#HASHTAGS & HOME RUNS

By Leo Gentile, Leon Ma, and Maya Singh



Research Question and Motivation

- NCAA athletes can now profit from NIL (since 2021)
  - Social media is a major driver of NIL value
- NIL may impact players' decisions to go pro or stay in college
  - NIL valuation data is often private or paywalled
- Social media following serves as a strong, accessible gauge
- Combining our interests in sports, data, and online influence

Research Question:

What drives the social media following of D1 college baseball players?

Data Set and Sources

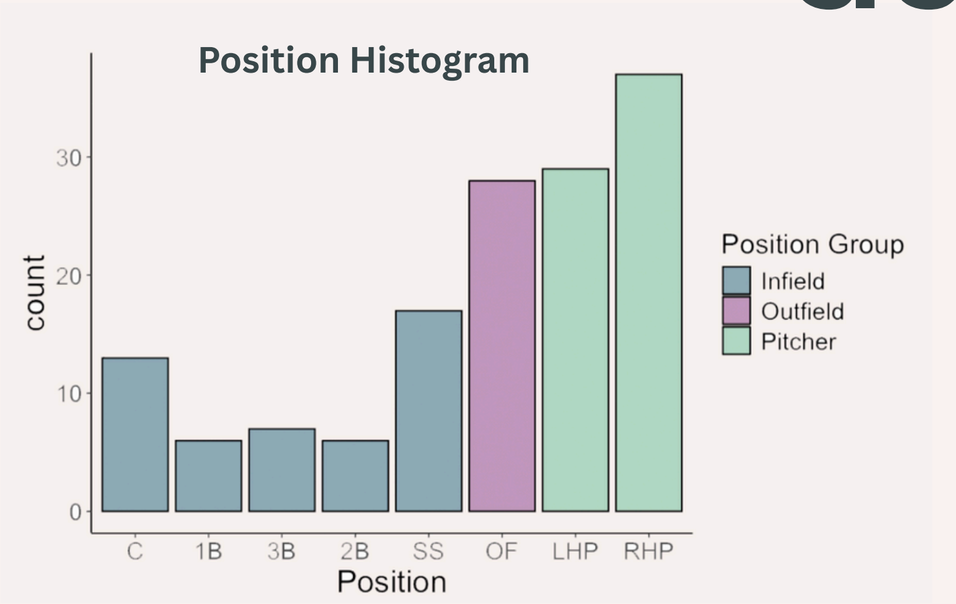
- 150 players from D1Baseball.com’s Top 150 MLB Draft Prospects (2024)
  - Final dataset: 144 players (removed private Instagram accounts)
- Stats from D1Baseball, Baseball-Reference, and team websites
- Chose D1Baseball for consistent, reputable rankings
- Used most recent full season for accurate performance data

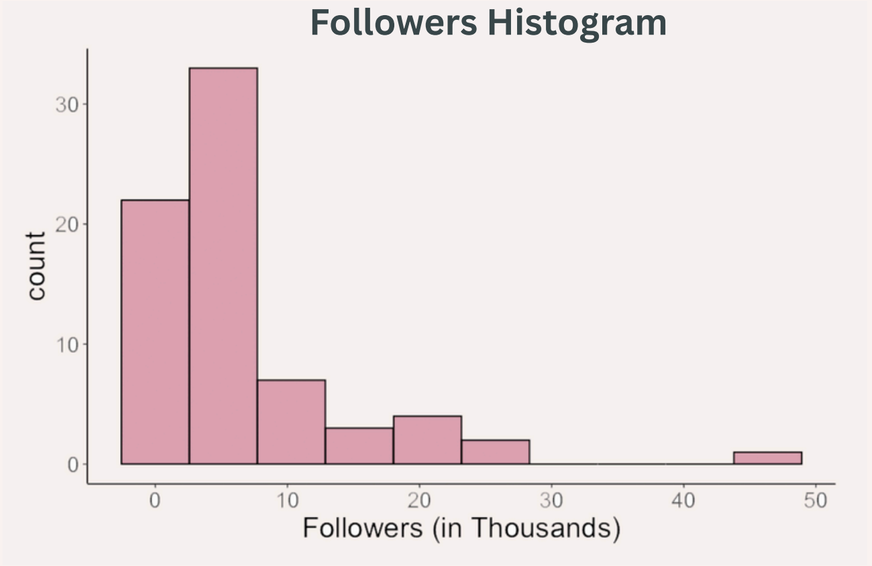
Variables & Cleaning Process

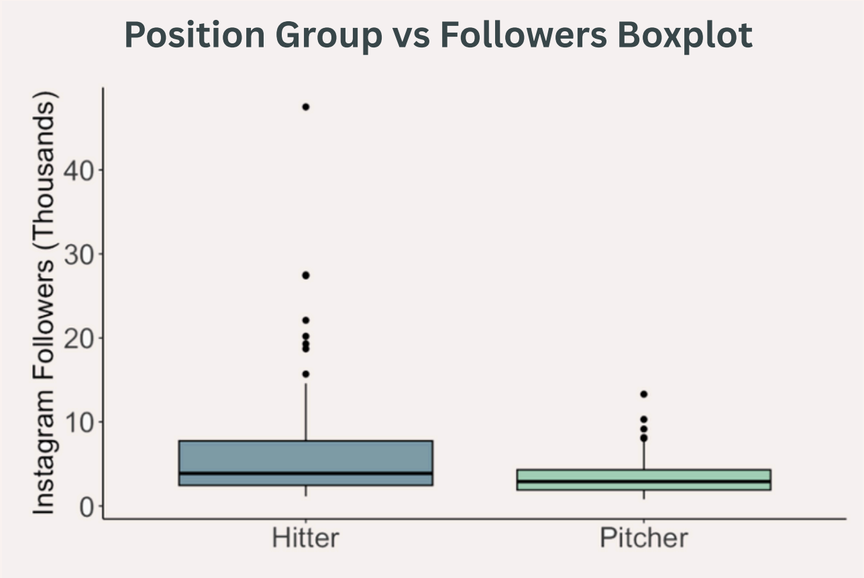
- On-field performance: BA, OPS, ERA, strikeouts, home runs, etc.
- Social Media: Instagram followers (in thousands), posts, account age
- Demographic: class year, team, Power 5 status, race, height
- Manually added missing data (e.g., height, race, transfer status)

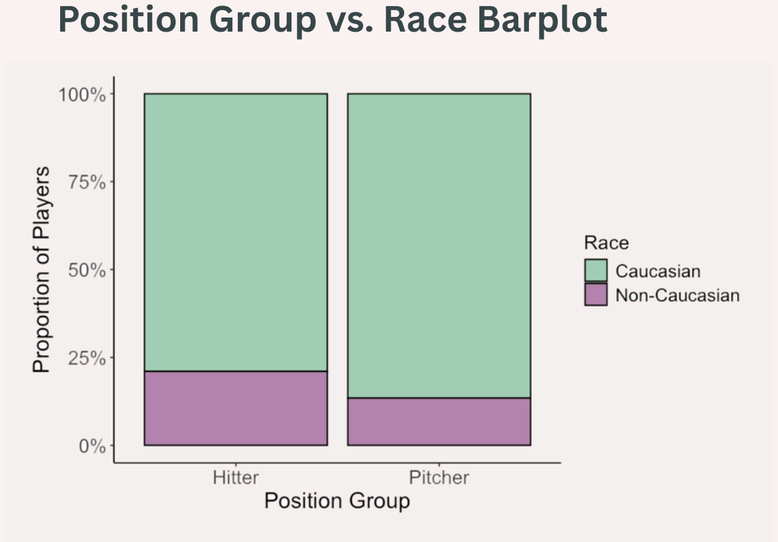
Final dataset: 23 variables across 144 players

General EDA


Position Histogram
count
30
20
10
0
C
1B
3B
2B
SS
OF
LHP
RHP
Position
Position Group
Infield
Outfield
Pitcher

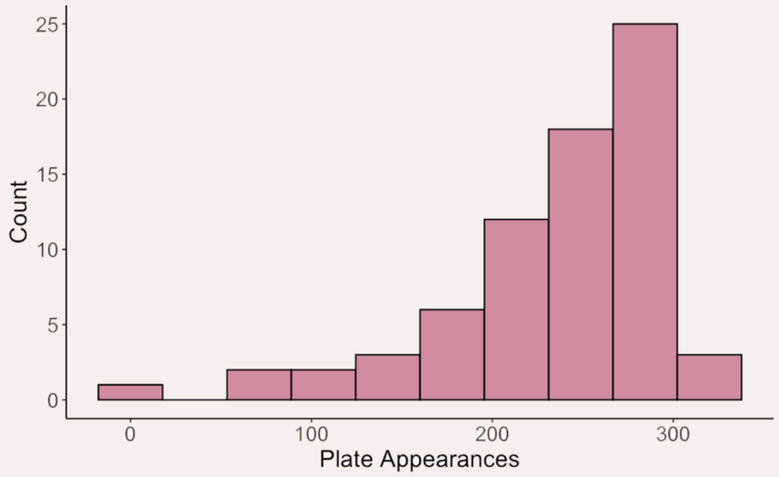

Followers Histogram
count
30
20
10
0
0
10
20
30
40
50
Followers (in Thousands)


Position Group vs Followers Boxplot
Instagram Followers (Thousands)
40
30
20
10
0
Hitter
Pitcher

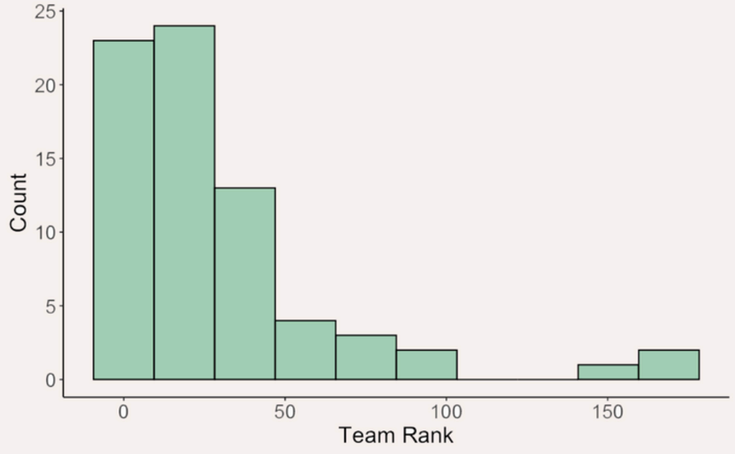

Position Group vs. Race Barplot
Proportion of Players
100%
75%
50%
25%
0%
Hitter
Pitcher
Position Group
Race
Caucasian
Non-Caucasian

Hitters EDA

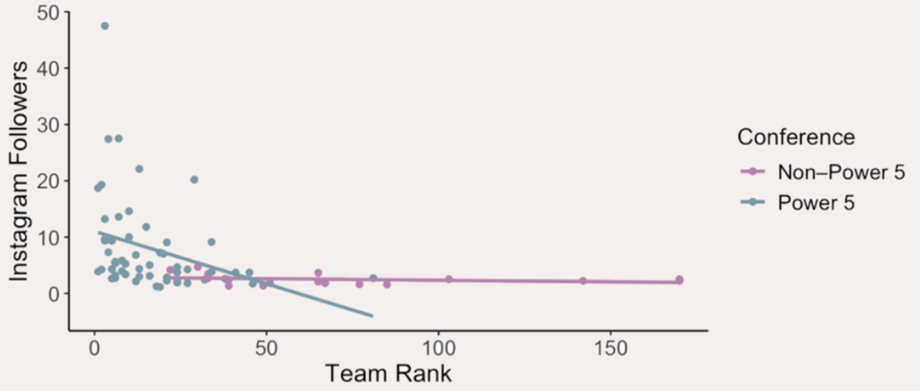
Plate Apperances Histogram



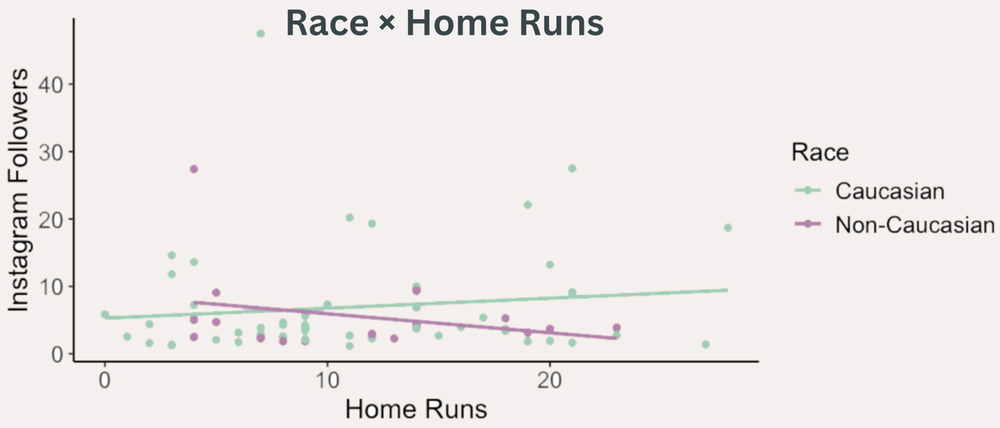
Team Rank Histogram



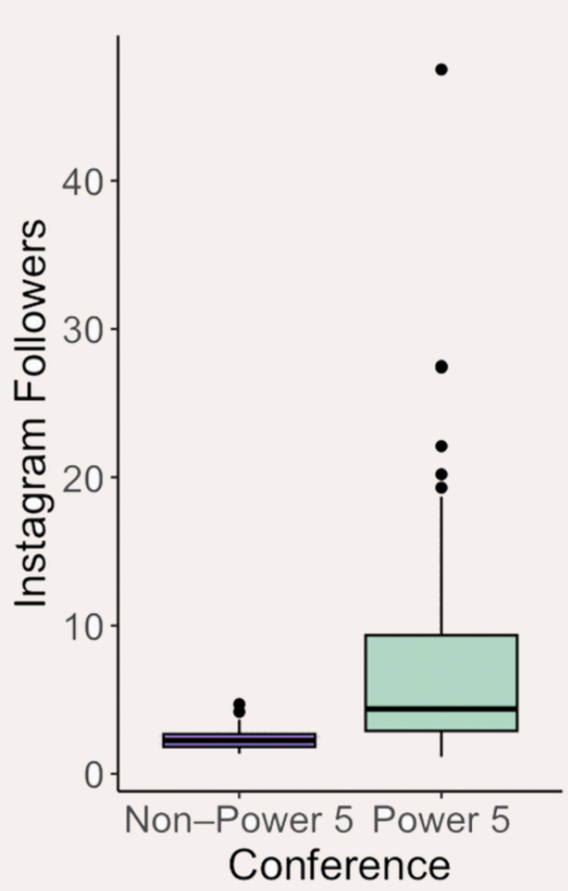
Team Rank × Conference Type



Race × Home Runs

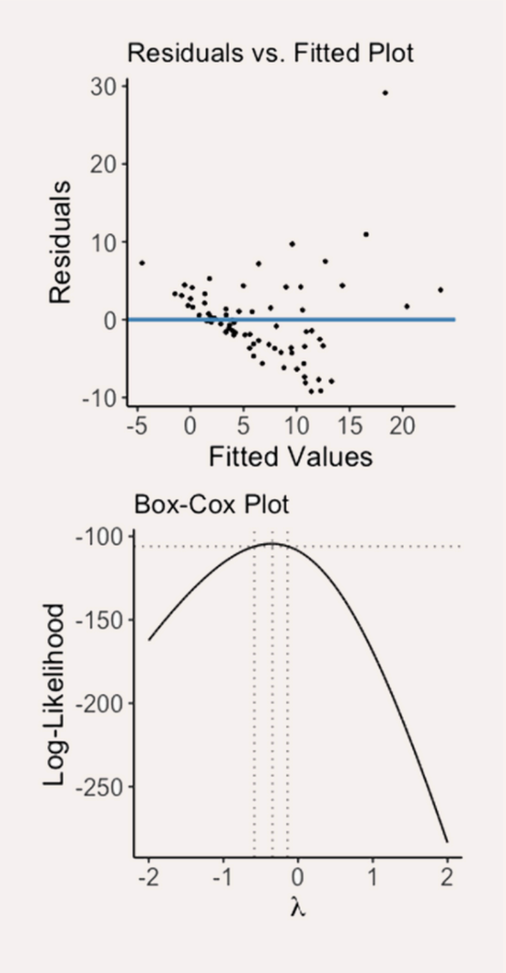


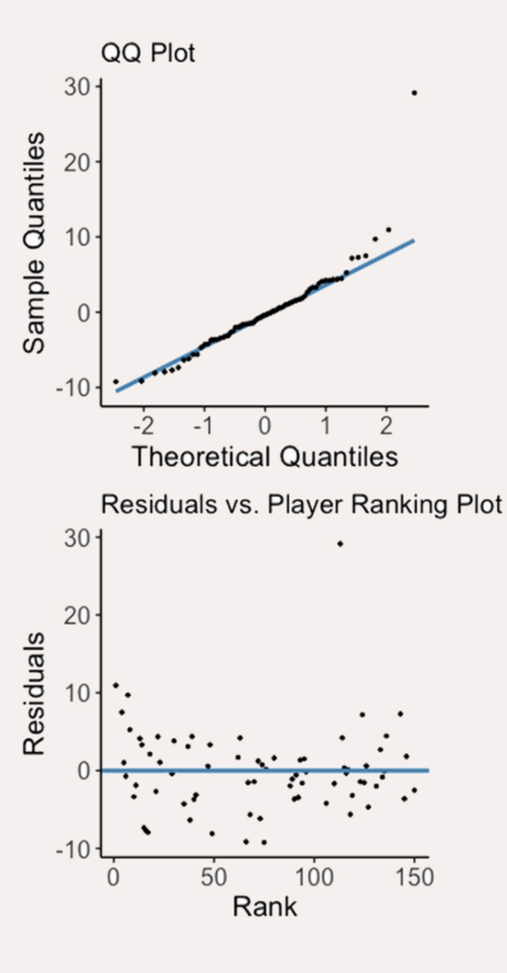
Conference vs. Followers Boxplot



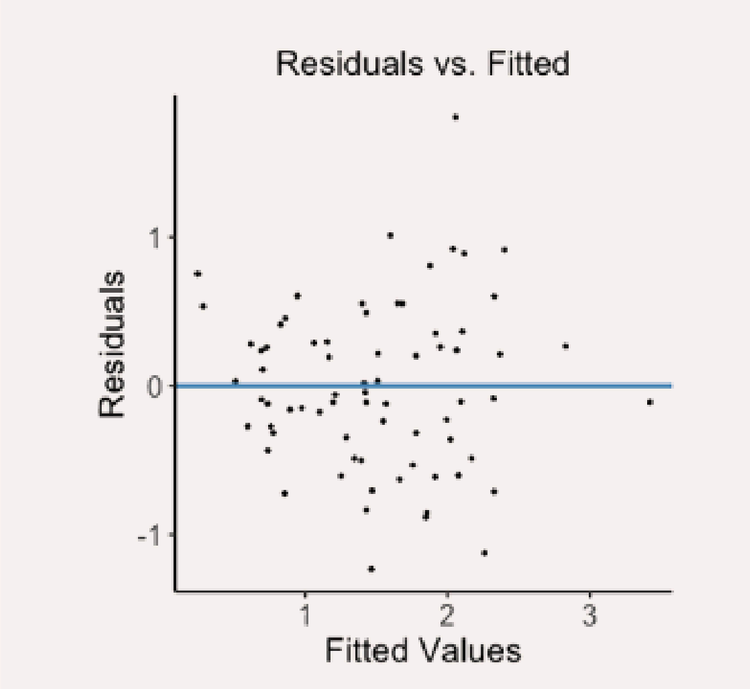
Hitters Initial Model

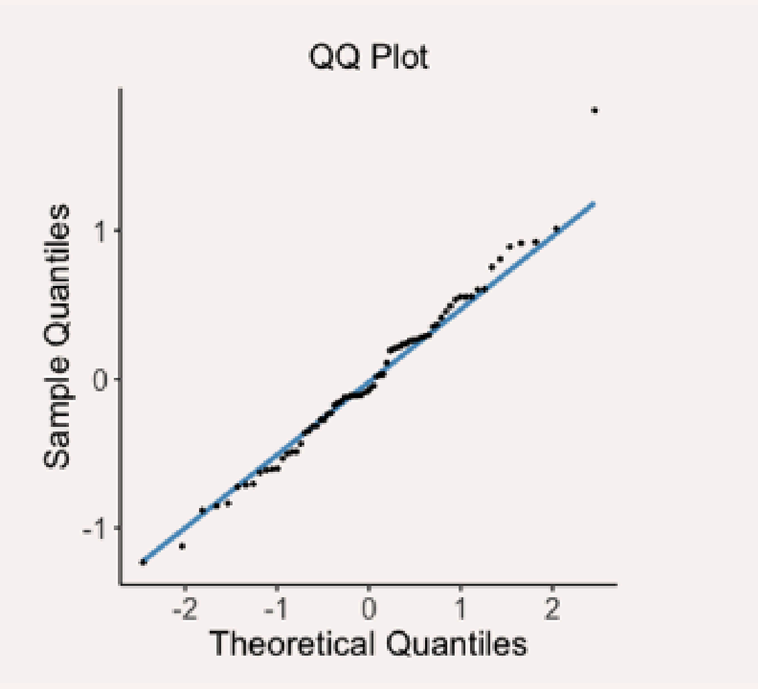
- Interactions Included
  - Caucasian × Home Runs
  - Height × Home Runs
  - Height ×Caucasian
  - Power5 × Team Ranking
- Collapsing Categorical Variables
  - Position: Infield vs. Outfield
  - Race: Caucasian vs. Non-Caucasian
  - Power 5 to make up for Conference
- Assumption Violations
  - Non-linearity in Prospect Rank vs. Residuals
  - Non-constant variance of errors





Diagnostics for Final Model





$$\log(\widehat{\text{Followers}}) = 13.152 + 0.889 \cdot \text{PowerFive}_{\text{Yes}} - 0.001 \cdot \text{TeamRank} - 0.160 \cdot \text{Height} - 12.625 \cdot \text{Caucasian}_{\text{Yes}} - 2.225 \cdot \text{BA} - 0.044 \cdot \text{HomeRuns} + 0.013 \cdot \text{Posts}$$

$$+ 0.003 \cdot \text{Months} + 0.056 \cdot (\text{Caucasian}_{\text{Yes}} \cdot \text{HomeRuns}) + 0.163 \cdot (\text{Height} \cdot \text{Caucasian}_{\text{Yes}}) - 0.016 \cdot (\text{PowerFive}_{\text{Yes}} \cdot \text{TeamRank})$$

Final Model

- After both direction stepwise selection and transformations
- Adjusted R^2 = 0.4825

Variable	Estimate	Std. Error	p-value	Lower	Upper
(Intercept)	13.152	6.969	0.063	-0.789	27.094
Power 5 (Yes)	**0.889**	**0.309**	**0.005**	**0.271**	**1.508**
Team Rank	-0.001	0.003	0.576	-0.008	0.004
Height	**-0.16**	**0.093**	**0.093**	**-0.347**	**0.027**
Caucasian (Yes)	**-12.625**	**7.491**	**0.097**	**-27.611**	**2.36**
BA	-2.225	1.67	0.187	-5.566	1.115
Home Runs	**-0.044**	**0.026**	**0.093**	**-0.096**	**0.007**
Posts	**0.013**	**0.003**	**0.0001**	**0.006**	**0.019**
Months	0.003	0.002	0.14	-0.001	0.009
Caucasian:Home Runs	**0.056**	**0.028**	**0.057**	**-0.001**	**0.114**
Height:Caucasian	0.163	0.101	0.113	-0.04	0.367
Power 5:Team Rank	**-0.016**	**0.006**	**0.022**	**-0.03**	**-0.002**

Followers Per Post Exploration

Variable	Estimate	Std. Error	p-value
(Intercept)	14.057	8.033	0.085
Power 5 (Yes)	0.233	0.268	0.387
Team Rank	-0.005	0.003	0.088
Transfer (Yes)	0.305	0.218	0.167
Height	-0.183	0.108	0.094
Caucasian (Yes)	-14.503	8.579	0.095
BA	-6.068	1.891	0.002
Home Runs	-0.051	0.03	0.095
Caucasian:Home Runs	0.068	0.033	0.044
Height:Caucasian	0.186	0.116	0.116

Adjusted R-squared = 0.2481

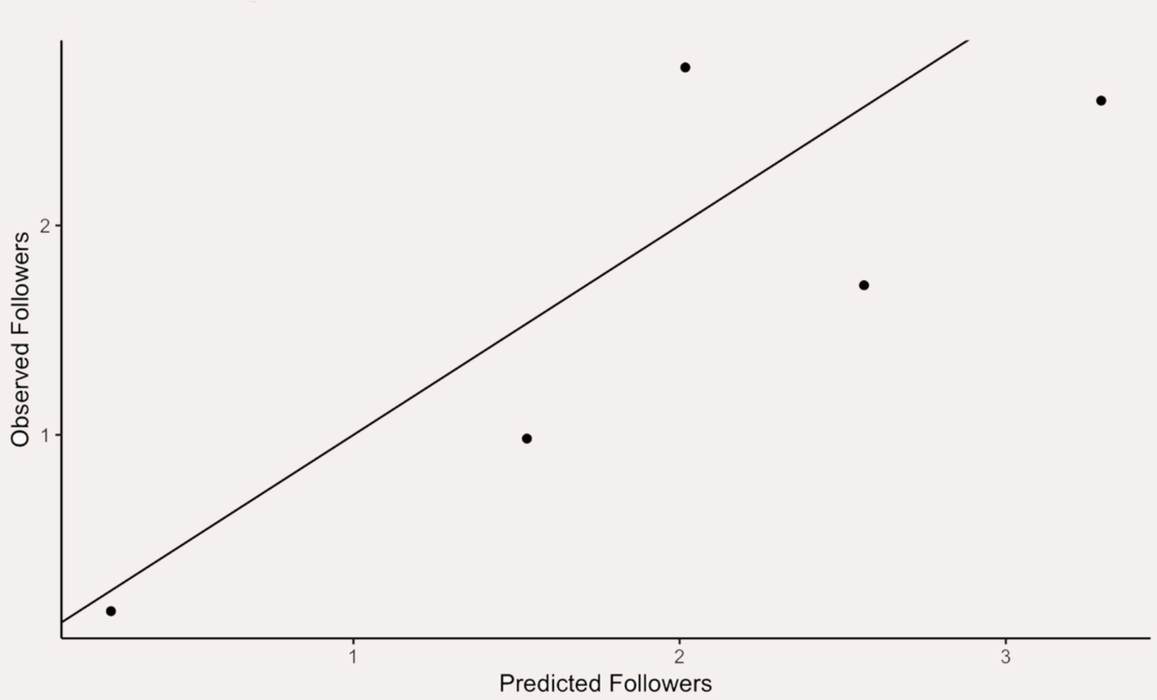
Response Variable: log(Followers/Post)

Testing Model Performance and Prediction: Hitters

Player	Model Prediction	Confidence Interval	Prediction Interval	Actual Point
Luke Stevenson	4.178	(3.092, 5.646)	(1.212, 14.396)	6.827
Blake Cyr	7.814	(5.573, 10.957)	(2.247, 27.33)	47.46

Standard Player: Luke Stevenson





$R^2 = 0.6012$

Outlier: Blake Cyr

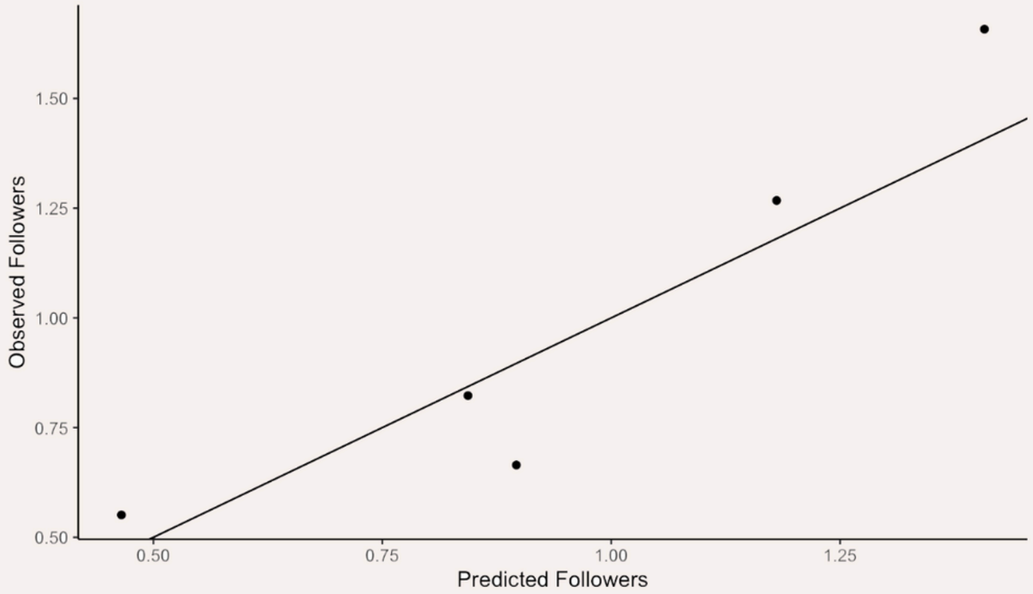


Testing Model Performance and Prediction: Pitchers

Player	Model Prediction	Confidence Interval	Prediction Interval	Actual Point
Matt Scott	4.125	(2.968, 5.725)	(1.490, 11.404)	3.54
Chase Shores	5.709	(3.838, 8.482)	(2.013, 16.167)	13.3

Standard Player: Matt Scott





$R^2 = 0.837$

Outlier: Chase Shores



Conclusion

- Significant predictors (0.05):
  - Posts
  - Power 5:Team Rank

- Significant predictors (0.1):
  - Height
  - Caucasian
  - Home Runs
  - Caucasian:Home Runs

- Testing models performed fairly well

- Final Adjusted-R^2 value of 0.4825 for hitters model and 0.4379 for pitchers model

- Some controversial effects

- Visibility is the main driver

Limitations and Potential Expansion

- Potentially important, but difficult to measure variables:
  - “Attractiveness”
  - Team USA Participation

- Expansion on Significant Effect of Posts:
  - We did Followers/Posts model already
  - Topics of Posts: Sponsored/Ads?
  - Collab Posts with larger accounts
  - Likes per Post /Engagement Rates

- Larger Data Set: Use More D1 Players

Thank You!