Jay Kwon | Metis Data Science and Engineering (flex program) | Module 4 – Classification (March, 2022)

**NBA Home Team Game Outcome Classification**

**Abstract**

The goal of this project was to use classification models to predict whether the home team wins upcoming NBA games in hopes of turning a profit when wagering on the games. I worked on a [dataset obtained from Kaggle](https://www.kaggle.com/nathanlauga/nba-games?select=games.csv) which had historical NBA game outcomes (binary target) and major statistics of each game for the past 18.5 NBA seasons. Using the previous 10 game averages of the statistics and win-rate, I fit logistic regression, random forest, and gradient boosted tree models and assessed its performance by comparing its accuracy, precision, and ROC AUC curves. After refining the logistic regression model to optimize precision, I assessed the viability of the model for profiting on NBA wagers. At a 0.5 threshold, the accuracy of the model was 65% with an AUC of 0.67. Profitable scenarios were extrapolated for when the threshold was high. At a threshold of 0.88, we need to be offered greater than 0.136 to 1 on our wager to be profitable. The win-rate history was the most promising feature to predict NBA game outcomes, while the previous average statistics of the game proved to have little predictive power.

**Design**

Sports betting allows us to wager on the outcome of NBA games (win or lose) to potentially turn a profit if one is successful in consistently beating the payout odds being laid. Anyone interested in the NBA, sports betting, and making money would benefit from this project. If the model is precise enough to “beat the odds”, then it may be deployed to generate profits. The interpretation of the model will reveal insights on important features and how to improve the model. I was most concerned about the precision of the model given I’d be using the model to wager on the home teams and the penalty for being wrong affects whether the model would be profitable.

**Data**

The raw data with historical game outcomes and statistics were obtained from a [dataset on Kaggle](https://www.kaggle.com/nathanlauga/nba-games?select=games.csv). The statistics for each game were points, rebounds, assists, FG%, 3PFG%, FT% for home and away teams. Rolling previous 10 game averages were derived for the statistics for all 30 teams. A final features DataFrame was created to include these averages for each game/row for the home and away teams. 4 sets of the 6 averages were created filtered based on home team games only (for the home team), away games only (for the away team), and all games for both teams. Additionally, the previous 10 game win-rates were added as features for both the home and away teams. The final DataFrame, had 24526 rows, each representing a unique game, and 26 features for each game. The data spanned 18.5 NBA seasons dating from October 2003 to November 2021.

**Algorithm**

- empirical probability: 0.59

- Logistic Regression:

test accuracy = 0.65, precision = 0.67, AUC = 0.67

- Random Forest: 0.63

test accuracy = 0.63, precision = 0.69, AUC = 0.64

- Grad Boosted Trees: 0.64

test accuracy = 0.64, precision = 0.66, AUC = 0.66

- At threshold of 0.88 the logistic regression model has a precision of 100%, but the low recall led to only 14 true positives out of over 24,000 games.

**Tools**

Sklearn to fit model, split data, evaluation metrics

Pandas to filter and manipulate data

Matplotlib and Seaborn for visualizations

**Communication**

- 5 minute presentation with Slides

- GitHub repo: https://github.com/jaykwon2/NBA\_outcomes\_classification