Environment Note

reportlab package is not installed in this environment.
PDF generated with matplotlib PdfPages fallback (same required content included).

Executive Summary

1. Executive summary
- Objective metric: RMSE on margin prediction (primary). MAE is secondary.
- Prior stated RMSE baseline: 279.39
- Best time-aware CV stage RMSE: 38.286
- Improvement vs prior baseline: 241.104
- Final derby output uses nonlinear mapping + constrained ensemble + calibration + uncertainty-aware shrinkage.

2. Data constraints (home vs neutral, no derby labels)
- Train games: 940 (2025-01-01 to 2025-06-30)
- Derby games: 75 on 2025-07-04
- No derby labels available. Validation uses chronological expanding folds of Train.csv only.
- Derby predictions are neutral-site (home advantage set to 0 in derby feature generation).

3. Feature engineering
- Pregame-only sequential Elo, Massey, off/def net, strength of schedule, EMA margin, rest/volatility retained.
- Elo HOME_ADV tuned via time-aware CV: 40 Elo points.
- Feature counts: linear=69, nonlinear=153.
- Nonlinear mapping implemented with polynomial + piecewise hinge/bin transforms of rating differentials.

Validation Design

Chronological folds:

```
 fold=1 train_n=125 val_n=193 train_end=2025-01-12 val=2025-01-23..2025-02-19
 fold=2 train_n=318 val_n=205 train_end=2025-02-19 val=2025-03-03..2025-04-02
 fold=3 train_n=523 val_n=151 train_end=2025-04-02 val=2025-04-14..2025-04-30
 fold=4 train_n=674 val_n=102 train_end=2025-04-30 val=2025-05-13..2025-05-29
 fold=5 train_n=776 val_n=164 train_end=2025-05-29 val=2025-05-30..2025-06-30
```

HOME_ADV tuning summary:

home_adv	brier_mean	brier_std	logloss_mean
40	0.207073	0.016372	0.602385
30	0.207074	0.016117	0.602619
50	0.207499	0.016722	0.602982
20	0.207513	0.015942	0.603697
60	0.208349	0.017143	0.604418
10	0.208391	0.015857	0.605621
70	0.209607	0.017630	0.606675
0	0.209715	0.015849	0.608400
80	0.211258	0.018179	0.609746
90	0.213288	0.018777	0.613619
100	0.215681	0.019418	0.618285
110	0.218422	0.020094	0.623747
120	0.221485	0.020797	0.629974

4. Model comparison table (all configs, CV metrics) (1/1)

config_name	model_name	mapping	rmse_mean	mae_mean	bias_mean	rmse_std	resid_skew_oof	resid_kurtosis_oof	outlier_freq_2p5sd_oof
histgb__nonlinear	histgb	nonlinear	38.198	30.303	-0.278	2.602	0.146	0.107	0.017
histgb__linear	histgb	linear	38.343	30.496	0.099	3.089	0.143	0.056	0.013
ridge__linear	ridge	linear	39.694	31.539	4.947	8.715	-0.072	0.501	0.018
elasticnet__linear	elasticnet	linear	39.951	31.708	5.173	9.699	-0.131	0.623	0.020
huber__linear	huber	linear	42.298	33.747	5.113	12.592	0.155	1.339	0.022
ridge__nonlinear	ridge	nonlinear	42.647	33.684	7.176	13.627	-0.287	1.786	0.025
elasticnet__nonline	elasticnet	nonlinear	42.948	33.850	7.434	15.047	-0.415	2.184	0.023
huber__nonlinear	huber	nonlinear	69.747	43.733	8.482	70.600	13.591	300.287	0.007

Modeling Highlights

4. Model comparison summary

Selection rule: lowest CV RMSE first, MAE second.

Nonlinear mapping comparison:

mapping	best_config	rmse_mean	mae_mean	bias_mean	n_features	rmse_delta_vs_linear	nonlinear_improves
linear	histgb__linear	38.343132	30.495563	0.098565	69	0.00000	True
nonlinear	histgb__nonlinear	38.197842	30.302522	-0.277977	153	-0.14529	True

7. Ensemble weighting results (linear vs nonlinear)

variant	rmse_progressive	mae_progressive	rmse_global_weighted_oof	mae_global_weighted_oof	w_ridge	w_elasticnet	w_huber	w_histgb
linear	41.428035	32.613651	38.654435	30.588890	0.254238	1.918637e-12	0.0	0.745762
nonlinear	50.825831	35.273374	38.630140	30.574031	0.111202	0.000000e+00	0.0	0.888798

Final nonlinear global ensemble weights (simplex constrained):

Ridge=0.1112

ElasticNet=0.0000

Huber=0.0000

HistGB=0.8888

7. Ensemble weighting results - nonlinear progressive fold weights (1/1)

fold	source	w_ridge	w_elasticnet	w_huber	w_histgb
1	uniform_warm_start	0.250	0.250	0.250	0.250
2	prior_folds	0.000	0.000	0.000	1.000
3	prior_folds	0.039	0.000	0.000	0.961
4	prior_folds	0.111	0.000	0.000	0.889
5	prior_folds	0.111	0.000	0.000	0.889

Calibration and Variance

5. Calibration analysis
- Fold calibration regression: actual = a + b * predicted (progressive prior-fold fit).
- Global calibration for final derby: a=1.0559, b=0.7717
- Average fold slope b=0.7892, average intercept a=0.9759
- Interpretation: average slope implies raw ensemble is overconfident.
- Calibration RMSE effect: raw=50.826 -> calibrated=38.313

6. Variance modeling
- squared_residual ~ |rating_diff| (using |elo_diff_pre|; nonnegative slope enforced).
- Global variance model: var = 1610.0192 + 0.000000 * |rating_diff|
- Shrink lambda selected by OOF RMSE of simulated trimmed mean: 0.100
- Variance reference scale for shrinking: 1703.319

8. Uncertainty-aware prediction
- Per-fold residual SD estimated from variance model.
- 2000 Normal draws per game, 5% trimmed mean candidate prediction.
- Compared raw, calibrated, variance-shrunk, simulated-trimmed stages via OOF RMSE.

5. Calibration fold slopes/intercepts (1/1)

fold	intercept_a	slope_b	source
1	0.000	1.000	identity_warm_start
2	-0.554	0.785	prior_folds
3	2.220	0.739	prior_folds
4	1.777	0.709	prior_folds
5	1.437	0.713	prior_folds

6. Variance model fold parameters (1/1)

fold	var_intercept	var_slope	source
1	1467.871	0.000	constant_warm_start
2	1813.067	0.000	prior_folds
3	1767.522	0.000	prior_folds
4	1703.319	0.000	prior_folds
5	1683.939	0.000	prior_folds

8. Shrink lambda sweep and simulation RMSE (1/2)

lambda_shrink	rmse_shrunk	mae_shrunk	rmse_sim_trimmed	mae_sim_trimmed	shrink_mean	shrink_min
0.100	38.315	30.239	38.282	30.211	0.910	0.904
0.000	38.313	30.278	38.307	30.267	1.000	1.000
0.050	38.301	30.250	38.324	30.248	0.953	0.949
0.150	38.349	30.245	38.338	30.227	0.871	0.862
0.200	38.398	30.266	38.408	30.259	0.835	0.824
0.250	38.458	30.300	38.480	30.327	0.802	0.790
0.300	38.525	30.341	38.497	30.317	0.772	0.758
0.350	38.599	30.388	38.611	30.409	0.743	0.729
0.400	38.676	30.437	38.698	30.440	0.717	0.701
0.450	38.756	30.490	38.769	30.478	0.693	0.676
0.500	38.838	30.548	38.882	30.567	0.670	0.653
0.550	38.921	30.611	38.988	30.680	0.648	0.631
0.600	39.003	30.671	39.008	30.657	0.628	0.610
0.650	39.085	30.735	39.129	30.771	0.609	0.591
0.700	39.167	30.800	39.163	30.785	0.592	0.573
0.750	39.247	30.866	39.207	30.842	0.575	0.556
0.800	39.326	30.934	39.303	30.901	0.559	0.540
0.850	39.404	31.002	39.429	31.023	0.544	0.525
0.900	39.480	31.069	39.538	31.128	0.530	0.511
0.950	39.554	31.135	39.558	31.112	0.517	0.497
1.050	39.698	31.266	39.666	31.217	0.492	0.472
1.000	39.627	31.201	39.689	31.261	0.504	0.484
1.100	39.768	31.330	39.724	31.303	0.480	0.461
1.150	39.835	31.391	39.866	31.429	0.469	0.450

8. Shrink lambda sweep and simulation RMSE (2/2)

lambda_shrink	rmse_shrunk	mae_shrunk	rmse_sim_trimmed	mae_sim_trimmed	shrink_mean	shrink_min
1.200	39.882	31.433	39.908	31.433	0.462	0.450
1.350	39.934	31.479	39.917	31.482	0.453	0.450
1.400	39.946	31.489	39.946	31.478	0.451	0.450
1.250	39.912	31.459	39.953	31.472	0.457	0.450
1.300	39.923	31.469	39.962	31.481	0.455	0.450
1.500	39.950	31.493	39.966	31.492	0.450	0.450
1.450	39.950	31.493	39.999	31.526	0.450	0.450

Stage comparison (OOF) (1/1)

stage	rmse	mae	bias	resid_skew	resid_kurtosis	outlier_freq_2p5sd
sim_trimmed	38.286	30.231	-0.481	0.083	0.123	0.016
calibrated_progressive	38.313	30.278	-0.045	0.095	0.133	0.013
variance_shrunk	38.315	30.239	-0.512	0.078	0.113	0.013
nsemble_raw_global_weigh	38.630	30.574	0.223	0.132	0.134	0.016
ensemble_raw_progressiv	50.826	35.273	3.905	2.526	34.960	0.015

Robust Regression

6. Robust regression / heavy tail control

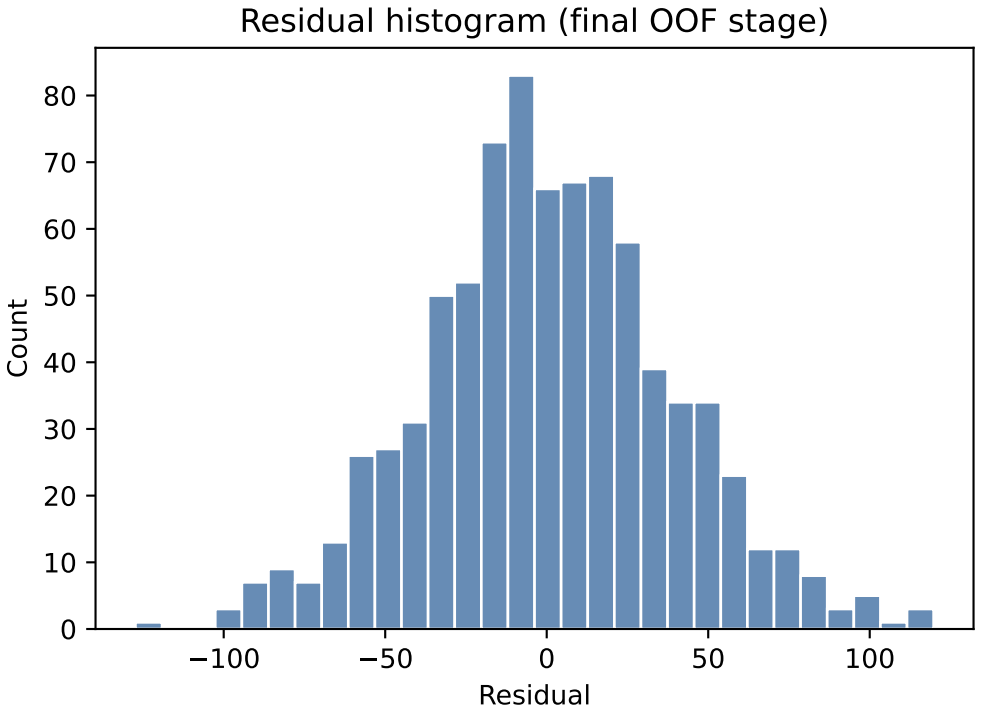
- Explicitly compared Ridge (L2) and Huber (robust loss) on nonlinear feature mapping.
- Residual skewness, kurtosis, and outlier frequency reported below.

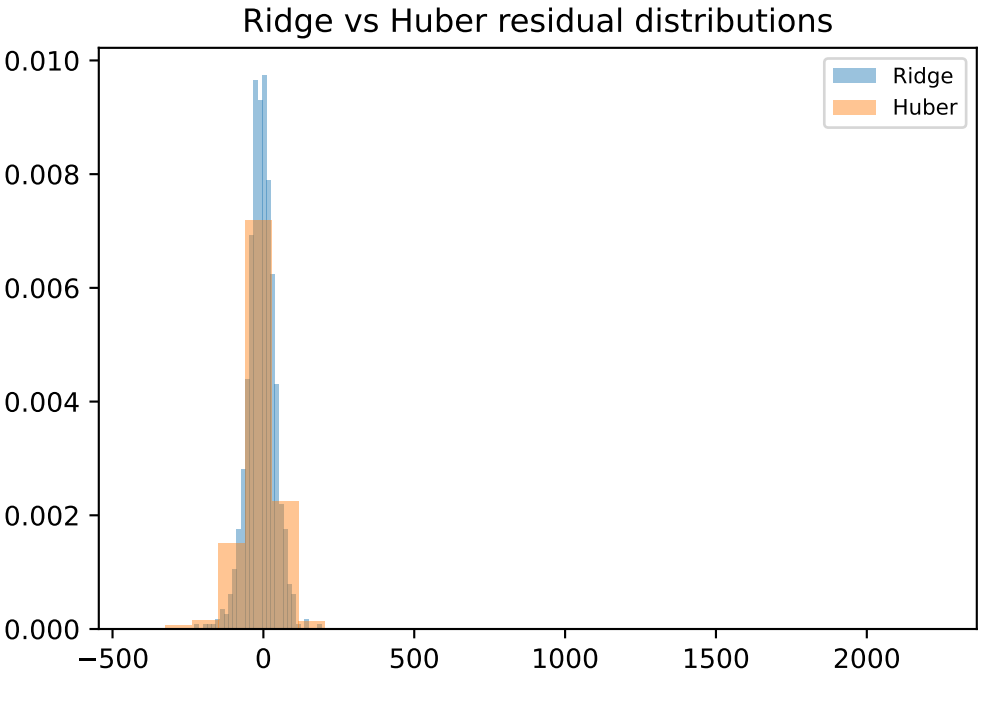
config_name	rmse_mean	mae_mean	bias_mean	resid_skew_oof	resid_kurtosis_oof	outlier_freq_2p5sd_oof
ridge__nonlinear	42.647210	33.683618	7.175799	-0.286608	1.785807	0.024540
huber__nonlinear	69.746838	43.733121	8.481543	13.590929	300.286800	0.007362

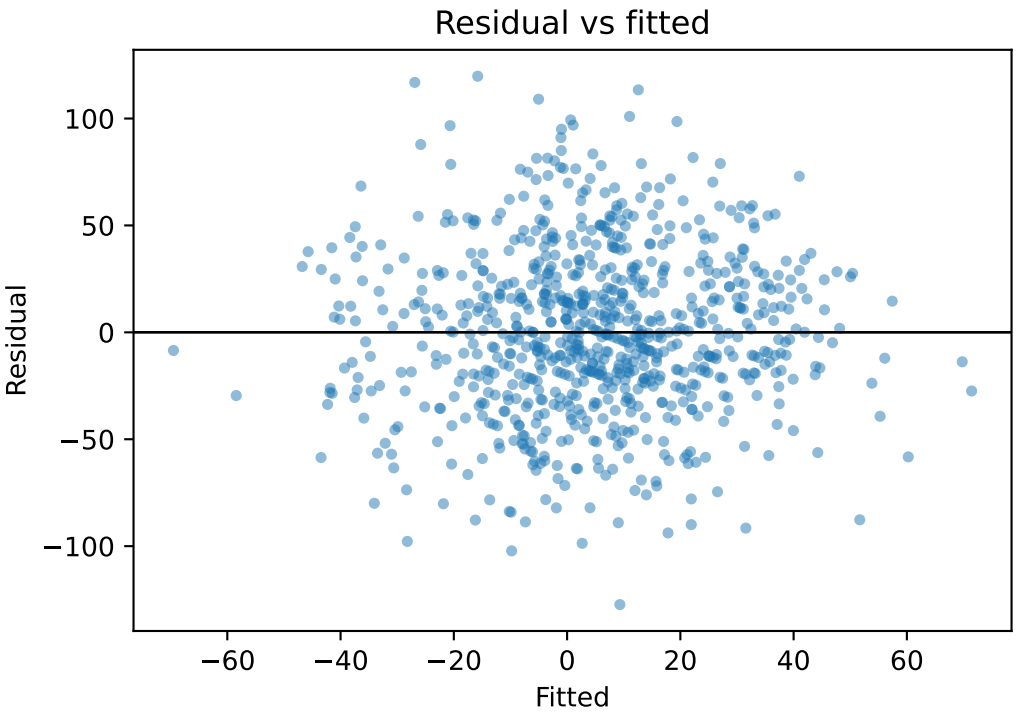
11. Key decisions & rationale

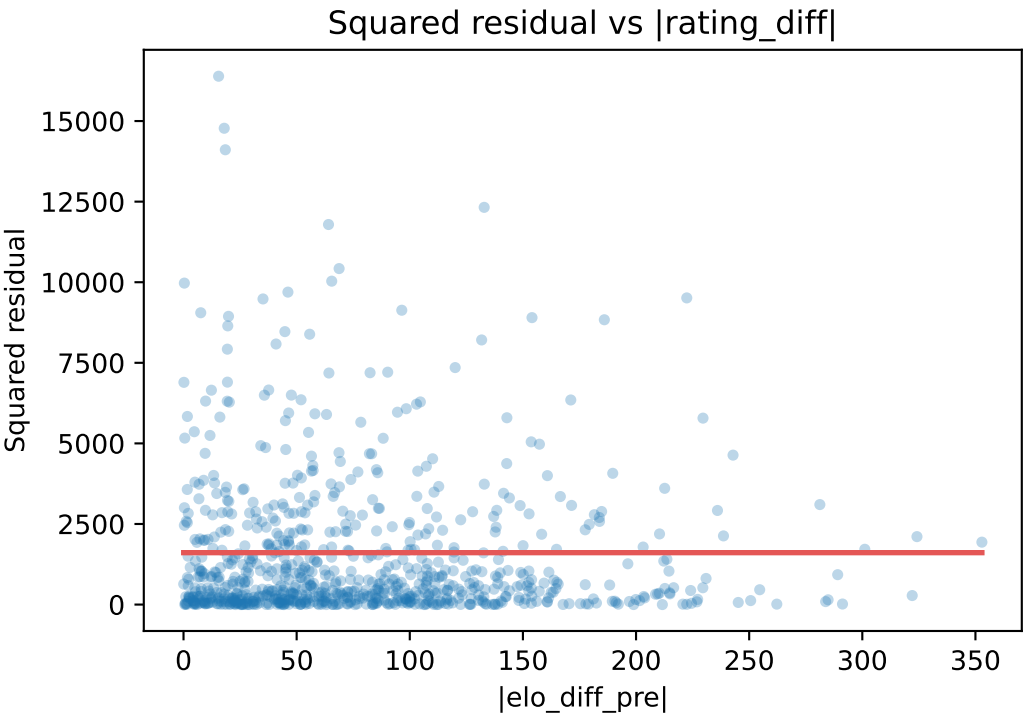
- RMSE is the governing selection criterion for configs and downstream stack comparisons.
- Simplex-constrained weights enforce stable convex blending across model families.
- Calibration and variance-aware shrinkage correct scale/bias and reduce extreme-margin risk.

8. Diagnostic plots - residuals and variance

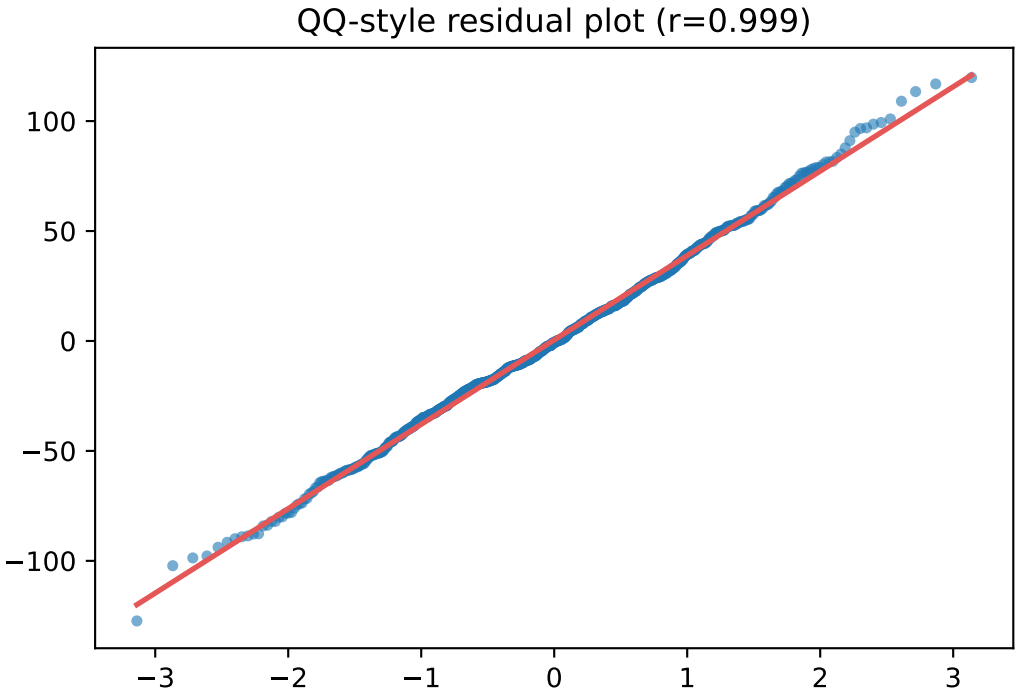


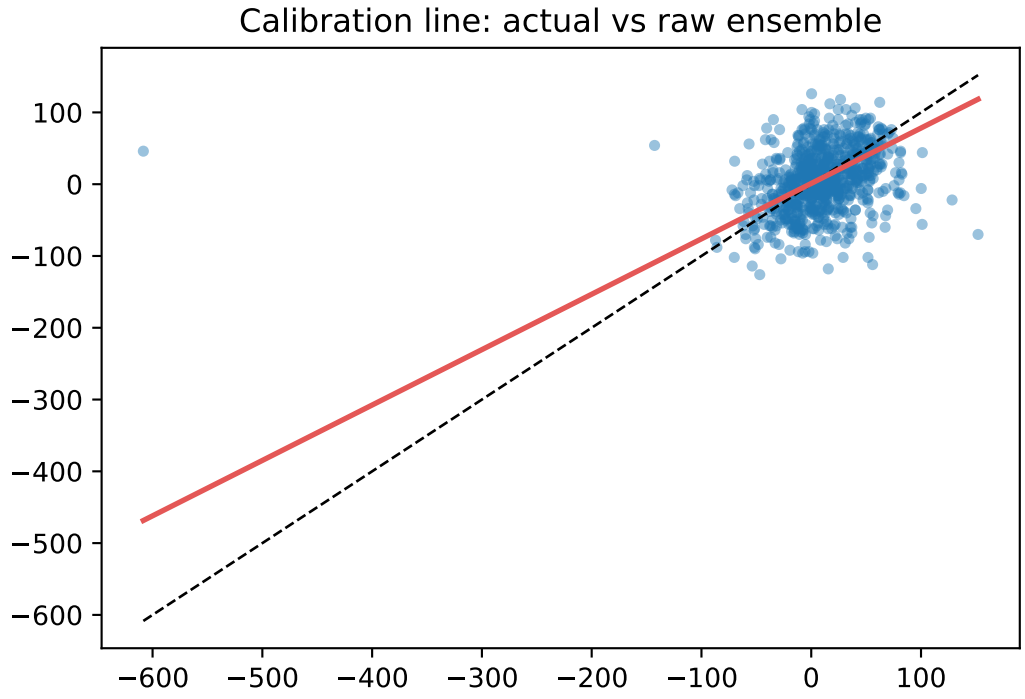


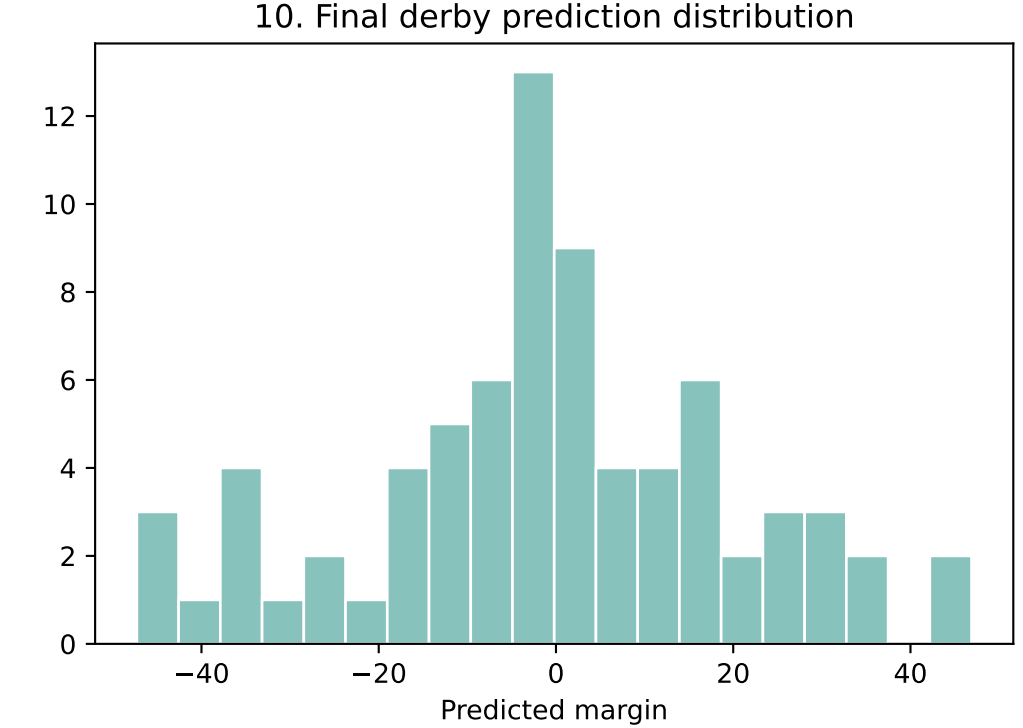


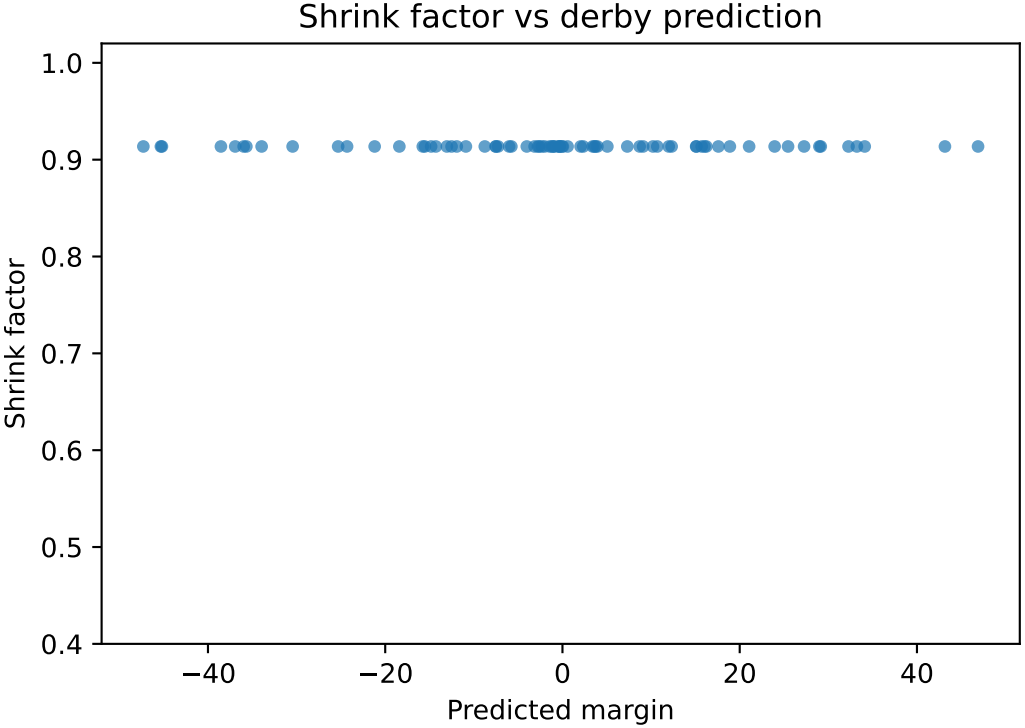


8. Diagnostic plots - QQ, calibration, derby distribution

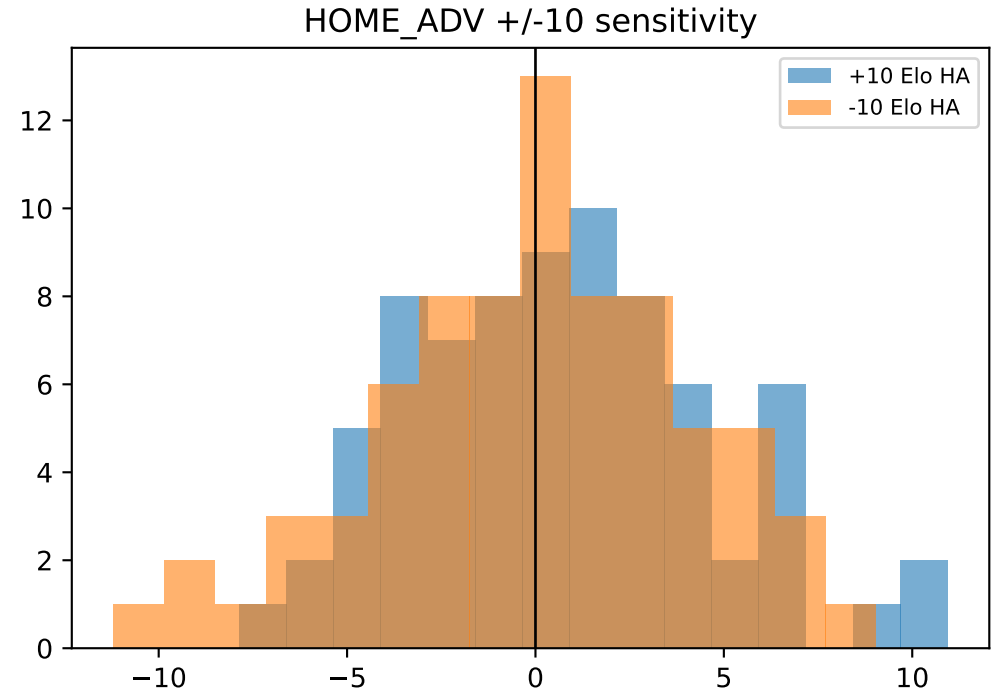


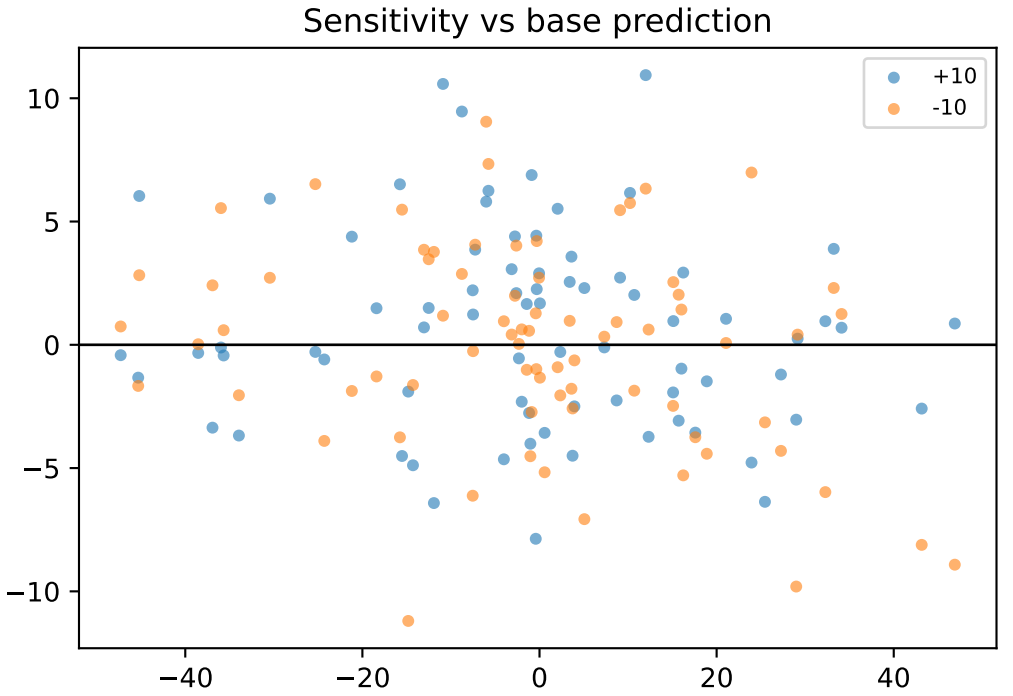


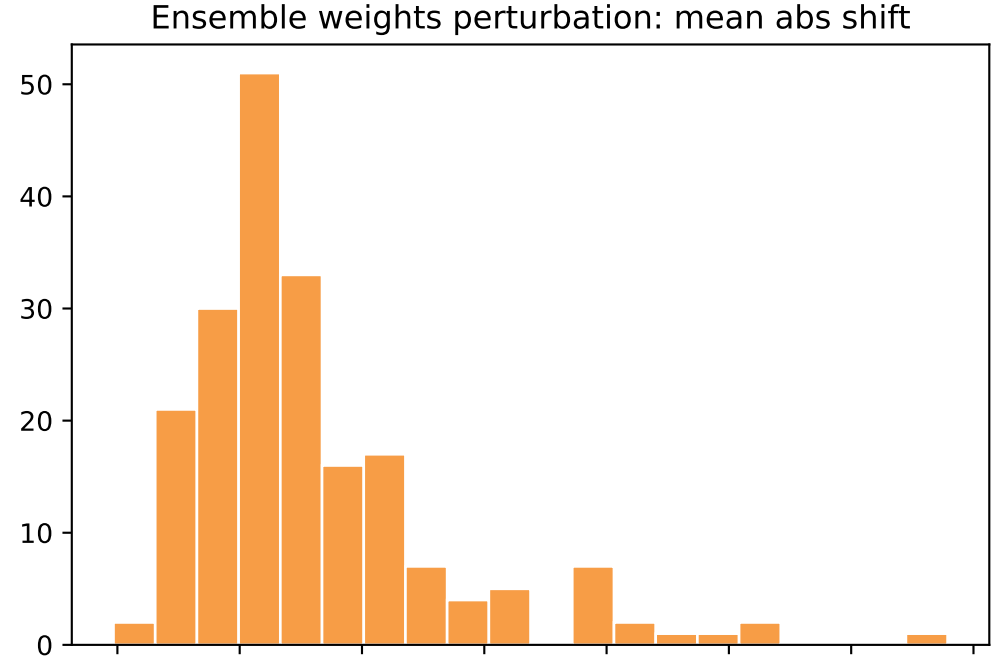


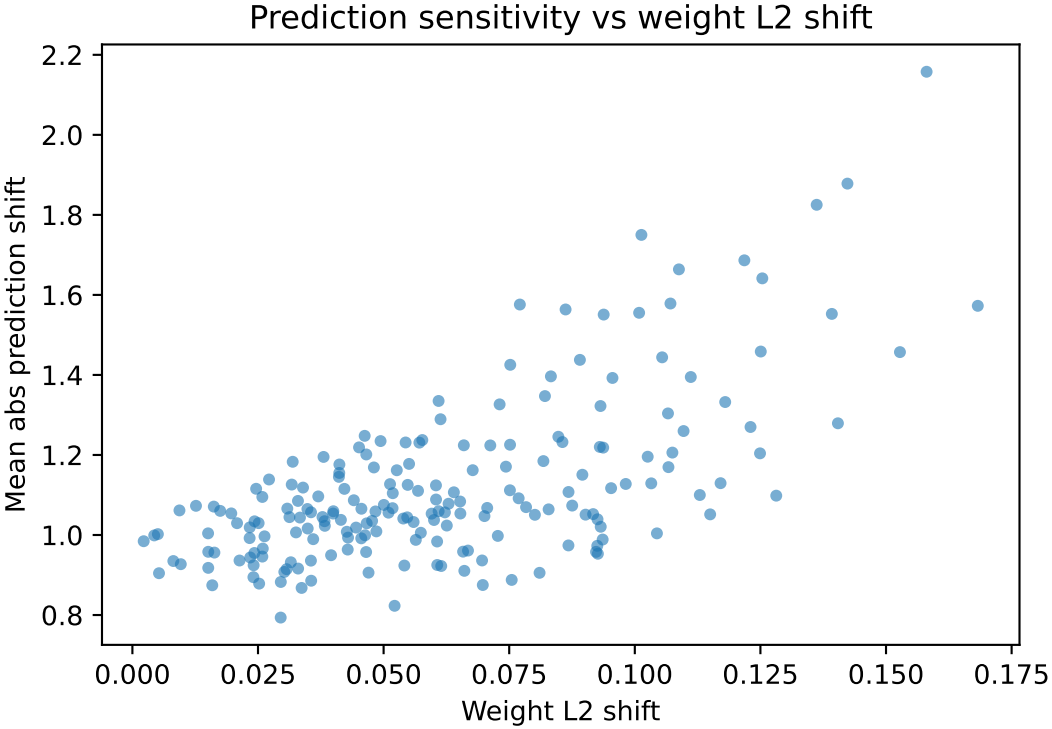


9. Sensitivity analyses









Final Decisions And Limitations

10. Final derby prediction distribution
- Count=75, mean=-0.773, std=20.570, min=-47, max=47
- Quantiles 1/5/50/95/99: -45.520, -37.600, 0.000, 32.300, 44.040

11. Key decisions & rationale
- Final derby stack keeps all required layers: nonlinear transform, ensemble weighting, calibration, uncertainty shrinkage/simulation.
- Winsorization uses Train margin 1st/99th percentiles before rounding.

12. Limitations and future improvements
- Derby labels unavailable; all tuning based on Train-only time-aware CV proxies.
- Global deployment calibration/weights are fit on pooled OOF predictions; nested tuning could further reduce optimism.
- No injuries/lineups/travel features included.
- QuantileRegressor comparison omitted to keep deterministic runtime bounded.

Rankings construction
- Ranking weights (Elo/Massey/Net): {'elo': 0.5090354682209606, 'massey': 0.4527681109252348, 'net': 0.03819642085380458}
- Component OOF ridge weighting proxy: {'elo': 0.5090354682209606, 'massey': 0.4527681109252348, 'net': 0.03819642085380458}

Appendix

Appendix

First 10 rows of predictions.csv:

GameID	Date	Team1_Conf	Team1_ID	Team1	Team2_Conf	Team2_ID	Team2	Team1_WinMargin
1941	7/4/2025	Purple	68	Idaho	Yellow	119	Houston	0
1942	7/4/2025	White	105	Helena	Yellow	120	Iowa	-3
1943	7/4/2025	Red	82	Jackson	Purple	67	Concord	-3
1944	7/4/2025	White	108	Montpelier	Yellow	129	West Virginia	-6
1945	7/4/2025	Yellow	121	Lansing	White	113	South Dakota	-15
1946	7/4/2025	Crimson	21	Lincoln	Yellow	118	Florida	-1
1947	7/4/2025	Yellow	117	Columbia	Crimson	23	Nashville	-8
1948	7/4/2025	Orange	63	New Jersey	Yellow	126	Pierre	10
1949	7/4/2025	Purple	77	Wilmington	Orange	58	Honolulu	-2
1950	7/4/2025	Crimson	22	Michigan	Green	49	Natchez	-1

Top 20 teams in rankings (sorted by Rank):

TeamID	Team	Rank
139	Boise City	1
94	Kentucky	2
140	Pennsylvania	3
20	Phoenix	4
38	Massachusetts	5
139	Tokyo	6
2	Baton Rouge	7
90	Big Sur	8
37	Minnesota	9
137	Milan	10
92	Illinois	11
88	Port St. Lucie	12
116	California	13
149	Seoul	14
46	Boston	15
54	Boise	16
19	Georgia	17
131	Belgrade	18
84	Mississippi	19
22	Michigan	20