box score dataset

- Contains more advanced metrics
- More convenient to analyze the “Four Factors”
- Used for RQ2

Model1: Evaluate the Three Pointers' Effects in Game Wins

FG3A & OPPONENT_DREB

FG3A & WIN_RTO

Factors affect winning from perspective of 3-Point

Data processing

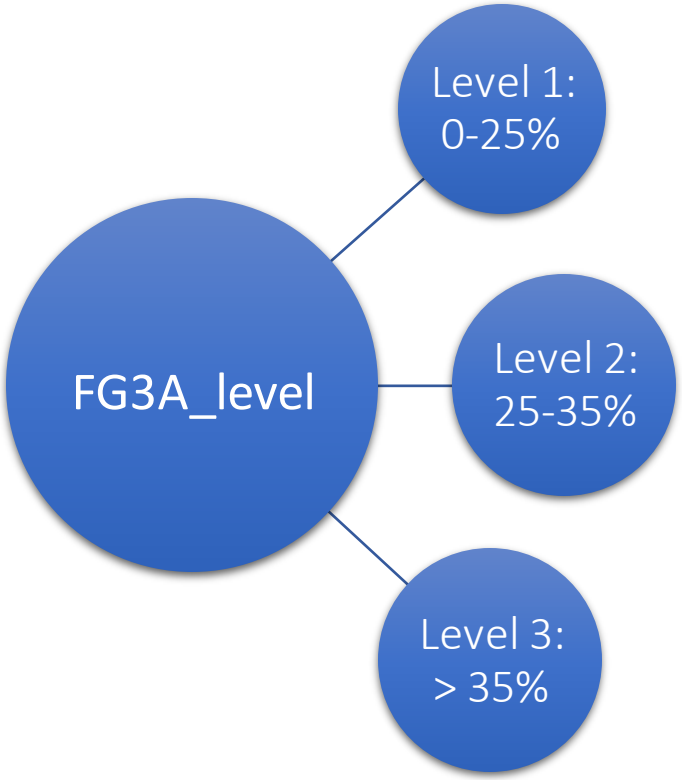
GAME_ID	TEAM_ID	FGM	FGA	FG_PCT	FG3M	FG3A	FG3_PCT	PTS	REB	OREB	DREB	SEASON	HOME_TEAM_ID	VISITOR_TEAM_ID	WIN_TEAM_ID
21900895	1610612749	3.0	11.0	0.273	2.0	7.0	0.286	8.0	8.0	4.0	4.0	2019	1610612766	1610612749	1610612749
21900895	1610612749	17.0	28.0	0.607	1.0	4.0	0.250	41.0	20.0	2.0	18.0	2019	1610612766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

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
10300003	2003	1610612739	1610612739	1610612765	1610612739	100.0	4.0	38.0	6.0	77.0	52.0	12.0	40.0

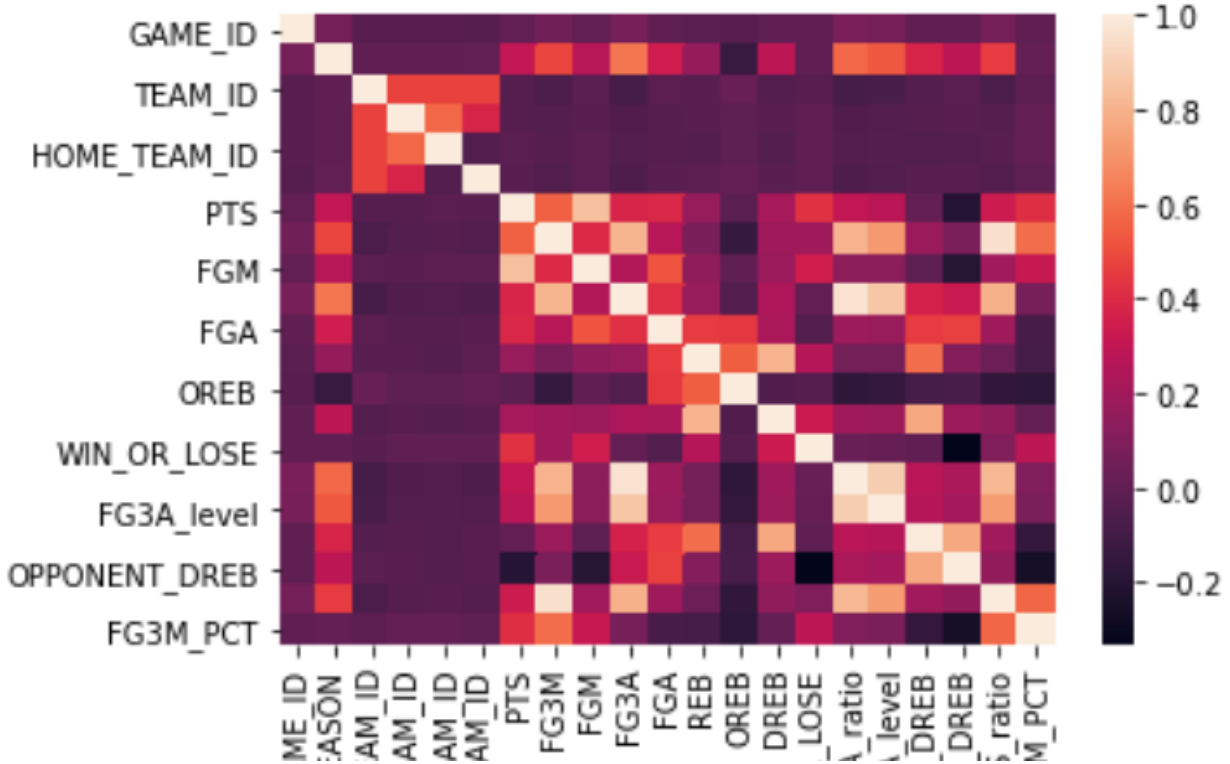
Data processing

- 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)



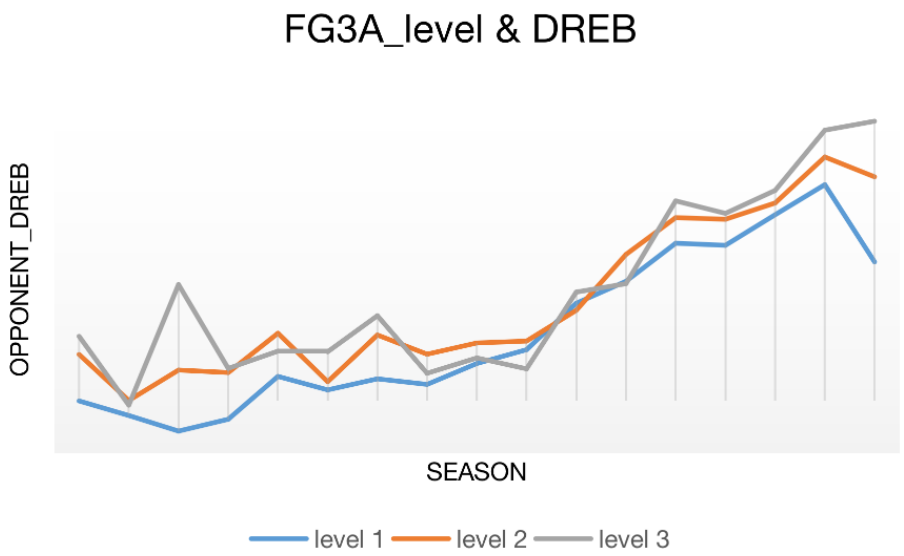
FG3A & OPPONENT_DREB

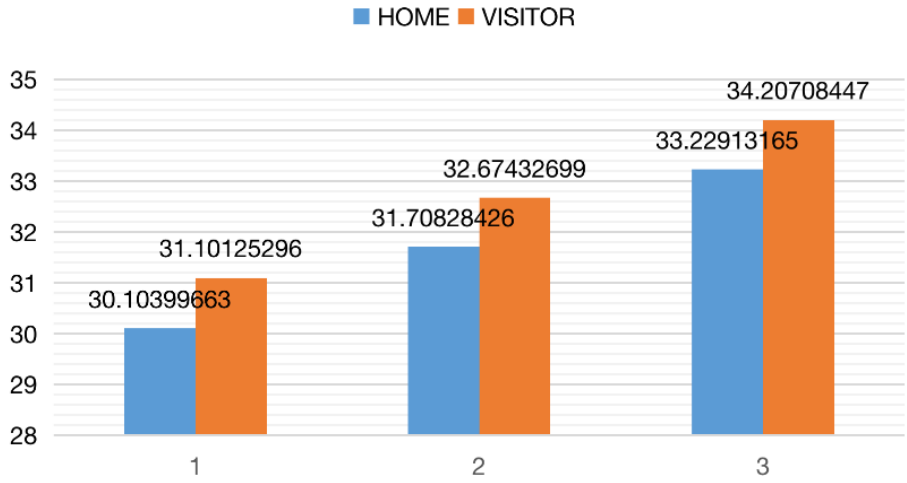


- Heatmap: the lighter color of each small matrix represents a higher correlation between two features.

- The opponent's defensive rebound has correlation with the 3-pointer attempts level.

FG3A & OPPONENT_DREB

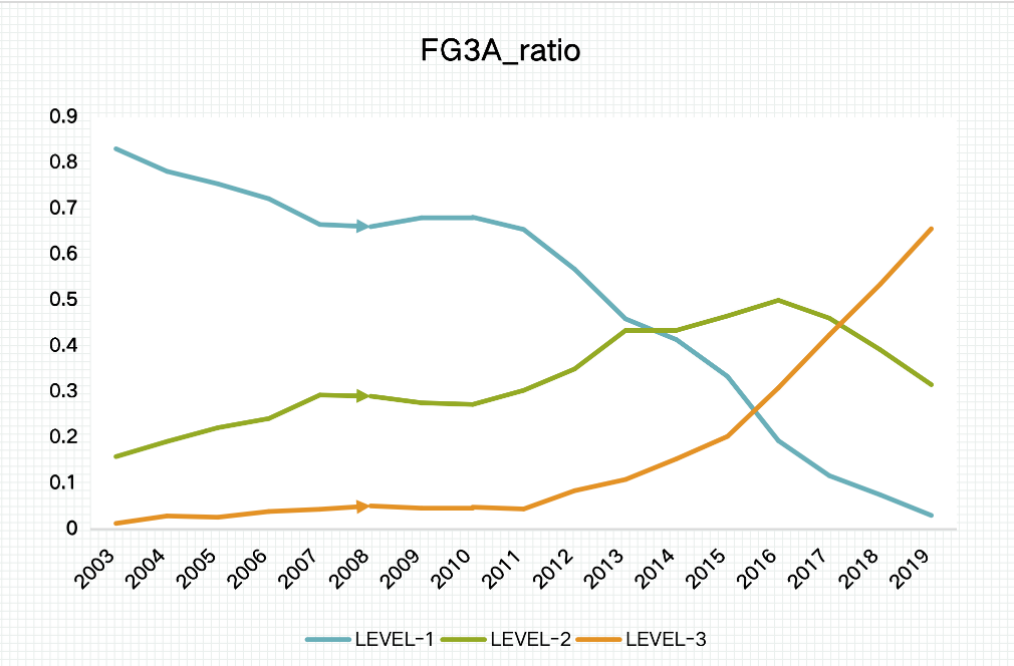


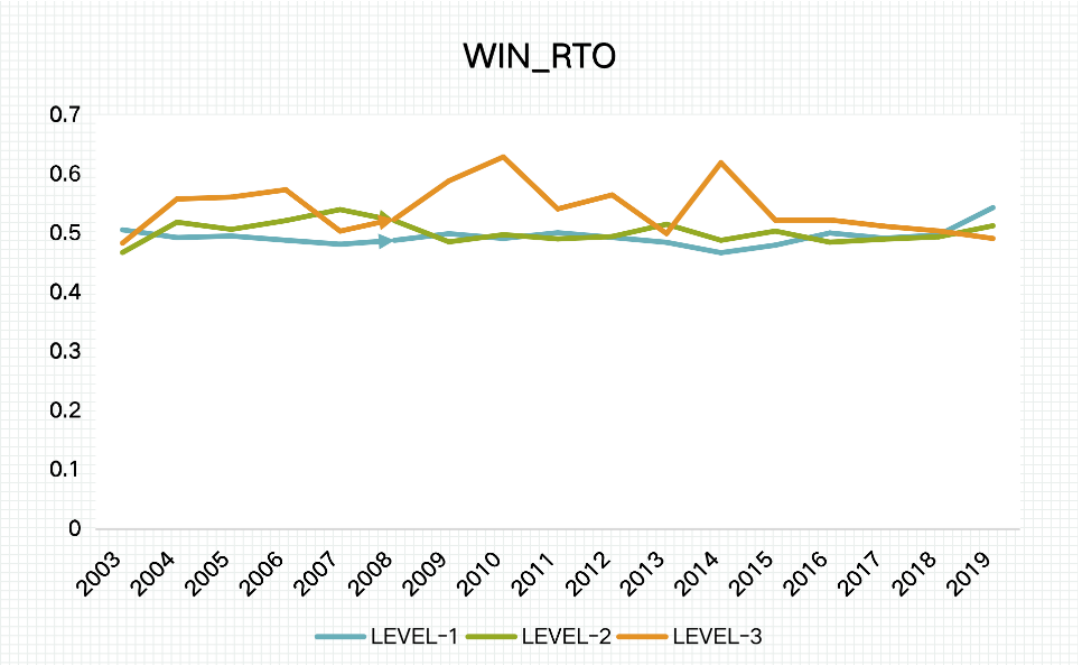


As the 3-Point attempts (FG3A) have increased over the years, opposing rebounding has essentially increased.

The number of home rebounds is slightly higher than the number of away rebounds.

FG3A & WIN_RTO





level 3 (3-Point attempts rate is more than 35%): increase;

level 1 (3-Point attempts rate is less than 25%): reduce

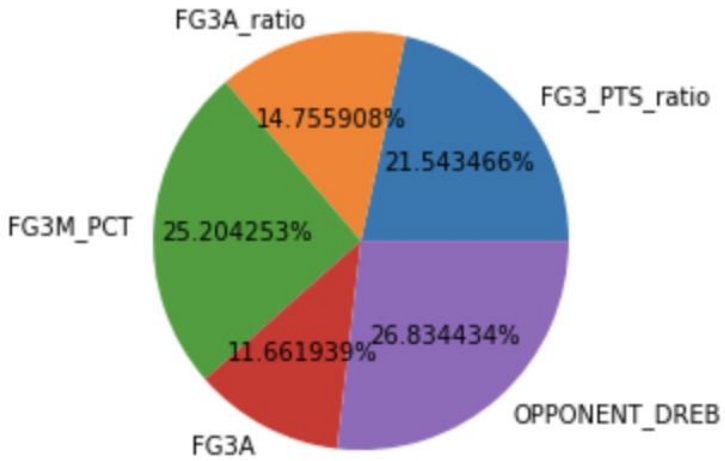
no obvious increase or decrease

Factors that influence winning from the perspective of 3-Point

GBDT (Gradient Boosting Decision Tree)

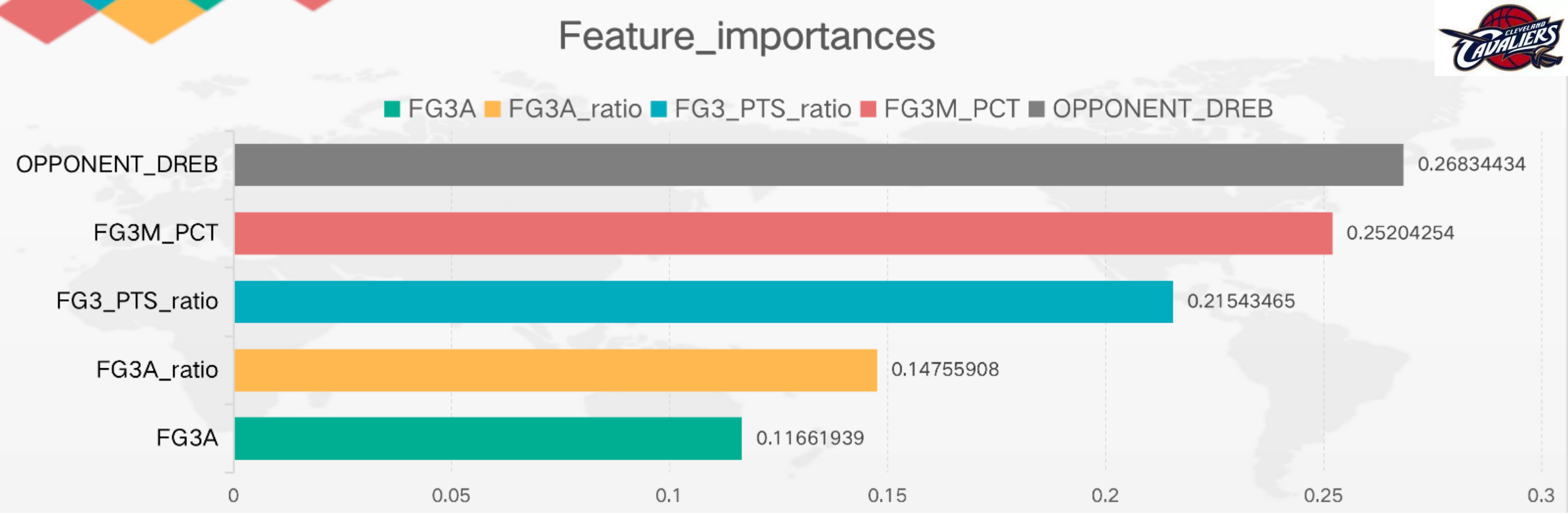
FG3_PTS_ratio, FG3A_ratio, FG3M_PCT, FG3A, OPPONENT_DREB

```
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)
```




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

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



Factors that influence winning from the perspective of 3-Point

- more 3-Points, the opponent's rebounds increase;
- more rebounds an opponent receives, the greater the opponent's winning percentage;
- the home team made a more consistent 3-Pointer;
- improve shooting percentage and get more rebounds to increase winning rate

Model2: Predict Game's Results Based on Oliver's Four Factors

Four Factors of Basketball Success (Dean Oliver)

Data Processing

- Shooting → eFG%: (FG + 0.5 * 3P) / FGA
- Turnovers → TOV%: 100 * TOV / (FGA + 0.44 * FTA + TOV)
- Rebounding → OREB%: teamOREB%= teamORB / (teamORB + opptDRB)

  teamDREB%= teamDRB / (opptORB + teamDRB)
- Free Throws → FTR%: teamFTR%= teamFTA / teamFGA

  opptFTR%= opptFTA / opptFGA

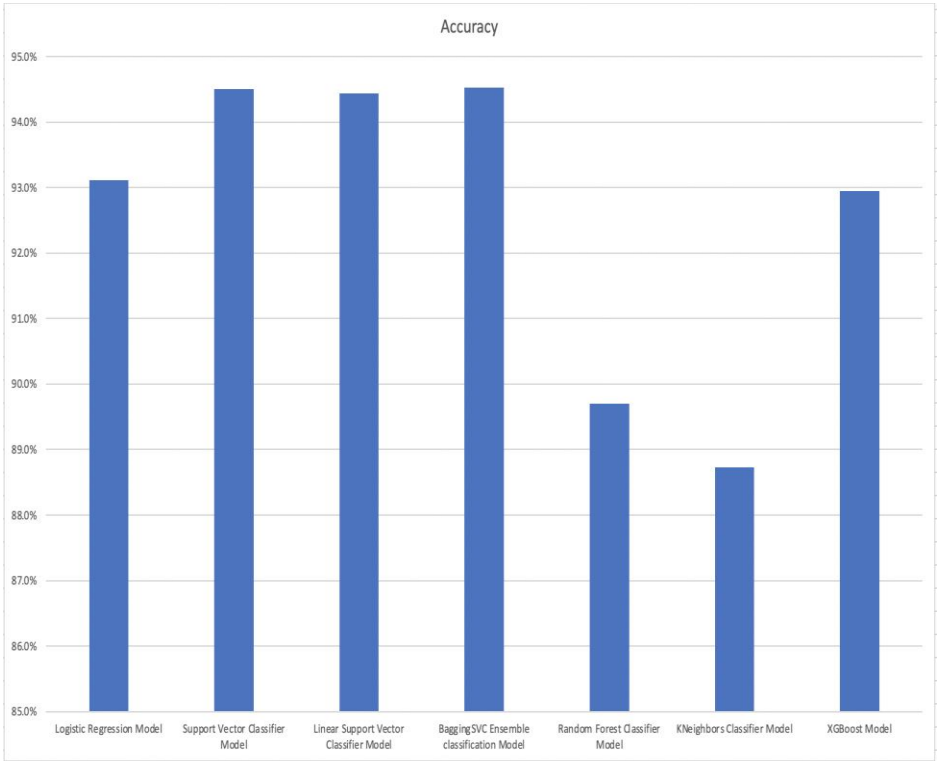
Model2: Predict Game's Results Based on Oliver's Four Factors

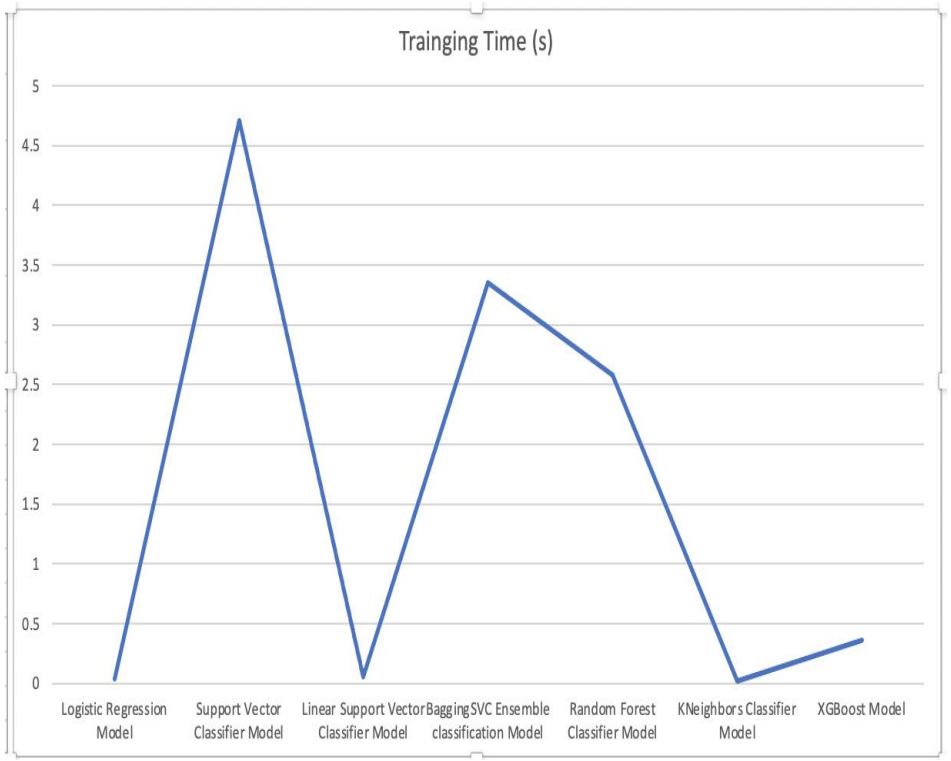
Models:

1. Logistic Regression Model
2. Support Vector Classifier Model
3. Linear Support Vector Classifier Model
4. BaggingSVC Ensemble classification Model
5. Random Forest Classifier Model
6. KNeighbors Classifier Model
7. XGBoost Model

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

Comparison of results





We can see from the models…

- Highest Accuracy: Support Vector Classifier Model & BaggingSVC Ensemble classification Model
- All models' accuracy are above 88%

```
# 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)
```

```
CPU times: user 4.42 s, sys: 271 ms, total: 4.69 s
Wall time: 4.71 s

0.9451219512195121
```

```
# 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)
```

```
CPU times: user 3.19 s, sys: 96 ms, total: 3.29 s
Wall time: 3.35 s

0.9453477868112015
```

Four Factors:

- Data analysis based on four factors can help a team to formulate and adjust its tactics both before and during the games
- Cannot guarantee the victory of a game, but can let us know the gap
- Training players' directional ability

Discussions & Conclusions

RQ1:

- 3-Point attempts are significantly increacing, but is not directly associated with team wins.
- More 3PA will lead to more opponent's defensive rebounds.
- 3PA with higher quality - effectively control opponent’s rebounds – prepare for opponent's more fast breaks

Discussions & Conclusions

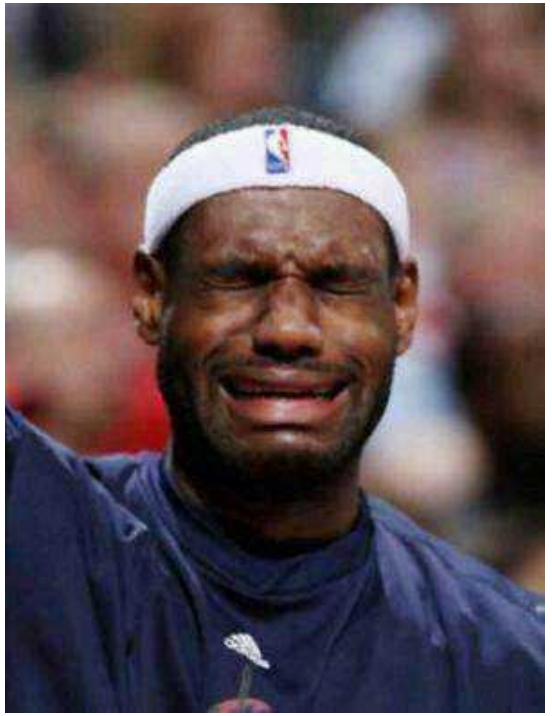
RQ2:

- Four Factors (Shooting, Turnovers, Rebounding and Free Throws)= high influence on team's win

- Game prediction and resource for tatical arrangements

Limitations

- Two datasets from different time ranges were used
- Fewer features were introduced into the GBDT model, not enough for comprehensive insights
- Errors and outliers were not included in the model comparison



Thank you