Brett Gordon

Final Report:
NBA Game Recommendation Analysis

1. Introduction

Problem Statement

With streaming becoming a larger and larger service over time, it is more important than ever to determine what services people want to watch as the competition becomes more fierce. Our goal is to help our client determine the best NBA game to stream for a premium price for each night of the NBA season.

Background

As of September 2022 there are about 113 million streaming services in the United States and studies show over a quarter of Americans prefer watching sports over a streaming service. New competitors need to come up with strategies to make themselves stand out from the crowd.

One such strategy is giving higher quality stream for what is expected to be the best game of any given night for an affordable price. In order to do so, the service needs to have a good strategy on how to predict what the best game will be, otherwise they will get lost in the crowd quickly.

Goal

Our goal is to look into the ways to determine what will be the best game for any NBA fan to watch. We are looking to base this on both teams playing in the game and how close we expect that game to be. We are going to use data from the 2020 NBA season as a baseline for this information and in the future hope to expand back to 2014 and up the current date.

2. Data Wrangling

There were two datasets used for this project, the first being an SQLite collection (https://www.kaggle.com/datasets/wyattowalsh/basketball) that collected data about individual games dating from 1946 to 2023. This dataset includes over 64,000 NBA games, 4,800 players and all 30 teams. This collection had the following datasets: Player, Team, Team Attributes, Team History, Player Attributes, Game Officials, Game Inactive Players, Team Salary, Player Salary, Draft, Draft Combine, Player Photos, Player Bios, Game, News, News Missing. This is a total of 16 datasets but fortunately a lot of them were irrelevant, such as anything related to team information or draft history. The datasets we used were: **Team, Player, Game and Player Attributes.**

The second dataset (https://www.kaggle.com/datasets/visalakshiiyer/nba-match-data) contained information about teams individual players from 2014 to 2020, broken down by each season. This information contained advanced metrics such as RAPTOR and WAR. For this case we only needed the modern_RAPTOR_by_team.csv.

Eventually we compiled the data into one dataframe that contained information from the 2020 NBA season. The games we chose were only from teams that had played at least 40 games that season as the high variability of win percentages in the early stages of the season can lead to unreliable data. We also focused on players with higher advanced metrics as these are the exciting players that fans normally prefer watching. The primary team data points we used included: Point differential, team winning percentages and total number of points scored in a game. For individual players, the data points we focused on were typical of the standard statistics: points scored, assists and rebounds. The shape of this final dataset was 5,417 rows and 30 columns.

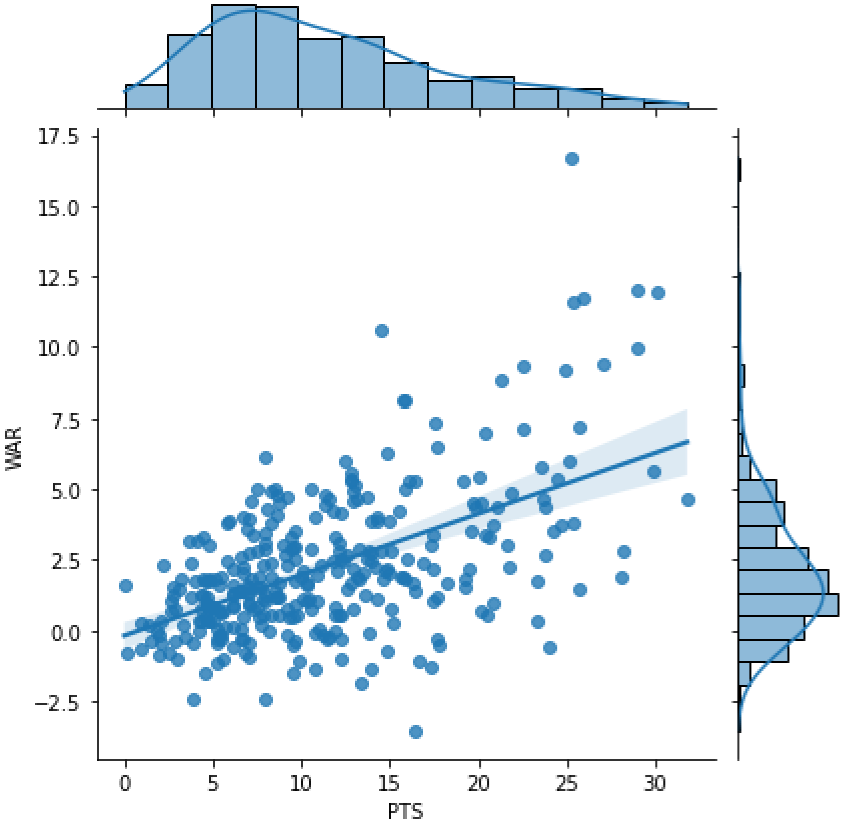


Figure 1. Individual Points per Game vs Individual WAR, frequency associated bar graphs.

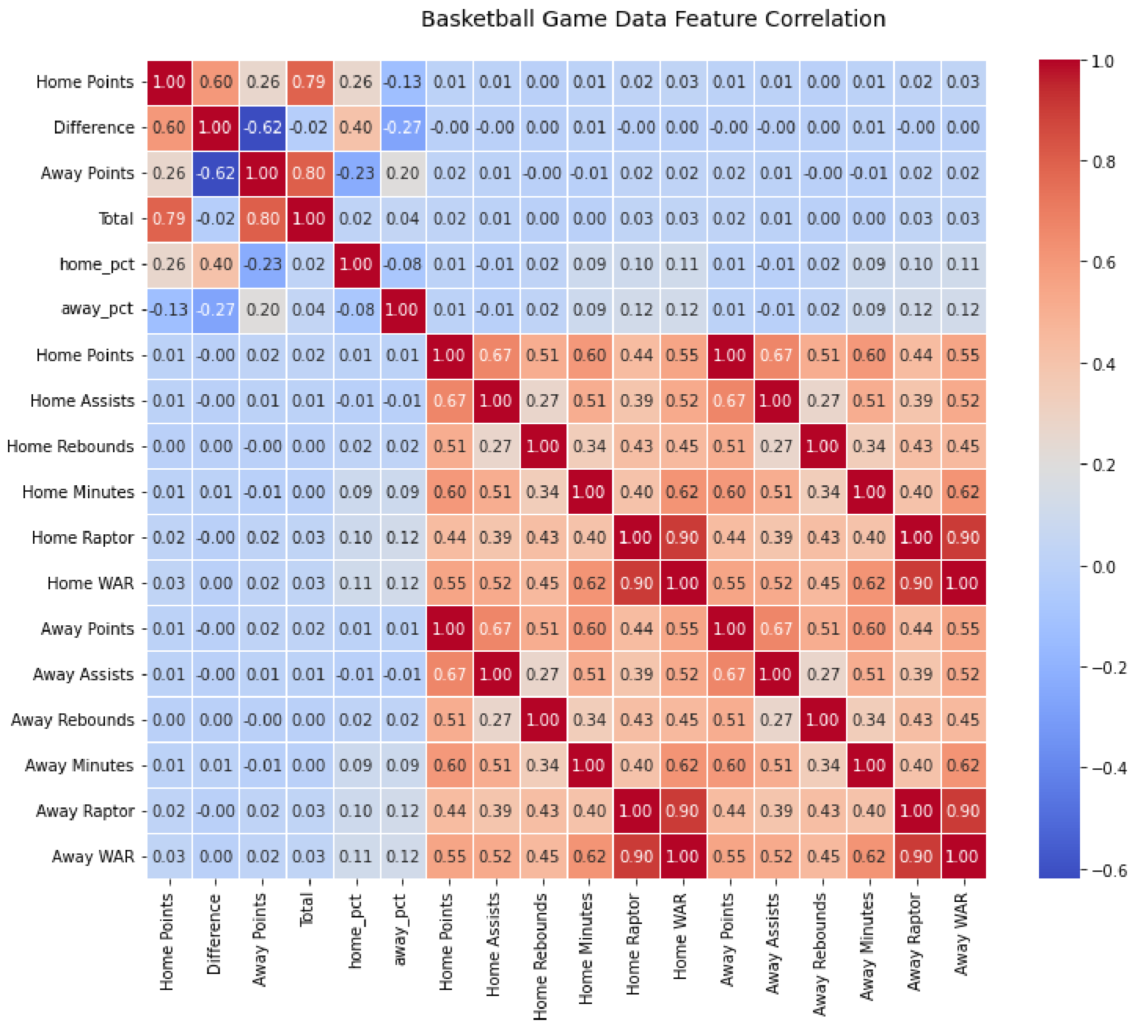


Figure 2 Heat Map exploring Team and Individual Data Points.

Exploratory Data Analysis and Modeling

With the given data it was confirmed that there is a correlation for points/assists/rebounds and RAPTOR and WAR. This is as expected considering RAPTOR and WAR are advanced metrics used to determine how valuable a basketball player is in the game. There was also a noticeable correlation between home winning percentage (home_pct) and point differential (winning team total points minus losing team total points) which alludes to the idea that teams tend to win more at home than on the road.

As you can see in Figure 2, there is a large correlation in most of the individual player statistics but they have no real correlation to the team statistics. This is determined to be due to the fact we chose the players with the higher advanced metrics and these tend to be top tier players that can score points and get assists and rebounds along with playing the most minutes.

Eventually, the Home Points and Away Points were discarded as they are directly correlated to Total Points and Point Difference and these would skew the models. With the models, the focus shifted to having the Point Difference be the dependent variable. Everything that happens in a game leads to the final score plus viewers will likely want to watch games that are closer and more exciting in the final minutes of the game.

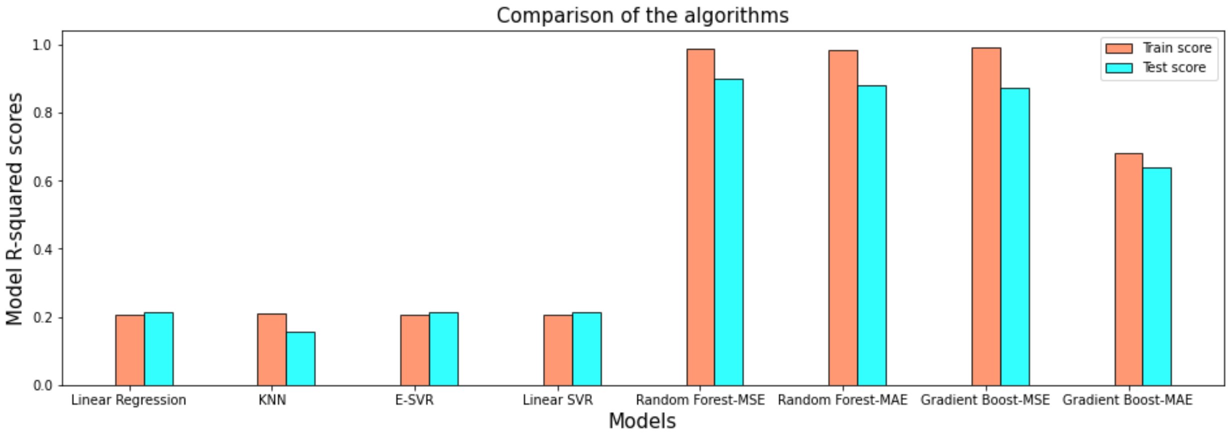


Figure 3: Train and Test scores of different model algorithms

Using eight different algorithms, we used GridSearchCV to determine the best fits for whatever models it was applicable to. As you can see in Figure 3, the Linear Regression, K-Nearest Neighbors, Epsilon-Support Vector Regression and Linear Support Vector Regression models scored very poorly and could be discarded as the focus turns to Random Forest and Gradient Boost.

With Random Forest we evaluated both Mean Squared Error Criterion (MSE) and Mean Absolute Error Criterion (MAE). We found the best fit was for n_estimators equal to 60 and minimum samples for a leaf to be just 1. Using these we found the following data:

Mean Squared Error

R^2 (train): 0.9888
R^2 (test): 0.9017
MSE (train): 2.5847
MSE (test): 20.9853
MAE (train): 0.9673
MAE (test): 2.5852

Mean Absolute Error

R^2 (train): 0.9858
R^2 (test): 0.8814

MSE (train): 3.2821
MSE (test): 25.3088
MAE (train): 1.1694
MAE (test): 3.1083

Between the two of these the MSE criterion had the better train and test numbers without it seeming to overfit.

In regards to Gradient Boosting we had to fit the MSE and MAE to separate models, one with the loss for absolute error (default) and one for squared error). For the MSE we found the following best model and training and test results:

Mean Squared Error

Estimators: 100
Minimum Samples for Leaf: 4
Learning Rate: 0.5
Subsample: 1
Max Depth: 9
Max Leaf Nodes: 25

R^2 (train): 0.9926
R^2 (test): 0.8732
MSE (train): 1.7156
MSE (test): 27.0632
MAE (train): 0.965
MAE (test): 2.9851

For the MAE model in Gradient Boosting, we got a different model fit and found considerably worse numbers for the training and test sets:

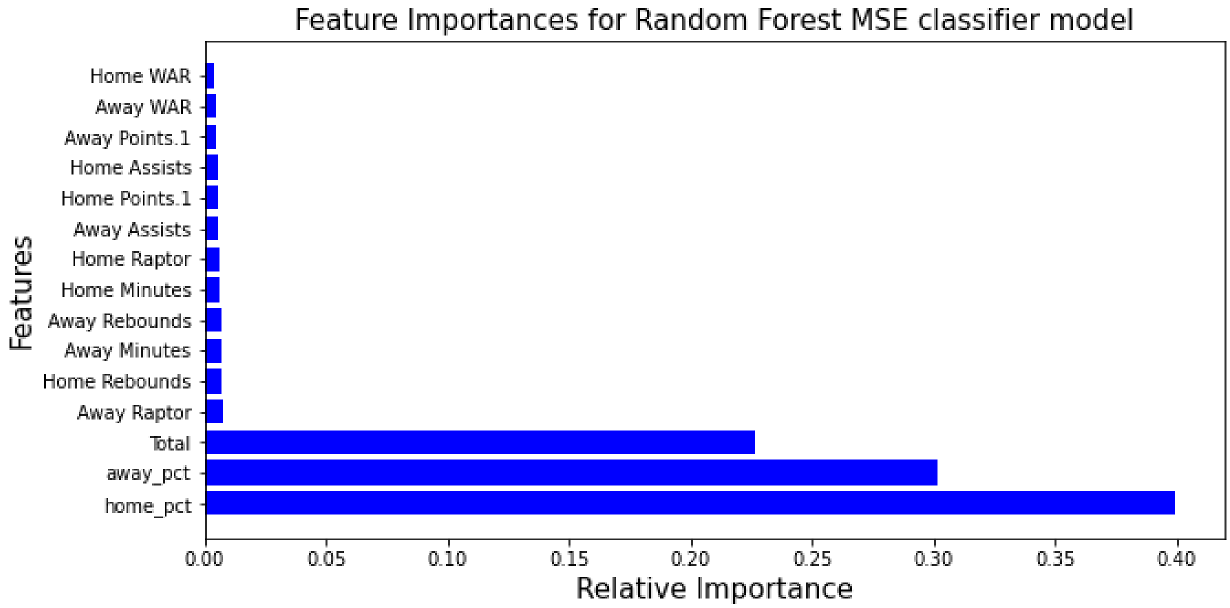
Mean Absolute Error

Estimators: 100
Minimum Samples for Leaf: 4
Learning Rate: 1
Subsample: 1
Max Depth: 9
Max Leaf Nodes: 25

R^2 (train): 0.683
R^2 (test): 0.6391
MSE (train): 73.1548
MSE (test): 77.0319
MAE (train): 4.6939

MAE (test): 5.1858

With this information we chose to look at the feature importance in the Random Forest MSE and Gradient Boosting MSE models as they seemed to have the best scores.



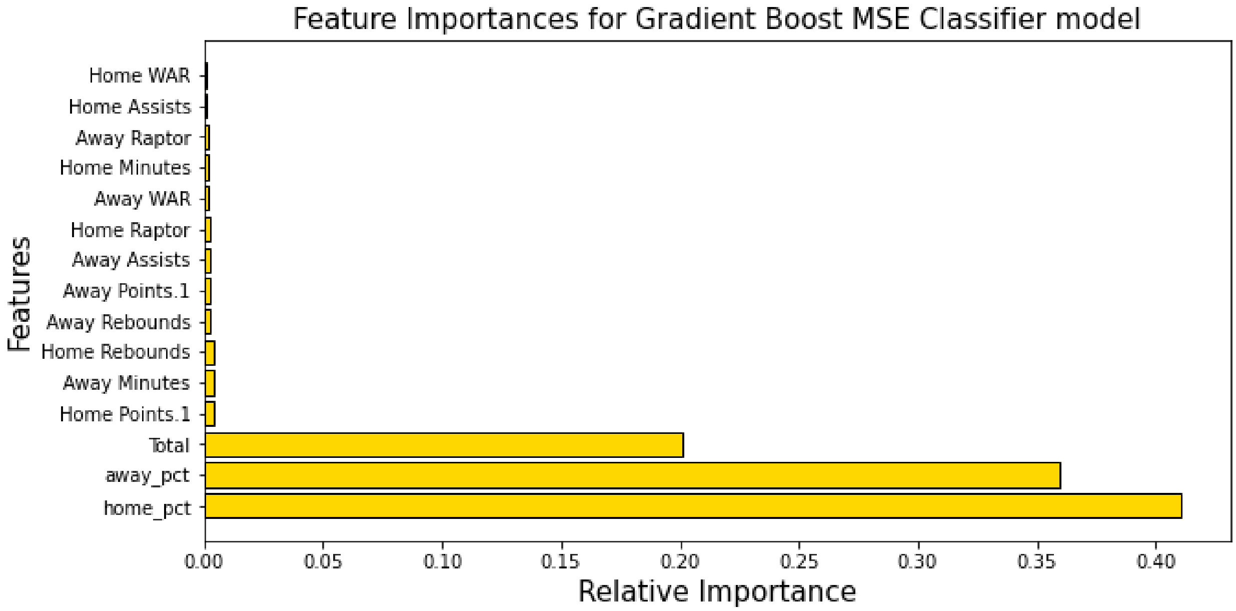


Figure 4 and 5: Feature Importance for Random Forest and Gradient Boosting MSE models

As we can tell from figures 4 and 5, home team win percentage seems to be the most important feature of point differential, followed closely by away team win percentage and point total. These should be used as the critical points going forward to determine what game should be chosen for the premium price each night. It is important to note that when choosing the game, the home and away win percentages are already known but the point total is predicated on how strong each teams offense and defense is leading up to that game.

Future Research

In future iterations, the plan is to run models for different dependent variables and weight them as point differential isn’t the only indicator of what viewers want to watch. Some viewers focus on how good both teams are or if there are multiple superstar players in the game. Sometimes viewers will just want to watch a high scoring, exciting game too as there are a lot of factors that can go into what is the most popular game for any given night.

It would also be of our best interest to expand the dataset to beyond the 2020 season as that would give us a greater sample size and also potentially find trends. Some of these trends could include time series analyses as certain games play bigger roles as the season comes to a close, such as games between two teams fighting to get into the playoffs.