Reds Hackathon: Predicting Playing Time

Kevin Baer

About Me

- Second-year student @ UCLA studying Data Theory
- Consultant with UCLA Football
- Presented volleyball research at CMSAC 2024
- Big Data Bowl projects have been read by NFL fr

  office executives and even a head coach!

Starting Out

- Use 2022 and 2023 data to predict 2024
- train/test using 2021 and 2022 data to predict 2023

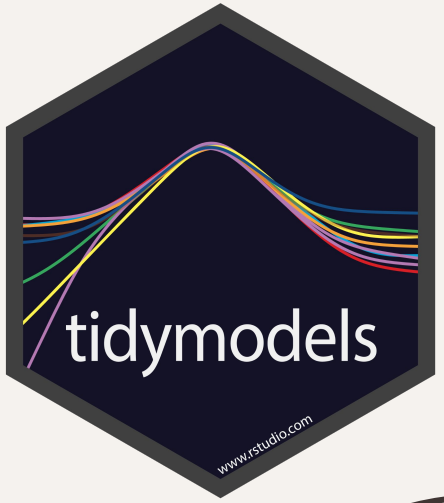
Pros:

- Typical ML process, creates nice tabular data
- No reason why this shouldn't work!

Cons:

- Data loss of 2021 when predicting 2024

ML Model Infrastructure

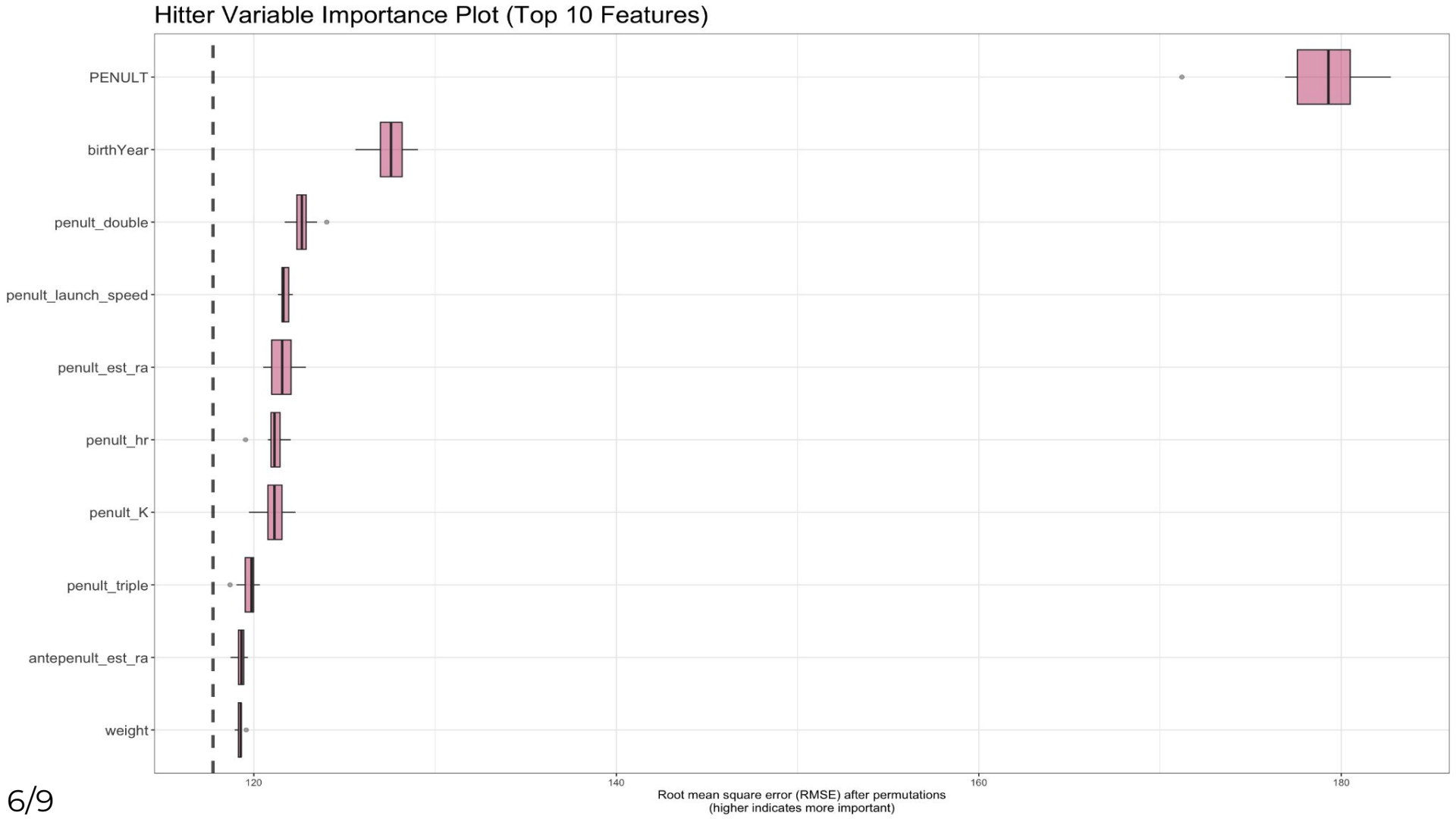


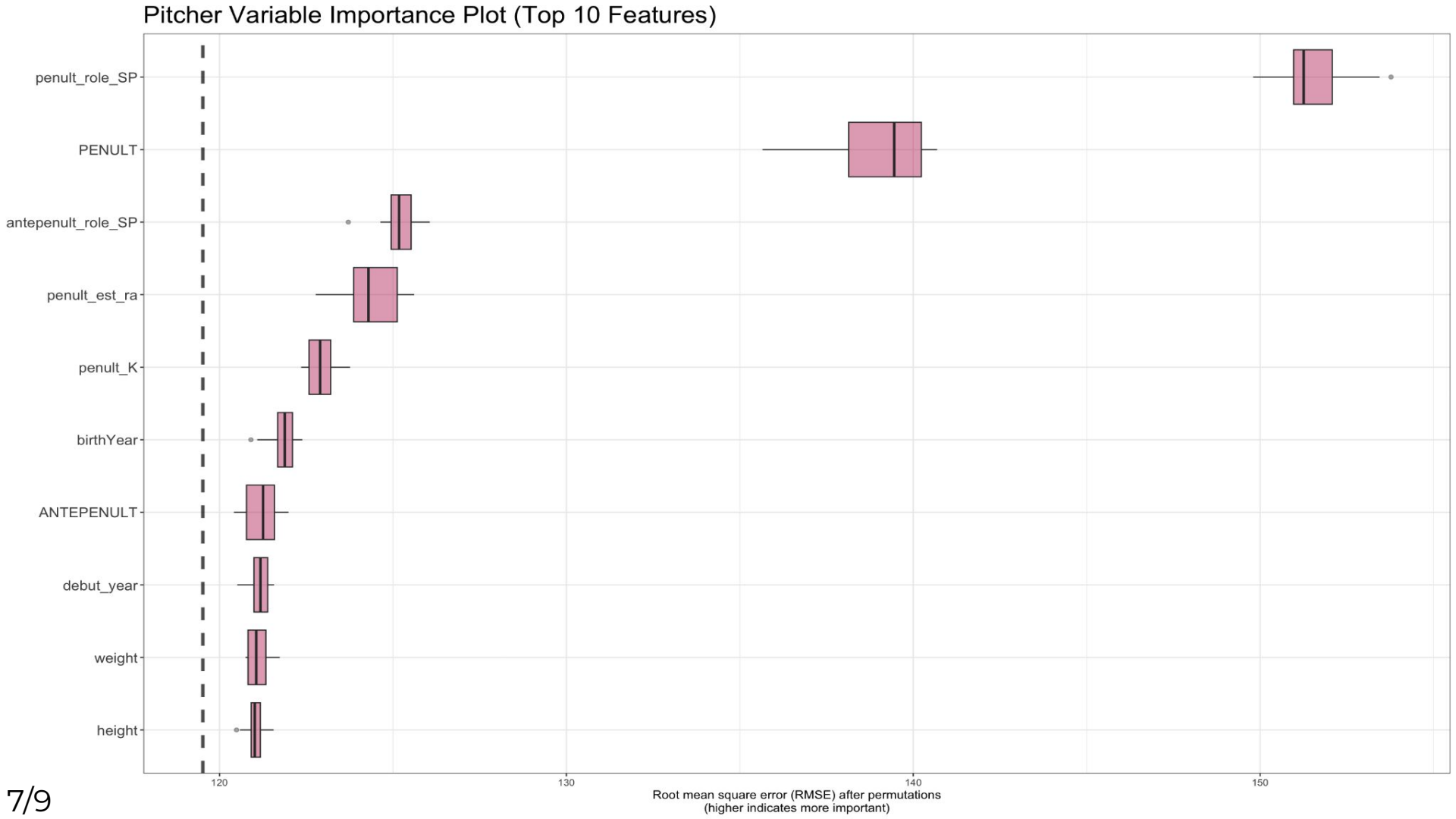
- Tidymodels (by Max Kuhn)
- Two models - one for hitters, one for pitchers
- XGBoost won out
- 75/25 split stratified on playing time
- 5 fold cross validation
- penult_ and antepenult_ prefixes for different years

Feature Engineering

- Some Lahman data (height, weight, throws/bats, debut year, birth year, birth country)
- Penultimate and antepenultimate numbers for various relevant performance statistics, as well as overall playing time
- Means of all numeric variables
- Frequency of dummy categories

Package ‘fastDummies’


Hitter Variable Importance Plot (Top 10 Features)
PENULT
birthYear
penult_double
penult_launch_speed
penult_est_ra
penult_hr
penult_K
penult_triple
antepenult_est_ra
weight
120
140
160
180
Root mean square error (RMSE) after permutations
(higher indicates more important)


Pitcher Variable Importance Plot (Top 10 Features)
penult_role_SP
PENULT
antepenult_role_SP
penult_est_ra
penult_K
birthYear
ANTEPENULT
debut_year
weight
height
120
130
140
150
Root mean square error (RMSE) after permutations
(higher indicates more important)

Name (Pos)	My Hackathon Predictions	FanGraphs Depth Chart Projection	Difference
Martinez, Nick (SP)	400	707	-307
Friedl, TJ (OF)	317	567	-250
Abbott, Andrew (SP)	404	652	-248
Richardson, Lyon (SP)	207	66	+141
Ashcraft, Graham (RP)	387	228	+159
Benson, Will (OF)	340	119	+221

names	prediction	Fangraphs	delta
Martinez, Nick	400	707	-307
Friedl, TJ	317	567	-250
Abbott, Andrew	404	652	-248
Hays, Austin	297	490	-193
Greene, Hunter	517	702	-185
Candelario, Jeimer	390	532	-142
Fraley, Jake	321	448	-127
De La Cruz, Elly	559	665	-106
Encarnacion-Strand, Christian	246	350	-104
Suter, Brent	168	272	-104
Santillan, Tony	159	251	-92
Lux, Gavin	415	490	-75
Rogers, Taylor	209	269	-60
Pagan, Emilio	207	267	-60
Barlow, Scott	199	249	-50
Trevino, Jose	215	262	-47
Diaz, Alexis	231	276	-45
Singer, Brady	678	717	-39
Moll, Sam	200	236	-36
Fairchild, Stuart	234	252	-18
Lodolo, Nick	507	522	-15
Stephenson, Tyler	505	490	15
Steer, Spencer	597	574	23
Espinal, Santiago	273	203	70
Marte, Noelvi	249	154	95
Spiers, Carson	329	199	130
Richardson, Lyon	207	66	141
Ashcraft, Graham	387	228	159
Benson, Will	340	119	221

Thank You!
Any Questions?

GitHub Repository with code and citations

LinkedIn and Email

Can find this all on my website

(very much in progress)

kevbaer.github.io