NBA Predictions

Lee Richardson, Daren Wang, Xiaofeng Yu, Chi Zhang

Carnegie Mellon University

Introduction and Goals

- The goal of this project was to predict the outcomes of NBA basketball games as accurately as possible
- Using game predictions, we can create a distribution of how many games each team is expected to win over the course of a season and compare with other projection systems.
- We also looked at which features were the most important in terms of accurately predicting games.

Data

Data collection and processing was a large portion of our project. We used web crawlers in the R and Python languages to extract data ftom three seperate data sources:

- ESPN.com
- Basketball-Reference.com
- stats-for-the-nba.appspot.com

And then used ETL and merging techniques to load them all into an SQLite database.

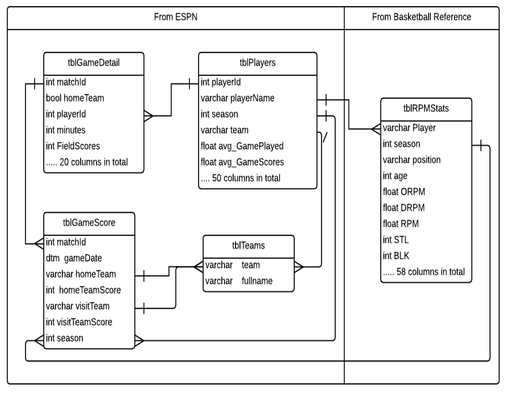

Figure 1: Entity-Relationship graph for SQLite database. The database follows the third normal form to ensure there is no redundancy. The lines between tables indicate the primary-foreign key relationship between dataset entities.

We have made our database publically available at `https://github.com/leerichardson/game_simulation/blob/master/data/nba/nba.db`

Real Plus Minus (RPM)

Throughout the poster we will refer to a feature called RPM. This is a new statistic which measures a player's estimated impact on efficiency per 100 possessions. It has become widely adopted in the basketball analytics community as the best measure of how good each individual player is both on offense and defense.

Features

Our final dataset had 44 different features, most of which were weighted averages for each team based on last seasons statistics. We also had features which exclusively considered the team performance in the previous season, but of course this does not account for player movement.

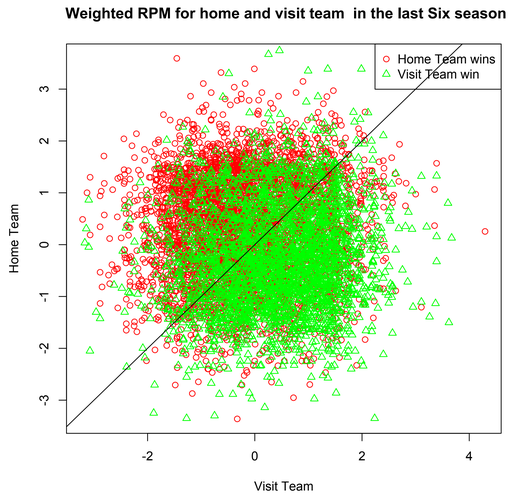

Figure 2: This plot shows Home and Away weighted RPM for each of the games in our dataset. This shows that our dataset is fairly messy, and certainly not linearly seperable.

Feature Selection

In order to choose which features were the most important, we used model selection techniques such as AIC. The following table shows the residual deviance of a logistic regression model if we were only to choose one covariate as a predictor.

Feature	Deviance
RPM Weight Away	9140.5
RPM Weight Home	9160.6
Weighted Minutes	9264.2
ORPM Weight Home	9303.3
ORPM Weight Away	9310.8
DRPM Weight Away	9329.1
DRPM Weight Home	9358.2
Previous Year Score Differential Home	9376.9
Weighted Assists	9384.7
Previous year Score Differential Away	9389.6

Table 1: This shows which single feature enhanced our model fit the most. Anything ending with "RPM" means Real Plus Minus, a new stat which is widely recognized as our best way to characterize the quality of which is quite clearly a superior predictor of performance compared with our othr features.

Algorithms

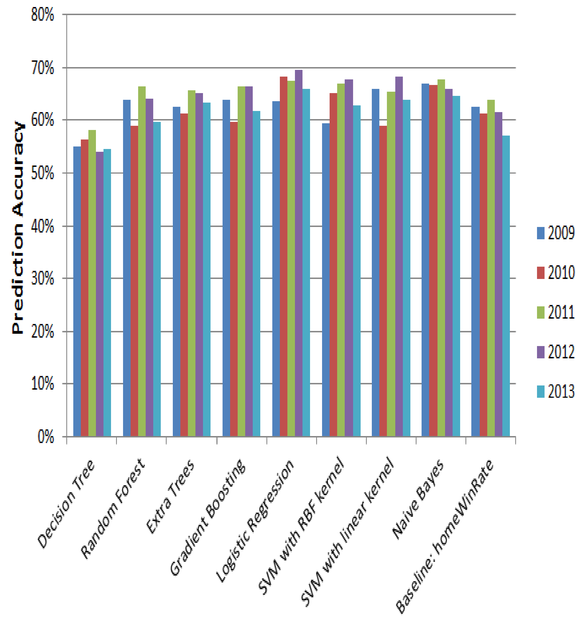

Figure 3: This plot shows the test error for our different algorithms in all of our years. From this, we see that Naive Bayes and Logistic Regression gave us the best overall test error. For each testing year, we used all the previous years as training data.

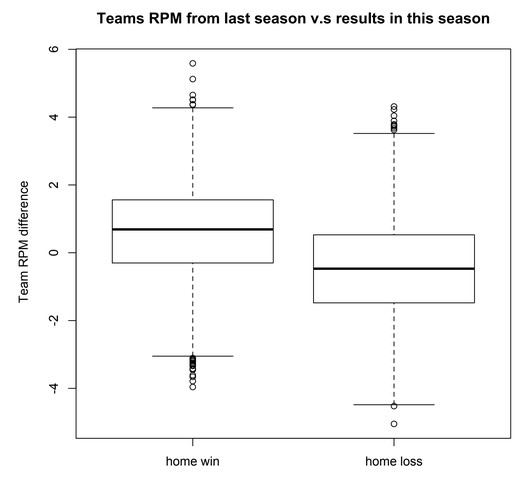

Figure 4: This shows the Difference in RPM between the home and away team in wins and losses. It's clear there is some signal here, as the difference in RPM is much larger in wins compared with losses

Simulations

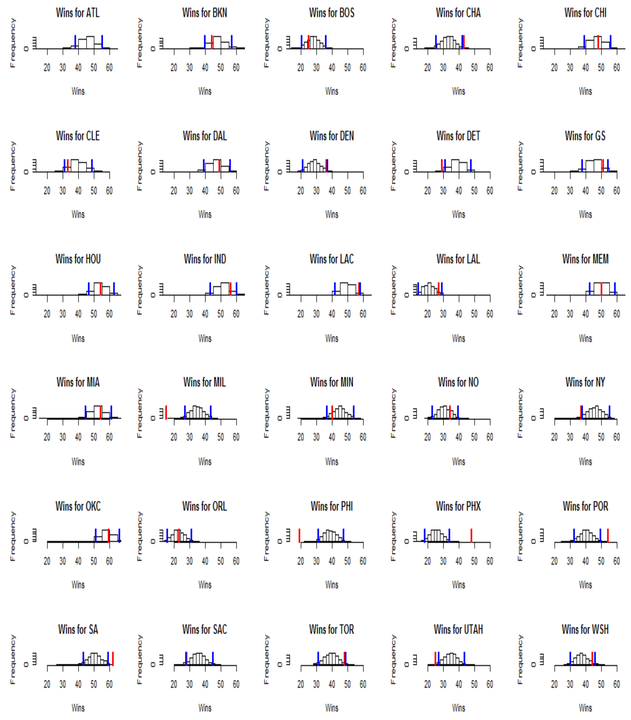

Figure 5: Results from simulating the 2013 NBA season 1000 times, using probabilities from our best Logistic regression model. The blue lines represent our confidence intervals whereas the Red lines represent the actual number of wins for each team. Our simulations trapped the true number of wins in 70 % of our intervals.

Our algorithms gave us probabilities that each team would win, one natural thing to do was to use these probabilities to simulate the season many times, and look at the distribution of wins. Our projections were in the top 15 in accuracy compared with many other projection systems.

Conclusion and Future Work

- One of the first things we noticed is that on average, home teams win about 60% of games. So just naively picking the home team without any information on how good the teams are gives a decent test error.
- From this, the question then becomes how much better does the away team have to be in order to pick them to win the game. We found that Weighted Adjusted Plus minus (WAPM)was the best indicator of how good a team is.
- In the future, we would like our WAPM to include more than just the previous season, and be a projection using more information. Also, we think that using current season data as a feature would be helpful and allow for predictions in the second half of the year to be stronger.