Jay Kwon

Metis Data Science and Engineering (flex program)

Module 4 – Classification (March, 2022)

Minimum Viable Product (MVP)

**NBA Home Team Outcome Classification**

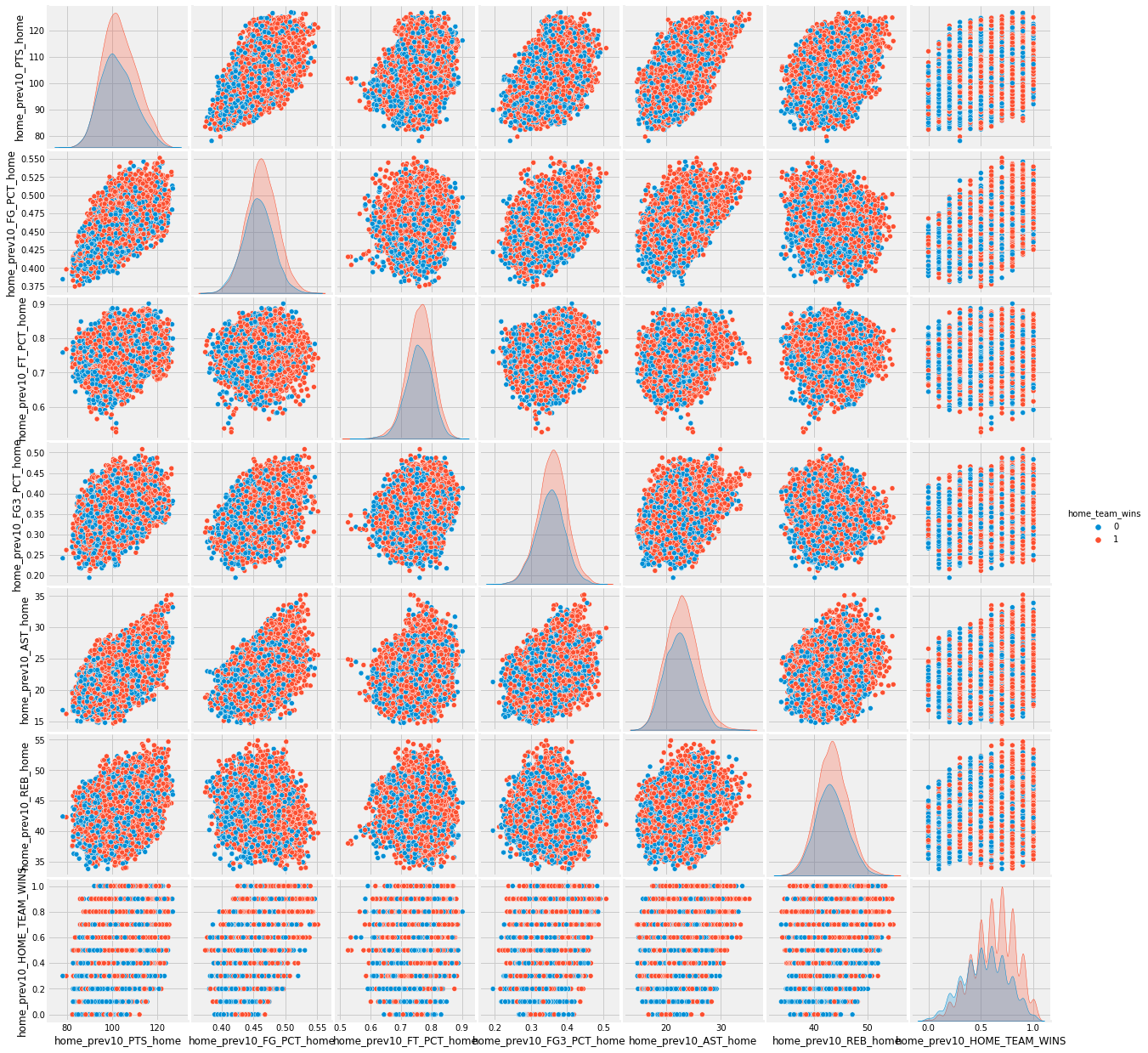
- **Goal:** The goal of this project is to create classification models that predict the outcome (win or lose) of the home team for NBA games, using historical NBA data including statistics for each game (points, rebounds, assists, FG%, 3PFG%, FT% for home and away teams). Different classification models will be assessed and compared.

- **EDA/Data Cleaning:** After some EDA, missing and inaccurate data points were appropriately removed, and the dataset was filtered to represent a complete history of NBA games for the past 18.5 seasons, including preseason and playoff games. Duplicate rows were also removed. The dtype of the estimate game date column was changed to datetime and the dataset was sorted in chronological order.

An important revelation was that the empirical probability (“baseline” probability) for the home teaming winning was 59%. This isn’t a very big discrepancy in distribution and class imbalance should not be a big problem. As a first approximation we are aiming for our models to have an accuracy greater than 59%, however, we are more concerned about precision. This is because if we were to wager/bet on the home team winning, we are concerned about how many of our picks are correct.

- **Feature Engineering:** Rolling previous 10 game averages were derived for the 6 major 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, has 24526 rows, each representing a unique game and 26 features for each game.

(continued on next page…)



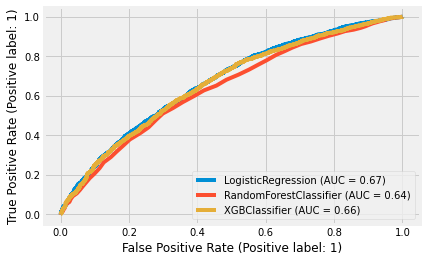
Above is a pair-plot of 7 of the features for the home team. The 6 main statistics were filtered to home games only. The previous 10 home games win-rate includes all games (home or away). Unfortunately, we don’t see very good “separation” between the outcomes for the statistics features. The win-rate feature looks more promising. Looking at the bottom right corner panel, we see that the distribution for the positive outcomes (red) is centered to the right of the distribution for negative outcomes (blue). If the previous 10 game win-rate was higher, the home team was more likely to win.

- **Preliminary Modeling:** Preliminary models were fitted on all of the features for logistic regression, random forest, and gradient boosted trees. The test accuracies for these models were:

🡪 logistic regression: 64%

🡪 random forest: 63%

🡪 gradient boosted trees: 64%



Above, the ROC AUC curves indicates that the logistic regression and random forest models are slightly better (more towards the top left).

- **Future Steps:**

🡪 add cross validation

🡪 tweak hyperparameters and add/remove features to optimize precision of the model

🡪 follow up AUC analysis

🡪 relate findings to practical application of model

(i.e. wagering on games and generating a profit)