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, as well as all advanced metrics directly obtained from game metrics (exclusive of the ELO rating).

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.

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 $E_{home} = \dfrac{1}{1 + \frac{R_{opponent} - R_{home}}{400}}$ $k = 20\dfrac{(MOV_{winner} + 3)^{0.8}}{7.5 + 0.006R_{difference}}$

Table 2: Advanced NBA metrics and their corresponding 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 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 scores, 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

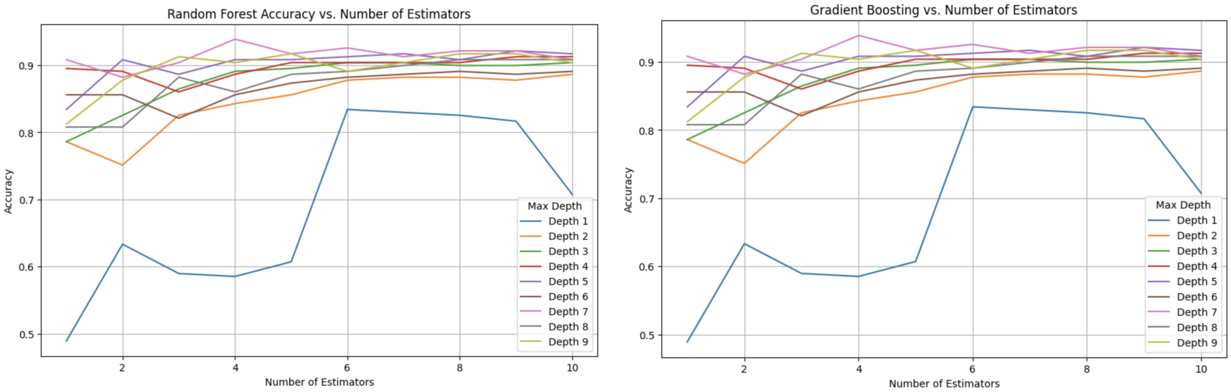
A win streak is an indicator of psychological advantage and historical competence of a team. Going into a game with a win streak may affect team performance. Hence, we incorporated this feature into our data.

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.

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

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.



(a) Random Forest

(b) Boosting

Figure 1: Best Depth and Number of Estimators

2.10 Removal of Game Results

As mentioned in section 2.2, transformation of features, the columns provided in the initial dataset (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 goals percentage, free throws made, free throws attempted, free throws percentage, offensive rebounds, defensive rebounds, total rebounds, assists, steals, blocks, turnovers, personal fouls, and plus/minus statistic) are all results of the game. We removed these columns from our dataset since we would not have this information when attempting to predict the game's outcome. Instead, we utilized the engineered features built from these initial columns to train our models.

2.11 Engineered Dataset

include some table (can it have really small margins?) of final dataset before we fed it into SVC

3 Feature Selection

We used a Support Vector Classifier (SVC) model to select features for our final model. This is because

4 Model Selection

Features	**Accuracy**	**L1 Penalty**
Home Advantage Win Streak Team Elo Opponent Elo Prev Matchup Losses Prev Comp Ratio Binary (Win/Loss) Stability	.9170	0.01
Home Advantage Win Streak +/- Rolling Average DREB% Rolling Average Team Elo Opponent Elo Prev Matchup Losses Prev Comp Ratio FTA Stability STL Stability BLK Stability PF Stability Binary (Win/Loss) Stability	.9214	0.02
Home Advantage Win Streak +/- Rolling Average 3PM Rolling Average 3PA Rolling Average OREB% Rolling Average DREB% 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	.9258	0.03

Table 3: Feature selection results from a Linear SVC.

Model	**Accuracy**
SVC ($\lambda = 0.01$)	.908
SVC ($\lambda = 0.02$)	0.5
SVC ($\lambda = 0.03$)	0.5
LDA	0.5
QDA	0.5
Random Forest	0.938865
Gradient Boosting	0.943231

Table 4: Comparison of results from various models.