Drafting new talent for the SF Giants 2023 season

Linear Regression of team win percentages from last 5 MLB regular seasons

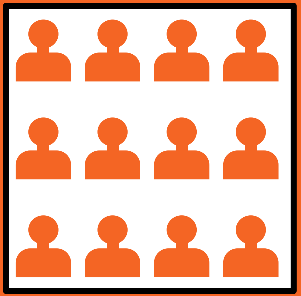
Business Problem

- Recruit new talent from collegiate and/or minor leagues.

- Predictions on team performance through the regular season.

Introduction

Introductio



Spoiler:

The 3 most predictive stats of increasing or decreasing a team's win percentage are:

1. **Runs**
2. **Strikeouts**
3. **Walks**

Reg Season Win %

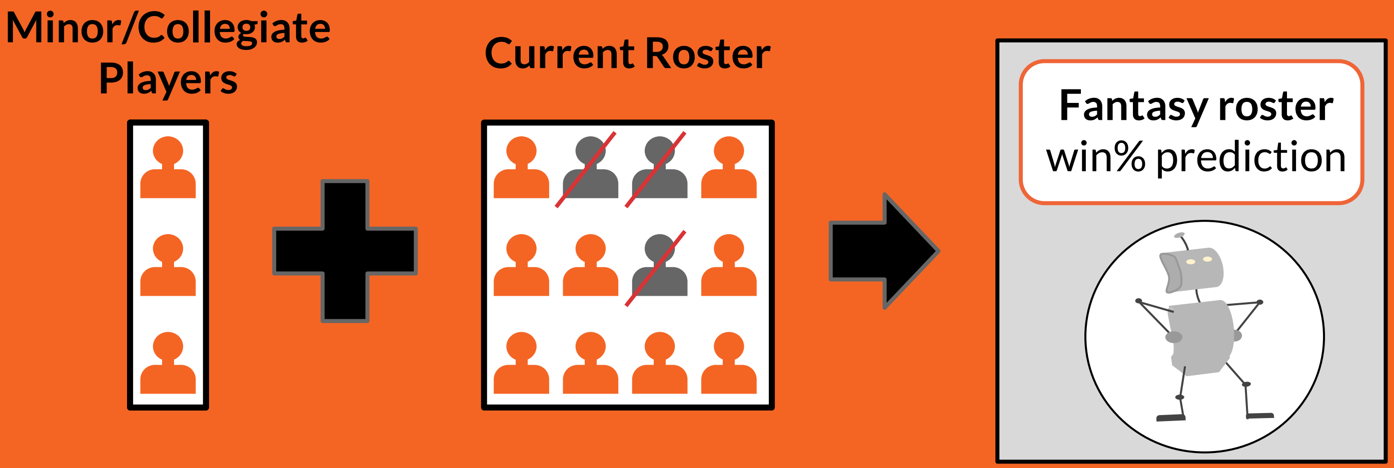
Introduction

Spoiler:

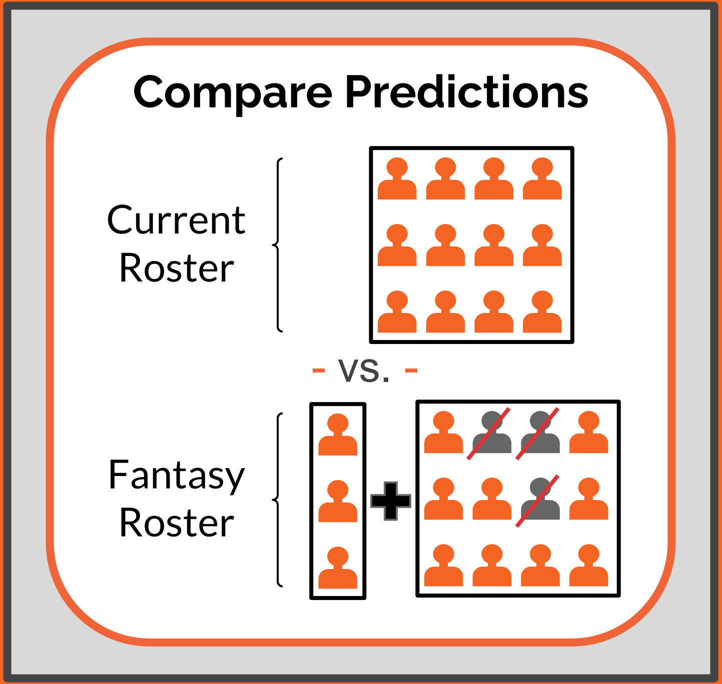
The 3 most predictive stats of increasing or decreasing a team's win percentage are:

1. **Runs**
2. **Strikeouts**
3. **Walks**

Introduction



Introduction



How do we

MAXIMIZE

Win %

3 Steps to Maximizing Win %

Step 1

Get Data

Step 2

Train a Model

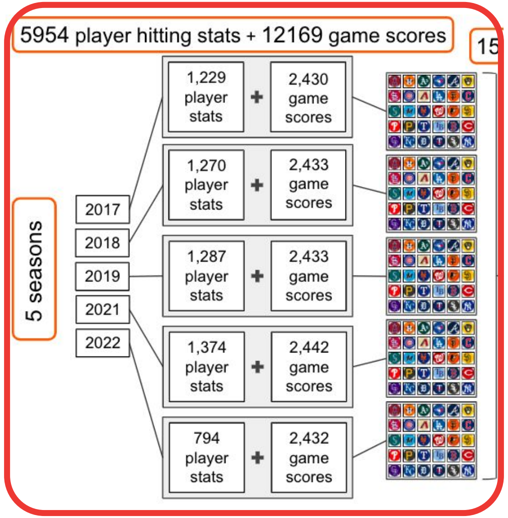
Step 3

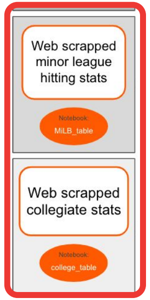
Utilize Trained Model

3 Steps to Maximizing Win %

Step 1

Get Data





Step 1

Get Data

Get Data - Web Scraping

- Major League Player and Game Stats
  - 2017, 2018, 2019, 2021, 2022

MLB.com

- Minor League Triple-A Player Stats
  - 2022

MiLB.com

- Collegiate Division-1 Player Stats
  - 2022

TheBaseballCube.com
& D1Baseball.com

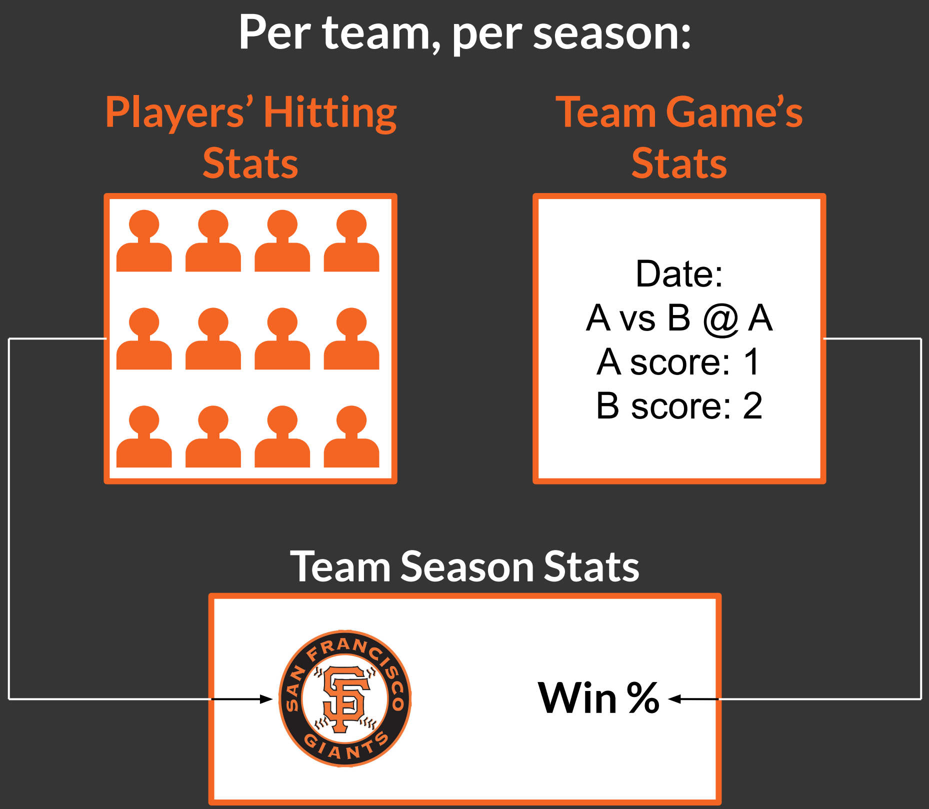
Get Data - Feature Engineering

Per team, per season:

- Major League Player and Game Stats
  - 2017, 2018, 2019, 2021, 2022

MLB.com

Get Data - Feature Engineering



Get Data - Feature Engineering

30 teams per season:

5 seasons:

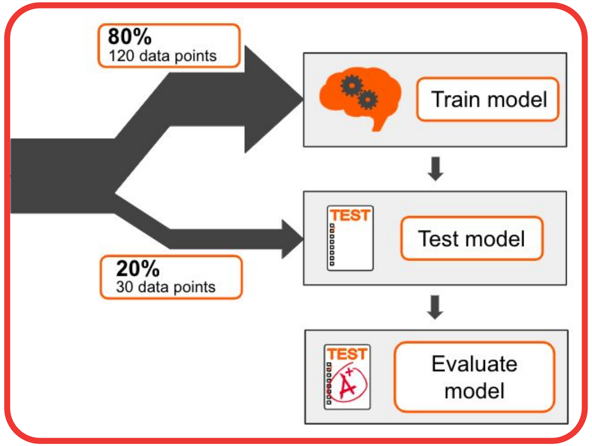
Get Data - Feature Engineering

150 Data Points

3 Steps to Maximizing Win %

Step 2

Train Model



Step 2

Train Model

Train Model

(and Test, and Evaluate)

Train model

80% of data (120 data points)

Test model

Remaining 20% of data (30 data points)

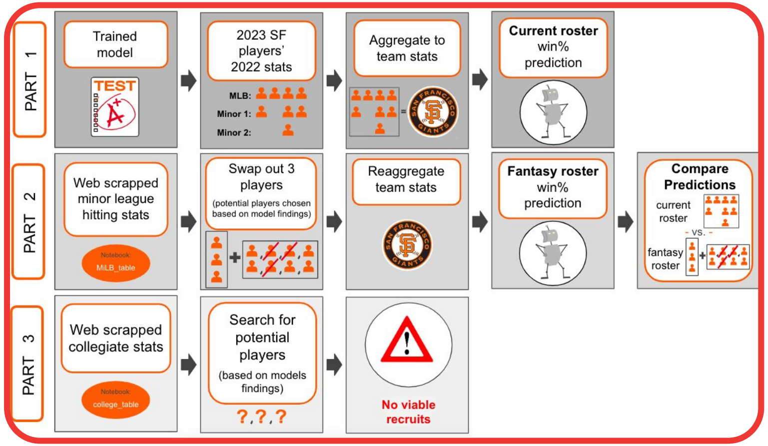
Evaluate model

Explains 72.7% of the variability observed in regular season win %

3 Steps to Maximizing Win %

Step 3

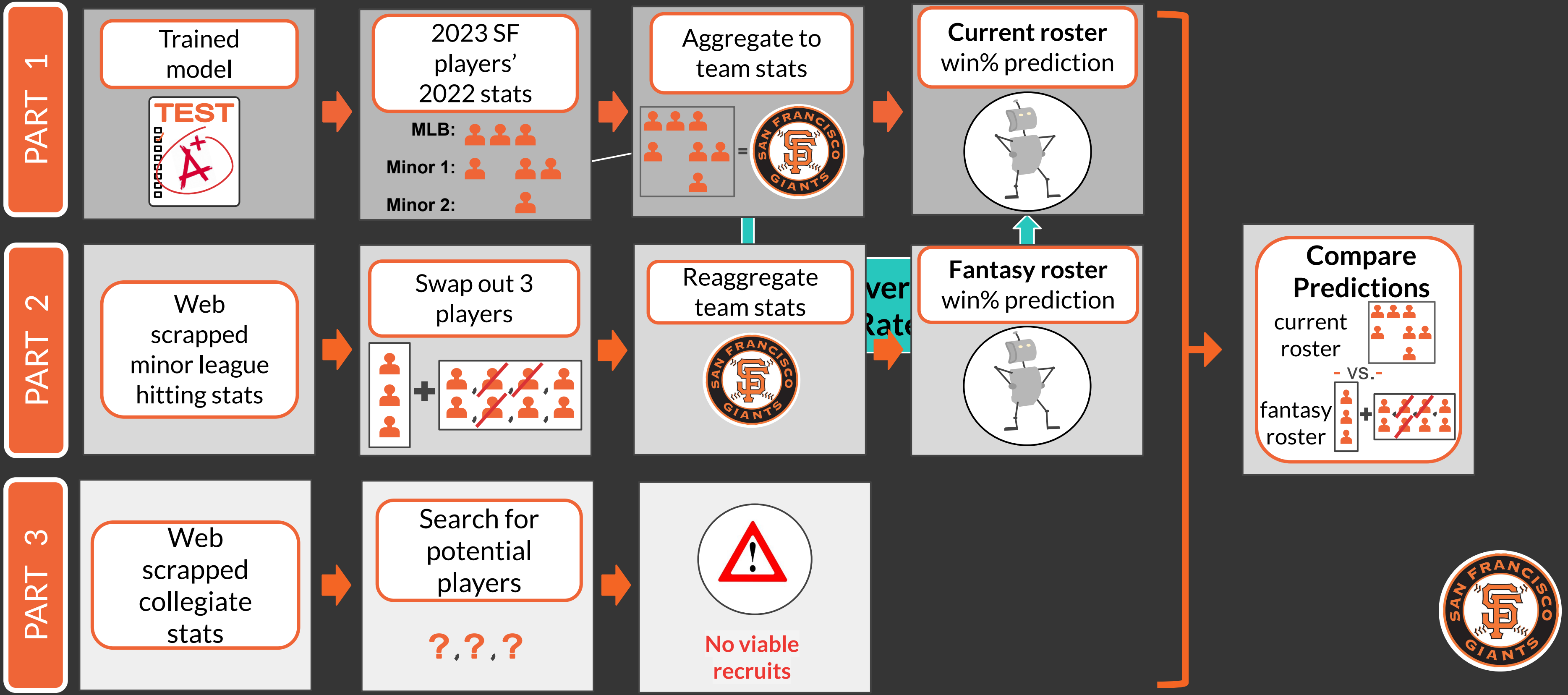
Utilize Model



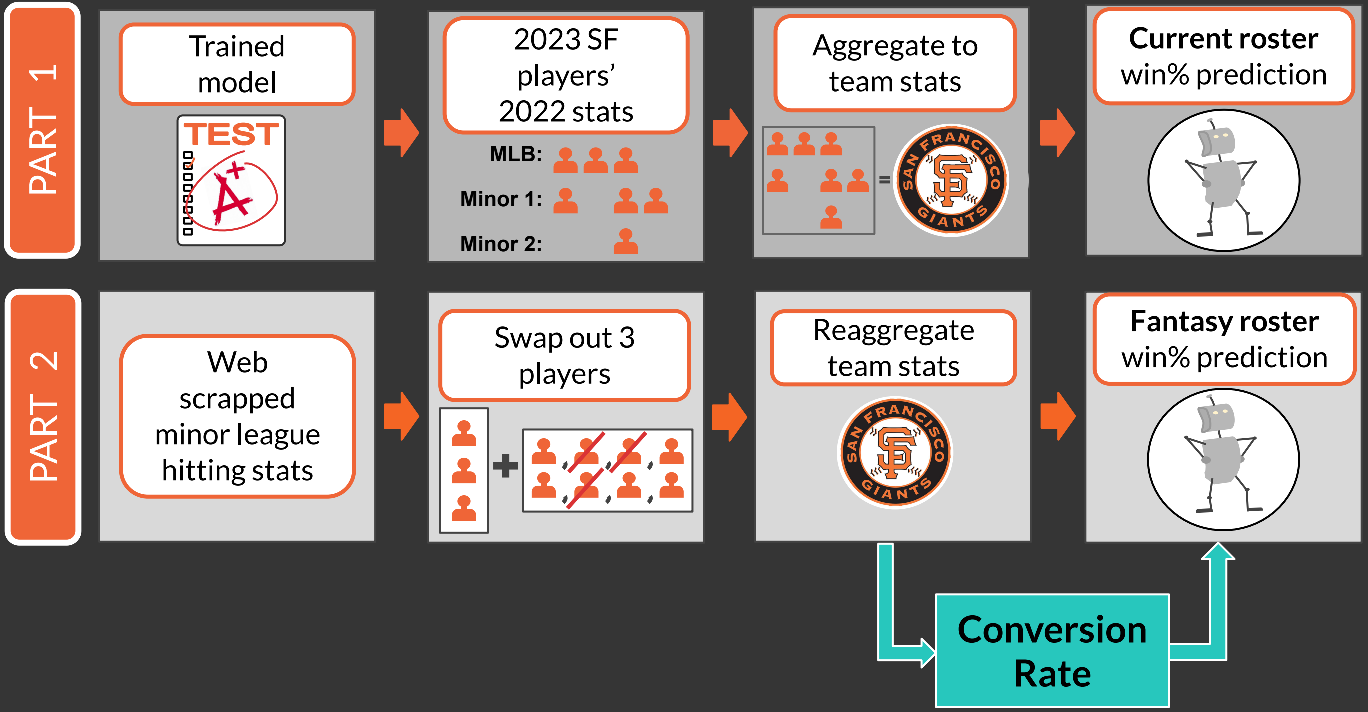
Step 3

Utilize Model

Utilize Model



Utilize Model



Utilize Model

PART 1

- Trained model
  - TEST
- 2023 SF players' 2022 stats
  - MLB:
  - Minor 1:
  - Minor 2:
- Aggregate to team stats
- **Current roster** win% prediction

PART 2

- Web scrapped minor league hitting stats
- Swap out 3 players
- Reaggregate team stats
- **Fantasy roster** win% prediction

PART 3

- Web scrapped collegiate stats
- Search for potential players
  - ?,?,?
- **No viable recruits**

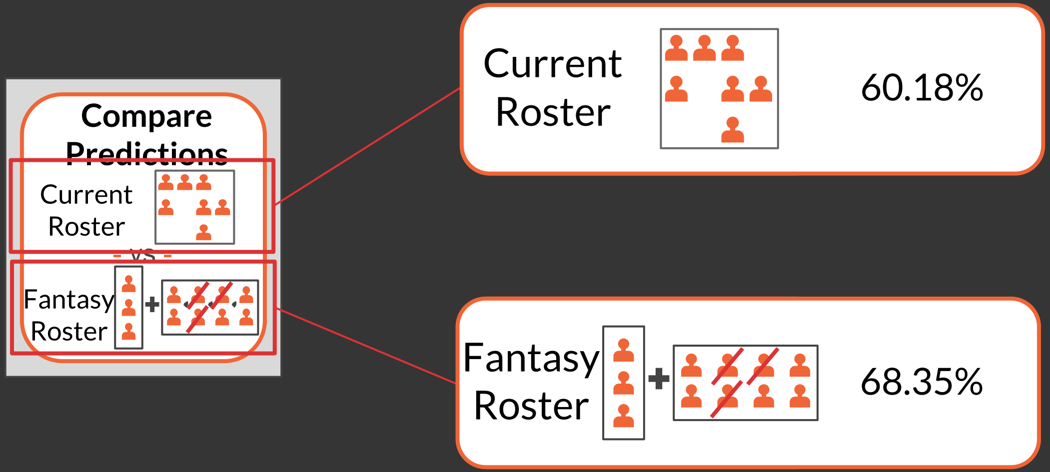
Compare Predictions

current roster

- vs.-

fantasy roster

Utilize Model



Spoiler

The 3 most predictive stats of increasing or decreasing a team's win percentage are:

1. **Runs**
2. **Strikeouts**
3. **Walks**

By how much do these 3 statistics affect win % ?

3 most predictive stats:

1 Run = 0.0491%

1 Strikeout = -0.0195%

1 Walk = 0.0159%

Limitations

There are 2 limitations to this model:

1. The amount of data
   - 150 data points

2. The statistics used
   - MLB hitting stats

Next Steps

Next steps for this project:

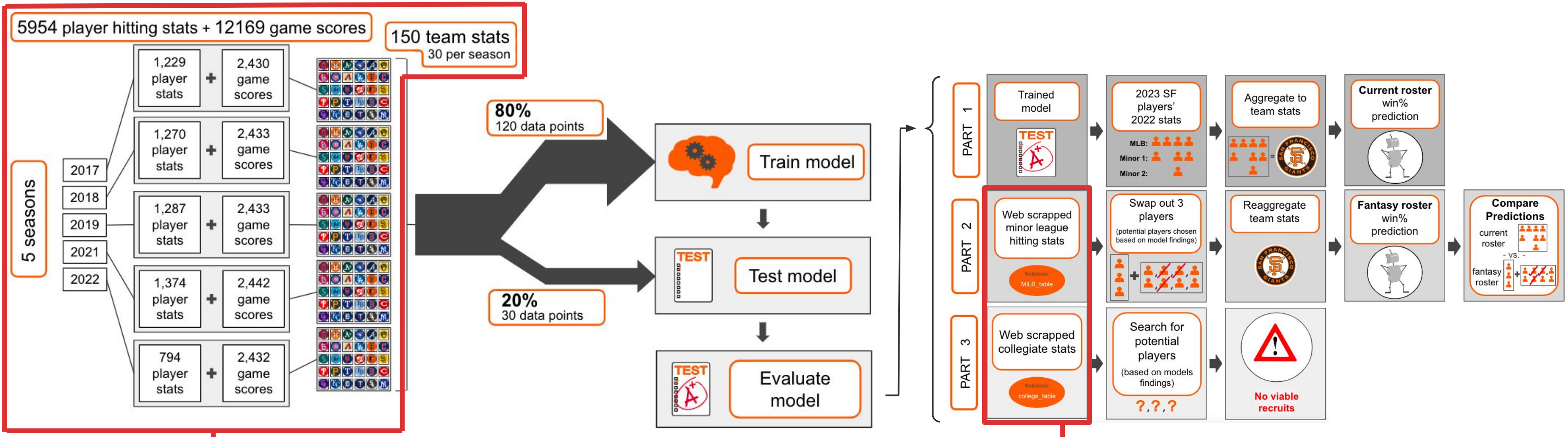
- Webscrap more seasons - more data points
- Include pitching and fielding stats - more information
- Combine minor and major stats - potentially more or less accurate information
- Use prior year stats to predict post year win percentage - for predictive purposes only
- Multiply each feature by it's percent of impact, then aggregate the three features to get the exact impact a player will have on a teams win percentage - may change recruitment strategy

Questions

Thank you

Let’s work together,
Cassarra Groesbeck

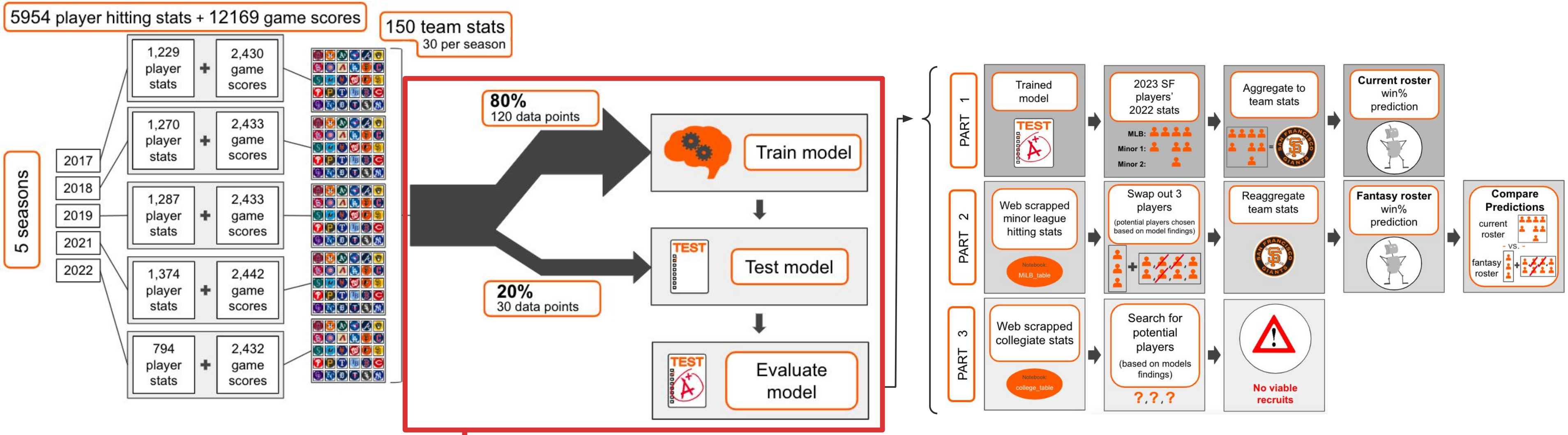
3 Steps to Maximizing Win %



Step 1

Get Data

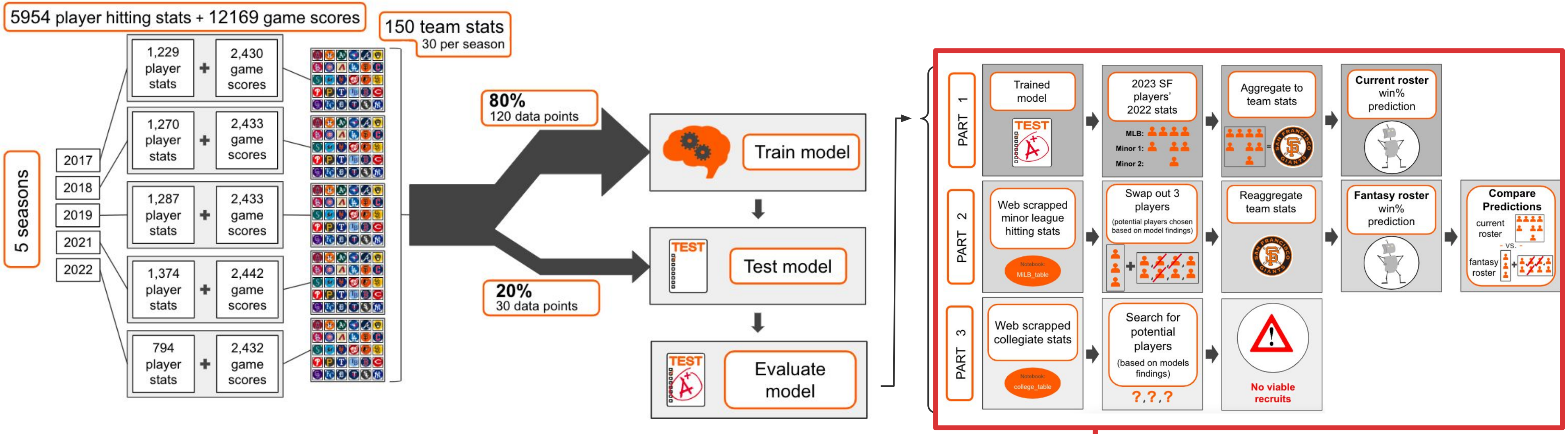
3 Steps to Maximizing Win %



Step 2

Train Model

3 Steps to Maximizing Win %



Step 3

Utilize Model

	Team	Year	Runs Sum	Walks Sum	Strikeouts Sum	% wins
6	HOU	2022	801	573	1271	65.64
17	SF	2022	678	543	1390	50.00
31	ATL	2021	949	662	1790	54.32
47	SF	2021	862	639	1540	66.05
71	WSH	2019	925	630	1402	57.41
77	SF	2019	625	421	1325	47.53
105	BOS	2018	940	606	1307	66.67
107	SF	2018	525	361	1276	45.06
126	HOU	2017	924	542	1127	62.35
137	SF	2017	612	463	1195	39.51

P Values

To determine if an observed outcome is statistically significant, we look at the P values; in this case they are all below .05 indicating that there *is* a relationship between the associated coefficient and a teams percentage of wins in their regular season games.

R squared

This final linear regression model is able to explain 72.7% of the variability observed in regular season percentage of wins.

Adjusted R squared

The adjusted r squared is essentially the same as the r squared, just 0.7% difference, so we can be confident, as stated above, in the 72.7% reliability of this model.

F statistic and Prob(f-statistic)

The Prob (F-statistic) of 1.46e-32 tells us, there is 0% chance that any experimentally observed difference is due to chance alone.

Cond. No

This summary report give a condition number of 3.65, well below 10, indicating there are no multicollinearity issues.

Swap these players in

	Team	Runs	Walks	Strikeouts	Player Name	Position	RSO
0	ABQ	95	39	67	Wynton Bernard	CF	28
54	OKC	100	71	76	Miguel Vargas	3B	24
22	ELP	73	38	58	Brett Sullivan	C	15

Swap these players out

	Team	Runs	Walks	Strikeouts	Player Name	Position	RSO
2	SF	49	40	89	Austin Slater	CF	-40
14	SAC	5	2	6	Joey Bart	C	-1
3	SF	46	39	122	J.D. Davis	3B	-76
10	SF	34	26	112	Joey Bart	C	-78

Final team used for win prediction

	Games Played	At Bats	Runs	Hits	Doubles	Triples	Home Runs	Runs Batted In	Walks	Strikeouts	Stolen Bases	Caught Stealing	Batting Average	On-Base Percentage	Slugging Percentage	On-Base Plus Slugging	Player Name
0	108.0	429.0	95.0	143.0	31.0	8.0	21.0	92.0	39.0	67.0	30.0	5.0	.333	.387	.590	.977	Wynton Bernard
54	113.0	438.0	100.0	133.0	32.0	4.0	17.0	82.0	71.0	76.0	16.0	5.0	.304	.404	.511	.915	Miguel Vargas
22	113.0	421.0	73.0	120.0	28.0	6.0	9.0	81.0	38.0	58.0	3.0	0.0	.285	.339	.444	.783	Brett Sullivan
0	134.0	380.0	57.0	104.0	19.0	3.0	23.0	70.0	42.0	100.0	3.0	2.0	0.274	0.353	0.521	0.874	Joc Pederson
1	136.0	454.0	88.0	118.0	25.0	2.0	36.0	106.0	73.0	151.0	1.0	2.0	0.506	0.735	1.072	1.807	David Villar
4	57.0	224.0	31.0	55.0	8.0	0.0	11.0	34.0	20.0	65.0	0.0	0.0	0.246	0.308	0.429	0.737	Mitch Haniger
5	140.0	488.0	71.0	127.0	22.0	2.0	14.0	62.0	33.0	89.0	21.0	6.0	0.26	0.322	0.4	0.722	Thairo Estrada
6	151.0	525.0	72.0	120.0	28.0	1.0	19.0	71.0	59.0	103.0	0.0	0.0	0.229	0.316	0.394	0.71	Wilmer Flores
7	148.0	485.0	73.0	104.0	31.0	2.0	17.0	57.0	61.0	141.0	5.0	1.0	0.214	0.305	0.392	0.697	Mike Yastrzemski
8	118.0	387.0	46.0	101.0	18.0	2.0	10.0	49.0	45.0	97.0	14.0	4.0	0.543	0.725	0.899	1.624	Luis Gonzalez
9	77.0	217.0	29.0	45.0	7.0	1.0	8.0	26.0	26.0	51.0	1.0	0.0	0.207	0.305	0.359	0.664	LaMonte Wade Jr.
11	120.0	411.0	50.0	94.0	15.0	2.0	9.0	52.0	40.0	99.0	1.0	1.0	0.231	0.508	0.344	0.852	Brandon Crawford
12	117.0	447.0	65.0	99.0	17.0	1.0	11.0	45.0	43.0	118.0	6.0	6.0	0.327	0.487	0.449	0.936	Heliot Ramos
13	123.0	447.0	74.0	127.0	26.0	6.0	19.0	75.0	55.0	129.0	10.0	2.0	0.577	0.773	1.029	1.802	Blake Sabol
14	87.0	296.0	60.0	78.0	12.0	2.0	23.0	61.0	45.0	90.0	7.0	2.0	0.275	0.669	0.574	1.243	Isan Diaz
15	65.0	227.0	33.0	61.0	12.0	0.0	11.0	36.0	26.0	58.0	0.0	0.0	0.581	0.783	1.004	1.787	Marco Luciano
16	117.0	451.0	87.0	123.0	25.0	6.0	15.0	58.0	63.0	110.0	32.0	11.0	0.512	0.632	0.793	1.425	Brett Wisely
17	93.0	376.0	58.0	81.0	15.0	1.0	12.0	47.0	28.0	66.0	11.0	3.0	0.64	0.775	1.344	2.119	Luis Matos
18	8.0	21.0	5.0	5.0	0.0	0.0	2.0	6.0	10.0	3.0	0.0	0.0	0.238	0.515	0.524	1.039	MitchHaniger
19	14.0	44.0	11.0	11.0	4.0	0.0	2.0	11.0	10.0	6.0	0.0	0.0	0.25	0.397	0.477	0.874	LaMonte Wade