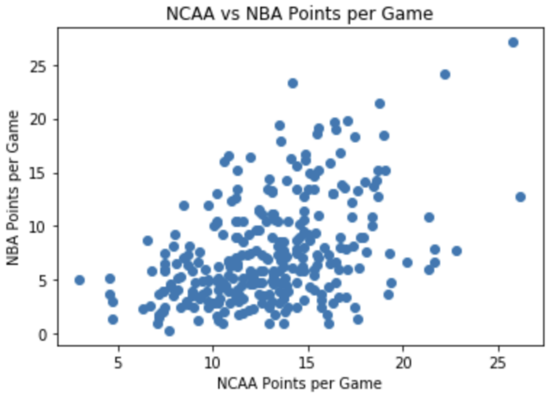
Making Wise Draft Picks

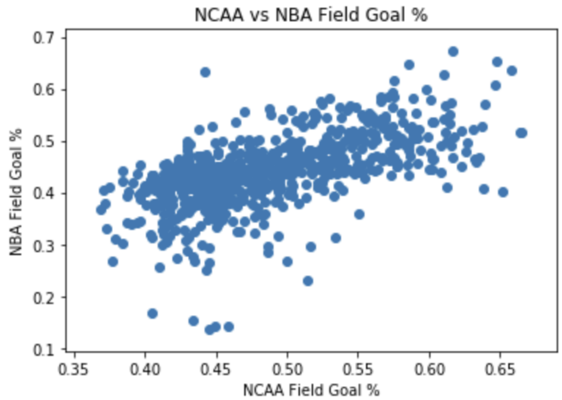
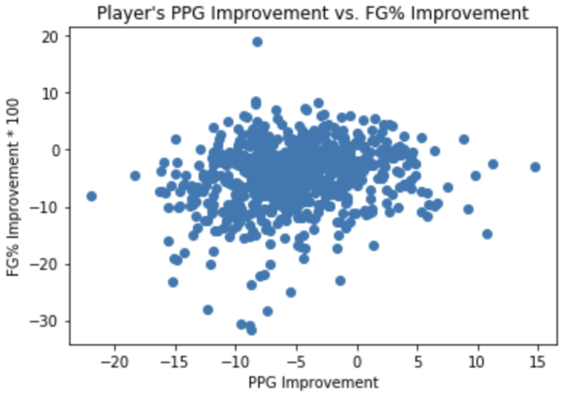
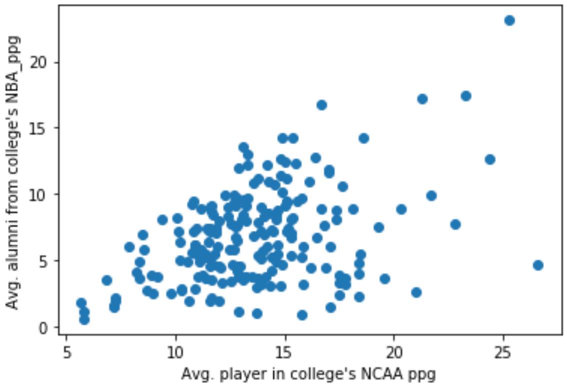
Team: Warren Wang, Kaycee Pham, Daniel Gultom

Abstract: While the skills of the team may determine the outcome of a basketball season, many other factors are also at play. In particular, the NBA draft is perhaps one of the most anticipated sport events every year and decisions are often scrutinized. This project attempts to explore how NBA teams are affected by their college recruits, and how a team might improve due to the addition of a player. Specifically, we will explore a few methods in which one can predict how a college player will impact an NBA team's record in their first season.

Make sure to number figures and tables and include informative captions. Specifically, you should address the following in the narrative. Note: There is a page limit of 6 pages, excluding figures and tables.

- What type of question are you trying to answer with the data.
  - Given some form of information on a player's or a team's past records, can we predict their performance in the future, and whether or not the team that drafted them will improve due to their addition?
  - We hypothesize that if a player performs well in his college career, he will then be drafted onto a team that will improve due to his addition. Inversely, if a player performed poorly in college, the team that drafts him will not improve in the following year.
- Perform exploratory data analysis (EDA) and provide data visualizations.
  - We began by trying to explore what stats may have strong relations between college and NBA careers so that we could determine what features would be relevant for our model.

	We compared PPG for both NCAA and NBA, but did not see too strong of a correlation.

	We compared Field Goal % for both NCAA and NBA, and noticed that no matter the FG% value in college, most people perform the same in the NBA—the correlation is quite subtle.
	We created two columns that displayed a player's raw improvement by getting the difference between their NCAA values, and NBA values (ppg & fg pct). We wanted to see if a player's improvement was linear across the points scored and accuracy in field goal percentage. We noticed that the field goal improvement stayed quite stagnant and had little to no correlation to the points per game improvement.
	We tried to also see if the college someone went to affected performance. We aggregated the college dataset by the name of the university, and took a look at some of the statistics. The following image is a graph in which each point represents a college, and the x-axis is the avg. player NCAA ppg and the y-axis is the avg. player NBA ppg.

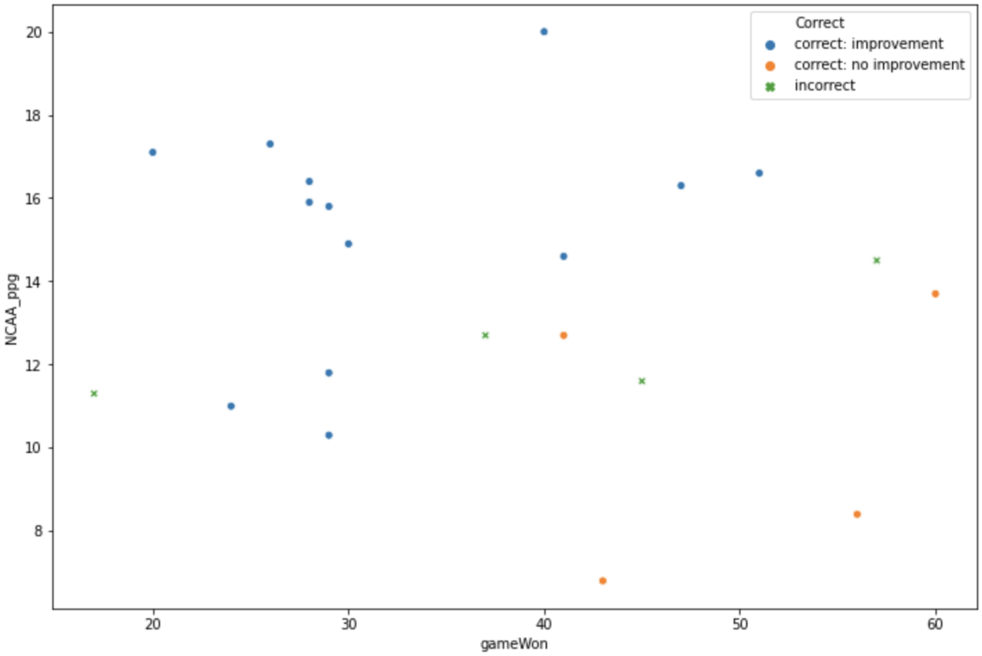
- Describe any data cleaning or transformations that you perform and why they are motivated by your EDA.
  - Data cleaning for player improvement from the NCAA to the NBA:
    i. First we filtered the college table so that it would have records only during and after the year 2012. We also made sure that there weren't any NaN values in the NCAA columns that we were keeping track of (NCAA_ppg & NCAA_fgpct). This was motivated by our EDA because any NaN values would cause errors when comparing the values for improvement.
    ii. We then filtered the table once more, to ensure that the rows only contained players who played over 10 games in each league, NCAA and NBA, so there would be a proper trend. This was motivated by our EDA because we noticed

several outliers that might have been caused by players who had very few games played, resulting in rare percentages of 100% and 0% for the FG%.
        iii. We tried to filter by playing position, since we are keeping track of points scored, and not all positions are designed to focus on scoring (i.e. defense), but did not see much of a shift in trend, thus decided not to focus on it too much. The same thing occurred for college attended, thus we left those two categories alone.
    - Data cleaning for which player got drafted onto which team, and what his impact is:
        i. First, find the first team that a player joined (aka drafted team) in 2012 or later.
        ii. Next, filter to only consider players that had significant playing time in their first season (at least 10 minutes per game).
        iii. We also noticed that there was a scenario in 2012 where the Boston Celtics played 81 games instead of the standard 82, therefore their gamesWon was a NaN value. We patched this by manually researching how many games were won in that year and filling in the NaN value.
- Carefully describe the methods you are using and why they are appropriate for the question to be answered.
    - We decided not to hot encode colleges or position, because during our EDA stage we realized that these two factors did not significantly affect how well a player performed in terms of PPG and FG%.
    - After cleaning our data and creating certain features (improvement, collegeValue) as described above and below, we decided to train multiple models (LinearRegression, Logistic Regression, and Random Forest) and compare their accuracy on the same training and testing data.
    - We discovered that a random forest was most appropriate for our question because we had clustering within our data. We saw a strong correlation between a player's PPG and the improvement that the team he joined the following year.
    - Inversely, there is a cluster of teams with low improvement if the prior season had many games won.
    - We also decided to incorporate minutes played as a feature, because we saw a strong correlation between how much a new player played on his drafted team and how much that team ended up improving in the season that he first joined.
- Create a complete statement of the model and assumptions they are using for inference.
    - Our random trees model was fitted using these features: 'college_value', 'playMin', 'gmDate', 'teamAbbr', 'year', 'name', 'gameWon', 'improvement', 'height', 'college', 'position', 'weight', 'NCAA_games', 'NCAA_ppg', 'NCAA_fgpct', 'NBA_ppg', 'NBA_fg%', 'active_to', 'active_from'.
    - This model predicts whether a team will improve if they draft a certain player. We defined improvement to be either 1 or -1, based on whether or not the team's record

was better during the first season the draftee played (compared to the team's previous season without the draftee) (1).

- Assumptions:
    i. All teams played 82 games per season starting at the first recorded date in 2012 of the table, with the exception of Boston in the 2012-2013 season.
    ii. Because we only had player information, we attributed team improvement solely on the addition of a new player from college. Just because a team improved/regressed does not necessarily mean that it was due to the player. There are many external factors that we couldn't train our model on, as we lacked the data (change of coach, injuries, free agents joining the team).

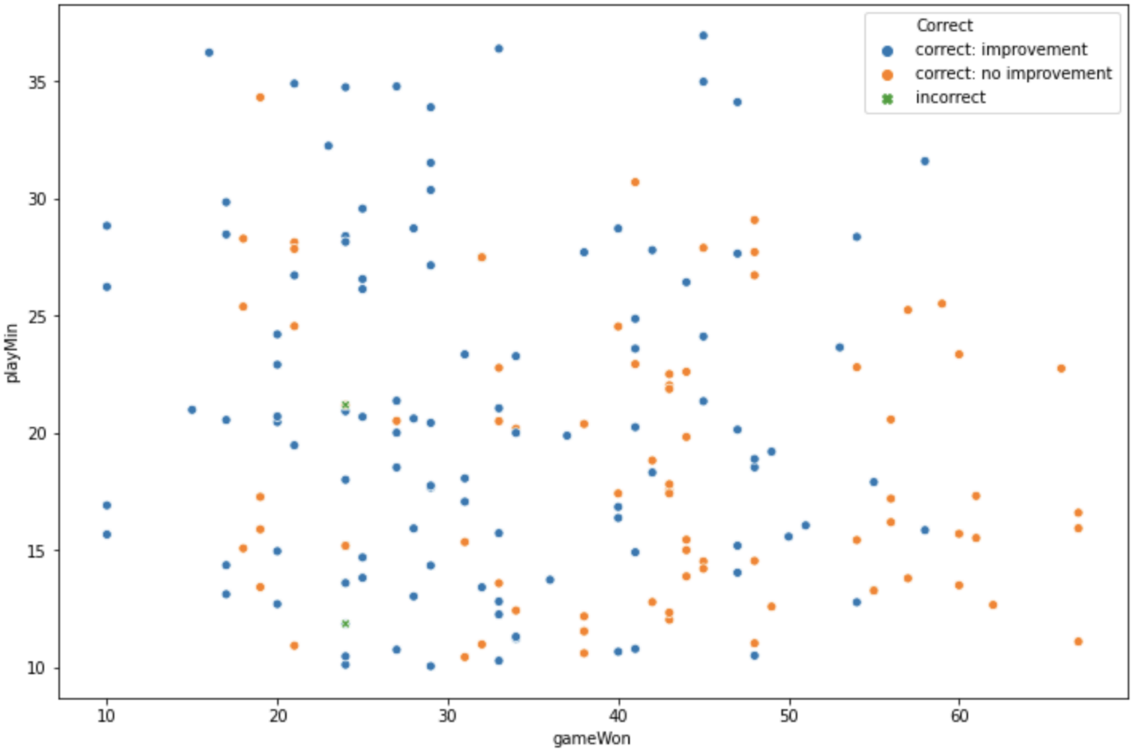
Summary and Results



This visualization of our test results demonstrates how two of our more important features, NCAA_ppg and gameWon, provided a clean separation of our data points.

Our hypothesis was that players who scored a lot in college and played for a bad NBA team probably helped to improve that teams record (and the opposite vice versa).

Here we see that our model correctly predicts a team to improve in that exact scenario (lots of points and low gameWon).



This is more of a fun visualization of our training set, specifically how minutes played and gamesWon affected our predictions. While the relationship isn't as clear as above, it does show that we tended to correctly predict no improvement when a team won a lot of games and drafted a player they did not give many minutes to. The opposite was true for bad teams that played their rookies lots of minutes.

Addressing the following seven specific questions.

1. **What were two or three of the most interesting features you came across for your particular question?**

- gamesWon: This proved to be a very important feature, as we noticed a trend in improvement from one season to the next is quite dependent on how many games were won to begin with. We realized that this probably has to do with the fact that teams who do poorly in a season are more likely to draft better players for their next season, since teams with lower records have higher priority in the draft.
- college_value: We created a metric called collegeValue that determined how competitive a player's college was. We did this by dividing their average college score with their average NBA score, to properly "scale" their college. This was done because we knew that different

colleges played in different divisions, thus we wanted to assign a weight to see which colleges were more competitive/more likely to produce a strong player.

2. **Describe one feature you thought would be useful, but turned out to be ineffective.**

- We thought that paying attention to the position of the players might help with accuracy, as we thought that some positions were less intended for scoring points (defense players), and our features most definitely depend on the points scored per game. However, upon categorically analyzing points scored per game for each position, we were surprised to see no significant differences.

3. **What challenges did you find with your data? Where did you get stuck?**

- A challenge that we encountered was finding enough data. Ideally, we would have information that went before 2012, but we're limited by the datasets we had. Because we have NBA data only after 2012, it caused us to only use people who went to college around the 4 years before 2012, thus causing another limit.

4. **What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?**

- A large limitation was that our college data set spanned over a much larger period of time than our box score data did. This meant that we were unable to incorporate a lot of the data from college.csv, thus reducing our dataset by a significant amount.
- Since basketball seasons are not exactly within one year, but span among two, we had to make the assumption that a season includes the current date that a game was played, up until 82 games into the next year. We also had to assume that a rookie played his first NBA season before the new year, otherwise we could potentially have been counting a few games in the following season. This is because each season has 82 games, so we felt like we could assume that the data was complete and split our seasons with this method. This could prove to be incorrect if a couple of games were not recorded in the dataset, thus throwing off our sectioning of seasons, and improvement data.

5. **What ethical dilemmas did you face with this data?**

- Some people did not show up in the college dataset, but they were also drafted by the NBA in the same season that other players got drafted. Since we wanted to specifically see how a draftee's college career affects NBA performance, we decided to ignore those people. This could skew the data because those new players who did not appear in the college dataset may have actually made significant contributions to how the drafting team does the next season.
- We also only used data for new players who played more than 10 minutes in a game. This may be considered too small of playtime to be considered "contribution" to a team, but we needed to have enough data to create significance. This might exaggerate data when a player just joined a team and they do better, when it may have had very little to do with that new player's contributions.

6. **What additional data, if available, would strengthen your analysis, or allow you test some other hypotheses?**

- We would try to get more detailed data on a player’s college career. It would be interesting to see how specific factors such as college coaches and in-game performance affects how someone would do in the NBA. We would also have wanted to check out assists and rebounds per game in the college dataset. Unfortunately, because those stats weren’t there for the college set, we couldn’t use those in the NBA stats because we can’t compare them.

7. **What ethical concerns might you encounter in studying this problem? How might you address those concerns?**

- As we know, causation is not the same as correlation. While trying to recruit the best player possible for your team is good, some factors may be more exaggerated than others. For example, if we were to find some correlation between the college a player attended and their future performance, this may cause a positive feedback loop, where teams are biased for or against certain colleges when it may not actually strongly impact a player’s performance.
- To address this bias, it may do well to look into a variety of factors that determine a player’s history and look at them holistically to ensure that players don’t get drafted based on things that don’t actually determine success.

Provide an evaluation of your approach and discuss any limitations of the methods you used.

- Our approach was largely guided by the data we were provided. For example, our biggest limitations were our features being limited by our dataset columns and our accuracy being limited by the amount of time recorded in our standings table (2012-2018). If we had access to more college stats (efg%, assists/game, rebounds/game, coach, stats for each year in college), we could’ve utilized more meaningful features.
- Our approach was also guided in part by intuition. We made educated guesses on what features/factors would impact improvement (ppg, minutes, etc) and backed up our guesses by looking at the data visualizations.

Describe any surprising discoveries that you made and future work

- One thing that actually surprised us was during data cleaning, when we discovered that the number of players that fulfilled the criteria that we wanted to consider is actually quite smaller than we anticipated (209 players). In future work, we would definitely look into broader and deeper datasets, broader as in more features such as rebounds, and deeper as in data not limited to a very few number of years (2012-2018) relative to the NBA’s history.