NCAA March Madness Insights & Predictions

Eddie Wu, Zhengchao Ni, Zhefu Peng

Abstract

Our team is building a classification model that can accurately predict the outcome of matchups in the march madness tournament, and at the same time we are investigating the impact of game statistics on the matchup outcomes. In total of nine classifiers are trained and tested using the regular season per-game data as the training set and the tournament per-game data as the testing set, and we have chosen Adaboost and Logistic Regression due to their high classification accuracies (94.2% and 97.2%), high F1 score (0.942 and 0.9721) and excellent performance on their learning curves without signs of overfitting. In addition, we have found that teams in the regular season with higher number of assists or defensive rebounds are more likely to win the game, and those with lower number of turnovers hardly lose the matchup. It is also shown that teams with higher number of steals & blocks in the tournament have greater probabilities to win any tournament matchup, according to our final models.

1 Introduction

Our group aims at analyzing the past data of Division I college men's basketball teams, including both the regular season and tournament statistics, from which we plan to discover aspects of the basketball game that have significant impacts on the outcomes and can likely lead to upset in the tournament. In short, we want to find one or some key factors which could help us to predict the outcome of a series of games. All of these past data(from years 2003 to 2017), is used as the training set, and the 2017-2018 season games and/or tournament games are used as the testing set for evaluation of our chosen classifiers. To solve the problem, the classifiers and the models we want to use include Logistic Regression, Random Forest, Gradient Boosted Trees, Adaboost among others. For these training models or classifiers, the first thing we did was built the feature vectors. The more specific talking about feature vectors would be seen in the section 2.

In general, the team with higher rank(so called strong seed) would defeat the team with lower rank. Nevertheless, sometimes there are some upsets in a tournament. For example, in 2018 NCAA March madness tournament, the 16# team UMBC(completely the weak seed) defeat the 1# team Virginia and reached to final 32, which really surprise everyone. It would be so amazing and exciting when we could know the result of games in advance. In one basketball game, the team with higher scores or points will win the game. However, besides the scores, there are still some statistics in one game, such as the number of turnovers, rebounds(offensive and defensive, or total), assists, personal fouls, etc. Maybe there is a relationship between the outcomes and these statistics. Which one has the best impact on a game? If we found it, we could predict the outcomes of each competition in the tournament. Therefore, we chose four classifiers, Adaboost, random forest, gradient boosting, logistic regression to build predicting and testing models separately. After that, we could find the statistics which has the best effect on a game from different four models. Then, we would use

the data that shows the average basketball statistics information for every team in the 2017-2018 regular season, picking up the certain key factor and applying it to our prediction models. At last, we compared the result we got from the four models and the real facts in 2018 NCAA March Madness Tournament, calculating the predicting accuracy. With the predicting accuracy, we could pick up the final key factor and the best classifier(s).

In this paper, section 2 describes the datasets we used, the process of cleaning and wrangling the data and modification of the dataset in detail, for example, adding the label to the result of each game, fitting to each classifier. Section 3 discusses the details of machine learning techniques, introduces the training and tuning procedures, evaluates the classifiers' performances, and provides a comparison between different models. Section 4 concludes the paper with demonstrating limitations and future interests.

2 Data Processing

2.1 Desciprtion of Data

We have two sources of data. First source is from kaggle website (https://www.kaggle.com/c/march-machine-learning-mania-2017/data#), which provides us with the outcomes of games from year 2003 to 2017, including regular seasons and tournaments. Besides the final outcome, the dataset contains the specific value such as the number of rebounds, assists, turnovers, etc, per game. Second source is from "https://www.sports-reference.com/cbb/seasons/2018-school-stats.html", which contains the average basketball statistics information for every team in the 2017-2018 regular season. After we established our models and classifiers, we wanted to use this dataset to predict the final outcome of 2018 tournament, comparing to the real result and calculating the prediction accuracy. The real results are collected and stored in a .csv file that notes the winner in each matchup.

We are using two files from the kaggle source, one of which contains detailed game by game statistic of two teams in the regular season from 2003 to 2017, and another contains detailed game by game statistic of two teams in the march madness tournament from 2003 to 2017. Both of these datasets are very clean and of high-quality, since they contain all the data and statistics we need in entirety without missing values, and we hardly need to clean them.

We are using one file from the second source and another file that is created by us manually, both of which will be described in length at the next section.

Note that there are some abbreviations in the following paragraphs and figures, and here we would like to explain them a little more so those terms can be better understood:
fga - field goals attempted; fgm - field goals made; fga3 - three pointers attempted; fgm3 - three pointers made; fta - free throws attempted; ftm - free throws made; or - offensive rebounds; dr - defensive rebounds; ast - assists; to - turnovers; blk - blocks; stl - steals; pf - personal fouls; Astvsto - the rate of assists to turnovers; ShootingAverage3 - the percentage of three pointers for whole team; ShootingAverage - the percentage of field goals for whole team.

2.2 Data Cleaning and Wrangling

For our choices of classifiers, we built the feature vectors consisting of the basic basketball statistics for every pairs of college team (for each year in the season and in the tournament) that played against each other, and the associated target variable would just be 0 (higher seeded team, or the team with better records, wins) or 1 (lower seeded, or the team with worse records, team wins). Although the data of past seasons are already recorded on the per-game basis, the available data for current (2017-2018) season is for the whole season per each team, thus we had to programmatically calculate the average statistic (field goals made per game, for example) for the teams before we use it as part of our testing dataset.

Moreover, we have to manually create the table containing all the 2018 March Madness tournament matchups, which would then be used to join with existing tables to get per-game statistics for each college team we are interested in (the ones that made it to the tournament). This manually generated table, after joining with others, would provide part of the testing labels in our testing dataset.

3 Implementation and Analysis

3.1 Models Selection and Training

Our project deals with a basic binary classification problem (win or lose), and we have chosen the following models for the task: Logistic Regression, Linear Discriminant Analysis classifier, Decision Tree, K-Nearest Neighbor (K=2), Support Vector Machine with Gaussian Kernel, Naive Bayes, Adaboost, Random Forest and Gradient Boosting Tree, all of which have been proven to be useful and effective in binary classification problems (even multiclass classification for a few of them). K-Nearest Neighbor works well if the data points are relatively far away from each other so they can be separated into different groups, and in our case it might work well as we expected winning team and losing team to have somewhat different basketball statistics, but it might not be as robust as some other more powerful classification algorithm; Logistic Regression is typically the go-to model for any kinds of classification problem, as it is very fast to compute and can generally lead to fairly accurate predictions (especially with binary labels) without overfitting much; One of the ensemble method used here is Random Forest, and it is a very popular method both in academia and in industry for classification problem, as it ensembles many decision tree classifiers together where each tree has randomly selected features (through bagging) as splitting nodes and can most likely lead to very accurate result without overfitting. Adaboost is another ensemble learning algorithm, which combines several weak learning algorithms and then updates and iterates their weights. It is always a strong learning algorithm with high performance in practice. There are definitely pros and cons associated with each of the method/model we have chosen, and performances of them will be evaluated later in the report.

Before we move on to the model fitting, we would like to specify what our training and testing set entails. In the college basketball world, the regular season performance of a year typically dictates the team's eligibility in that year's march madness tournament and the team's seeding in the tournament if selected, therefore our team has decided to use all the regular season per game statistics and the associated game outcomes as the training set, and the tournament per game statistics with the outcomes are subsequently used as the testing set. Doing so will help us determine how predictive the regular season performances are on teams' tournament outcomes, as we will assess a few accuracy metrics (accuracy, precision, and etc) of our classifiers and observe if the regular season statistics predict the tournament

match-up results relatively accurately. Note that the season average data for 2017-2018 season is used as part of the testing set corresponding to the 2018 march madness tournament.

3.2 Model Comparison

In this paper, we explore basic machine learning algorithms to complete the task, including Logistic Regression, Linear Discriminant Analysis classifier, Decision Tree, K-Nearest Neighbor (K=2), Support Vector Machine with Gaussian Kernel, Naive Bayes, Adaboost, Random Forest and Gradient Boosting Decision Tree, the performance of which are presented in Table 1.

Table 1 Model Performance Comparison

Model	Test Accuracy	Precision	Recall	F1
Logistic Regression	97.20%	97.23%	97.20%	97.21%
Linear Discriminant Analysis	93.06%	93.07%	93.06%	93.06%
Naïve Bayes	88.31%	88.30%	88.30%	88.30%
K Nearest Neighbour	85.19%	86.21%	85.19%	85.05%
Support Vector Machine	89.44%	89.47%	89.44%	89.44%
Decision Tree	85.40%	85.41%	85.40%	85.40%
Random Forest	90.79%	90.80%	90.79%	90.79%
Adaboost	94.20%	94.21%	94.20%	94.20%
Gradient Boosting	92.44%	92.48%	92.44%	92.44%

From the above table, we can see that Logistic Regression and Adaboost are the two best performing classifiers as they not only produce the highest overall test accuracy, but also are associated with the highest F1 score, which suggests that these two classifiers are making accurate predictions without favoring any of the two binary classes. Moreover, the fitting time to these two classifiers are not very long, making them very ideal for the goal of our project.

We are using both of these models instead of only the best performing Logistic Regression as our chosen classifiers for the following reason: Logistic Regression works very well when there are not that many data as it was the case for our project, but with increasing amount of basketball game data such that in spans over a greater range of seasons (such as from all the way back in 1940s to now) it may not work as well due to possibility of more noises, and that's why we are also including a second classifier, Adaboost, which is an ensemble method

that can be robust to noises and can scale up with even very large amounts of data, and it is also quite powerful to fit the data without overfitting.

Figure 1 and 2 below compare the learning curve between a few well-performing classifiers and the error rates between two ensemble methods. As we can see from figure 1, Logistic Regression achieves both 0 training error and 0 testing error due to low amount of noises in our not-so-large dataset, while Random Forest achieves the highest testing error even with 0 training error therefore it's safe for us to eliminate this method for our project. Both Adaboost and Gradient Boosting Tree achieve similar training and testing error at around 0.05, and there's no huge discrepancy between training and testing performance, so we decide to investigate performances of these two ensemble methods further, as depicted in figure 2. With increasing number of estimators, we can see that Adaboost achieves both lower training error and testing error that the Gradient Boosting method does, therefore we have chosen Adaboost among the two here, as well as the near-perfectly-performing Logistic Regression, confirming with the models chosen from the accuracy & F1 score figure above.

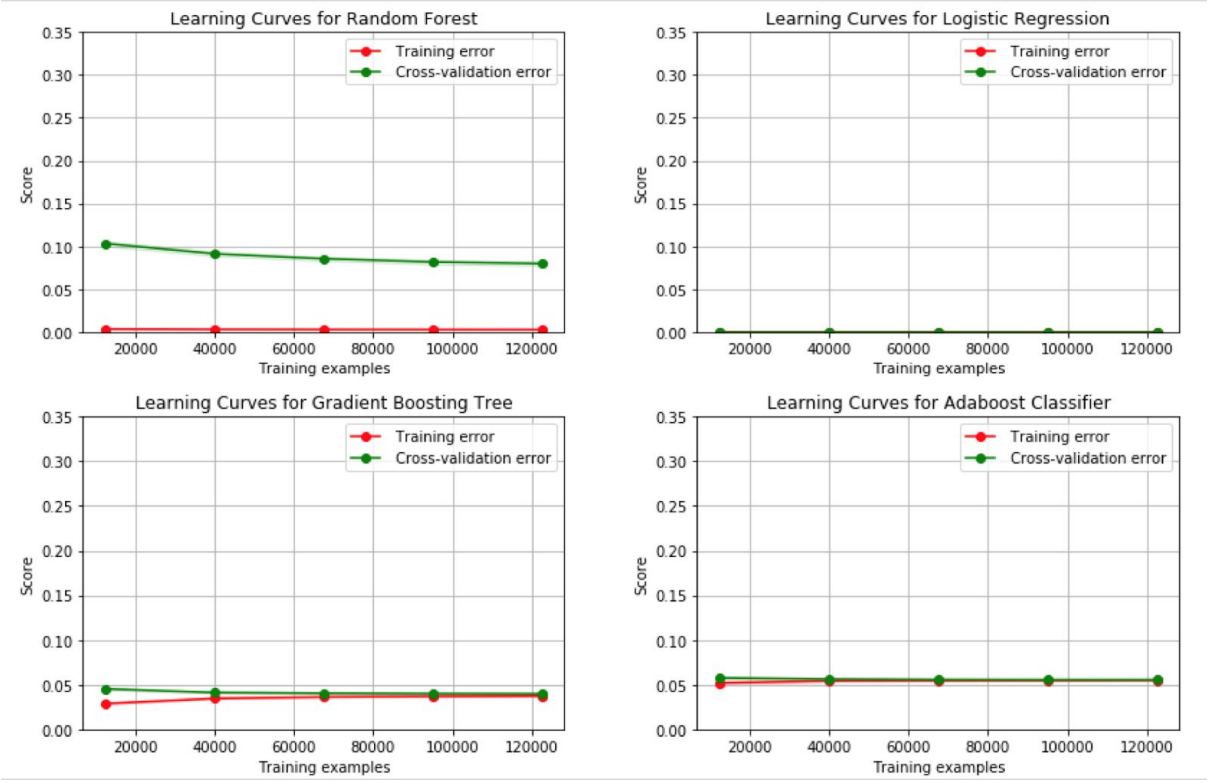

Fig.1 Learning Curves

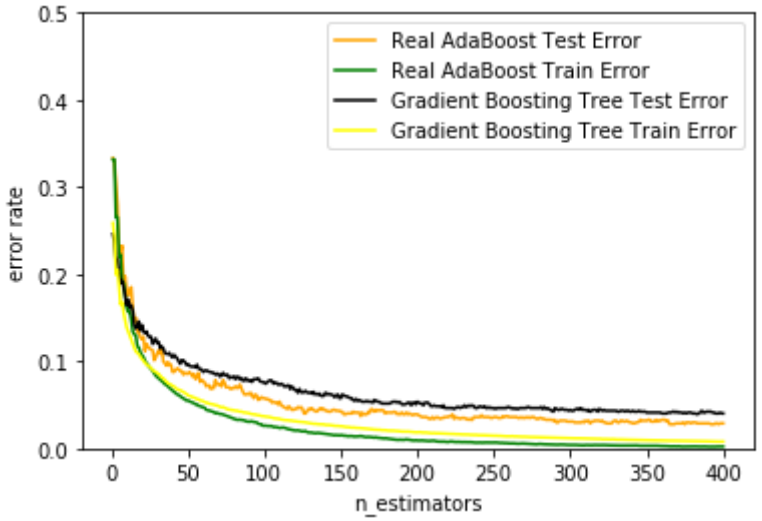

Fig.2 Error rate curves

3.3 Interesting Insights

Below are figures which show some important statistics of the winning teams along with losing teams. Specifically, Fig.3 shows the comparison of field goal attempts (fga), field goal made (fgm), field goal of 3-pointer attempts (fga3) and field goal of 3-pointer made (fgm3); Fig.4 free throw attempts (fta), field throw made (ftm), offensive rebounds (or) and defensive rebound (dr); Fig.5 assistants (ast), turnovers (to), blocks (blk) and steals (stl); and Fig.6 personal fouls (pf), assistant-turnover ratio (ASTvsTO), shooting average (excluding 3-pointers) and shooting average of 3-pointers. In those figures, the red points stand for the statistics of winning team, and the green points stand for that of losing teams.

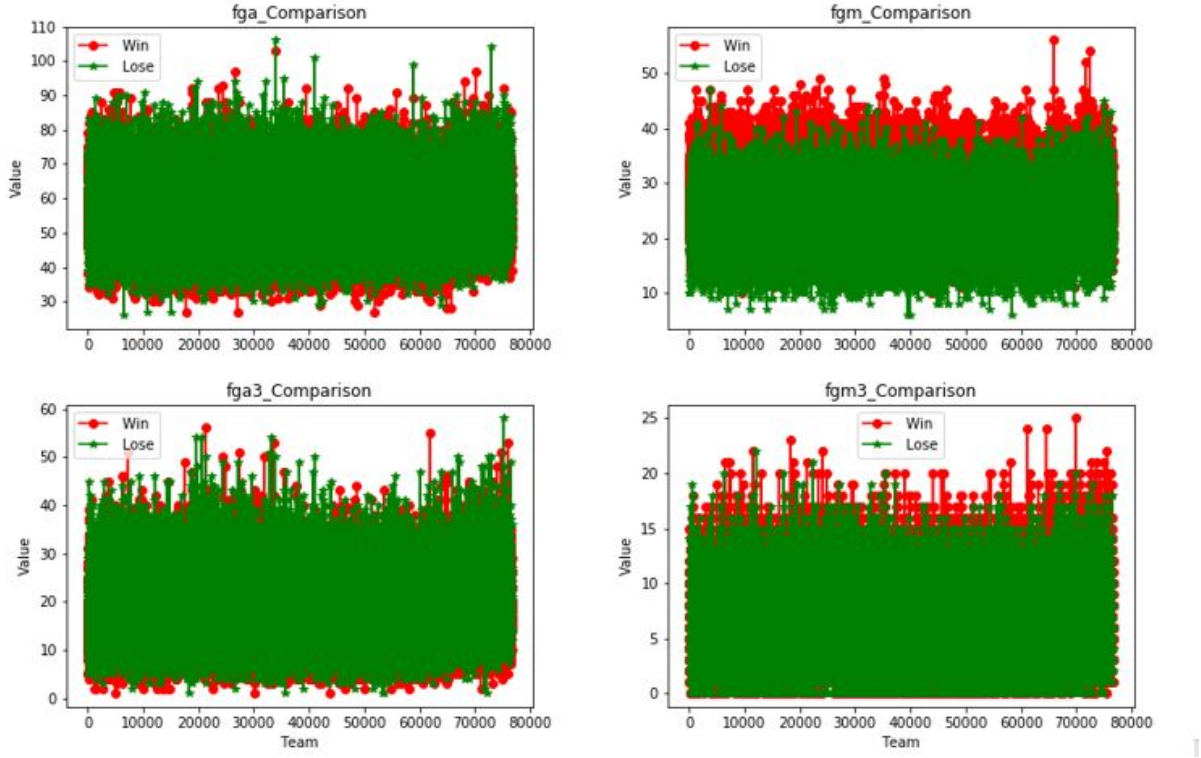

Fig.3 Comparison

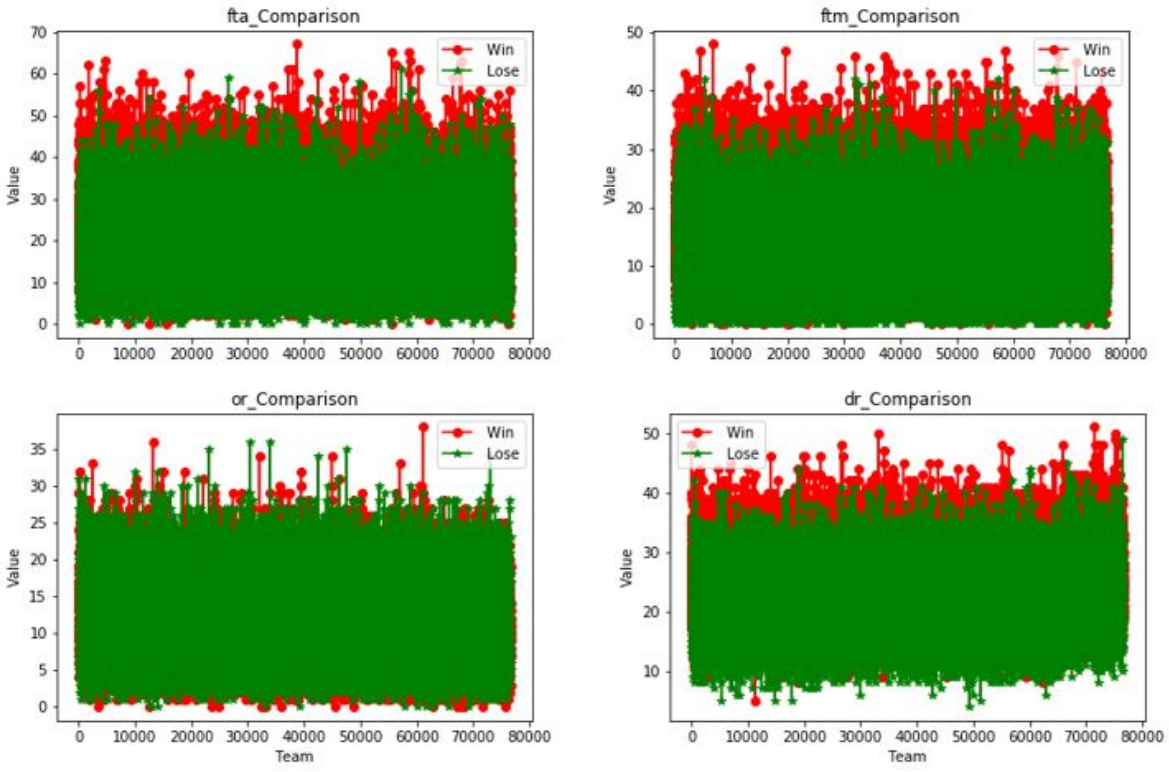

Fig.4 Comparison

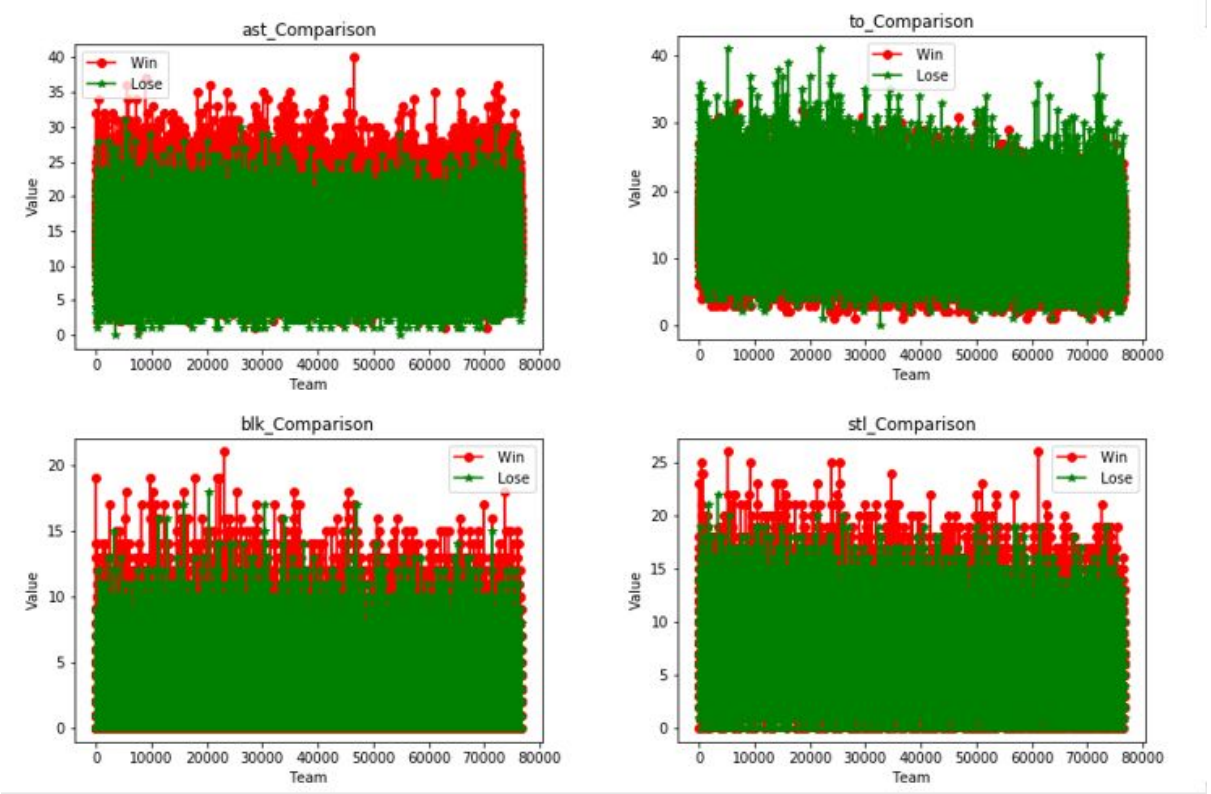

Fig.5 Comparison

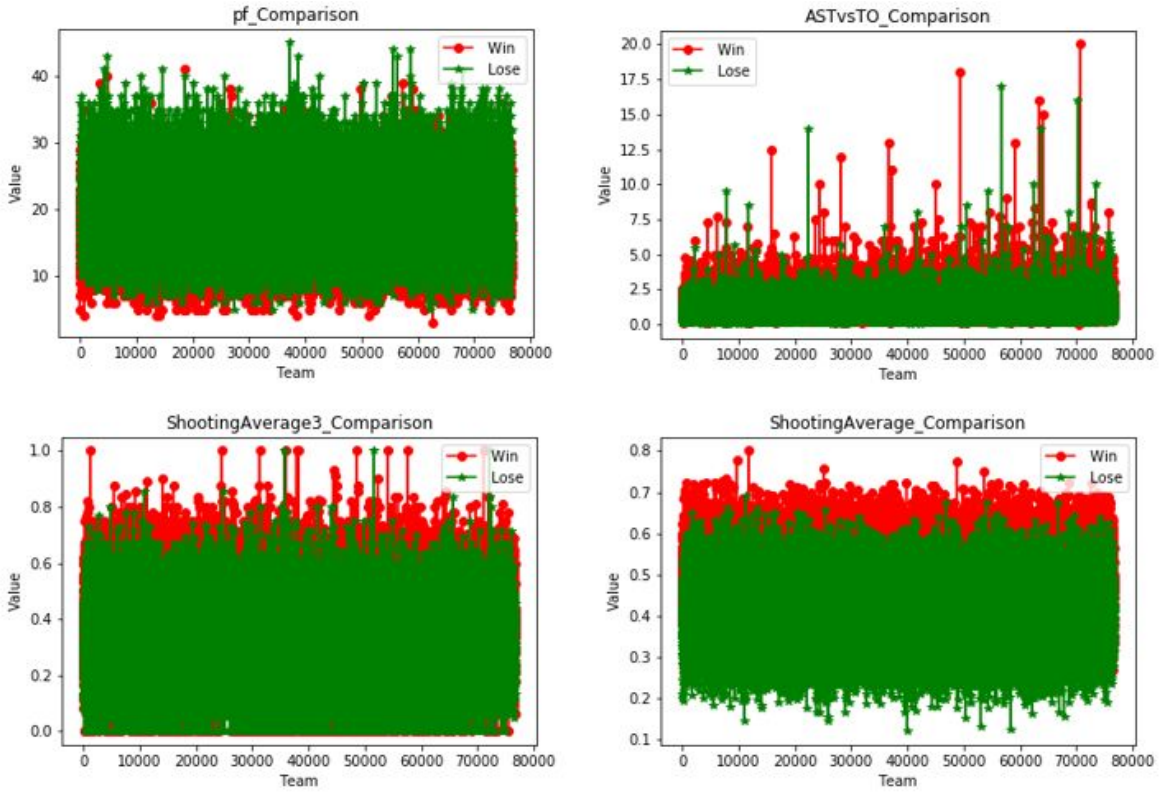

Fig.6 Comparison

From the above figures, we can easily find among all the statistics, the shooting average of winning teams are completely higher than that of losing teams. It seems intuitive, for those who shoot more accurately will definitely have higher chance of winning a game. Meanwhile, the number of assistants is another crucial one, indicating that if a team can move the ball more smoothly and move the ball better, they have more chance to win. Also, the number of defensive rebound of winning team is largely higher than that of losing team, because if one cannot control its defensive rebounds properly, it will give chances and maybe points to its opponents, leading the team to lose.

According to our best performing classifier, logistic regression, the probabilities generated for each team show some interesting yet straightforward trends with a few basic aspects of the basketball games as shown in Figure 7 below: teams are less likely to win with more combined turnovers and fouls compared to the opponents, but are more likely to win with higher assist-to-turnover ratio than the opponents, or higher total rebounds (defensive and offense) than the opponents, or higher combined steals and blocks than the opponents, which should not be surprising to anyone who knows a little bit about basketball but it's helpful to visualize these impacts through our machine learning classifier generated outcomes/probabilities. More specifically, these plots tell us the following: if all tournament teams' regular season per-game performance are taken into account, if these teams have lower amounts of turnovers & fouls, higher amounts of assist-to-turnover ratio, higher amounts of total rebounds, or higher number of steals & blocks in the march tournament, then they are more likely to win any match up (higher winning probabilities) as predicted by our classifier.

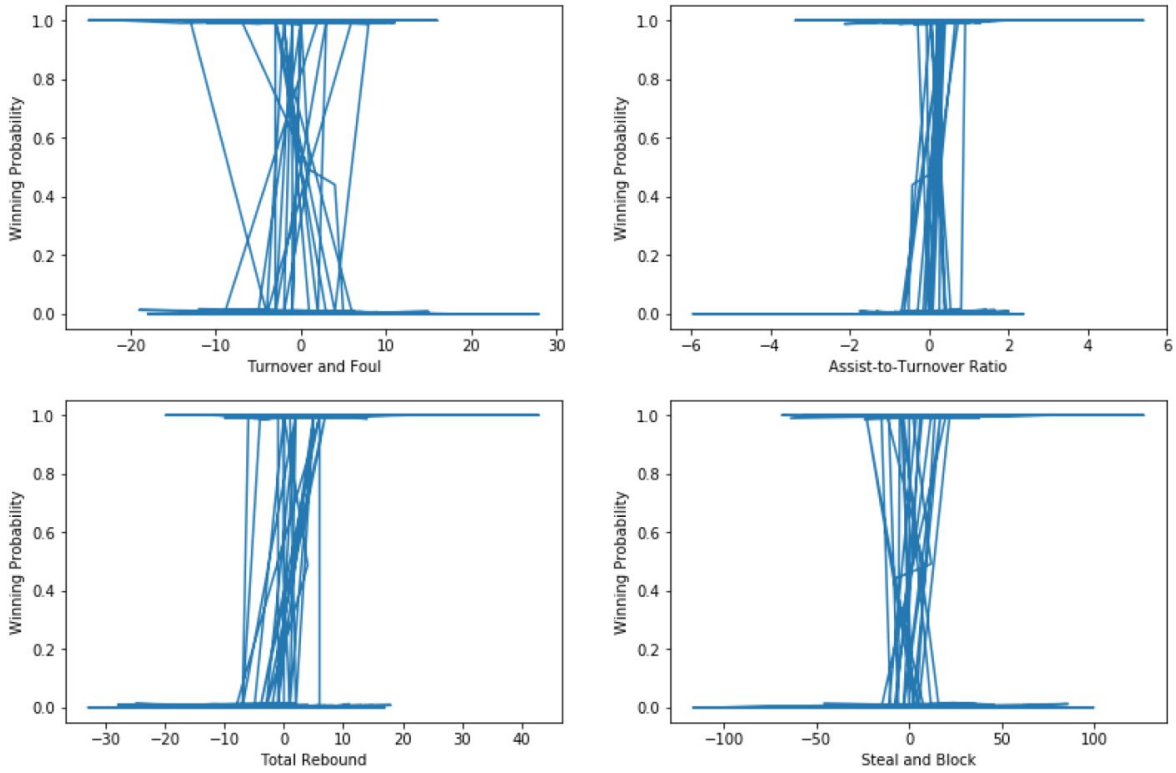

Fig.7

Our team has also used our final classifier on a few individual games/upsets in this year's march madness tournament, and it has worked remarkably well in predicting the outcomes: 16th seed University of Maryland-Baltimore County upsetting the 1st seed University of Virginia is predicted correctly as the first instance of such upset in the history, 13th seed University of Buffalo eliminating the 4th seed University of Arizona filled with NBA talent is also predicted correctly, and a closer match up between University of Missouri (8th seed) and Florida State University also produces a correct prediction. This reinforces the statement earlier that how college teams perform in the regular season can significantly impact how they fare in that year's march madness run.

4 Conclusion

In this project, we worked on the prediction of the NCAA games outcome and investigated some of the key factors that would affect the performance of each team. We tried nine different classifiers and built the models accordingly, finally we chose Adaboost and Logistic Regression as the set of optimal classifiers, according to the test accuracy metrics, which include classification accuracies of 94.2% and 97.2%, respectively, and their learning curves displaying minimal training and testing error as well as their F1 scores (0.942 and 0.9721). With these kinds of performances, we can announce that we are able to accurately predict the outcome of a series of games with our two chosen classifiers.

Besides the accuracy measures of our trained classifiers, we selected the key factor(s) that can impact the game results. From figure 3- 6, we could find that teams with higher number of assists or defensive rebounds are more likely to win the matchup. What's more, teams with

lower number of turnovers hardly lose the matchup. In addition, it is intuitively straightforward that team with higher percentage shooting average would win any match up, like shown in the graph. We have also investigated combinations of some factors, rather than using single factor, on their impact on game outcomes, and as shown from figure 7, we could conclude that if teams in the march tournament have lower amounts of turnovers & fouls, or higher amounts of total rebounds, assist-to-turnover ratio and steals & blocks, they are more likely to win any match up (higher winning probabilities) as predicted by our classifier.

5 Acknowledgement

We would like to show our greatest appreciation to Kaggle.com for providing comprehensive datasets of historical NCAA games. Also, we will thank Professor Zachary Ives and Professor Lyle Ungar for giving us useful and practical advises and teaching this great course.