Predicting NBA Match Results

Ivan Karp, [redacted], [redacted], [redacted], [redacted], [redacted]

1 Introduction

The purpose of this report is to analyze NBA game outcomes using a dataset containing game statistics, team performance metrics, and outcomes. By utilizing advanced basketball statistics, temporal features such as weighted averages, and contextual factors such as home-court advantage and win-loss records, we aim to create predictive models for forecasting game winners. We engineered features to reflect both cumulative performance and game-specific dynamics, utilized backward feature selection to identify key predictors, and evaluated models including Logistic Regression, Support Vector Machine (SVM), Quadratic Discriminant Analysis (QDA), XGBoost, and Random Forest. Among these models, Random Forest with low-decay weighted averages achieved the best performance, with a testing accuracy of 67.9%.

A detailed description of the data preprocessing, feature engineering, model evaluation, and insights into predictive performance are presented in this report, which also discusses the challenges associated with forecasting upsets and the inherent randomness in basketball.

2 Data Preprocessing

2.1 Investigating the Target Variable

We first examined the binary target variable—the W/L column. The W/L column contained equal numbers of wins and losses as expected, since the dataset split each game into two rows, where each row contained statistics for each team that competed. This makes it much easier for us to create a meaningful predictive model. One issue is that all the variables within our given dataset are dependent on time. The line plot below showing the total W/L difference for each team over time highlights the importance of how previous game results correlate to the outcome of the specific game in question. A solution to this is to incorporate weighted averaged variables for better prediction results.

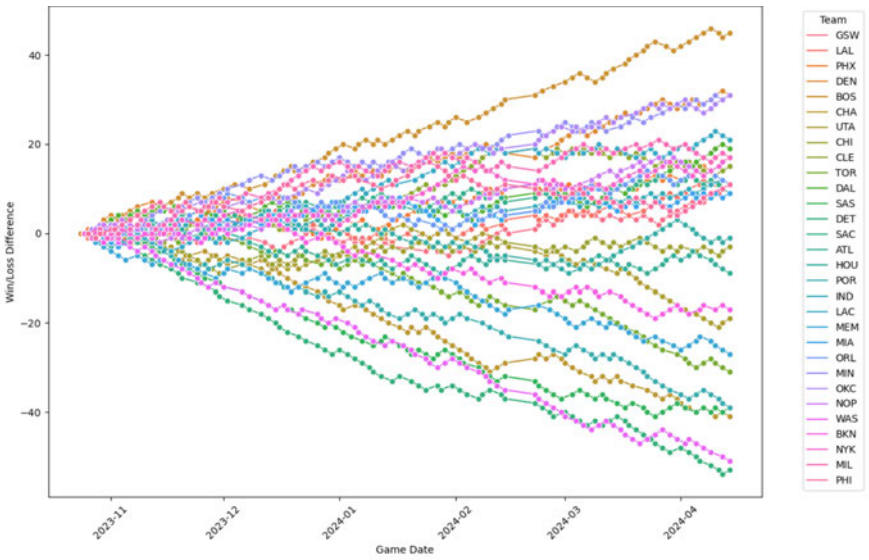

Fig. 1: Change in Win_Loss_Diff Over Time for each team over the season.

2.2 Home & Opponent

The Match Up column in the original dataset notes the two teams that played in the game specified in each row. Based on the description of the Match Up column given in the guidelines, we were able to extract 2 potentially valuable predictors - Home and Opponent. The Home column signifies whether the game was held in the team's arena, and the Opponent column holds the opponent that faced up against the team for that specific match.

2.3 Overtime

The original dataset records the number of minutes each game took. Usually, each game would take 240 minutes. In some cases, the game may take up to 265 or 290 minutes. To count for the situations where certain teams may have players with higher endurance strength, we decided to make a boolean column Overtime that indicates whether the game ends within 240 minutes or not (we did not distinguish between 265 and 290 as no game took 290 minutes before November 13, 2023). We then dropped the column Min, as it has become redundant.

2.4 Possessions

We added another feature called Possessions, which estimates the total number of possessions a team had during a game. This metric has been widely used in basketball analytics, as it forms the basis for many other advanced statistics, such as offensive and defensive ratings. We used the formula

$$Poss. = FGA - OREB + TOV + 0.44 \times FTA.$$

By including Possessions as a feature in our model, we aim to capture the pace and dynamics of each game [1].

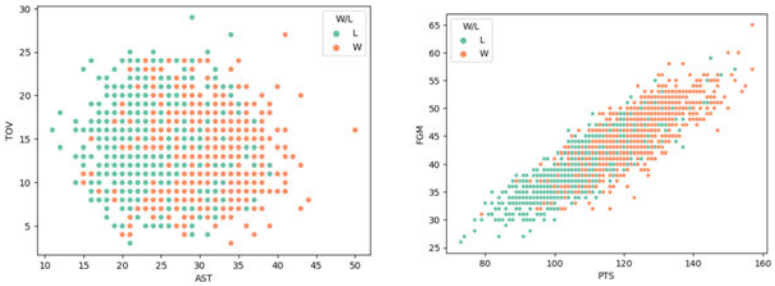

Fig. 2: Example of uncorrelated variables AST and TOV in each game colored by Win or Loss (left), and correlated variables PTS and FGM.

2.5 Advanced Basketball Statistics

Basketball is a complicated game—simple statistics tend to mislead the viewer into hasty conclusions without providing the full context of the game. As avid viewers of NBA basketball, we believed adding advanced basketball statistics would provide a better understanding of how the teams were performing each game. Therefore, we decided to incorporate the following four advanced basketball statistics:

Column	Description	Formula
ATR	Assist-to-Turnover Ratio	AST/TOV
eFG	Effective Field Goal Percentage	$FGM + 0.5 \times 3PM/df.FGA$
OFR	Team Offensive Rating	$PTS/Possessions$
DFR	Team Defensive Rating	$PTS_allowed/Possessions$

Table 1: Table of advanced basketball statistics and their formulae.

2.6 Weighted Features

As mentioned previously, the prediction accuracy of our dataset relies heavily on time, where recent games played by the team are more relevant to the outcome than previous ones. All NBA fans can agree that momentum is one of the most important aspects of the game. This is why oftentimes basketball analysts believe a winning streak increases the probability of that team winning the game. In a statistical sense, basketball games are not independent—each game has a high correlation to the game directly preceding it. Therefore, we believed it was important to scale and average the numerical columns with appropriate weights corresponding to their dates.

2.7 Weighted Average Formula

We decided to use exponential decay to calculate the weighted averages of features. Exponential decay emphasizes recent games while still accounting for past performances. The weights are determined using the following formula:

$$WEIGHT = \exp(-\lambda t)$$

where t is the time difference between the current game and the game being considered, and λ is the decay rate that controls how quickly the influence of past games diminishes.

2.8 Calculating Weighted Averages

To ensure that games occurring after the game of interest had no influence on the prediction, we iterated over each row and calculated each column's weighted average based on games that occurred prior to the row.

We applied the exponential decay function to each of the quantitative game statistics. For the choice of λ, we first began with $\lambda = 0.1$. As the decay rate increases, the emphasis put on more recent games also increases. During modeling, we may adjust the value of λ to evaluate model performances to find the most appropriate λ. Weighted averaged statistics were then added into the dataset as new columns, named in the format of "stat_w".

2.9 Current and Relative Results

We calculated four additional features to enhance the dataset's predictive capability: Total_Wins, Total_Loses, Win_Loss_Diff, and Relative_Result. These features provide insight into team performance trends and their historical match-ups with specific opponents. The first three features capture cumulative statistics for each team prior to the current game, which could reflect the performance of each team up to the current game in this season. Teams with more cumulative wins are more likely to win the games. In addition, to highlight whether a team tends to outperform or under-perform against their specific opponent based on past games, we added Relative_Result as a feature, which calculates cumulative wins and losses for a team against the opponent before the current game. These features address both empirical evidence and current statistics to help create a more context-aware predictive model for game outcomes.

2.10 Home and Away Games

Since the data was initially formatted to have each game listed twice (once for the home team and again for the away team) we split the data into home and away teams and join them back into a single table. Thus, each row includes the statistics for the home and away team, with the Win-Loss column representing whether the home team won that game. All variables about the present game were removed so that each column only includes information about past performance. Finally, each team's first game of the season was removed since there is no historical data on which to predict the outcome.

3 Experimental Setup

3.1 Feature Selection

The final data set of 1,215 observations includes 49 total, and 25 distinct features with 24 providing information about the home team, 24 providing the same information but for the away team, and an additional feature giving the head-to-head Win-Loss difference whose value for the home team would simply be the negative of that for the away team. Of the distinct features, 7 provide information about the Win-Loss record of each team while the rest provide on-the-court statistics (3-points made, defensive rating, etc.).

Such a large selection of features brings with it obvious concerns regarding overfitting and high correlation between statistics. However, the 49 starting features have already undergone a form of model selection, as they are nearly all regularly used in modern basketball analysis. With a game as lucrative as basketball for both successful teams and sports bettors, it is only to be expected that numerous useful statistics have been discovered. It is therefore important that the proposed selection procedure take only model performance into account and not needlessly remove useful features. Thus, our chosen procedure involves repeatedly performing 50-fold cross-validation on random forest models with variables chosen by backward selection based on importance. The results of this procedure are shown in Fig. 3. According to 50-fold cross-validation

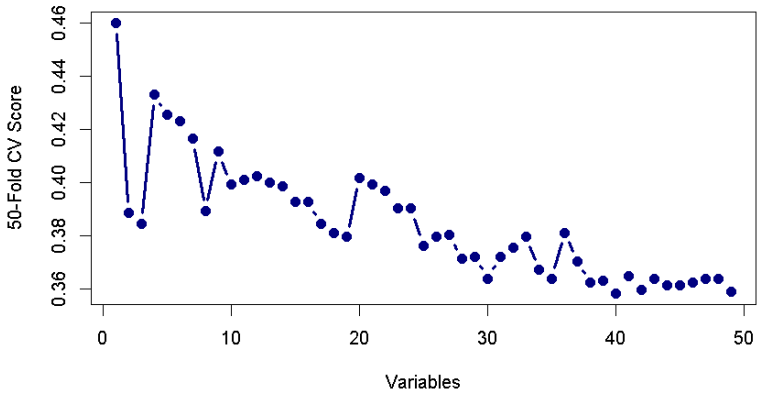

Fig. 3: Results of 50-Fold Cross Validation of a Random Forest model.

of competing Random Forest models the optimal number of predictors is 40. One of the notable features removed is points-per-game which is highly correlated with multiple other variables including points differential, expected field goals, and offense rating. In this case, it appears that the inclusion of more precise offensive data overrides the need for the general points variable. Likewise, offensive statistics such as 3-points made/allowed, assists-per-turnover, and free throw attempts were deemed unimportant when compared to the wide selection of alternative metrics.

Another interesting exclusion is the head-to-head win-loss difference which is highly correlated with the overall win-loss record. It seems that there is no special relationship between individual teams that cannot be otherwise explained by their overall performance. Thus, observed phenomena such as certain teams being uniquely good against certain opponents may simply be products of random chance. For example, it appears unlikely that an unimpressive team like the Sacramento Kings would have a 4-0 record against a slightly higher ranked Los Angeles Lakers unless they were uniquely favored against them. However, the chance of at least one team in the entire league over-performing against a higher ranked opponent is rather high, so if the statistic is only brought up when it's interesting it may create a false perception of importance. After acquiring the final model, the importance of each remaining variable was calculated using the mean decrease in Gini Purity and plotted in Fig. 4.

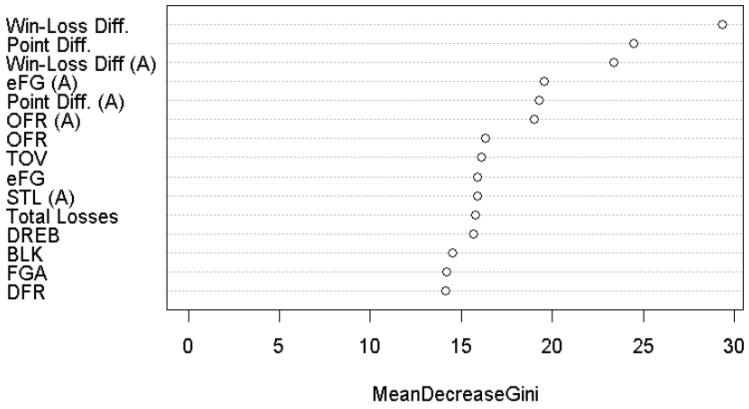

Fig. 4: Fifteen most important variables by mean decrease in Gini Purity in the Random Forest model.

We find that the home team's Win-Loss and Points Differences going into a game are the most important variables for predicting the outcome, while the away team's win-loss and points differences are the third and fifth most important, respectively. As expected, it appears that the most important factor in predicting whether a team will win or lose a game, is how many games they have won or lost in the past and how many points they won or lost by. In terms of on-field statistics, the most important are effective field goal percentage, offense rating, and number of turnovers. The only purely defensive statistics in the top fifteen are blocks in thirteenth place and defensive rating in fifteenth. Looking further into it, it appears that both these statistics are correlated with multiple higher influence offensive statistics. Thus, teams with a strong offense likely have a strong defense too, and the strength of that defense is easier to capture in statistics. It's much easier to count points scored than points not scored.

3.2 Model Evaluation

With the optimal subset of features chosen, we now check the performance of various potential models and average weightings. As discussed in Section 2.5, an exponential weighting was applied to each statistic with various exponents λ which decreases the importance of a game's statistics the further back in the season it is. A large value of λ indicates that recent performance is significantly more important, while a small value indicates that long-term performance is more informative, with a value of 0 representing an unweighted average. To evaluate each model and weighting, the data with the 40 most important features was split into training and testing sets in an 80—20 split of 972 training and 243 testing observations. The results of the testing data are given in Table 2.

λ	**RF**	**XGBoost**	**Log Reg**	**SVM**	**QDA**	**Avg.**
0	62.1	60.9	63.0	63.8	61.7	62.3
0.1	**67.9**	53.9	64.1	66.2	65.4	**63.5**
0.2	65.8	56.0	65.4	63.4	53.8	60.9
0.5	61.7	58.0	67.1	65.8	61.3	62.8
1	61.7	58.8	66.2	64.6	60.9	62.4
Avg.	63.8	58.2	**65.0**	64.2	60.7	

Table 2: Results of model evaluation for different models and exponential decay weighting.

Overall, it appears that the best models on average are those fitted using logistic regression, followed closely by the support vector machine and random forest. Additionally, we find that the weighting with the highest performance over all models is that with decay parameter $\lambda = 0.1$. Such a small decay value indicates that long-term performance is far more important than recent form. With this weighting, the past 5 games make up only about 40% of the weighted average, while the 20 games before that make up 53%.

4 Results and Analysis

4.1 Final Model

The final chosen model is the random forest model with an exponential weight with parameter $\lambda = 0.1$ applied to the weighted data. This model had the highest accurate prediction rate on the testing data of 67.9%. The confusion table of this model is given in Table 3. We see that our model has a significant preference for predicting home wins, with a predicted 59% home team win percentage compared to the 54% present in the overall data set. We further see that if the model predicts an away team win, it has a 72% accuracy rate, compared with only 64% for a home team win.

	Pred. Win	Pred. Loss
Actual Win	94	27
Actual Loss	51	71

Table 3: Confusion table of predicted outcomes against actual.

4.2 Performance Over Time

One complication of the weighted modeling procedure is that proper weights for each team generally aren't established early in the season, and thus our model cannot be expected to perform as well until later. Ordering the test data by date and plotting the cumulative prediction accuracy in Fig. 5 demonstrates this dynamic.

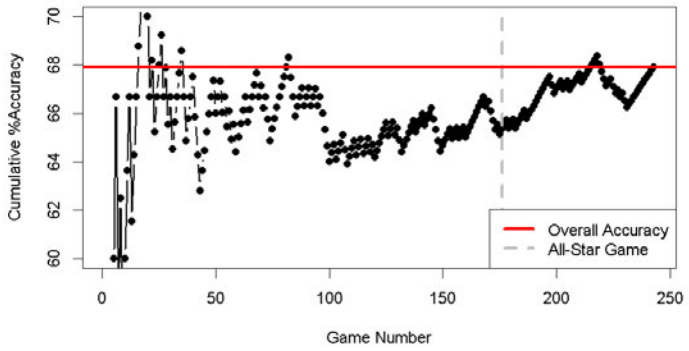

Fig. 5: Cumulative prediction accuracy rate over the ordered test data.

After some turbulence in the first 100 games of the test data set, the model's prediction accuracy is only about 64%. It then climbs to the overall accuracy of 67.9% over the rest of the games. If we only consider the 67 games in the training set which occur after the All-Star game, we find that the model has an almost 75% accuracy rate. Thus, we may assume that about 7% of the error in our modeling predictions can be attributed to a lack of weighting data, while the rest can be attributed to insufficiency in the modeling procedure and random chance. In a more robust modeling procedure, the weighting of the previous season could be used as the starting point for the next to partially bridge this gap.

4.3 Biggest Misses

We may examine where the model made its greatest mispredictions to see whether it is behaving as expected. We do this by finding the incorrect predictions in the testing data in which the model had the greatest certainty. Carrying this out gives the five largest mispredictions compiled in Table 4. It appears that all of these cases represent significant upsets of teams with highly negative win rates beating top teams

Game	Predicted Winner	Probability	Winner W-L	Loser W-L
CHI vs. MIL 11/30	Milwaukee Bucks	84.0%	6-14	13-6
SAS vs. NYK 3/29	New York Knicks	81.6%	18-56	44-29
BOS vs. LAC 1/27	Boston Celtics	80.4%	30-14	35-11
MEM vs. MIL 2/15	Milwaukee Bucks	80.0%	20-36	35-21
DAL vs. MEM 1/09	Dallas Mavericks	80.0%	14-23	22-16

Table 4: Most confident missed predictions given by the model.

in the league. Looking at the biggest miss, we see that our model failed to predict the Chicago Bulls breaking a five-game losing streak against a superior Bucks side. That upset required Nikola Vucevic to put up 29 points, his second-best performance that season, and only went to overtime thanks to an Alex Caruso buzzer beater. Thus, even when our model is incorrect, we can expect its predictions to be reasonable.

Fig. 6: Alex Caruso about to damage the integrity of our model.

4.4 Upset Performance

Our model clearly performs poorly against major upsets, but it remains to be seen whether it can predict these upsets at all. We define a major upset as a team winning against an opponent with a win-loss difference 10 points or greater than themselves. Overall, there were 135 games in our testing set in which an upset could have occurred, 35 of which actually were upsets. Our model was able to correctly predict 8 of the 35, or only 23% of these upsets. Further, it predicted 16 upsets in total, meaning that if it predicts a major upset it will occur 50% of the time. These results aren't particularly inspiring, with our model only predicting major upsets at half the frequency that they truly occur and with coin-flip accuracy. However, one team beating another despite essentially every statistic putting it as an impossibility is the worst case for

any model, so this under-performance is not too concerning. Overall, this may be more an indication of basketball's randomness rather than a true failure of the model.

One of the biggest correctly predicted upsets includes the San Antonio Spurs notching their first and only win against the Jazz that season, with a difference of 12 wins and losses between the two teams. Up to that point in the season, the Spurs outperformed the Jazz in multiple on-field statistics, especially expected field goals and offense rating, but were unable to convert these numbers into wins. The next biggest upset which our model predicted correctly comes from the below-average Atlanta Hawks beating Oklahoma City, the best team in the Western Conference that season, who at the time were looking to build on a five-game winning streak. Despite a recent poor run of form, the Hawks still looked promising in multiple offensive statistics, and thus it isn't surprising that our model would put them as slight favorites.

5 Conclusion

In this project, we successfully approached data analysis and model development for predicting NBA game results. By using various statistical indicators and team performance data, we implemented a Random Forest Model with high accuracy in predicting game results. Variables such as Win-Loss and Points Differences emerged as the most important predictive factors.

The model showed stable performance with prediction accuracy of 67.9% on testing data, reach up to 75% towards the end of the season. However, it showed limited results in predicting exceptional games, limiting itself mostly to "reasonable" predictions. This seems to be a result of the inherent variability and randomness of basketball games.

This project demonstrated that the data-driven analysis approach can provide practical value in sports game prediction, and it further suggests the possibility of further improving the performance of the model by including additional contextual information. This approach can be utilized not only in sports analysis but also in various data-based decision-making fields.

References

[1] Pomeroy, K.: The Possession. https://kenpom.com/blog/the-possession/: :text=The(2004)