Primary Questions

- What team variables are important to a team's long-term success (the next season)
- What team variables are most important to a team's short-term success (a single game)
- How these variables differ, overlap, and why

Long-term Success Team Variables

- **Current Season Win Percentage**
  - The percentage of games won this season
  - Win % = (Wins / Total Games Played) * 100
  - 50% is average; 60%+ is playoff-caliber; 70%+ is championship contender status
- **Offensive Rating**
  - Measures how many points a team scores per 100 possessions
  - Allows us to compare teams that play at different speeds
  - League average is about 115
- **Defensive Rating**
  - Measures how many points a team allows per 100 possessions
- **Net Rating**
  - The difference between offensive and defensive ratings
- **Assist %**
  - The % of team field goals that were assisted
- **Assist-Turnover Ratio**
  - Dividing a team's total assists by their total turnovers
- **Offensive Rebound %**
  - The percentage of time a team secures an offensive rebound
- **Defensive Rebound %**
  - The percentage of rebounds collected on the defensive end
- **Effective Field Goal %**
  - Measures a player's or team's shooting efficiency by adjusting for the extra value of three-point shots
  - EFG % = (2pt_FGM + 1.5 * 3pt_FGM) / FGA
- **True Shooting %**
  - Measures scoring efficiency by incorporating field goals, three-point shots, and free throws
  - TS % = Total Points / 2 * (FGA + 0.44 * FTA)
  - Players with 60+ TS % are typically considered efficient
  - Accounts for free throws unlike EFG %
- **Pace**

    - How many possessions a team has in a game
- **Opponent Field Goal %**
    - The average opponent FG percentage a team allows
- **Opponent 3pt %**
    - The average opponent 3PT percentage a team allows
- **Opponent Turnovers**
    - The average amount of turnovers a team forces onto the opponent
- **Opponent Blocks**
    - The average amount of blocks a team allows opponents to have
- **Plus Minus**
    - In our data, this is expressed as the opponent's point differential, meaning negative values are better

Short-term Success Team Variables

**Unlike the Long-term Variables, these are calculated as rolling statistics meaning only up to the game we are trying to predict*

- **Point Differential**
    - The difference in points scored
- **Field Goal Percentage Differential**
    - The percentage of all shots made compared to total shot attempts
    - FG % = (FGM/FGA) * 100 where FGM is Field Goals made and FGA is attempts
    - Teams typically average around 47% with 50%+ considered "good"
- **Three Point Percentage Goal Differential**
    - The percentage of three pointers made compared to three point attempts
    - 3PT % = 3PM/3PA * 100 where 3PM is three pointers made and 3PA is attempts
    - Teams typically average around 37% with 40%+ considered "good"
- **Free Throw Differential**
    - The percentage of free throws made compared to free throw attempts
    - FT % = (FTM/FTA) * 100 where FTM is Free Throws Made and FTA is attempts
    - Teams typically average around 78% with 80%+ considered "good"
- **Defensive Rebound Differential**
    - Teams typically average around 33-35 defensive rebounds per game
- **Steal Differential**
    - The number of times a defensive player legally takes the ball from the offense
    - Teams typically average around 7-8 steals per game
- **Block Differential**
    - The number of shots legally deflected by a defensive player
    - Teams typically average around 4-5 blocks per game
- **Turnover Differential**

    - The number of times a team loses possession of the ball to the defense
- **Cumulative Winrate Differential**
    - The percentage of games won over the current season
    - Win % = (Wins / Total Games Played) * 100
    - 50% is average; 60%+ is playoff-caliber; 70%+ is championship contender status
- **Last Season's Winrate Differential**
    - Winning percentage from the previous season
- **Last 10 Games Winrate Differential (Momentum)**
    - Winning percentage over the most recent 10-game stretch
    - Indicates current momentum or "hot/cold" streaks entering the matchup
- **Rest Days Differential**
    - The number of days off a team has had since their last game
    - Rest Days = Date of Current Game - Date of Last Game - 1
    - Teams playing on 0 rest (back-to-back) typically perform worse due to fatigue
- **Game Number Differential**
    - The total number of games played in the season so far
    - Accounts for schedule imbalances where one team has played more than the other at that point of the season
- **Average and Standard Deviations**
    - For each variable listed above, the differential is between the average of each team
        - Ex: Home's (Average FG %) - Away's (Average FG %) = Average FG % Differential
    - However, there are also differentials between each team's standard deviation Ex: Home's (FG % Standard Deviation) - Away's (FG % Standard Deviation) = FG % Standard Deviation Differential
        - This allows us to compare how consistent each team is
        - Between two teams with similar averages, the expectation is that the more consistent team should win

The Methodology

1. Good teams score well on advanced metrics
2. For most teams, changes between seasons are minor so good teams stay good
3. Meanwhile, short-term success is volatile and susceptible to off-court factors
    a. Variables such as momentum and rest days become more important
4. We suspect the weight of these off-court factors will lessen the coefficients of the long-term success variables identified

The baseline for our short-term models is 58%, not 50%. Even between two identical teams, the home team tends to perform better. Therefore, selecting home as the winner in every match-up historically results in approximately 58% accuracy.

The Approach

Long-term

1. Compile advanced team offensive and defensive stats along with next season's winning percentage
2. Feed this data to our models to identify feature importance

Short-term

1. Compile data for each regular season game for seasons 2015 to 2022
2. For each game, compute the difference between the Home and Away team averages and standard deviations for each of our variables
3. Feed the differentials to our models and optimize the parameters for accuracy

The Logic

Long-term

The first part of our project follows our last midterm project where we questioned which features are most important to a team's success in the following season. This exploratory data analysis leverages the NBA API for its stats. After creating a master DataFrame for seasons 2015 to 2022, we run Linear Regression, Random Forest, and Gradient Boost Regressor models to see if there exists a consensus feature importance.

Short-term

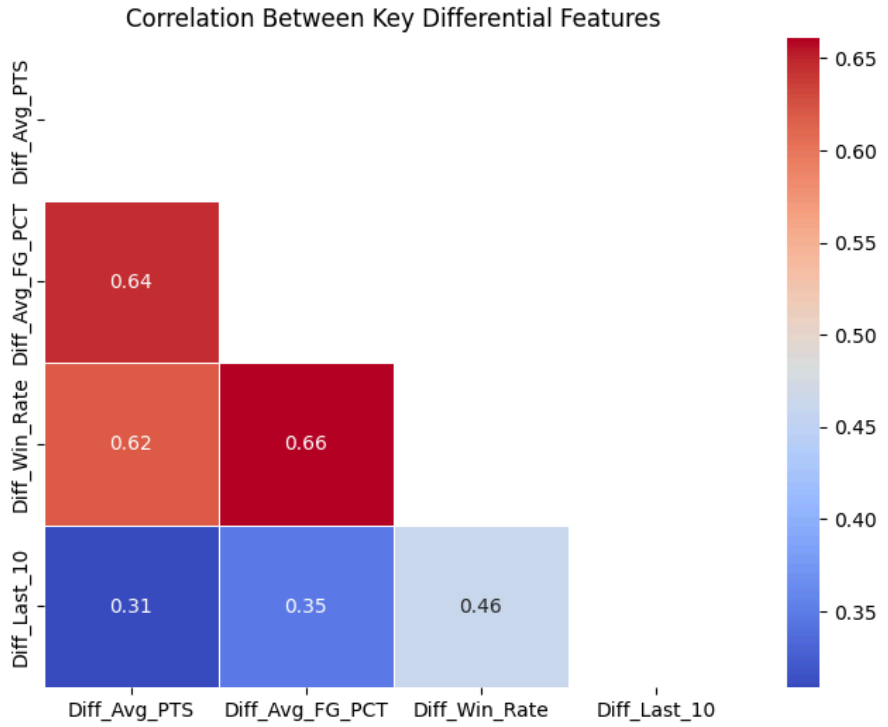
Our models assign the weightings to each variable that best fit our dataset's actual results. No manual adjustments are necessary. However, the majority of our variables focus on the ongoing

season. So if a team is playing its first game of the season, all stats other than “Last Season’s Winrate” are missing. To mitigate this, our models bake in the stats for: Average Points, Average Field Goal Percentage, Average Three Point Percentage, and Last 10 Games Winrate. Rather than setting these equal for each team, the home team is given slightly better numbers to ensure home court advantage is incorporated.

Our models utilize data from the 2015 to 2022 seasons. To prevent leakage, the training is done on only the first 80% of game data. The remaining 20% is the test set. 2015 is chosen as a starting point rather than earlier seasons because this is where the “three point” era cemented itself. The shot was increasing in popularity up to this point, but Steph Curry’s first MVP season and the Golden State Warrior’s championship undoubtedly put the league on notice. Suddenly, even Centers (normally relegated to the paint) were practicing and shooting three pointers. Successful teams leveraged the threat of threes to dramatically increase their spacing on the court and employed new offensive schemes. For example, 2018 Rockets and their infamous Game 7 against the Warriors which today is remembered for the team’s overreliance on threes. Including games from earlier seasons would only detract from our model’s predictive abilities due to the structural differences between how teams approached the game before the three point era.

Multicollinearity

When evaluating coefficients, multicollinearity is a slight issue. For instance, our variable tracking the difference between each team’s cumulative win rate has a correlation of 0.66 with the difference between each team’s field goal percentage. Intuitively, this is expected as the team that shoots more efficiently likely has the higher winrate. However, 0.66 does not breach the 0.7 threshold that many consider for multicollinearity so our coefficients are still usable for interpretation. Multicollinearity does not conflict with our models' predictive abilities.



The Models Employed

Logistic Regression (LR)

Works similarly to linear regression, except the probabilities are bound between 0% and 100%. The model optimizes the coefficients for each parameter according to the dataset we provide. It settles on the combination of coefficients that makes its predictions most accurate.

Random Forest

Utilizes many different decision trees to arrive at a prediction. Each tree is assigned to a random subset of stats. With this approach, hidden patterns and relationships between variables are more salient.

K-Nearest Neighbors (KNN)

Plots each game on a graph and separates them into similar groups. Its prediction is made based on these similar games and their outcomes.

Gradient Boosting Regression

Builds each decision tree sequentially rather than independently. While Random Forest takes the aggregate decision of its trees, this seeks to improve on each iteration and correct errors. This captures non-linear patterns other miss.

XGBoost

Essentially a highly optimized Gradient Boosting Regression. Minimizes the complexity and noise to deliver superior results.

The Results

Long-term

Similarities:

- Models show that the three most important features were the teams' Net Rating, their Win Percentage and having a positive Plus Minus.
- Features such as the TS_Pct, Dreb_PCT and the OPP_TOV were ranked the same in the Random Forest and Gradient Boosting Models.

Differences:

- Linear Regression Model differs most compared to the Random Forest and Gradient Boosting Model.
- The Linear Regression Model has some interesting outliers such as the EFG_PCT and the TS_PCT as its most influential coefficients.
- The Gradient Boosting model seems to weight the Net_rating as the most important feature while the Random Forest has the W_PCT and the NET_Rating almost equally as important.

Determining Next Season's Win Percentage

We were also intrigued to find out how different models would perform in predicting next season's win rate. Specifically, the 2022 season win rate. R-squared and Mean Squared Error are used as our goodness of fit measures.

The best model was undoubtedly the XGBoost Model. This model was able to explain 80% of the variance and had a Mean Squared Error of only 6%. In comparison, the other models (Random Forest Regressor and KNN), had an R-squared of 68 and 66 percent respectively. This similar efficiency is likely due to how they group variables. However, the Random Forest Regressor had a MSE of 14%, nearly double the KNN's MSE. This discrepancy is likely due to our relatively small dataset. Less points lead to a higher variance when training each tree.

Short-term

--- MODEL COMPARISON RESULTS ---

Model	Train Accuracy	Test Accuracy
Logistic Regression	**63.2**	**62.6**
KNN	**100.0**	**58.8**
XGBoost	**65.4**	**62.8**

--- TOP 5 FEATURE COMPARISON ---

LR_Feature	LR_Coeff	XGB_Feature	XGB_Importance
Diff_Last_10	**0.24**	**Diff_Last_10**	**0.19**
Diff_Prior_Win_Rate	**0.12**	**Diff_Win_Rate**	**0.07**
Diff_Win_Rate	**0.11**	**Diff_Prior_Win_Rate**	**0.05**
Diff_Avg_FG_PCT	**0.06**	**Diff_Avg_FG_PCT**	**0.05**
Diff_Std_DREB	**0.05**	**Game_Number_Home**	**0.04**

K-Nearest Neighbors

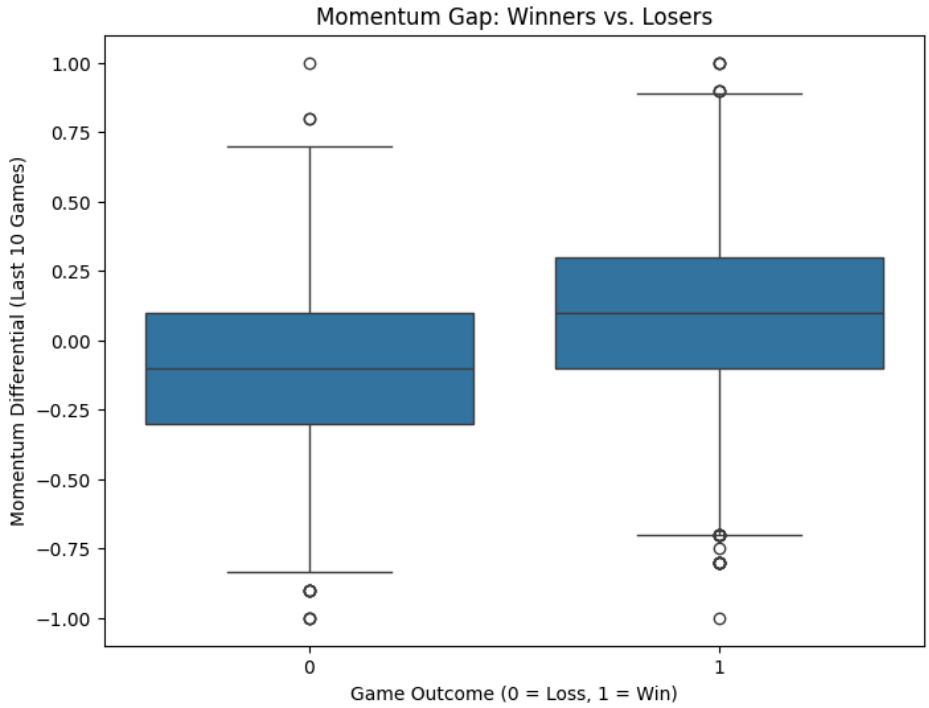
Our KNN model performed the worst, achieving only a 58.8% accuracy rate, less than 1% better than the baseline. The KNN approach is better suited for predictive exercises with less variables than ours. Here, when the KNN model attempts to find similarities between games, it struggles immensely. The most similar games often vary significantly because of the number of variables.

Additionally, KNN assigns an equal weight to all variables. Not even casual basketball fans view all stats equally. Defensive rebounds are incredibly valuable, but they are also expected. A team is never acclaimed for its "phenomenal defensive rebounding ability." Meanwhile, a team notorious for forcing opponents to turnover the ball at a high rate will be applauded for their defensive effectiveness.

Logistic Regression and XGBoost

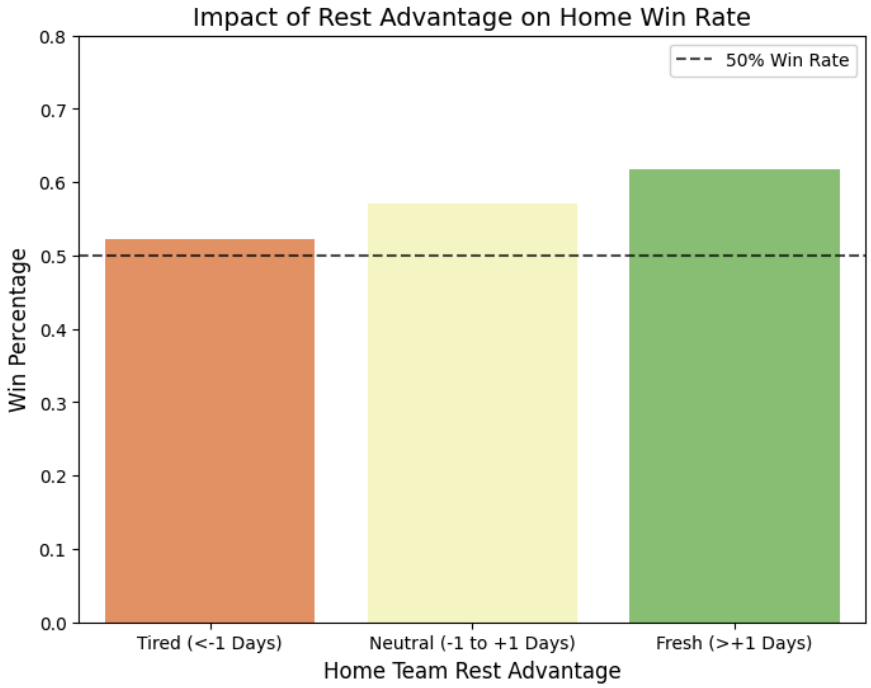
Both models performed effectively the same and better than KNN. The similar results are attributable to our data. Specifically, variables such as field goal differential have a linear relationship to winning. Otherwise, we lacked the sufficient data to improve further with XGBoost. This model thrives on complex interactions and nuance. More detailed data such as match-up specific stats (how well each team performed in their most recent contest against one another) could propel XGB further. Without this, its ceiling is comparable to LR.

The Role of Momentum



In both our Logistic Regression and XGB Boost models, the “Last 10 Games Winrate Differential” variable was identified as the most important by a large margin. This result aligns with our expectations that a team being “hot” does show up statistically.

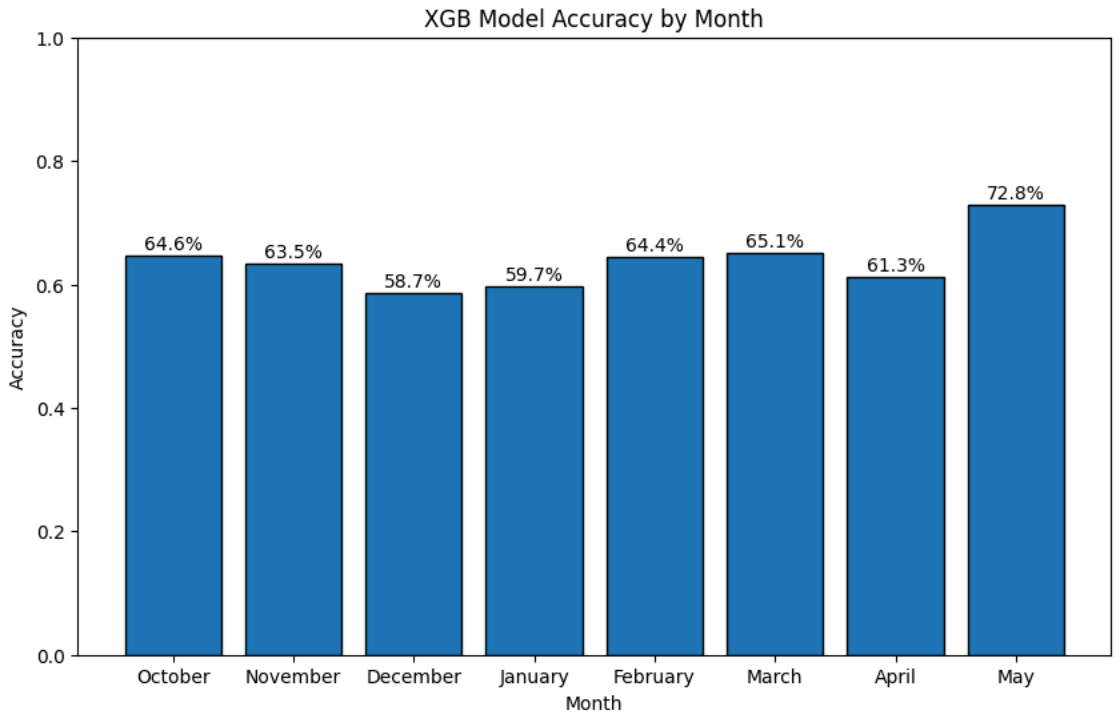
The Importance of Rest Days



While this was not identified within the top five most important coefficients for either model, data does indicate its relevance. Fresh teams have close to a 10% higher win rate than tired teams. Simply put, playing basketball at a top level even two days in a row is physically grueling. Few teams can perform as well as they would wish on a back-back. Thus, when comparing two similar teams, the rest day differential is worth considering.

Limitations

Roster Changes



The short-term model's accuracy decreases in December and January. This decline is likely associated with increased injuries as the season progresses. Our model does not account for team injuries in its head-head matchup predictor. However, if the injury is significant and long-term, it should be reflected within the "diff_last_10" variable. The accuracy increases significantly beginning in February which coincides with the NBA trade deadline. At this point, team rosters are well-established, the effects of long-term injuries are noticeable within the data, and teams have decided whether they are "tanking" (purposely worsening their record) or competing seriously.

Culture and Coaching

Even though these variables are non-quantifiable, most basketball fans agree they are extremely important. Some franchises have a "winning" culture such as the Celtics, while others like the Wizards always struggle. We hope to capture some of this through our "Last Season's Winrate" differential. Unless going through a rebuild, a team with a good coaching staff and culture always performs adequately.

The long-term models likely suffer more from roster changes and less than culture and coaching changes. Since they examine next season's success from a wider timeframe, the culture and coaching are inherently factored more into predictions. However, between season trades are frequent and numerous. The long-term models do not have an avenue to attain this information or consider its effects.

Data Availability, Complexity, and Computing Power

Vegas models are more accurate than anything we may ever achieve because of the mesmerizing number of factors they consider. A model like XGBoost, can feast off this information and find hidden interactions one would not consider. Then, even if one manages to compile this data, the computational needs grow exponentially. Small things like the weather and a player's social media posts all play a role, albeit tiny. These details give superior models an edge that we cannot replicate due to our limited capacity. Even so, professional models are capped at 65%-70%. The remaining error is unfilterable noise and truly random chance that no model can ever compensate for.

Conclusion

Both the long-term and short-term success models place an emphasis on holistic stats rather than granular ones. A team may have a poor shooting night, but compensate defensively. This nuance is reflected on the team's win rate or point differential since they measure outcomes rather than components. Additionally, these broader figures are more reflective of unquantifiable measures such as culture and coaching staff. Overall, we can conclude that the most predictive variables are the most encompassing ones.