Eliza van Hamel Platerink
Stefan Schlotter
UC Berkeley

Data 100 Final Project

1. Project Set-Up

The open-ended, start-to-finish analysis of a dataset for this final project left us, although originally daunted, with an opportunity to express our curiosity and demonstrate understanding of this semester's core concepts. We wanted to use this unique opportunity for a final, given to us in a time of a looming global pandemic, in order to delve deeper into modeling and prediction, and to explore a dataset that we not only were interested in, but could be used to make a difference in the future.

After some investigation into the different data sets, we finally settled on exploring the Basketball data set. We picked this set simply because it sparked more interest for us than COVID-19 and Contraception sets. We felt that there was a lot of interesting information to sift through in the basketball set and it provided the best opportunity to find something insightful.

2. Question Framing

In order to decide upon a question, we spent some time looking through the columns and deciphering what many of them meant. We quickly realized that we had access to an endless amount of statistics. We wanted to use these statistics in order to glean something meaningful. We wanted to use what we have learned in order to create a classifier that can help predict a team's winning percentage for a season based on different features of the team. In sports, basketball in particular, a winning percentage is the fraction of games or matches a team or individual has won. In our case, we focused mainly on the teams.

The set includes 5 tables -- a college player's statistics, the box score information for NBA players over the last 7 years, the NBA team box score information over the last 7 years, an official box score for teams over the last 7 years, and a standings table of rankings of each team. We ended up deciding that the college table would not be valuable in answering our question about the data; in addition, the table helped many NaN values and this made it unappealing. There was an obvious connection between the NBA tables, and we wanted to exploit that connection in order to see if we could predict the winning percentage of a team based on the players statistics. Our official question is:

> *What factors can we use from the given data in order to predict the winning percentage of teams in the NBA?*

This would require investigation into the 3 NBA based tables, a clear understanding of the player statistics and their meanings, and removal of columns that would either cause overfitting, or would generally cause outliers and be useless. We would then create a model (or three!) that we could use to predict winning percentages.

3. Exploratory Data Analysis (EDA)

We wanted to highlight a clear understanding and interpretation of the two major tables we used, before we properly cleaned the two sets.

	Official Box Score	Team Box Score	Standings
rows	155713	14758	29520
columns	51	123	39
# qualitative ordinal	13	10	3
# qualitative nominal	10	8	2
# quant. discrete	21	43	18
# quant. continuous	7	62	16
Each row represents...	A player	A team	A game
Any other notes.	Coarse grain Foreign key in this case is the 'teamAbbr'	Primary key is 'teamAbbr'	Primary key is 'teamAbbr'

Beyond the granularity of the data, this phase of the project highlighted that we would likely be grouping together the tables in order to have one succinct dataset. It also became clear that the columns we would use as features in our model would have to be hand picked carefully in order to achieve the best results.

4. Data Cleaning

We then cleaned our data. After some investigation, we quickly realized that we would not have use for all of the columns. We also realized that there were no places where the NBA data was missing (i.e. no 'NAN' values) that could mess up the computation/model.

Cleaning the Standings table:

To start, we need to make sure that for every season between 2012 and 2018, we had the win-loss record of every team after 82 games. 82 games is the specific number because that is how many games are in the regular season.We removed all values that were not regular season games, because in the playoffs, the best teams are playing the best teams so those losses can tarnish the winning percentages of the best teams. One problem we encountered while doing this was that many teams finished playing 82 games on different days, so we had to find the specific day when every team had played 82 games or risk adding up one team's win-loss record several times. We then selected the day for each season where everyone had played 82 games and grouped by team. This gave us the cumulative win-loss records for each team from 2012-2018 in the regular season (from this point it was easy to divide wins by total games played to get winning percentage).

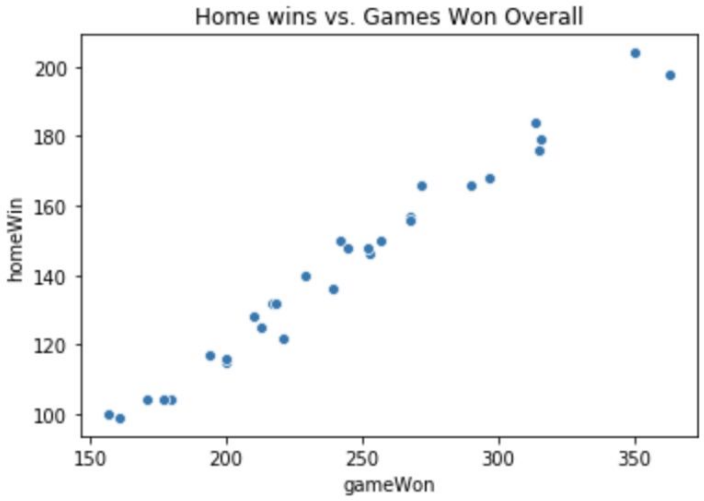
Cleaning the officialBoxScore and playerBoxScore tables:

The officialBoxScore table consisted of a team's statistics for every game they played between 2012 and 2018. The playerBoxScore table consisted of each player's statistics on each team for every game

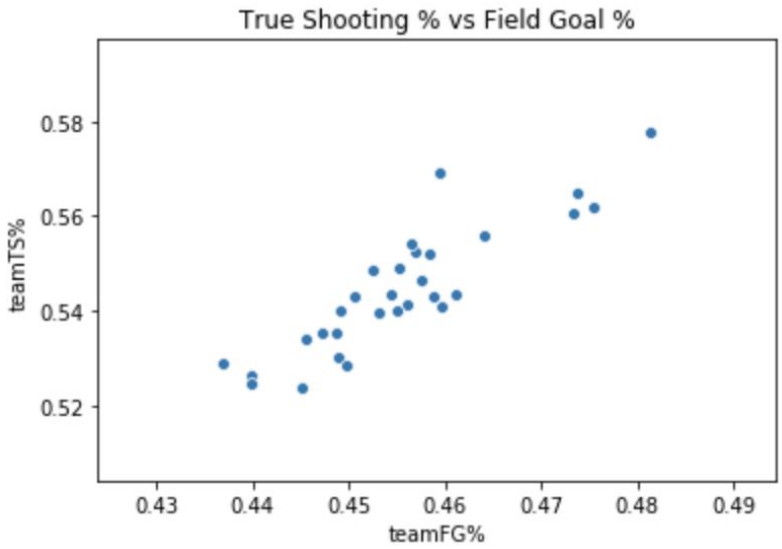
between 2012-2018. It may seem weird that we decided to clean and use playerBoxScore at all, but the table gave us access to features such as the average age of a player, which we felt might be indicative of a team's record. For both tables, we essentially grouped them by team. However, we quickly realized that this would not be as easy as previously thought because many columns needed to be grouped by mean in order to be viable statistics, and many had to be grouped by sum. To solve this problem, we created functions to quickly drop columns we did not need and to split our data set into the columns by which we would be aggregating by sum, and those by which we use the mean. We then grouped them and recombined them.

After cleaning the tables, we grouped them together based on team (the "teamAbbr" column). It was then time to begin selecting which columns we would use as features in our model. Outlined below is a walk-through of our thinking on how we selected some features along with visuals that supported our opinions:

- The following graph illustrates that games won and home wins are directly correlated. This intuition is the reason that we removed awayWin, awayLoss, homeWin, and homeLoss from the data. GameWon and GameLost were removed because we already have calculated a winning percentage column, and the other features were clearly related to winning percentage.

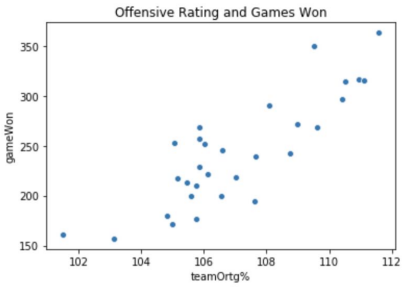


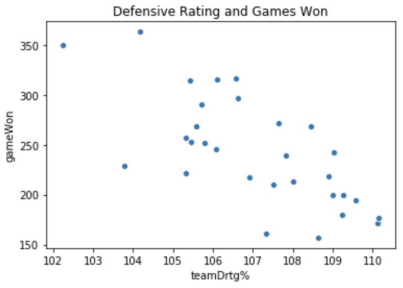
  -
- Using several shooting percentage statistical markers would also jeopardize the accuracy of the model. As the plot below illustrates, stats such as FG% and TS% are very correlated. We decided to only select the true shooting percentage stat, which combines together three point shooting percentage, two point percentage, and overall field goal percentage.



- As can be seen below, the interquartile ranges for OREB% and DREB% average out to roughly TREB%, leading us to only use TREB% as a feature.

- As can be seen below, offensive rating and defensive rating seem to correlate well with winning percentage. I decided to keep both these features because they tell you about the efficiency of a team offensively and defensively, but do not directly relate to winning percentage or any other variable in any way.





After this arduous process of sifting and cleaning, and then more sifting and cleaning, we standardized our data. It was then time to perform our train-test split, so that we could preserve some data for our testing. The data we had downloaded was all the data we had available for both training models and testing the models that we train, so we therefore needed to split the training data into separate training and testing datasets. We decided to split the data into 80% training data and 20% test data. For our purposes, this meant that we were training our model on 24 of the teams and testing the model on 6 of the teams.

5. Methods and Experiments

When it came to modeling our data, it became clear the Linear Regression would be the best modeling option. Linear Regression models the relationship between variables and a observed variable by fitting a linear equation to the data. While exploring our data, we found that many statistics such as offensive rating seemed to be linearly correlated with the data, leading us to believe Linear Regression would be a natural and accurate choice. Since all our features were numerical, tricks such as one hot encoding and regularization were not necessary, so we decided to use the skLearn LinearRegression model without a pipeline or any additions. Obviously, for each model, we fit the data and then used the model to predict winning percentage.

We went through three model processes, before we finally settled on one we thought would result in the best prediction.

Model One:

Through the data cleaning process, it was brought to light that games won and home wins are directly correlated. This intuition is the reason why we removed awayWin, awayLoss, homeWin, and homeLoss from the datatable. GameWon and GameLost were removed because we already have calculated a winning percentage column, and the other features were clearly related to winning percentage. Using several shooting percentage statistical markers would also jeopardize the accuracy of the model. Stats such as FG% and TS% are very correlated. We decided to only select the true shooting percentage stat, which combines

together three point shooting percentage, two point percentage, and overall field goal percentage. We used the same reasoning to select only true rebounding percentage, and not all the rebounding stats.

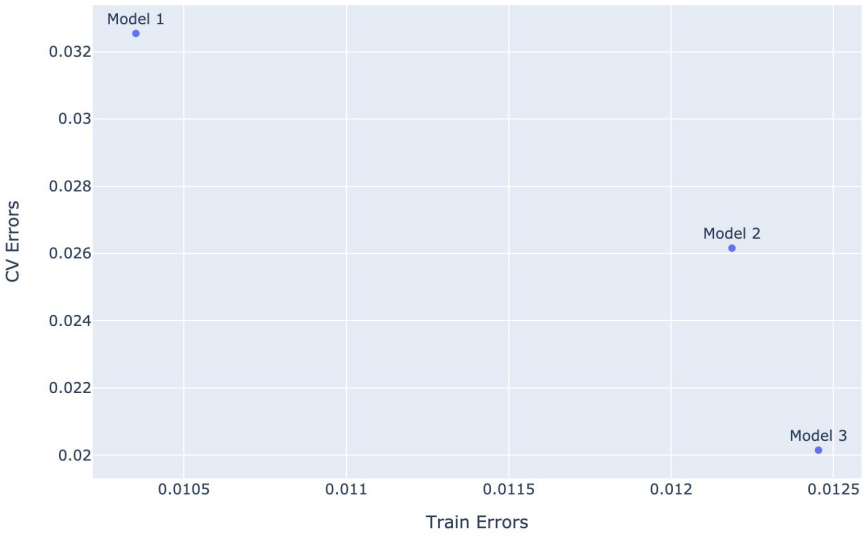
Model Two:

We took out average age, pace, and possession. We both agreed that the style of how the team plays and the age of the player could influence the team's winning percentage, but it was also just as likely that it was of no use at all. For the sake of the second model, we made the assumption that the average age of a team does not seem to be indicative of a team's winning percentage because a team's star players could be young or old. Pace and possession times/averages also felt like they could be unimportant; a team can win games playing a fast or slow style.

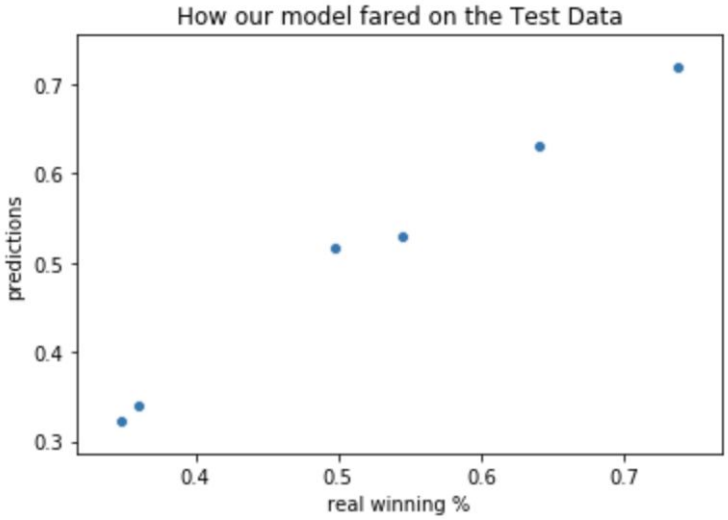
Model Three:

After further thought, we decided that points for and points against may not be helping increase the efficiency of the model. Statistics such as assist percentage, offensive rating, defensive rating, etc. help to emphasize the efficiency of a team. The amount of points scored by a team or the amount of points scored on a team could be related to pace of play and may not point to efficiency. If a team scores very easily and often, they are likely going to give up more points through more defensive possessions. For these reasons, in our third model we decided to eliminate the "ptsFor" and "ptsAgn" features.

As can be seen below, the first model has the lowest training error, but also the most cross validation error, which led us to believe that the training data was being overfitted by our features. As a result, we removed several features (as explained earlier) and tried a few more models until we were comfortable with the balance of the CV and training errors. It was acceptable for us that the training data went down because we were concerned with how the model would work on the test features and not overfitting the training data.



Once we decided that we were satisfied with model three, we ran our test data on the model and discovered that our predictions were no more than a couple wins/losses off from the correct winning percentage for any of the teams in the test part of our data, as shown below:



Seven Questions:

1. The two most interesting features we came across were the offensive rating and the defensive rating of the teams. According to *Basketball Reference* (https://www.basketball-reference.com/about/ratings.html), offensive and defensive rating are efficiency metrics that essentially measure how many points a team scores or gives up every one hundred possessions. We initially struggled with how such a metric may be correlated to all our other features, but ultimately decided that because it takes so many factors such as pace of play into account that it would not directly be related to any other statistic. We removed the points score and pace of play features throughout our models just to be safe, and it ended up improving the model, but we felt that features such as assist percentage, turnover percentage, etc. would be safe to keep in our model because there are many ways to have a high offensive or defensive rating; it does not just come down to shooting percentage or assists or turnovers individually.
2. One feature we thought would be effective, but ended up not being as helpful as we wanted, was the age column. We thought that having younger, healthier, and potentially fitter athletes would allow the team to out-perform other teams. Alternatively, maybe having an older team with more experience might contribute to winning percentage. We thought this feature would be so effective that we even went through the process of extracting this information from another table and joining it as its own column to the team box score. However, it proved itself to be unhelpful for our overall model, likely because age ends up balancing out for success between younger players and experienced veterans.
3. The biggest challenge we faced while exploring our data was during the cleaning stage of the data analysis. There were so many columns to sift through and consider when selecting important features, and it often took a while to decipher the meaning of some columns. In addition, cleaning the standings table proved to be tricky, as described earlier in the report. It was difficult to find the exact dates where every team has played 82 games exactly each season.
4. Some of the limitations of the analysis were that none of the tables gave us access to financial information such as total player salaries; such figures may have helped improve accuracy of our model as well. In addition, coaches, general managers, and other staff were not included in the

model, and such people could definitely impact winning percentage. Another limitation is that if we were to predict a team's winning percentage based on their stats, it would ruin the model if players got injured and changed the teams numbers moving forward. An assumption we made that could prove to be incorrect is that only cumulative team stats matter when predicting win percentage. While we did have access to a player box score table, it would not have been very difficult to track how having a star player (or two or three) impacted wins and losses given that we grouped by teams. We ended up assuming that it would be minimally impactful to perform such analysis, and that could be wrong.

5. An ethical dilemma we faced is that we had to drop many of the columns, because we thought they would not be useful and there were simply too many columns to work with. There could potentially be some very valuable information nestled in the columns we decided not to use. Because we did not have to fill any NaN values, we did not encounter this ethical dilemma, however, if we were to expand our data set, it's highly possible that we could have.
6. An example of some additional data that would strengthen our analysis would be if we had access to 2019 teams as opposed to just up through 2018 (or teams prior to 2012). It would also be interesting to have similar datasets for teams at the college level and see which statistics remain important in such an analysis (the college dataset we were given was just for players and has many NaN values and was therefore not usable).
7. An ethical concern we might encounter in studying this problem is whether or not people could use this information for illegal sports betting. This information could be used on more recent data points in order to predict how a team will do throughout the season. This could be used to automatically generate March Madness brackets, should the analysis be done on an NCAA data set, or could be used to potentially determine the outcome of the playoffs. We might address these concerns by not releasing this model to the general public, or not allowing people to have access to the statistics for the current season in order to predict this season's winning percentage.

6. Final Thoughts

When applying our work to the future, it is easy to see how valuable it could really be. For one, a NBA team could use our model halfway through the season to predict their final winning percentage based on their current statistics. In addition, with more data collected, they could apply our model to winning in the playoffs. We had considered modeling playoff win percentages as well but decided that the sample size of 2012-2018 was too small given how few playoff games are played by each team. In the future, weighting in the value of coaches, how much money a team spends on its players, and how many individual star players a team has are very achievable extensions of what we have accomplished.