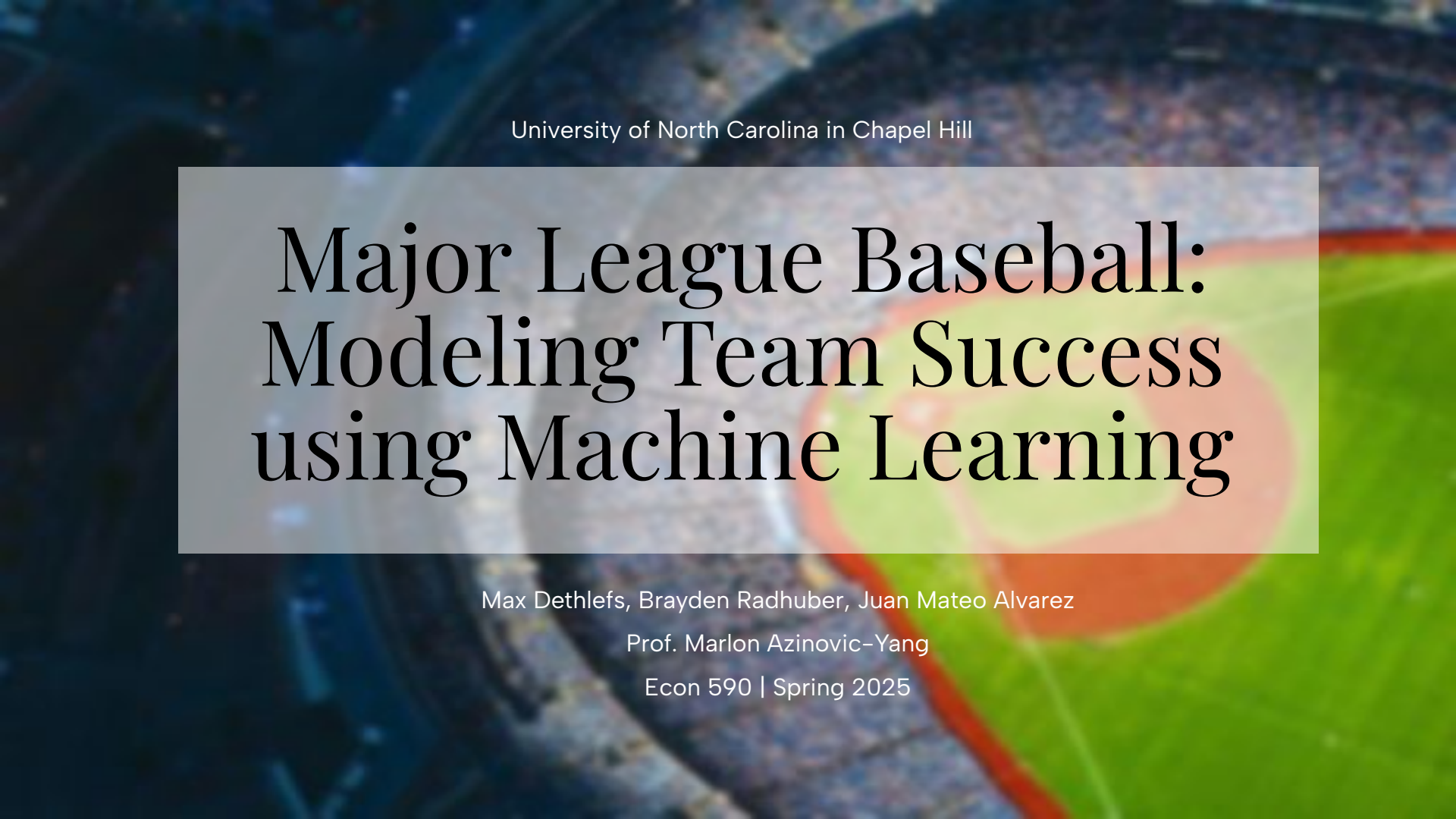


University of North Carolina in Chapel Hill

Major League Baseball: Modeling Team Success using Machine Learning

Max Dethlefs, Brayden Radhuber, Juan Mateo Alvarez

Prof. Marlon Azinovic-Yang

Econ 590 | Spring 2025

Presentation Roadmap

1

Research Outline and Data Exploration

2

Model Design and Comparison

3

Results + Model Predictions

Research Topic

1. Research Question

Which team-level statistics are most predictive of MLB team success? How can these be used to predict future team success?

2. Research Relevance

- Analytics departments of MLB teams would be interested in which player statistics are most likely to drive team success. (Ex. Valuing slugging percentage over batting average.)
- Find disparities between model and sportsbook predictions of team wins, place bets on largest difference for highest value.

Data Collection and Overview

Data Collection and Overview

- Source: Data was retrieved from **Baseball Reference** using **StatHead**, a comprehensive database of baseball history including every player, team, season, league, award, record and score dating back to 1871.

- Scope: Our dataset was a join between batting and pitching statistics of each individual season for all 30 teams between 1974-2024.

- Variables: In total there were 58 variables which represented a team level statistic for each given year.

- Note: All statistics were calculated by total team performance, not by individual player.

Team Batting Season Stats Finder - Baseball

Find individual seasons, combined seasons, or number of seasons matching your criteria. • Video Tutorial • Sample Searches • Data Coverage

Current Search

For single seasons, from 1994 to 2024, in the regular season, sorted by descending Wins.

≡ Show Criteria ➦ Share Results

Query Results

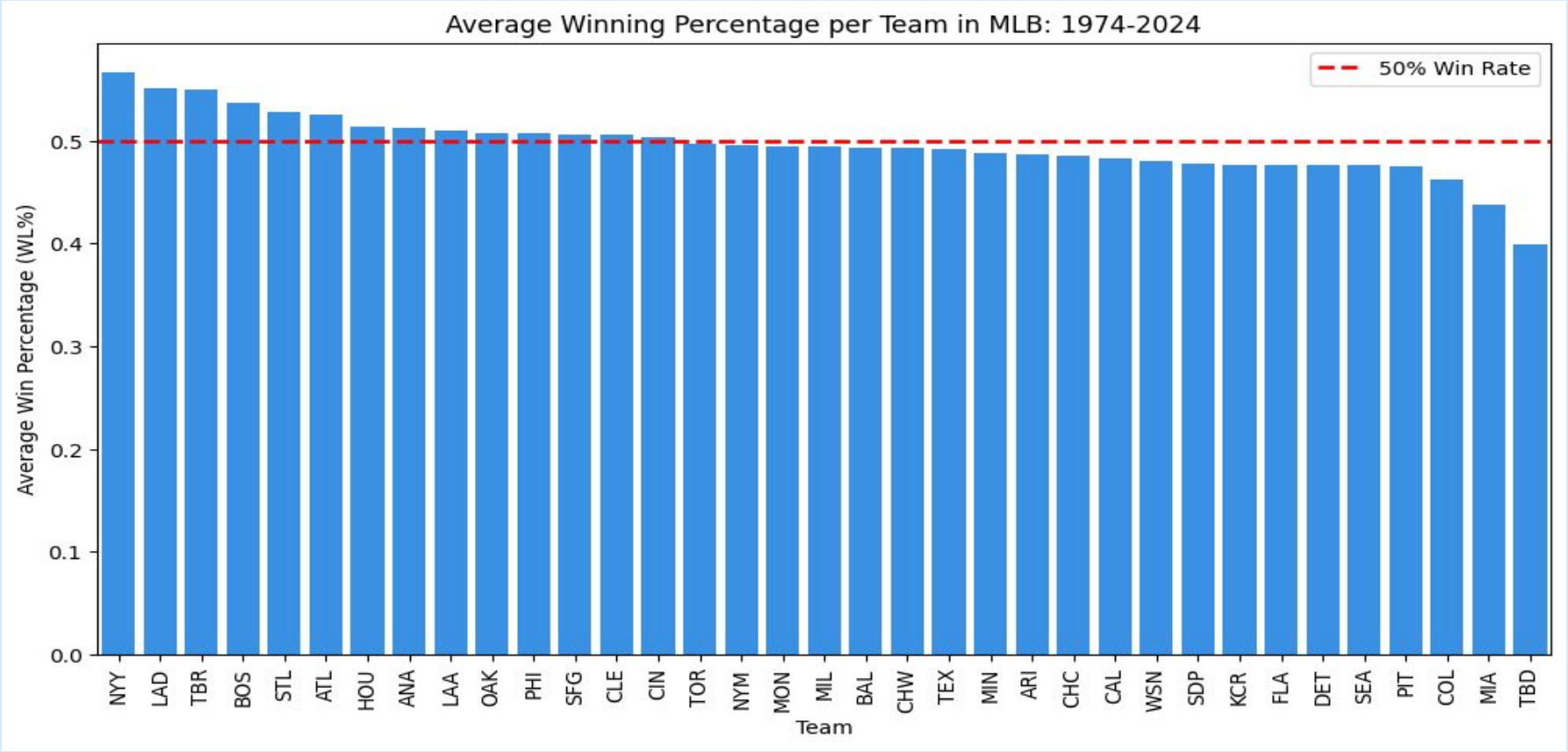
Export Data ▾ Glossary

Rk	Season	Team	Lg	W	GP	W	L	WL%	Bat#	PA	AB	R	H	1B	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GIDP	HBP	SH	SF	IBB	LOB	R/Gm
1	2001	SEA	AL	116	162	116	46	.716	32	6474	5680	927	1637	1120	310	38	169	881	174	42	614	989	.288	.360	.445	.805	117	2530	112	62	48	70	54	1257	5.7
2	1998	NYY	AL	114	162	114	48	.704	32	6444	5643	965	1625	1097	290	31	207	907	153	63	653	1025	.288	.364	.460	.825	116	2598	145	57	32	59	34	1203	6.0
3	2022	LAD	NL	111	162	111	51	.685	29	6247	5526	847	1418	850	325	31	212	812	98	18	607	1374	.257	.333	.442	.775	115	2441	85	56	3	53	22	1159	5.2
4	2018	BOS	AL	108	162	108	54	.667	40	6302	5623	876	1509	915	355	31	208	829	125	31	569	1253	.268	.339	.453	.792	112	2550	130	55	7	48	38	1124	5.4
5	2021	SFG	NL	107	162	107	55	.660	54	6196	5462	804	1360	823	271	25	241	768	66	14	602	1461	.249	.329	.440	.769	107	2404	117	64	36	30	45	1138	5.0
6	2019	HOU	AL	107	162	107	55	.660	36	6394	5613	920	1538	899	323	28	288	891	67	27	645	1166	.274	.352	.495	.848	119	2781	146	66	10	57	17	1168	5.7
7	2022	HOU	AL	106	162	106	56	.654	23	6054	5409	737	1341	830	284	13	214	715	83	22	528	1179	.248	.319	.424	.743	111	2293	118	60	9	42	18	1068	4.5
8	2021	LAD	NL	106	162	106	56	.654	61	6239	5445	830	1330	822	247	24	237	799	65	17	613	1408	.244	.330	.429	.759	101	2336	96	104	32	45	36	1169	5.1
9	2019	LAD	NL	106	162	106	56	.654	46	6282	5493	886	1414	813	302	20	279	861	57	10	607	1356	.257	.338	.472	.810	111	2593	100	81	55	45	47	1124	5.5
10	1998	ATL	NL	106	162	106	56	.654	42	6217	5484	826	1489	951	297	26	215	794	98	43	548	1062	.272	.342	.453	.795	107	2483	104	61	76	46	37	1148	5.1
11	2004	STL	NL	105	162	105	57	.648	37	6297	5555	855	1544	987	319	24	214	817	111	47	548	1085	.278	.344	.460	.804	107	2553	121	51	73	70	64	1156	5.3
12	2023	ATL	NL	104	162	104	58	.642	23	6249	5597	947	1543	920	293	23	307	916	132	27	538	1289	.276	.344	.501	.845	126	2803	128	67	2	43	20	1062	5.8
13	2017	LAD	NL	104	162	104	58	.642	52	6191	5408	770	1347	794	312	20	221	730	77	28	649	1380	.249	.334	.437	.771	104	2362	119	64	31	38	41	1146	4.8
14	2019	NYY	AL	103	162	103	59	.636	39	6245	5583	943	1493	880	290	17	306	904	55	22	569	1437	.267	.339	.490	.829	118	2735	113	49	10	33	18	1039	5.8
15	2018	HOU	AL	103	162	103	59	.636	36	6146	5453	797	1390	889	278	18	205	763	71	26	565	1197	.255	.329	.425	.754	106	2319	156	61	14	45	19	1052	4.9
16	2016	CHC	NL	103	162	103	58	.640	45	6335	5503	808	1409	887	293	30	199	767	66	34	656	1339	.256	.343	.429	.772	104	2359	107	96	42	37	45	1217	5.0

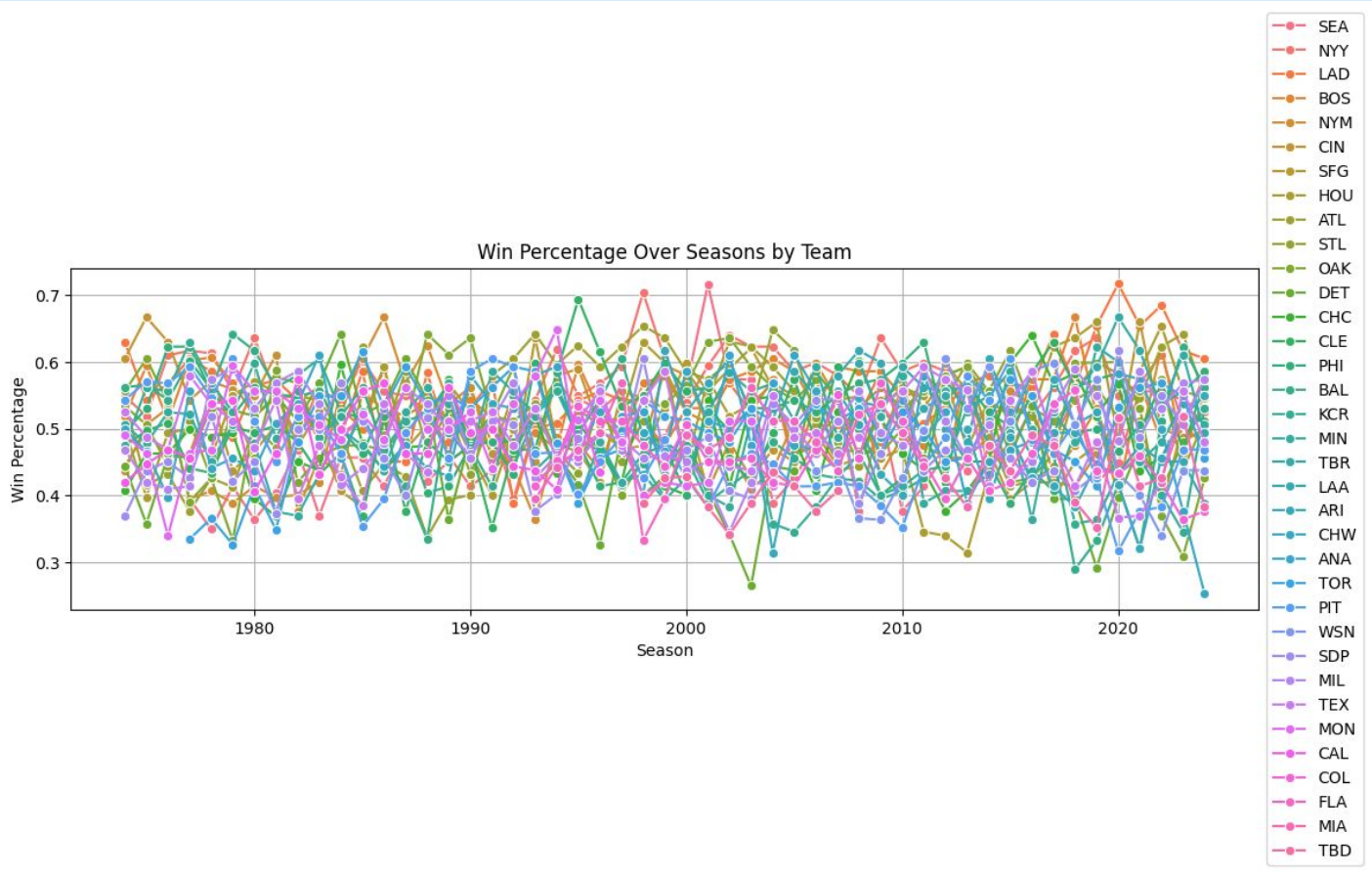
Limitations of Data

- Teams change year over year, you cannot predict free agent signings, injuries or trades given this data.
- Differences in ballparks can skew data for some teams. (Ex. Shorter right fields will lead to more homeruns for lefty batters)
- There is no salary cap in baseball, leading to larger market teams such as LA and NY spending a disproportionate amount of money on payroll.
- There are several instances in the data of teams which are no longer in existence, changed names or became a team later on.

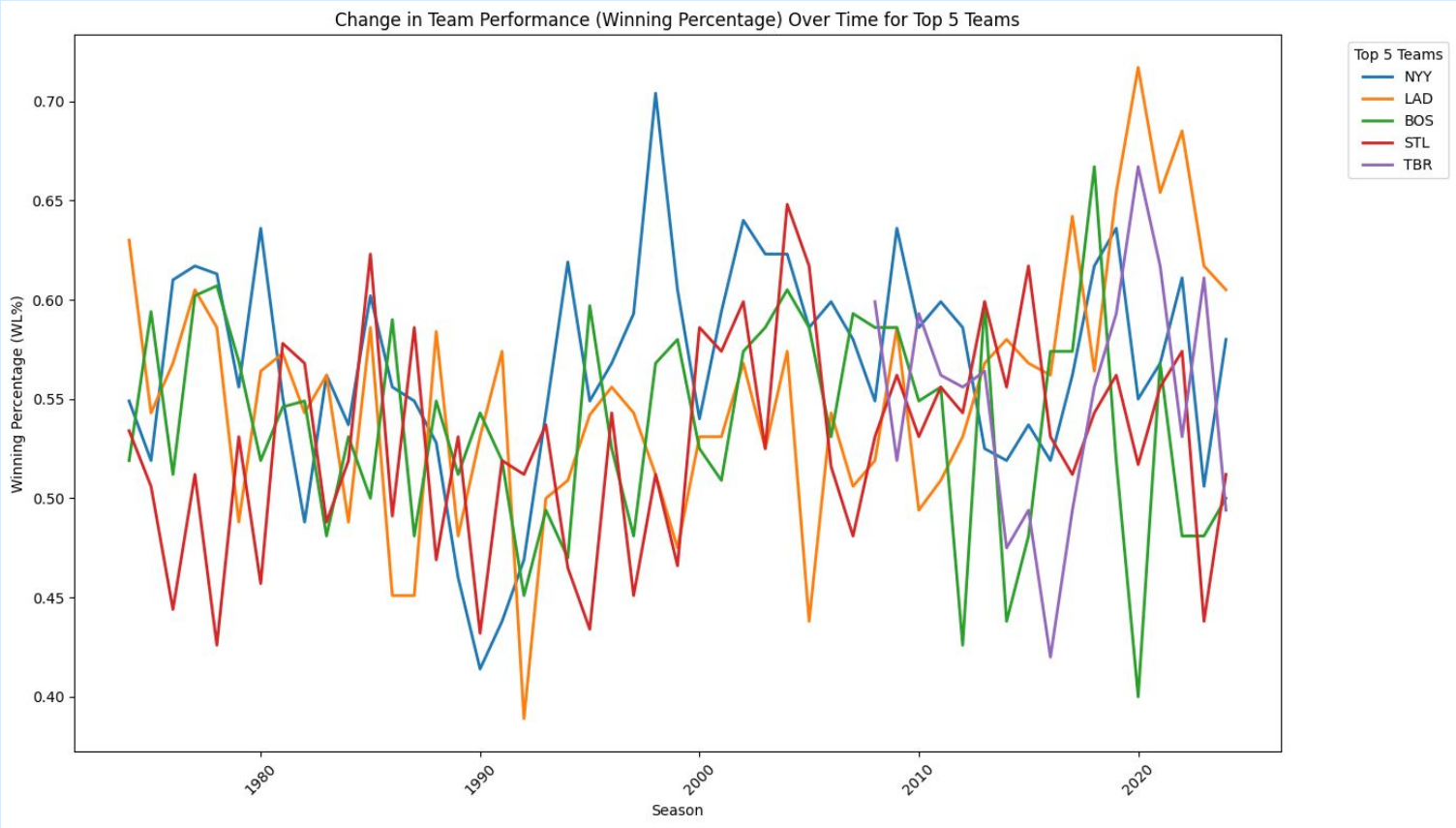
Data Exploration



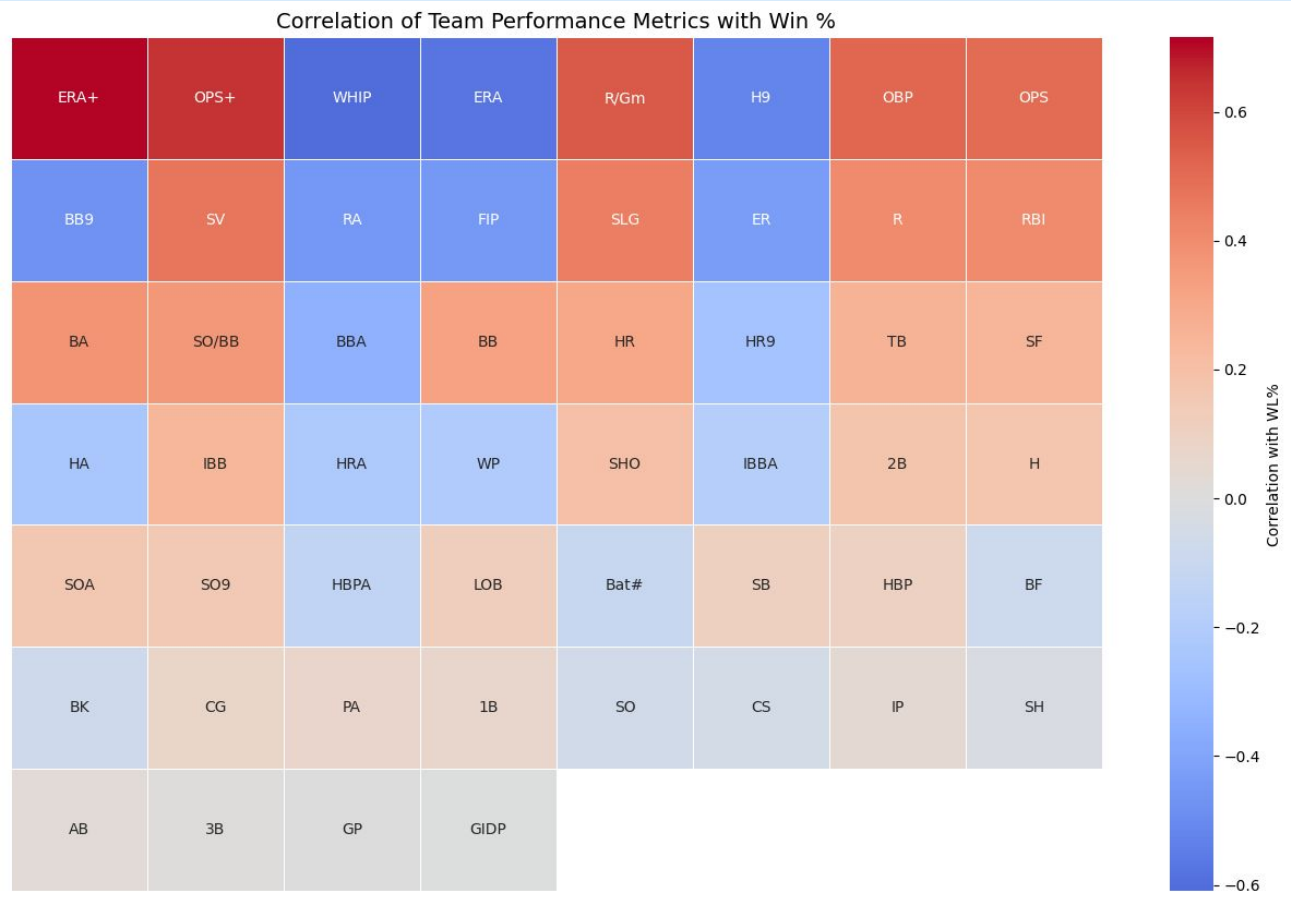
Historically, certain teams outperform the league average



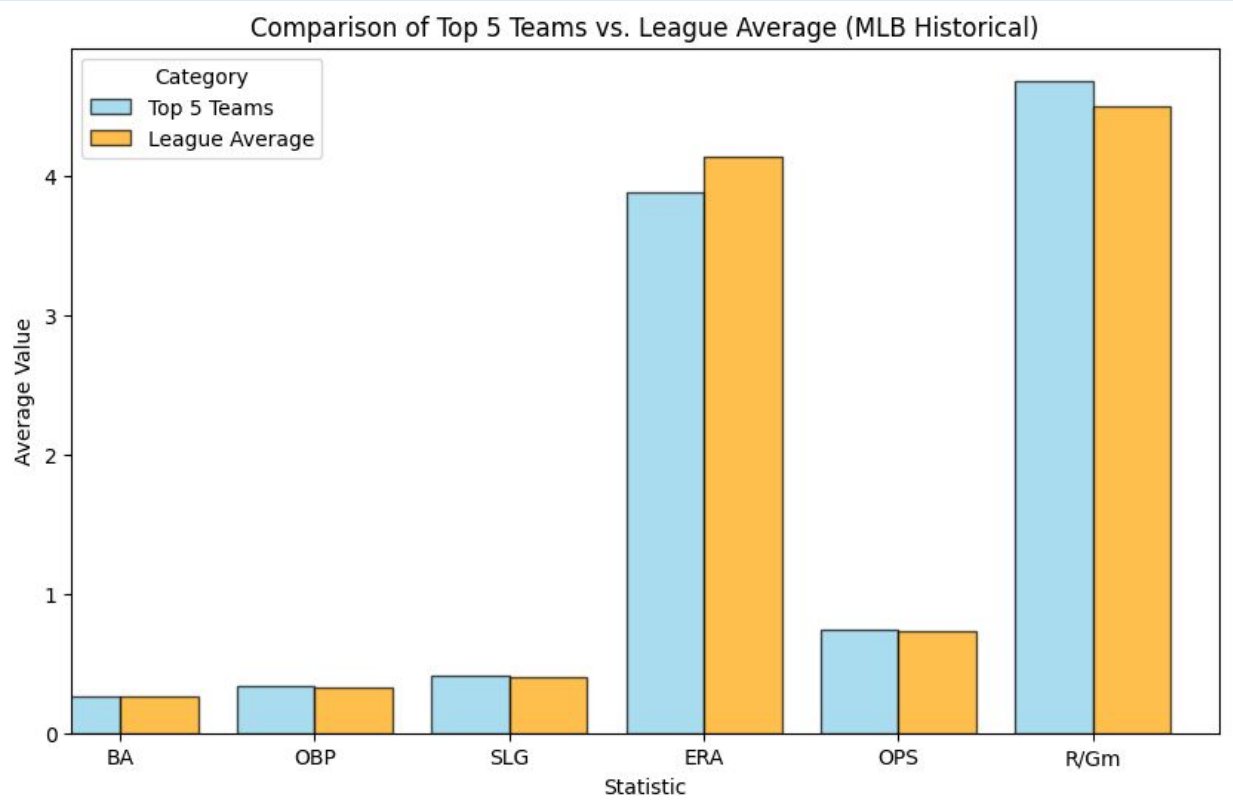
Winning is hard



However, even winners do not win consistently



Lots of factors are correlated with winning



Using team wide statistics, we see top teams outperform in certain categories

Rolling Lagged Model: Randomforest

Benefits

1. Handles complex non-linear relationships
2. Minimizes effect of outliers

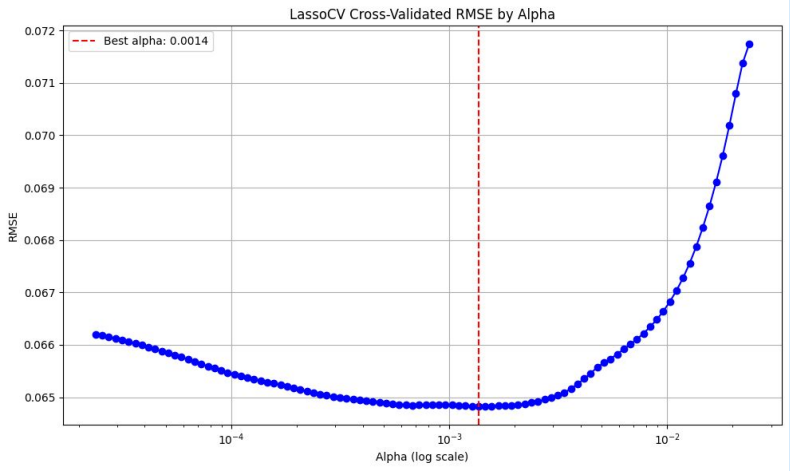
Lagging Periods

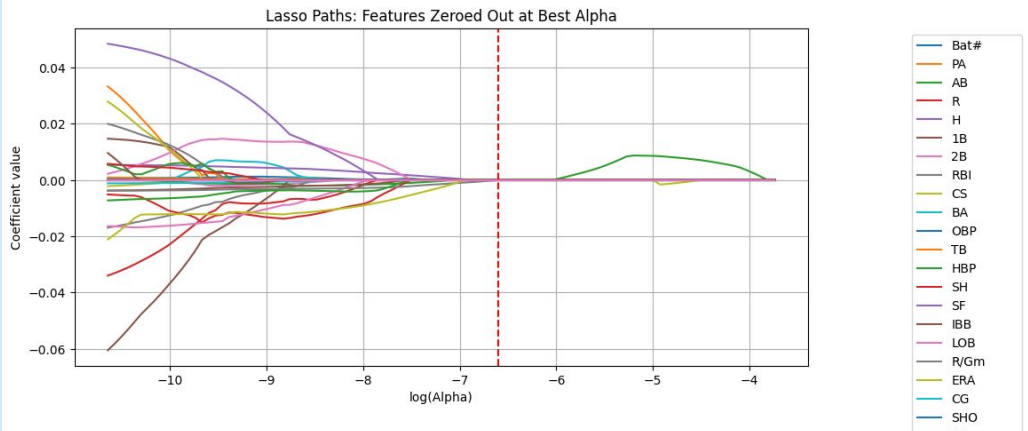
- Captures recent trends
- Prevents overfitting
- Smoothes noises

Team	Predicted WL%	Projected Wins
LAD	0.582	94.2
ATL	0.569	92.1
PHI	0.552	89.5
NYY	0.55	89.1
HOU	0.544	88.1
TOR	0.532	87.7
SEA	0.541	87.6
CHC	0.535	86.7
ARI	0.533	86.4
BAL	0.533	86.3
SDP	0.532	86.2
MIN	0.535	86.0
BOS	0.531	86.0
NYM	0.53	85.8
CLE	0.523	84.7
STL	0.52	84.2
TEX	0.506	81.9
MIL	0.489	79.2
SFG	0.477	77.3
CIN	0.477	77.2
LAA	0.47	76.2
PIT	0.47	76.1
OAK	0.467	75.7
KCR	0.464	75.1
DET	0.455	73.6
CHW	0.448	72.6
MIA	0.445	72.1
WSN	0.444	72.1
COL	0.417	67.6

Lagged Lasso Regression Model

Lasso Model Design





14

Variables Kept Under Lasso

3B,HR, SB, BB, SO, SLG, OPS, OPS+, GIDP, HRA, HBPA, WHIP, H9, HR9

37

Variables Dropped via Coefficient Shrinkage

Include notable statistics: ERA, ERA+, PA, R/GM

Top 5

Most important predictors by absolute coefficients

- WHIP: -0.0113
- OPS: 0.0104
- HR9: -0.0088
- OPS+: 0.0071
- SLG: 0.0063

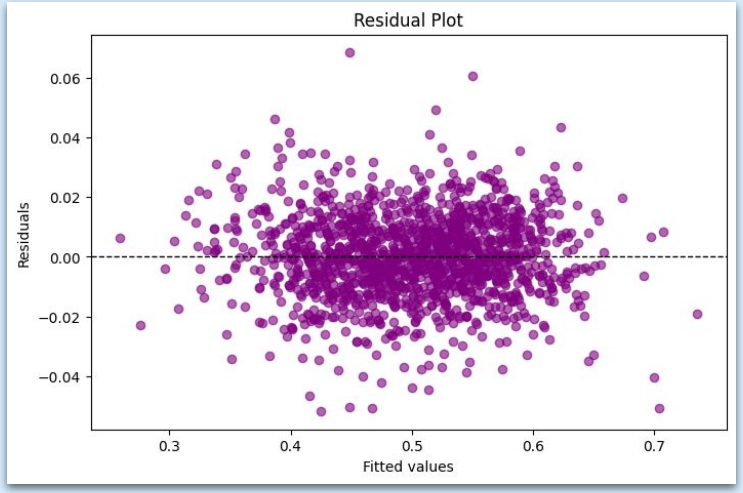
Linear Framework

Ordinary Least Squares

- **Objective**: Forecasting future win percentage
- **Approach:** Use Intratemporal data (1994-2024), season-level performance data
- **Goal:**
  1. Ensure strong generalization
  2. Predict Using performance metrics only
- **Design Choice:**
  - Drop team identifiers to avoid overfitting
  - Retain year and league indicators (Season, Lg)
- **Exclusions:** W (Wins), L (losses), and GP (Games Played) due to data leakage and redundancy.

In-Sample Results

- Model explains 96.1% of the variance in win percentages
- Several significant predictors:
  - Runs (R) and LOB (Left on Base) positively correlate with Win % → teams with better offense and on-base efficiency win more.
  - ERA and RA (Runs Allowed) negatively correlate with Win % → stronger pitching limits opponent scoring, boosting win rate.



Residuals exhibit strong model fit: symmetry, constant variance, and no major violations of linear assumptions..

Out-of-Sample Testing & Multicollinearity

- **Train/Test Split:** 80/20 chronological → simulates forecasting with historical data
- **Test RMSE:** 0.0873 → on average, predictions deviate from actual Win % by ~8.7 percentage points
- **Diagnostic Check:** Correlated Variables → Variance Inflation Factors (VIF)
  - Found 40 variables with VIF above the conventional threshold of 10 → strong evidence of multicollinearity
  - Multicollinearity inflates standard errors and undermines coefficient interpretability

OLS with Rolling-Averages

- Applied 3-year rolling averages to reduce season-to-season volatility and smooth trends.
- **Train/Test Split:** 80/20 chronological
- **Test MSE:** 0.00621 → improved generalization compared to raw-data OLS (0.0873)
- Lower-in sample R^2 (0.0248) reflect reduced variance and fewer overfit predictors
- Many predictors lost significance → less noisy input features.
- Multicollinearity remains a concern → **Action Taken:** Reduced Predictors using Lasso, Best Subset, and Tree-based methods

Variable Selection

OLS via LASSO-Regressors

- Applied Lasso regression to full standardized data set (α = 0.01) → Shrinks coefficients and selects only the most predictive features
- 7 predictors selected from 51 total: ERA+, OPS+, R/Gm, ERA, SV, WHIP, OBP

Model	R^2 (Train)	RMSE (Test)	MAE (Test)
Full OLS (49 vars)	0.962	0.0171	0.0131
Lasso-Reduced OLS	0.889	0.0250	0.0193

- Ful model is more accurate but LASSO-selected model is more interpretable
- All selected variables are intuitively meaningful

LASSO with Rolling-Averages

- Applied LASSO regression to standardized 3-year rolling average data (α = 0.01) to smooth short-term variability and reduce overfitting.
- Selected 16 predictors out of 51 → includes offensive and pitching metrics like OPS, SO, 3B, HBP

Model	R^2 (Train)	RMSE (Test)	MAE (Test)
Full OLS (Rolling)	0.248	0.0621	0.0500
Lasso-Reduced OLS	0.254	0.0613	0.0497

- Slight performance gain over full model → LASSO effectively filtered noise
- Suggests LASSO works well even with smoothed data

Best Subset Selection

- Exhaustively evaluated all combinations of predictors for model sizes 1 -10
- Selection guided by Adjusted R^2, AIC, BIC, and Cross-Validation

Model	Predictors	RMSE (Test)	MAE (Test)	Adj. R^2 (Train)
Best Subset	**PA, R, LOB, IP**	**0.0211**	**0.0148**	**0.922**

- Achieves high accuracy with just 4 variables
- Features span both offensive pressure and pitching endurance
- Despite not incorporating time-dependent structure or modeling autocorrelation, the best subset models attain competitive predictive accuracy

Linear Framework Summary & Takeaways

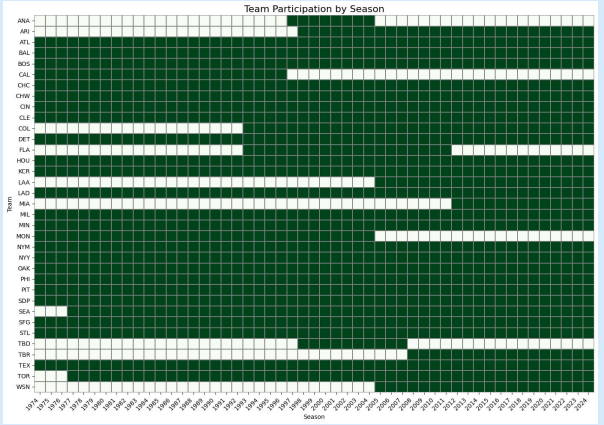
- Raw-season models outperformed rolling-average models in predictive accuracy → likely due to larger sample size and less information loss
- Best Subset achieved the best balance of accuracy and simplicity, using just 4 predictors
- LASSO effectively reduced dimensionality while maintaining solid out-of-sample performance.
- Rolling-average models, while slightly less accurate, better simulate real-world forecasting by only using past data
- These results motivate a shift toward time series methods, which explicitly model temporal dependencies

Time Series Analysis

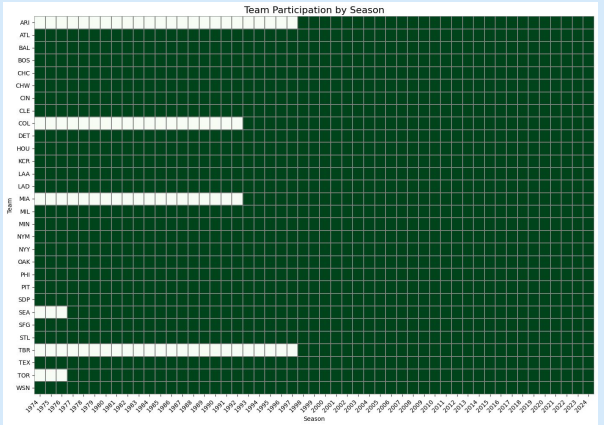
Team Name Standardization

- Several MLB franchises have changed names or locations → SARIMAX requires uninterrupted team-level data
- We standardize team codes by merging historical identifiers

Historical Code	New Standard
MON	WSN
FLA	MIA
ANA	LAA
CAL	LAA
TBD	TBR



↓



Rolling Window Regression

- Performed 10-year rolling regressions using multiple predictors at once
- Adds a time-aware dimension to variable selection → highlights robustness across eras
- **Goal:** evaluate how stable each predictor is over time in a realistic, multi-variable setting
- Variables such as OPS, OBP, and ERA become unstable when when included together due to multicollinearity
- Results suggest that SARIMAX should focus on small set of stable, non-redundant predictors → Ideally one offensive and one pitching variable

Predictor Selection: VIF analysis

- Ran Variance Inflation Factor (VIF) checks to assess multicollinearity
  - Full model → severe multicollinearity
  - Reduced model with OPS and ERA had VIF's close to 1 → low correlation and unique information
- Selected OPS (offensive) and ERA (pitching) as final predictors for SARIMAX models based on low multicollinearity and strong temporal performance

SARIMAX

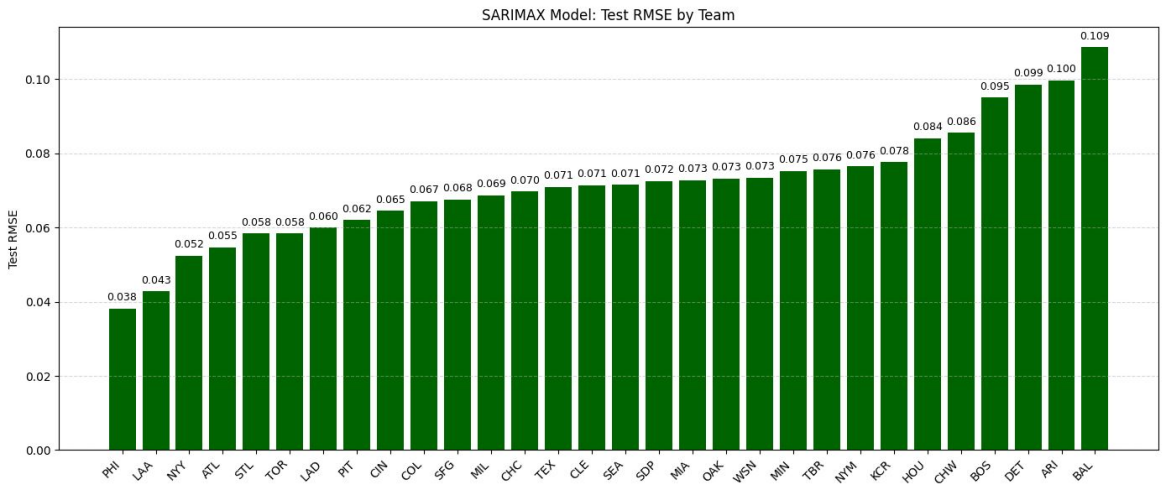
- Seasonal AutoRegressive Integrated Moving Average with Exogenous variables (extension of ARIMA)
- Models both trends and seasonality in time series → win percentage over seasons
- **Goal:** Forecast next year's win percentage using their own past win percentages and their own past stats
- **Limitations:**
  - Requires one model per team
  - Exogenous variables must be lagged
  - Seasonality must be carefully handled
  - Missing exogenous values must be dropped (or interpolated)

SARIMAX Modeling: Yankees

- Shifted ERA and OPS back by one year to avoid data leakage
- Augmented Dickey-Fuller (ADF) test to check that win % was stationary
- Baseline configuration → captures essential time series dynamics without overfitting
    - p = 1: use last year's win %
    - d = 0: no differencing needed (our data is stationary)
    - q = 1: adjust based on last year's prediction error
- OPS has a positive (significant) effect while ERA has a negative (significant) effect

Generalizing SARIMAX across all teams

- Built a separate Sarimax model for each team using a 80/20 chronological split
- Lagged predictors: ERA and OPS
- Results suggest the model captures core team dynamics
- Average **Test RMSE** across all teams: 0.0715



Results highlight strong performance for teams like PHI, LAA, and NYY, while teams with less consistent historical data (e.g., BAL, ARI) show higher prediction error.

Final Results and Betting Advice

Prefered Model Comparison

Model Name	RMSE	R^2	Predictors
Best Subset Selection	0.0211	0.922	PA, R, LOB, IP
Random Forest Rolling Averages	0.0639	0.2804	All predictors
Lasso Rolling Averages	0.063	0.2873	WHIP, OPS, HR9, OPS+, SLG
OLS Rolling Averages	0.0621	0.248	All predictors

Projected Wins vs. Sportsbook Predictions with Lasso Rolling Averages

Team	Projected Wins	Sportsbook Wins	Win Residual
CHW	71.5	54.5	+17.0
COL	68.8	53.5	+15.3
MIA	74.9	63.5	+11.4
PIT	76.2	68.5	+7.7
STL	83.7	76.5	+7.1
MIN	85.2	78.5	+6.7
TBR	86.5	80.5	+6.0
HOU	90.0	84.5	+5.5
ATL	93.4	89.5	+3.9
TOR	84.3	80.5	+3.8
MIL	85.3	81.5	+3.8
SEA	88.2	84.5	+3.7
BAL	86.6	83.5	+3.1
LAA	78.5	75.5	+3.0
WSN	71.8	69.5	+2.3
OAK	72.4	70.5	+1.9
CLE	82.2	80.5	+1.8
NYY	89.5	90.5	-1.0
CIN	76.6	79.5	-2.9
TEX	84.0	86.5	-3.0
SFG	82.3	85.5	-3.2
BOS	82.0	85.5	-3.5
PHI	89.0	93.5	-4.5
ARI	82.6	87.5	-4.9
SDP	86.3	91.5	-5.2
CHC	83.2	88.5	-5.3
KCR	77.1	82.5	-5.4
NYM	86.1	93.5	-7.4
DET	77.9	86.5	-8.6
LAD	96.9	105.5	-8.6

Using Our Model to Beat Vegas

- The top 5 and bottom 6 win residuals represent the largest differences between our model and the Vegas' Sportsbook prediction.

- Our recommendation is to bet over season wins on the **White Sox, Rockies, Marlins, Pirates and Cardinals.**

- Our recommendation is to bet under season wins on the **Padres, Cubs, Mets, Tigers and Dodgers.**

- Looking for student loan help? Bet a mere $100 on a parlay with these 11 predictions for a payout of $107,699.43.