Stat 451 Project

Matthew Nelson, Matt Baxley, Raymond Zhao, Matthew Attanasio, Rahul Vyas

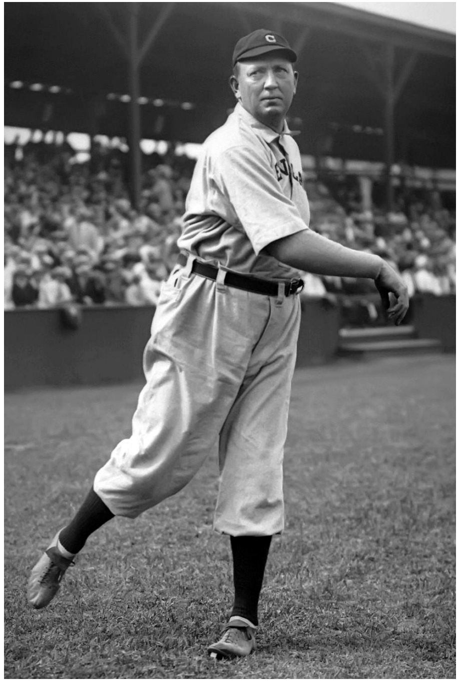
What is a Cy Young?

Award is named after Cy Young

- Won 511 games in his career
- Best season, 1901, 33 wins, 158 Ks, 1.62 ERA
- 3 no-hitters
- Hall of Fame in 1937

Cy Young Award

"...the Cy Young Award is voted upon by the Baseball Writers' Association of America prior to the beginning of the postseason. From 1956 to 1966, the award was given to one pitcher, but has been given to one pitcher per league since 1967. The vote totals are based on a weighted points system." (MLB)



The Data

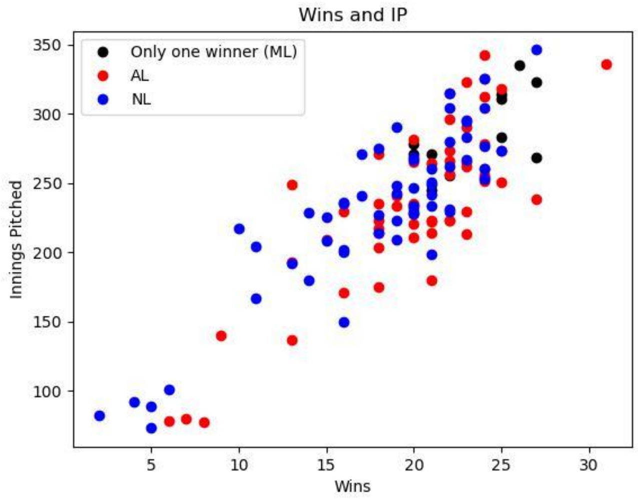
2 main datasets were combined for analysis:

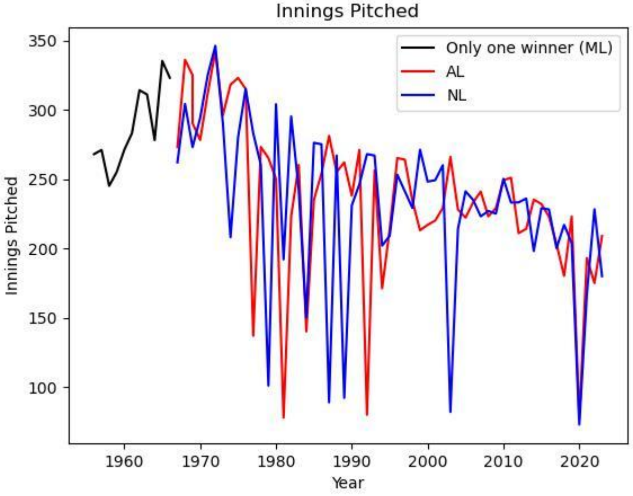
- All Cy Young winners since 1956
- Statistics of every pitcher in both the National League and American League (which make up the MLB) for the last 50 years (1974-2023).
    - Different types of statistics, including many advanced ones - 4 separate CSVs for each league's season (400 CSVs total)
        - Basic Statistics
        - Value
        - Advanced Statistics
        - Win Probability
    - Team Win % that year also added to dataset

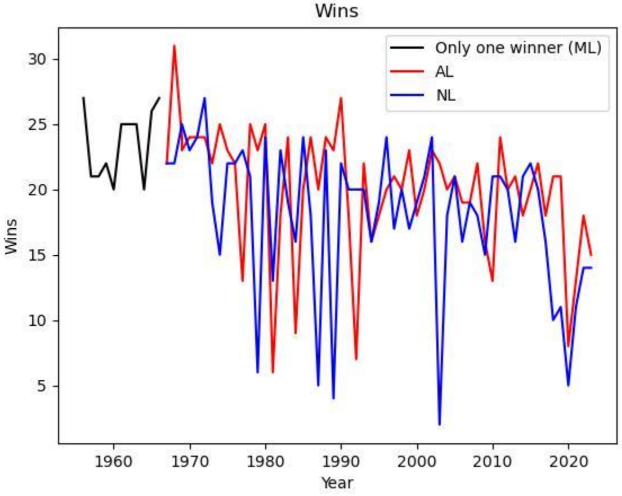
All data taken from baseball-reference.com

Correlation in the Data

1. Find correlating factors
   a. W and IP: 0.85
   b. SO and WAR: 0.71
2. Plot factors over time







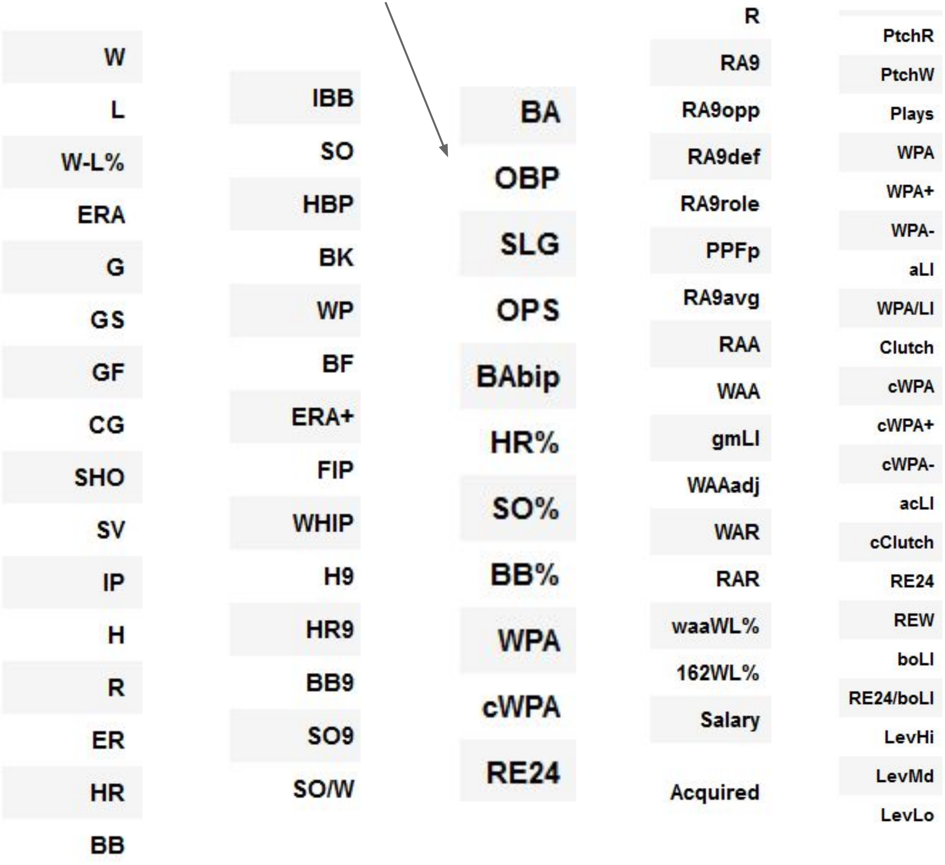
The Plan

- Perform logistic regression on the dataset, to determine which person each year has the **highest likelihood** of being a Cy Young winner
    - Since there are many, many pitchers, we filtered the pitchers to those with 70 or more innings pitched that season.
- Since there are 2 Cy Young winners per year (1 in NL and 1 in AL), we must **oversample** the data, as it is very imbalanced
- 50/50 Train-Test split, oversample training data and test on non-oversampled testing data
- Observe model performance on entire history of MLB pitching

What Features to Use?

- Advanced statistics on top of many basic statistics leads to a horrifying amount of features
- How do we figure out which ones to use?

(for anyone who knows baseball: these are averages of batters against the pitcher)



W
L
W-L%
ERA
G
GS
GF
CG
SHO
SV
IP
H
R
ER
HR
BB

IBB
SO
HBP
BK
WP
BF
ERA+
FIP
WHIP
H9
HR9
BB9
SO9
SO/W

BA
OBP
SLG
OPS
BAbip
HR%
SO%
BB%
WPA
cWPA
RE24

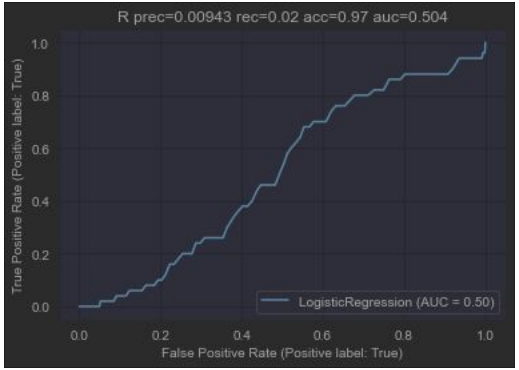
R
RA9
RA9opp
RA9def
RA9role
PPFp
RA9avg
RAA
WAA
gmLI
WAAadj
WAR
RAR
waaWL%
162WL%
Salary
Acquired

PtchR
PtchW
Plays
WPA
WPA+
WPA-
aLI
WPA/LI
Clutch
cWPA
cWPA+
cWPA-
acLI
cClutch
RE24
REW
boLI
RE24/boLI
LevHi
LevMd
LevLo

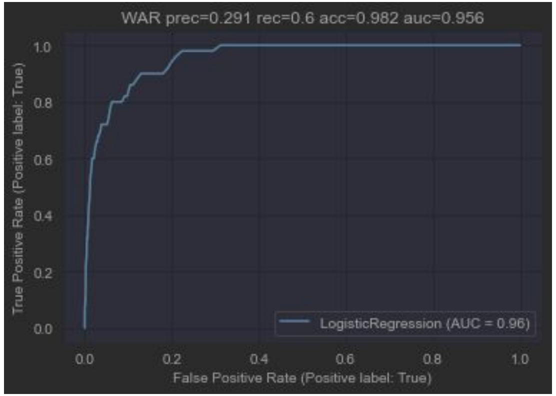
Feature Selection

We selected features using precision/recall and ROC area under curve for a model fitted using only that feature, along with some knowledge about baseball.

R (Runs Allowed)
Probably not getting used
Precision=0.00943, recall=0.02



WAR (Wins Over Replacement)
Probably getting used
Precision=0.291, recall=0.6



Features Used

- **ERA** - Earned Runs Average (Average runs allowed per 9 innings pitched)
- **GS** - Games Started
- **SO** - Strikeouts
- **SV** - Saves (When the team is up by less than 3 runs and the pitcher finishes the game, that’s counted as a save)
- **SHO** - Shutouts (Pitched 9 innings, opposing team scored 0 runs)
- **W** and **L** (Wins and Losses)
    - Only one win and loss is credited to a pitcher on each side each game (usually starter)
- Team Win Rate
- WAR (Wins Over Replacement)
- WPA (Win Percentage Added)
- PtchR (An advanced statistic that estimates a pitcher’s contribution to their team’s total runs)

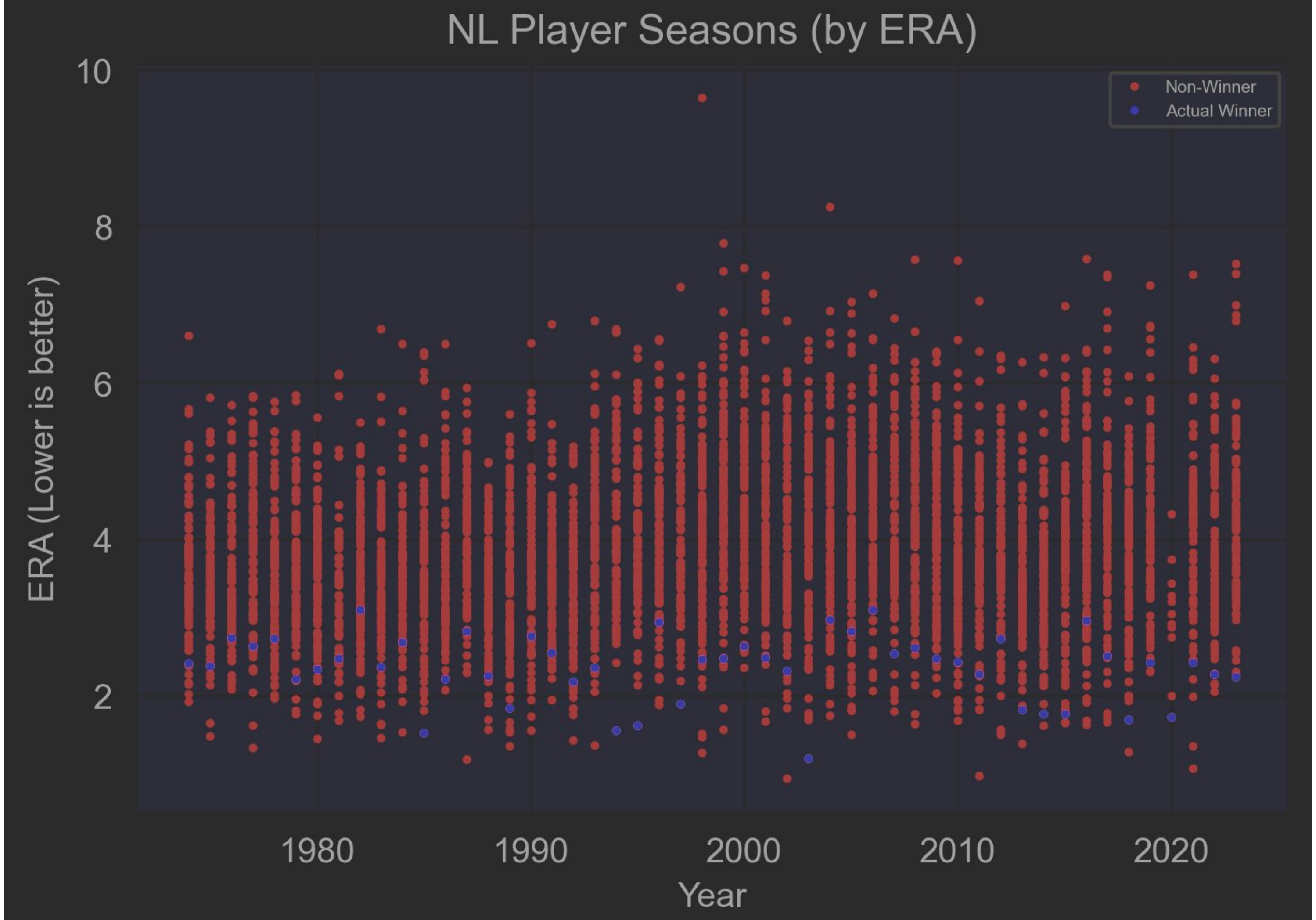
WAR

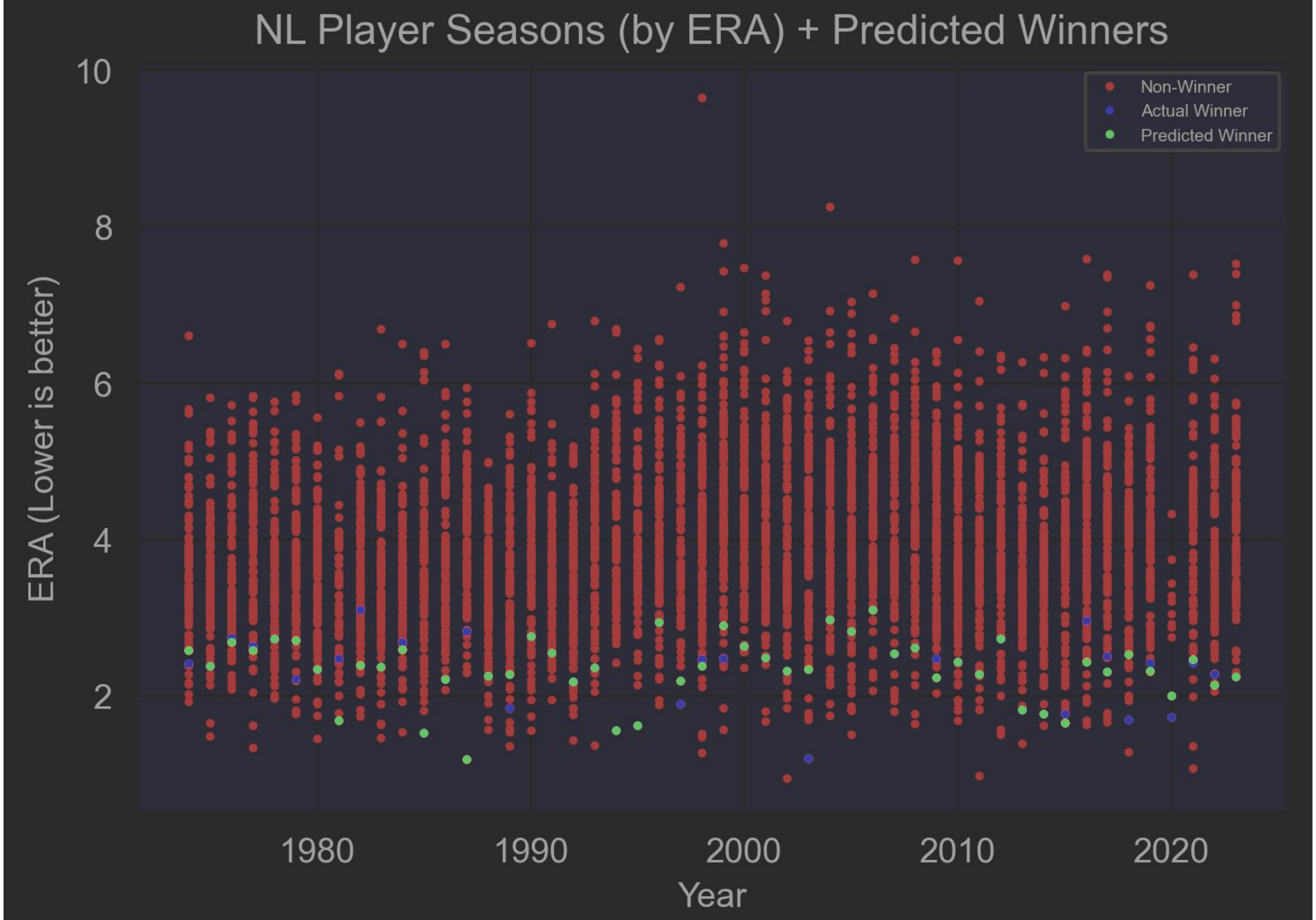
- “Wins Above Replacement”
  - How many wins a given player has added to the team above what an available replacement player would have added
- Considered one of the most important advanced statistics in baseball
- A very good measure of how much value a player is bringing to a team
- Meaning (according to baseball-reference):
  - 8+: MVP
  - 5+: All-Star
  - 2+: Starter
  - 0-2: Reserve
  - <0: Replace this player
- Similar to WPA (Win Percentage Added)
  - How much a pitcher added to their team’s win chance

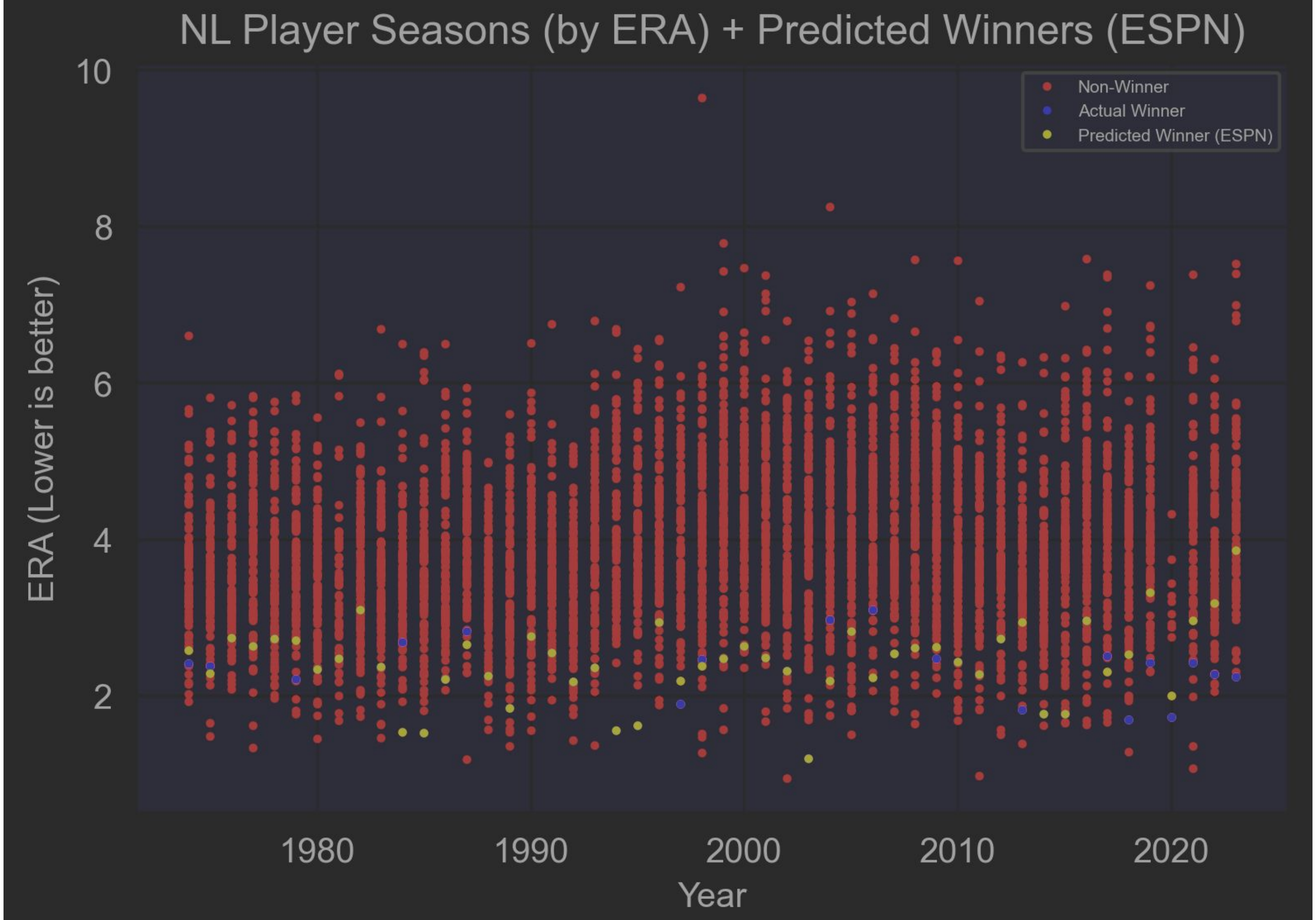
What about other people's attempts?

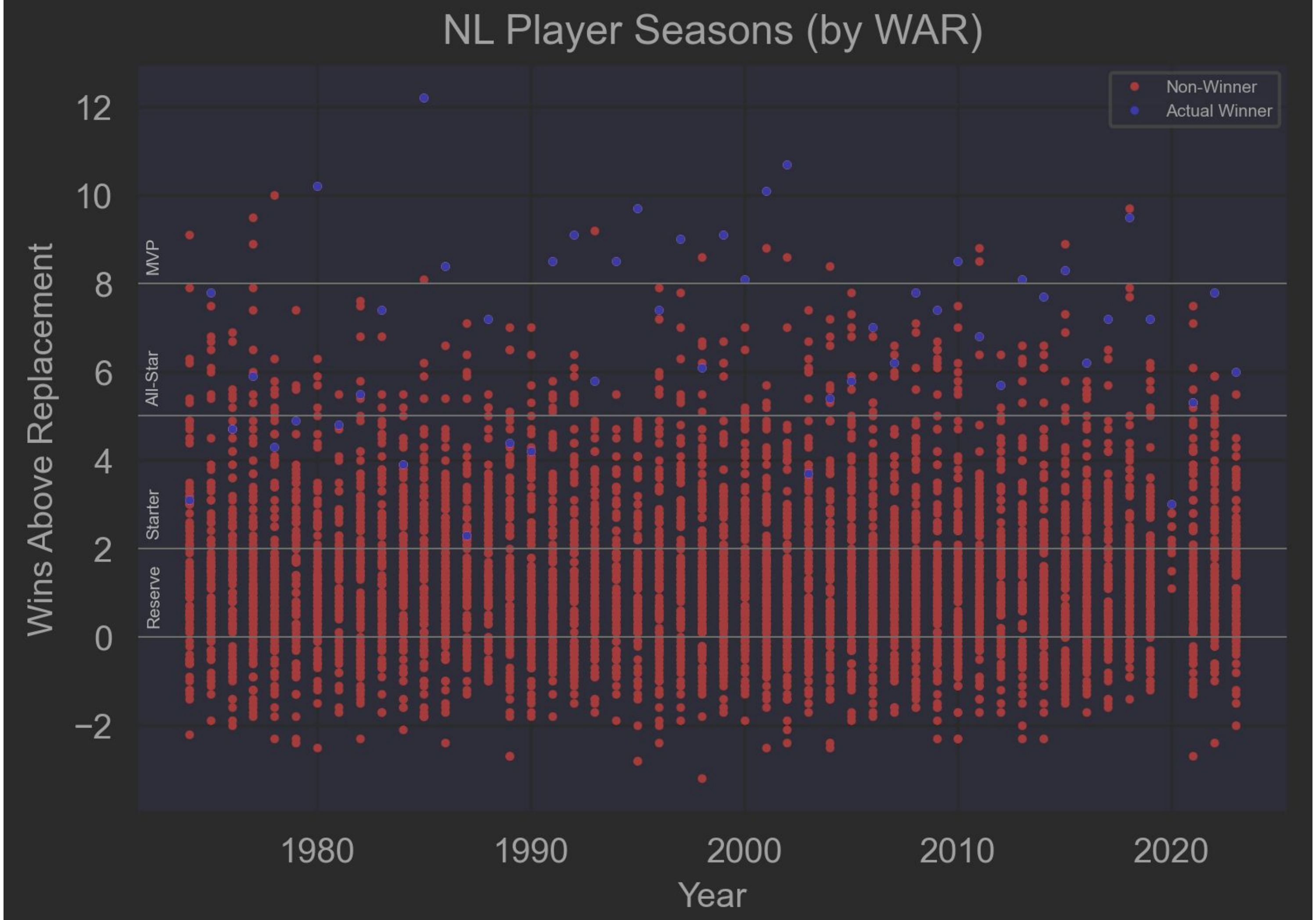
- ESPN has a Cy Young predictor
- Citing a formula from The Neyer/James Guide to Pitchers, they use the following formula to calculate a “Cy Young” score **(highest score wins)**
  - **5/9 * (Innings Pitched) - Earned Runs + (Strikeouts / 12) + (Saves * 2.5) + Shutouts + (Wins * 6 - Losses * 2) + Victory Bonus (12 if won division, 0 otherwise)**

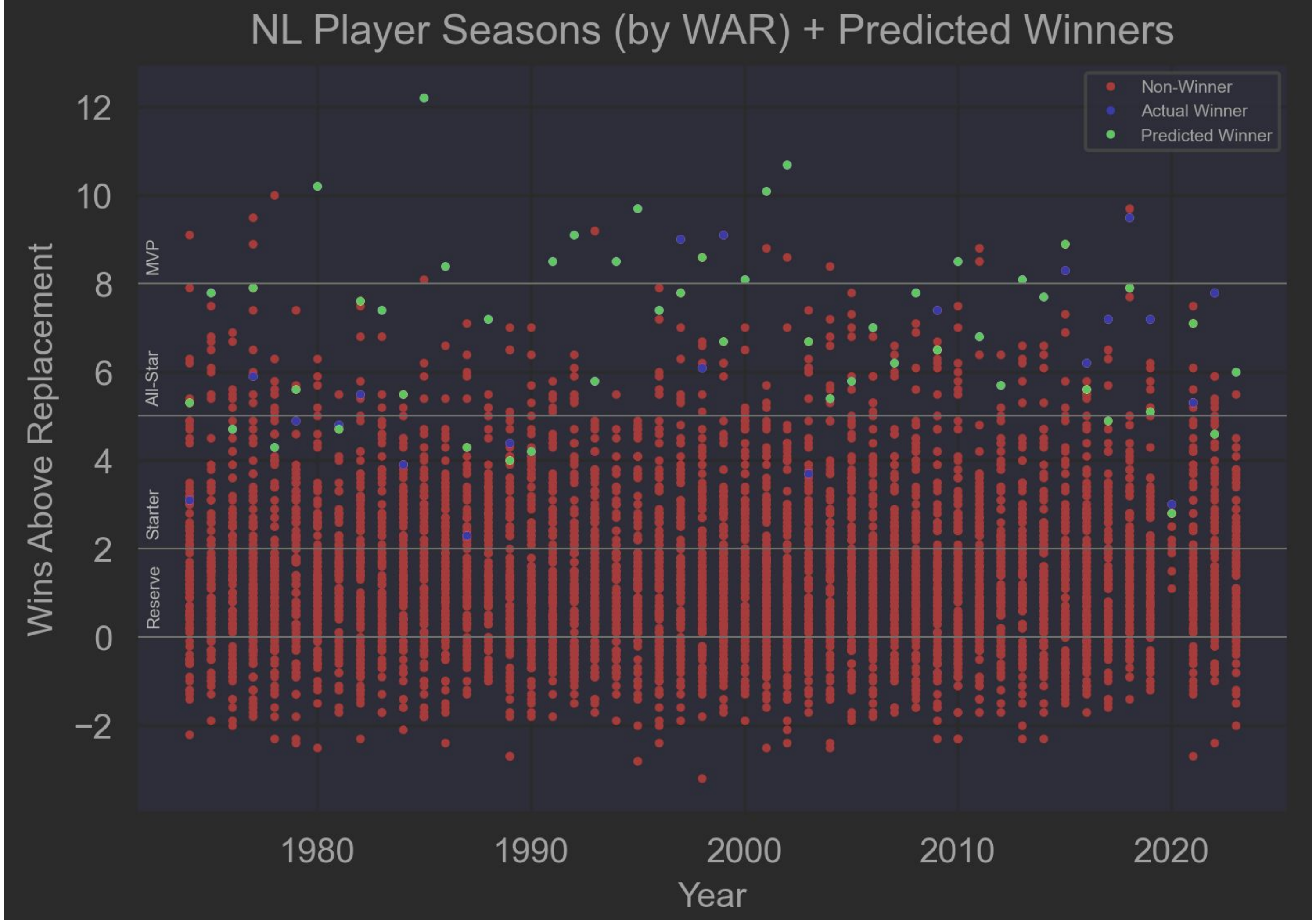
    Notably, this formula does not use any advanced statistics
- We replicate this formula, but replace the “Victory Bonus” with a bonus of 12 if they are in the top 3 teams in their league (there are 3 divisions per league)
  - No good data source available on who won each division
- We now have 2 models to compare the performance of

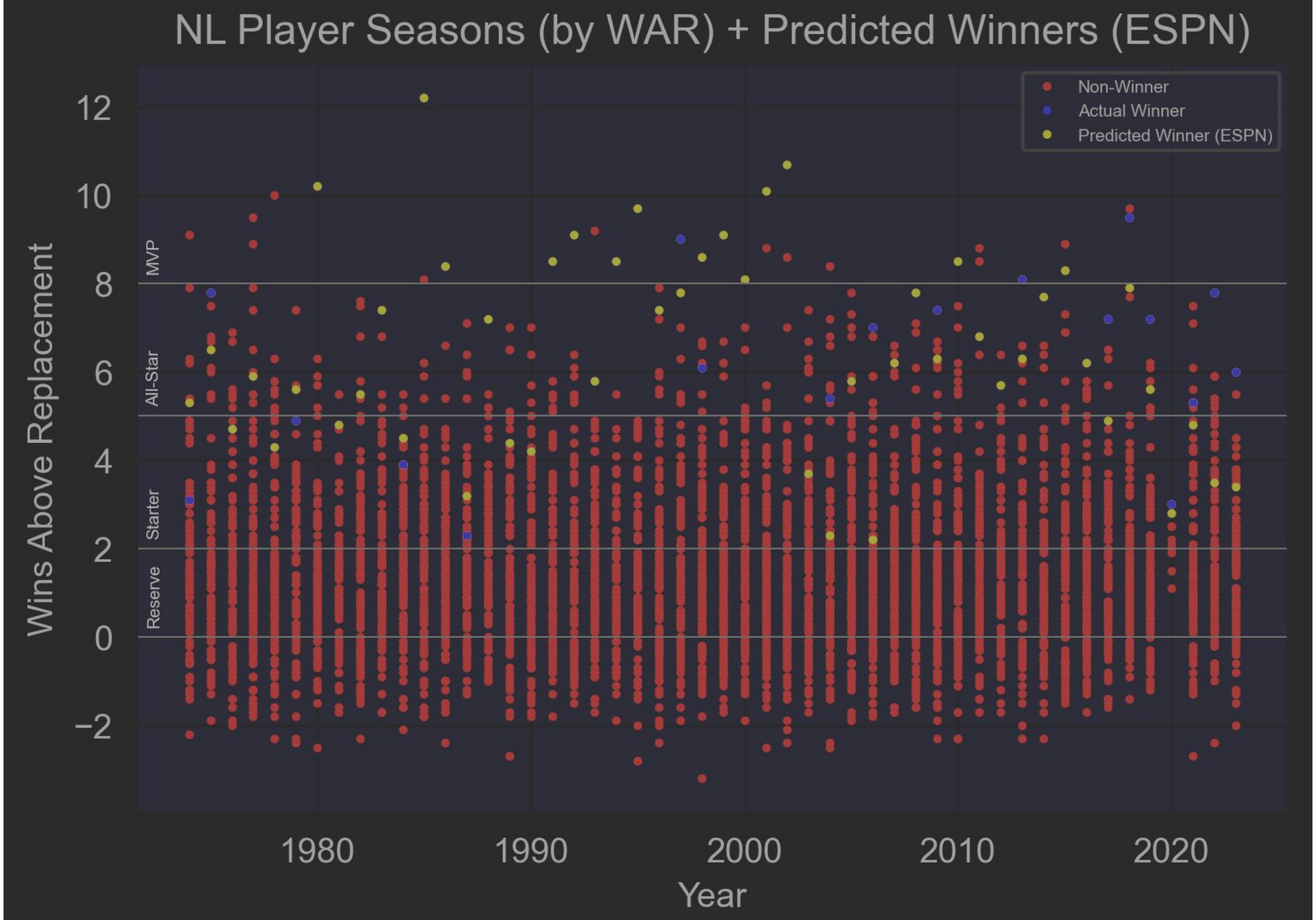

NL Player Seasons (by ERA)
ERA (Lower is better)
10
8
6
4
2
1980
1990
2000
2010
2020
Year
● Non-Winner
● Actual Winner

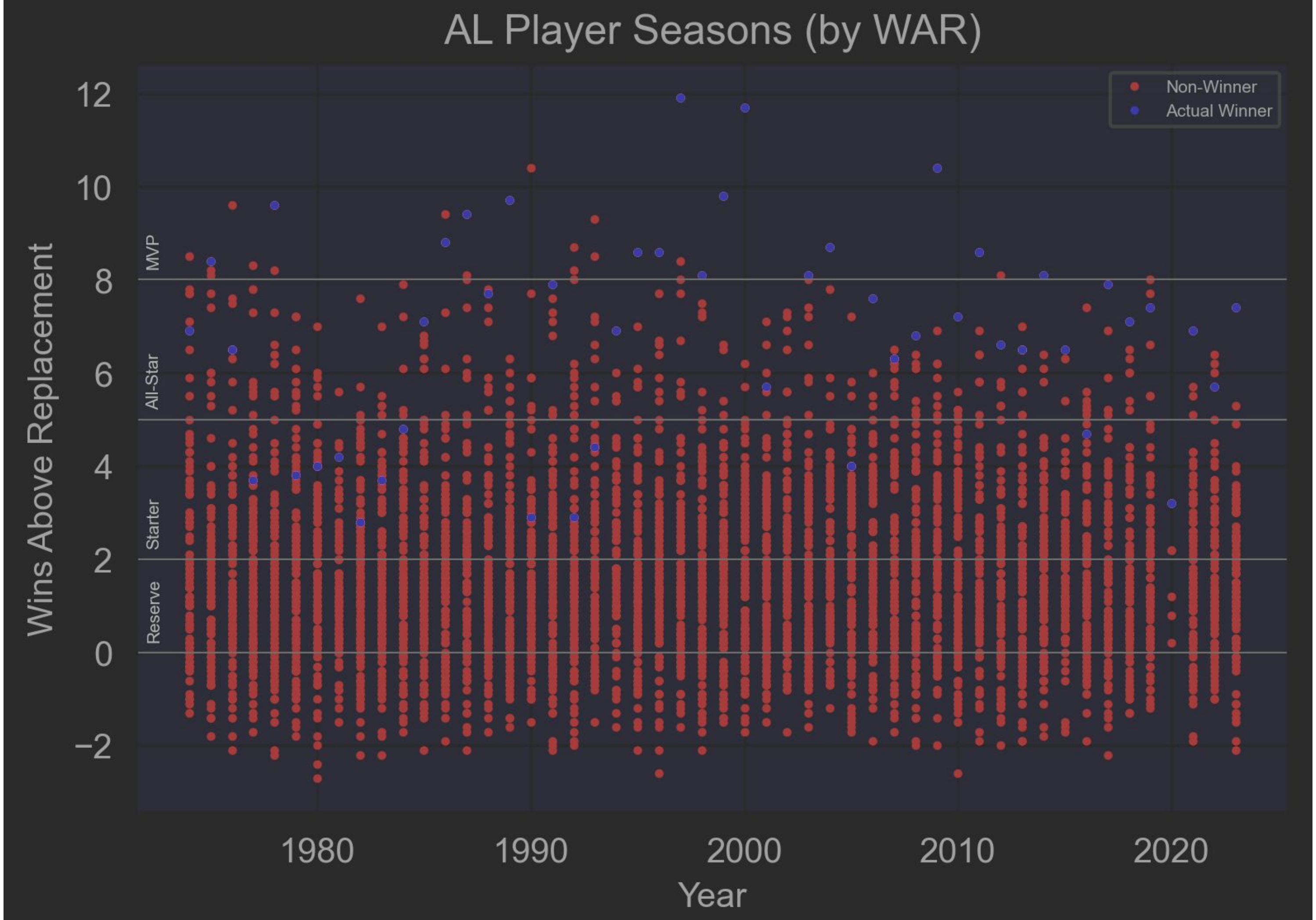

NL Player Seasons (by ERA) + Predicted Winners
Non-Winner
Actual Winner
Predicted Winner
ERA (Lower is better)
10
8
6
4
2
1980
1990
2000
2010
2020
Year

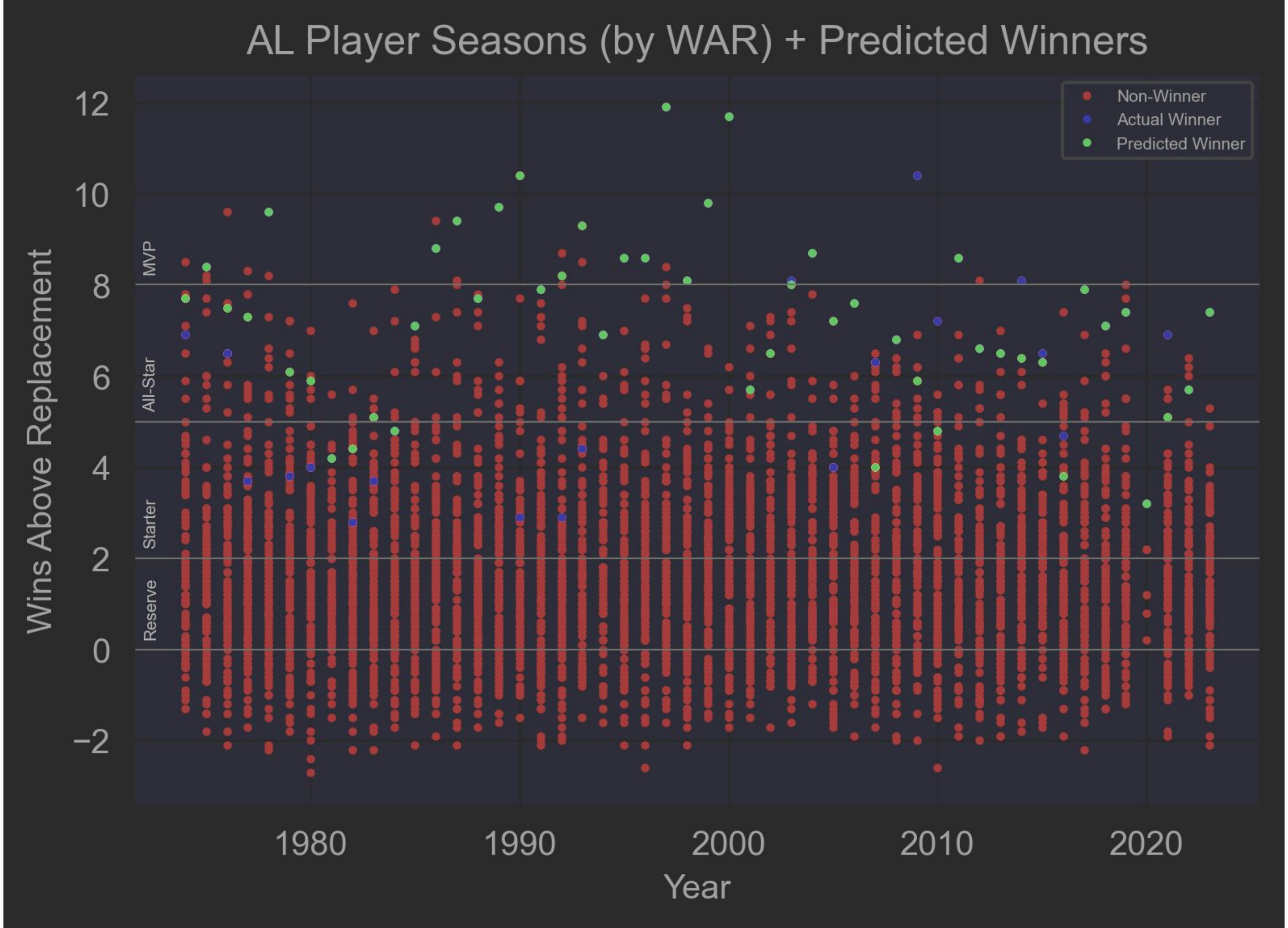

NL Player Seasons (by ERA) + Predicted Winners (ESPN)
ERA (Lower is better)
Year
10
8
6
4
2
1980
1990
2000
2010
2020
● Non-Winner
● Actual Winner
● Predicted Winner (ESPN)

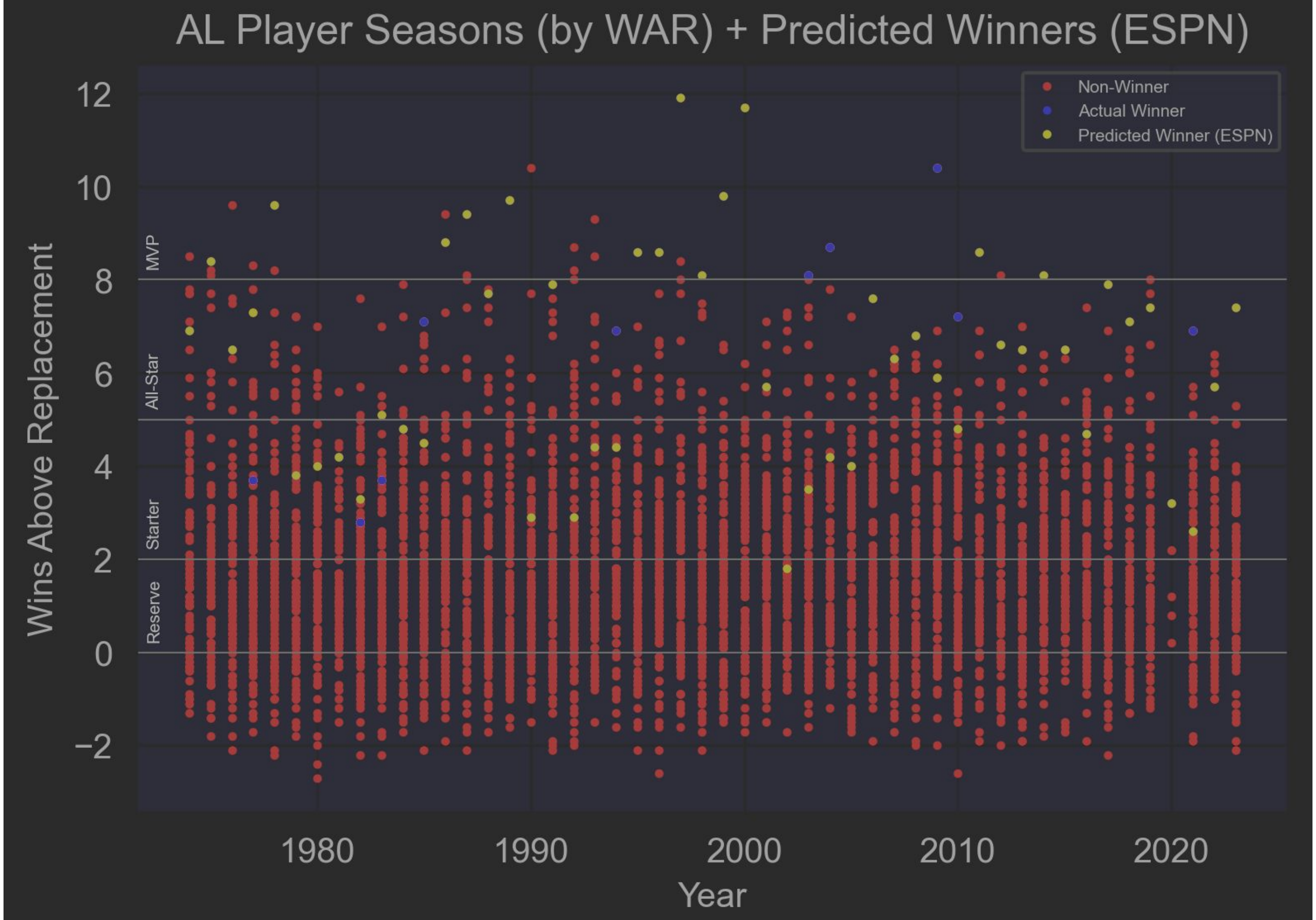

NL Player Seasons (by WAR)
Non-Winner
Actual Winner
Wins Above Replacement
MVP
All-Star
Starter
Reserve
12
10
8
6
4
2
0
−2
1980
1990
2000
2010
2020
Year


NL Player Seasons (by WAR) + Predicted Winners
Non-Winner
Actual Winner
Predicted Winner
Wins Above Replacement
MVP
All-Star
Starter
Reserve
12
10
8
6
4
2
0
−2
1980
1990
2000
2010
2020
Year


NL Player Seasons (by WAR) + Predicted Winners (ESPN)
Non-Winner
Actual Winner
Predicted Winner (ESPN)
Wins Above Replacement
MVP
All-Star
Starter
Reserve
12
10
8
6
4
2
0
−2
1980
1990
2000
2010
2020
Year


AL Player Seasons (by WAR)
Non-Winner
Actual Winner
Wins Above Replacement
12
10
8
6
4
2
0
−2
MVP
All-Star
Starter
Reserve
1980
1990
2000
2010
2020
Year


AL Player Seasons (by WAR) + Predicted Winners
Non-Winner
Actual Winner
Predicted Winner
Wins Above Replacement
MVP
All-Star
Starter
Reserve
12
10
8
6
4
2
0
−2
1980
1990
2000
2010
2020
Year


AL Player Seasons (by WAR) + Predicted Winners (ESPN)
Non-Winner
Actual Winner
Predicted Winner (ESPN)
Wins Above Replacement
MVP
All-Star
Starter
Reserve
12
10
8
6
4
2
0
−2
1980
1990
2000
2010
2020
Year

Model Performance

Our model (on unseen test data):

- Precision: 0.36
- Recall: 0.72
- Accuracy: 0.985 (mostly meaningless because data is imbalanced)
- Area under ROC curve: 0.984

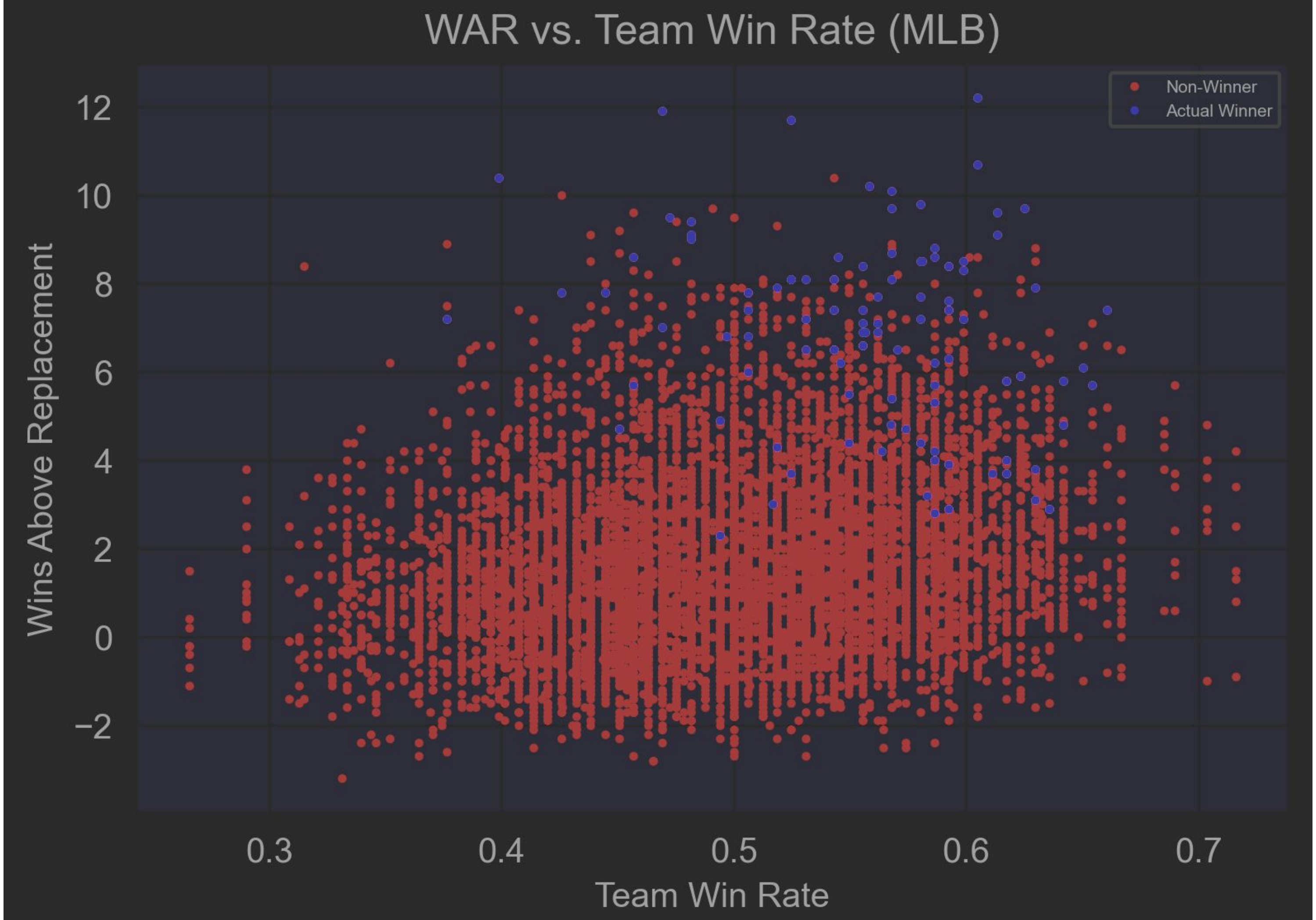
ESPN (on unseen test data):

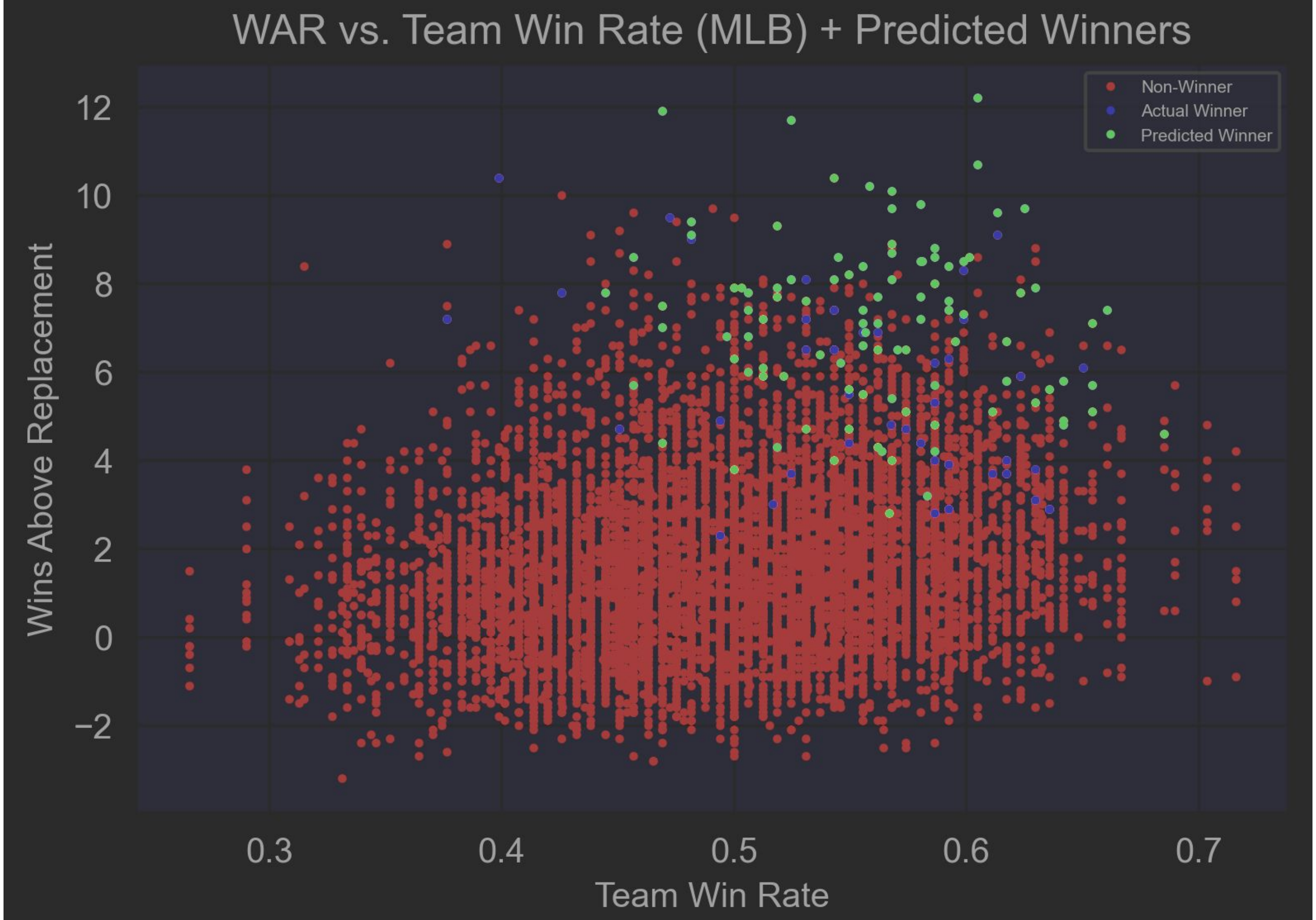
- Precision: 0.36
- Recall: 0.72
- Accuracy: 0.985
- Area under ROC curve: 0.973

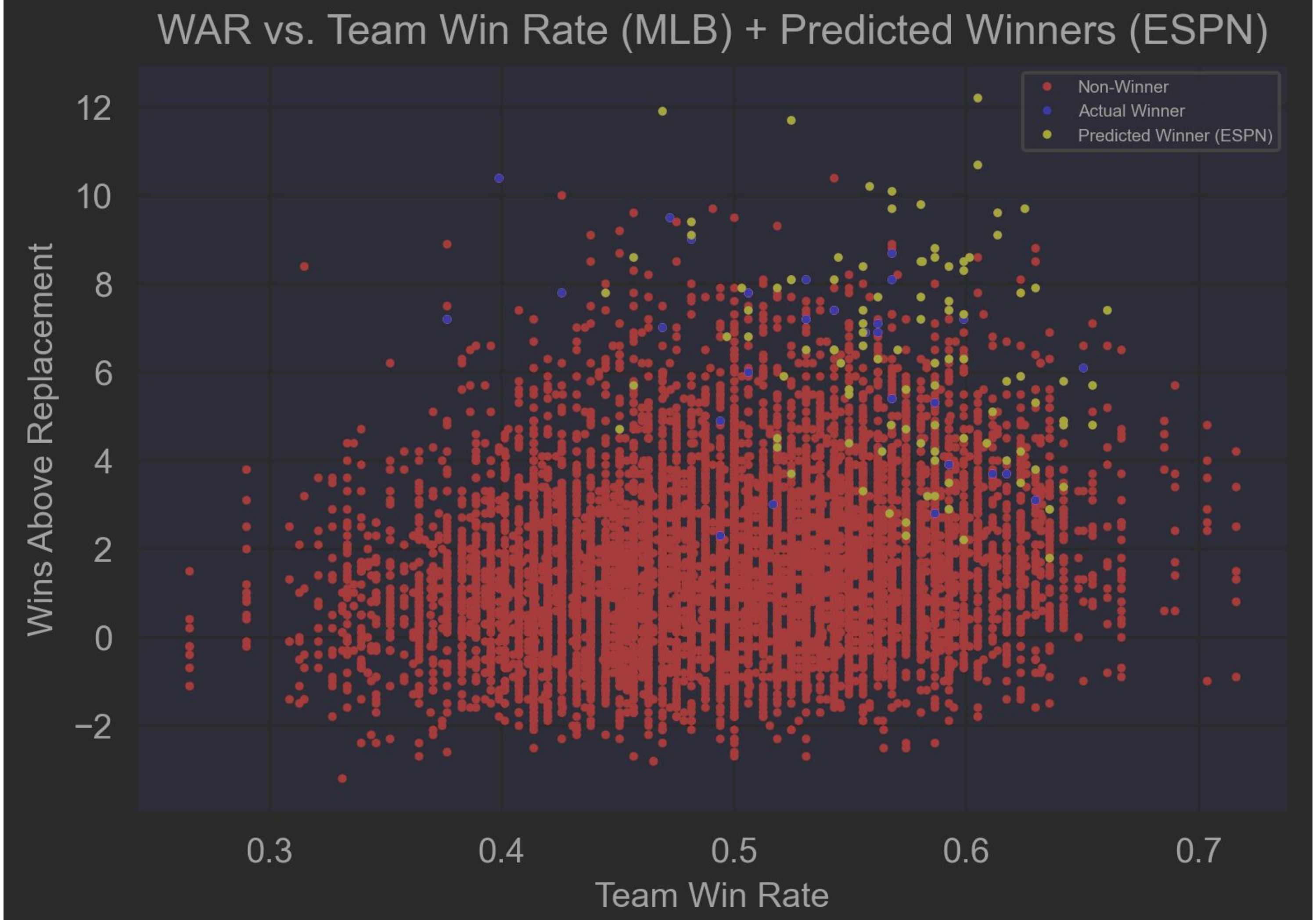
Performance was incredibly similar

One more question:

Does performance of the team a pitcher is on unfairly influence who wins the Cy Young award?


WAR vs. Team Win Rate (MLB)
Non-Winner
Actual Winner
Wins Above Replacement
12
10
8
6
4
2
0
−2
Team Win Rate
0.3
0.4
0.5
0.6
0.7

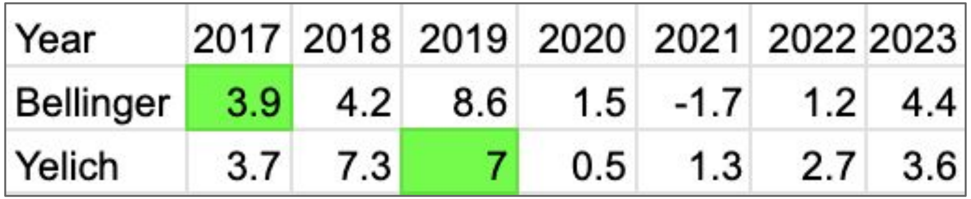

WAR vs. Team Win Rate (MLB) + Predicted Winners
Non-Winner
Actual Winner
Predicted Winner
Wins Above Replacement
12
10
8
6
4
2
0
−2
Team Win Rate
0.3
0.4
0.5
0.6
0.7


WAR vs. Team Win Rate (MLB) + Predicted Winners (ESPN)
● Non-Winner
● Actual Winner
● Predicted Winner (ESPN)
Wins Above Replacement
12
10
8
6
4
2
0
−2
Team Win Rate
0.3
0.4
0.5
0.6
0.7

Answers:

- Cy Young Voters aren’t very good at picking high-value pitchers that are on bad teams
- They like voting for pitchers that are on teams that are winning
- Since models can only reflect the results of these votes, they also both emphasize team win rate

“Taking it Home”



Year	2017	2018	2019	2020	2021	2022	2023
Bellinger	3.9	4.2	8.6	1.5	-1.7	1.2	4.4
Yelich	3.7	7.3	7	0.5	1.3	2.7	3.6

- Baseball is random
  - “No two plays are the same”
- Baseball is volatile
  - Green highlight is MVP year
  - One of the most telling statistics of a season, WAR, is unpredictable between seasons
- WAR and WPA are great predictors of player impact
  - Their effectiveness at predicting who wins Cy Young is good, but limited, because voters like pitchers that are on winning teams
- Cy Young voters cannot be completely modeled using just statistics
  - There are likely intangibles that lead them to voting one way or another, that cannot be quantified through a model like this