Data Analytics

MoneyBall Watson

By: Aryan Mishra, Brian Sohn, Eli Salk, Yiyao Qu, Ziyi Liu

Abstract

As sports betting continues to rise in popularity, we are curious to see how accurately we can construct a binary classification model to predict wins and losses during the regular season games in the NBA. Our goal is to capture useful information from game statistics, matchups, and engineering features to predict the winner in each game through machine learning. We built multiple machine learning models to determine the strongest model and how important the features we added are in determining the outcome of games. The findings will have useful insights for various stakeholders, especially the NBA teams. They can use our analysis to identify potential factors that can help them win games and provide valuable insights for NBA sports gambling and fantasy basketball participants.

Data Acquisition

We acquired the original datasets from Kaggle: games.csv, games_details.csv, and teams.csv deriving from *basketball-reference.com*. The games.csv included all games, pre-season, regular-season, and playoffs from 2003 to 2020 for each team with the outcome and team statistics. Games_details.csv contained player statistics for each game. Teams.csv had all the team IDs and other information about the organization, which is irrelevant.

Data Preprocessing

After reviewing the datasets, we decided it would be best to only account for games during the regular season for two reasons: pre-season games do not accurately represent a team's strength, since starters do not usually get many minutes; playoff games only account for the teams who make the playoffs, which is only 16 of the 30 teams in the NBA.

We scraped the regular season dates from Wikipedia and set conditions for our data frame to only show those within the dates after converting this metric to a DateTime function. This made our data frame account for a total of 82 games played by each team in each season.

Furthermore, instead of analyzing the entire dataset (due to time and computational constraints), we chose to focus our analysis on 2015-2016. This season was significant to us personally since it was a historic record-breaking season for the Golden State Warriors with 73 wins and only nine losses.

We reorganized the dataset so that it could be used in later phases. First, individual statistics in games_details.csv were all summed up into game statistics, since our model required its inputs to be on the game level. Second, the two rows for each game (one for home, the other for away) were combined into a single row so that we had exactly one row for every game, allowing us to join this table to games.csv. Thus, our new table consisted of 1230 rows, one for each regular season game, each of which had columns for both home team and away team's statistics in the game. These raw statistics could finally be used in computing engineered features, which would eventually become the independent variables in our models.

Feature Research and Engineering

After researching background sports literature on basketball and analyzing NBA team success, we developed metrics to include in our dataframe, as seen below. Our four main focus areas were strength, momentum, efficiency, and home-court advantage. Feature engineering is an essential process in data analysis to reap utmost benefits from the data available. When feature engineering processes are executed well, our resulting dataset will be optimal and contain all the essential factors that would help with the quality of our model and drawing insights. (See appendix for formulas).

Strength	The strength attribute was measured using an elo rating (see appendix for formula) which starts at 1500 and increases with wins and decreases with losses over time. The procedure shows that the points gained or lost were related to the probability of winning or losing the game. If a team had a low chance of winning a game and won, they would be given a more significant amount of points towards their elo than if they were favored, and vice versa with losses.
Momentum	In sports, momentum is something that many fanatics believe in heavily, and others say is irrelevant. However, recently many studies have found that momentum for a player or team does affect performance. Therefore, we found it necessary to include a feature that took the running win rate of the past five games to determine if their momentum would help them accomplish a win.
Efficiency	As sports analytics becomes increasingly popular within the NBA, efficiency seems to be the pillar. Efficiency was reviewed both in terms of offense and defense (See appendix for formulas). A higher value for offensive efficiency indicates that a team is offensively stronger. This is opposite in the case of defensive efficiency, where a lower value indicates stronger defensive power.
Home Court Advantage	We implemented the concept of home-court advantage by taking the percentage of home games won within all home games and dividing that by the rate of all games won within all games. (see appendix for formula) This formula cleared up any teams that performed overall bad or terrific, which would not accurately represent the strength of their home or away rate, and measured the value of home-court advantage as well as how well teams perform on the road.

Exploratory Data Analysis

We decided to use five teams in our data analysis, representing all teams' performance during the 2015-16 regular season: Golden State Warriors (best), LA Clippers (good), Houston Rockets (mediocre), New Orleans Pelicans (bad), Los Angeles Lakers (worst). After creating a correlation heatmap (See Figure 1 - appendix) to detect any multicollinearity, we discovered that the home variables are correlated, and the away variables are also correlated. The dependent variable is not highly correlated to any independent variable because variables for both the home team and visitor team should be compared. In other words, you can not predict the winner with only one team's data.

Afterward, we drew distributions for all variables to visualize the patterns and confirm that our data was on the right track. Analyzing the home over overall win rate (See Figure 9 - appendix), we can see that GSW remains first place since they had an incredible season, and their home games vs. away games records were more or less the same. Although they weren't any good that year, they always performed better at home. When analyzing other teams who performed poorly during the season, we can see that the mean is significantly larger than one.

Similarly, when reading the plots for away win rate over total win rate, we see the mean for the five teams besides the warriors is less than one, indicating that playing on the road is hurtful to most groups. Furthermore, the best teams also appeared to have higher offensive efficiency and lower defensive efficiency throughout the season (See Figure 7, 8 - appendix): the Warriors had the best in both aspects, and the Lakers had the worst in both parts. Although offense and defense are not directly related, a team that gives up the ball many times in the game has more opportunities to get scored on and less control.

Our last metric, which appeared to be the most impactful when running our machine learning models, was the elo rating. ELO is a metric that evolves and changes each game (See Figure 6 - appendix). All teams start with a rating of 1500 and either increase by wins or decrease by losses accounting for the probability of the expected result.

Modeling & Results

We proceeded to the modeling stage after making sense of our data. Our independent variables were the features that we engineered: Elo Ratings, Offensive Efficiency, Defensive Efficiency, Average win rate of the previous 5 games, and the proportion of the home win rate over the overall win rate. Our dependent variable was a binary variable, with 1 indicating the home team winning, and 0 indicating the away team winning, thus ensuring we have a binary classification problem.

Due to the nature of our dataset, we had separate statistics for home teams and away teams, and subsequently separate features (e.g. Offensive Efficiency for the home team, and Offensive Efficiency for the away team). We ended up taking the difference between the home features and the away features to combine the home data and away data. Next, we proceeded to the model selection phase. Given the situational constraints, evaluation metrics, and data sources available, we asked ourselves what types of models make the most sense to try? Some factors we considered when selecting a model included training and prediction speed, volume and dimensionality of data, and explainability. As a result, we ended up training and testing the following models: Logistic Regression, Random Forest, Neural Network, Naive Bayes, and Support Vector Machines (SVMs). Each of these models have their pros and cons. For example, when considering training and prediction speed, logistic regression was much quicker than neural networks for the same amount of data. Considering the volume of data, neural networks are usually better at handling such kinds of data, whereas when we consider explainability, Random Forests and Logistic Regression are more interpretable than Neural Networks and SVMs.

After narrowing down the types of models to the selected candidates, we proceeded to the training phase, which involved splitting the dataset into the training set (75%) and test set (25%), and subsequently training each model on the training set and validating the trained models on the test set. We also know that model training has much more than just the train/test split. Indeed, we further performed cross validation (specifically K-Fold Cross Validation) and hyperparameter tuning for each selected model. Cross validation is critical because it assesses the performance of an algorithm in several subsamples of training data, thus helping us avoid training and testing on the same subsets of data points, which would lead to overfitting. Furthermore, hyperparameters are vital because they impact a model's training time, compute resources needed (and hence cost), and ultimately are crucial in a model's performance. We tuned our hyperparameters using grid search, which involves forming a grid that is the Cartesian product of those parameters and then sequentially trying all combinations and seeing which yields the best results. Once we performed the necessary computations and steps needed in thoroughly training the models, we ended up with the following final results: 67.93% for Logistic Regression, 65.72% for Random Forest, 68.39% for Neural Network, 64.1% for Naïve Bayes, and 67.45% for SVMs.

The accuracies obtained by each model were the final accuracies obtained after training/testing the data and performing hyperparameter tuning (See Table 2 & Figure 10 - appendix). The accuracies were consistent because no one model vastly outperformed the other. The model with the highest accuracy was the Neural Network, followed by SVMs. This is unsurprising because Neural Networks are great at learning and modeling complex relationships, which is really important because in basketball many of the relationships between inputs and outputs are not alway linear and can be complex. This also applies to

SVMs, since they are usually the model of choice in sports literature due to their ability to express complex relationships in a computationally efficient manner.

To give us better interpretability of the results, and subsequently the data, we also calculated feature importance scores. Feature importance assigns the score of input features based on their importance to predict the output, thus helping us better understand which features had the highest predictive power. For this project, we calculated the feature importance scores for Logistic Regression, Random Forest, and SVM models. The model coefficients act as feature importance scores for the Logistic Regression and SVM models whereas for the Random Forest model, feature importance was computed as the (normalized) total reduction of the criterion brought by that feature (known as Gini importance). After calculating the scores and visualizing them on a plot (see Figure 11 - appendix), we found that the elo_rating feature had the highest predictive power. This is unsurprising given the amount of background literature and research available that has gone behind creating this feature.

Conclusion

Throughout this project, our team successfully acquired the data, applied conditions to filter out data and reconstruct the dataframe, and engineered new features that can capture the impact of multiple existing features at once. Then, with the preprocessed data, we were able to build various models including Logistic Regression, Random Forest, Neural Network, Naive Bayes, and SVMs, to get the prediction we want. Our ultimate finding indicates that Neural Network performs best to our data, with an accuracy score approaching 70%.

Possible Improvements

To improve our analysis, we could have incorporated data of dominant players and events (for example injuries and trades), which could have had a significant impact on the outcome of the game. Another avenue of improvement is feature engineering. For instance, we might functionalize the constant k in the ELO equation so that it becomes a dynamic parameter in order to take the time of the season into account (for example some teams might naturally be stronger at the end of the season rather than the beginning). Moreover, we could have combined different features with different models (instead of using the same features for each model), potentially increasing model performance and robustness. Considering that teams change their roster and coaches every season, we could have also compared differences between different NBA seasons to capture seasonal variation (e.g. by calculating the average accuracy of our models across multiple seasons), thus providing predictions based on each season. Finally, in addition to simply predicting the wins and losses, we could also construct new models predicting the score spread in games to give gamblers a clearer picture of the strength differences between teams.

Appendix

Table 1
Summary of features

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

Table 2
Model Results

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%	68.38%
Naive Bayes	64.10%	8.12%	64.10%
Support Vector	66.88%	5.67%	67.45%

Machine			

Formula for Features

1. ELO (strength): $R_{i+1} = R_i + k * (S - E)$
   a. S = 1 if win, 0 if loss
   b. E = Probability of winning = $\frac{1}{1 + 10^{\frac{R_{i,\,opponent} - R_i}{400}}}$
2. Win_avg5 (momentum): $\frac{\#\ of\ games\ won\ in\ last\ 5\ games}{5}$
3. Offensive Efficiency: $\frac{Points\ Scored}{Possessions}$
   a. Possessions: Field Goals Attempted - Offensive Rebound + Turnover + 0.4 * Free Throws Attempted
4. Defensive Efficiency: $\frac{Points\ Conceded}{Possessions}$
5. Home (away) over overall (home court advantage): $\frac{\frac{\#\ of\ Home\ (Away)\ \ Games\ Won}{\#\ of\ \ Home\ (Away)\ \ Games}}{\frac{\#\ of\ All\ Games\ Won}{\#\ of\ All\ Games}}$

Figure 1

(Absolute) Correlation of Features

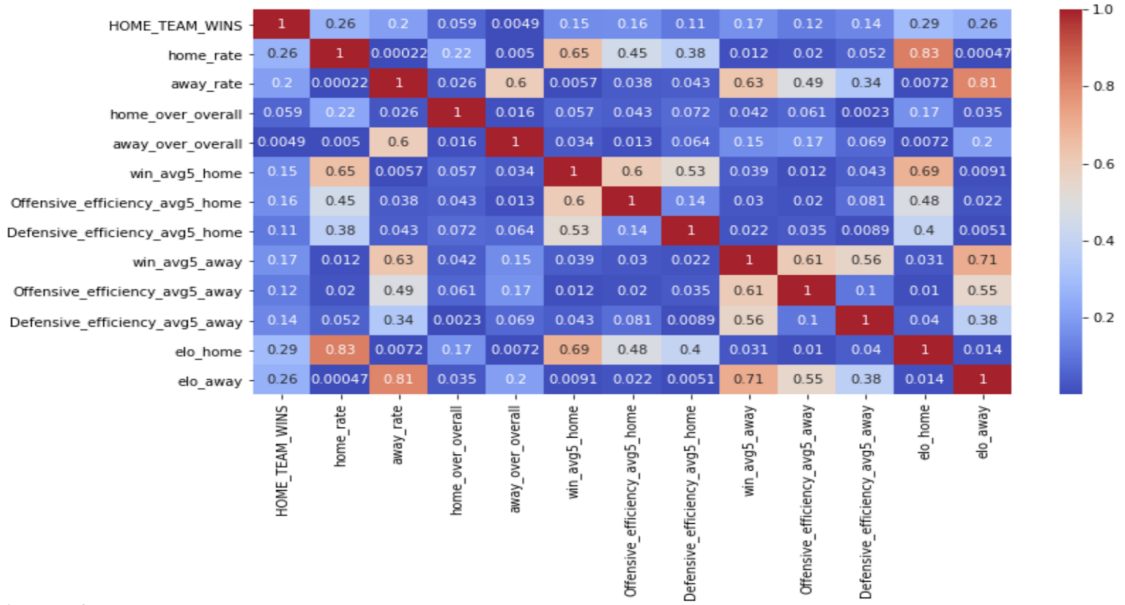


Figure 2

Distribution of Home Win Rate

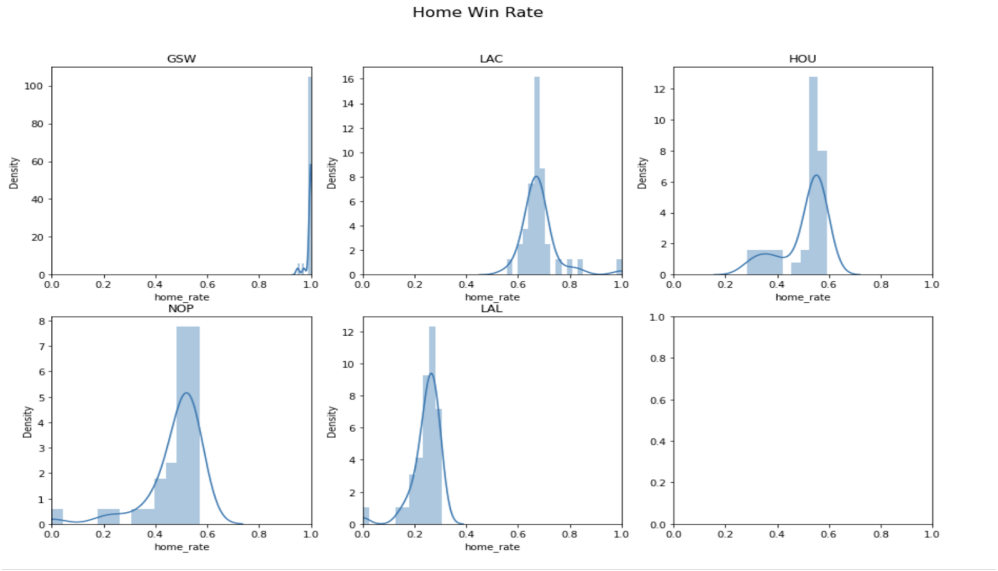


Figure 3
Distribution of Away Win Rate

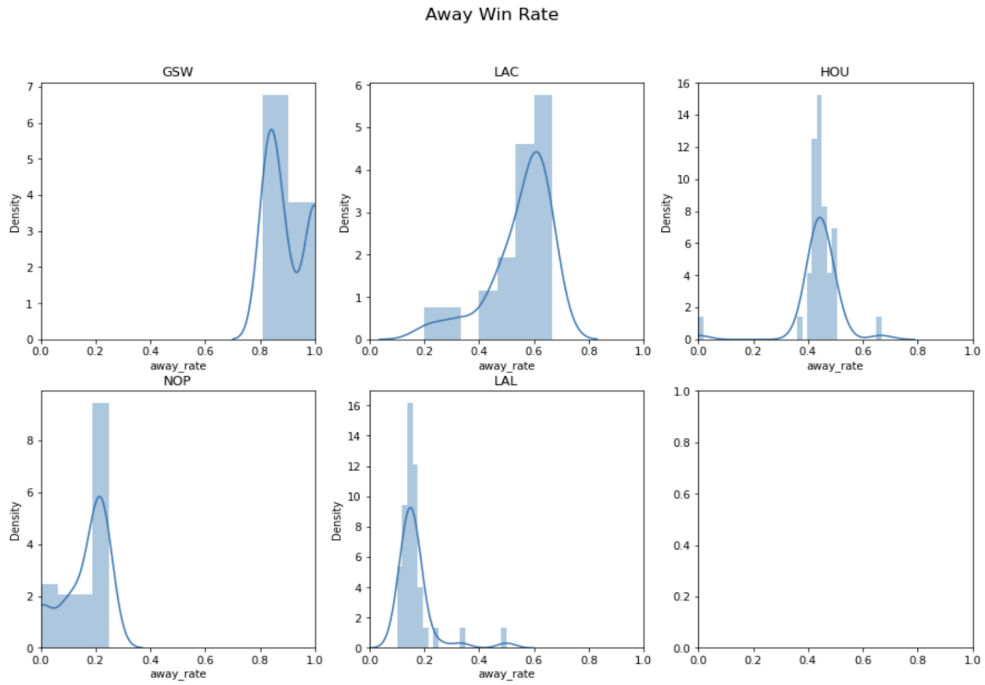


Figure 4

Distribution of Running Win Rate in 5 games

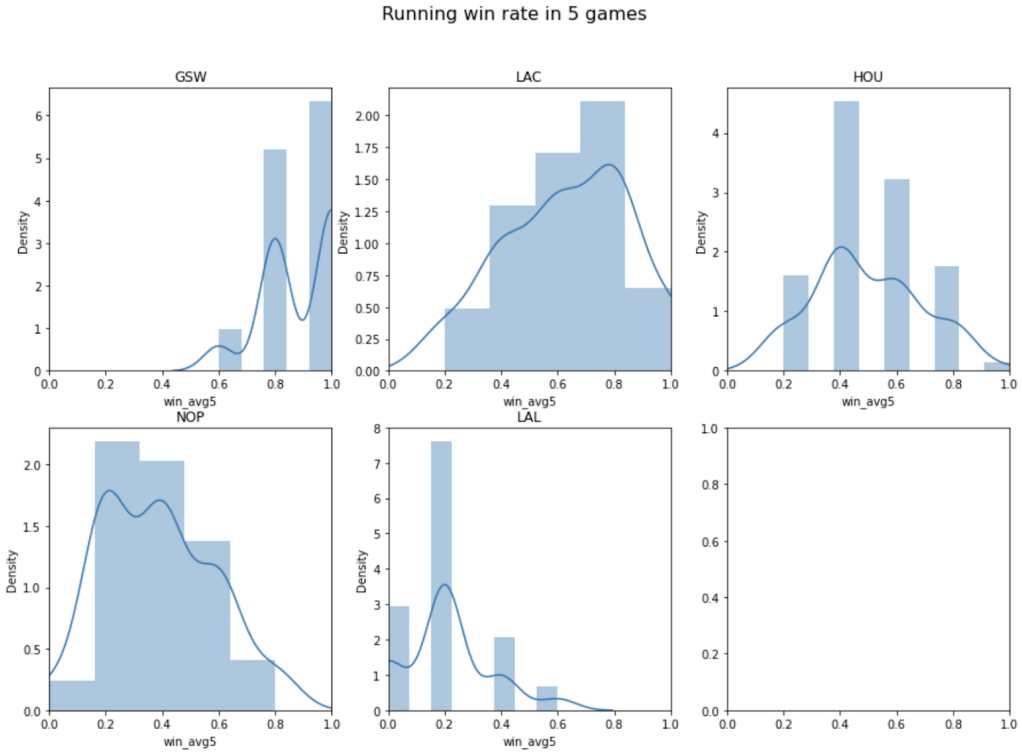


Figure 5

Player Statistics Comparison: Stephen Curry vs Average of All Players

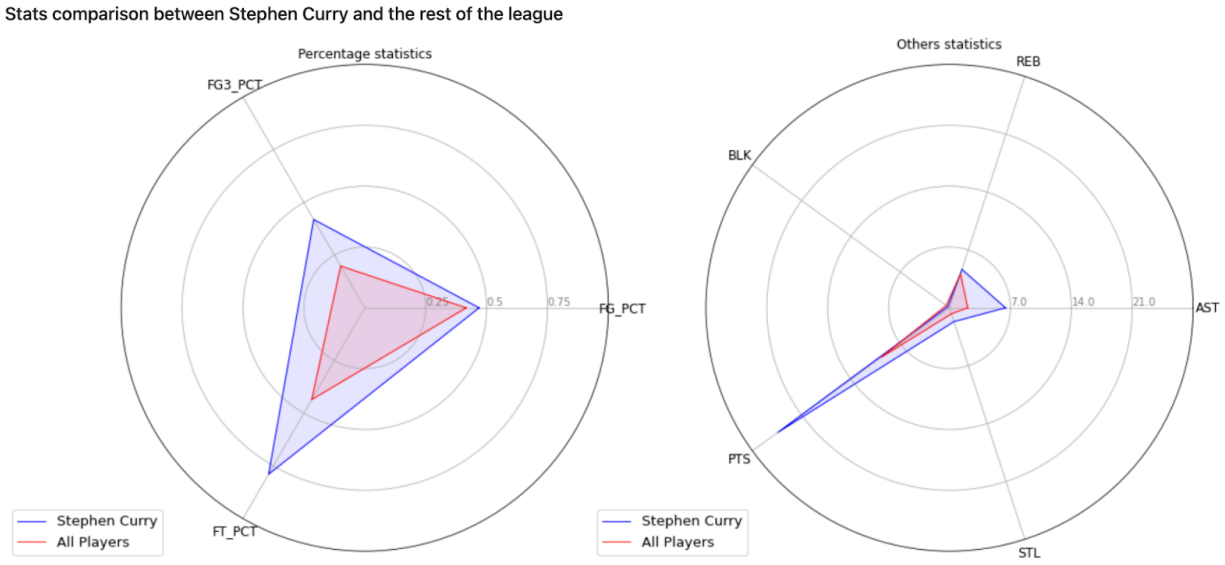


Figure 6
ELO over Time

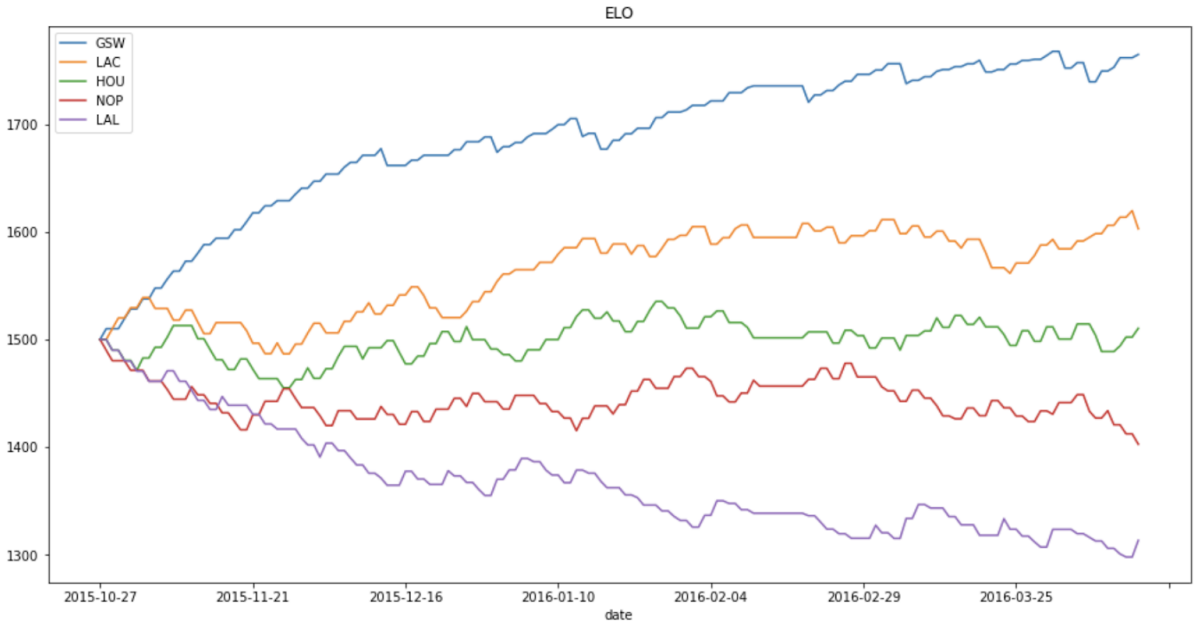


Figure 7
Offensive Efficiency over Time

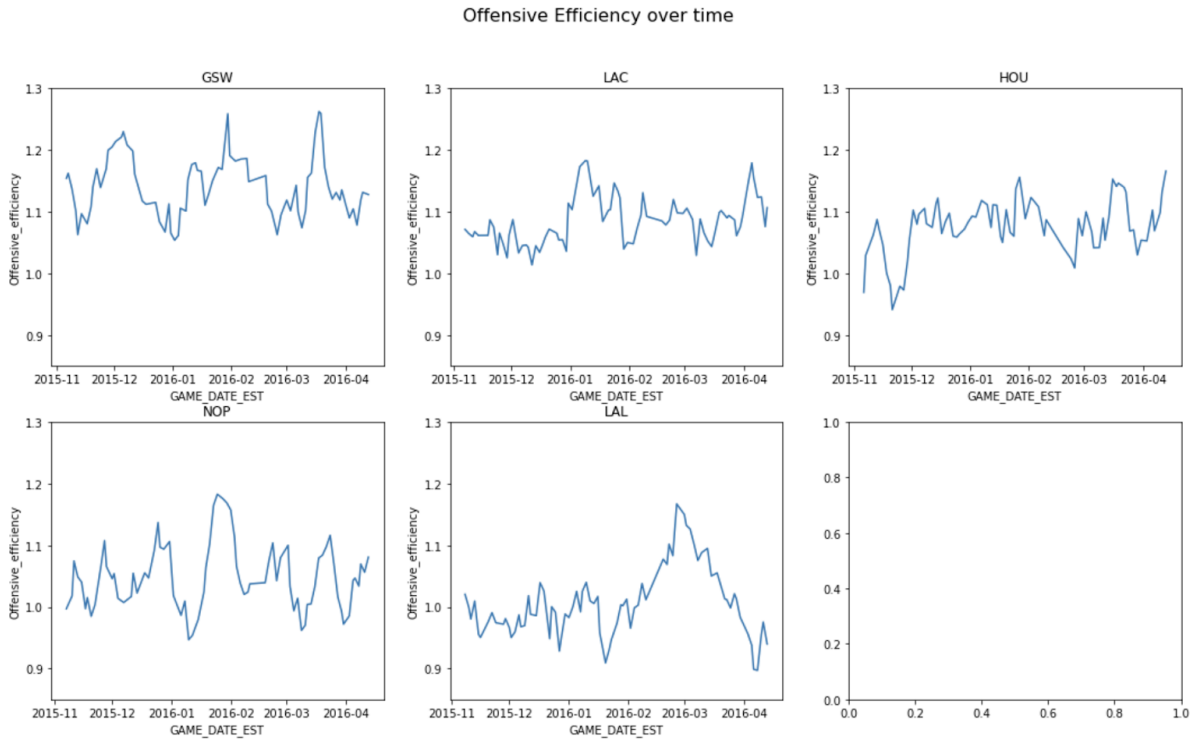


Figure 8
Defensive Efficiency over time

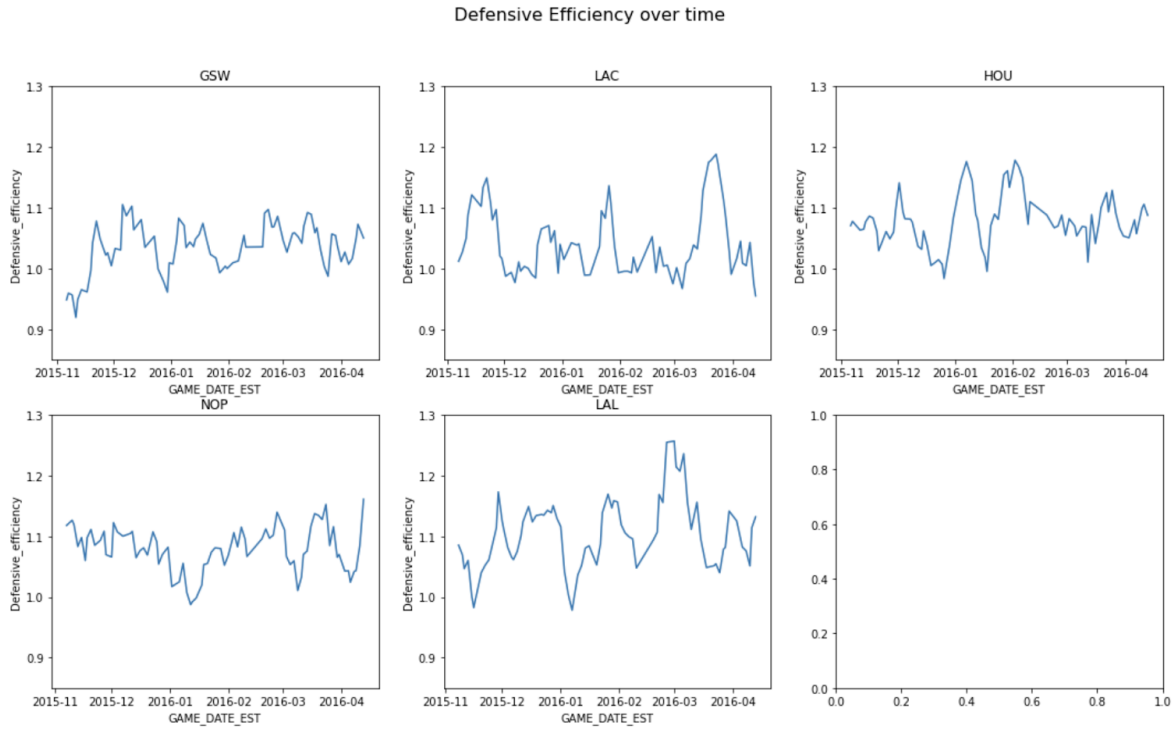


Figure 9
Home over Overall Win Rate and Away over Overall Win Rate over Time (see home court advantage)

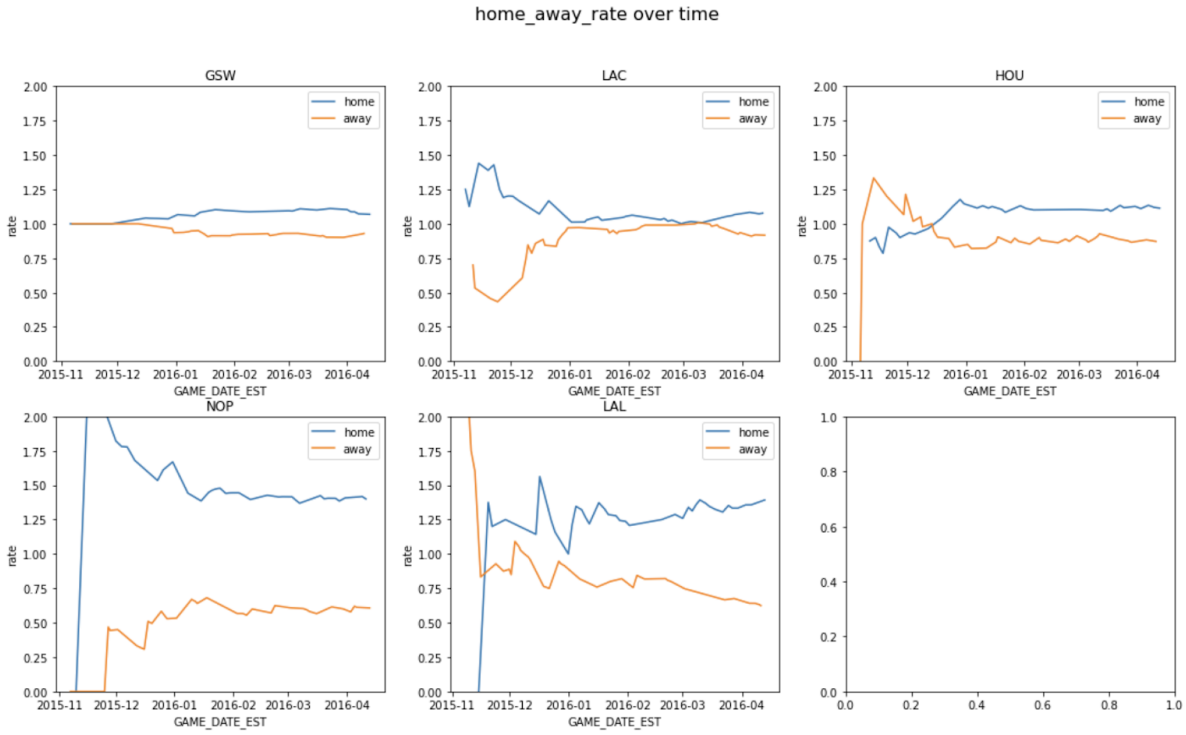


Figure 10
Accuracy of Each Model

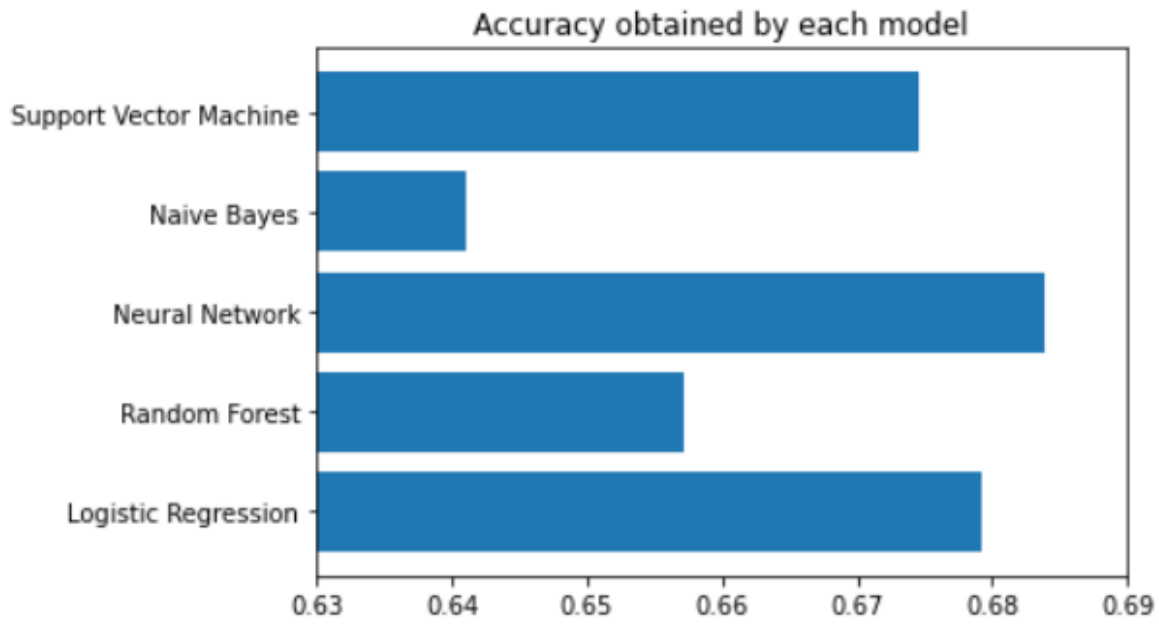
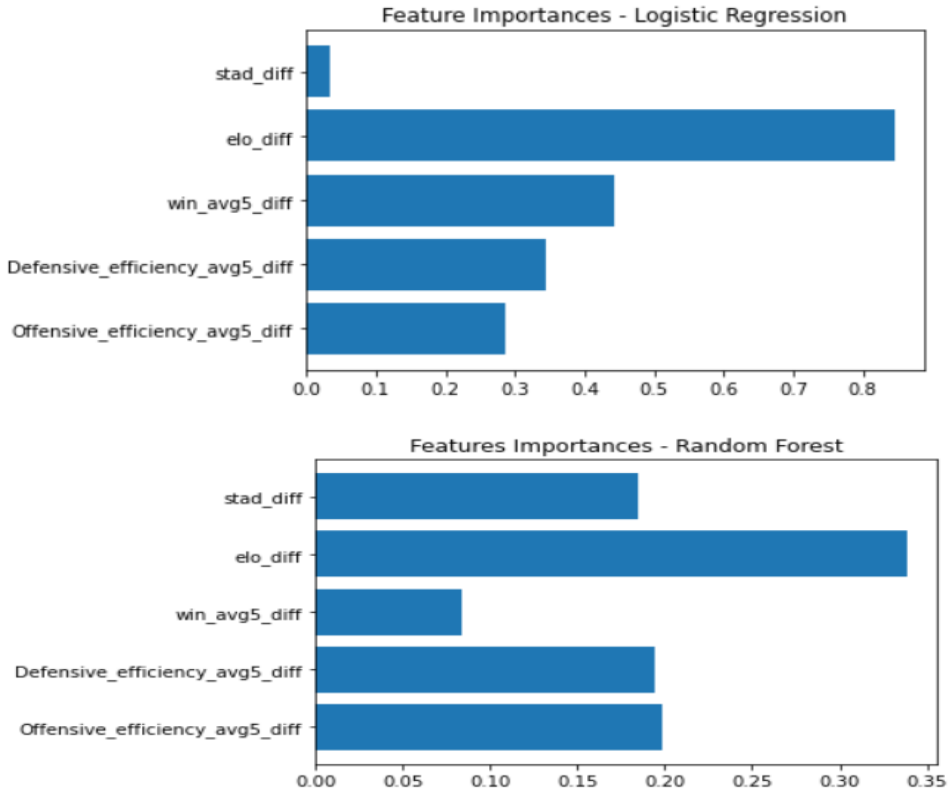
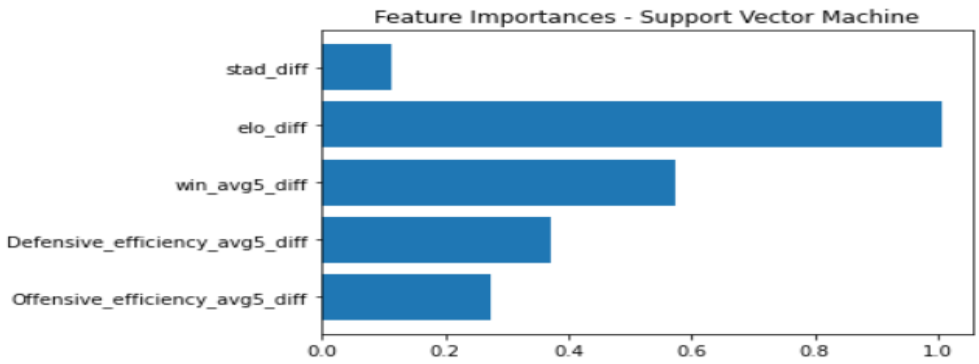


Figure 11

Feature Importances from Each Model





References

Arkes, Jeremy. "Core.ac.uk." *Finally, Evidence for a Momentum Effect in the NBA*, Calhoun: The NPS Institutional Archive, 2011, https://core.ac.uk/download/pdf/156711222.pdf.

"Elo Rating Algorithm." *GeeksforGeeks*, 20 Mar. 2018, https://www.geeksforgeeks.org/elo-rating-algorithm/.

Fein, Zach. "Cracking the Code: How to Calculate Hollinger's per without All the Mess." *Bleacher Report*, Bleacher Report, 3 Oct. 2017, https://bleacherreport.com/articles/113144-cracking-the-code-how-to-calculate-hollingers-per-without-all-the-mess.

Lauga, N. L., *NBA Games Data* (Version 8) [Dataset], 2021, November 18, https://www.kaggle.com/nathanlauga/nba-games

Mikołajec, Kazimierz, et al. "Game Indicators Determining Sports Performance in the NBA." *Journal of Human Kinetics*, Akademia Wychowania Fizycznego w Katowicach, 5 July 2013, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3796832/.

natesilver538. "How We Calculate NBA Elo Ratings." *FiveThirtyEight*, FiveThirtyEight, 21 May 2015, https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/.

Neil_Paine. "How Our NBA Predictions Work." *FiveThirtyEight*, FiveThirtyEight, 18 Dec. 2018, https://fivethirtyeight.com/methodology/how-our-nba-predictions-work/.

Weiner, Josh. "Predicting the Outcome of NBA Games with Machine Learning." *Medium*, Towards Data Science, 7 Jan. 2021, https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20.