# Mid Term Project Report

Group 3 - Alfred, Grace, Hiroaki

Predicticion of  NBA game outcomes using machine learning.

## Summary

In this project, we read the paper which predicts game winners and losers of the NBA, and reproduce methods introduced by the paper. After that, we try other machine learning algorithms and update model parameters to improve accuracy.

## Motivation

Predicting NBA game outcomes is challenging because of the complex relationships between player performance and team stats. This study aims to improve predictions by using machine learning models like Logistic Regression, SVM, and Random Forest. By comparing these models, we hope to find the most important factors, such as field goal percentage, that influence game results. This project will help many people, including coaches, players, sports analysts, and sports bettors.

## Research paper details

We chose this paper. The goal of this paper is to use various variables of NBA games to predict the winner of the game. The data is the all NBA games from the 2004 season to December 2020. There are a total of 25797 games.

### Feature Selection

The author used 5 type features for the home and away teams, so the number of all features are 10.

- FG-PCT -> field goal percentage
- FT-PCT -> free throw percentage
- FG3-PCT -> three-point field goal percentage
- AST -> assists
- REB -> rebound

## Model Performance

The author evaluated 4 machine learning algorithms(Logistic regression, SVM, LSTM, Random forest) with various metrics(Accuracy, Precision, Recall, F1 Score, and AUC value).

**Table 3: the Accuracy, Precision, Recall, F1 Score, and AUC value. The values of Accuracy, Precision, Recall, and F1 Score are based on the value of TP, TN, FP, and FN. The value of AUC is based on ROC curve.**

| Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| LR | 0.8164 | 0.8251 | 0.8671 | 0.8455 | 0.90 |
| SVM | 0.7481 | 0.76 | 0.8288 | 0.7929 | 0.82 |
| LSTM | 0.7531 | 0.7502 | 0.8631 | 0.8027 | 0.83 |
| RF | 0.8378 | 0.8501 | 0.8758 | 0.8627 | 0.92 |

ROC curve of each model is as follows.

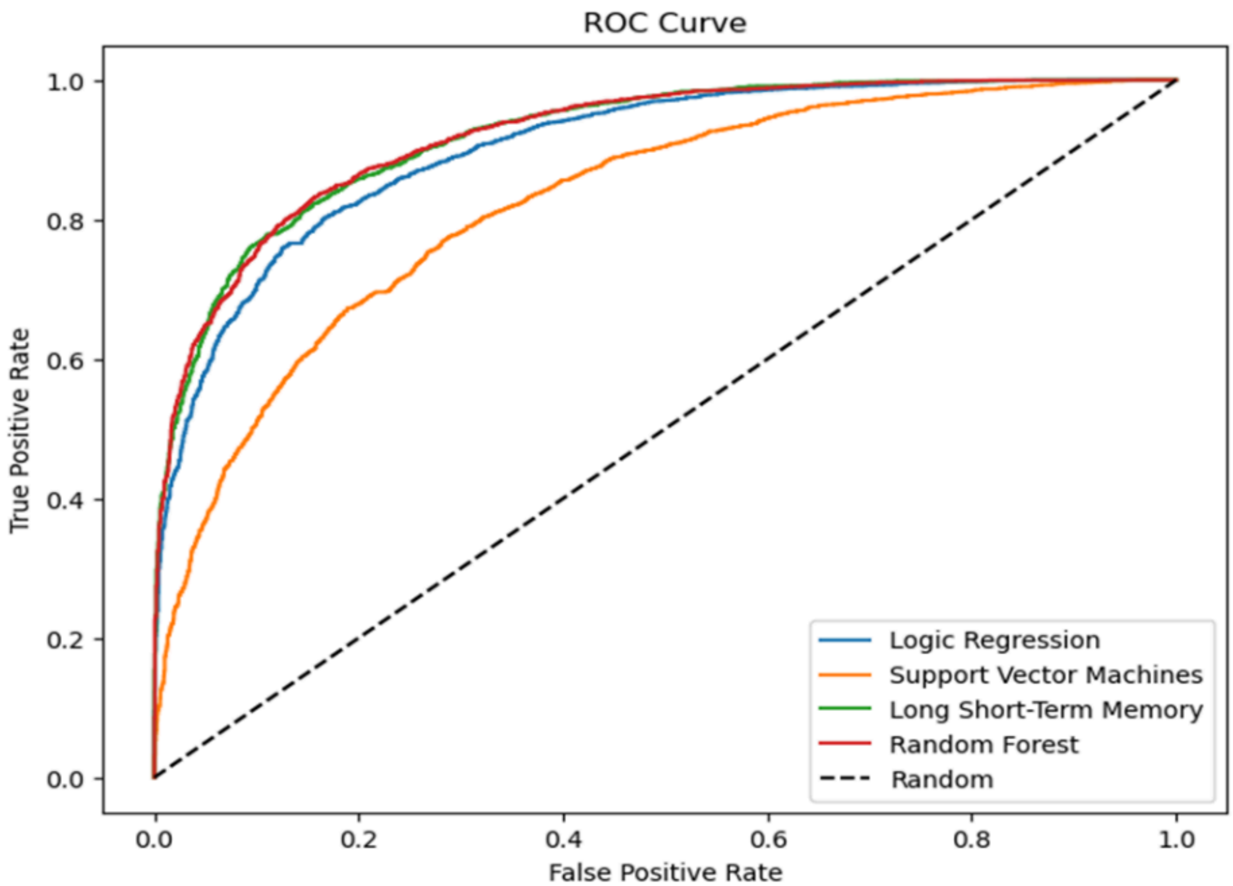



**Figure 6: the ROC curves for the four models**

# Feature Importance

The feature importance chart highlights that **field goal percentage (FG_PCT)** is the most influential factor in predicting game outcomes. Among all features, **FG_PCT_home** and **FG_PCT_away** have the highest importance scores, suggesting that a team's shooting accuracy plays a critical role in determining the final result. This finding emphasizes that **a team's ability to convert field goal attempts is a stronger predictor of success than other statistics such as rebounds, assists, or free throw percentage.**

.

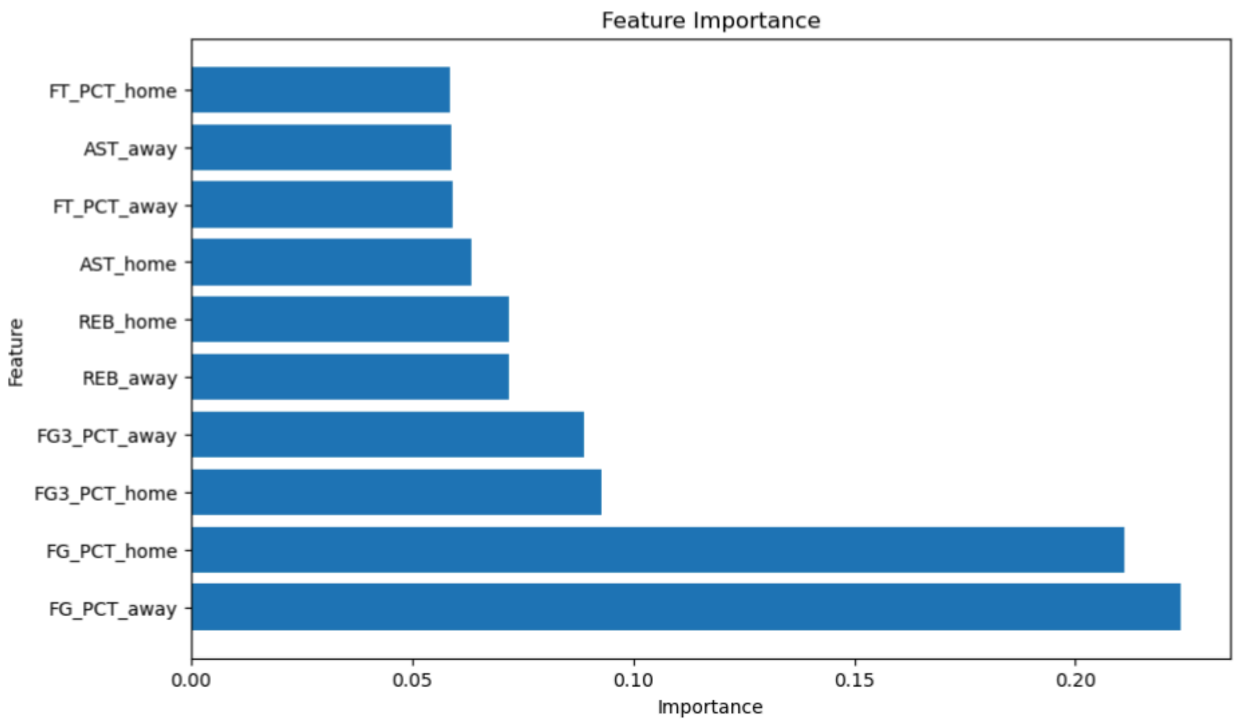



**Figure 8: the importance of each features**

# Dataset details

We used [this dataset] to predict the result of the NBA. This dataset includes all NBA games from the 2004 season to Dec 2020.

We have 4 csv files. Explanation of each file is as follows. If you want to know the details of the dataset, please check this page.

All the detailed information on the nomenclature of the different Datasets is described below. Each element of the table is detailed so that it is easier for the reader to understand, and can easily identify the reference and definition of each element of the datasets.

The study applied several steps to clean and filter the data before modeling:

*Data Collection: NBA game data from the 2004 season through December 2020 was obtained, including player and team statistics and game details.

Dataset Structure
The dataset includes the following variables:

1. Input Variables (Features)
These are the factors used to train the prediction models:

FG-PCT (Field Goal Percentage) → Field goal percentage.

FT-PCT (Free Throw Percentage) → Free throw percentage.

FG3-PCT (Three-Point Field Goal Percentage) → Three-point goal percentage.

AST (Assists) → Number of assists made by the team.

REB (Rebounds) → Number of rebounds obtained.

HOME TEAM → Whether the team is playing at home or away.

Each of these features is collected for both the home and visiting teams.

2. Output Variable (Target)
This is the variable that want to predict:

HOME-TEAM-WINS (0 or 1)
1 → If the home team wins.

0 → If the visiting team wins.

## games.csv

All games from the 2004 season to the last update with the date, teams and some details like number of points, etc.

This dataset provides detailed statistics for basketball games, including performance metrics for both home and away teams, and the outcome of each game.

The dataset contains the following columns:

GAME_DATE_EST: The estimated date of the game.

GAME_ID: A unique identifier for the game.

GAME_STATUS_TEXT: The status of the game (e.g., "Final").

HOME_TEAM_ID: The ID of the home team.

VISITOR_TEAM_ID: The ID of the visiting team.

SEASON: The season in which the game was played.

TEAM_ID_home: The ID of the home team (repeated for convenience).

PTS_home: Points scored by the home team.

FG_PCT_home: Field goal percentage for the home team.

FT_PCT_home: Free throw percentage for the home team.

FG3_PCT_home: Three-point field goal percentage for the home team.

AST_home: Assists by the home team.

REB_home: Rebounds by the home team.

TEAM_ID_away: The ID of the away team.

PTS_away: Points scored by the away team.

FG_PCT_away: Field goal percentage for the away team.

FT_PCT_away: Free throw percentage for the away team.

FG3_PCT_away: Three-point field goal percentage for the away team.

AST_away: Assists by the away team.

REB_away: Rebounds by the away team.

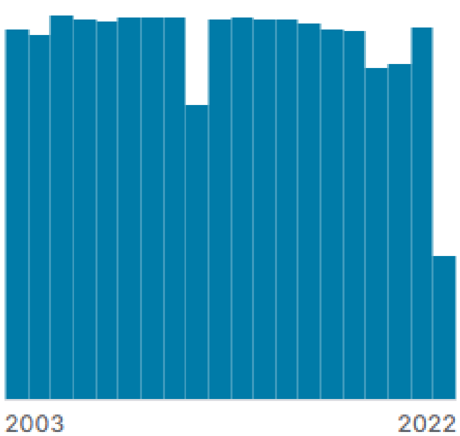
HOME_TEAM_WINS: A binary indicator (1 or 0) showing whether the home team won (1) or lost (0).

| | GAME_DATE_EST | GAME_ID | GAME_STATUS_TEXT | HOME_TEAM_ID | VISITOR_TEAM_ID | SEASON | TEAM_ID_home | PTS_home | FG_PCT_home | FT_PCT_home |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2022-12-22 | 22200477 | Final | 1610612740 | 1610612759 | 2022 | 1610612740 | 126.0 | 0.484 | 0.926 |
| 1 | 2022-12-22 | 22200478 | Final | 1610612762 | 1610612764 | 2022 | 1610612762 | 120.0 | 0.488 | 0.952 |
| 2 | 2022-12-21 | 22200466 | Final | 1610612739 | 1610612749 | 2022 | 1610612739 | 114.0 | 0.482 | 0.786 |
| 3 | 2022-12-21 | 22200467 | Final | 1610612755 | 1610612765 | 2022 | 1610612755 | 113.0 | 0.441 | 0.909 |

| AST_home | REB_home | TEAM_ID_away | PTS_away | FG_PCT_away | FT_PCT_away | FG3_PCT_away | AST_away | REB_away | HOME_TEAM_WINS |
|---|---|---|---|---|---|---|---|---|---|
| 25.0 | 46.0 | 1610612759 | 117.0 | 0.478 | 0.815 | 0.321 | 23.0 | 44.0 | 1 |
| 16.0 | 40.0 | 1610612764 | 112.0 | 0.561 | 0.765 | 0.333 | 20.0 | 37.0 | 1 |
| 22.0 | 37.0 | 1610612749 | 106.0 | 0.470 | 0.682 | 0.433 | 20.0 | 46.0 | 1 |

# SEASON

Season when the game occured




| | | |
|---|---|---|
| Valid | 26.7k | 100% |
| Mismatched | 0 | 0% |
| Missing | 0 | 0% |
| Mean | 2.01k | |
| Std. Deviation | 5.59 | |
| Quantiles | 2003 | Min |
| | 2007 | 25% |
| | 2012 | 50% |
| | 2017 | 75% |
| | 2022 | Max |

## games_details.csv

Details of games dataset, all statistics of players for a given game

This dataset provides detailed statistics for each player in various games, including their performance metrics such as points, rebounds, assists, and more.

The dataset contains the following columns:

GAME_ID: Unique identifier for the game

TEAM_ID: Unique identifier for the team

TEAM_ABBREVIATION: Abbreviated team name

TEAM_CITY: City of the team

PLAYER_ID: Unique identifier for the player

PLAYER_NAME: Name of the player

NICKNAME: Player's nickname

START_POSITION: Player's starting position in the game

COMMENT: Additional comments about the player's performance

MIN: Minutes played

FGM: Field goals made

FGA: Field goals attempted

FG_PCT: Field goal percentage

FG3M: Three-point field goals made

FG3A: Three-point field goals attempted

FG3_PCT: Three-point field goal percentage

FTM: Free throws made

FTA: Free throws attempted

FT_PCT: Free throw percentage

OREB: Offensive rebounds

DREB: Defensive rebounds

REB: Total rebounds

AST: Assists

STL: Steals

BLK: Blocks

TO: Turnovers

PF: Personal fouls

PTS: Points scored

PLUS_MINUS: Player's impact on the game score while on the court

| | GAME_ID | TEAM_ID | TEAM_ABBREVIATION | TEAM_CITY | PLAYER_ID | PLAYER_NAME | NICKNAME | START_POSITION | COMMENT | MIN | ... | OREB | DREB | REB | AST | STL | BLK | TO | PF | PTS | PLUS_MINUS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 22200477 | 1610612759 | SAS | San Antonio | 1629641 | Romeo Langford | Romeo | F | NaN | 18:06 | ... | 1.0 | 1.0 | 2.0 | 0.0 | 1.0 | 0.0 | 2.0 | 5.0 | 2.0 | -2.0 |
| 1 | 22200477 | 1610612759 | SAS | San Antonio | 1631110 | Jeremy Sochan | Jeremy | F | NaN | 31:01 | ... | 6.0 | 3.0 | 9.0 | 6.0 | 1.0 | 0.0 | 2.0 | 1.0 | 23.0 | -14.0 |
| 2 | 22200477 | 1610612759 | SAS | San Antonio | 1627751 | Jakob Poeltl | Jakob | C | NaN | 21:42 | ... | 1.0 | 3.0 | 4.0 | 1.0 | 1.0 | 0.0 | 2.0 | 4.0 | 13.0 | -4.0 |
| 3 | 22200477 | 1610612759 | SAS | San Antonio | 1630170 | Devin Vassell | Devin | G | NaN | 30:20 | ... | 0.0 | 9.0 | 9.0 | 5.0 | 3.0 | 0.0 | 2.0 | 1.0 | 10.0 | -18.0 |

### TEAM_CITY

City where the game was played

| | |
|---|---|
| Los Angeles | 5% |
| Miami | 4% |
| Other (607917) | 91% |

| | | |
|---|---|---|
| Valid | 669k | 100% |
| Mismatched | 0 | 0% |
| Missing | 0 | 0% |
| Unique | 33 | |
| Most Common | Los Angeles | 5% |

## players.csv

Players details (name)

This dataset contains information about basketball players, season,etc..

The dataset contains the following columns:

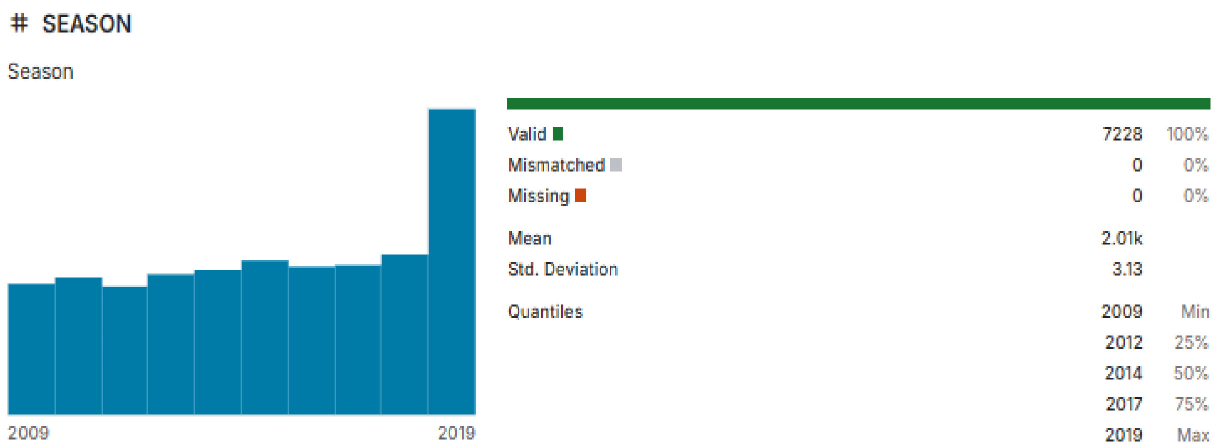
PLAYER_NAME: The name of the player.

TEAM_ID: The ID of the team the player is associated with.

PLAYER_ID: The unique ID of the player.

SEASON: The season year.

| | PLAYER_NAME | TEAM_ID | PLAYER_ID | SEASON |
|---|---|---|---|---|
| 0 | Royce O'Neale | 1610612762 | 1626220 | 2019 |
| 1 | Bojan Bogdanovic | 1610612762 | 202711 | 2019 |
| 2 | Rudy Gobert | 1610612762 | 203497 | 2019 |
| 3 | Donovan Mitchell | 1610612762 | 1628378 | 2019 |




## ranking.csv

Ranking of NBA given a day (split into west and east on CONFERENCE column)

The dataset provides a detailed record of the standings for teams in the Western Conference of the NBA during the 2022-2023 season, updated on various dates. It includes information on the number of games played, wins, losses, winning percentage, and home/road records for each team.

The dataset contains the following columns:

TEAM_ID: Unique identifier for the team

LEAGUE_ID: Identifier for the league

SEASON_ID: The season in which the data was recorded

STANDINGSDATE: Date of the standings

CONFERENCE: The conference the team belongs to

TEAM: Name of the team

G: Games played

W: Wins

L: Losses

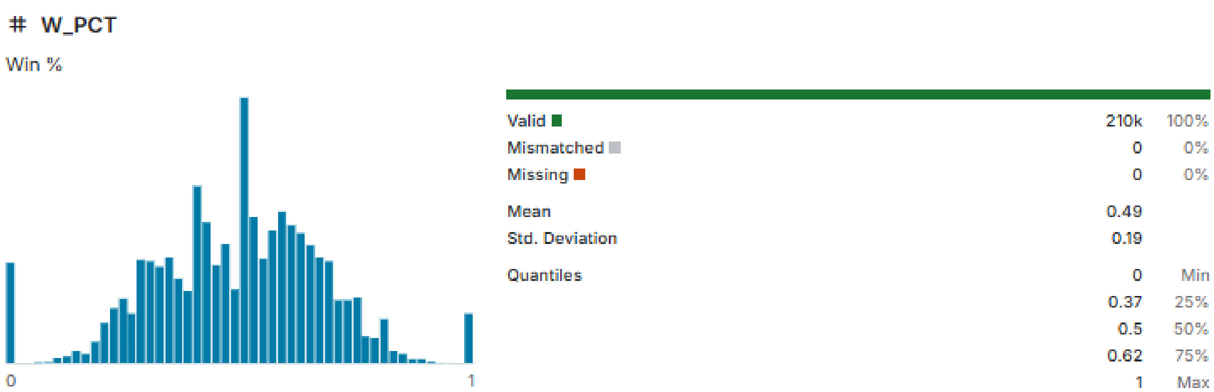
W_PCT: Win percentage

HOME_RECORD: Win-loss record for home games

ROAD_RECORD: Win-loss record for away games

RETURNTOPLAY: Indicates if the team returned to play after a break

| | TEAM_ID | LEAGUE_ID | SEASON_ID | STANDINGSDATE | CONFERENCE | TEAM | G | W | L | W_PCT | HOME_RECORD | ROAD_RECORD | RETURNTOPLAY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1610612743 | 0 | 22022 | 2022-12-22 | West | Denver | 30 | 19 | 11 | 0.633 | 10-3 | 9-8 | NaN |
| 1 | 1610612763 | 0 | 22022 | 2022-12-22 | West | Memphis | 30 | 19 | 11 | 0.633 | 13-2 | 6-9 | NaN |
| 2 | 1610612740 | 0 | 22022 | 2022-12-22 | West | New Orleans | 31 | 19 | 12 | 0.613 | 13-4 | 6-8 | NaN |
| 3 | 1610612756 | 0 | 22022 | 2022-12-22 | West | Phoenix | 32 | 19 | 13 | 0.594 | 14-4 | 5-9 | NaN |




## teams.csv

All teams of NBA

This dataset provides detailed information about NBA teams, including their history, location, management, and affiliations.

LEAGUE_ID: The ID of the league (e.g., 00 for the NBA).

TEAM_ID: The unique ID for each team.

MIN_YEAR: The first year the team was active.

MAX_YEAR: The last year the team was active (e.g., 2019 in this dataset).

ABBREVIATION: The abbreviation of the team's name (e.g., ATL for Atlanta Hawks).

NICKNAME: The nickname or the main name of the team (e.g., Hawks, Celtics).

YEARFOUNDED: The year the team was founded.

CITY: The city where the team is based.

ARENA: The name of the arena where the team plays its home games.

ARENACAPACITY: The seating capacity of the arena (if available).

OWNER: The owner(s) of the team.

GENERALMANAGER: The general manager of the team.

HEADCOACH: The head coach of the team.

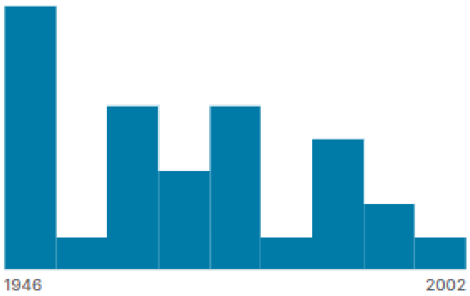
DLEAGUEAFFILIATION: The affiliated G League (formerly D-League) team, if any.

| | LEAGUE_ID | TEAM_ID | MIN_YEAR | MAX_YEAR | ABBREVIATION | NICKNAME | YEARFOUNDED | CITY | ARENA | ARENACAPACITY |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1610612737 | 1949 | 2019 | ATL | Hawks | 1949 | Atlanta | State Farm Arena | 18729.0 |
| 1 | 0 | 1610612738 | 1946 | 2019 | BOS | Celtics | 1946 | Boston | TD Garden | 18624.0 |
| 2 | 0 | 1610612740 | 2002 | 2019 | NOP | Pelicans | 2002 | New Orleans | Smoothie King Center | NaN |
| 3 | 0 | 1610612741 | 1966 | 2019 | CHI | Bulls | 1966 | Chicago | United Center | 21711.0 |

| OWNER | GENERALMANAGER | HEADCOACH | DLEAGUEAFFILIATION |
|---|---|---|---|
| Tony Ressler | Travis Schlenk | Lloyd Pierce | Erie Bayhawks |
| Wyc Grousbeck | Danny Ainge | Brad Stevens | Maine Red Claws |
| Tom Benson | Trajan Langdon | Alvin Gentry | No Affiliate |

## # YEARFOUNDED

Founded Year




| | | |
|---|---|---|
| Valid | 30 | 100% |
| Mismatched | 0 | 0% |
| Missing | 0 | 0% |
| Mean | 1.97k | |
| Std. Deviation | 16.4 | |
| Quantiles | 1946 | Min |
| | 1949 | 25% |
| | 1970 | 50% |
| | 1980 | 75% |
| | 2002 | Max |

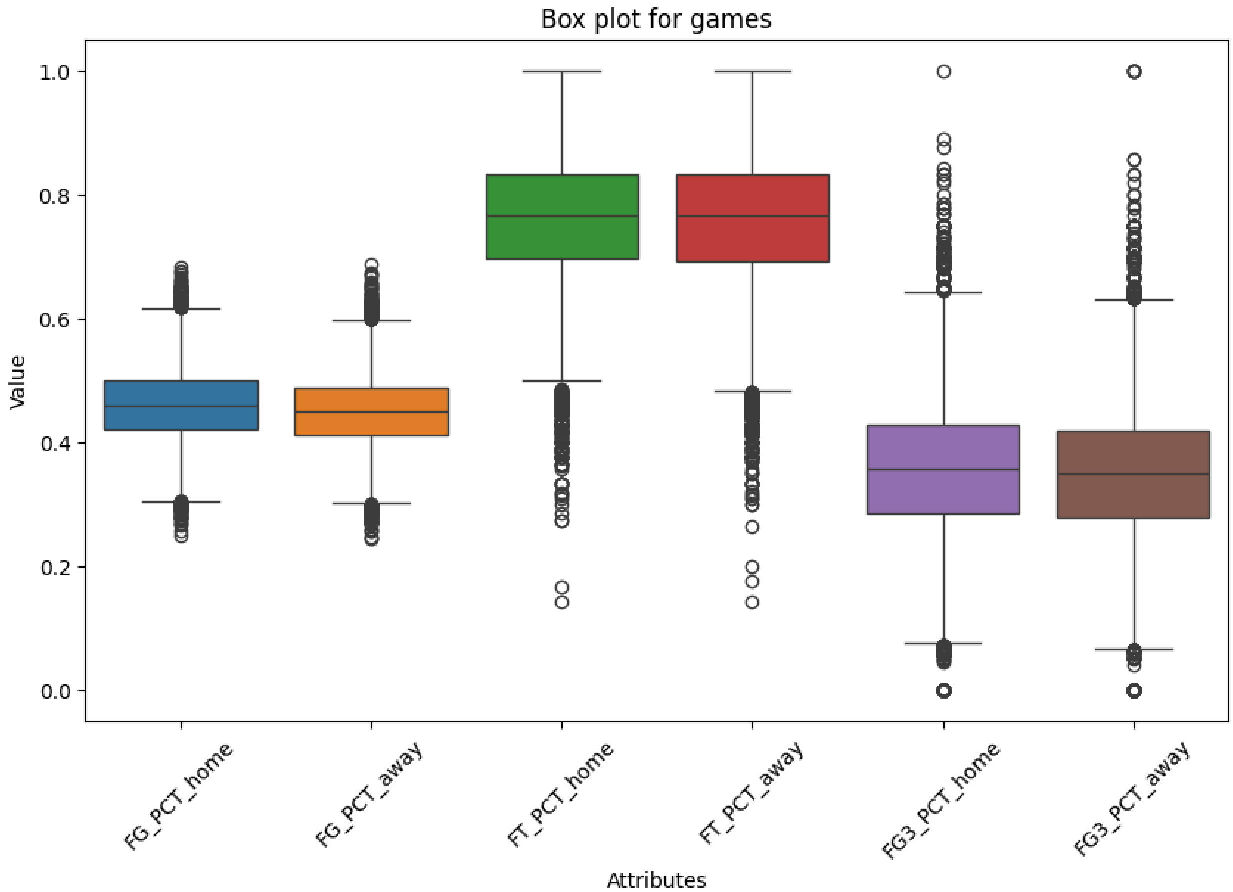
# Data preprocessing and feature engineering

Rows including missing values are empty about features that we want to use. Those rows are totally useless, so we dropped them. We used RobustScaler of scikit-learn to scale data in a way that is resistant to outliers.

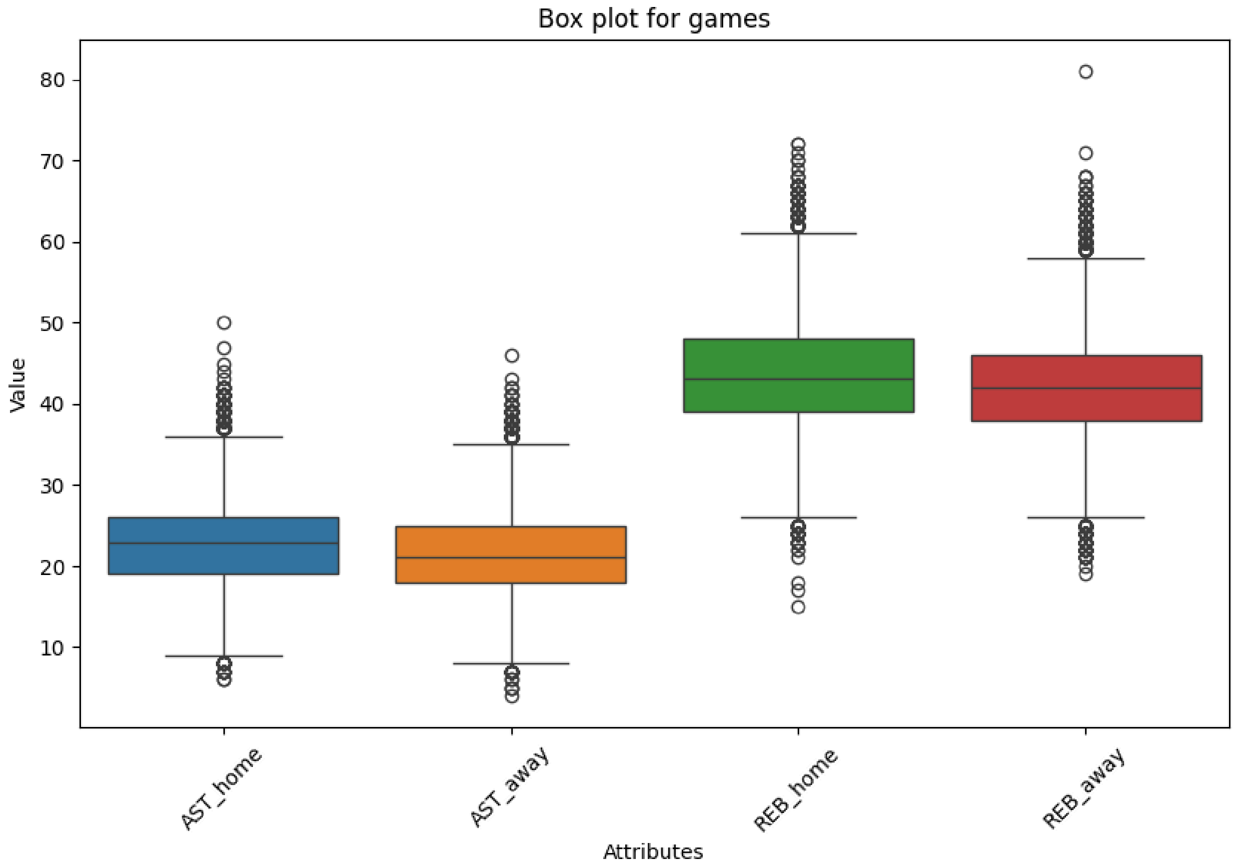
```
games.isnull().sum()
✓ 0.0s

GAME_DATE_EST        0
GAME_ID              0
GAME_STATUS_TEXT     0
HOME_TEAM_ID         0
VISITOR_TEAM_ID      0
SEASON               0
TEAM_ID_home         0
PTS_home            99
FG_PCT_home         99
FT_PCT_home         99
FG3_PCT_home        99
AST_home            99
REB_home            99
TEAM_ID_away         0
PTS_away            99
FG_PCT_away         99
FT_PCT_away         99
FG3_PCT_away        99
AST_away            99
REB_away            99
HOME_TEAM_WINS       0
dtype: int64
```

```
       SEASON  TEAM_ID_home  PTS_home  FG_PCT_home  FT_PCT_home  ...  \
19175    2003    1610612753       NaN          NaN          NaN  ...
19176    2003    1610612737       NaN          NaN          NaN  ...
19177    2003    1610612738       NaN          NaN          NaN  ...
19178    2003    1610612759       NaN          NaN          NaN  ...
19179    2003    1610612749       NaN          NaN          NaN  ...
...       ...           ...       ...          ...          ...  ...
19269    2003    1610612743       NaN          NaN          NaN  ...
19270    2003    1610612757       NaN          NaN          NaN  ...
19271    2003    1610612759       NaN          NaN          NaN  ...
19278    2003    1610612747       NaN          NaN          NaN  ...
19279    2003    1610612747       NaN          NaN          NaN  ...
```

This is a box plot about the percentage of each stat. Home team’s stats are better than away team's.



This is a box plot about the points of each stat. Home team’s stats are better than the away team's.

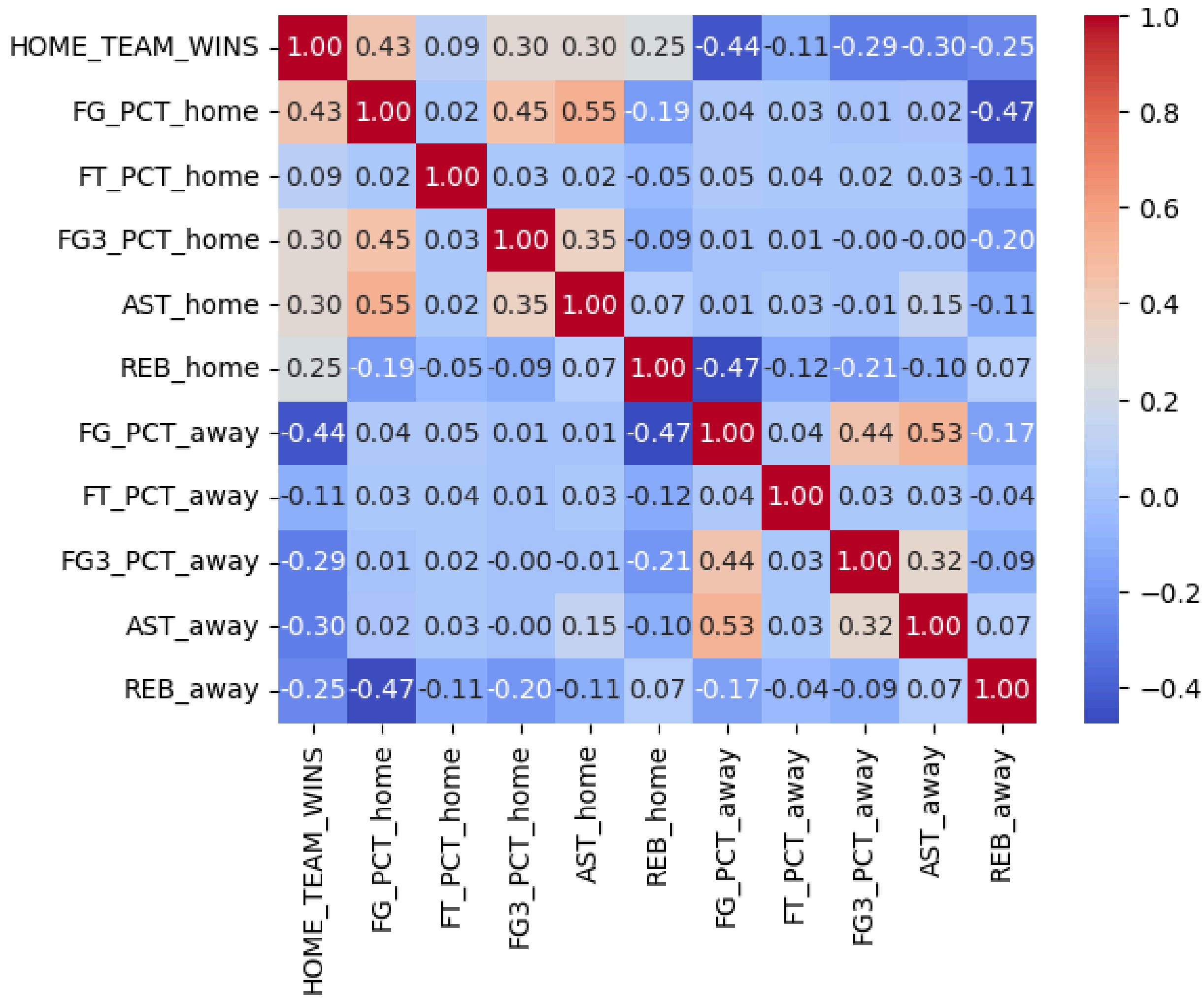


The following image shows static data of each feature. As you can see, home team’s stats are better than the away team's.

| | FG_PCT_home | FT_PCT_home | FG3_PCT_home | AST_home | REB_home |
|---|---|---|---|---|---|
| count | 26552.000000 | 26552.000000 | 26552.000000 | 26552.000000 | 26552.000000 |
| mean | 0.460735 | 0.760377 | 0.356023 | 22.823441 | 43.374284 |
| std | 0.056676 | 0.100677 | 0.111164 | 5.193308 | 6.625769 |
| min | 0.250000 | 0.143000 | 0.000000 | 6.000000 | 15.000000 |
| 25% | 0.422000 | 0.697000 | 0.286000 | 19.000000 | 39.000000 |
| 50% | 0.460000 | 0.765000 | 0.357000 | 23.000000 | 43.000000 |
| 75% | 0.500000 | 0.833000 | 0.429000 | 26.000000 | 48.000000 |
| max | 0.684000 | 1.000000 | 1.000000 | 50.000000 | 72.000000 |

| | FG_PCT_away | FT_PCT_away | FG3_PCT_away | AST_away | REB_away |
|---|---|---|---|---|---|
| count | 26552.000000 | 26552.000000 | 26552.000000 | 26552.000000 | 26552.000000 |
| mean | 0.449732 | 0.758816 | 0.349489 | 21.496271 | 42.113249 |
| std | 0.055551 | 0.103429 | 0.109441 | 5.160596 | 6.533039 |
| min | 0.244000 | 0.143000 | 0.000000 | 4.000000 | 19.000000 |
| 25% | 0.412000 | 0.692000 | 0.278000 | 18.000000 | 38.000000 |
| 50% | 0.449000 | 0.765000 | 0.350000 | 21.000000 | 42.000000 |
| 75% | 0.487000 | 0.833000 | 0.419000 | 25.000000 | 46.000000 |
| max | 0.687000 | 1.000000 | 1.000000 | 46.000000 | 81.000000 |

The following image shows correlation among each feature. As you can see, HOME_TEAMS_WIN has strong correlation among FG_PCT, FG3_PCT, AST and REB.



# Steps reproduced from the paper

1. Get NBA data from Kaggle
   We get data from kaggle.
2. Extract 10 features from csv data
   The paper used 10 features for training, so we extract those features from csv data.
3. Clean dataset
   Target features include missing values, so we clean the dataset by removing values.
4. Train model
   The paper used machine learning models like SVM, Random forest and Logistic regression. Therefore, we train the same models.
5. Evaluate model
   The paper used accuracy, precision, recall, F1 score, and AUC value for evaluation. Therefore, we use the same evaluation metrics.

# Contributions

We increased accuracy by changing algorithms and model parameters compared to the author's method. Roles of team members are as follows.

- Alfred
    - Data cleaning
    - Data exploration
- Grace
    - Data cleaning
    - Data exploration
    - Model training
    - Evaluation

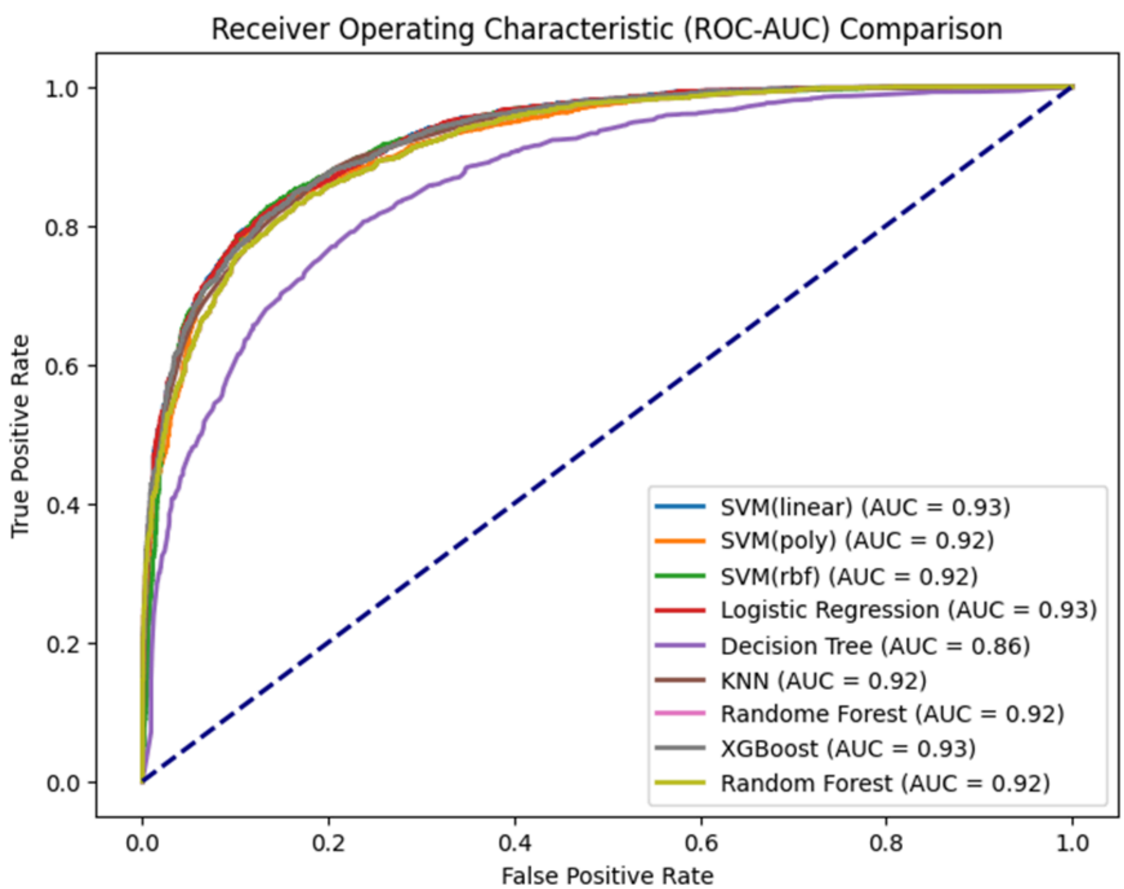
- Hiro
    - Data cleaning
    - Data exploration
    - Model training
    - Evaluation

# Significant improvements

## Try another machine learning algorithms

We tried other machine learning algorithms that the author of the paper didn't use. We tried KNN, Decision tree and XGBoost. Comparison between author's algorithms and our algorithms is as follows.

| Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| LR | 0.84 | 0.78 | 0.84 | 0.81 | 0.93 |
| SVM(linear) | 0.84 | 0.81 | 0.79 | 0.80 | 0.93 |
| RF | 0.88 | 0.86 | 0.83 | 0.85 | 0.92 |
| KNN(k=37) | 0.84 | 0.83 | 0.79 | 0.81 | 0.92 |
| DT | 0.84 | 0.81 | 0.80 | 0.80 | 0.86 |
| XGB | 0.87 | 0.84 | 0.83 | 0.84 | 0.93 |



## Try different model parameters

We tried different model parameters to check whether we can increase accuracy of models.

### Logistic regression

We added the following parameters.

- C=0.1
    - Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.
- max_iter=500
    - Maximum number of iterations taken for the solvers to converge.
- solver='liblinear'
    - Algorithm to use in the optimization problem.

| Model | Accuracy of test data |
| --- | --- |
| Logistic regression | 0.8392016569384296 |
| Logistic regression with different parameters | 0.8401430992280173 |

### Random Forest

We added the following parameters.

- n_estimators=200
    - The number of trees in the forest.

| Model | Accuracy of test data |
| --- | --- |
| Random forest | 0.8361890416117492 |
| Random forest with different parameters | 0.8390133684805121 |

# Challenges

- The scores of both teams in a match are directly related to the outcome (target) of the match, so the model may rely too heavily on them. This could lead to the model memorizing (overfitting) the training data, without learning generalizable patterns.
- By removing PTS_home and PTS_away, the model is forced to focus on other features, that are not directly tied to the game outcome.

- Model performance improved after applying GridSearchCV and RandomSearchCV to Decision Tree, Random Forest, and XGBoost. While Decision Tree model still showed signs of overfitting, there was an overall improvement.
- After experimenting with the optima k value in KNN, the model performance became more balanced.
- In the initial version of the ROC curve graph, the SVM model was not displayed alongside the other models. The issue was resolved by correctly updating the dictionary function that was used to store and plot the models.
- Class Imbalance If one team wins much more often than another, the dataset could be unbalanced.
  Models can become biased and always predict the stronger team will win.
  This could be addressed with techniques such as oversampling or undersampling.

# Conclusion and Future Scope

In this study, we reproduced the methodology presented in the original paper and explored additional machine learning algorithms, including KNN, Decision Tree, and XGBoost. We also fine-tuned model parameters to enhance performance. Our optimized models consistently outperformed the approach outlined in the paper.

For future work, we aim to incorporate deep learning techniques to further improve prediction accuracy. The original paper predicts game outcomes using in-game statistics, such as field goal percentage. However, in many real-world scenarios, this data is unavailable before the game starts. To address this limitation, we plan to develop models that predict future game results using only historical performance data and team names, making the approach more applicable to real-time predictions.