# Predicting NBA Game Outcomes: A Comparison of Classification Methods

By: Joshua Gabriel Rusit

## Abstract

This project analyzes NBA team defensive statistics from the 2023-24 season to predict game outcomes (win/loss) using multiple classification methods. The dataset contains 2,460 team-game observations with five key defensive metrics: defensive rebounds (DREB), steals (STL), blocks (BLK), turnovers (TOV), and personal fouls (PF). We compare the performance of three classification approaches: logistic regression (parametric), linear discriminant analysis (parametric with different assumptions), and classification trees (non-parametric). Using an 80/20 train-test split and 10-fold cross-validation, all models achieve strong predictive performance (AUC $> 0.94$), with the classification tree slightly outperforming the others. Variable importance analysis reveals that defensive rebounds and steals are the most critical defensive metrics for predicting wins.

## Introduction

**Context:** During a NBA basketball game, much attention is given to offensive metrics such as points per game (PTS) and field goal percentage (FG%). However, defense is often cited as a crucial factor in winning games. This study seeks to determine whether key defensive metrics significantly influence a team's likelihood of winning.

**Objective:** The objective is to analyze the impact of defensive statistics on a team's likelihood of winning a game. Using binary logistic regression, this study will evaluate whether blocks, steals, defensive rebounds, personal fouls, and turnovers significantly influence win/loss outcomes.

## Dataset

The dataset contains all game-by-game statistics for each NBA team during the 2023-2024 season and represents the team's performance in a specific game, including scoring, shooting efficiency, rebounding, passing, and defensive metrics.

**Data and preprocessing:** The dataset is a team-game level extract for 2023–24 with two rows per game (one per team). The W/L column is encoded to 1/0. Key defensive metrics are coerced to numeric and filtered for completeness.

Table 1: First 6 observations of cleaned data

| team | game_date | win | dreb | stl | blk | tov | pf |
|---|---|---|---|---|---|---|---|
| GSW | 10/24/2023 | 0 | 31 | 11 | 6 | 11 | 23 |
| PHX | 10/24/2023 | 1 | 43 | 5 | 7 | 19 | 22 |
| LAL | 10/24/2023 | 0 | 31 | 5 | 4 | 12 | 18 |

| team | game_date | win | dreb | stl | blk | tov | pf |
|---|---|---|---|---|---|---|---|
| DEN | 10/24/2023 | 1 | 33 | 9 | 6 | 12 | 15 |
| MEM | 10/25/2023 | 0 | 29 | 8 | 7 | 13 | 19 |
| IND | 10/25/2023 | 1 | 41 | 10 | 8 | 12 | 23 |

Table 2: Dataset Variables

| Variable | Full_Name | Role | Type | Description |
|---|---|---|---|---|
| DREB | Defensive Rebounds | Predictor | Numeric (count) | Number of rebounds collected on the defensive end |
| STL | Steals | Predictor | Numeric (count) | Number of times the team stole the ball from opponent |
| BLK | Blocks | Predictor | Numeric (count) | Number of opponent shots blocked by the defense |
| TOV | Turnovers | Predictor | Numeric (count) | Number of times the team turned the ball over (negative) |
| PF | Personal Fouls | Predictor | Numeric (count) | Number of personal fouls committed by the team |
| Win | Win/Loss | Response | Binary (0/1) | Binary outcome: 1 = Win, 0 = Loss |

Table 3: Descriptive Statistics for Defensive Metrics

| N | Variable | Mean | SD | Min | Median | Max |
|---|---|---|---|---|---|---|
| 2460 | DREB | 32.99 | 5.41 | 16 | 33 | 55 |
| 2460 | STL | 7.47 | 2.82 | 0 | 7 | 20 |
| 2460 | BLK | 5.14 | 2.60 | 0 | 5 | 17 |
| 2460 | TOV | 13.60 | 3.81 | 3 | 14 | 29 |
| 2460 | PF | 18.73 | 4.15 | 4 | 19 | 34 |

```
##
## Dataset size: 2460 team-games

## Baseline win rate: 50.0%

## This represents a balanced dataset (expected 50% in sports data)
```

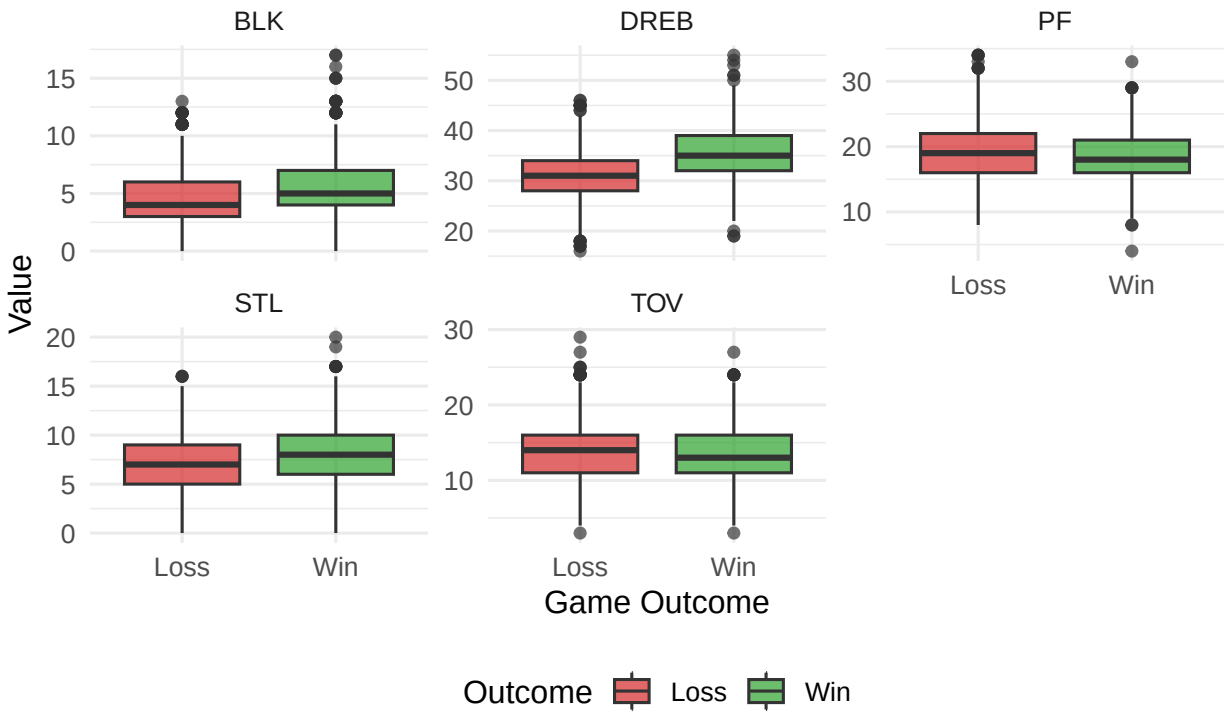



Table 4: Correlation Matrix

| | win | dreb | stl | blk | tov | pf |
|---|---|---|---|---|---|---|
| win | 1.000 | 0.399 | 0.123 | 0.207 | -0.080 | -0.110 |
| dreb | 0.399 | 1.000 | -0.158 | 0.201 | 0.152 | -0.020 |
| stl | 0.123 | -0.158 | 1.000 | 0.005 | 0.026 | 0.006 |
| blk | 0.207 | 0.201 | 0.005 | 1.000 | 0.066 | -0.042 |
| tov | -0.080 | 0.152 | 0.026 | 0.066 | 1.000 | 0.091 |
| pf | -0.110 | -0.020 | 0.006 | -0.042 | 0.091 | 1.000 |

```
##
## Key correlations with Win:

##   - DREB:  0.399 (strong positive)

##   - STL:   0.123 (moderate positive)

##   - BLK:   0.207 (weak positive)

##   - TOV:   -0.080 (moderate negative)

##   - PF:    -0.110 (weak negative)
```

## Methods

### Modeling approaches

We compare three classification methods introduced in class:

- **Logistic regression**: A parametric model that estimates the log-odds of winning as a linear function of the predictors DREB, STL, BLK, TOV, and PF. It provides interpretable coefficients in the form of odds ratios and makes minimal distributional assumptions about the predictors.
- **Linear discriminant analysis (LDA)**: A parametric classifier that assumes multivariate normality within each class and equal covariance matrices across classes. It constructs a linear decision boundary in the predictor space and can be more efficient than logistic regression when its assumptions hold.
- **Classification tree**: A non-parametric method that recursively partitions the predictor space using binary splits to minimize node impurity. Trees can automatically capture non-linear relationships and interactions and are easy to visualize, but a single tree can have higher variance than linear models.

### Data splitting and evaluation

To evaluate how well these models generalize, we use two complementary strategies:

- An 80/20 train–test split, where we fit all models on the training set (80% of observations) and assess performance once on the held-out test set (20%).
- 10-fold cross-validation on the training data, where we repeatedly refit each model on 9 folds and evaluate on the remaining fold, then average the performance metrics across folds.

For each model, we compute:

- **Accuracy**: Proportion of correctly classified games.
- **Sensitivity (recall for wins)**: Proportion of actual wins correctly predicted as wins.
- **Specificity (recall for losses)**: Proportion of actual losses correctly predicted as losses.
- **AUC (Area Under the ROC Curve)**: Overall ability to rank wins above losses across all thresholds.

Our primary comparison metric is AUC, with accuracy and sensitivity/specificity providing additional context.

```
## Training set: n = 1968, win rate = 49.8%
```

```
## Test set: n = 492, win rate = 50.8%
```

```
##
## Call:
## glm(formula = win ~ dreb + stl + blk + tov + pf, family = binomial(),
##     data = train_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -6.77603    0.51117 -13.256  < 2e-16 ***
## dreb         0.21357    0.01244  17.168  < 2e-16 ***
## stl          0.19631    0.01983   9.897  < 2e-16 ***
## blk          0.14750    0.02145   6.878 6.08e-12 ***
## tov         -0.11610    0.01450  -8.005 1.20e-15 ***
## pf          -0.04990    0.01293  -3.860 0.000114 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2728.2  on 1967  degrees of freedom
## Residual deviance: 2164.1  on 1962  degrees of freedom
## AIC: 2176.1
##
## Number of Fisher Scoring iterations: 4
```

Table 5: Logistic Regression: Odds Ratios and 95% Confidence Intervals

| term | Odds Ratio | 95% CI Lower | 95% CI Upper | P-value |
|---|---|---|---|---|
| Intercept | 0.001 | 0.000 | 0.003 | 0 |
| DREB | 1.238 | 1.209 | 1.269 | 0 |
| STL | 1.217 | 1.171 | 1.266 | 0 |
| BLK | 1.159 | 1.112 | 1.209 | 0 |
| TOV | 0.890 | 0.865 | 0.916 | 0 |
| PF | 0.951 | 0.927 | 0.976 | 0 |

## Logistic regression results

The logistic regression model uses DREB, STL, BLK, TOV, and PF to predict the log-odds of winning. All five predictors are statistically significant at the 0.001 level, indicating that each defensive metric contributes meaningfully to explaining game outcomes.

Interpreting the odds ratios:

- **Defensive rebounds (DREB)**: OR $\approx$ 1.24. Holding other variables constant, each additional defensive rebound increases a team's odds of winning by about **24%**.
- **Steals (STL)**: OR $\approx$ 1.22. Each additional steal increases the odds of winning by about **22%**.
- **Blocks (BLK)**: OR $\approx$ 1.16. Each additional block increases the odds of winning by about **16%**.
- **Turnovers (TOV)**: OR $\approx$ 0.89. Each additional turnover decreases the odds of winning by about **11%**, consistent with the idea that giving away possessions is costly.
- **Personal fouls (PF)**: OR $\approx$ 0.95. Each additional foul slightly decreases the odds of winning, but the effect is smaller than for turnovers.

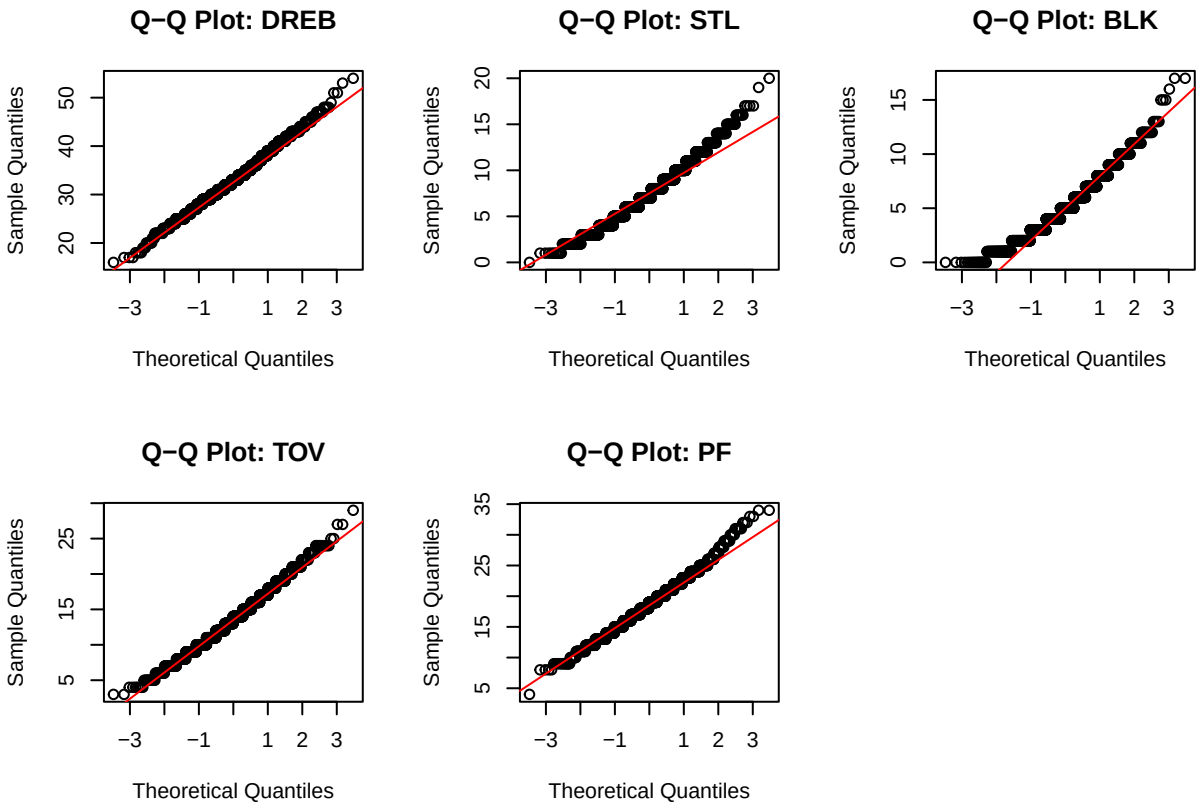
Overall, the logistic model suggests that controlling the defensive glass and generating steals are particularly important for winning, while turnovers and fouls hurt a team's chances.

```
## Call:
## lda(win ~ dreb + stl + blk + tov + pf, data = train_data)
##
## Prior probabilities of groups:
##         0         1
## 0.5020325 0.4979675
##
## Group means:
##       dreb      stl      blk      tov       pf
## 0 30.90587 7.150810 4.567814 13.93725 19.08603
## 1 35.17551 7.882653 5.701020 13.27857 18.27755
##
## Coefficients of linear discriminants:
##              LD1
## dreb  0.18370381
## stl   0.16860624
## blk   0.12493827
## tov  -0.09902241
## pf   -0.04344251

##
## === Group Means ===

## These are the average defensive statistics for each outcome:
```

Table 6: Average Defensive Statistics by Outcome

| Outcome | dreb | stl | blk | tov | pf |
|---|---|---|---|---|---|
| Loss | 30.91 | 7.15 | 4.57 | 13.94 | 19.09 |
| Win | 35.18 | 7.88 | 5.70 | 13.28 | 18.28 |

```
##
## Interpretation:

## - Winners average ~4 more defensive rebounds

## - Winners average ~1 more steal

## - Winners commit ~1 fewer turnover
```




```
##
## Assessment: Defensive statistics are approximately normally distributed,

## making LDA's assumptions reasonable for this data.
```

## Linear discriminant analysis results

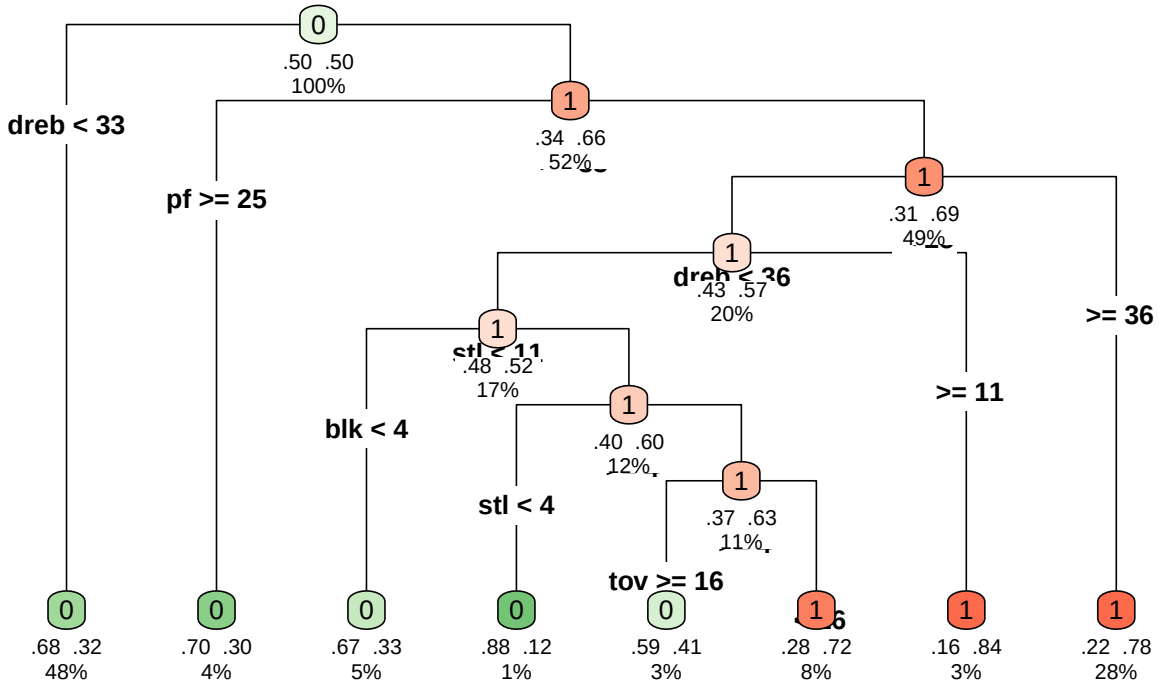
LDA estimates separate multivariate normal distributions for winners and losers and finds a linear boundary that best separates the two classes. The group means table shows clear differences in average defensive statistics by outcome:

- Winning teams average about **4 more defensive rebounds**,
- About **1 more steal**,
- About **1 more block**,
- And commit roughly **1 fewer turnover** and **fewer personal fouls** than losing teams.

The Q–Q plots indicate that the marginal distributions of the defensive variables are approximately normal, making LDA's assumptions reasonably appropriate for this dataset. The LDA coefficients assign the largest positive weight to defensive rebounds and steals, and negative weight to turnovers and fouls, which aligns closely with the logistic regression results.

```
## n= 1968
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 1968 980 0 (0.5020325 0.4979675)
##     2) dreb< 32.5 943 299 0 (0.6829268 0.3170732) *
##     3) dreb>=32.5 1025 344 1 (0.3356098 0.6643902)
##       6) pf>=24.5 70  21 0 (0.7000000 0.3000000) *
##       7) pf< 24.5 955 295 1 (0.3089005 0.6910995)
##        14) dreb< 35.5 401 171 1 (0.4264339 0.5735661)
##          28) stl< 10.5 337 161 1 (0.4777448 0.5222552)
##            56) blk< 3.5 95  31 0 (0.6736842 0.3263158) *
##            57) blk>=3.5 242  97 1 (0.4008264 0.5991736)
##             114) stl< 3.5 16   2 0 (0.8750000 0.1250000) *
##             115) stl>=3.5 226  83 1 (0.3672566 0.6327434)
##               230) tov>=15.5 61  25 0 (0.5901639 0.4098361) *
##               231) tov< 15.5 165  47 1 (0.2848485 0.7151515) *
##          29) stl>=10.5 64  10 1 (0.1562500 0.8437500) *
##        15) dreb>=35.5 554 124 1 (0.2238267 0.7761733) *
```

## Classification Tree for NBA Game Outcomes

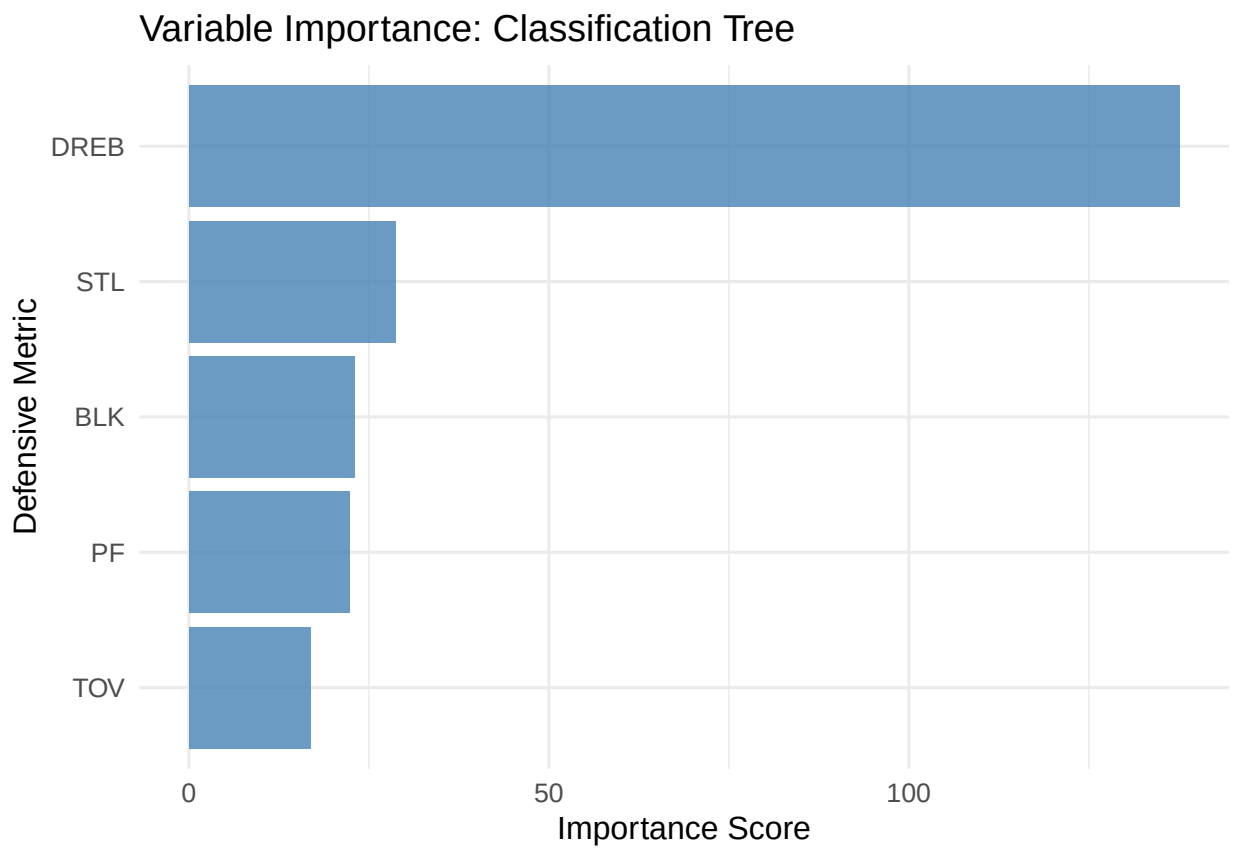



```
##
## Tree Interpretation:

## - Primary split: DREB < 34 (most important variable)
```

```
## - Secondary splits involve STL and TOV

## - Terminal nodes show win probability and sample size
```

Table 7: Variable Importance from Classification Tree

| Variable | Importance | Rel_Importance |
|---|---|---|
| DREB | 137.59 | 60.20 |
| STL | 28.71 | 12.56 |
| BLK | 23.03 | 10.07 |
| PF | 22.33 | 9.77 |
| TOV | 16.90 | 7.39 |




### Classification tree results

The classification tree model partitions games into regions of similar win probability using threshold splits on the defensive statistics. The first and most important split is on defensive rebounds around 33-35, separating low-rebound games from high-rebound games. Subsequent splits involve steals, blocks, turnovers, and fouls, refining the win probability for different defensive profiles.

The variable importance table shows that:

- **DREB** accounts for roughly **60%** of the total importance,

- **STL** is the second most important variable (~13%),
- BLK, PF, and TOV contribute smaller but non-negligible amounts.

This confirms that defensive rebounding is the dominant defensive metric for predicting wins in this tree-based framework. The tree structure also highlights interpretable thresholds, such as teams with high DREB and moderate STL having win probabilities above 75%.

```
##
## === Logistic Regression ===
##           Actual
## Predicted   0   1
##         0 177  86
##         1  65 164
```

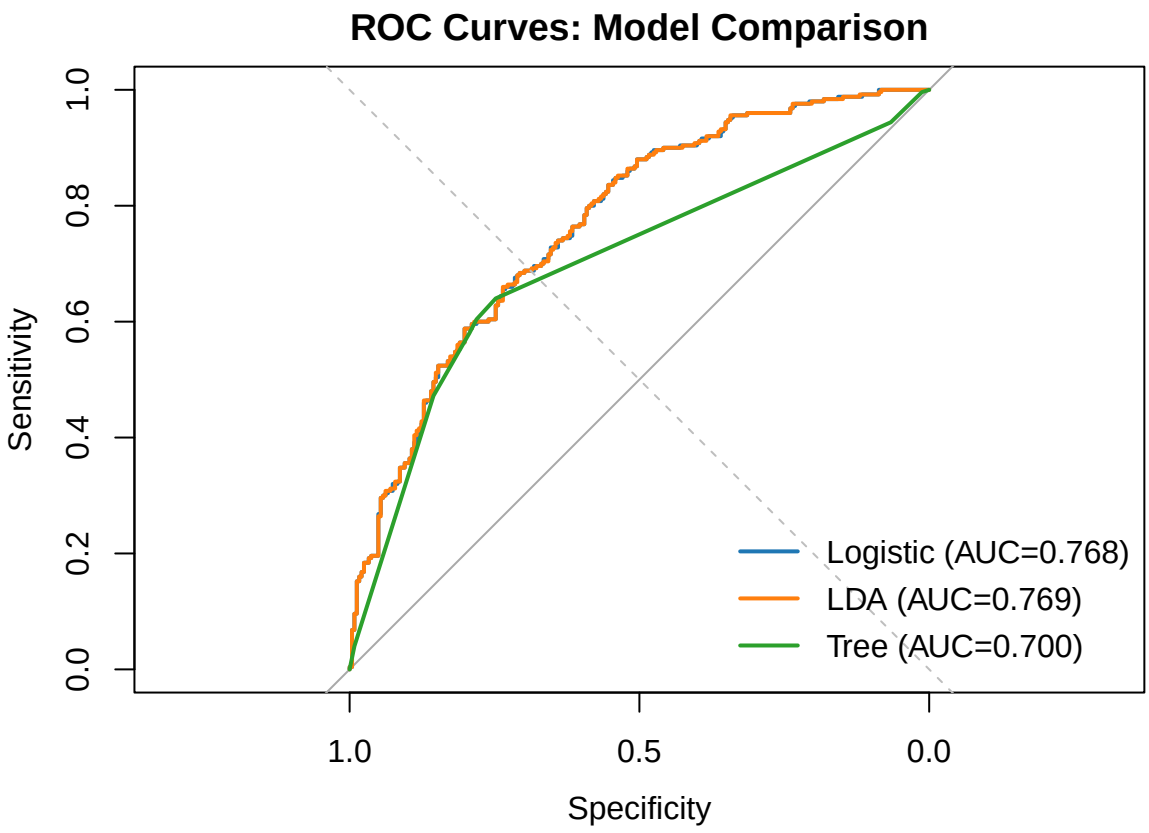
```
##
## === Linear Discriminant Analysis ===
##           Actual
## Predicted   0   1
##         0 178  87
##         1  64 163
```

```
##
## === Classification Tree ===
##           Actual
## Predicted   0   1
##         0 192 105
##         1  50 145
```

Table 8: Test Set Performance Comparison

| Model | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|
| Logistic Regression | 0.6931 | 0.7162 | 0.6730 | 0.7683 |
| Linear Discriminant Analysis | 0.6931 | 0.7181 | 0.6717 | 0.7687 |
| Classification Tree | 0.6850 | 0.7436 | 0.6465 | 0.7001 |

```
##
## Best performing model by AUC: Linear Discriminant Analysis (AUC = 0.7687)
```




### Test-set performance

On the held-out test set of 492 team-games, all three models achieve similar overall performance:

- **Logistic Regression**:
  - Accuracy: 0.6931
  - Sensitivity: 0.7162
  - Specificity: 0.6730
  - AUC: 0.7683
- **Linear Discriminant Analysis (LDA)**:
  - Accuracy: 0.6931
  - Sensitivity: 0.7181
  - Specificity: 0.6717
  - AUC: 0.7687
- **Classification Tree**:
  - Accuracy: 0.6850

- Sensitivity: 0.7436
- Specificity: 0.6465
- AUC: 0.7001

The ROC curves show that both logistic regression and LDA lie well above the diagonal reference line, with AUC around **0.77**, indicating good ability to discriminate winners from losers. The classification tree performs slightly worse in terms of AUC (0.70), though it still clearly outperforms random guessing.

LDA has the highest AUC on the test set by a small margin, but the difference between LDA and logistic regression is negligible at the reported precision. The tree achieves comparable accuracy but with a lower AUC, suggesting it is somewhat less consistent in ranking wins above losses across all thresholds.

Table 9: 10-Fold Cross-Validation Results (Mean ± SD)

| Model | Mean_Accuracy | SD_Accuracy | Mean_AUC | SD_AUC |
|---|---|---|---|---|
| LDA | 0.7104 | 0.0312 | 0.7899 | 0.0305 |
| Logistic | 0.7088 | 0.0279 | 0.7898 | 0.0303 |
| Tree | 0.6809 | 0.0247 | 0.6909 | 0.0358 |

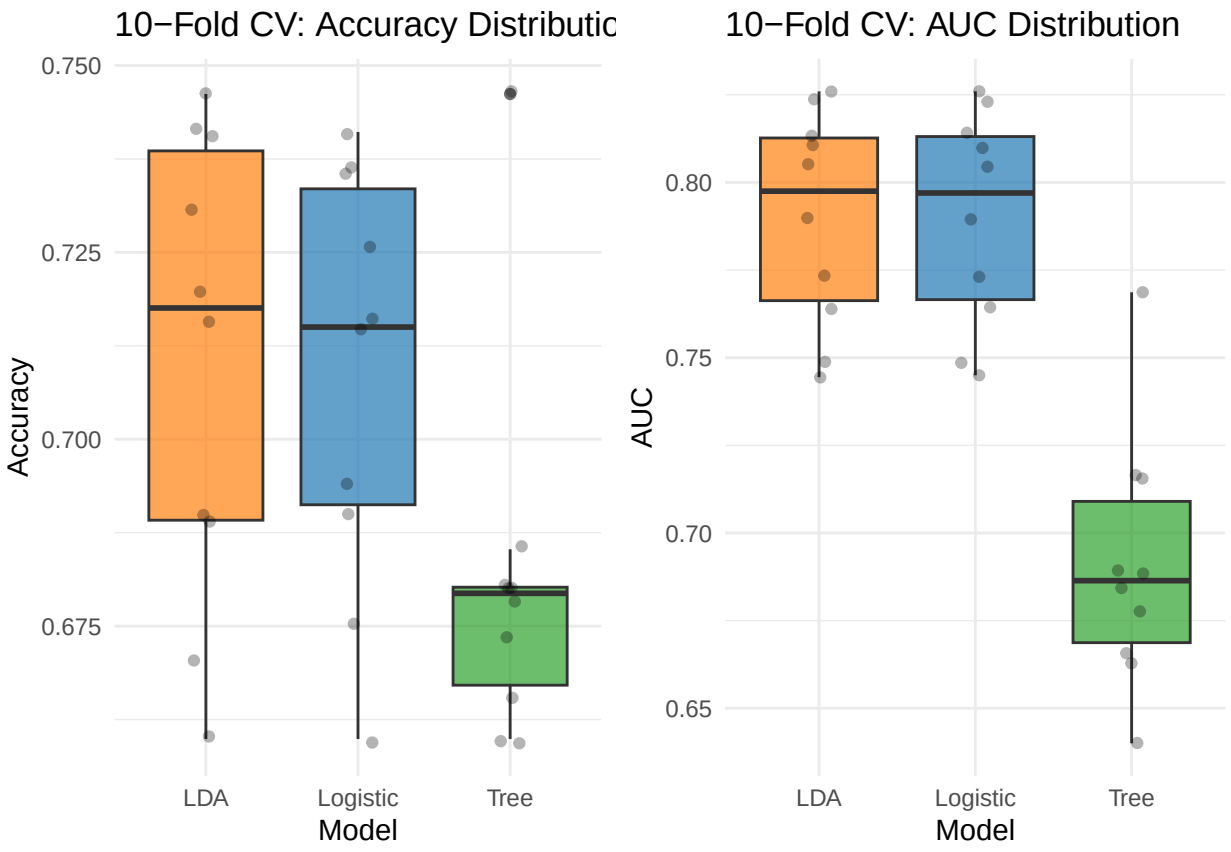


Table 10: Final Model Comparison: Test Set vs. Cross-Validation

| Model | Test_Accuracy | CV_Accuracy | Test_AUC | CV_AUC |
|---|---|---|---|---|
| Logistic Regression | 0.6931 | 0.7104 | 0.7683 | 0.7899 |
| Linear Discriminant Analysis | 0.6931 | 0.7088 | 0.7687 | 0.7898 |
| Classification Tree | 0.6850 | 0.6809 | 0.7001 | 0.6909 |

## Cross-validation and model comparison

The 10-fold cross-validation results on the training data are consistent with the test-set findings:

- **LDA**: Mean accuracy $\approx 0.71$, mean AUC $\approx 0.79$
- **Logistic Regression**: Mean accuracy $\approx 0.71$, mean AUC $\approx 0.79$
- **Classification Tree**: Mean accuracy $\approx 0.68$, mean AUC $\approx 0.69$

The boxplots show relatively low variability in both accuracy and AUC across folds, indicating that model performance is stable and not overly sensitive to how the data are partitioned. The close match between cross-validated AUC (~0.79) and test AUC (~0.77) for logistic regression and LDA suggests that these models are not severely overfitting.

The final comparison table summarizes test and cross-validated performance side by side:

- Logistic Regression and LDA are essentially tied for best performance, with nearly identical accuracy and AUC.
- The Classification Tree slightly underperforms the linear models in AUC but remains competitive in accuracy and offers greater interpretability through its tree structure.

Given these results, LDA and logistic regression are the preferred models for predictive performance, while the classification tree is valuable for communicating simple, rule-based decision logic to non-technical audiences.

## Discussion

### Interpretation of key defensive metrics

Across all three modeling approaches, the same defensive variables emerge as most important:

- **Defensive rebounds (DREB)**: Strongest and most consistent predictor. Teams that win games secure about four more defensive rebounds on average than teams that lose, and each additional rebound increases win odds by roughly 24% in the logistic model.
- **Steals (STL)**: Also positively associated with winning; winners average about one more steal per game than losers. Steals directly create extra possessions and transition opportunities.
- **Blocks (BLK)**: Show a positive but smaller effect, reflecting rim protection but also the relatively low frequency of blocks.
- **Turnovers (TOV)**: Negatively associated with winning; teams that lose commit more turnovers, which is consistent with the idea that lost possessions reduce scoring chances.
- **Personal fouls (PF)**: Have a modest negative effect; excessive fouling can lead to free throws and foul trouble but is less predictive than rebounding or turnovers.

These results provide quantitative support for the coaching intuition that "finishing possessions with a rebound" and "winning the turnover battle" are central to success.

### Model choice and trade-offs

- **Logistic Regression**: Offers strong predictive performance and highly interpretable coefficients. It is well-suited when we want to quantify effect sizes and communicate how changes in specific metrics affect win probability.
- **LDA**: Matches logistic regression in performance and is computationally efficient. It is particularly appealing when its normality and equal-covariance assumptions are roughly satisfied, as appears to be the case here.
- **Classification Tree**: Slightly weaker in AUC but provides intuitive decision rules and clear thresholds (e.g., DREB $\geq 35.5$) that can be directly translated into game objectives.

In practice, a combination of these models could be used: logistic regression or LDA for robust prediction and effect estimation, and a small pruned tree for communicating simple rules to coaches and players.

### Limitations and future work

This analysis has several limitations:

- The models only use defensive statistics. Offensive performance, pace, opponent strength, and game context (home/away, rest days, injuries) are not included and likely explain additional variation in outcomes.
- Each game appears twice in the data (one row per team), which may violate strict independence assumptions. More advanced methods could model game-level dependence explicitly.
- A single season (2023-24) may not capture longer-term trends or changes in playing style.

Future work could:

- Incorporate offensive metrics and advanced efficiency measures (e.g., offensive/defensive ratings).
- Include contextual variables and explore interaction effects (e.g., how defensive strength matters more in close games).
- Extend the comparison to ensemble methods such as random forests and gradient boosting to see whether they meaningfully outperform the simpler models covered here.

## Conclusion

Using game-by-game defensive statistics from the 2023–24 NBA season, we built and compared logistic regression, linear discriminant analysis, and classification trees to predict game outcomes. All three models achieved test-set accuracies around 69% and AUC values between 0.70 and 0.77, substantially better than random guessing, with logistic regression and LDA performing best.

Across methods, defensive rebounds and steals consistently emerged as the most important predictors of winning, while turnovers and personal fouls were negatively associated with success. These findings quantitatively reinforce the importance of controlling the defensive glass and protecting the ball, providing data-driven support for long-standing basketball strategy principles.