Group Members:
Anokh Ambadipudi aa699
Charlie Konen ctk13
Josh Ye jty7
Edmond Niu ewn7
Brian Young by64

Predictive Analytics for NBA Game Outcomes

Part 1: Introduction and Research Questions (15 points)

Given the popularity and large fan base surrounding the sport of basketball, our project idea is to investigate the contributors to the outcome of an NBA basketball game. We believe this data is important to analyze as identifying the factors that are more important and the factors that are less influential in the outcome of a game would be useful information for basketball teams to analyze and consider. We plan on using historical data to identify the factors that best predict the outcomes of NBA basketball games. Some factors we will investigate include player statistics and team performance. Our group is interested in this topic as many of us are passionate about basketball and the potential factors that are involved in the outcome of a basketball game. Some of us are also involved in sports betting and would benefit from a tool that could give us leverage to make smarter bets. Our research questions are as follows:

1. The NBA has seen an increasing number of 3-pointers per game being attempted. Given this increase, we want to investigate whether teams should work on their accuracy of making 3 point shots. Is a higher 3-point field goal percentage correlated with a higher win rate for NBA basketball teams?
2. Are NBA basketball teams more likely to win basketball games at “home” or “away”? Is there such a thing as “home-field advantage”?
3. What other team-related performance ratings (offensive ratings, defensive ratings, turnover percentage, etc.) have the greatest impact on the outcome of an NBA basketball game?

These research questions are relevant to us and also to the broader sports fan base and sports world. As stated above, our group is interested in this topic as many of us are passionate about basketball and the potential factors that are involved in the outcome of a basketball game. Some of us are also involved in sports betting and would benefit from a tool that could help us bet smarter. These research questions could also help the sport of basketball to be more competitive as the findings from our research questions provide suggestions for improvement for each NBA team. Given the trend of increasing 3 point attempts made by NBA basketball teams, we wanted to provide insight as to whether teams should value a higher 3 point percentage, as a form of advice. Diving into the second research question could also reveal points of unfairness in a league that officials could consider moving forward. Lastly, by identifying which other team-related performance factors are most predictive of game outcome, NBA teams can utilize that to improve their odds at winning, making the game more exciting and competitive.

Part 2: Data Sources (15 points)

Our goal is to identify and gather historical game data, player statistics, and team performance metrics. For our project, we elected to use the data contained on Basketball-Reference.com. The main reason why Basketball Reference was chosen is due to its high quality and quantity of data. Data for nearly every game, player, and team can be found on this site, meaning we can gather a large volume of data quickly and efficiently to answer our questions. It also has a simple, yet comprehensive nature, cleanly covering each NBA franchise’s total wins, losses, and various performance ratings compiled over an entire season. Furthermore, the data is easily exportable and encapsulates deep statistics and analytics beyond simply “scoreboard” statistics - i.e., statistics that can simply be gathered from sitting in an NBA game. In addition to the standard statistics like field goal percentage, average game blocks, assists, steals, etc. Basketball Reference also had an “Advanced Statistics” dataset, which included metrics such as Offensive and Defensive ratings, Point Differentials, and Strength of Schedule statistics—these were exactly the team-related performance metrics that we wanted to investigate. For a comprehensive picture

of each team in the 2022-2023 NBA season, we combined three datasets from Basketball Reference: Advanced Statistics, League Standings, and Per-game average statistics. The datasets we used can be found in our Github repository's data folder: https://github.com/EddieTGH/cs216_project/tree/main/data.

In order to combine each dataset, we merged the datasets together by "Team". In certain datasets, the team name was followed by an asterisk. Prior to merging, we used a regular expression string to extract only the team names without special characters. Finally, after the final dataset was merged, we also created three new columns - overall win/loss percentage (WLP), away WLP, and home WLP. Because only data regarding the number of wins for all games, solely home games, and solely away games were available, we had to create these 3 new columns in order to analyze the differences among home, away, and overall win rate percentages. Here is the formula we used: (number of games won / total number of games played) * 100. Lastly, we also converted "3P%", which was a decimal proportion, to "3 Point Percentage", which was an actual percentage through this formula: 100*3P%.

Part 3: What Modules Are You Using? (15 points)

Module 4: Data Wrangling: The data we collected did not exactly have all the variables that we wanted. As such, in order to generate a final data set in which we can base our statistical models and analysis on, concepts from the data wrangling module were used, such as handling missing data through dropna(), importing our CSV data into pandas dataframes through read_csv(), manipulating column names, changing the order of columns, creating new columns that combined and did math on existing columns, etc. We also did some casting with astype(int) for certain columns that we wanted to convert originally string data of numbers to int data. We mostly used the contents of this module in the data gathering/cleaning phases. We used this module to import our data, clean our data, combine columns, create columns, and more to prepare our data for analysis.

Module 7: Statistical Inference: We also use concepts from Module 7. We computed two confidence intervals utilizing a bootstrap distribution using np.mean() to take the mean aggregation value of a column of data. Bootstrap resampling and conducting a 95% confidence interval were concepts in this module that we implemented in analyzing our 2nd research question. We also used Module 3 briefly to visualize this bootstrap resampling distribution. In addition to these concepts, we also performed a hypothesis test investigating for a difference of two means. We utilized stats.ttest_ind() from scipy to determine if there exists a difference in means between home win percentage and away win percentage across NBA teams. We created our own null and alternate hypothesis and interpreted our p-value. These were all concepts core to Module 7. We used this module to answer our second research question through conducting a hypothesis test (difference of two means) and computing two confidence intervals (with bootstrap resampling). We mostly used the contents of this module in the data investigation/analysis phases.

Module 6: Combining Data: We merged 3 data frames (advanced, per_game, and standings) in this project to create our final dataframe. We had 3 datasets imported that all had different information, but they all pertained to the same 30 teams. In order to summarize and ensure that our data is still structured and properly usable, we had to combine the datasets through a merge. We did not want to lose any data so we conducted an outer join of these 3 data frames, on the "team" column. This combined all the data from each of the 3 individual data frames into one dataframe with one entry for each team. We mostly used the contents of this module in the data cleaning phase. We used this module to merge our 3 dataframes into our merged, summarized data frame, with one entry for each of 30 NBA teams that we used for further data cleaning and analysis.

Module 8: Prediction & Supervised Machine Learning: Our data analysis used both linear regression as well as random forest modeling. The linear regression is a model that we learned how to use in Module 8, but the random forest is a model that we learned outside the scope of this course. The linear regression model helped us understand the relationship between our dependent and independent variables, highlighting the direct impact that various team-related performance statistics had on win rate percentage. On the other hand, the random forest model provided insights on variable importance and was crucial for handling non-linear relationships and interactions among variables. We used this module to create our

many linear regression models, we used it to analyze the performance of each through metrics including root mean square error (RMSE), pearson r correlation coefficient squared (r^2), and adjusted r^2. This model also taught us how to create a train-test split that we utilized in many of our linear regression models as well as k-fold cross validation to determine the best hyperparameters for our random forest model. We mostly used the contents of this module in the data investigation/analysis phases.

Part 4: Results and Methods (15 points):

Research Question 1: The NBA has seen an increasing number of 3-pointers per game being attempted. Given this increase, we want to investigate whether teams should work on their accuracy of making 3 point shots. Is a higher 3-point field goal percentage correlated with a higher win rate for NBA basketball teams?

First, we trained a linear regression model that utilized an NBA basketball team's 3 point percentage (percentage of 3 point shots that were successful), to predict the win percentage of that team over the course of a full NBA basketball season. We utilized the open source machine learning Python framework, sklearn, and trained the model with a 80-20 train-test split. The model's statistics are shown below:

```
RMSE: 6.462901457526928
R-squared: 0.5649400054851862
Coefficient: 4.895390751774447
Intercept: -126.0329827777706
Predicted win percentage for a 3-point percentage of 35%: 45.305693534335035
```

Figure 1: Linear Regression Model Output Predicting Win Percentage from 3 Point Percentage

Our model has a root mean squared error of 6.46. Our model's r-squared value is 0.56. 56% of the variability in an NBA team's win percentage over a season can be explained by that team's 3 point percentage over that same season. Our model has a slope of 4.90. All else held constant, for every percent increase in 3-point percentage, the win percentage of an NBA team is predicted to increase by 4.90%. The intercept is -126, which means, all else held constant, when an NBA team has a 3-point percentage of 0%, they have a predicted win percentage of -126%, which, in the context of our data, is impossible and is an error of extrapolation. We also use our linear regression model to predict the win percentage of an NBA team based on their 3-point percentage. An NBA basketball team with a 3-point percentage of 35% is predicted to have a win rate of 45.3% over the course of a season.

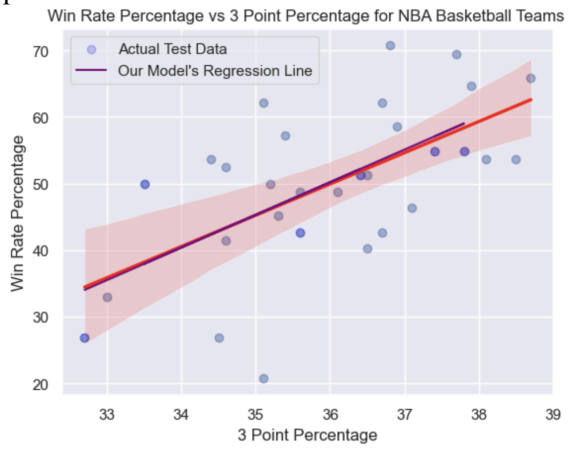

Figure 2: Win Percentage versus 3 Point Percentage for NBA Basketball Teams

We plot the relationship between a team's 3 point percentage and its win rate for the season. As can be seen, we plot our model's regression line (shown in purple) and it is exactly the same as the regression line that Seaborn provides (shown in red). The actual test data is shown as blue dots in the graph. This

plot clearly showcases the moderately strong, positive, linear relationship between a team's 3 point percentage and their win rate percentage.

Conclusion #1: An NBA team's 3-point percentage is moderately strong at predicting a team's win percentage over the course of a season. A team that ends a season with a higher 3-point field goal percentage is predicted to end the season with a higher win percentage than a team with a lower 3-point field goal percentage.

Research Question 2: Are NBA basketball teams more likely to win basketball games at "home" or "away"? Is there such a thing as "home-field advantage"?

We analyze the win rate percentage of basketball games that are played at a team's "home" in comparison to the win rate of basketball games that are played "away". We conduct a hypothesis test that analyzes the difference in means between these two groups:

1. Null Hypothesis: There is no difference in win percentage of games played at home vs away for NBA basketball teams in any given season.
2. Alternate Hypothesis: There is a discernible difference in win percentage of games played at home vs away for NBA basketball teams in any given season.

```
TtestResult(statistic=4.422537243822545, pvalue=4.3475370182099444e-05, df=58.0)
```

Figure 3: Results of t-test for the means of two independent samples of scores

We utilized scipy's stats.ttest_ind to conduct this hypothesis testing. We assume that the populations (home win rate percentage vs away win rate percentage) have identical variances by default. We use this two-sided test for the null hypothesis that the two independent samples (home win rate percentage vs away win rate percentage) have identical average (expected) values. The p-value is 0.0000435. We fail to reject the null hypothesis. There is convincing evidence that there is a discernible difference in win percentage of games played at home versus away for NBA basketball teams in any given season.

The data we were provided included the win rate percentage for home game and away games for 30 NBA basketball teams in 2022-2023. We assume this data to be a representative sample of these 30 NBA basketball teams in any recent past year. We utilize a bootstrap resampling (because n is small) with 10000 resamples in order to simulate the win percentage for these NBA basketball home games and away games for any given season (with these 30 teams) given solely these 30 teams results in 2022-20223. Here are the results:

```
95% Confidence Interval (Away Games): (37.72357723577236, 46.34146341463415)
95% Confidence Interval (Home Games): (52.52032520325203, 63.57723577235772)
```

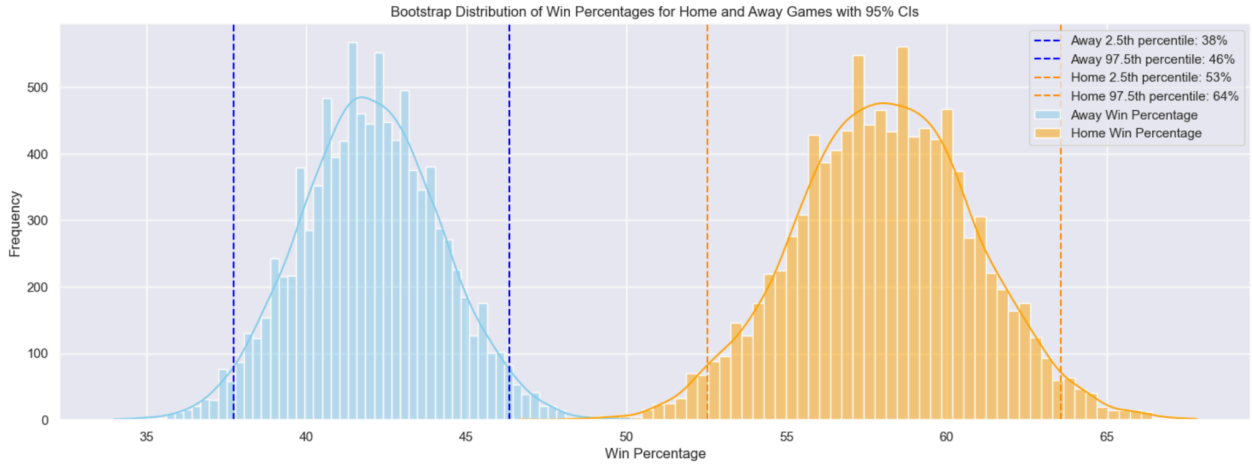

Figure 4: 95% Confidence Interval and Bootstrap Distribution for Home vs Away Games

1. We are 95% confident that the true win percentage of away games of NBA basketball teams in the 2022-2023 season is between 38% and 46%
2. We are 95% confident that the true win percentage of home games of NBA basketball teams in the 2022-2023 season is between 53% and 64%.

Because there is no overlap between the two 95% confidence intervals, we can conclude that there is a statistical difference between the mean win percentage of NBA basketball games played at home versus away for these 30 NBA teams in any season in proximity to the 2022-2023 season.

Conclusion #2: We conducted a hypothesis testing that tested two independent samples (home win rate percentage vs away win rate percentage for NBA basketball games) and whether they have identical average (expected) values. This resulted in a p-value < 0.05. We failed to reject the null hypothesis; there is convincing evidence that there is a discernible difference in win percentage of games played at home versus away for NBA teams in any given season. We also computed two 95% confidence intervals, one for NBA games played at home, and one for games played away. These confidence intervals did not overlap, which also provided evidence that there is a statistically significant difference between the mean win percentage of games played at home versus away for NBA basketball teams in any given season.

Research Question 3: What other team-related performance ratings (offensive ratings, defensive ratings, turnover percentage, etc.) have the greatest impact on the outcome of an NBA basketball game?

We analyzed various different team-related performance ratings and measured the strength of their correlations and their power in predicting a team's win rate percentage. We fit multiple linear regression models, each with one predictor, in search of the highest r^2 value. Here are the results:

```
RMSE: 5.2673932206499305
R-squared: 0.711008171595721
Coefficient: -3.8491166289841754
Intercept: 492.6915971374646
```

Figure 5: Linear Regression Model utilizing **Defense Rating** (an estimate of points allowed per 100 possessions) to predict **Win Rate Percentage**

We can conclude that 71% of the variability in the win rate percentage can be explained by the model.

```
RMSE: 5.695905088802006
R-squared: 0.662075585613555
Coefficient: 3.736501079913609
Intercept: -378.6355159879895
```

Figure 6: Linear Regression Model utilizing **Offense Rating** (an estimate of points produced or scored per 100 possessions) to predict **Win Rate Percentage**

We can conclude that 66% of the variability in the win rate percentage can be explained by the model.

```
RMSE: 3.4996355679399564
R-squared: 0.8724325937619011
Coefficient: 3.0346749440600944
Intercept: 50.51586287514208
```

Figure 7: Linear Regression Model utilizing **Simple Rating System** (a rating that takes into account average point differential and strength of schedule) to predict **Win Rate Percentage**

We can conclude that 87% of the variability in the win rate percentage can be explained by the model.

```
RMSE: 3.428621730824907
R-squared: 0.8775572067085717
Coefficient: 2.950541362302527
Intercept: 50.41960261510763
```

Figure 8: Linear Regression Model utilizing **Net Rating** (an estimate of point differential per 100 possessions) to predict **Win Rate Percentage**

We can conclude that 88% of the variability in the win rate percentage can be explained by the model.

We also analyzed other variables and their predictive power, but these were among the most predictive of win rate percentage. Amongst these 5 variables, we took the 3 with the highest r^2 values, those that were most correlated with the win rate percentage. These 3 variables were: (1) Defense Rating (2) Simple Rating System and (3) Net Rating. We then combined and utilized these 3 variables as the 3 predictors for another linear regression model. This model's results are shown below:

```
RMSE: 3.470349552882498
R-squared: 0.8745587061312551
Adjusted R-squared: 0.8600847106848615
Coefficient: -0.1452640548153662
Intercept: 67.06025139421207
```

Figure 9: Linear Regression Model utilizing **Net Rating, Defense Rating,** and **Simple Rating System** to predict **Win Rate Percentage**

Utilizing adjusted r^2 now that we have more than one predictor, we can conclude that 86% of the variability in the win rate percentage can be explained by the model (adjusted for the number of predictors). We can conclude that using more variables as predictors does not add to the model's predictive power (with respect to these specific variables).

We went for a different approach and utilized a random forest model. Because we know that random forests are more effective at modeling complex datasets with non-linear relationships between the independent and dependent variables, we thought of giving it a test. We utilized Simple Rating System (SRS) and Net Rating (NRtg) as our two predictors in our training data, an 80-20 train-test split, and 5-fold cross validation optimizing hyperparameters including number of trees (n_estimators), maximum depth of the tree (max_depth), minimum number of samples to split a node (min_samples_split), and minimum number of samples to be a leaf node (min_samples_leaf). Here are the results:

```
Best Parameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
RMSE: 2.9214561536847032
R squared: 0.9111018287435437
Feature Importances: [0.54009429 0.45990571]
```

Figure 10: Random Forest with **Simple Rating System** and **Net Rating** to predict **Win Rate Percentage**

91% of the variability in NBA win percentage over the 2022-2023 season can be predicted by our model.

Conclusion #3: An NBA team's net rating (point differential per 100 possessions) and simple rating system (average point differential and strength of schedule) can be utilized to best predict teams' win rate percentages over the course of a season. These ratings have the greatest impact on the outcome of an NBA basketball game in comparison to the rest of the performance team ratings that we analyzed.

Link to our Github Repo that houses all of our code (written in Python):
https://github.com/EddieTGH/cs216_project (finished code can be found in FinalDataAnalysis.ipynyb)

Part 5: Limitations and Future Work (10 points):

We faced some data limitations, particularly our data analysis only lasting one season. Our analysis assumes the data from a single season (2022-2023) to be representative for all seasons; this may not accurately reflect the true dynamics of the sport. Another limitation of the project was the assumption that the underlying relationship between the predictors (3-point percentage, defensive ratings) and the outcome variable (win percentage) is linear. This assumption is inherent in the use of linear regression models and might not accurately represent the true dynamics of basketball performance, where relationships could be more complex and non-linear. Furthermore, the heavy reliance on quantitative team performance metrics may not fully encompass the qualitative and unpredictable elements of basketball, such as team chemistry and player mental health. Future studies could enhance the robustness of findings by incorporating data from multiple seasons to better account for inter-season variability. Basketball reference also tracks a large amount of advanced team-based performance statistics like Player Efficiency Rating (PER) or True Shooting Percentage (TS%) that would be useful when performing an individual player-based analysis. With this, further questions can be raised: in terms of true shooting percentage, is it more advantageous to have a balanced but mediocre team or an unbalanced team with a few players with ridiculously high true shooting percentages? As statisticians come out with more advanced statistics, this study could be repeated with metrics that might reveal more insights into the game than we were able to uncover. Finally, the incorporation of qualitative analyses regarding player psychology could yield interesting findings. For example, a future study could track press favorability of a certain player and determine if that affects performance over time.

Part 6: Conclusion (5 points):

Our study addressed multiple research questions focusing on NBA basketball team performances. Firstly, we examined whether an increase in 3-point shooting accuracy correlates with NBA team win percentages. By employing linear regression analysis, our findings substantiated a moderately strong, positive relationship between a team's 3-point shooting percentage and its win percentage. Specifically, we observed that a 1% increase in 3-point shooting accuracy could predict a nearly 4.90% increase in win percentage, suggesting teams that excel in 3-point accuracy tend to perform better overall. Secondly, our investigation into whether NBA teams have a home-field advantage revealed significant differences in win percentages between home and away games. Utilizing a t-test and bootstrap resampling, we found compelling evidence that teams perform better at home, with home game win percentages ranging confidently between 53% and 64%, compared to 38% to 46% for away games. This statistical significance underscores a consistent home-field advantage across the league. Lastly, we analyzed the impact of various team performance ratings—defense rating, offense rating, net rating, simple rating system, etc.—on win percentages. Our results indicated that the simple rating system (SRS) and net rating (NRtg) were most predictive of win outcomes, with the random forest model including these variables explaining up to 91% of the variability in win rates. These ratings, reflecting a combination of point differentials and strength of schedule, emerged as the strongest predictors among those we tested.