Forecasting NBA Games: A model and feature set comparison

Divya Parmar

George Washington University

Abstract

This work investigates modeling and forecasting NBA games through feature engineering and model selection.

- When various model types are tested, they perform extremely similarly and return highly correlated outputs, suggesting that feature choices are more important than model choices.
- The betting spreads are found to be the most effective feature in predicting game results, displaying the skill of oddsmakers. Spread related and game outcome related measures perform better than using game-specific information.
- There is a ceiling of model performance when simply using box score statistics, and to improve there is a need to model both game-to-game player ability and opponent specific characteristics.

Literature Review

- Historically, modeling the outcome of sporting events has been done a few different ways. One way is either modeling the effect of each game action, such as a Markov chain approach. Another is to model each individual player's effect on winning through player partial effects and to combine such metrics into a larger model.
- A more common method is using production functions that focus on factors that determine the outcome of a game, such as scoring points or runs. This second method can turn to creating proxy variables for the strength of each team. Proxy metrics include Elo (taken from chess) and Massey rankings, which have been shown to have predictive power.
- One set of commonly used proxy metrics for basketball is the four factors. Introduced by Dean Oliver in his 2004 book Basketball on Paper, the four factors are effective field goal percentage, turnover ratio, free throw rate, and offensive rebounding percentage. These four factors (especially effective field goal percentage) are correlated with winning games, both in the regular season and postseason.
- Maral Haghighat, Hamid Rastegari, and Nasim Nourafza studied the application of classification models to various sports including NBA. They reviewed the use of ANN (multi-layer perceptron neural network), SVM, naïve bayes, decision trees, Fuzzy system, and logistic regression to find that accuracy ranged from 65 to 67 percent. Alexander Fayad studied similar models using such as shooting, rebounding, assists, and turnovers, finding that accuracy ranged from 68 to 73 percent.
- Zifan Shi, Sruthi Moorthy, and Albrecht Zimmermann applied predictive models to NCAA March Madness and found that MLP and naïve bayes outperformed random forest and decision trees, but they still hit a glass ceiling of around 75 percent in prediction accuracy. However, Jordan Gumm found that models such as logistic regression and a residual neural network (RNN) were able surpass the theoretical 75 percent accuracy threshold surmised by the sports betting literature, although not with regularity.

Feature Engineering and Results

Here are feature groups that were compared in analysis:

Feature Group	Features Included
Spread Only	(1) Home spread points
Spread Comprehensive	(1) Home average spread in last 10 games, (2) Away average spread in last 10 games
Spread Coefficient	(1) Home team spread-based coefficient controlling for opponent, (2) Away team spread-based coefficient controlling for opponent
Spread Weighted Coefficient	(1) Home team spread-based coefficient controlling for opponent and weighted by recency, (2) Away team spread-calculated coefficient controlling for opponent and weighted by recency
Elo	(1) Home team Elo value going into the game, (2) Home team win probability according to Elo formula, (3) Away team Elo value going into the game Note that away team Elo probability is not included, as it is simply 1 minus home team probability
Four Factors	(1) Home team EFG, (2) Home team Off Reb Rate, (3) Home team FT Rate, (4) Home team TO Rate, (5) Away team EFG, (6) Away team Off Reb Rate, (7) Away team FT Rate, (8) Away team TO Rate
Four Factors Moving Avg	(1) Home team EFG moving average, (2) Home team Off Reb Rate moving average, (3) Home team FT Rate moving average, (4) Home team TO Rate moving average, (5) Away team EFG moving average, (6) Away team Off Reb Rate moving average, (7) Away team FT Rate moving average, (8) Away team TO Rate moving average
NBA 2K video game	(1) Home team average NBA 2K rating, (2) Home team weighted average 2K rating, (3) Home team best player NBA 2K rating, (4) Away team average NBA 2K rating, (5) Away team weighted average 2K rating, (6) Away team best player NBA 2K rating
Point Differential and Rest	(1) Home team days rest, (2) Home team average point differential in all prior games, (3) Home win percentage in last 10 games, (4) Home team back-to-back game flag, (5) Home team average point differential in last 10 games, (6) Home team average point differential in last 10 home games, (7) Away team days rest, (8) Away team average point differential in all prior games, (9) Away win percentage in last 10 games, (10) Away team back-to-back game flag, (11) Away team average point differential in last 10 games, (12) Away team average point differential in last 10 home games, (13) Playoff game flag

Log-loss for each feature group and model type combination :

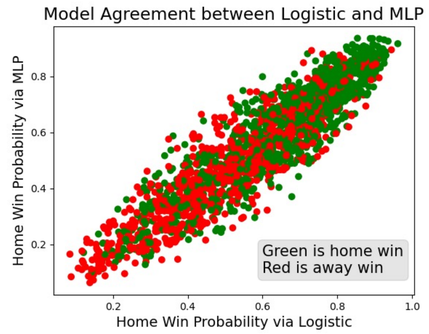
	Model Type			
Feature Set Name	Logistic Regression	Logit MLE	Naïve Bayes	MLP
All Features or N/A	0.613	0.613	1.258	0.614
Spread Only	0.599	0.599	0.6	0.687
Spread Comprehensive	0.605	0.605	0.606	0.605
Elo	0.613	0.613	0.668	0.625
Spread Coefficient	0.609	0.609	0.61	0.639
Spread Weighted Coefficient	0.607	0.607	0.608	0.617
Four Factors	0.644	0.644	0.655	0.654
Four Factors Moving Avg	0.656	0.656	0.663	0.661
NBA 2k	0.637	0.637	0.668	0.635
Point Differential and Rest	0.618	0.618	0.834	0.625

F1-score for each feature group and model type combination:

	Model Type			
Feature Set Name	Logistic Regression	Logit MLE	Naïve Bayes	MLP
All Features or N/A	0.733	0.733	0.713	0.72
Spread Only	0.737	0.737	0.737	0.746
Spread Comprehensive	0.737	0.737	0.729	0.728
Elo	0.725	0.725	0.71	0.721
Spread Coefficient	0.729	0.729	0.73	0.742
Spread Weighted Coefficient	0.73	0.73	0.733	0.717
Four Factors	0.723	0.723	0.717	0.72
Four Factors Moving Avg	0.706	0.706	0.709	0.71
NBA 2k	0.719	0.719	0.706	0.71
Point Differential and Rest	0.721	0.72	0.691	0.712

Model Explanation

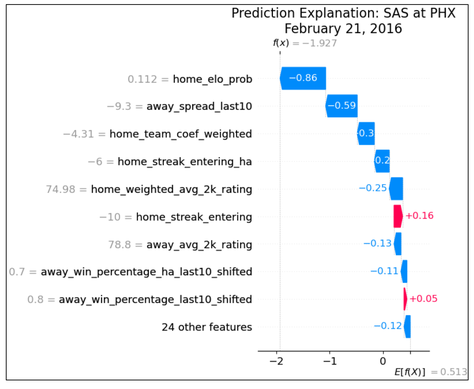
An MLP model and a logistic model trained on the same dataset returned predicted probabilities with a correlation coefficient of 0.924.



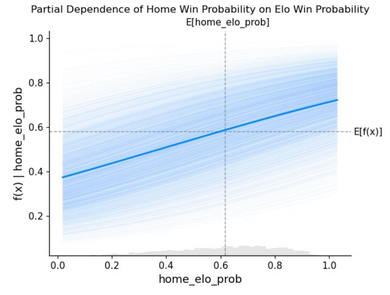
On February 21st, 2016, the model predicts an 87 percent chance of victory for the visiting Spurs over the host Suns (when the raw value of -1.927 is put into the logistic sigmoid function). This is compared to a base guess of 37 percent.

This is driven by:

- The home team's Elo win probability of 11 percent
- The away team being an average favorite of 9.3 points in their last 10 games (laying points means negative spread)
- The home team being on average 4.3 points worse than the average NBA team when controlling for opponent



By plotting the partial dependence of the target variable on Elo win probability, we can see a clear relationship.



Conclusion

- Closing spread is the best feature at predicting game outcomes, demonstrating the skill of oddsmakers using the most advance data and models. Game specific information performs worse than features created from spreads and past win/loss outcomes.
- Various model types (logistic regression, naïve bayes, and multi-layer perceptron) all come to approximately the same performance metrics given the same feature set. In addition, these models return highly correlated predicted win probabilities. This suggests that feature engineering and pre-processing are more important than model selection.
- Although this is time ordered data, fitting a time series model on points scored was not effective. Time series analysis may help with forecasting sub-components of the data, but this finding suggests that game results have much more day-to-day variance and model complexity than daily temperatures or climate (datasets commonly modelled with time series).
- With the data available here, there seems to be a ceiling of an F1-score of around 75 percent and floor on the Brier score loss of around 0.2. At this level, these models are not precise enough to use in betting, as missed classifications are very costly and there is an oddsmaker cut in priced into each bet.
- To take the next step in model performance, there is a need for data beyond simple box scores statistics. Examples of this would be:
  - Team-specific matchups (how does Team A match up with Team B)
  - Player ability fluctuations within a season as opposed to fixed value, and substituting advanced metrics such as Box Plus-Minus for subjective value like NBA 2k rating.
  - Referee effects on specific team's chance of winning.

Full Report and Further Contact

See full report on GitHub:

Connect on LinkedIn:

Acknowledgements

First, the author would like to thank Spencer Siegel of Big League Advance for his mentorship and advice through his understanding of the sports betting industry. Second, the author would like to thank Professor Amir Jafari for his teachings and encouragement throughout the research and drafting process. Last but not least, the author would like to thank Bipin Parmar for assistance with obtaining the data and ensuring data quality.