Multivariable Analysis Final Report: Predicting Game Outcomes on NBA Team Starting Lineup

Max Baxley, Jenna Odom, and Vaishnavi Pulla

Department of Statistics, The University of Georgia

STAT 4250: Multivariable Analysis

April 21, 2025

The National Basketball League (NBA) provides an ideal setting for real-time statistical modeling and predictive analysis. In this project, we will be investigating how in-game metrics influence game outcomes and whether final scores and win/loss results can be accurately predicted based on performance features available during the game. Through this research, we are seeking to uncover meaningful insights into the trajectory of NBA games and assess the predictive power of different modeling approaches.

We collected data from publicly available sources, including Kaggle NBA datasets and the official NBA API. These datasets provided extensive coverage of both player-level and team-level statistics. For each game, we extracted time-stamped events such as scoring plays, substitutions, fouls, turnovers, and assists. We also gathered metadata such as game location, home or away designation, team matchups, and the type of game (regular season, play-in, or playoff). This rich structure allowed us to analyze performance dynamics across varying game contexts.

At the player level, we included features such as points per game, rebounds, assists, shooting percentages, and plus-minus scores. These variables reflected each player's contribution to team performance and enabled us to construct summary features for each team based on its lineup. Before modeling, we preprocessed the data to address missing values, resolve formatting inconsistencies, and properly encode categorical variables. This ensured a clean dataset and minimized bias or noise in the downstream analysis.

We began with Exploratory Data Analysis (EDA) to understand the distribution of game statistics and uncover potential relationships between variables. Using visualizations such as histograms, boxplots, and correlation heatmaps, we explored the connections between features like field goal percentage, three-point shooting, turnovers, and total game score. Descriptive

statistics helped us identify average performance levels, outliers, and key indicators across teams and seasons. EDA also allowed us to assess the influence of contextual variables such as home-court advantage and game type.

Next, we applied regression modeling to examine how in-game statistics predicted final scores. We started with simple linear regression models to explore individual variable relationships and then extended our approach to multivariate linear regression, which captured joint effects. To improve prediction performance and address multicollinearity, we implemented Ridge Regression and LASSO. These regularization techniques constrained coefficient estimates, reduced overfitting, and facilitated automatic feature selection. We evaluated models using cross-validation to select optimal penalty parameters and ensure that results generalized beyond the training data.

We also incorporated dimensionality reduction techniques to simplify our high-dimensional feature space. Using Principal Component Analysis (PCA) and Factor Analysis, we transformed original variables into fewer uncorrelated components that captured core performance traits such as offensive strength, defensive pressure, and efficiency. These components revealed latent structures in team behavior and streamlined input for subsequent classification models.

To predict whether a team won or lost, we approached the task as a binary classification problem. Our baseline method was Linear Discriminant Analysis (LDA), which assumed normally distributed features and linear boundaries between classes. LDA provided an interpretable projection of the feature space and highlighted variables most responsible for separating wins and losses.

However, in many cases, game outcomes did not follow linear separability, particularly in close games. To address this, we implemented Support Vector Machines (SVM), which used

kernel functions to model nonlinear decision boundaries. SVMs relied on a hinge loss function and were effective at creating maximum-margin classifiers. We encoded game outcomes as 1 (win) and -1 (loss) to align with the algorithm's requirements. SVMs offered flexibility and robustness, especially in high-dimensional settings.

We also explored ensemble learning methods to enhance prediction accuracy. Specifically, we implemented Random Forests, which aggregated predictions from multiple decision trees trained on bootstrapped samples, and Gradient Boosting, which built trees sequentially to minimize residual error. Finally, we integrated LDA, SVM, and Random Forest outputs into a Voting Classifier. This ensemble model selected the majority class prediction across its base models, leveraging their combined strengths to improve overall performance. Ensemble methods proved particularly valuable for handling edge cases, such as games with shifting momentum or minimal score margins.

We assessed model performance using metrics appropriate to both regression and classification tasks. For regression models that predicted numeric outcomes such as final scores or point differentials, we used Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics quantified prediction accuracy and penalized large deviations to reflect real-world error impacts.

For classification models, we evaluated accuracy, precision, and recall. Accuracy captured the proportion of correct predictions, while precision and recall provided insight into how well models identified wins without generating false positives or negatives. We applied k-fold cross-validation to partition data into multiple training and validation sets, which helped ensure robustness and mitigated overfitting.

To simulate real-world prediction conditions, we used temporally separated training and testing sets. For example, we trained models using games from earlier in the season and tested them on later games. This approach mimicked practical forecasting scenarios, where only past data is available to predict future events. It also helped us observe how models behaved under evolving player performance trends and strategic adjustments throughout the season.

As an initial implementation, we built an SVM classifier that predicted whether a team won or lost based on its starting five players. We aggregated key metrics for each starting lineup, including points per game, shooting efficiency, and player impact measures such as PER (Player Efficiency Rating). The model used these features to construct a feature vector representing a team's strength at the beginning of each game.

We encoded the outcome as 1 for a win and -1 for a loss to ensure compatibility with the SVM's hinge loss function. Preliminary results suggested that the model successfully identified meaningful patterns based on lineup configurations. Even in the absence of in-game statistics, the model achieved reasonable accuracy, indicating that player combinations alone held predictive value. These early findings provided a solid foundation for scaling the analysis to incorporate real-time performance metrics and more complex interactions.

We hypothesized that halftime score differential, shooting percentage, turnover count, and total assists were the most predictive variables for final outcomes. These features captured a team's offensive efficiency and ball-handling capabilities. We expected that models using regularization—particularly LASSO—would outperform basic linear regression by reducing noise and emphasizing the most informative predictors.

In classification tasks, we anticipated that SVMs and ensemble classifiers would yield superior accuracy, especially in games with narrow margins or volatile pacing. Dimensionality

reduction methods such as PCA were expected to uncover clusters of similar team behavior and improve model discrimination. We also expected ensemble models, especially voting classifiers, to demonstrate increased resilience to misclassification by integrating multiple model perspectives.

This project develops a comprehensive analytical pipeline for predicting NBA scoreboard trends and outcomes. By integrating exploratory analysis, regression, dimensionality reduction, and ensemble classification, we aim to uncover the most impactful performance metrics and build predictive models with real-world utility. The models we construct align with core statistical concepts covered in STAT 4250, including regularized regression, classification, and model evaluation using cross-validation