NBA Games

Jackson Cramer, Joshua DeRuyter,
Nathaniel Donahue,
Siti Aleeya Nuha Roslee, Yingying Xie

Research Question

What NBA statistics are most valuable for predicting whether a home team wins?

Purpose

- Use machine learning techniques to highlight statistics most correlated to an NBA team winning
- Advise coaches to prioritize practice related to these statistics to maximize their chances of winning
- Make recommendations to NBA analysts for how they should should conduct similar future analyses

DATA

Data Description

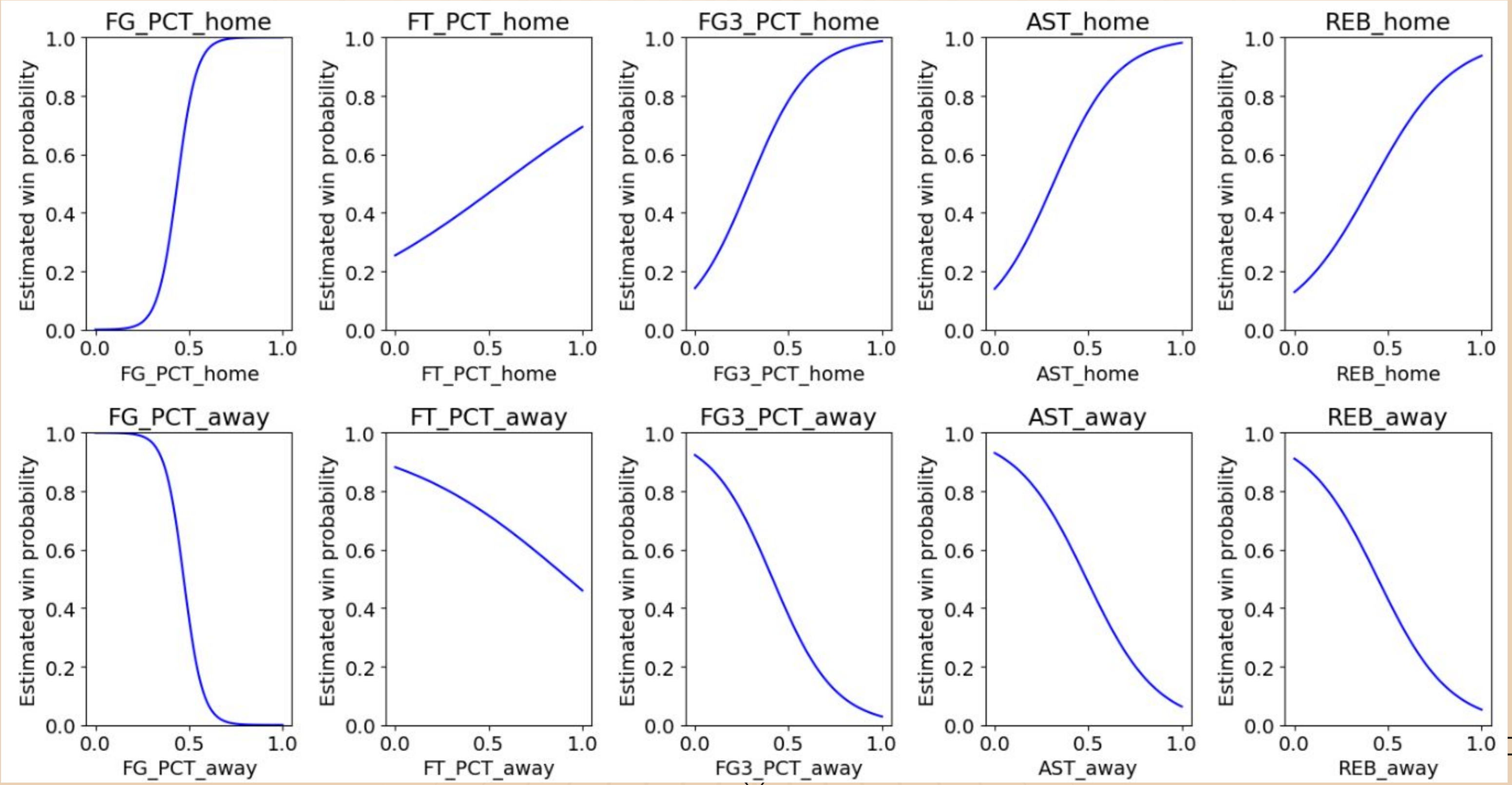
- Acquired via Kaggle, titled “NBA Games Data”
- Each row corresponds to a game, each column represents a statistic related to that game
- Games span from 2003 to 2022, totalling 26,552 NBA games
- Features: Field goal percentage | Free throw percentage | Three point percentage | Total assists | Total rebounds
- Target: Home team wins

	FG_PCT_home	FT_PCT_home	FG3_PCT_home	AST_home	REB_home	FG_PCT_away	FT_PCT_away	FG3_PCT_away	AST_away	REB_away	HOME_TEAM_WINS
0	0.484	0.926	0.382	25.0	46.0	0.478	0.815	0.321	23.0	44.0	1
1	0.488	0.952	0.457	16.0	40.0	0.561	0.765	0.333	20.0	37.0	1
2	0.482	0.786	0.313	22.0	37.0	0.470	0.682	0.433	20.0	46.0	1
3	0.441	0.909	0.297	27.0	49.0	0.392	0.735	0.261	15.0	46.0	1
4	0.429	1.000	0.378	22.0	47.0	0.500	0.773	0.292	20.0	47.0	0

Normalize features not originally spanning from 0-1 with min-max normalization

	FG_PCT_home	FT_PCT_home	FG3_PCT_home	AST_home	REB_home	FG_PCT_away	FT_PCT_away	FG3_PCT_away	AST_away	REB_away	HOME_TEAM_WINS
0	0.484	0.926	0.382	0.431818	0.543860	0.478	0.815	0.321	0.452381	0.403226	1
1	0.488	0.952	0.457	0.227273	0.438596	0.561	0.765	0.333	0.380952	0.290323	1
2	0.482	0.786	0.313	0.363636	0.385965	0.470	0.682	0.433	0.380952	0.435484	1
3	0.441	0.909	0.297	0.477273	0.596491	0.392	0.735	0.261	0.261905	0.435484	1
4	0.429	1.000	0.378	0.363636	0.561404	0.500	0.773	0.292	0.380952	0.451613	0

Exploratory Data Analysis



METHODS

Logistic Regression

- Create a probability spanning from 0-1 for a home team winning
- Why Logistic Regression?

  -Binary Target: great for handling binary outcomes like "win" (1) or "loss" (0).

  -Probability Output: provides probabilities for outcome, helping assess win likelihood given game statistics

  -Requires Minimal Assumptions: doesn’t require features to be normally distributed or have a linear relationship with target

Hyperparameter Tuning

- Minimize the cost function

$$\|\mathbf{w}\| + C\left[-\sum_{i=1}^{N}\left(y_i \ln f_{\mathbf{w},b}(\mathbf{x}_i) + (1-y_i)\ln\left[1 - f_{\mathbf{w},b}(\mathbf{x}_i)\right]\right)\right]$$

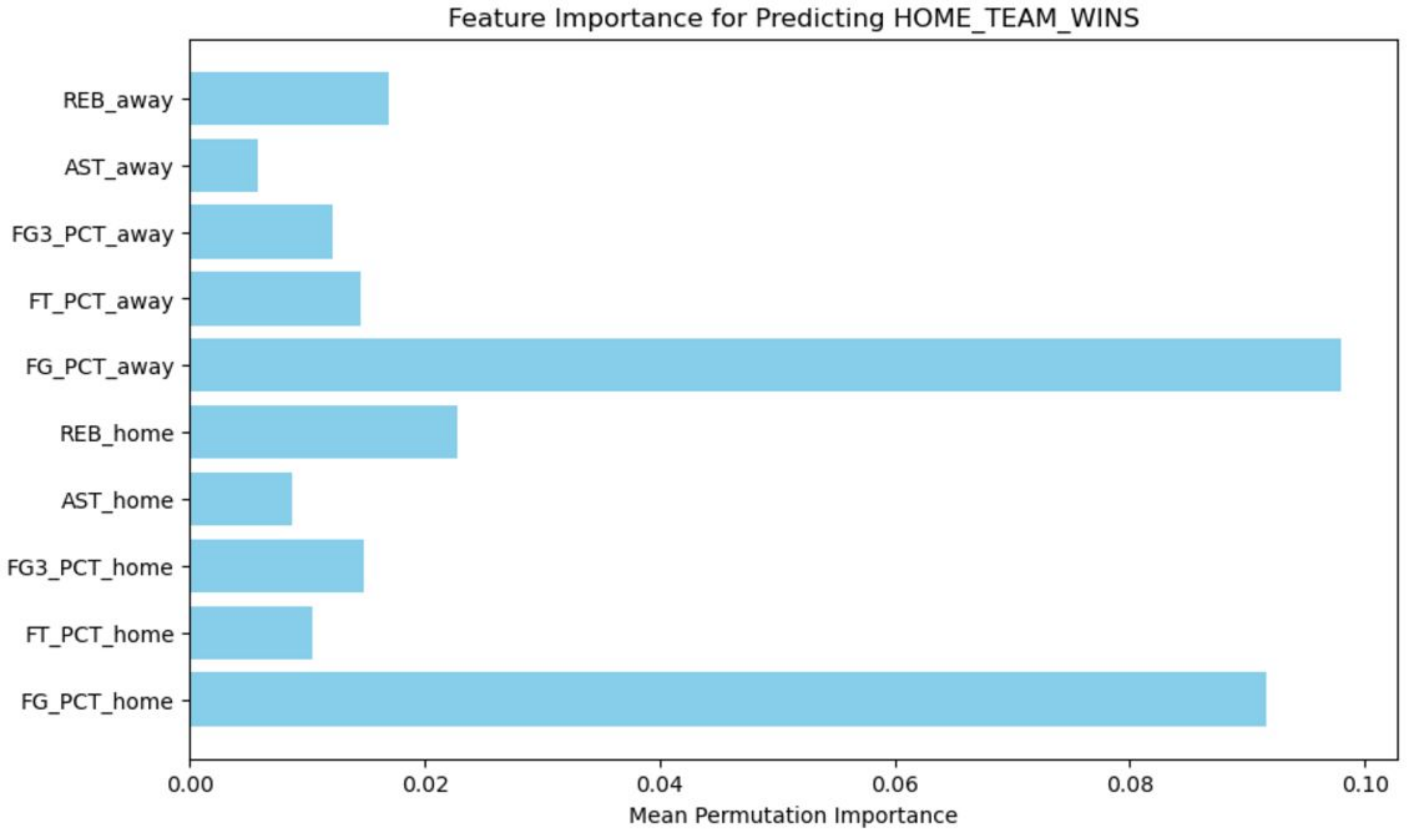
- ↑ C : emphasizes fitting the data
- ↓ C : prevents overfitting
- Use grid search to find “best C”

  -Result: C=20
- Using C=20 and lasso regression, all features were included

Feature Selection

- Lasso regression: eliminate features less important for predicting whether a home team wins
- Started from “best C” (C=20)
- Lowered C until at least one feature was removed (C = 0.01)
- FT_PCT_home and FT_PCT_away removed
- Reasoning: limited point value, strategic fouling

Permutation Feature Importance



Scoring

- Data not imbalanced: HOME_TEAM_WINS: 1 - 58.9%, 0 - 41.1%

	Model 1: Logistic Regression	Model 2: Lasso Regression
Parameter	C = 20	C = 0.01
Feature Selection	None	No Free Throw Percentage for Home & Away
Mean Squared Error	0.16	0.18
Accuracy	0.84	0.81
Precision	0.86	0.81

CONCLUSION

Coach Recommendations

- Practice game-like shooting situations as a team (high priority)
- Practice rebounds and assists in small group setting (medium priority)
- Practice free throws individually (low priority)

Analyst Recommendations

- **Parameter selection**:
  -a higher C may give better predictions, but risks overfitting
  -a lower C can achieve feature selection, which may save computational resources
- **Accuracy is not always the goal of modeling**
  -we used modeling as a means of exploring which statistics a coach should prioritize, not accurately predicting a win
- “All models are wrong, some are useful.” - George Box

Thank you!