Statistics 101C: Final Project

Classification of Winners in NBA Games:

[Our method is Lasso Regression]

Bowen Zheng, Jacob Kelman, Jason Scruggs, Jinhyo Lee, Tamara Pina

Departments of Mathematics and Statistics, UCLA

Contents:

1. Abstract

This study aims to develop a predictive model for forecasting the outcomes of NBA games using historical game data, in this case the 2023-24 Season. Using the first 80% of matchups several models were trained and those models were tested on the remaining 20% of games. Of the models tested, Lasso Regression demonstrated the most effectiveness with a game prediction accuracy of nearly 70%, outperforming Random Forest, Gradient Boosting, Logistic, and Ridge Regression. With the given NBA statistics and the statistics that were engineered it was found that the most influential predictors were difference in win percentage between the two opposing teams, each teams variability of points scored per game, and the difference between the amount of points each team typically concedes in a game. The results suggest that teams with better historical records, stable scoring patterns, and strong defenses are more likely to win.

2. Introduction

Our Problem

In competitive sports, both fans and teams care about the performance of their respective teams. Winning or losing a game could mean the gain or loss of quite a sum of money. The goal of this study is to build a model to predict whether a team will win a game based on historical data from not only the individual team but the entirety of the NBA.

Variables of the study

Given these variables, the study seeks to explore their interrelationships and assess their potential impact on game outcomes. The relationships identified during this exploratory analysis will inform the subsequent analysis and development of more sophisticated features.

Variables	**Categorical vs Numerical**	**Description**
Team	Categorical	Team name
Match up	Categorical	Description of the Match ex) GSW vs PHX, GSW is the home team and PHX is the guest team. GSW @ PHX , GSW is the guest team and PHX is the home team.
Game Date	Categorical	Date of the game ex) MM/DD/YYYY
W/L	Categorical	Win or Loss
MIN	Numerical	Minutes played
PTS	Numerical	Points Scored
FGM	Numerical	Field Goals Made
FGA	Numerical	Field Goals Attempted
FG%	Numerical	Field Goal Percentage
3PM	Numerical	3-Point Shot Made
3PA	Numerical	3-Point Shot Attempted
3P%	Numerical	3-Point Shot Percentage
FTM	Numerical	Free Throw Made
FTA	Numerical	Free Throw Attempted
FT%	Numerical	Free Throw Percentage
OREB	Numerical	Offensive Rebounds
DREB	Numerical	Defensive Rebounds
REB	Numerical	Total Rebounds
AST	Numerical	Assists

STL	Numerical	Steals
BLK	Numerical	Blocks
TOV	Numerical	Turnovers
PF	Numerical	Personal Fouls
+/-	Numerical	Plus/Minus statistic

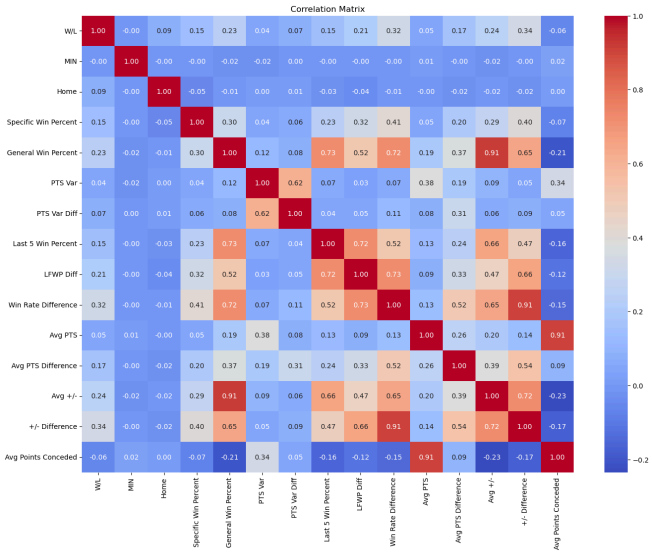
The raw dataset was found to be generally well-structured, with nearly all columns containing complete data and no missing values. A single instance was identified where a team did not attempt any free throws in a game; in this case, an imputed value was assigned to ensure consistency with the analysis requirements. Given the overall completeness and quality of the data, minimal preprocessing and data wrangling were required.

3. Exploratory Analysis

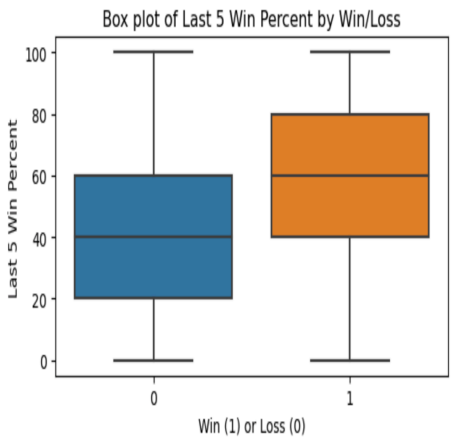
The initial step of exploratory analysis involved examining the already existing variables to identify those that are significant and could potentially contribute to feature engineering. A preliminary look at a correlation heatmap confirms the expectation that points (PTS) and field goals made (FGM) would be strong predictors. Interestingly, field goal percentage (FG%), defensive rebounds (DREB), and assists (AST) had also demonstrated notable correlation, suggesting their potential relevance.

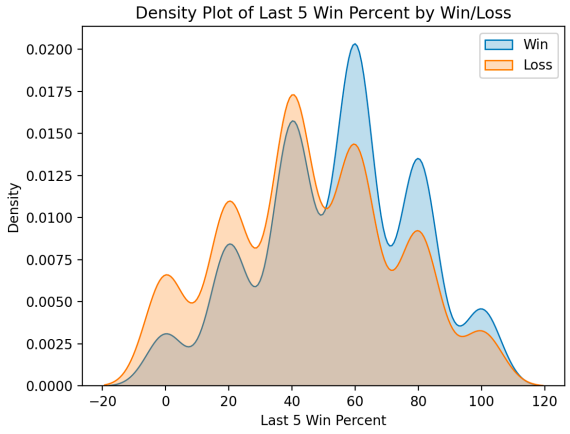
Following our initial analysis, new variables were engineered to capture additional patterns within the data. A correlation matrix was used to visualize these engineered variables and their relationships. Strong correlations once again indicate that these variables may serve as

contributing predictors in the predictive models. This exploratory analysis provides a framework for our final variable selection with which the models will be tested.



With evaluation of the correlation matrix one of the variables which piqued interest was the 'Last 5 Win Percent' feature as this feature resulted in a correlation percentage of over 70% with both the 'Last Five Win Percentage Difference' and most interestingly the 'General Win Percentage'.



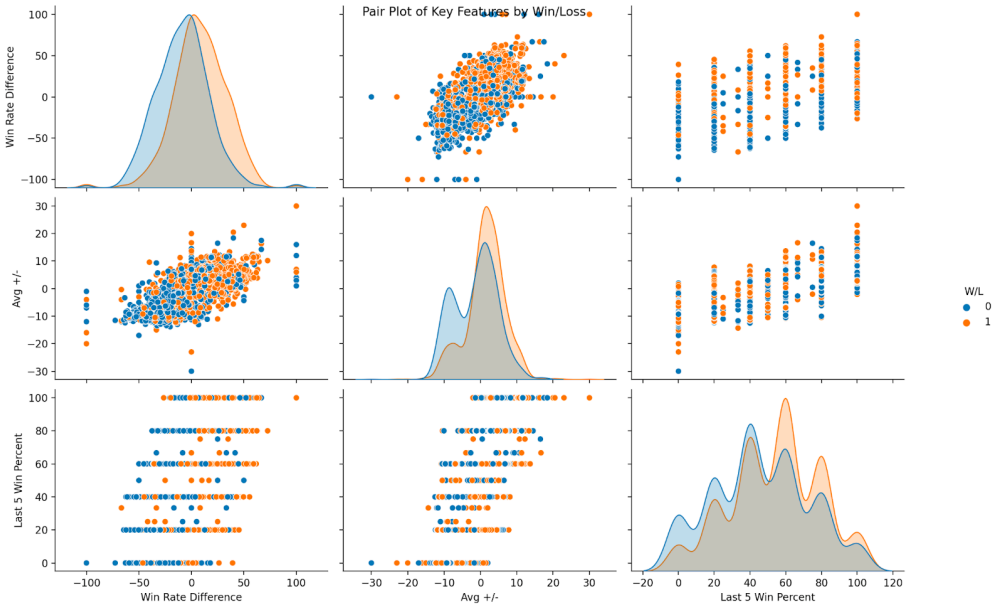


This trend suggests that a team's latest five games are a strong indicator of the probability of winning their next match. Observing this, we decided to allocate more weight to the results of a team's most recent games compared to their earlier games. The weight is defined using an exponentially decreasing sequence, where the weight assigned to each game is calculated as:

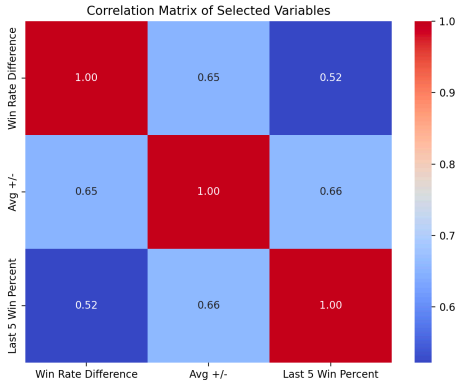
$$w_j = \frac{\alpha^{n-j}}{\sum_{k=1}^{n} \alpha^{n-k}}$$

Here, w_j represents the weight for game j, α = 0.2 is the decay factor, n is the total number of prior games, and j is the position of the game in the sequence, with j=1 corresponding to the earliest game and j=n the most recent. This approach ensures that recent games have a proportionally greater impact on the weighted win percentage while older games contribute less influence.

Based on an exploration of all the features of interest, a deeper exploration on the correlation between 3 specific features was conducted. The correlation between 'Win Rate Difference', 'Avg +/-', 'Last 5 Win Percent' was found to be higher than random chance would suggest.



The figure above shows a visual representation of the correlation between the three features and with further analysis we can deduce a percentage of correlation between said three features.



The correlation between ‘Win Rate Difference’ and ‘Avg +/-’ suggests a team with a better win rate is typically more dominant in their games. This indicates higher offensive skill is more indicative of a team’s likelihood of winning. For this reason, we gave more importance to the points scored rather than any of the rebounds, blocks, or overall defensive metrics.

4. Analysis

To develop a predictive model for NBA game outcomes, a new set of features were engineered and used to create a comprehensive dataset for training and testing. The dataset was split into two parts: the first 80% of the season's games were used for training the models, while the remaining 20% were reserved for testing. This separation ensured that the models were evaluated on data they had not seen during training, preserving the validity of the results.

Several machine learning models were tested, including Logistic Regression, Ridge Regression, Random Forest, Gradient Boosting, and LASSO Regression. Among these, LASSO Regression (Least Absolute Shrinkage and Selection Operator) demonstrated the highest predictive accuracy, establishing it as the best-performing model in our analysis.

LASSO applies an L1 penalty to the coefficients of a linear model, which eliminates less relevant features by shrinking their coefficients to zero. This approach enables the model to focus on the most informative predictors while reducing complexity, thus preventing overfitting and enhancing generalizability to unseen data.

The final features used to train our model included the difference in general win percentages between the two teams, the average points scored per game, the average points allowed per game, the difference in average points conceded by the two teams, and the variance in points scored by the teams in prior games. These features were selected to capture critical aspects of team performance and competitiveness. The dataset was constructed by aggregating historical data for each team, emphasizing recent performances through exponentially decreasing weights, and calculating performance differences between opposing teams. This structure ensured that the model prioritized recent trends, which are often more indicative of future outcomes.

5. Results

Summary

In this analysis, the performance of several different machine learning models to predict the outcome of each NBA game were evaluated. Among the models tested, LASSO Regression was the most accurate, achieving an accuracy of 69.34%, this is due to Lasso's ability to penalize the less informative predictors and selecting only the most relevant features for the model. Logistic Regression and Ridge Regression had slightly lower accuracies of 67.5%, followed by Random Forest with 65% and Gradient Boosting exhibited the lowest predictive accuracy at 60.3%.

An analysis of feature importance revealed that the difference in general win percentage between the two teams was the most predictive variable, showing a correlation of 0.324 with game outcomes. Within the LASSO Regression model, features that were labeled as influential included the difference in win percentage, the variability of points scored per game, and the difference in average points conceded by each team. These findings highlight the significance of team consistency, scoring variability, and defensive strength in predicting game outcomes.

Interpretation

Teams that score consistently are more likely to win a game. Teams that have won more games in the past, are more likely to win games in the future, especially if they play a team that has a lower record of wins. Teams with strong defenses, notable by conceding less points, are

more likely to win. A combination of all these factors, each with varying levels of importance, is the foundation of our model's predictive strategy.

6. Conclusions

Challenges of the study

When reflecting on the process of our project, the most challenging aspect of training our model was knowing what had the potential to make our model stronger. In essence, we had to find a way to "make the most of" our original data. Combining an understanding of what makes a basketball team "good" with the primarily numerical data was the first obstacle in our path. Once we had generated our ideal features from the data, the next challenge was selecting which combination of features allowed for the greatest accuracy. With the risk of overfitting or overloading our model with negatively correlated features, we accessed our features and ran multiple models to find the highest return in accuracy. With all these combined we experimented with limiting the inclusion of the earlier points in the data set. This reinforced our idea that granting a higher weight to the more recent games was more indicative of a team's chance of success.

Recommendations for the future

In the future, we need to pay closer attention to the original provided predictors before we feature engineer. It seems that our engineered features did not benefit our analysis too greatly. We can also look further into testing even more hyperparameters in the models.