Team Final Report

Section – C, Team 16

Aaryansh Vaish, Arthur Bazil, Jasmine Duong, Ryan Rumao, Tanisha Agarwal, Zheng Yang

NBA Background Information:

The National Basketball Association ('NBA') is a professional basketball league composed of 30 teams across the United States and Canada. As of the 2025 season, the average valuation of each team is about 4.6 billion dollars (CNBC).

Each NBA team plays 82 games per season, and the top 8 teams from both the Eastern and Western Conferences qualify for that year's NBA playoffs. There are 4 rounds of playoffs where teams play a best-of-7 game: First Round, Conference Semi-Finals, Conference Finals, and NBA finals. These games generate millions of dollars of revenue and immeasurable loyalty to the brand.

Business Understanding

One of the most compelling aspects of sports is the uncertainty: it is why the most competitive games always draw the most attention. Some fans are willing to spend thousands of dollars to cheer on their favorite team, because to them, the team is a part of their identity. Sports teams bring unity and a sense of belonging to people from all walks of life, and can connect people to their homes, childhoods, and communities. It is important to understand the emotional connection to the game for millions of people to understand how team performance and successful marketing is powerful in capturing revenue and brand loyalty.

However, while uncertainty excites fans, it can also cost NBA teams millions of dollars in lost revenue. This often happens in the NBA playoffs, where high performing teams in the regular season perform poorly in the playoffs. Every NBA season produces

millions of data points that are aggregated into player and team statistics. If this data is properly analyzed, it could show how performance is linked to playoff success.

We built a predictive model that inputs regular season NBA team and player data and forecasts a ranking of their position in the playoffs. This model attempts to address the challenges of managing the large volume of statistics for NBA management teams, and guide their marketing, team training, and contracts decision making to enhance competitiveness and revenue growth.

How data mining will help the business problem: Data mining provides a systematic approach to handling the large and complex datasets and helps quantify the overall impact of the statistics. In the NBA, this would enable teams to identify key metrics in which they would have to focus on to gain a competitive edge.

Data Understanding:

We found an online source for NBA data for teams and players statistics - https://www.nbastuffer.com/. We decided to use data from the 2020-2021 NBA season to the 2024-2025 season as the format for playoff qualification was changed in the 2020 season with the introduction of the play-ins tournament.

Due to NBA playoffs format changing over the years, we were restricted to data from the 2020-2021 season and onwards. This led us to have around 150 rows for team statistics in total, out of which 30 were going to be used for the test split and 120 for the train split. For this reason, we had reduced amounts of data in comparison to similar projects conducted.

Another issue we came across involved filtering out players that transferred in the middle of the regular season, to keep solely the statistics for the team they are currently on. We were not able to conduct this task. Manual lookups would be needed, which would be time consuming and would leave room for manual error. Furthermore, our data only consists

of in-game statistics and as a result, does not account for injuries, personal issues that a player might be going through which could affect their performance, drastically affecting playoff results.

Data Cleaning and Preparation:

After manually scraping the data from this website for players and teams, we modified columns such as Team and Rank to be consistent with the formatting across different seasons, and also for the player data and team data. As part of the data preparation and with the objective of training our model, we attributed categorized variables to each team from a range of 0 to 6, within our team dataset. The categorized variables represent a past team's performance for each season, in the playoffs, and are used in order to best predict future results. This is the feature variable we will aim to train our models on and use for prediction for the 2024-2025 season. The feature variable's breakdown is as follows:

- 0 = The team did not make the play-ins or playoffs
- 1 = The team made the play-ins but did not make the playoffs
- 2 = The team made it to the first round of playoffs
- 3 = The team made it to the conference semi-finals
- 4 = The team made it to conference finals
- 5 = The team made it to NBA finals
- 6 = The NBA champions

Moreover, as part of our data cleaning process within the players' dataset, we are looking at removing any player that may not have a significant impact on a team's performance. As a result, we decided to exclude any player that has played less than 5 minutes per game and has played less than 10 games per season. Lastly, we aggregated the player data by calculating weighted averages across each season for each team to get

important metrics such as Average Offensive Rating (Avg_Ortg) and Average Turnover % (Avg_TO) that we merge with the team data to potentially use in our modelling approach.

Modeling:

Linear Regression: We use a multiple linear regression model as well as linear regression with interaction variables. After testing a variety of different combinations, we settled on a simple model that uses some of the aggregated player metrics as well as some team statistics for multiple linear regression. Similarly, for linear regression with interaction variables, we decided to use fewer variables that might have a lot of effect on rankings but include interactions amongst them to see how it affects playoff standings.

Alternative: LASSO Regression, Multinomial Logistic Regression

Pros: Simple model that is easy to interpret. Coefficients directly relate to how one feature affects the chances of making playoffs.

Cons: Assumes a linear relationship and independence between variables, difficult to expand to non-linear interactions. Susceptible to multicollinearity. Difficult to get significant variables in terms of p-values as the dependent variable is something we added to the dataset.

The model helps identify predictors that are relevant to success in the playoffs, and also informs us of the incremental effect these predictors tend to have on playoff rankings which could be useful in business scenarios. It provides information regarding what aspects should the team focus heavily on and what might not matter as much. This could guide strategic development and resource allocation for the teams.

LASSO (Least Absolute Shrinkage and Selection Operator): After testing a bunch of different combinations of features, we find that the LASSO model is most effective when provided many features with interactions amongst all of them and it can decide which ones to

use and which ones to not. We make predictions with minimum lambda as well as 1SE lambda.

Alternative: Ridge Regression

Pros: Handles multicollinearity and reduces overfitting compared to linear regression. Performs feature selection to automatically identify the best features needed to make the predictions.

Cons: There is a potential to ignore variables that might actually be significant, and the model is much more difficult to interpret compared to linear regression.

The LASSO model identifies influential predictors of playoff performance and success for our case, while automatically filtering out variables that may not be relevant. This allows the model to solely focus on factors that do drive performance for the team, including efficiency or consistency metrics. LASSO can assist team managers and coaches in conducting strategic decision-making and resource allocation, therefore investing in key areas for optimized performance.

CART: We use a Regression Tree to identify useful features and split points to make predictions. We run cross validation to identify the best size for the tree. Unfortunately, due to the heavy impact certain variables such as eDIFF (Difference in offensive and defensive efficiency) have, the tree just decides to split across the 1 variable.

Alternative: Random Forest, Gradient Boosted Trees (XGBoost)

Pros: Easy to interpret and identify features that strongly impact the dependent variable. Does not need to assume linearity in variables.

Cons: Prone to overfitting and it tends to be unstable, with small changes in data affecting split points and optimal size heavily, which in turn, affects the accuracy of the model. A limitation of the model is it tends to split test data points into buckets with data points in the same bucket having the same prediction. As a result, the predictions the model makes are adjusted by adding random noise to it as a tiebreaker for us to be able to use them effectively.

The Regression Tree is essentially characterized as a simple rule-based decision model for quick and effective interpretation. The model's utilization can be useful in understanding cutoff points with team and player variable ranges that drive teams to either making the playoffs or not making the playoffs. NBA organizations can help turn simple data insights into quick actionable items and strategies. An example could be a team on the verge of qualifying for playoffs.

Random Forest: We run a collection of regression trees using the Random Forest algorithm (500 trees) and with a mtry parameter value of 7. We decided on this mtry value after testing multiple values and settling on the one that got us the least RMSE.

Alternative: CART, Gradient Boosted Trees (XGBoost)

Pros: Usually has quite high predictive accuracy compared to most models, able to deal with non-linear interactions and is quite good at handling outliers and overfitting.

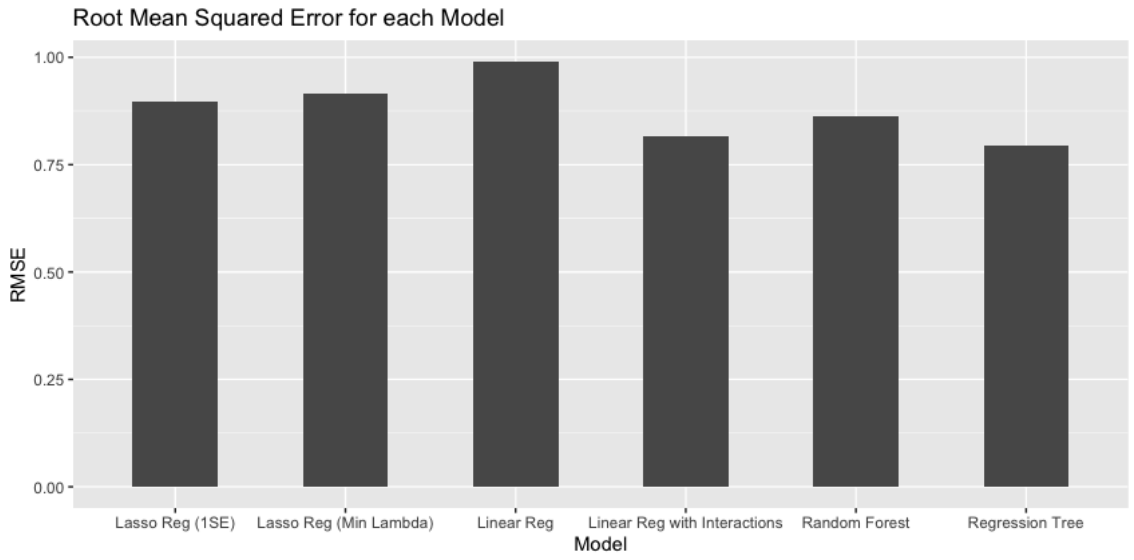
Cons: Loss of interpretability of feature variables and split points and also interpretability of a single tree. Requires testing and tuning of parameters across different combinations.

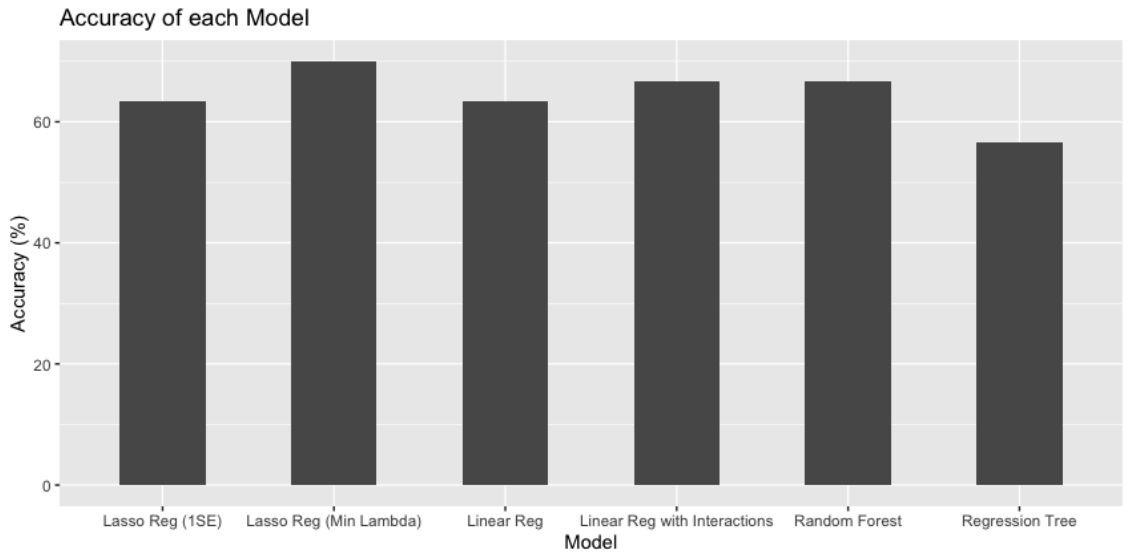
The Random Forest captures complex interactions between variables such as player metrics and team depth through stats, to provide a more balanced view of playoff success predictors. By ranking them by importance, the model is accurate and stable to desired results. As a result, the organizations are able to receive data-backed insights that can provide them guidance in managing trades, player development, and coaching strategies.

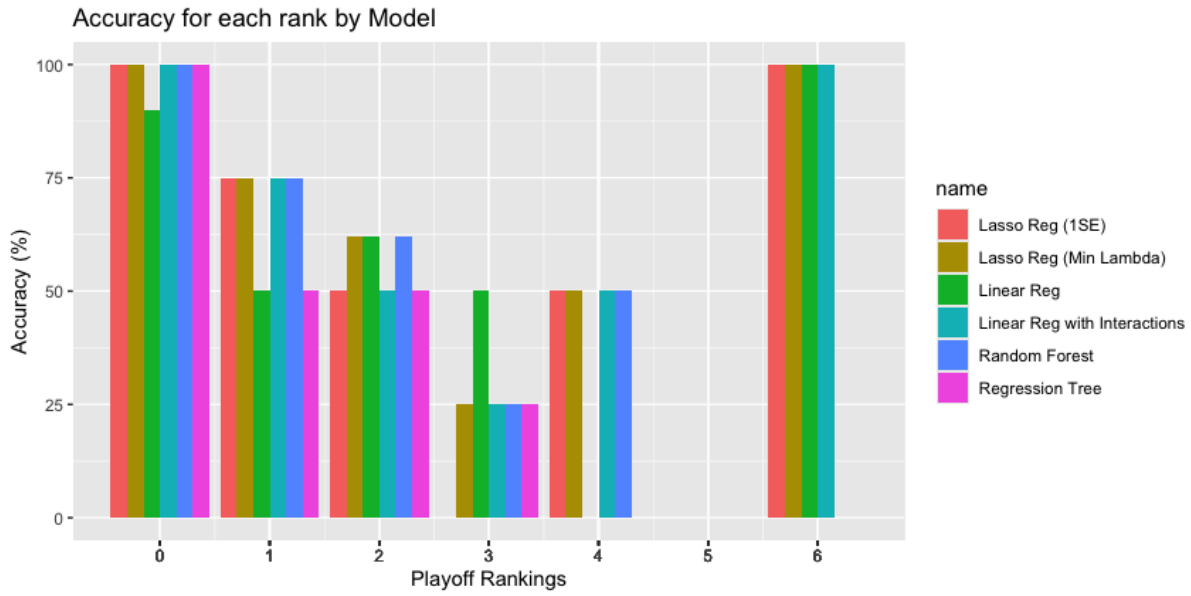
Evaluation + Predictions:

We evaluate all models using the out of sample Root Mean Squared Error and Accuracy. We calculate accuracy for models in 2 different ways. Firstly, we simply compute for all predictions, how many the model was able to get right and then we compute accuracy within each playoff ranking to see what the model struggles with predicting. RMSE tends to range from 0.79 to 0.99, indicating that predictions are usually within 1 round of the actual playoff rankings. Accuracy values tend to hover in the 60–66.7% range, with a low of 56.67% and a high of 70% for Regression Tree and Lasso with minimum Lambda respectively. Accuracy across ranks differ quite a bit, with models being able to predict teams who did not make playoffs as well as the playoff championship quite well. Play-in teams and teams that went out in 1st round (Ranking = 1 or 2) tend to still have decent accuracy across most models, but accuracy tends to get worse as playoff rankings increase until the championship (Ranking = 6) with none of the models correctly predicting the runner-ups (Ranking = 5). Overall, Linear Regression with interaction terms, Random Forest and Lasso Regression with the minimum Lambda value perform the best, balancing a low RMSE score and a decent accuracy score.

	Model	RMSE	Accuracy
1	Linear Reg	0.9902938	0.6333333
2	Linear Reg with Interactions	0.8161584	0.6666667
3	Lasso Reg (Min Lambda)	0.9166020	0.7000000
4	Lasso Reg (1SE)	0.8955779	0.6333333
5	Regression Tree	0.7941406	0.5666667
6	Random Forest	0.8624865	0.6666667


Root Mean Squared Error for each Model
RMSE
1.00
0.75
0.50
0.25
0.00
Lasso Reg (1SE)
Lasso Reg (Min Lambda)
Linear Reg
Linear Reg with Interactions
Random Forest
Regression Tree
Model


Accuracy of each Model
Accuracy (%)
60
40
20
0
Lasso Reg (1SE)
Lasso Reg (Min Lambda)
Linear Reg
Linear Reg with Interactions
Random Forest
Regression Tree
Model


Accuracy for each rank by Model
Accuracy (%)
100
75
50
25
0
0
1
2
3
4
5
6
Playoff Rankings
name
Lasso Reg (1SE)
Lasso Reg (Min Lambda)
Linear Reg
Linear Reg with Interactions
Random Forest
Regression Tree

From a business case perspective, the models could be used to simulate and find areas of improvement. For instance, we could consider how increased consistency in the regular season would affect our playoff performance, or how sacrificing a bit of our defensive rating to focus more on the offensive rating leads us to performing better in the playoffs.

In terms of ROI, it is not viable to use such a model to find a monetary value for the return a team might get based on increased playoff performance as that would vary heavily based on the popularity of the team, fanbase, sponsorships and also the star power of their players. Furthermore, external factors such as injuries, last-minute trades and bad matchups could affect playoff results heavily. However, the model could be used to provide non-monetary ROI in terms of increased decision making quality, strategic changes and competitive advantage by focusing on specific areas common to teams that tend to perform well in playoffs.

Deployment:

The model could be continuously refreshed throughout the season, with the final model updated after all of the data from the regular season is accounted for. This model is intended to be used as a decision support tool to better budget marketing resources, more efficiently train, and to start planning current contracts earlier in preparation for the NBA Draft.

Marketing: Ranking predictions can help marketing teams allocate promotional budgets more effectively. Teams projected to perform well can increase investment in playoff-related campaigns, ticket pricing strategies, and merchandise pushes, while lower-ranked teams can focus on fan engagement and community-building efforts. The Marketing focus should be around the possibility of winning and qualifying for the play-offs, fueling excitement to grow a larger fan base.

Training: Coaches can use performance indicators from the model to identify weak areas to train more on, and adjust accordingly. Teams should use the model in support of their current training regime.

Contracts: Teams can use these insights to plan negotiations and budget accordingly. This will help with forecasting salary demands and help with the draft budget.

Potential Issues: Teams must keep in mind that actions taken with respect to the model, especially with gameplay and training, may have opposing effects. For example, if a team predicted to place high decides to rest their key players more to prepare for later, more important games, they might be offset by a lower ranked team that puts in fighting efforts earlier on. Additionally,

Ethical considerations: One major ethical consideration pertains to any reliance on the model for direct monetary gains, specifically, sports betting. While the model is designed to assist with decision making regarding the team, it should never be taken to an extent where actions require absolute certainty. Overconfidence in the model can encourage unhealthy betting behaviors, especially among fans and stakeholders who might misinterpret forecasts as guarantees.

Risks**:** A potential risk for the deployment of our models in the long-term would be related to potential rule changes within the playoffs/play-ins. Our model would have to be adapted in order to run accurate predictions. Our goal to mitigate these risks would involve developing flexible models that can be quickly updated and trained to best meet our expectations. Moreover, complex model output interpretability may be challenging for managers and coaches with a lack of analytical experience. In order to mitigate these issues, we would share complementary tools including visualizations to communicate key drivers and trends for performance optimization.

APPENDIX:

Contributions

Ryan Rumao: Data Scraping, Data Cleaning, Modeling, Evaluation, Visualization

Arthur Bazil: Data Scraping, Data Formatting, Modeling, Data Cleaning and Preparation, Powerpoint Presentation

Jasmine Duong: Problem Definition, Project Sections: Background, Business Understanding, Deployment, PowerPoint Presentation

Tanisha Agarwal: Background Research, Business Understanding

Aaryansh Vaish: Modeling, Evaluation

Zheng Yang: Data Understanding, Data Preparation

Information Sources:

https://www.cnbc.com/2025/02/14/cnbcs-official-nba-team-valuations-2025.html\

https://news.mit.edu/2025/basketball-analytics-investment-nba-wins-and-other-successes-0325

Code Sources:

https://www.rdocumentation.org/packages/HDCI/versions/1.0-2/topics/Lasso

https://www.rdocumentation.org/packages/tree/versions/1.0-45/topics/tree

https://stackoverflow.com/questions/65396035/adding-noise-to-a-column-in-dplyr

https://stackoverflow.com/questions/72221596/how-to-calculate-weighted-average-with-r

https://ggplot2.tidyverse.org/reference/geom_bar.html

Data Source:

https://www.nbastuffer.com/nba-stats/player/

https://www.nbastuffer.com/2024-2025-nba-team-stats/

Data Glossary:

GP: Games played

PPG: Points Scored Per Game

oPPGPoints Allowed Per Game

pDIFF: Points Differential = [(Total Points Scored) – (Total Points Allowed)]/ (Games Played)

PACE: Pace, an estimate of Possessions Per 48 Minutes

oEFF: Offensive Efficiency, points scored per 100 possessions

dEFF: Defensive Efficiency, points allowed per 100 possessions

eDIFF: Efficiency Differential = [(Total Offensive Efficiency) – (Total Defensive Efficiency)]/ (Games Played)

SoS: Strength of schedule for the games played. Average of opponent efficiency differential is used as an indicator of the strength of the schedule. The higher the SoS rating, the tougher the schedule where zero is average.

rSoS: Strength of schedule for the remaining games. The higher the rSoS rating, the tougher the schedule where zero is average.

SAR: Schedule Adjusted Rating. A team evaluation metric based on team's efficiency differential and strength of schedule.

CONS: Consistency Rating. Consistency is calculated with game-by-game efficiency differential variations. The lower the rating, the more reliable and stable the team's performance.

a4F: Adjusted Four Factors, calculated by applying weights to the differentials of offensive and defensive four factors. A4F explains the specified proportion of variability in wins

W: Wins: The most important goal in sports unless your team is not tanking

L: Losses: Total number of games lost

W%: Winning percentage

eWIN%: Correlated Gaussian Expected Winning Percentage, indicates the ideal winning percentage based on offensive and defensive performance

pWIN%: Projected Winning Percentage, each point differential translates to 2.7 wins over the course of the season

ACH: Achievement Level In Terms of Wins, this metric is based on the differential between actual and expected winning percentages. Positive figures indicate overachievement while negative figures indicate the team should have won more games.

STRK: Current Streak, winning or losing streak for the season

CUR: Denoted by "*", CUR indicates the team that the player is currenty playing for.

USG%: Usage rate, a.k.a., usage percentage is an estimate of the percentage of team plays used by a player while he was on the floor

TO%: A metric that estimates the number of turnovers a player commits per 100 possessions

eFG%: With eFG%, three-point shots made are worth 50% more than two-point shots made. eFG% Formula=(FGM+ (0.5 x 3PM))/FGA

TS%: True shooting percentage is a measure of shooting efficiency that takes into account field goals, 3-point field goals, and free throws

PpG: Points per game

RpG: Rebounds per game

ApG: Assists per game

SpG: Steals per game

BpG: Blocks per game

TOpG: Turnovers per game

P+R: Player Props = (Points + Rebounds) total per game

P+A: Player Props = (Points + Assists) total per game

P+R+A: Player Props = (Points + Rebounds + Assists) total per game

VI: The versatility index is a metric that measures a player's ability to produce in points, assists, and rebounds

ORTG: Individual offensive rating estimates the number of points produced (including assists) by a player per 100 offensive possessions

DRTG: Individual defensive rating estimates the number of points allowed (including blocks, steals) by a player per 100 defensive possessions