Executive Summary -Predicting NBA Outcomes with Different Methods

By Keegan and Subhan

Background:

Many sports are a game of inches, meaning that predicting the outcome of a game is immensely difficult. Where one dribble, shot, steal, or dunk can change the outcome of the game, basketball is no different. Not only is predicting the outcome of a game challenging, but it is also essential to many, especially with the rise of sports betting when people are putting millions of dollars on the line.

Methods:

Before starting our project, we spent lots of time data wrangling to ensure we would not get any leakage, meaning that we were predicting games based on the stats from the current game specifically whether or not the home team loses. To combat this, we made five-game and season averages for every variable we were interested in testing with a lag to ensure current game stats were not used to predict games.

In our project, we attempted to tackle the problem of predicting the outcomes of NBA games based on past stats. To do this, we used three different types of models, including Logistic Regression, Random Forests, and Gradient Boosting.

We used Logistic Regression because the model is intended to be used for binary outcome predictions, making it suitable for predicting the outcomes of NBA Games by analyzing relationships between various variables.

We also used Random Forests, which were suitable for their ability to reduce the risk of overfitting while increasing the model's accuracy. The other benefit of random forests is their ability to consider interactions between variables through the construction of the trees.

Lastly, we used gradient boosting which we thought could be good due to their strong ability to capture relationships between variables.

Results:

In evaluating the performance of three predictive models—Logistic Regression, Random Forest, and Gradient Boosting we observe distinct outcomes based on their respective metrics. The Logistic Regression model demonstrated a decent ability to distinguish between classes with a ROC of 0.7159, alongside a high sensitivity rate of 0.7996, indicating it is good at identifying away wins. However, it scored lower in specificity (0.4735), telling us the model struggles in correctly identifying away losses.

The Random Forest model, with an accuracy of 0.6630 at its best setting (mtry = 7), showed a moderate performance compared to the other models. The random

On the other hand, the Gradient Boosting model surfaced as the superior model in terms of accuracy (0.6761) and the highest sensitivity of 0.8700. Despite this, similar to the Logistic Regression model, it struggled with low specificity (0.3952), suggesting a tendency to misclassify away losses.

Conclusion:

Overall, while Gradient Boosting leads in accuracy and sensitivity, it, alongside Logistic Regression, faces challenges with specificity. Random Forest, while the least accurate, presents a balanced profile that might be improved with further tuning. The selection between these models should be guided by the consideration of the trade-offs between accurately detecting true away wins and avoiding false positives for away wins. However, all of the models could use more tuning.