

A basketball is positioned in the center of the frame, resting on a light-colored wooden floor with red court lines. The ball is slightly out of focus, creating a soft, dreamy atmosphere. The background is a warm, light beige color.

MoneyBall Watson

Team Members:

Aryan Mishra; Brian Sohn; Eli Salk; Yiyao Qu; Ziyi Liu

Table of Contents

- Goal
 - Data Acquisition
 - Data Preprocessing
 - Feature Engineering
 - Exploratory Data Analysis
 - Modeling
- 
- A basketball is positioned in the center of the slide, slightly to the left. It is a standard orange basketball with black lines. The background is a light beige color with a subtle pattern of red and yellow lines, suggesting a basketball court floor. The basketball is casting a soft shadow on the floor below it.

Goal

- Create a ML algorithm that would predict the wins and losses of all NBA teams during the regular season
- Provide insights for sports gambling/fantasy basketball



Data Acquisition

raw data: <https://www.kaggle.com/nathanlauga/nba-games>

- games.csv  (game level)
- games_detail.csv  (player level)
- players.csv
- ranking.csv
- teams.csv 

Data Preprocessing

- regular games
- 15-16 season

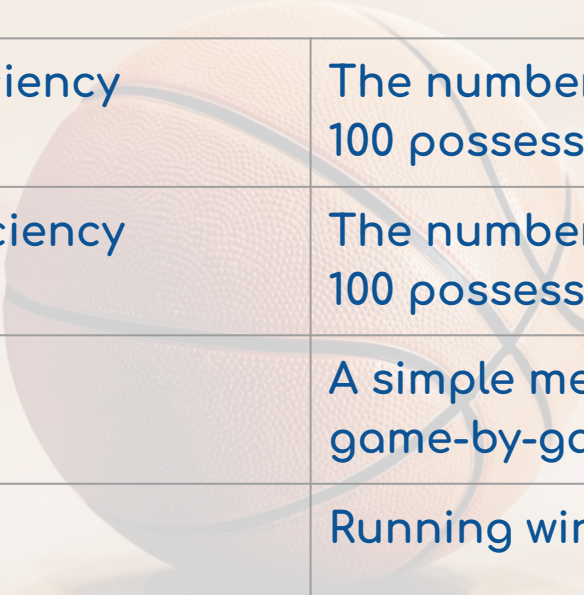
Game_id	Team_id	Player_id	Stats
1	1	1	A
1	1	2	B
1	2	1	C
1	2	2	D

Game_id	Team_id	Stats
1	1	A+B
1	2	C+D

Game_id	Home_team_id	Visitor_team_id	Home_features	Visitor_features
1	1	2	F(A+B)	F(C+D)



Feature Research



Offensive Efficiency	The number of points a team scores per 100 possessions
Defensive Efficiency	The number of points a team allows per 100 possessions
Strength	A simple measure of strength based on game-by-game using ELO Rating
Momentum	Running win rate of the past 5 games
Home Advantage	Percentage of home games won over percentage of total games won

Feature Engineering

1. **home_rate:** $\frac{\text{\# home games won}}{\text{\# all home games}}$ for home team
2. **away_rate:** $\frac{\text{\# away games won}}{\text{\# all away games}}$ for visitor team
3. **home_over_overall:** $\frac{\frac{\text{\# home game won}}{\text{\# all home games}}}{\frac{\text{\# all game won}}{\text{\# all games}}}$ for home team
4. **away_over_overall:** $\frac{\frac{\text{\# away game won}}{\text{\# all away games}}}{\frac{\text{\# all game won}}{\text{\# all games}}}$ for home team

Feature Engineering

5. win_avg5_home: $\frac{\text{\# games won in 5 previous games}}{5}$ for home team

6. win_avg5_away: $\frac{\text{\# games won in 5 previous games}}{5}$ for away team

7. Offensive_efficiency_home: $\frac{\text{PTS home}}{\text{Total_possessions_home}}$ for home team

8. Offensive_efficiency_away: $\frac{\text{PTS away}}{\text{Total_possessions_away}}$ for visitor team

Feature Engineering

9. Defensive_efficiency_home: $\frac{\text{PTS away}}{\text{Total_possessions_home}}$ for home team

10. Defensive_efficiency_away: $\frac{\text{PTS home}}{\text{Total_possessions_away}}$ for visitor team

11. elo_home: $r_{i+1} = r_i + k * (S_{\text{home}} - E_{\text{home}})$

12. elo_away: $r_{i+1} = r_i + k * (S_{\text{visitor}} - E_{\text{visitor}})$

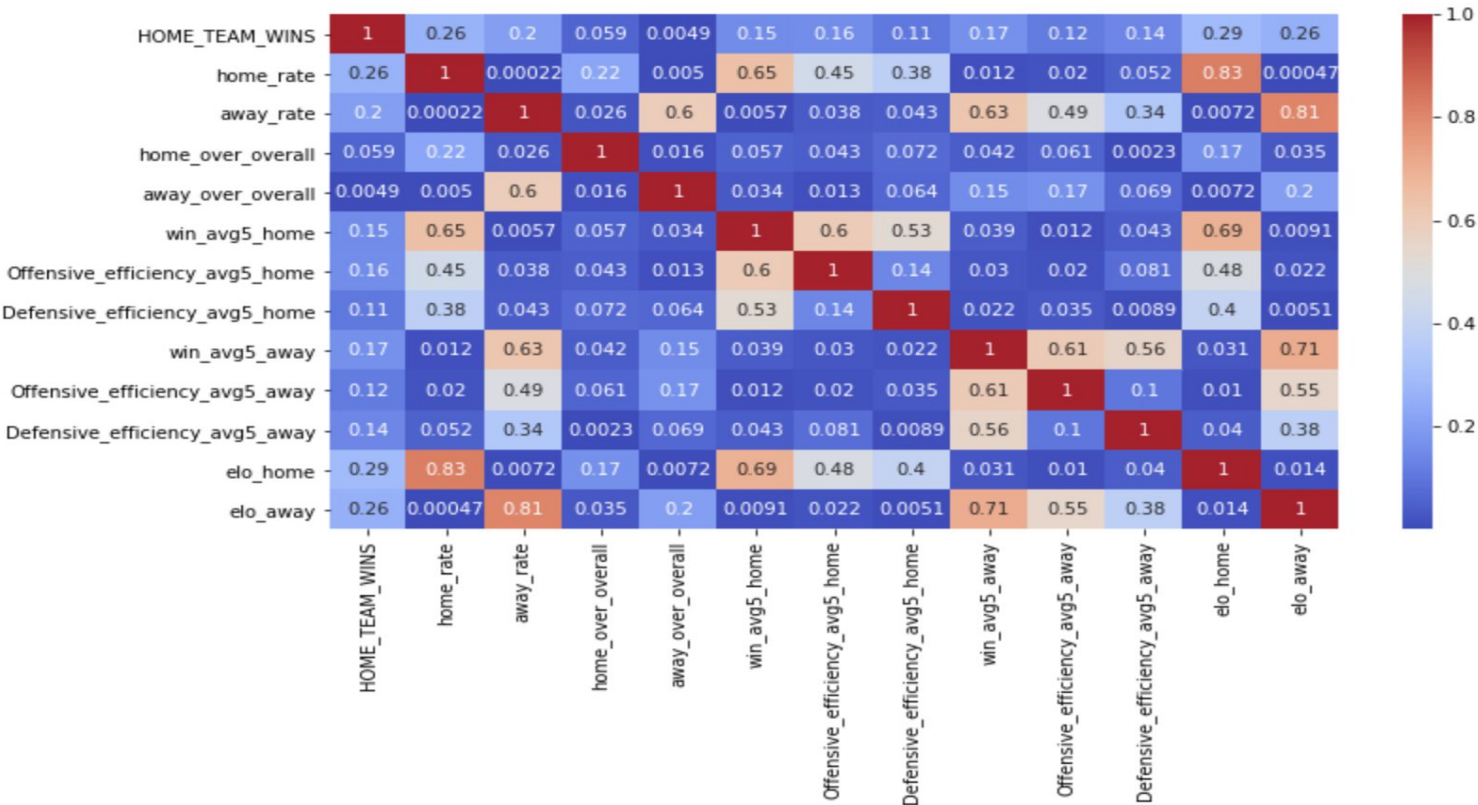
Here, S_{team} is a state variable: 1 if the team wins, 0 if the team loses. E_{team} represents the expected win probability of the team.

Exploratory Data Analysis

We picked five teams to represent teams of all levels:

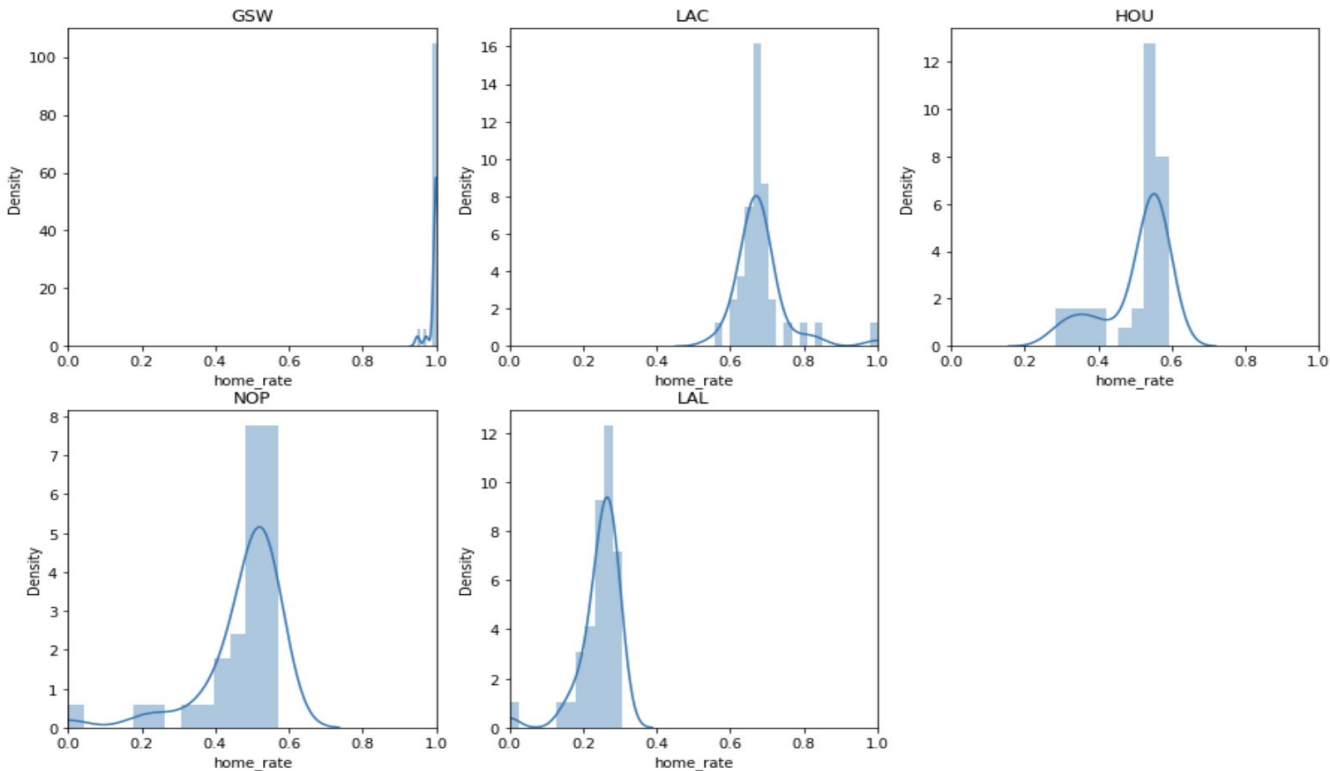
- 1: best: Golden State Warriors (GSW)
- 2: good: LA Clippers (LAC)
- 3: mediocre: Houston Rockets (HOU)
- 4: bad: New Orleans Pelicans (NOP)
- 5: worst: LA Lakers (LAL)

EDA - Correlation Heatmap



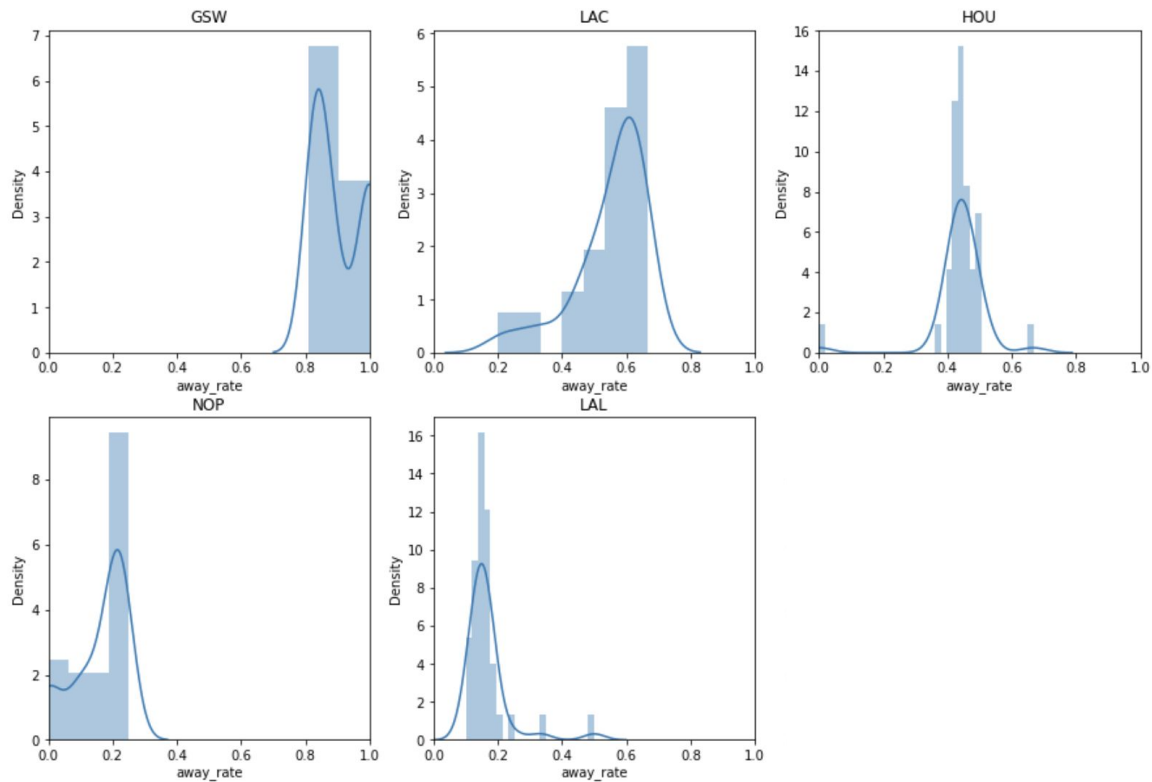
EDA - Distribution Plots (home_rate)

Home Win Rate



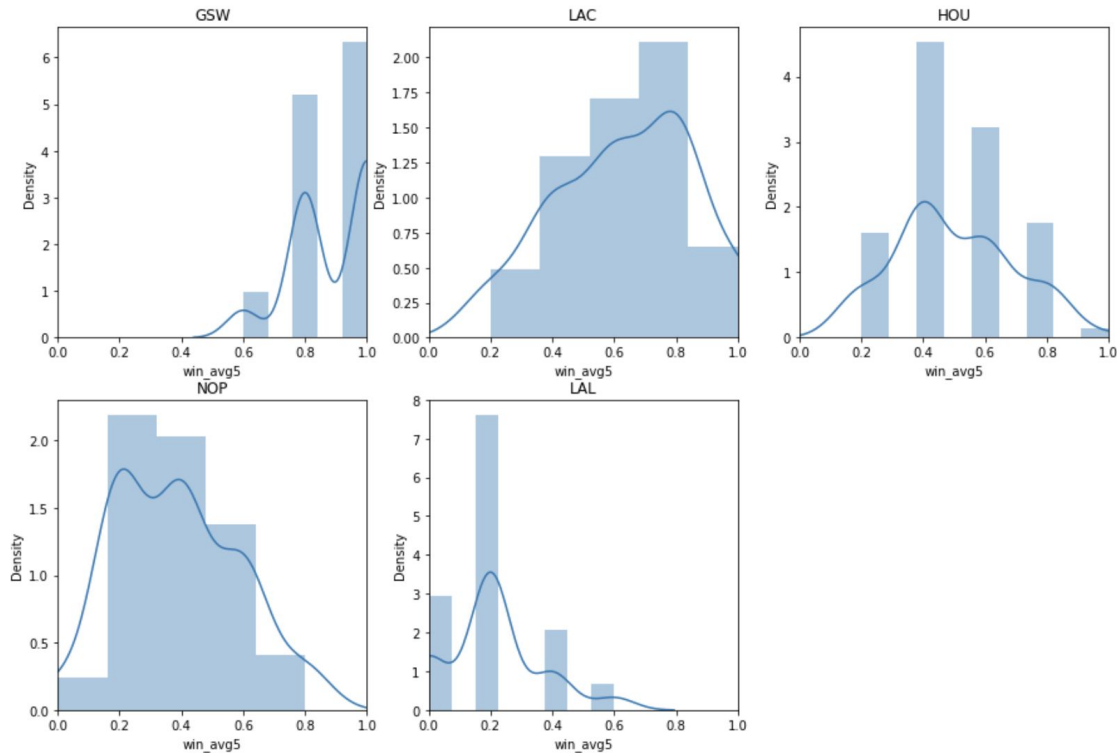
EDA - Distribution Plots (away_rate)

Away Win Rate



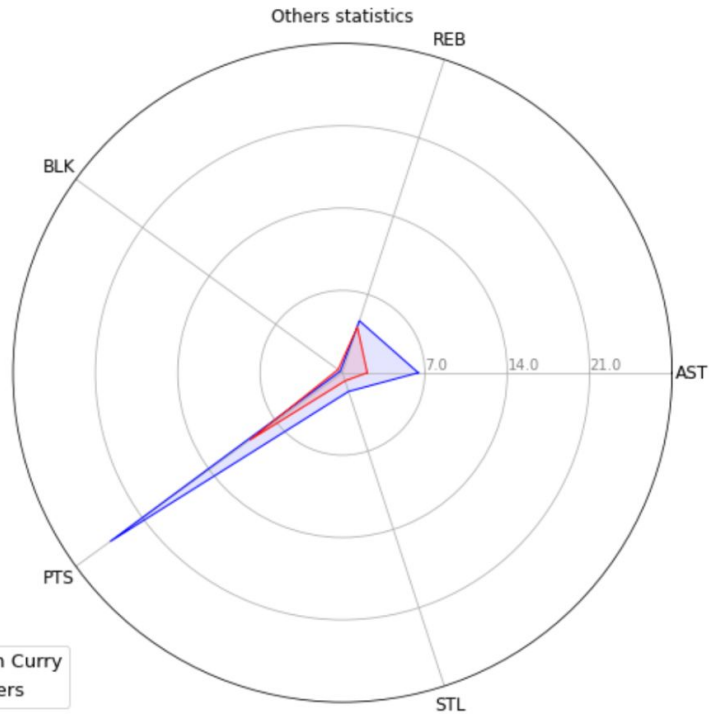
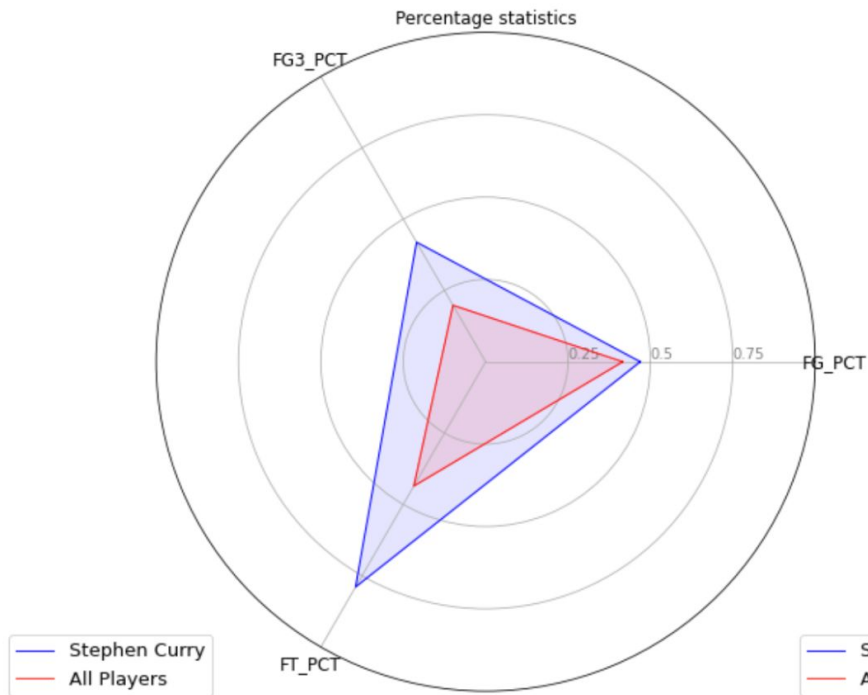
EDA - Distribution Plots (win_avg5)

Running win rate in 5 games

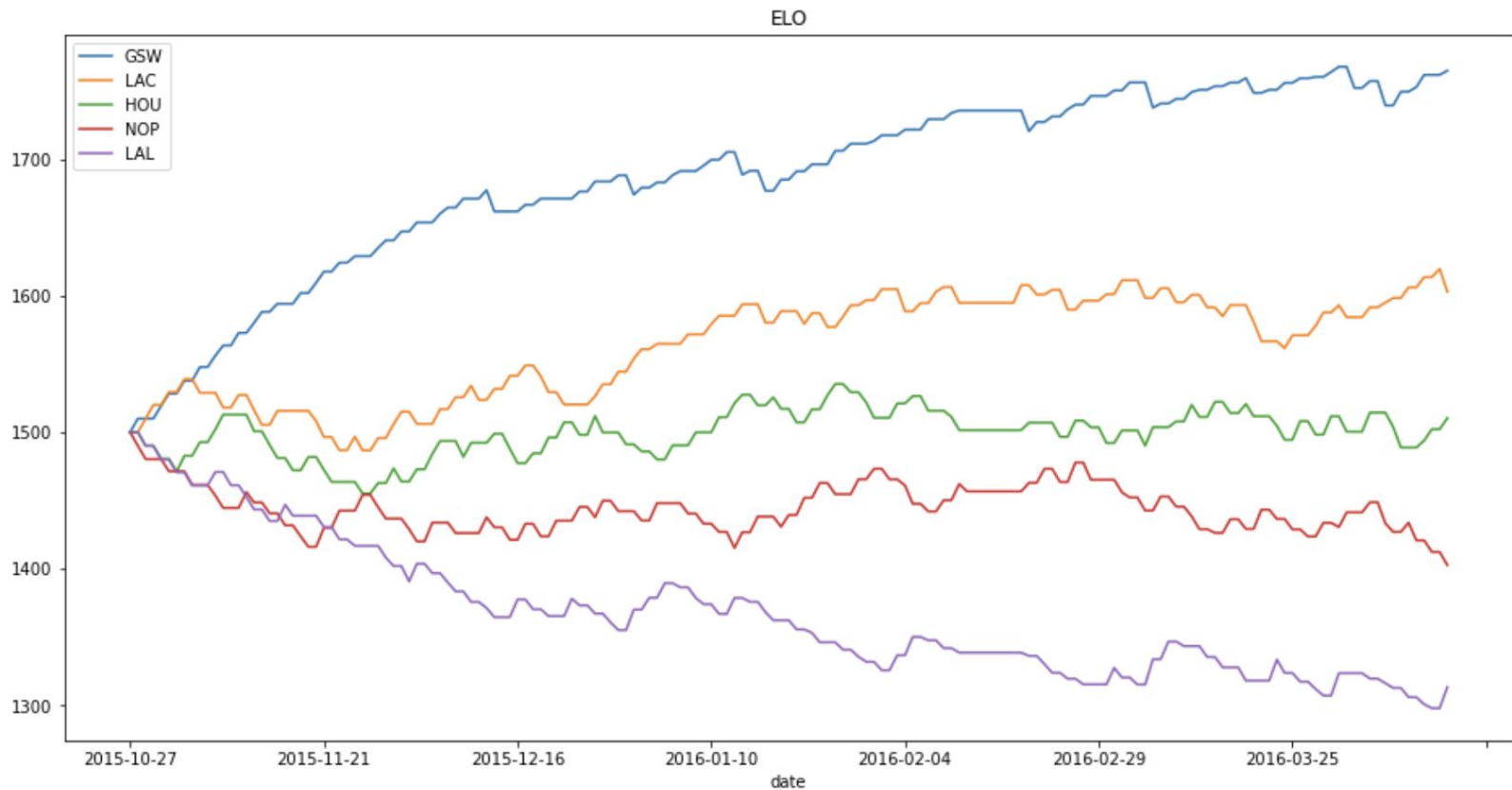


Player EDA - Stephen Curry vs. All Players avg

Stats comparison between Stephen Curry and the rest of the league

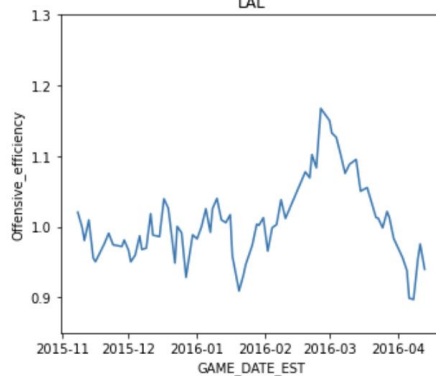
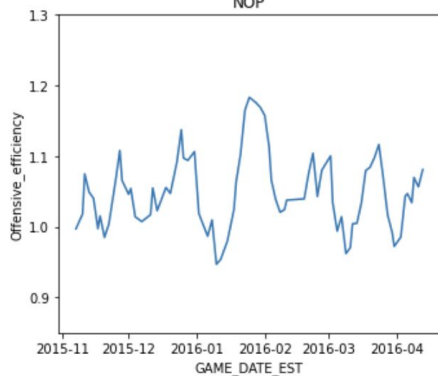
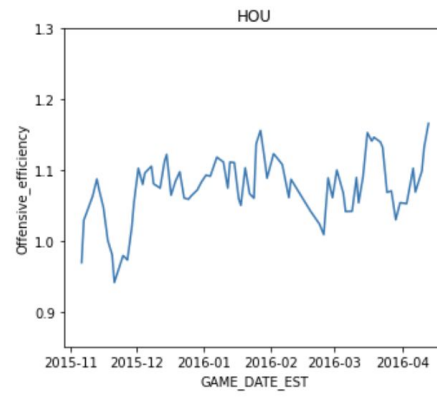
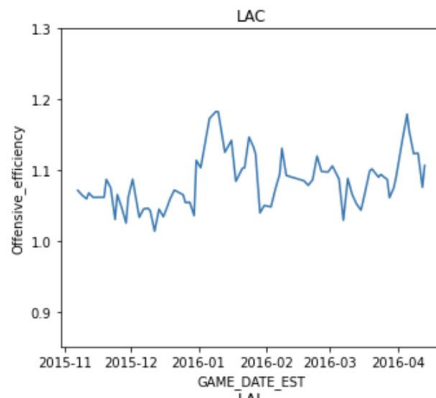
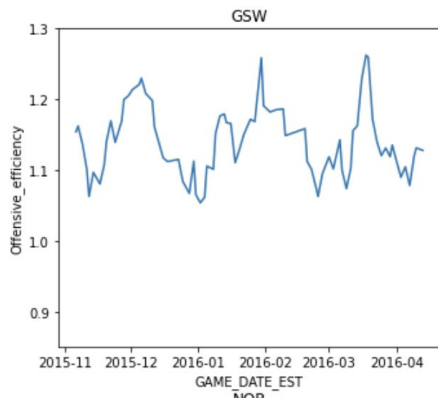


EDA - Elo over time



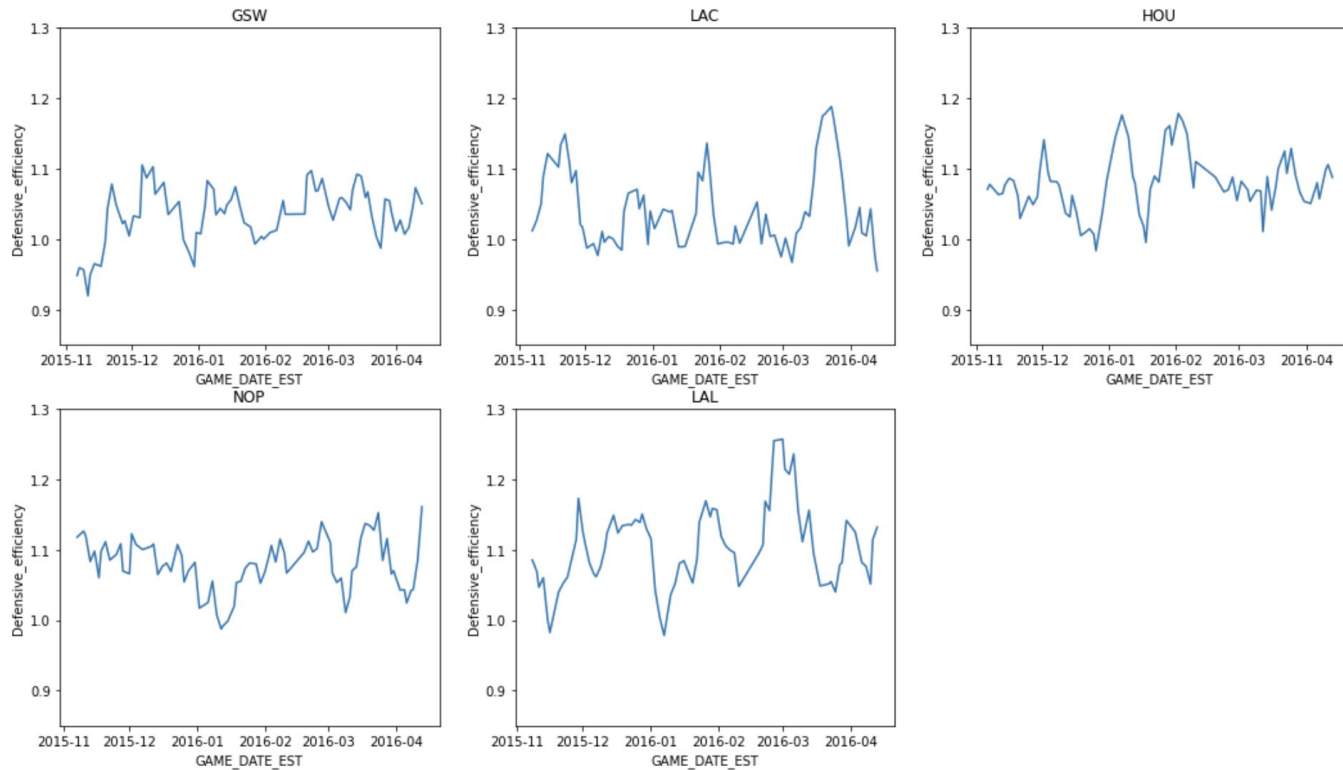
EDA - Offensive_efficiency over time

Offensive Efficiency over time



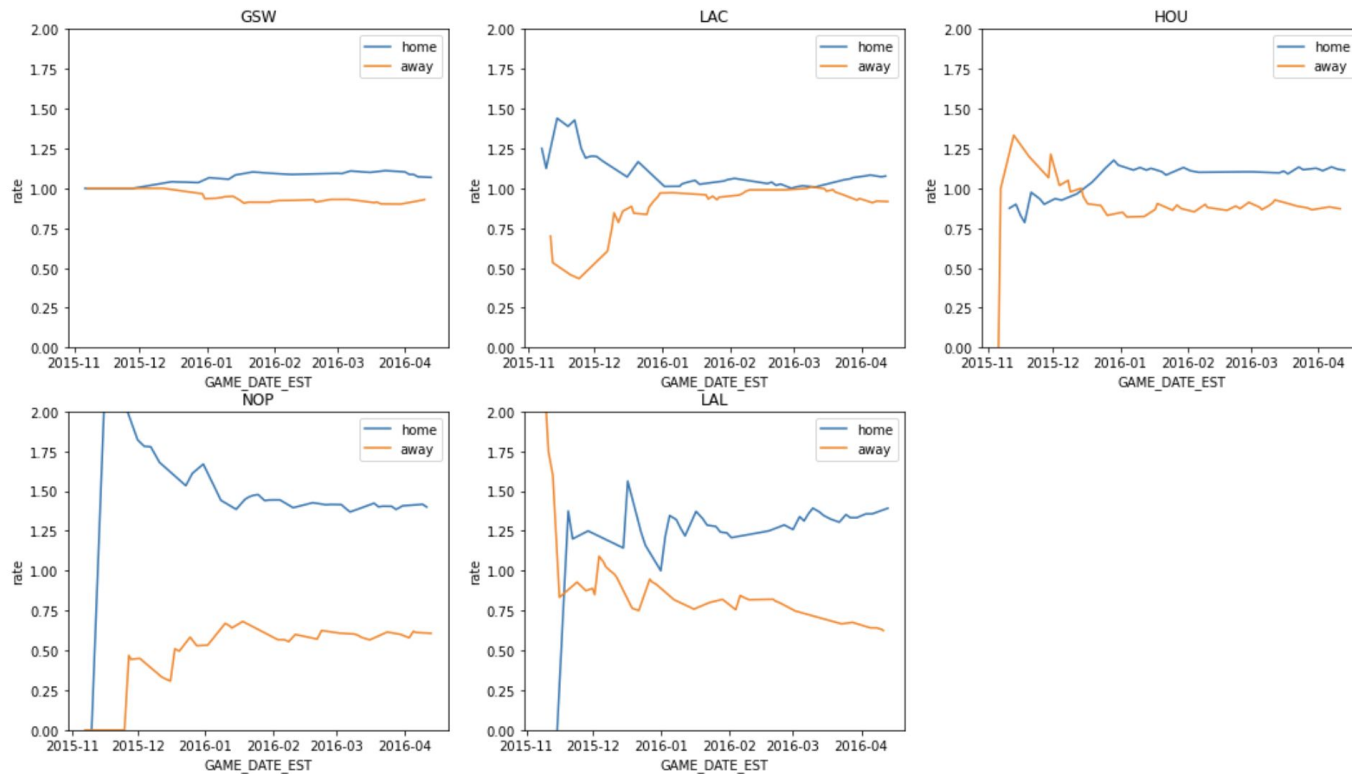
EDA - defensive_efficiency over time

Defensive Efficiency over time



EDA - home_away_rate over time

home_away_rate over time



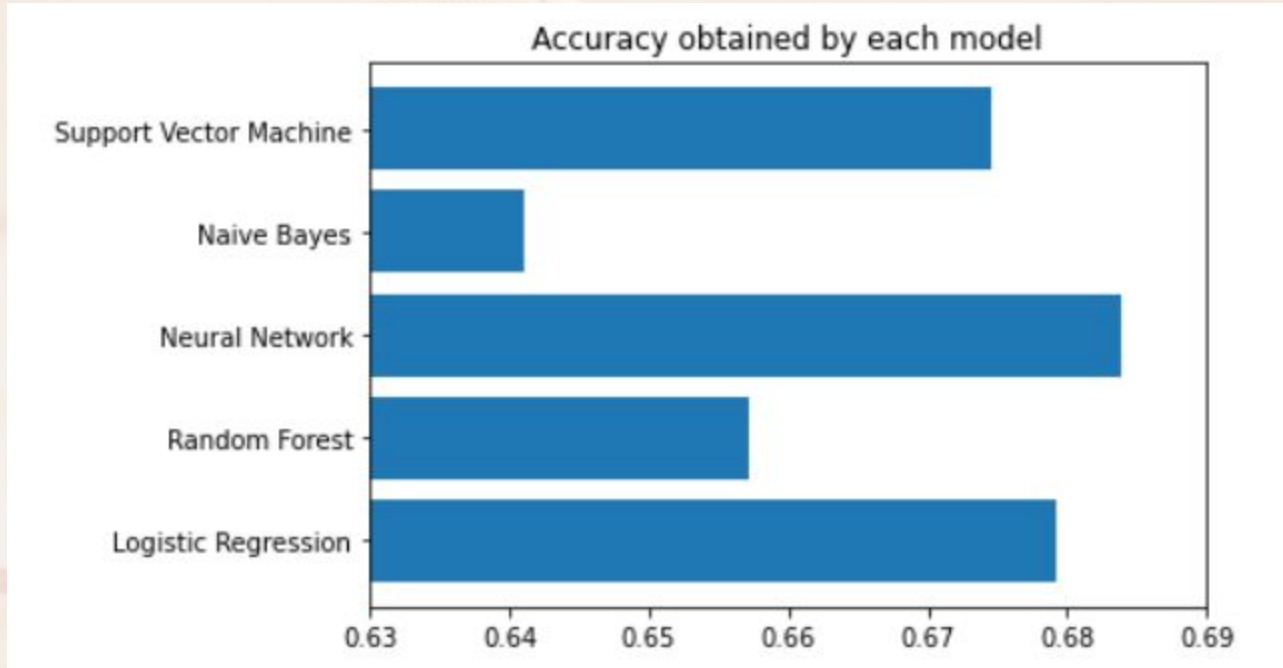
Modeling - Independent variables

- Offensive_efficiency_avg5_diff
 - Defensive_efficiency_avg5_diff
 - win_avg5_diff (momentum)
 - elo_diff (strength)
 - stad_diff (home_over_overall - away_over_overall)
- 
- A basketball is positioned in the center of the slide, slightly behind the list of variables. It is an orange ball with black lines, resting on a light-colored wooden floor with red boundary lines. The background is a soft, out-of-focus image of a basketball court.

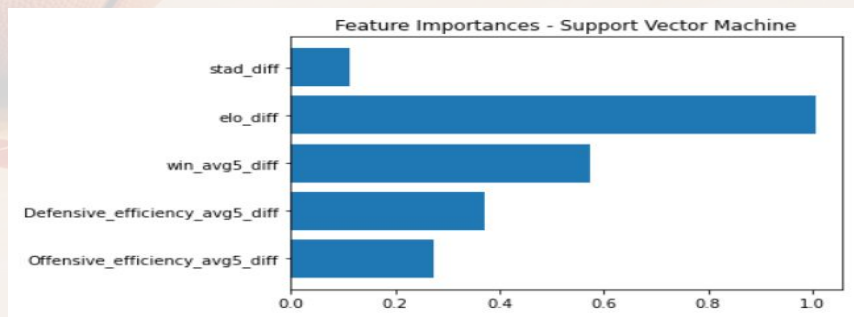
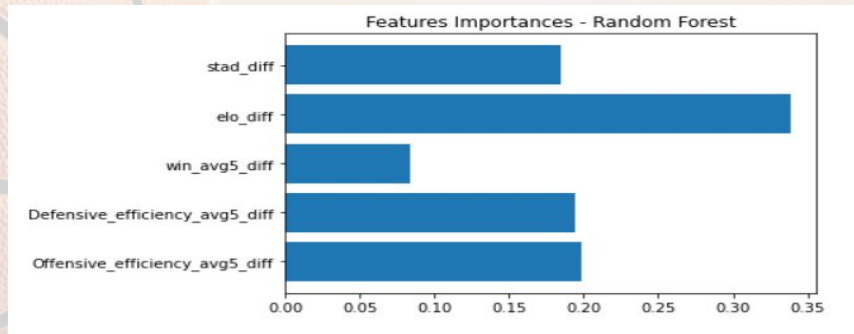
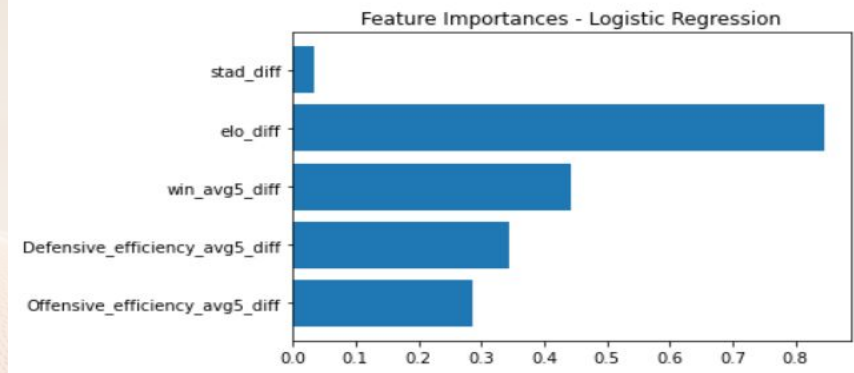
Modeling - Baseline: 50%

Modeling	K-Fold Cross Validation Accuracy	K-Fold Cross Validation Standard Deviation	Hyperparameter Tuning using GridSearchCV
Logistic Regression	67.58%	6.21%	67.93%
Random Forest	64.1%	5.80%	65.72%
Neural Network	68.04%	5.90%	68.39%
Naive Bayes	64.10%	8.12%	64.10%
Support Vector Machine	66.88%	5.67%	67.45%

Modeling - Baseline: 50%



Modeling - Feature Importances



Conclusion

Results

- Improved accuracy from 50% to 69% (with highest accuracy recorded at 72%)
- Identified ELO rating and momentum as the most important factors

Further Applications & Improvement

- Incorporating data of dominant players & events (e.g. injuries, transactions)
- Compare differences between different NBA seasons to capture seasonal variation
- Modify ELO equation so number of games is taken into account when predicting games
- Implement a spread prediction model in order to use in real world gambling

A 3D-rendered orange basketball with black lines is positioned on a light-colored wooden basketball court. Red court lines are visible on the floor. The text "Thank you" is written in a dark blue, sans-serif font across the middle of the basketball.

Thank you