Capstone Project: Prediction of NBA Game Winners

Problem to solve:

NBA is the second most popular sport in North America. Being able to predict game result is important for general managers of each team especially when it's closer to the end of a season. They can use that to decide if they have a shot to the playoffs or if they can rest their star players to avoid them getting injured. And for teams that are making it to playoffs, it's very important for them to predict game results to see which seeds they are likely to be. Also, NBA attracts many people to bet on the game results, so the sport betting market is quite huge. Therefore, it's interesting to know if we can find a way to beat the house.

Approach:

Using game states from the games already played by each team to build a model to predict the game result

Findings:

The highest accuracy of my model is at 68.4% comparing with the baseline at 55.0%. Both svc and logistic regression models tied at that percentage. Hyperparameter tuning didn't help on accuracy. Also, selecting more import features to build a new model didn't help either.

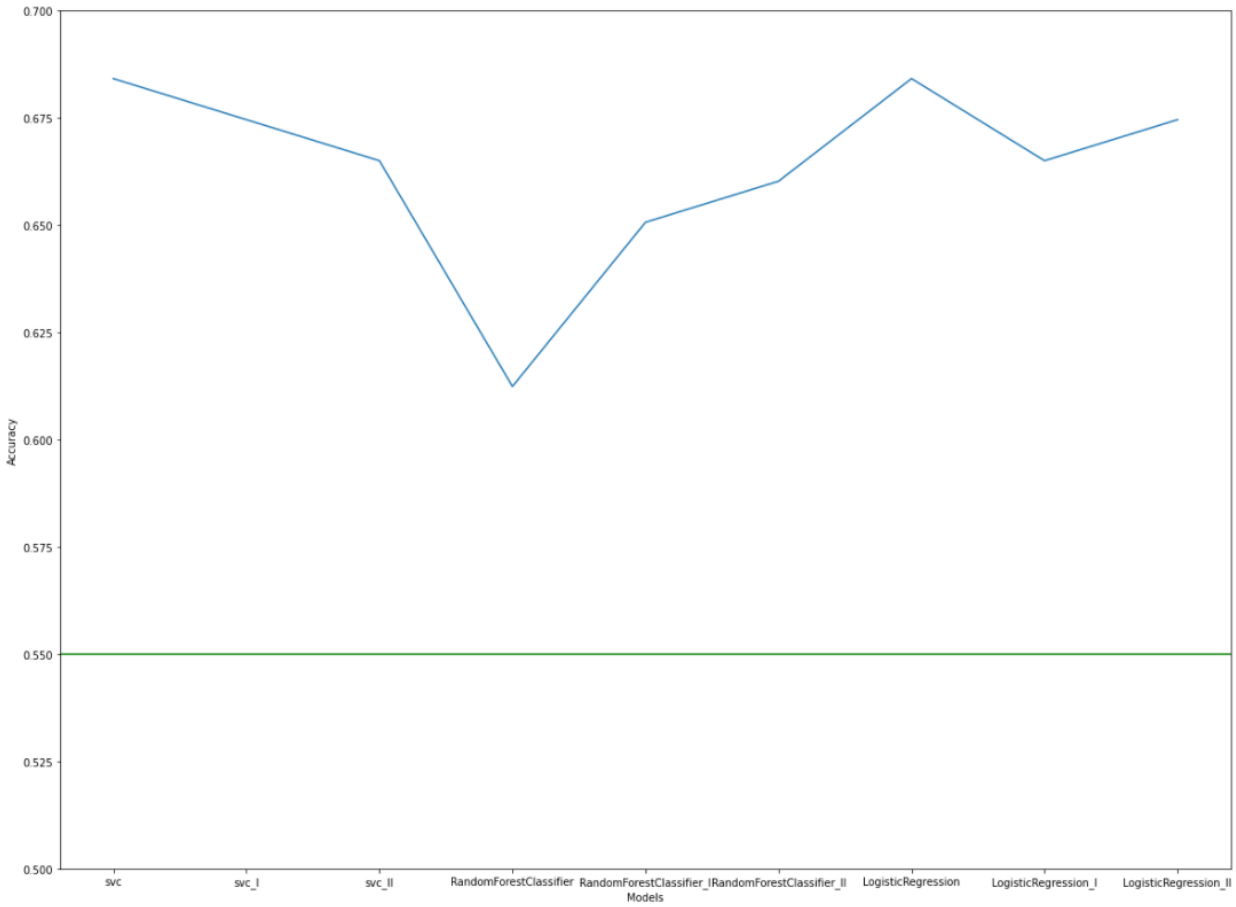

Fig 1: Model Accuracy: The column names ended with "_1" represented the result of hyperparameter tuning using grid search method. The row index ended wtih "_2" represented the result of using top 10 most important feature , selected through the method of permuation importance in "_1".

As for which features are more important, there are three features that worth noticing:

1. **season_PIE_Home**: It is significantly more important than other features in logistic regression model and it is also top 5 in svc and random forest method.

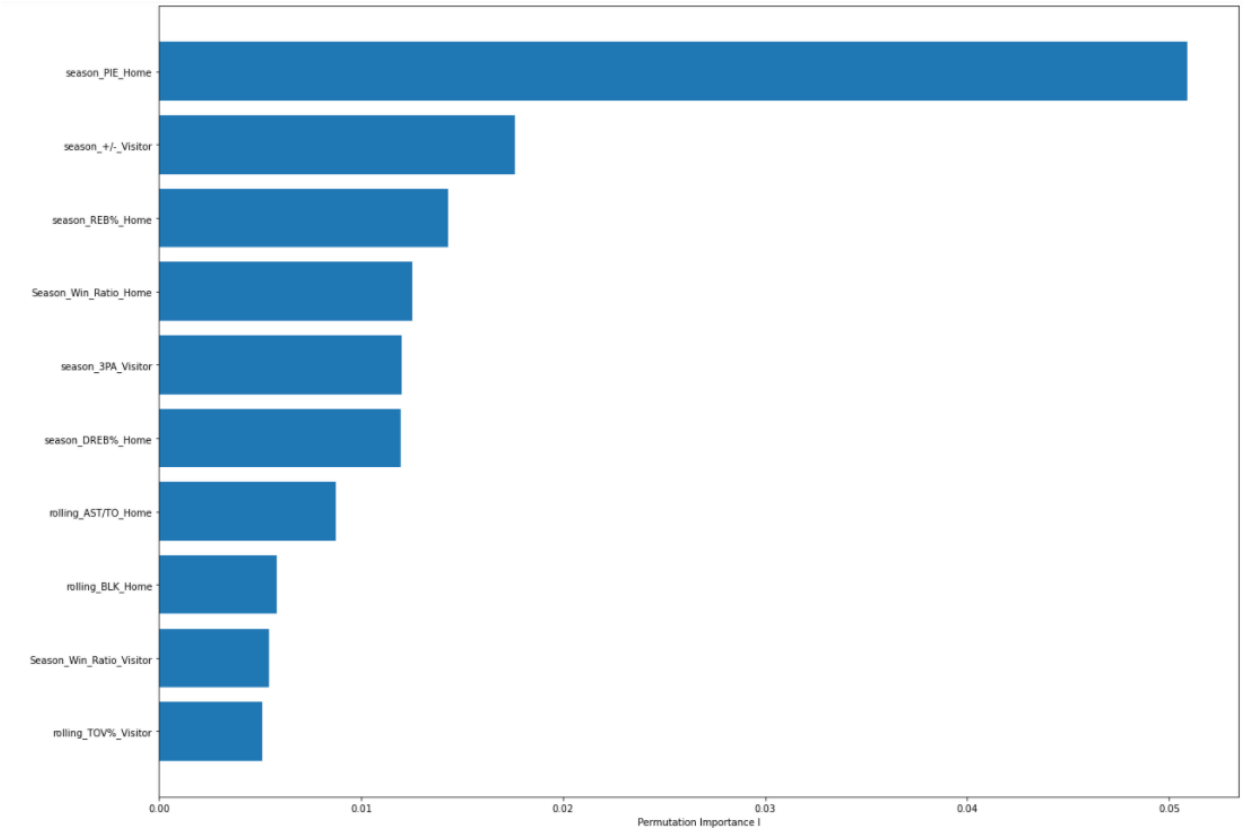

Fig 2: Permutation importance in logistic regression I

2. **season_Win_Ratio_Home:** The most important feature in random forest model and it is also top 7 in 2 other methods.
3. **season_DREB_Home**: The most important feature in svc and it is also top 4 in random forest method.

Furthermore, “home” stats are more important than “visitor” stats, and “season” stats are more important than “rolling” stat (rolling window set at 5).

Recommendations for clients to use the findings:

1. Season’s stats are more important comparing with rolling stats, so for a team to predict their game result, they should look at the whole season’s stats instead of recent stats.

2. Home team's stats are more important than visitor team's stats. And therefore, when a team play a home game, whether they win or not are more dependent on their own stats instead of their opponent's.
3. The best accuracy I could get was still less than 70%, which means historical data still couldn't explain game result more than 30% of the time. Players and coaches should get frustrated knowing the odds are against them because there are still rooms to turn things around. And people who put wager on game should pay closely attention on the first half of the game, they may consider bet on a different team at half time because if players are playing significantly different from their previous games, they are likely to change the game result.

Ideas for further studies:

Future studies should look at each player's stat as well. Rooster can vary from game to game due to injury or other reason and that should be put into consideration. And players combination seems to matter as well, different players combination can contribute significantly different depend on whether they have good or bad atmosphere. Furthermore, officials matter sometimes because the way they officiate games may lead to different game result.