**NBA Dataset Analysis Using Machine Learning**

**Introduction**

The 2024 National Basketball Association (NBA) season provided a rich dataset, comprising over 200,000 rows of game data and more than 15 predictors. Analyzing such data offers an opportunity to explore critical questions in basketball analytics. In this project, we focus on three primary research questions:

• Is there any impact of game environment (home or away teams) and player positions on the likelihood of making a shot?

• Which shot types and action types are most valuable for teams and players to attempt?

• What factors most significantly influence the probability of a successful shot attempt?

By deploying machine learning models, including logistic regression, decision trees, and neural networks, this study aims to uncover actionable insights. The findings will inform players and coaches, enhancing game strategies and decision-making processes. Furthermore, this project demonstrates the potential of data-driven approaches in sports analytics, offering a reproducible methodology for similar studies.

**Data Understanding and Preparation**

**Data Source and Variable Explanation**

The dataset used in this project contains information from the NBA 2024 season, including game details, player actions, and shot outcomes. Key variables include:

**• Categorical Predictors**: Home team, away team, player position, shot type, zone range, etc.

**• Numerical Predictors**: Shot distance, game time variables (minutes and seconds left), and quarters.

**• Dependent Variable**: Whether a shot was made (binary).

**Data Processing**

**Cleaning**

Rows with missing values in critical variables (e.g., position, position group) were removed. This step reduced the dataset but ensured the integrity of the analysis. Categorical variables were transformed using one-hot encoding and target encoding where appropriate.

**Feature Engineering**

To improve model performance, several features were created or transformed:

**• Factorization**: Categorical variables such as action type, zone range, and shot type were encoded numerically.

**• Home Game Indicator**: A binary variable was introduced to indicate if the game was played at the player’s home court.

Additionally, exploratory data analysis (EDA) revealed correlations between features, such as a positive correlation between shot distance and success probability for specific zones. Visualizations from EDA, including histograms and scatter plots, are included in Appendix D. Below is an example of the initial dataset preview:

These transformations ensured the dataset was suitable for machine learning models.

**Modeling**

**Model Building**

**Logistic Regression**

Logistic regression models were built with various feature sets to evaluate the relationships between predictors and the probability of making a shot. Encoding methods were used to handle categorical variables, and correlation analysis was conducted to avoid multicollinearity.

**Decision Tree**

Two decision tree models were tested:

• A depth-limited tree with 5 levels, achieving an accuracy of 62.4%.

• An optimized tree with 7 levels, slightly improving accuracy to 62.6%.

Figure 1 below shows the visualization of the decision tree with 5 levels, illustrating how different features influence shot-making outcomes.

**Neural Network**

Two neural network architectures were evaluated:

• A 5,5 configuration yielding 62.9% accuracy.

• A 3,3 configuration performing marginally better at 62.98%.

**Results**

**Comparative Analysis**

The three models demonstrated similar accuracy scores:

• Logistic Regression: 62.1%

• Decision Tree: 62.6%

• Neural Network: 62.9%

Despite its marginally higher accuracy, the neural network offered limited interpretability. Decision trees, in contrast, provided clear insights into feature importance, aligning better with the research objectives.

**Feature Importance**

Decision tree analysis highlighted key factors influencing shot success:

**• Top Features**: Shot distance, action type, and zone range were consistently ranked as the most influential.

**• Insights**: Short to mid-range shots (8-16 ft) and specific zones (e.g., corners) increase success probabilities.

Figure 2 below displays the feature importance rankings derived from the decision tree model:

Another visualization of feature importance for an alternative model iteration is shown below:

**Evaluation**

**Performance Metrics**

The primary evaluation metric was accuracy, complemented by precision and recall from classification reports. While the models’ performances were comparable, decision trees provided the best trade-off between accuracy and explainability.

**Addressing Research Questions**

**• Impact of Game Environment**: Home-court advantage did not emerge as a significant factor.

**• Valuable Shot Types**: Running layup shots, jump shots, and 3-point shots were identified as high-value attempts.

**• Key Influencing Factors**: Action type, shot distance, and zone range were the most predictive features.

**Recommendations**

Based on the analysis, the following recommendations are proposed:

**• Training Focus**:

• Emphasize practice on running layup and jump shots.

• Encourage players to attempt 3-point shots from the left and right corners.

**• Shot Selection Strategies**:

• Avoid shots within 8 feet unless in the restricted area.

• Target shots between 8-16 feet for higher success rates.

**• Positioning Insights**:

• Players should focus on the left and right-side center zones for optimal scoring opportunities.

**• Data-Driven Coaching**:

• Leverage player-specific data to tailor training programs based on individual performance trends.

Implementing these recommendations can improve team productivity and individual player performance.

**Conclusion**

Our study leveraged machine learning models to analyze the 2024 NBA season dataset. Decision trees emerged as the most interpretable model, revealing critical factors influencing shot success. The findings provide actionable insights, such as focusing on specific shot types and zones. This project demonstrates the potential of data-driven strategies to enhance basketball analytics and optimize player training and game strategies.

Future work could explore integrating additional contextual features, such as player fatigue or defensive positioning, to further refine shot prediction models.

**References**

• APA-style citations for data sources, tools, and academic references used throughout the project.

**Appendices**

**Appendix A: Feature Importance Table**

**(DF2)**

**DF 3**

**DF 4**

**DF 5**

**Appendix B: Code Snippets**

**Decision Tree (DF4)**

**Neural Network:**

**Logistic Regression**

**Poster:**