ai-nba-predict: Predicting the NBA Champion

Project Members

Name	Organization	Email
Youssef Ben Khelil	Paris Institute of Digital Technology	youssef.ben-khelil@eleve.isep.fr
Leo Bertazzoli	Zurich University of Applied Sciences	leobe@gmx.ch
Chang Won Jung	Hanyang University	richardj0916@hanyang.ac.kr

I. Introduction

The National Basketball Assosciation (NBA) is a North American basketball league made up of 30 teams. Every year from October to April, the teams compete in a regular season tournament of 82 games each, in order to determine their seeding in the playoffs. The top 8 teams in each conference are placed in the playoff bracket, where the eventual winner is crowned the NBA Champion.

Because we follow the NBA it would be great if we could predict the champions of the future seasons. On the one hand its just for our interest. On the other hand, if our prediction succeeds, we could start betting on the prediction. The final goal is to predict the champion of the next season by analyzing the past 40 years and find patterns, with the help of AI, with which we can predict future champions.

II. Dataset

Like many sports leagues, the NBA tracks and records various statistics from all of their games. These statistics range from simple statistics such as win percentage, to advanced statistics such as effective field goal percentage. Baksetball Reference provides a massive collection of NBA statistics from team stats to individual player statistics.

The dataset that we used for our project was provided by JK-Future-Github, who used a web crawler to collect data from Basketball Reference. He also added additional features such as Top_3_Conference, which describes whether the team finished within the top 3 in their respective conference. We removed seasons before 1980, as that was the year when the 3 point line was introduced, and the 2023 season, as the data collected from that season was from an incomplete season.

III. Methodology

i. Algorithms

We decided to use different regression algorithms and compare the results of each algorithms predictions. The two algorithms that we chose were Gradient Boosting and Random Forest.

We used the libraries provided by XGBoost and scikit-learn:

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
import numpy as np
import random
from xgboost import XGBRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
import shap
```

ii. Training and Test Data

First, we divided the dataset into training and test data. Instead of a complete random splitting of the dataset, we decided to randomly select two seasons from each decade. The reasoning behind this method is because the NBA game has evolved with time, causing changes in playstyles which altered which skillsets were considered valuable to winning. For example, the modern NBA has seen the rise and absolute necessity of having good three point shooting, whereas in previous eras the three point shot was considered an inefficient shot.

```
#Read CSV and drop empty values
nba_stats = pd.read_csv('nba_team_data.csv')
nba_stats = nba_stats.dropna()

#Create feature columns
exclude_columns = ['season', 'name', 'conference', 'Champion_Win_Share']
feature_columns = [col for col in nba_stats.columns if col not in
exclude_columns]

random_seed = 12345
random.seed(random_seed)

#Split by randomly selecting 2 Seasons each decade and save in train_set
and test_set
nba_stats['season'] = nba_stats['season'].astype(str)
nba_stats['Decade'] = nba_stats['season'].apply(lambda x: x[:3] + '0')
unique_decades = nba_stats['Decade'].unique()
train_set = pd.DataFrame()
test_set = pd.DataFrame()
for decade in unique_decades:
    decade_data = nba_stats[nba_stats['Decade'] == decade]

    # Randomly select two seasons for the test set
    test_seasons = random.sample(list(decade_data['season']), k=2)
    test_set = pd.concat([test_set,
decade_data[decade_data['season'].isin(test_seasons)]])
    train_set = pd.concat([train_set,
decade_data[~decade_data['season'].isin(test_seasons)]])

train_set = train_set.drop(columns=['Decade'])
test_set = test_set.drop(columns=['Decade'])

#Save test season list
test_seasons = test_set['season'].unique()
```

iii. Target Feature

The target feature is `Champion_Win_Share`. This is defined to be the total wins that a team gets in the playoffs divided by the max number of wins that a team can get (16 wins). A team with a greater number of playoff wins will have a higher `Champion_Win_Share`, and be considered to be more successful. The championship team will have a value of 1.

The end goal will be use a single season's data on our model to estimate each team's `Champion_Win_Share`. The team with the highest `Champion_Win_Share` will be considered as the winner of that NBA season.

```
#Create the feature and target sets
X_train = train_set[feature_columns]
y_train = train_set['Champion_Win_Share']
X_test = test_set[feature_columns]
y_test = test_set['Champion_Win_Share']
```

iv. Model Fitting and Predictions

We then trained our model and made predictions with the test dataset.

```
# Set random seed for models
random_seed = 21
random.seed(random_seed)

xgb_model = XGBRegressor(random_state=random_seed)
xgb_model.fit(X_train, y_train)

rf_model = RandomForestRegressor(random_state=random_seed)
rf_model.fit(X_train, y_train)

xgb_predictions = xgb_model.predict(X_test)
rf_predictions = rf_model.predict(X_test)

test_set['XGB_Predicted_Champion'] = xgb_predictions
test_set['RF_Predicted_Champion'] = rf_predictions
```

IV. Evaluation and Analysis

i. Root Mean Squared Error

Calculations of the Root Mean Square Error of XBG and Random Forest suggested that the Random Forest algorithm was slightly more accurate than Gradient Boosting.

```
xgb_rmse = np.sqrt(mean_squared_error(y_test, xgb_predictions))
rf_rmse = np.sqrt(mean_squared_error(y_test, rf_predictions))
```

```
print("XGBoost Mean Squared error:")
print(xgb_rmse)
print("Random Forest Mean Squared error:")
print(rf_rmse)
```

```
XGBoost Mean Squared error:
0.16948661337862525
Random Forest Mean Squared error:
0.14994427109053007
```

ii. Prediction Results

We tested each NBA season from our test dataset for both models, and received the following results:

```
for season in test_seasons:
    print(f"\nResults for {season} Season:")

    xgb_output = test_set[(test_set['season'] == str(season))][['name',
'XGB_Predicted_Champion','Champion_Win_Share']].sort_values(by='XGB_Predic
ted_Champion', ascending=False).head(8)
    print("XGBoost Predictions:")
    print(xgb_output)

    rf_output = test_set[(test_set['season'] == str(season))][['name',
'RF_Predicted_Champion','Champion_Win_Share']].sort_values(by='RF_Predicte
d_Champion', ascending=False).head(8)
    print("\nRandom Forest Predictions:")
    print(rf_output)
```

```
Results for 2022 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
21  Golden State Warriors                0.932403              1.0000
3            Phoenix Suns                0.725190              0.4375
23        Milwaukee Bucks                0.653609              0.4375
26              Miami Heat                0.576241              0.6875
16      Memphis Grizzlies                0.569955              0.3750
19         Boston Celtics                0.395949              0.8750
27     Philadelphia 76ers                0.303881              0.3750
24              Utah Jazz                0.274240              0.1250

Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
21  Golden State Warriors                0.732250              1.0000
23        Milwaukee Bucks                0.642542              0.4375
3            Phoenix Suns                0.619958              0.4375
26              Miami Heat                0.516333              0.6875
16      Memphis Grizzlies                0.504500              0.3750
27     Philadelphia 76ers                0.484708              0.3750
19         Boston Celtics                0.469417              0.8750
24              Utah Jazz                0.280000              0.1250

Results for 2021 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
32       Brooklyn Nets                0.636510              0.4375
30        Phoenix Suns                0.503888              0.8750
41      Boston Celtics                0.492383              0.0625
49           Utah Jazz                0.436725              0.3750
48     Milwaukee Bucks                0.429333              1.0000
34          Miami Heat                0.410697              0.0000
37  Philadelphia 76ers                0.407649              0.4375
38      Denver Nuggets                0.396213              0.2500

Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
37    Philadelphia 76ers                0.637000              0.4375
32       Brooklyn Nets                0.581208              0.4375
30         Phoenix Suns                0.566417              0.8750
49            Utah Jazz                0.559917              0.3750
34           Miami Heat                0.541542              0.0000
48       Milwaukee Bucks                0.499833              1.0000
38        Denver Nuggets                0.456292              0.2500
31  Los Angeles Clippers                0.397750              0.6250

Results for 2014 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
249  Oklahoma City Thunder                0.883442              0.6250
253             Miami Heat                0.777698              0.8125
260  Los Angeles Clippers                0.681905              0.3750

Results for 2013 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
271             Miami Heat                0.815632              1.0000
290  Oklahoma City Thunder                0.803429              0.3125
298          Denver Nuggets                0.554232              0.1250
```

```
266   Los Angeles Clippers                0.681905                0.3750
246          Indiana Pacers               0.470598                0.6250
269            Chicago Bulls              0.289826                0.0625
266          Toronto Raptors              0.237718                0.1875
254        San Antonio Spurs              0.224729                1.0000
250   Golden State Warriors               0.203595                0.1875
```

```
Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
249  Oklahoma City Thunder               0.725958              0.6250
254      San Antonio Spurs               0.692792              1.0000
253             Miami Heat               0.691500              0.8125
260   Los Angeles Clippers               0.615167              0.3750
246          Indiana Pacers               0.474833              0.6250
269            Chicago Bulls              0.353542              0.0625
259          Houston Rockets              0.239667              0.1250
258            Brooklyn Nets              0.217167              0.3125
```

```
Results for 2006 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
488       Detroit Pistons                0.639705              0.6250
503            Miami Heat                0.624961              1.0000
484     San Antonio Spurs                0.606858              0.4375
481      Dallas Mavericks                0.529078              0.8750
480           Phoenix Suns               0.525985              0.6250
501  Cleveland Cavaliers                 0.525163              0.4375
489       New Jersey Nets                0.264907              0.3125
507         Indiana Pacers               0.263731              0.1250
```

```
Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
484     San Antonio Spurs               0.723625              0.4375
503            Miami Heat               0.626250              1.0000
488       Detroit Pistons               0.586500              0.6250
481      Dallas Mavericks               0.491250              0.8750
501  Cleveland Cavaliers                 0.470292              0.4375
480           Phoenix Suns               0.425833              0.6250
495         Chicago Bulls                0.207917              0.1250
492     Memphis Grizzlies               0.192333              0.0000
```

```
Results for 1997 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
759            Chicago Bulls              1.066725            1.000000
756          Houston Rockets              0.759611            0.600000
753         New York Knicks               0.682844            0.400000
749              Miami Heat               0.632136            0.533333
768               Utah Jazz               0.501059            0.866667
757     Seattle SuperSonics               0.443911            0.400000
748  Portland Trail Blazers               0.175713            0.066667
762       Los Angeles Lakers              0.174035            0.266667
```

```
Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
759            Chicago Bulls              0.890500            1.000000
756          Houston Rockets              0.653250            0.600000
768               Utah Jazz               0.635667            0.866667
757     Seattle SuperSonics               0.538000            0.400000
753         New York Knicks               0.479125            0.400000
749              Miami Heat               0.450958            0.533333
744          Detroit Pistons              0.275458            0.133333
748  Portland Trail Blazers               0.168583            0.066667
```

```
Results for 1988 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
997            Boston Celtics             0.842371            0.600000
1007      Los Angeles Lakers               0.637079            1.000000
998          Dallas Mavericks              0.451329            0.666667
989            Denver Nuggets              0.410720            0.333333
1001            Atlanta Hawks              0.373463            0.400000
992           Detroit Pistons              0.358115            0.933333
1010  Portland Trail Blazers               0.334198            0.066667
995       Seattle SuperSonics              0.217511            0.133333
```

```
Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
1007      Los Angeles Lakers               0.712250            1.000000
997            Boston Celtics             0.683917            0.600000
992           Detroit Pistons              0.430375            0.933333
989            Denver Nuggets              0.423125            0.333333
1001            Atlanta Hawks              0.414167            0.400000
998          Dallas Mavericks              0.385917            0.666667
995       Seattle SuperSonics              0.199250            0.133333
1010  Portland Trail Blazers               0.194667            0.066667
```

```
298          Denver Nuggets               0.554222                0.1250
297        New York Knicks                0.535287                0.3750
284           Brooklyn Nets               0.393295                0.1875
282          Boston Celtics               0.354804                0.1250
285   Los Angeles Clippers                0.354613                0.1250
278     Los Angeles Lakers                0.245232                0.0000
```

```
Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
271             Miami Heat               0.733208              1.0000
290  Oklahoma City Thunder               0.717292              0.3125
291      San Antonio Spurs               0.661292              0.9375
298          Denver Nuggets              0.564042              0.1250
297        New York Knicks               0.559750              0.3750
284           Brooklyn Nets              0.420250              0.1875
282          Boston Celtics               0.369875              0.1250
285   Los Angeles Clippers                0.280000              0.1250
```

```
Results for 2005 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
510            Miami Heat                0.729509              0.6875
512     Dallas Mavericks                 0.573328              0.3750
533   San Antonio Spurs                  0.572287              1.0000
514     Detroit Pistons                  0.561145              0.9375
528          Phoenix Suns                0.437242              0.5625
516         Chicago Bulls                0.343127              0.1250
520      Houston Rockets                 0.323569              0.1875
527       Denver Nuggets                 0.236380              0.0625
```

```
Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
510            Miami Heat               0.727042              0.6875
514       Detroit Pistons               0.594667              0.9375
533     San Antonio Spurs               0.551083              1.0000
512      Dallas Mavericks               0.504583              0.3750
528           Phoenix Suns              0.485750              0.5625
516         Chicago Bulls                0.433583              0.1250
520      Houston Rockets                 0.213708              0.1875
521   Seattle SuperSonics                0.192792              0.3750
```

```
Results for 1992 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
897            Chicago Bulls              1.015030            1.000000
891          Boston Celtics               0.546900            0.400000
902  Portland Trail Blazers               0.446584            0.866667
885      Cleveland Cavaliers              0.406120            0.600000
898    Golden State Warriors              0.391522            0.066667
907               Utah Jazz               0.348364            0.600000
882          Indiana Pacers               0.256973            0.000000
889            Phoenix Suns               0.201952            0.266667
```

```
Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
897            Chicago Bulls              0.866208            1.000000
891          Boston Celtics               0.680583            0.400000
902  Portland Trail Blazers               0.549750            0.866667
907               Utah Jazz               0.486292            0.600000
898    Golden State Warriors              0.433833            0.066667
885      Cleveland Cavaliers              0.331000            0.600000
882          Indiana Pacers               0.218875            0.000000
889            Phoenix Suns               0.214500            0.266667
```

```
Results for 1985 Season:
XGBoost Predictions:
                    name  XGB_Predicted_Champion  Champion_Win_Share
1073           Boston Celtics             0.833634            0.866667
1060      Los Angeles Lakers               0.601457            1.000000
1072      Philadelphia 76ers               0.573557            0.533333
1075         Milwaukee Bucks               0.555150            0.200000
1071         Houston Rockets               0.448750            0.133333
1069           Denver Nuggets              0.264647            0.533333
1076  Portland Trail Blazers               0.183150            0.266667
1064         New Jersey Nets               0.169489            0.000000
```

```
Random Forest Predictions:
                    name  RF_Predicted_Champion  Champion_Win_Share
1060      Los Angeles Lakers               0.799750            1.000000
1073           Boston Celtics             0.771625            0.866667
1075         Milwaukee Bucks               0.499000            0.200000
1071         Houston Rockets               0.461542            0.133333
1072      Philadelphia 76ers               0.445292            0.533333
1069           Denver Nuggets              0.412042            0.533333
1076  Portland Trail Blazers               0.252167            0.266667
1074        San Antonio Spurs              0.245167            0.133333
```

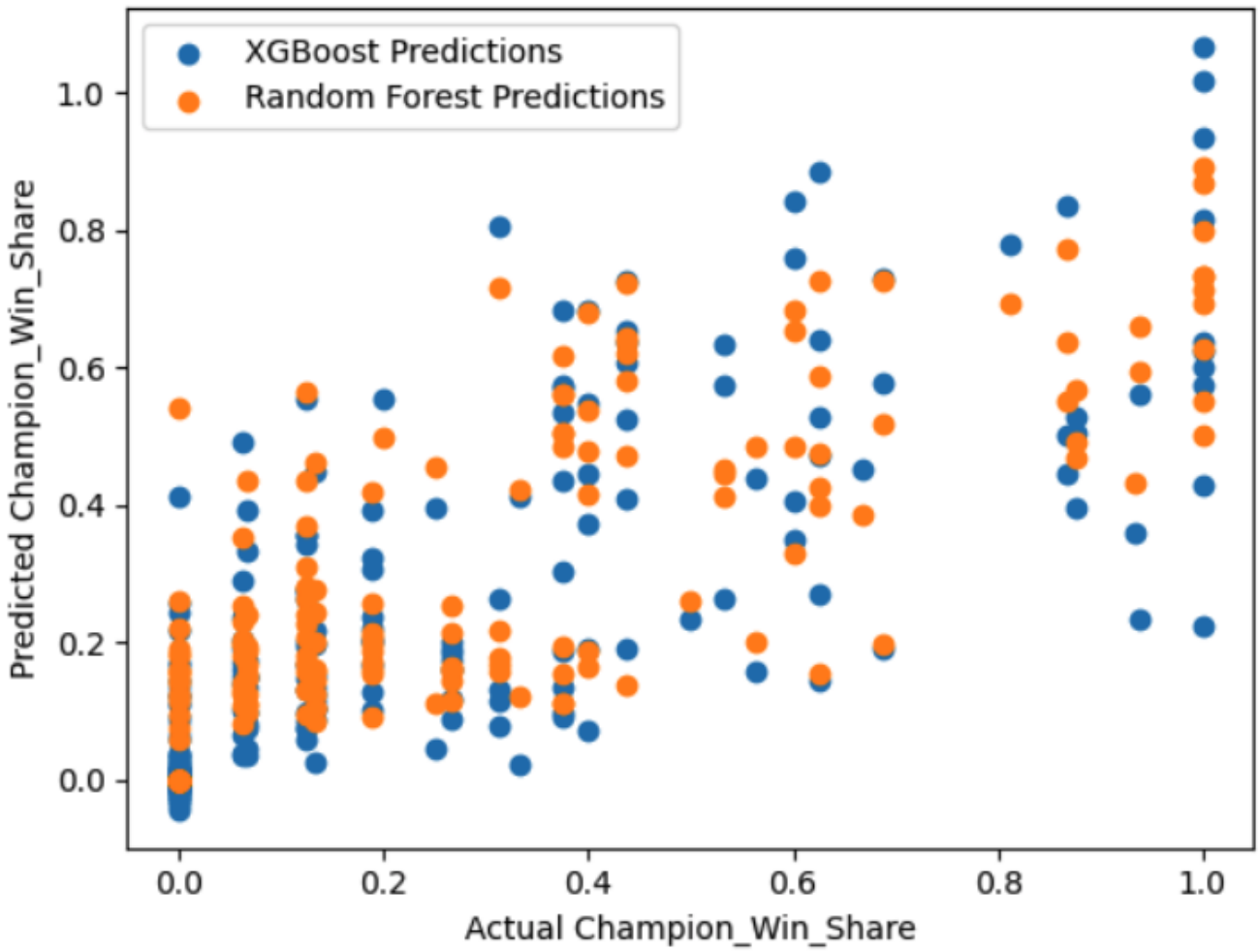
Although both the Gradient Boosting model and the Random Forest model yielded relatively similar results, Random Forest was found to be slightly more accurate.

Seeding Offset	XGBoost	Random Forest
0 (Correct)	4	6

Seeding Offset	XGBoost	Random Forest
1	3	2
2	1	1
> 3	2	1

Below is a scatterplot of XGBoost and Random Forest predictions compared to the actual `Champion_Win_Share`.

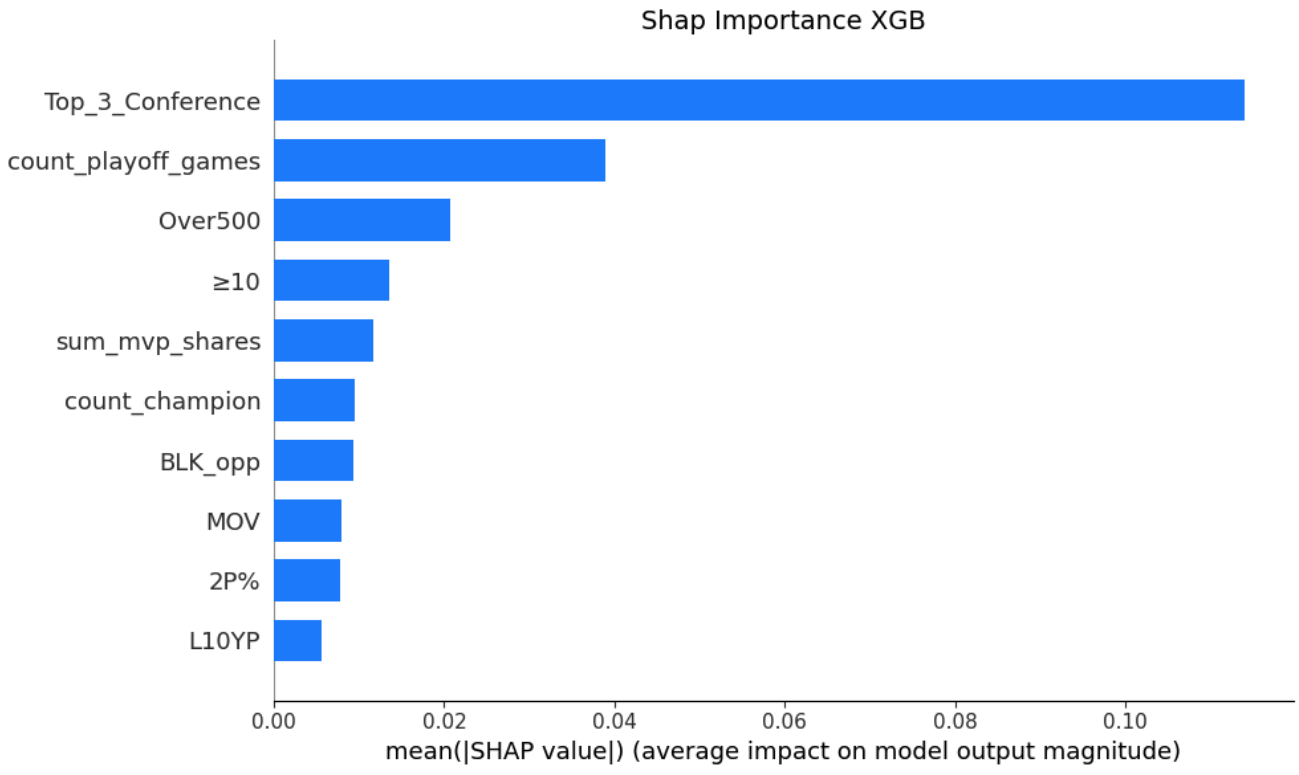
```
# Visualization
plt.scatter(y_test, xgb_predictions, label="XGBoost Predictions")
plt.scatter(y_test, rf_predictions, label="Random Forest Predictions")
plt.xlabel("Actual Champion_Win_Share")
plt.ylabel("Predicted Champion_Win_Share")
plt.legend()
plt.show()
```

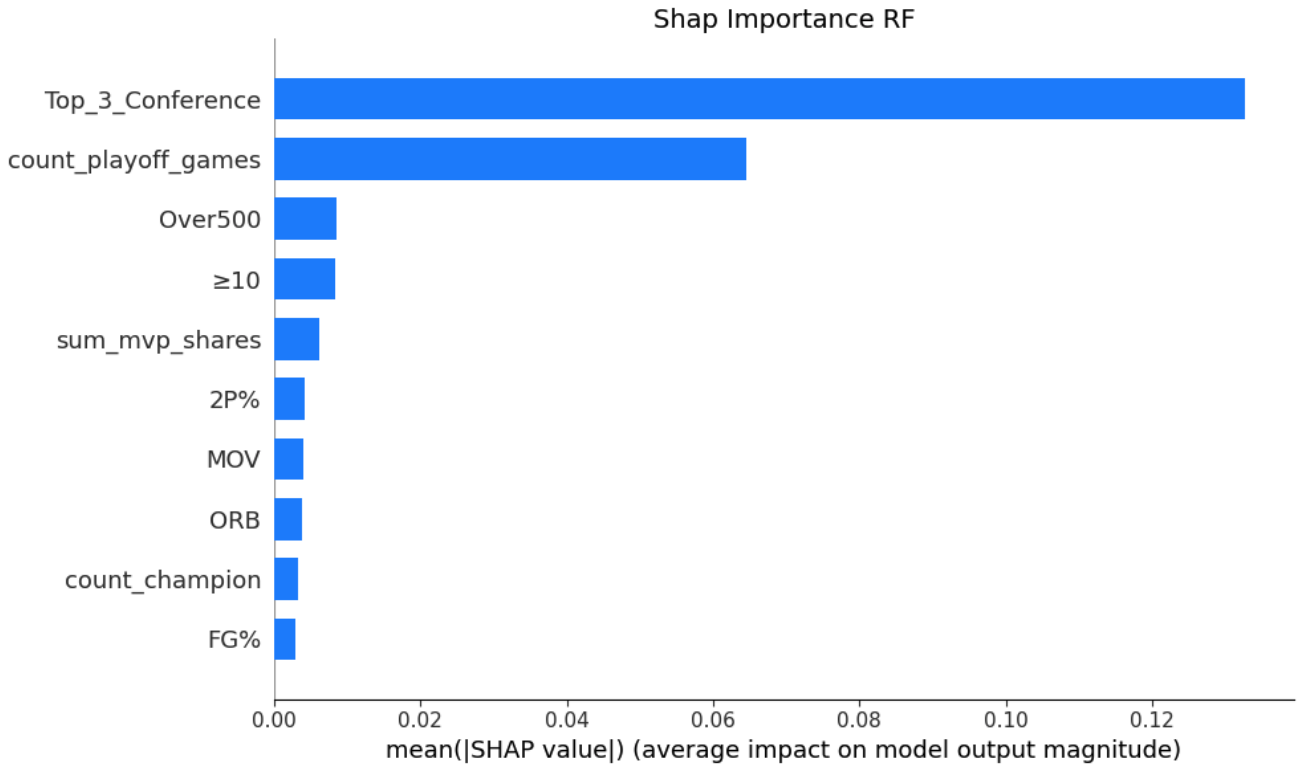


iii. Shap Importance

According to the creators of SHAP, "SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model." Using the SHAP library, we analyzed our results to see which features were the most important in determining `Champion_Win_Share`.

```
# Explain XGBoost predictions with SHAP
explainer_xgb = shap.Explainer(xgb_model)
shap_values_xgb = explainer_xgb.shap_values(X_test)

# Explain Random Forest predictions with SHAP
explainer_rf = shap.Explainer(rf_model)
shap_values_rf = explainer_rf.shap_values(X_test)

# Summary plot for XGBoost
shap.summary_plot(shap_values_xgb, X_test, feature_names=feature_columns,
plot_type="bar", show=False, max_display=10)

fig = plt.gcf()
fig.set_size_inches(10, 6)
plt.title("Shap Importance XGB", fontsize=14, loc="center")
plt.tight_layout()
plt.show()

# Summary plot for Random Forest
shap.summary_plot(shap_values_rf, X_test, feature_names=feature_columns,
plot_type="bar", show=False, max_display=10)

fig = plt.gcf()
fig.set_size_inches(10, 6)
plt.title("Shap Importance RF", fontsize=14, loc="center")
plt.tight_layout()
plt.show()
```





As shown in the graphs above, the Gradient Boosting and Random Forest algorithm determined a similar set of features to be the most important in determining whether or not an NBA team wins a championship. The five most important features of both models are not only identical but also in the same order. Furthermore, the two models also share three other features within the top 10 most impactful features.

Feature	Explanation
Top_3_Conference	True/False value on whether the team finished within the top 3 seeds in the regular season.
count_playoff_games	The total number of playoff games played by the players on the team.
Over500	Winrate against teams with a positive winrate
>= 10	Winrate with score differentials over 10 points
sum_mvp_shares	Indication of whether the team has a super star player
count_champion	Number of NBA champion players on the team
2P%	Two point percentage
MOV	Margin of victory

V. Related Work

ML Libraries/Tools:

- XGBoost (Gradient Boost)
- scikit-learn (Random Forest)
- SHAP (ML Model Analysis)

VI. Conclusion

Although