Statistics 101C Final Project

Predicting NBA Game Outcomes from Historical Season Data (2023-2024)

Emma Morrison (406011062), Melissa Chang (805915408), Olivia Motmans (405952209)
Allison Lynn (006002253), Cassidy Sadowski (806003871), Anna Dupree (806145960)

Department of Statistics & Data Science, The University of California, Los Angeles

December 2024

Contents

1 Introduction

In this paper, we present our analysis of the NBA dataset, which contains a detailed record of NBA basketball games throughout previous seasons including team performance, game statistics, and outcomes. This dataset contains 2,460 entries and includes 24 features such as team name, match-up details, game date, win/loss outcomes, minutes played, points scored, field goals made (FGM), and other performance metrics. Our objective was to evaluate the prediction accuracy of various machine learning models in forecasting game outcomes based on these features.

To achieve this, we explored different feature engineering techniques and applied a variety of predictive models, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Linear Support Vector Classifier (SVC), Logistic Regression, Gradient Boosting, and Random Forest.

Through this analysis, we aim to assess the effectiveness of these methods and identify the most influential features for improving predictive accuracy in NBA game analytics. In the following sections, we explain our methods of data preprocessing, experimental setup, and results and analysis.

2 Feature Engineering

2.1 Data Description

Data from the 2023-2024 NBA season was provided to us with the following features, described in Table 1 below.

2.2 Data Preprocessing

We imported the our given data set into Python and converted the Game Date and W/L columns into datetime and binary formats, respectively, for easier machine reading. Information from the Match Up column was extracted into a new Home Advantage (binary) feature, and a new Opponent Team feature for further processing.

2.3 Transformation of Features

Since the data features provided for each game are statistics for the game itself, it would behoove us to exclude the direct game outcomes for the purpose of prediction. We decided to incorporate a rolling average of the most recent 10 games for all game metrics in our dataset (all features in Table 1, excluding Team, Match Up, and Game Date), as well as all advanced metrics directly obtained from game metrics (exclusive of the ELO rating). These and additional engineered features were then used to train our models.

2023-2024 NBA Dataset	
Feature	**Description**
Team	Team name (e.g. LAL, BOS)
Match Up	Team match up (e.g. LAL vs. BOS if LAL is playing a home game. LAL @ BOS if LAL is playing an away game)
Game Date	Date of game in MM/DD/YYYY format
W/L	Win or Loss indicator (W or L)
MIN	Minutes played
PTS	Points scored
FGM	Field goals made
FGA	Field goals attempted
FG%	Field goal percentage
3PM	Three-point field goals made
3PA	Three-point field goals attempted
3P%	Three-point field goal percentage
FTM	Free throws made
FTA	Free throws attempted
FT%	Free throw percentage
OREB	Offensive rebounds
DREB	Defensive rebounds
REB	Total rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal fouls
+/-	Plus/minus statistic (point difference with other team at the end of the game)

Table 1: Description of our given dataset.

2.4 Advanced Metrics

Sports analytics has had a long history, with well-established means of measuring team and player performance. We used established NBA metrics in our feature engineering, constrained by the available data in our data set. Table 2 below describes our list of advanced metrics and their formulas.

2.5 ELO Rating

Elo ratings are a measure of a team's relative skill level, with higher ratings indicating stronger teams. The ratings are updated after each game based on the outcome, margin of

Advanced NBA Metrics	
Metric	**Formula**
True Shooting Percentage	$\text{TS\%} = \frac{\text{PTS}}{2(\text{FGA}+0.44\text{FTA})}$
Effective Field Goal Percentage	$\text{eFG\%} = \frac{\text{FGM}+0.53\text{PM}}{\text{FGA}}$
Assist Percentage	$\text{AST\%} = \frac{\text{AST}}{\text{FGM}}$
Turnover Percentage	$\text{TOV\%} = \frac{\text{TOV}}{\text{FGA}+0.44\text{FTA}+\text{TOV}}$
Offensive Rebound Percentage	$\text{OREB\%} = \frac{\text{OREB}}{\text{OREB}+\text{Opponent\_DREB}}$
Defensive Rebound Percentage	$\text{DREB\%} = \frac{\text{DREB}}{\text{DREB}+\text{Opponent\_OREB}}$
Possessions	$\text{POSS} = \text{FGA} - \text{OREB} + \text{TOV} + 0.475\text{FTA}$
ELO Rating	$R_{i+1} = k(S_{home} - E_{home} + R_i)$ where $E_{home} = \frac{1}{1+\frac{R_{opponent}-R_{home}}{400}}$ $k = 20\frac{(MOV_{winner}+3)^{0.8}}{7.5+0.006R_{difference}}$

Table 2: Advanced NBA metrics and their corresponding formulas.

victory, and the relative strength of the opponents. We computed Elo ratings by using an iterative process, starting with a baseline score of 1500 and subtracted or added points based on the results of each game. The formula used to update the ratings takes into account the expected outcome (based on the teams' current Elo ratings), the actual result of the game, and a dynamic K factor that adjusts based on the margin of victory. This feature provides insight into the performance trends of teams, which is important for accurately predicting game outcomes.

2.6 Stability

We engineered a feature which demonstrates the consistency of a team in all given features of the dataset. To do this, we calculated the variance of each given feature in a rolling

window of the 10 most recent games that a team played. The features for which stability was calculated include: binary win/loss, minutes played, points scored, field goals made, field goals attempted, field goal percentage, three-point field goals made, three-point field goals attempted, three-point field goal percentage, free throws made, free throws attempted, free throw percentage, offensive rebounds, defensive rebounds, total rebounds, assists, steals, blocks, turnovers, personal fouls, and plus/minus.

2.7 Team Matchup History

We included a rolling feature on a scale from 0-1, which measured the previous wins of each matchup. For instance, if the Boston Celtics had played the Atlanta Hawks three previous times, and won two out of three games, their previous matchup rating for the fourth game would be 66%, whereas the Hawks rating for the fourth game would be 33%. If no previous games have been played between the two teams, the rating is 0%.

2.8 Win Streak

A win streak is an indicator of psychological advantage and historical competence of a team. Going into a game with a win streak may affect team performance. Hence, we incorporated this feature into our data.

2.9 Weighting

In order to weigh recent wins more heavily than less recent games, we implemented a weighting feature utilizing the exponential decay function. We used cross validation based on log loss to find the best lambda in the exponential decay function. This gave us a metric that has a higher weighting for more recent games as that more accurately reflects how a team would perform in the next game.

$$w_i = e^{-\lambda(t_{current} - t_i)} \quad \text{where} \quad \lambda = \text{decay rate}$$

The exponential decay function penalizes less recent games not weighting them as high. We also choose a relatively high lambda from cross validation so it would penalize less recent games.

2.10 Data Selection for Training and Testing

Note that because we utilized a historical rolling average for many game metrics, the first 10 games for every team have a NaN entry in their rolling average metrics. Due to this, we removed the first 336 chronological rows from our training set, so that there would be no NaN values in our training data. Since game prediction relies on training on previous games for prediction of future games, we selected a cutoff date of 04/01/2024 (row 2230) for our training data. This left 230 rows for validation testing and 1893 rows for training, or approximately a 10/90% testing/training split.

2.11 Summary of Features

Below, we give a summary of all engineered features. A full dataset of these features were created.

Summary of Engineered Features	
Final Features	**Description**
Home Advantage	Binary designation of home team
Win Streak	Win streak of team at the start of the game
Advanced Metrics	All features described in Table 2
Rolling Average	Rolling average of team metrics from most recent 10 games (MIN, PTS, FGM, FGA, FG%, 3PM, 3PA, 3P%, FTM, FTA, FT%, OREB, DREB, REB, PF, +/-, TS%, eFG%, AST%, TOV%, OREB%, DREB%, POSS)
Previous Match Up	Wins, losses, and ratio of win/loss history between the playing teams
Weight	Exponential decay function penalizing less recent games
Stability	Rolling window of variance of each given metric over previous 10 games (W/L, MIN, PTS, FGM, FGA, FG%, 3PM, 3PA, 3P%, FTM, FTA, FT%, OREB, DREB, REB, AST, STL, BLK, TOV, PF, +/-)

Table 3: Summary of features fed into Linear SVC model.

3 Feature Selection

We chose Support Vector Classifier (SVC) with an L1 penalty (λ) to select features for our final model due to its efficiency and effectiveness in identifying the most relevant predictors in high-dimensional datasets. Linear SVC is particularly well-suited in this context because the L1 regularization encourages sparsity, which successfully eliminates less significant features by setting coefficients to zero. After preprocessing the data and standardizing it, we trained the Linear SVC model on the training dataset to evaluate the predictive importance on each feature. Using SelectFromModel, we took the most influential features based on their non-zero coefficients, resulting to only including the most predictive variables in our next analyses. Table 4 below outlines our results. Some feature selections were excluded for brevity. **The highest accuracy model used a penalty of $\lambda = 0.03$, and our selected features are shown in Table 4 below.**

Feature Selection	**Accuracy**	**L1 Penalty λ**
Home Advantage Win Streak Team Elo Opponent Elo Prev Matchup Losses Prev Comp Ratio Binary (Win/Loss) Stability	.917391	0.01
(features excluded for brevity)	.921739	0.02
Home Advantage Win Streak +/- Rolling Average 3PA Rolling Average OREB% Rolling Average DREB% Rolling Average POSS Rolling Average Team Elo Opponent Elo Prev Matchup Losses Prev Comp Ratio FTA Stability STL Stability BLK Stability PF Stability +/- Stability Binary (Win/Loss) Stability	.926087	0.03
(features excluded for brevity)	.921739	0.04

Table 4: Selected results from a Linear SVC.

4 Model Selection

We ran several different models to determine the best fit. These include:

1. **LDA:** Linear Discriminant Analysis (LDA) is a generative model that assumes that the shared covariance structure of each class of features results in linear decision boundaries.

2. **QDA:** Quadratic Discriminant Analysis (QDA) is a generative model that allows each class to have its own covariance matrix, enabling it to model more complex, non-linear relationships between features.

3. **Logistic Regression:** Logistic Regression is a discriminative model used for binary

classification tasks. It estimates the probability of an outcome by modeling the relationship between the features and the response with a logistic function.

4. **Random Forest**: Random Forest is an ensemble method that enhances overall accuracy by combining the predictions of multiple decision trees. It employs bootstrap aggregation, or bagging, where each tree is trained on a randomly selected subset of the data with replacement. Unlike traditional bagging methods, Random Forest introduces an additional layer of randomness by selecting a random subset of features at each node split. This ensures greater diversity among the trees, reduces correlation between them, and improves the model's robustness and performance.

5. **Gradient Boosting**: Gradient Boosting is an ensemble method that builds decision trees sequentially, with each tree focusing on correcting misclassifications in the previous tree. Unlike Random Forest, which trains trees independently, Gradient Boosting builds trees sequentially, gradually increasing accuracy. It assigns greater importance to misclassified data points, allowing the model to adapt to challenging cases and minimize errors . Because of this, Gradient Boosting is highly effective in capturing complex patterns and delivering high prediction accuracy.

5 Results and Analysis

5.1 Model

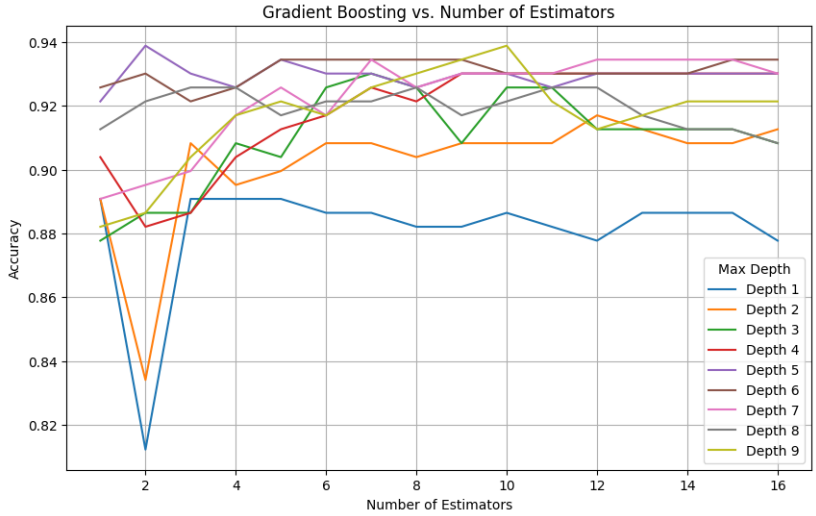

Figure 1: Boosting Accuracy based on Number of Estimators (n) and Depth (d).

Our results tell us that Gradient Boosting and Random Forest models give us similarly high prediction accuracy. The highest prediction accuracy of 93.87% was achieved by both models, using 10 and 2 estimators with a depth of 9 and 5, respectively, for Random Forest, and using 11 and 1 estimator with a depth of 7 and 8, respectively, for Gradient Boosting.

Model	**Accuracy**
SVC ($\lambda = 0.01$)	.917391
SVC ($\lambda = 0.02$)	.921739
SVC ($\lambda = 0.03$)	.926087
SVC ($\lambda = 0.04$)	.921739
LDA	.890830
QDA	.908297
Logistic Regression	.912664
Bagging (n = 17)	.925764
Random Forest (n = 17)	.0.921397
Random Forest (best accuracy) (n = 10, depth = 9)	.0.938865
Random Forest (best accuracy) (n = 2, depth = 5)	.0.938865
Gradient Boosting (n = 17)	.934498
Gradient Boosting (best accuracy) (n = 11, depth = 7)	.0.938865
Gradient Boosting (best accuracy) (n = 1, depth = 8)	.0.938865

Table 5: Comparison of results from various models.

This suggests that that there may be few highly determinant factors that predict a majority of team outcomes.

5.2 Selected Features

An interesting result of our Linear SVC model is that a large proportion of our final selected predictors were based on team-to-team interactions for each game. For example, Prev Matchup Losses and Prev Comp Ratio both measure the historical performance of Team A and Team B for a match between A and B. Elo ratings are a measure of a team's relative skill, but that our model decided to incorporate both Team Elo and Opponent Elo indicates that the opponent team's skill is just as important as the playing team's skill at determining game outcome. The metrics for +/-, OREB%, and DREB% also inherently incorporate the team's historical performance against other teams. Together, this paints a picture of basketball games outcomes as a highly interactive sport, and highly opponent dependent.

6 Conclusion

By leveraging our engineered features, we have improved the predictive power of our models. Among the tested methods, Gradient Boosting and Random Forest models achieved the highest accuracy in forecasting game results. The success of these can be attributed due to its effective capturing of both linear and nonlinear relationships in the data.