Predicting National Basketball Association Winners

Jasper Lin, Logan Short, and Vishnu Sundaresan

***Abstract*—We used National Basketball Associations box scores from 1991-1998 to develop a machine learning model for predicting the winner of professional basketball games. Initially, we found that simply choosing the team with a higher win percentage was correct 63.48% of the time. We implemented 5 different supervised learning classification models: Logistic Regression, SVM, aDaboost, Random Forest, and Gaussian Naive Bayes. Using points score, field goals attempted, defensive rebounds, assists, turnovers, overall record, and recent record as features, we found that Random Forest could accurately predict the result 65.15% of the time. Dividing up a season into quartiles, resulted in an improvement to 68.75% with logistic regression in the final quartile. Additionally, testing without using the teams current winning record resulted in a 2-3% decrease in prediction accuracy for most algorithms.**

1 Introduction

PREDICTING the outcomes of sporting events and the performance of athletes is a natural application for machine learning. Many professional sports have easily accessible data sets that tend to be random in nature and are attractive to predict. Predicting the outcomes of National Basketball Association (NBA) games is particularly interesting because basketball is a sport that is viewed as especially player driven. The current stigma is that it is necessary to have a superstar to win games. More and more advanced statistics are being adopted and used by teams. In this project, we use various classification algorithms to try to predict the winner of a matchup between two teams in addition to determining what are actually the most important factors to determining the outcome of a game without looking at individual player statistics.

2 Data

Our dataset consisted of all box score statistics for NBA games played beginning with the 1991-1992 season and ending with the 1997-1998 season. The statistics contained in the box score are discussed in Section 4. In line with our goal of predicting the results of a seasons games using past data, we defined the 1997-1998 season to be our test set and let the rest of the seasons be our training set.

3 Benchmarks

In order to establish the scope of the accuracies our model should achieve, we first developed benchmarks. We defined two naive win prediction methods: 1) Predict that the team with the greater difference between average points per game and average points allowed per game will win and 2) Predict that the team with the greater win rate will win. We then ran each of these benchmarks on the games of the 1997-1998 season to obtain our benchmark win prediction accuracies. In addition, we considered a third benchmark based on the win prediction accuracies of experts in the field which is generally around 71%. [1]

	$PredictionAccuracy$
Point Differential	0.635
Win-Loss Record	0.608
Expert Prediction	~ 0.71

Performing better than these benchmark accuracies indicates that our methods are able to capture certain abilities of a team not shown in their overall record, such as whether or not they are better defensively or offensively compared to the rest of the league, or if their recent performance has any effect on future games. It is noted, however, that experts do not predict a winner in games that are deemed too close to call. Thus, the reported accuracy of expert predictions is likely inflated and could prove difficult to surpass. [1]

4 Feature Selection

The standard NBA box score includes 14 statistics measuring each teams performance over the course of a game. These statistics are:

- *Field Goals Made (FGM)*
- *Field Goals Attempted (FGA)*
- *3 Point Field Goals Made (3PM)*
- *3 Point Field Goals Attempted (3PA)*

- *Free Throws Made Made (FTM)*
- *Free Throws Attempted (FTA)*
- *Offensive Rebounds (OREB)*
- *Defensive Rebounds (DREB)*
- *Assists (AST)*
- *Turnovers (TOV)*
- *Steals (STL)*
- *Blocks (BLK)*
- *Personal Fouls (PF)*
- *Points (PTS)*

Additionally, there are three stats that are important aggregations over the course of the season.

- *Average Points Allowed (aggregate)*
- *Average Points Scored (aggregate)*
- *Win / Loss Record (aggregate)*

Using the statistics contained in the box score, we constructed a 16-dimensional feature vector for each game, containing the difference in the competing teams net: [win-lose record, points scored, points allowed, field goals made and attempted, 3pt made and attempted, free throws made and attempted, offensive and defensive rebounds, turnovers, assists, steals, block, and personal fouls]. To the feature set given by the box score we decided to add an additional feature which quantified the recent performance of a team using the teams win record over their most recently played games. This feature and our motivations for believing it could contribute to better game winner predictions are discussed in section 4.1.

Initially, we trained and tested all of our learning models on the aforementioned feature vectors. We quickly realized, however, that besides logistic regression, which performed well, all of the other models suffered from overfitting and poor test accuracies. In order to curb our overfitting we decided to instead construct our models using a small subset of our original features consisting of the features that best captured a teams ability to win. In choosing a specific set of features to utilize in our learning models, we ran three separate feature selection algorithms in order to determine which features are most indicative of a teams ability to win. Two of the feature selection algorithms used were forward and backward search, in which we utilize 10-fold cross validation and add or remove features one by one in order to determine which features result in the highest prediction accuracies. In addition, we ran a heuristic feature selection algorithm to verify that the features selected tended to be those that are most informative about whether a team will win. The results of the three methods are shown in the table below.

Forward Search	Backward Search	Heuristic
Points Scored	Points Scored	Points Scored
Points Allowed	Field Goals Attempted	Field Goals Attempted
Field Goals Attempted	Defensive Rebounds	Free Throws Made
Defensive Rebounds	Assists	Defensive Rebounds
Assists	Turnovers	Assists
Blocks	Overall Record	Overall Record
Overall Record	Recent Record	Recent Record

The features selected by backward search were almost the exact same features as those selected by heuristic search. This indicated that the backward search features captured the aspects of a teams play that best indicated whether that team would win and thus that these features would likely yield good results. Our preliminary results showed that backward search did in result in the best cross-validation accuracy. The features selected by backward search also agree with the experts view of the game, that prediction is most accurate when considering the offensive and scoring potential of a team compared to its opponent. Each of the selected statistics are related to scoring, even turnovers and defensive rebounds as they essentially give the team possession of the ball.

4.1 Prediciton Accuracy Performance of Recent Record

There has been much debate in the past decade over whether a teams recent performance is an indicator of how likely a team is to win the next game. This phenomenon that players and teams doing well will continue to do well is known as the Hot Hand Fallacy, and it has been shown in the past that to have no correlation to how a team will do in the future. In order to explore this for ourselves (on a team level), we decided to test the accuracy of our model using only the a teams win-loss record in the past μ games. The graph below shows our accuracy varying μ from 1 to 20, where we notice that the accuracy peaks around 66.3% accuracy using cross-validation and a logistic regression model. Comparing this result to the results obtained in Section 5.1, we see that there is no noticeable increase in accuracy thus supporting the notion of the hot hand fallacy. Recently though, research shown at the MIT Sloan sports conference has shown that the fallacy may in fact be true, utilizing a new approach that takes into account the difficulty of shots taken by a player playing exceptionally well. [2] While we do not have the necessary data to test this claim, future work could include obtaining more data on the types of shots and positions that players attempt them.

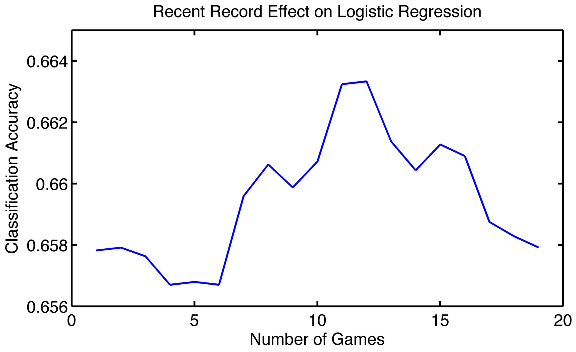

Fig. 1. Accuracy for Recent Win Percentage Window Lengths

5 Models, Results, and Discussion

Since our goal was to evaluate whether the outcome of games in a current season could be predicted using historical data, we constructed our machine learning models using the 1997-98 season as our test set and all other seasons as our training set. In addition, statistics regarding a teams performance from previous seasons were not factored into a teams statistics for the current season. This is because variables such as trades, injuries, experience, or management changes can cause high variance in the strength of a team from year to year. In order to make sure that we are consistent in our evaluation of a team, we therefore only base our feature vectors on the teams performance in games taking place in the current season. Naturally this means that at the start of the season our evaluation of a teams strength will be less accurate since we do not have as much information about the team. As the season progresses we will obtain more data about how a team performs and our evaluation of teams should become more accurate. As this occurs we can also expect that our game outcome predictions will become more accurate.

We used five machine learning algorithms in evaluating our dataset to predict game winners. These included logistic regression, support vector machine (SVM), adaptive boost, random forests, and Gaussian Naive Bayes. In order to optimize these models for our data, we used 10-fold cross validation to determine algorithm parameters. Logistic regression attempts to train coefficients for each feature in the feature vector in order to obtain probabilities that a team will win a game. SVM attempts to find a hyperplane that separates games resulting in a loss from games resulting in a win based on the feature vectors of games. Our SVM was implemented with a Gaussian radial basis function (RBF) kernel. Since our data was not linearly separable we also added a cost parameter of 10. These methods work well under the assumption that games resulting in wins tend to reside in a different section of the dimension defined by the feature vectors than games resulting in losses.

Adaptive Boosting and Random Forests attempt to accurately classify games as wins or losses by averaging the results of many weaker classifiers. Adaptive boosting performs multiple iterations that attempt to improve a classifier by attempting to correctly classify data points that were incorrectly classified on the previous iteration. This makes the boost sensitive to outliers as it will continually try to correctly predict the outliers. In the NBA season, outliers manifest themselves as upsets, games where a much weaker team defeats a stronger team. Our adaptive boost ran for 65 iterations, a parameter discerned by running cross-validation. Random forests constructs decision trees that each attempt to decide the winner of a game. The classifications of each of these trees are then averaged to give a final accurate prediction of the winner of a game. Our random forest was implemented with 500 trees and constructed decision trees up to a depth of 11, again with parameters determined using cross-validation. The strength of the random forest algorithm is that it can account for particularly complex decision boundaries, possibly resulting in a very high training accuracy.

We also tested our data on Gaussian Naive Bayes which will assume that the likelihood of our statistics is Gaussian and try to fit a model using this assumption. This model would perform well on our data if the statistics do indeed prove to follow Gaussian functions.

5.1 Win Classification Algorithms

Algorithm	Training Accuracy	Test Accuracy
Benchmark	-	0.635
Logistic Regression	0.661	0.647
SVM (RBF kernel, Cost = 10)	0.658	0.651
AdaBoost (65 iterations)	0.664	0.641
Random Forest(500 trees, Depth = 11)	0.809	0.652
Gaussian Naive Bayes	0.631	0.633

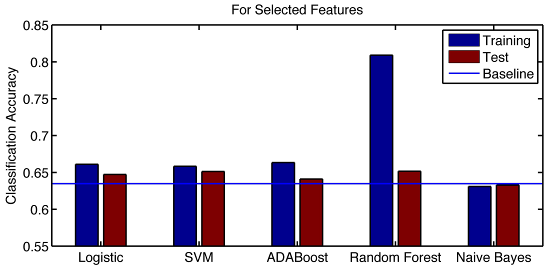

Fig. 2. Results of Classification Algorithms

An interesting observation of this data is that by simply comparing the win percentage of the two teams, we can accurately predict the result of the game 63.48% of the time. This is only 2% less accurate than including an additional 15 features. This result can be attributed to the fact that a teams win percentage inherently has information about how the team has been doing. Additionally, we have found that basketball is a very easy sport to predict, when compared to a sport such as baseball where chance plays a larger role and teams finish the season with win records much closer to 0.5, thus resulting in a very high baseline to surpass.

After training and testing each of our classification models on the data, the resulting accuracies essentially all outperformed the baseline set by our benchmark, but by only a small margin. Some of the algorithms such as adaptive boosting and especially random forest also seemed to overfit our data, as seen by the high training accuracy. This is possibly due to the nature of our algorithms having robust decision boundaries, as well as the possibility that past season statistics and games are not a good indicator of how a team performs and how the game in general works in the test season.

5.2 Accuracy of Win Classifications Over Time

One aspect of our learning models that we wanted to explore is how they performed over time. Intuitively, our statistics should vary much more at the start of the season, and slowly converge as we obtain more data to reflect a teams true ability to win a game. In order to explore this theory, we partitioned our test season into 4 equal sized chronological blocks, and tested our algorithm on games occurring within each of these 4 sections using all games up to that point to calculate a teams statistic feature vector.

Season Quarter	Log. Regression	SVM	AdaBoost	Random Forest
Quarter 1	0.588	0.588	0.559	0.559
Quarter 2	0.641	0.645	0.618	0.673
Quarter 3	0.662	0.657	0.624	0.638
Quarter 4	0.688	0.678	0.678	0.644

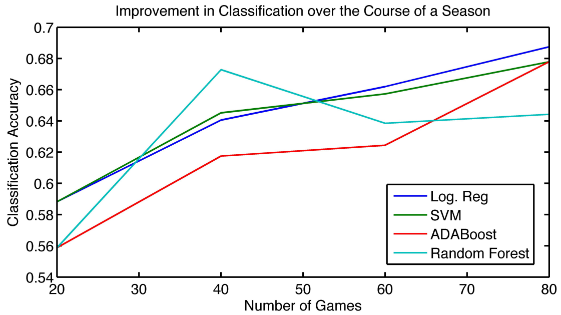

Fig. 3. Classification Accuracy Over the Course of a Season

As seen above, the result was what we expected, with the accuracy during the first quarter of the season being extremely low compared to our results in Section 5.1, and with the accuracy during the final quarter of the season being significantly higher. The accuracy at the end of the season is much higher than the baseline utilizing simply the win-loss record, indicating the the data does show potential to represent an aspect of a team not captured by one statistic alone. While our overall prediction accuracies in Section 5.1 seem only marginally better than the baseline, this trend shown by looking at each individual quarter of the season gives an indication that the accuracy can indeed be improved. Utilizing the law of large numbers, we believed that a teams long term performance would eventually regress to the mean and reflect the true ability of each team. As a result, the result of each game would likely be dictated by the difference in their average statistics each game.

5.3 Win Classification Without Win/Loss Record

Feature selection and our baseline has shown us that a team's win-loss record is clearly the best indicator of how well the team will do in future games, but is it possible to still attain that level of accuracy in prediction by simply using the NBA box score aggregates? We tested this hypothesis by re-testing each of our learning models on a feature vector containing all 14 of the original box score statistics.

Algorithm	Training Accuracy	Test Accuracy
Benchmark	-	0.635
Logistic Regression	0.663	0.645
SVM	0.661	0.637
AdaBoost	0.672	0.618
Random Forest	0.886	0.628
Gaussian Naive Bayes	0.560	0.599

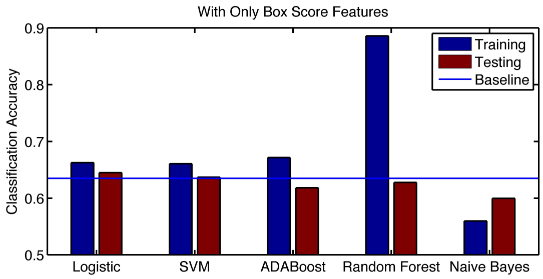

Fig. 4. Results of Classification Algorithms Using Only Box Score

As seen above, results show that the accuracies obtained from only using feature vectors containing the historical NBA box score aggregates performs reasonably well, but fall short of the benchmark for all models except for logistic regression and SVM. This indicates that box scores alone are not enough to represent a teams ability to win, and that further data is needed to increase our accuracy.

6 Conclusion

We found that a basketball teams win record plays a central role in determining their likeliness of winning future games. Winning teams win more because they have the ingredients for success already on the team. However, we were surprised that removing the winning record significantly changed classification accuracy. If we consider a teams win record as representative of that teams ability to win, then this implies that the box score statistics fail to completely represent a teams success on the court. This result points to the need for advanced statistics that go beyond the boxscore in order to potentially improve prediction accuracy for close games and upsets. This need explains the growing popularity on advanced statistic sport conferences like the MIT Sloan conference.

7 Future Work

Another interesting application of this project could be understanding how the game of basketball has evolved over time and if the features selected for the 1991-1997 seasons are the same features selected for modern day basketball. With modern basketball, there are far more advanced statistics outside of the box score that are available which could result in significantly better features to learn on. Additionally, we can further expand our data to tackle the notion that the hot hand fallacy is indeed true, by looking at the difficulty of a players shots given that they are performing well recently.

References

[1] M. Beckler, H. Wang. NBA Oracle

[2] A. Bocskocsky, J. Ezekowitz, C. Stein. The Hot Hand: A New Approach to an Old ”Fallacy”. 2014