Predicting NBA Game Outcomes by Analyzing Home-Court Advantage with Data Driven Modeling

David Gourley
Luddy School of Informatics, Computing, and Engineering
Indiana University
Fishers, United States
davgourl@iu.edu

***Abstract*—**This study uncovers the impact of home-court advantage in the NBA using data-driven methods. By using historical game data starting in 1946-47 from Wyatt Walsh's "NBA Database" on Kaggle, scoring patterns and win-loss distributions for both home and away teams are explored. Using both logistic regression and XGBoost models, the research predicts whether the home team will win or lose based on different variables provided by the dataset, such as points scored by home and away teams. For the data preprocessing step, null values and outliers were removed and both feature engineering and exploratory data analysis were employed. To evaluate the models, different methods were employed such as a confusion matrix, ROC curve, and precision-recall curve. These findings offer valuable insights for teams and players to improve strategic planning, sports gamblers to improve predictions about game outcomes and progress their betting strategies, and participants in fantasy basketball contemplating on starting one player over the other, dependent on if one is playing a home game or away.

Keywords—NBA statistics, home-court advantage, predictive modeling, logistic regression, XGBoost, data visualization

INTRODUCTION

Home-court advantage is a game-changer in basketball and significantly impacts the outcomes of games. Factors that influence this are fan support, reduced travel fatigue, and familiarity with the court, but it lacks quantification. This study aims to dig in deeper to this phenomenon and give numbers to back it up instead of assumptions. By exploring trends and utilizing predictive analytics, I aim to discover insights into how home teams outperform teams on the road. Not only does this analysis validate the home-court advantage, but also underscores the power of data-driven approaches in sports strategy development.

METHODS

The dataset was acquired through Kaggle created by Wyatt Walsh and contains comprehensive game statistics from the beginning of the NBA. During the preprocessing step, the data was cleaned by firstly removing rows with null values in critical columns, such as win-loss records and "wl_home" and "wl_away". Outliers are addressed in scoring data by examining points scored by home and away teams via IQR (interquartile range) method. To ensure data consistency, outliers were excluded. Feature selection was utilized where I added temporal features "year" and "month" for trend analysis. Relevant columns, such as points scored and win-loss records for both home and away teams, were selected and used for analysis. And for feature engineering, a binary target variable "wl_home_binary" was created indicating home team win (1) or team loss (0).

Exploratory Data Analysis is the next step which features several visualizations that help understand the data better. Boxplots and scatterplots were created to visualize point distributions and relationships/trends. Both histograms and KDE (kernel density estimates) highlighted scoring trends of home and away teams. Lastly, win-loss ratios were analyzed monthly to observe different trends over time.

For predictive modeling, an XGBoost model was employed, since an XGBoost model is able to consider more complex, non-linear relationships between features. The same preprocessing steps mentioned earlier were applied here. An XGBClassifier is trained on the scaled training data. For its evaluation, the XGBoost model predicts the labels for the test sets and reveals the accuracy of the model. It provides a classification report which includes precision, recall, f1-score, and support for each class. To reveal true positives/negatives and false positives/negatives, a confusion matrix is created. An ROC curve and AUC (area under the curve) are also made to explain how well the model can discriminate between classes. And lastly, it extracts and visualizes how important different features used in the model are and how impactful they are to the final prediction.

And for my other model, I used logistic regression which was trained first to predict home team outcomes. Relevant columns were extracted to see home court advantage: team ID's, points scored at home vs away, win/loss at home vs away, and the date of each game. The dataset was then split into training and testing subsets, with training at 80% and testing at 20%. StandardScaler was applied with goal of normalizing the input features. The next step was training the logistic regression on the preprocessed dataset. And of course, the model was evaluated using accuracy, confusion matrix, classification report, and ROC-AUC score.

RESULTS

The preprocessing of the dataset included scaling the features and splitting the data into training and testing sets. XGBoost and Logistic Regression models were both trained on the processed data. The XGBoost model achieved an accuracy of 99.99%, producing 1 false positive/negative, while the Logistic Regression model produced a perfect 100% accuracy with no false positive/negatives. The classification reports for each model showed near-perfect precision, recall, and F1-scores for both classes, with very slight discrepancies in misclassifications for XGBoost.

Figure 1: Win-loss distribution for teams that play at home vs on the road. A clear difference is visible, with home teams winning significantly more.

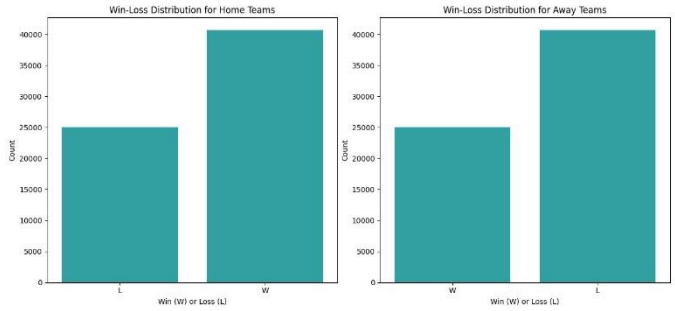


Figure 2: Points scored distribution for home vs. away teams. Similar to the figure above, home teams perform better than away teams, with them typically scoring more.

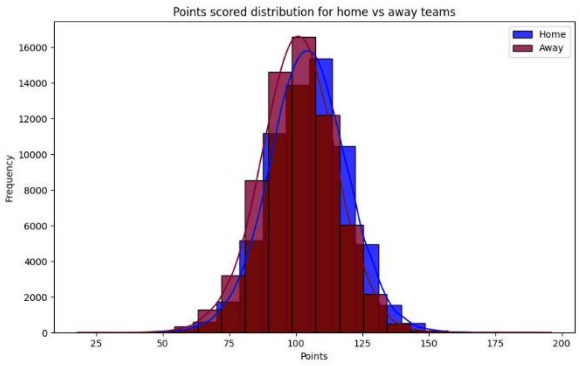


Figure 3: Monthly win ratios for home and away games. The blue (home teams) on a monthly average, win more than the red (away teams), although of course, there are a few noticable outliers that the models will remove for higher accuracy.

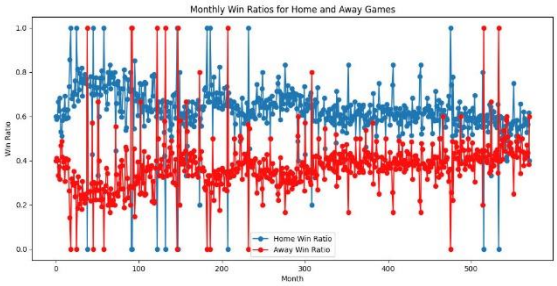


Figure 4: Monthly average points for home and away teams. The blue (home teams) consistently ranks higher than the red (away teams) meaning on a monthly average, they score more.

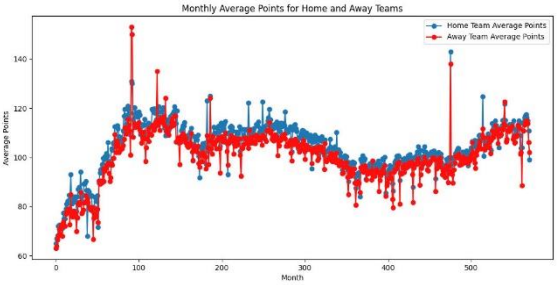


Figure 5: Confusion Matrix for the XGBoost model. There is 1 false positive/negative out of 12,838 predictions, meaning it has an accuracy of 0.999922106247079. A confusion matrix for both models is necessary as it evaluates how well the predictions from each model aligns with actual outcomes.

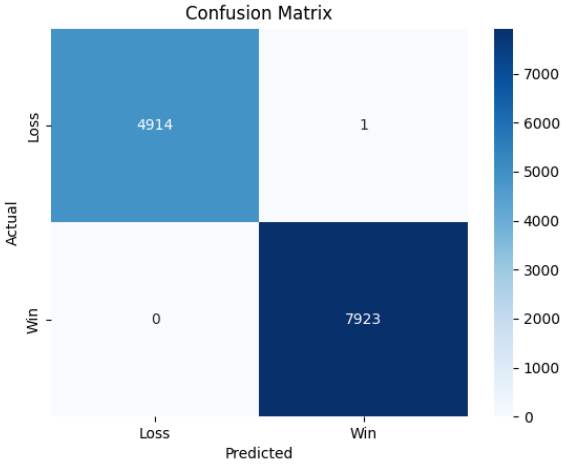


Figure 6: ROC curve for the XGBoost model. This was included in my project because it provides an important visualization that interprets the performance of binary classification models while assessing the trade-off between TPR (True Positive Rate) and FPR (False Positive Rate).

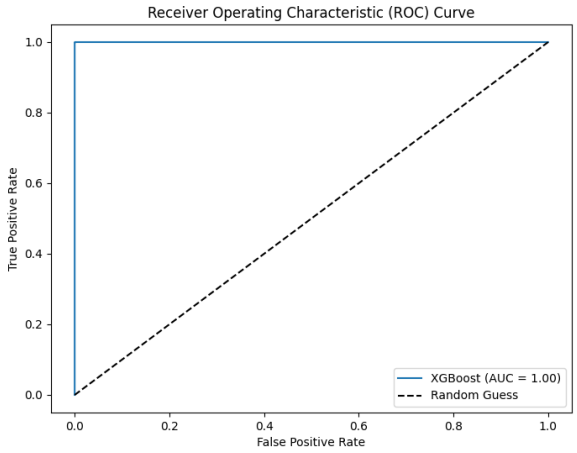


Figure 7: Confusion Matrix for the Logistic Regression model. Out of 12,838 predictions, there were 0 false positives/negatives, making it have 100% accuracy and the most optimal model, whereas the XGBoost model (Figure 5) produced 1 false positive/negative holding it back from 100% accuracy. The simplicity of Logistic Regression is likely why it performed better as it was able to capture trends without overfitting.

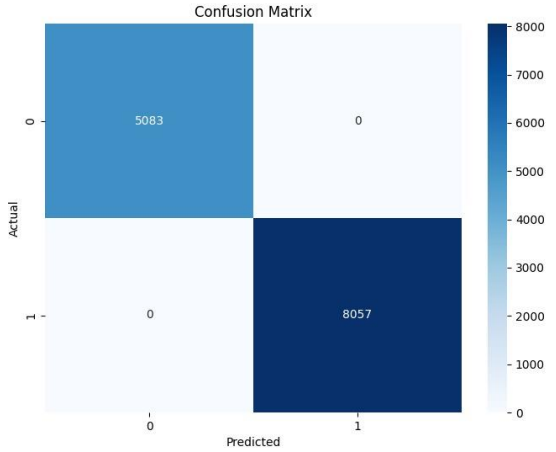


Figure 8: ROC Curve for the Logistic Regression model. Comparing this to Figure 6, we can see this ROC's curve is exactly at 1.00, suggesting the model is consistently classifying all positive instances correctly across all thresholds and not generating any false positives.

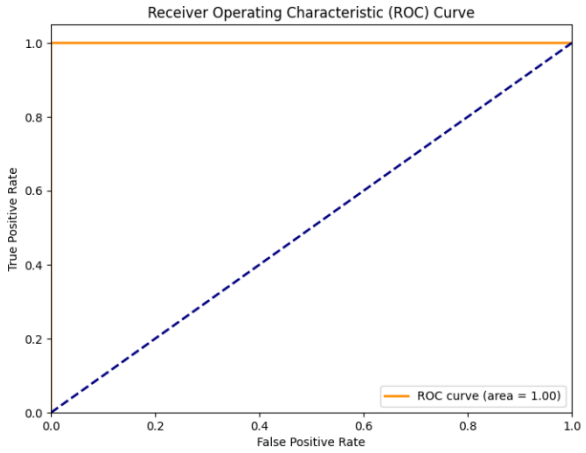
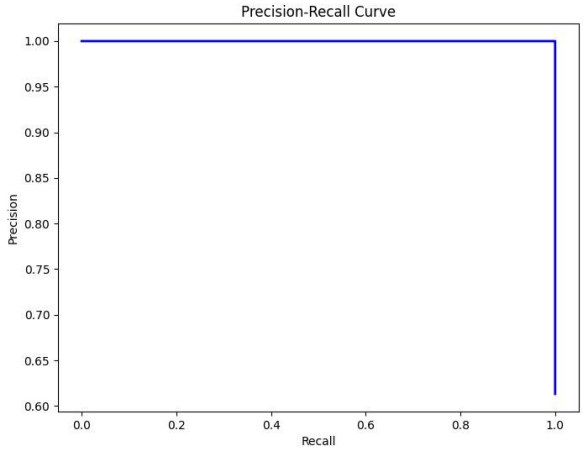


Figure 9: Precision-Recall Curve for the Logistic Regression model. Useful when focusing on the model's performance on the positive class and want to see how it balances precision and recall.



Discussion

For this project, I used both XGBoost and Logistic Regression models to examine and analyze patterns in the dataset to accurately guess whether an NBA team will win or lose when playing at home based various factors such as player statistics and team performance. It uses these patterns to make predictions on new data enabling future predictions and scenario analysis. Both models detect trends, such as the impact of home court advantage which was also analyzed before the creation of the two models and incorporates it into both the XGBoost and Logistic Regression model to enhance accuracy of predictions. Using these predictive models allowed for a better and deeper understanding of these patterns while providing a reliable method for predicting game outcomes based on home or away status.

Both models used in this project proved to be extremely accurate, with the XGBoost producing only 1 misclassification out of 12,838 predictions (0.9999%), and the Logistic Regression algorithm predicting 100%, making it a perfect model. This suggests shows that the models were extremely effective at predicting historical game outcomes, suggesting that home-court advantage is a substantial factor in determining game outcomes. With the models' near-perfect precision, recall, and F1-scores, they indicate that they classify home team wins accurately but also can handle misclassifications with minimal errors.

Exploratory Data Analysis captured insightful trends, such as the consistent advantage home teams have had over away teams historically. Visualizations of both win-loss distributions (Figure 1) and point differentials (Figure 2) showed that home teams score more and win more, supporting the idea that playing at home provides a significant advantage. And when analyzing monthly win ratios (Figure 3) and point distributions Figure 4), these trends remained consistent, highlighting the robustness of home-court advantage across different time periods.

One of the key observations during the modeling phase was that the Logistic Regression model outperformed the XGBoost model in terms of misclassifications, able to achieve perfect accuracy with no misclassifications. But, on the other hand, XGBoost only had one misclassification out of the 12,838 predictions. The XGBoost algorithm is known for its ability to generalize well and capture more complex patterns in the data, so why did the Logistic Regression model perform better? It was more than likely because the problem was simple enough for a linear model to identify the patterns with ease. The one misclassification in the XGBoost model holding it back from a 100% accuracy could because of the XGBoost's inherent complexity and its sensitivity to subtle details in the data. It's likely that this model overfitted to some minor nuances in the dataset or was impacted by small random variations in the data, while Logistic Regression, being simpler and less sensitive t details like that, was able to classify all instances correctly.

These findings from this project offer practical insights for a variety of people. For one, NBA teams and coaches can use this information to adjust game strategies based on home and away games. Sports bettors can also sharpen their predictions by factoring in home-court advantage. And lastly, fantasy basketball users might consider the home game status of players when making crucial lineup decisions, as the home-court advantage can and most likely will influence individual performance.

Despite the incredibly high accuracy achieved in this project by the two models, there are areas for improvement and to look deeper into. Additional factors that might influence game outcomes, such as roster changes and player injuries, could be incorporated to make more accurate predictions. Also, using player-specific performance metrics and more advanced statistics (such as RAPTOR, LEBRON, etc) could enhance the models further. Furthermore, experimenting with more complex models, such as Neural Networks, might be able provide deeper insights into the dynamics of game outcomes.

Comparing this project to other work in sports analytics, such as studies on player performance or overall game predictions, I aimed to focus on home-court advantage specifically. Few studies have tackled the concept of home-court advantage and its correlation to team performance using predictive modeling in the NBA, and instead focus more on team performance and individual player stats only. This project confirms the impact of home-court advantage and underscores the power of data-driven methods in sports analytics.

Future research could expand this model by including more advanced data, such as advanced player performance metrics and team-specific dynamics like bench strength and coaching strategies. Testing the model with more diverse datasets, such as different basketball leagues in Europe or Asia, could offer insights into how home-court advantage compares across different contexts. And lastly, exploring deep learning techniques such as RNNs (Recurrent Neural Networks) might allow for the modeling of temporal patterns, further improving prediction accuracy with more context and factors.

Author Contribution Statement

The author conducted this study independently. All aspects of the study, from data collection, preprocessing, feature selection, model implementation, hyperparameter tuning, and results analysis were conducted and completed by the author. The use of AI was not used in this project.

References

[1] C. M. Miller and L. Bornn, "Factorized point process intensities: A spatial analysis of professional basketball," *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1797–1806, 2019. DOI: 10.1080/01621459.2018.1537928.

[2] J. Brownlee, "Develop your first XGBoost model in Python with scikit-learn," Machine Learning Mastery, 2016. [Online]. Available:

https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/

[3] D. Cervone, A. D'Amour, L. Bornn, and K. A. Goldsberry, "A multiresolution stochastic process model for predicting basketball possession outcomes," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 585-599, 2016. DOI: 10.1080/01621459.2016.1141685.

[4] F. Caruana, "XGBoost parameter tuning*," Kaggle Kernel*, 2019. [Online]. Available: https://www.kaggle.com/learn/advanced-machine-learning

[5] J. Brownlee, *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End*, Machine Learning Mastery, 2016. [Online]. Available: https://machinelearningmastery.com/machine-learning-mastery-with-python/

[6] L. Bornn, S. Goldsberry, D. Cervone, and A. D'Amour, "Beyond points per game: Predicting basketball player impact with tracking data," in *MIT Sloan Sports Analytics Conference*, 2015, pp. 1–10. [Online]. Available: https://www.sloansportsconference.com/

[7] M. G. Hughes, R. Sharda, and M. J. D. A. Saluja, "Data preprocessing for predictive modeling: A case study with sports data," *International Journal of Data Science and Analytics*, vol. 10, no. 3, pp. 257-275, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s41060-020-00213-7

[8] N. G. Psarakis, P. Fitsilis, and N. Tselikas, "Prediction of NBA players' performance using supervised machine learning algorithms," in *Proceedings of the 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2018, pp. 1–6. DOI: 10.1109/IISA.2018.8633660.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207–216. [Online]. Available: https://dl.acm.org/doi/10.1145/170036.170072.

[10] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, 3rd ed. OTexts, 2021. [Online]. Available: https://otexts.com/fpp3/

[11] S. Shmueli, "Predictive Modeling: Evaluating the Performance of Your Models," *Journal of Business Analytics*, vol. 30, pp. 218-235, 2022. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/21674711221022642

[12] Walsh, W. O. (2023). *Basketball Data* [Data set]. Kaggle. Available: https://www.kaggle.com/datasets/wyattowalsh/basketball