CS 506 final project
NBA predictions

Yue Sun, YuNing Qiu, HeMing Huang

1. Problem Formulation

In recent years, lots of talk have been risen in NBA circles regarding the rise of statistical analysis [1], additionally, due to the stability of the data has been improved recently, most notably play by play data which provides a record of each event. So, our team decided to do some work trying to predict the match outcomes based on the available NBA data. The basic idea is to split each game into miniature games [2].

In this project, we have used linear and logistic regression, naive Bayes, neural networks, and Support Vector Machine's and trees as the classifiers to predict the match outcomes.

2. Data Analysis

Several challenges we encountered in dataset preparation were:

- The first challenge we encountered was that the data we need were from different website, different source, so we need to find a way to combine all the data.
- For each website, the player's name may have small difference, so we need to correct this small challenge.

To solve these problems, we wrote our own data content scrape script. The two main sources we used were ESPN's [3] NBA website, basketball reference. Our dataset includes game score, the home and away teams, the players involved and their individual statistics, and also 50 individual statistics for each player in each season from ESPN database (ESPN data provide basic statistics such as the average number of points per game or the average number of rebounds per game for a given season). After collect all the data, what we need to do was trying to merge all the data to a single dataset, and we need to use this specific dataset to train and test our prediction model.

3. Methodology

- **Prepare the dataset for training and testing**

  To construct features, we spent lots of time to merge the data can be used to train the model. The features are constructed using of both the ESPN data and the Regularized Adjusted Plus Minus (RAPM) dataset. To create the features, we considered each match and listed the players on each team first, after that, the players' statistics were merged from the previous season with the results in the matches of the current season. What's more, the players' statistics has been added to form statistics for home teams and away teams. In the computation of the offensive and defensive RAPM statistics of each team, the players' average minutes per game from the previous season were used as weights.

  For example, suppose that there are n players in each team, with the i-th player playing mi minutes per game and having RAPM score ri in the previous season. Then, the weight for the i-th player and the team offensive RAPM are determined as:

$$w_i = \frac{m_i}{\frac{1}{n}\sum_{i=1}^{n} m_i}$$

- **Feature and algorithms selection**

  As we presented before, in this project we tried different models to do the training and testing, including Decision Tree Classifier, SVM, Random Forest Classifier, Gradient Boosting Classifier, Naïve Bayes, and Logistic Regression Classifier. For each model, we split our dataset into a training set and a test set. For example, to predict the game results for the 2013 season, we trained the model using data from seasons 2008 to 2012.

- **Algorithms discussion**

  Before the training of the model, we set the game results are denoted as 0 and 1 (with 1's denoting home team wins).

- **Logistic regression:**

  Logistic regression is particularly useful when the outcome is binary. As before, we use 0 and 1 to denote wins and losses, and we predict that the home team wins a new game if the predicted value for that game is greater than or equal to 0.5. Predicted values from a logistic regression are estimates of the probability that the home team wins.

- **SVM:**

  A Support Vector Machine is a non-probabilistic classifier. We tried different kernels, one this should be noticed was that even though different kernels generate different decision boundaries, none allows to exactly classify all of the data points. For this reason, we use soft margin to train the model.

- **Decision tree:**

  A decision tree is useful when one needs a good way to split a dataset with many features. In our case, there are 44 features overall. we choose the subtree that minimized the regularized training error to avoid the overfitting.

4. Experimental Results

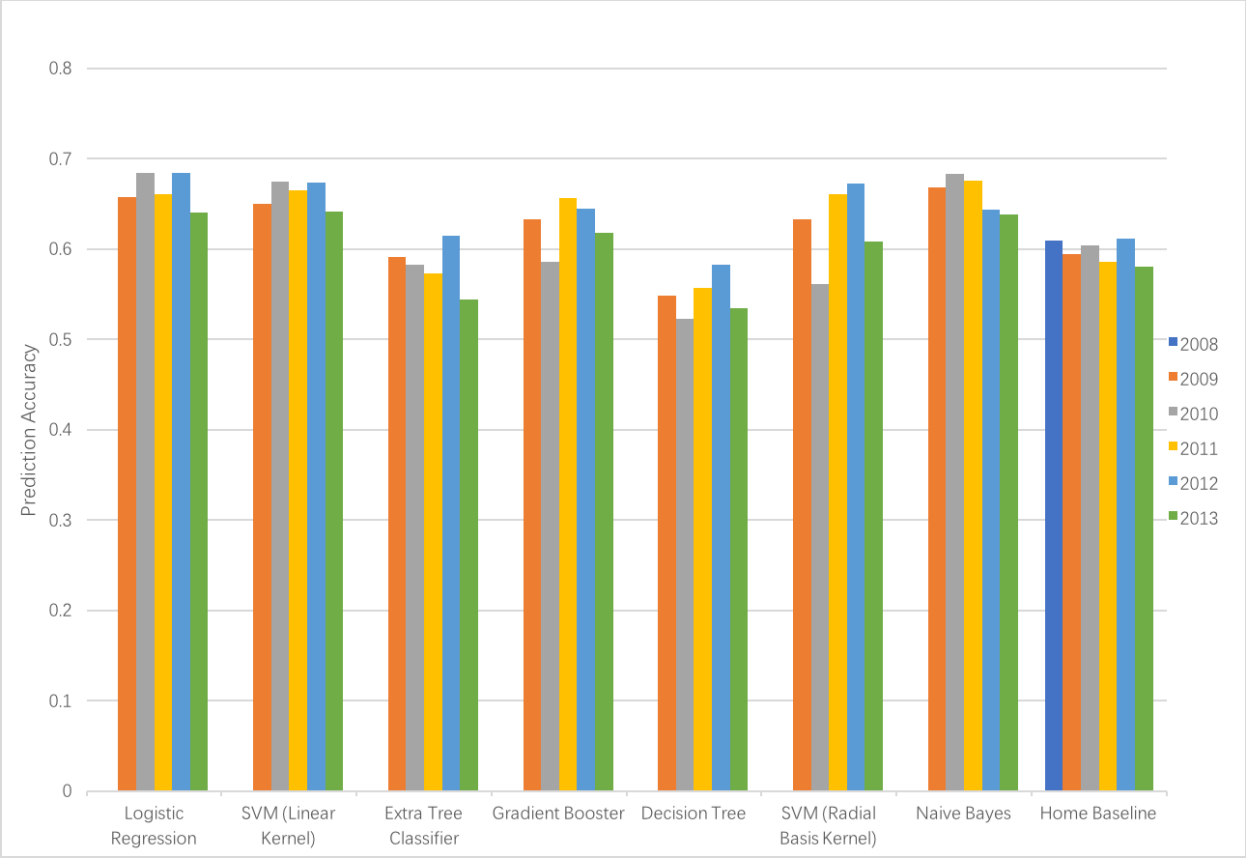

Figure 1. Results of each model

According to figure 1, the testing results for each model can be obtained, it can be seen that nearly all of the models can predict correctly higher that 50% (higher than random guess), which means the features we selected can illustrate the match outcomes significantly. Additionally, it can be seen that Logistic regression, and Naïve Bayes have the best performance, both of them are higher than 60% for each season, especially for logistic regression, the average accuracy for logistic regression was nearly 70%.

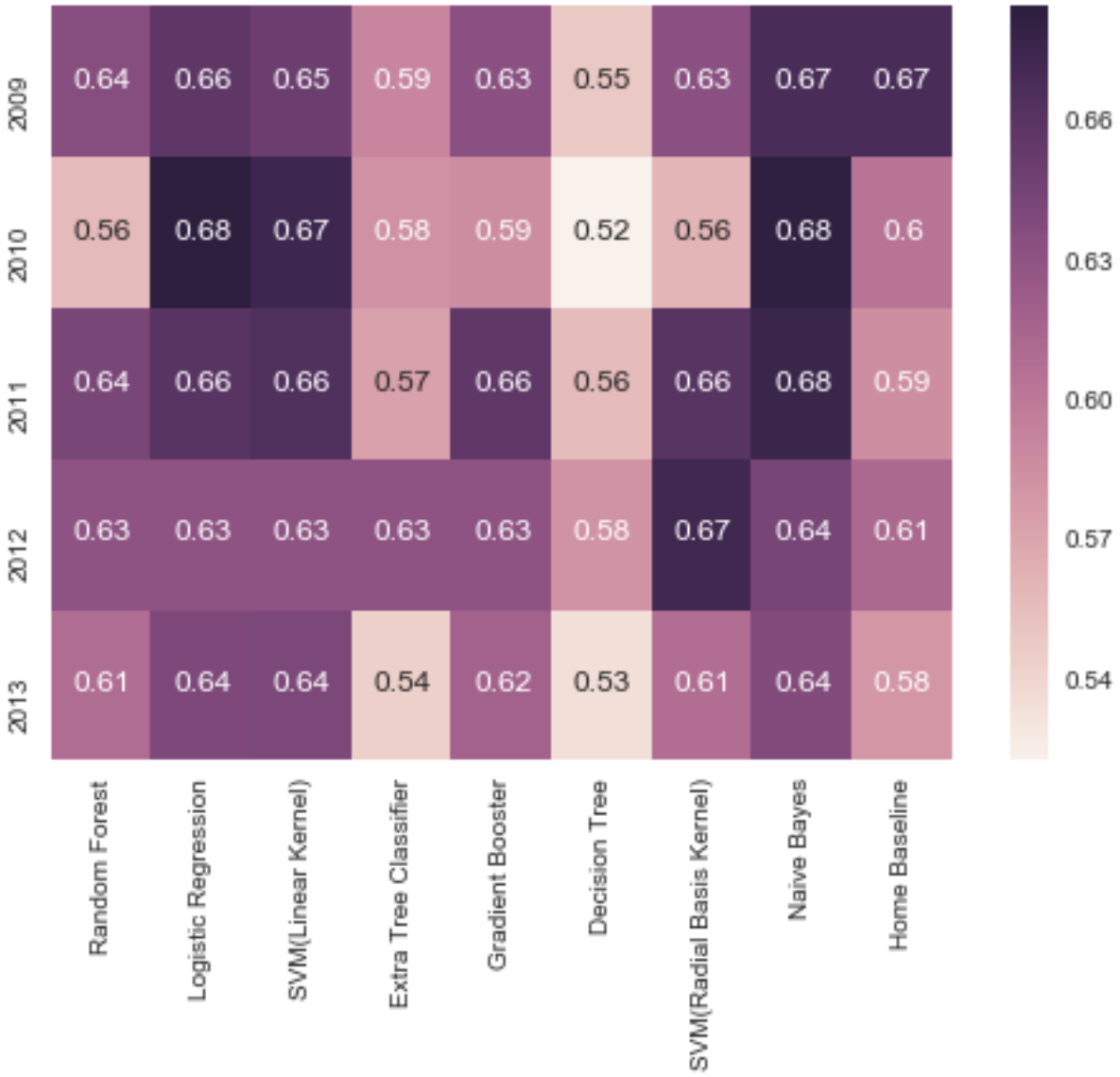

Figure 2. Heat map for the prediction results

According to the figure 2, the heat map of the prediction results can be obtained, it can be seen that the deeper color demonstrates the higher accuracy, therefore, it can be seen that Logistic regression always have the higher accuracy.

5. Conclusions

In conclusion, based on our machine leaning results, we can see that based on several basic statistic features, we can do some correct predictions (nearly 70%), however, as we can see, even though the best performance from logistic regression, the accuracy still was not higher than 70%, one main reason for this maybe is that no matter what

algorithm/features we used, upsets and unexpected results always happened in the NBA. There should have many ways to improve the results in the future, for example, we could not only be used individual player data from the previous season in order to predict the current season, which can avoid the situation if a player misses the previous season. Additionally, we could consider a weighted average of the past 3-5 seasons to get an estimate for the quality of each player, not only the previous season.

6. References

[1] Sam Hinkie and the Analytics Revolution in Basketball. Nilkanth Patel. http://www.newyorker.com/news/sporting-scene/sam-hinkie-and-the-analytics-revolution-in-basketball

[2] Paul Fearnhead, Benjamin M. Taylor On Estimating the Ability of NBA Players. 2010: http://arxiv.org/pdf/1008.0705.pdf.

. [2] ESPN. http://espn.go.com/nba/.