DS3000 Final Project

12/9/24

Benjamin Welsh, Lee Tamir, Jingcheng Tao

Abstract:

In this project, we used National Basketball Association (NBA) player and team data to answer two questions: can we use team statistics to predict the amount of wins a team will get, and can we use player statistics to classify all star and non all star players? To answer the first question, we used a multiple linear regression model taking in offensive rating, defensive rating, and turnover percentage to predict total wins. This model proved to be very effective, predicting with a high r^2 and low MSE. We answered the second question with a random forest classification model taking in player points per game, total rebounds per game, and games played to predict all star status. This model yielded a high accuracy in predicting all star status, but when taking a closer look there we discovered slight problems.

Introduction and Data Description:

We designed two models to help us better understand the relationships between player/team statistics and player/team success. To measure player success, we decided the best metric for this was all star selections, and for team success we decided the best metric was total wins.

We chose the features for our models, we decided to opt for three variables for each model, hoping to show that just a few statistics can lead us to solid predictions on player/team value.

The data pipeline for this project began with scraping the data from basketball reference, a highly reputable site for all things NBA statistics. We read the html tables with pandas, which gave us an uncleaned dataframe. For player data, we decided to just use the last season, yielding us a dataset of over 700 players and 30 all stars (will be talked about later). However for the team data, we went back 9 seasons for a total of 270 teams.

To clean the player data we used one hot encoding to separate the player awards (all stars, MVPs, …) to numerical values. Cleaning the team data was slightly more tedious. We removed many unimportant statistics such as attendance numbers and home stadium. Additionally, we standardized all team names to abbreviations, which in retrospect was unnecessary.

To predict team wins, we chose to use offensive rating (how many points a team scores per 100 possessions), defensive rating (how many points a team gives up per 100 possessions), and turnover percentage (what percentage of possessions end as turnovers) as our three features.

This was because offensive rating provides a good sense of a teams offensive production, likewise for defensive, and turnovers are very costly for a team.

To predict all star status, we chose points per game (amount of points a player scores per game), total rebounds (offensive and defensive rebounds per game), and games played as our features. We understood going into this good looking statistics are very important in earning an all star vote, and games played is also something that is considered.

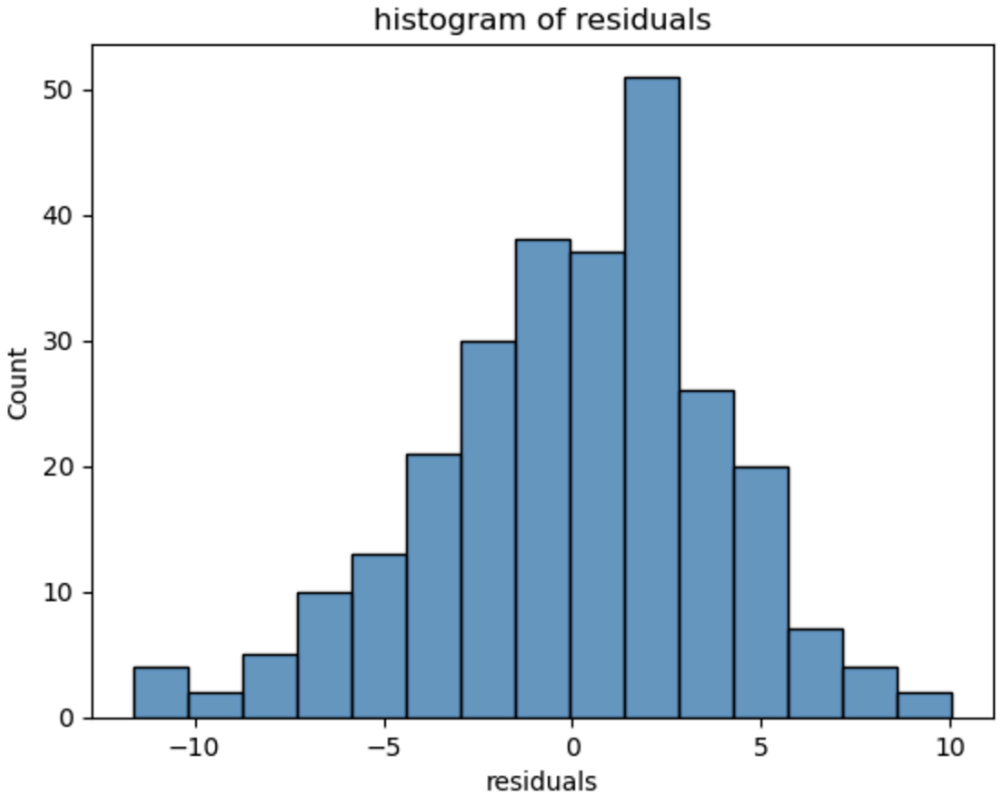
Method:

To ensure efficiency and accuracy, we decided to use multiple linear regression and random forest classification as our two models. We used multiple linear regression on our first goal, which was predicting the amount of wins a team had based off of their stats that season. We thought that multiple linear regression would be a good fit for this data because it was clear to us there was a linear relationship between our several features and team wins. To be more specific, we looked at three stats in particular: Offensive Rating, Defensive Rating and Turnover percentage. We believed that these three stats are strong indicators of how well a team would play, which directly translates to wins. We would do this by observing already finished seasons, and using the model to predict the wins based on those seasons stats, and then compare at the end to see how accurate the model was.

For our second method, we will use the random forest classifier in the sklearn library. We have chosen to focus on Points, Total Rebounds and Total Games Played, as we look to identify All Stars on each NBA team. The idea is to see if our model can take in these three stats, and identify All Star caliber players. We hoped to use a supervised machine learning model for classification. Additionally, because we were limited in our dataset size, we were slightly worried about overfitting, which the random forest classifier is good at preventing. The random forest classifier is the best option to ensure that we can properly predict All Stars based on the stats of the players.

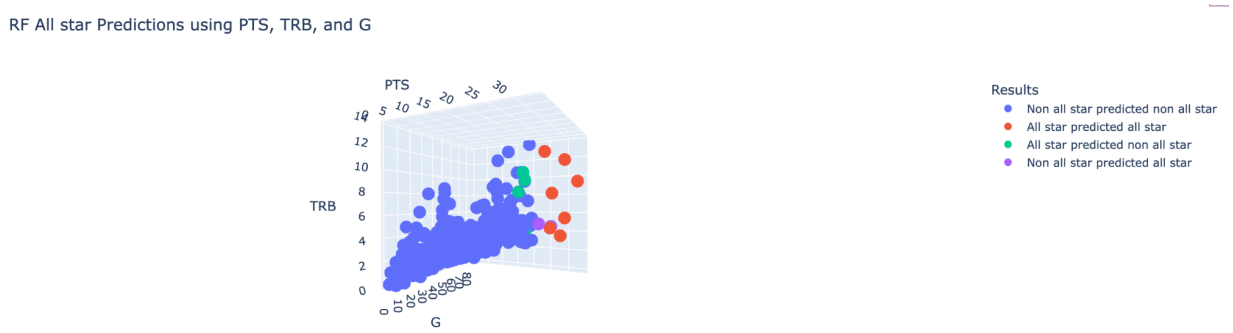
Results:

Fig 1 (Model 1):



As can be seen in the above figure, the histogram roughly follows a symmetrical, normal distribution, ideal for a regression model. Due to the majority of the residuals being close to 0, this indicates that our model's predictions were quite close to the actual values for most of the data, although at times the observed value was slightly larger than the predicted value.

Fig 2 (Model 2):



This figure is a 3d graph of the player data, with color data showing the relationship between actual and predicted all star status. As you can see, there is a vast majority of non all-star players in the NBA, and of players with high statistics (top right), there were several incorrect predictions (colored green and purple). This figure made us question the accuracy of our model, as will be discussed later.

Discussion:

Model 1

For our first model, we clearly saw that picking ORTG, DRTG, and TO% of a team were very strong indicators of how many wins a team had. For example, our ML model correctly predicted the 2024 Boston Celtics, who won the championship, to be the best team, and predicted the 2016 Philadelphia 76ers (who had a record of 10 wins to 72 losses) to be the worst team. Furthermore, our ML model generated an R^2 of 0.8958476288965742, showing that our ML model is quite accurate in predicting the actual data. Our ML model also generated an MSE of 14.706814990817946, which shows that there is a low variance in errors in our ML model. Although our model is quite accurate, it is not perfect. Thus, you could take its predictions at face value, but they might not always be correct. Due to this, we might want to see if any other team analytics correlate more to the success of a team, and add that analytic to our model to make it more accurate. Something that arose from our work, is that we noticed that two seasons (2020 and 2021) were shortened due to COVID, so we had to figure out how to navigate through this, as teams had less wins and losses because they played less games, but their Ortg, Drtg, and TOV% stayed relatively normal. To ensure the number of games didn't matter, we could have had our model instead predict team win percentage instead of total wins. Lastly, we determined that this model does not have any ethical concerns, as there is no moral or physical harm that is raised from predicting the number of wins of a team (unless it hurts the team's feelings).

Model 2

For our second model, we saw that using a player's total points, total rebounds, and total games poised to be very accurate in predicting if a player made an All-Star team that season. Our model had an accuracy of 98.4%. Although this accuracy may appear to be high, it is mainly due to the fact that there are a lot more players that aren't All-Stars than are All-Stars, so it can easily

predict that players with poor stats aren't All-Stars. In reality, we likely can't accept these results at face value, as when it came to predicting players who had very high stats, it wasn't nearly as accurate as the accuracy score suggests. Due to this, we may have to change some of the data points we used. Along with this, some of the inaccuracy could be due to injuries players attained around the time of the All-Star game. On the same topic, while working we noticed that this was a concern we had, and in the future we would have to find a way to not include players that missed the All-Star game due to injury in our learning dataset to make our model more accurate. For our model, ethically, it should not be used by the NBA to actually determine what players make or miss the All-Star game, as in the real world, there are a lot more things that go in to a player making the All-Star team than only the data we used, so some players may wrongfully miss the All-Star game with our model.