NBA Advanced Metrics Project

Predicting Home-Win Outcomes Using Advanced Statistics

Problem Statement & Data Preparation

DATA SOURCE & RAW STRUCTURE

NBA Stats API via hoopR (57k+ rows)

Raw team-level box scores (one row per team per game)

Needed to convert into game-level dataset

CLEANING & FEATURE CREATION

Paired home/away rows

Computed advanced stats (ORTG, DRTG, eFG%, TOV%, rebound%)

Lagged data to avoid leakage

MODELING & RESULTS

Logistic regression (inferential)

Penalized logistic (predictive)

Accuracy: 0.64–0.67 (baseline: 0.553)

Net rating differential strongest signal

ROLLING METRICS

Rolling 3-, 5-, 10-game averages

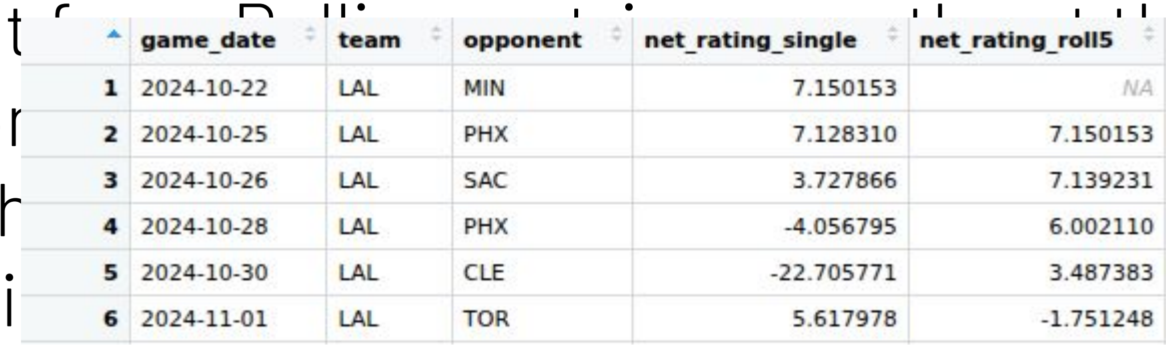
Captures recent performance trends

Removes game-to-game noise

ELO rating to capture longer time frames

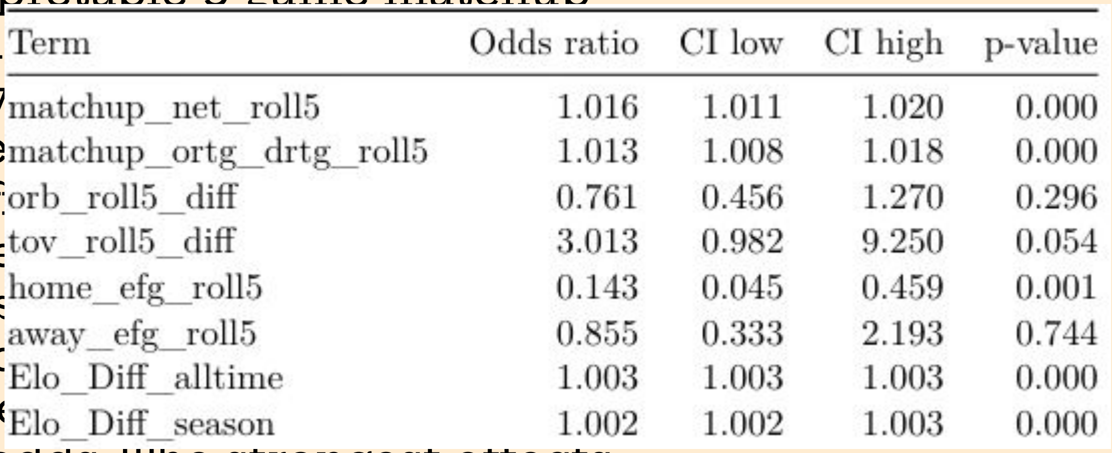
Advanced Stats & Rolling Feature Engineering

Creating the rolling averages to capture a team's
recent
randor and
give th w a
team i es are
what ultimately drive most of the predictive
power in my models.



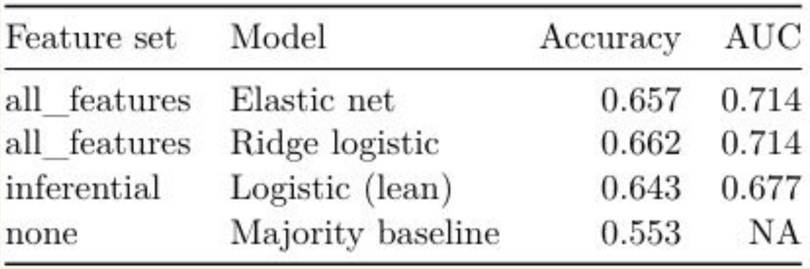
	game_date	team	opponent	net_rating_single	net_rating_roll5
1	2024-10-22	LAL	MIN	7.150153	NA
2	2024-10-25	LAL	PHX	7.128310	7.150153
3	2024-10-26	LAL	SAC	3.727866	7.139231
4	2024-10-28	LAL	PHX	-4.056795	6.002110
5	2024-10-30	LAL	CLE	-22.705771	3.487383
6	2024-11-01	LAL	TOR	5.617978	-1.751248

To understand which rolling metrics actually matter, I fit a lean logistic regression model using only a small set of interpretable 5-game matchup
differential
efficiency v
turnover pe
shooting ef
features are
prior games
leakage and
how much e
home-win odds. The strongest effects come from rolling net rating and rolling offensive-versus-defensive efficiency gaps, while turnover and rebounding differentials show smaller but directionally reasonable impacts.



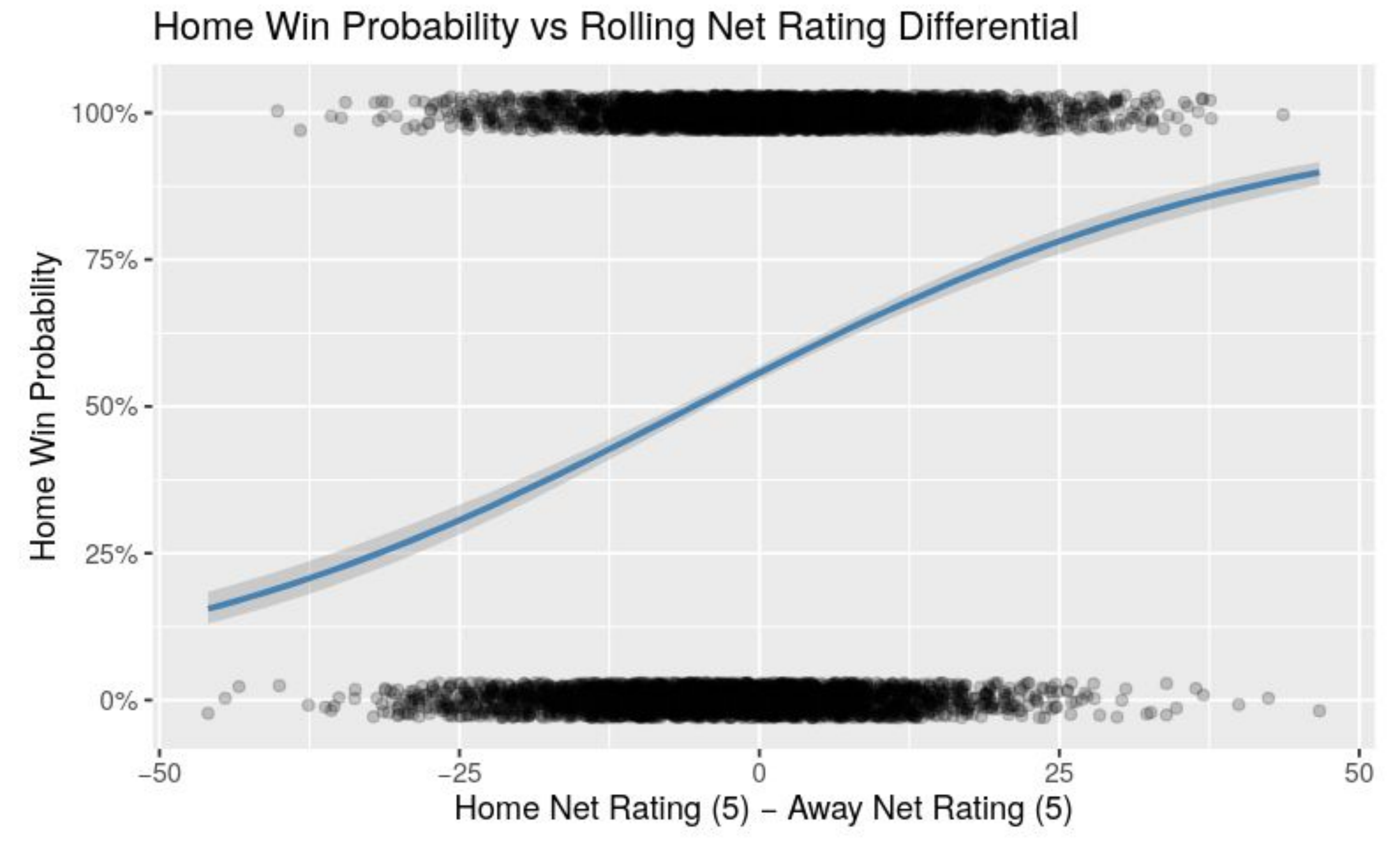
Term	Odds ratio	CI low	CI high	p-value
matchup_net_roll5	1.016	1.011	1.020	0.000
matchup_ortg_drtg_roll5	1.013	1.008	1.018	0.000
orb_roll5_diff	0.761	0.456	1.270	0.296
tov_roll5_diff	3.013	0.982	9.250	0.054
home_efg_roll5	0.143	0.045	0.459	0.001
away_efg_roll5	0.855	0.333	2.193	0.744
Elo_Diff_alltime	1.003	1.003	1.003	0.000
Elo_Diff_season	1.002	1.002	1.003	0.000

To evaluate predictive power, I trained several pre-game models on all seasons prior to 2025 and tested them on the 2025 holdout season. Ridge logistic and elastic net used the full rolling feature set, including shooting efficiency, net rating, rebounding, turnover rates, and Elo differentials. These models achieved holdout accuracies between 0.64 and 0.67, outperforming the majority baseline of 0.553. The lean inferential logistic regression reached 0.643 accuracy, showing that even a small feature subset performs competitively.



Feature set	Model	Accuracy	AUC
all_features	Elastic net	0.657	0.714
all_features	Ridge logistic	0.662	0.714
inferential	Logistic (lean)	0.643	0.677
none	Majority baseline	0.553	NA

Key Visual Insights



QUESTIONS
?

THANK YOU