QTM 347 Machine Learning I

Predicting MLB Gold Glove Winners

ANNA CHOI

Problem and Relevance

The MLB Gold Glove Award is given annually to baseball players who are considered the best defensive players at their respective fielding positions in each league (American and National).

The project goal is to **predict the winners of the MLB Gold Glove Award** for each position across both the American and National Leagues by utilizing position-specific and league-specific stats and metrics.

This problem involves:

- understanding complex fielding data and translating them into predictive models that can highlight defensive excellence
- benefiting teams in scouting, development, and strategic planning
- customizing predictions for different fielding positions and leagues

Approaches

Initial Objective – **Deterministic Classification**

- Predicts whether a player wins the Gold Glove Award as a binary outcome: 1 (Winner) or 0 (Non-Winner)

Changed the objective to predict probabilities instead of labels owing to **class imbalance**

- Only ~1% of rows represent winners, making it difficult for the models to predict "1" accurately
- The models overwhelmingly predict "0," resulting in high accuracy but poor utility for predicting winners

New Objective - **Probabilistic Classification**

- Captures the likelihood of each player winning rather than forcing a binary decision
- Provides more nuanced insights, allowing me to rank players based on their chances and identify top candidates beyond just the winner
- Even if a player is not predicted as a winner, a high probability (relative) highlights them as a strong candidate

Data Collection

Source of Data: Baseball-Reference

Collected fielding metrics for all MLB players across all 9 positions

- Pitcher, Catcher, First Base, Second Base, Third Base, Shortstop, Left Field, Center Field, Right Field
- Data from 2013 to 2024 since fielding metrics became more comprehensive since 2013

Created separate player and league datasets for each position

- Player Datasets: Individual player fielding metrics (e.g. putouts, assists, errors)
  - Each row represents a player's fielding performance for a single season
- League Datasets: Average fielding metrics for both National and American Leagues by season

Added the following columns for the classification task

- Season: The year in which the player's fielding performance is recorded
- Champion: If the player's team won the league championship of that season
- **Win: If the player won the Gold Glove Award for that position**

Data Overview

Player datasets

Approximately 30,000+ rows, spanning 12 seasons (2013–2024)

- Player Metrics
  - Games (G), Games Started (GS), Complete Games (CG), Innings (Inn), Chances (Ch), Putouts (PO), Assists (A), Errors (E), Double Plays (DP), etc.
  - Metrics differ slightly by position
- Player/Team Information
  - Age, Team (Tm), League (Lg), Season, Champion
- Target Variable
  - Win: Gold Glove Winner, represented as a binary outcome (0/1)

  * Three players were awarded for the first base position in 2018

League datasets

24 rows per position, representing league averages for 12 seasons

Data Cleaning and Preprocessing

Rows with missing or null values were removed to ensure only complete and accurate data were used for training and testing

- The fielding metrics are specific to each player's performance for a particular season
- Using mean or median values would introduce artificial data that does not accurately reflect the individual player's abilities or contributions

Duplicate player entries caused by FAs or trades during the season **were removed**

- The team assignment was prioritized by selecting the team with the most innings played, followed by the most games played if innings were equal, and finally the last team listed in the record if both were equal.

Dropped fielding metric columns not available in both player and league datasets for normalization

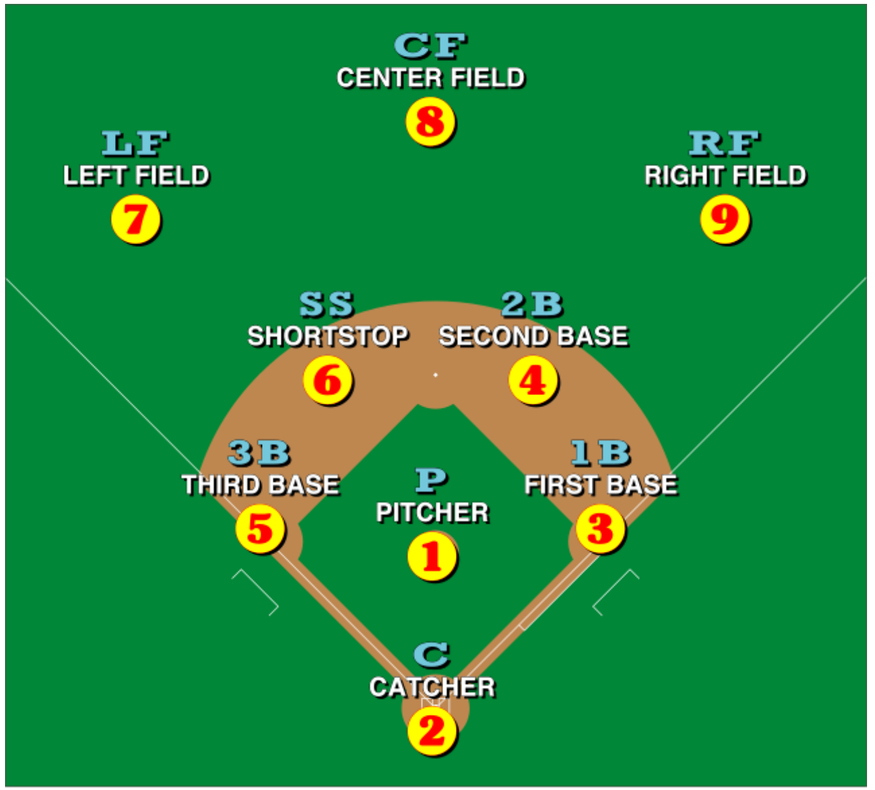
Used league datasets for normalization to account for seasonal trends and league disparities

- Fielding performance might have improved in recent years (e.g. better equipment, analytics)
- Metrics might vary between the National and American Leagues
- Retained the original value for normalization where division by zero occurred

Modeling Pt. 1

Used **5 classification models** for each of the 9 positions

- Random Forest
- Logistic Regression
- Gradient Boosting
- k-Nearest Neighbors (kNN)
- Support Vector Machine (SVM)



<u>Experimental Setup</u>

- Training Data (2013–2023)
  - Used data from previous 11 seasons to train the models
- Test Data (2024)
  - Reserved the most recent season’s data to evaluate real-world predictive performance

Modeling Pt. 2

Hyperparameter Tuning

- Used **GridSearchCV** to find the optimal parameters for each model
  - Random Forest: n_estimators, max_depth, min_samples_split, min_samples_leaf, bootstrap
  - Logistic Regression: penalty, max_iter
  - Gradient Boosting: n_estimators, learning_rate, max_depth, min_samples_split, min_samples_leaf
  - kNN: n_neighbors, weights, power parameter (p)
  - SVM: kernel, regularization parameter (C), gamma

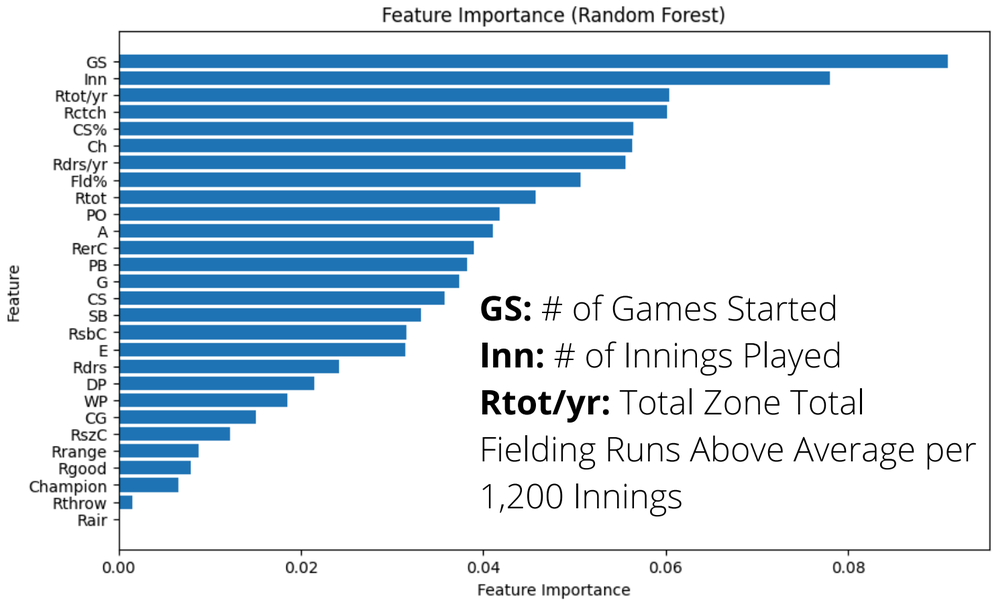
Model Evaluation Metric

- Focus on **AUC-ROC** for a better reflection of model performance on the minority class to address the issue of class imbalance
- Use AUC-ROC to measure the model's ability to distinguish between winners and non-winners
  - Unlike accuracy, which can be skewed by the majority class, AUC-ROC focuses on the model's ability to distinguish between classes

Results - Catcher

ROC AUC Scores

- **Random Forest: 0.9866**
- Logistic Regression: 0.9032
- Gradient Boosting: 0.9677
- k-Nearest Neighbors (kNN): 0.7151
- Support Vector Machine (SVM): 0.9086



Top Players in AL League for Random Forest:

	Name	Tm	Lg	Win	RF_Probability_Class_1
66	Cal Raleigh	SEA	AL	1	0.290833
45	Shea Langeliers	OAK	AL	0	0.140000
23	Freddy Fermin	KCR	AL	0	0.114167
33	Jonah Heim	TEX	AL	0	0.070833
42	Alejandro Kirk	TOR	AL	0	0.049167
61	Logan O'Hoppe	LAA	AL	0	0.016667
64	Salvador Perez	KCR	AL	0	0.016667
19	Yainer Diaz	HOU	AL	0	0.010000
91	Austin Wells	NYY	AL	0	0.010000
46	Korey Lee	CHW	AL	0	0.005000

Top Players in NL League for Random Forest:

	Name	Tm	Lg	Win	RF_Probability_Class_1
73	Keibert Ruiz	WSN	NL	0	0.175000
3	Patrick Bailey	SFG	NL	1	0.114167
80	Will Smith	LAD	NL	0	0.049167
12	William Contreras	MIL	NL	0	0.030000
83	Tyler Stephenson	CIN	NL	0	0.015833
2	Miguel Amaya	CHC	NL	0	0.012500
1	Francisco Alvarez	NYM	NL	0	0.012500
54	Gabriel Moreno	ARI	NL	0	0.007500
18	Elias Díaz	COL	NL	0	0.005000
55	Sean Murphy	ATL	NL	0	0.002500

Results - Pitcher

ROC AUC Scores

- **Random Forest: 0.9792**
- Logistic Regression: 0.9618
- Gradient Boosting: 0.9626
- k-Nearest Neighbors (kNN): 0.4942
- Support Vector Machine (SVM): 0.9776

The actual NL League Gold Glove winner (Chris Sale) for 2024 **did not appear** in the top predictions for any models except SVM

Chris Sale **ranked ninth** among the top predictions made by the SVM model

Top Players in AL League for SVM:

	Name	Tm	Lg	Win	SVM_Probability_Class_1
319	Seth Lugo	KCR	AL	1	0.187070
559	Michael Wacha	KCR	AL	0	0.041378
143	Zach Eflin	TBR	AL	0	0.033596
239	Tanner Houck	BOS	AL	0	0.029755
44	Tanner Bibee	CLE	AL	0	0.022530
492	Brady Singer	KCR	AL	0	0.022430
494	Tarik Skubal	DET	AL	0	0.020947
165	Chris Flexen	CHW	AL	0	0.020123
433	Cole Ragans	KCR	AL	0	0.019779
292	Dean Kremer	BAL	AL	0	0.019323

Top Players in NL League for SVM:

	Name	Tm	Lg	Win	SVM_Probability_Class_1
570	Logan Webb	SFG	NL	0	0.036745
251	Jake Irvin	WSN	NL	0	0.019481
525	Ranger Suárez	PHI	NL	0	0.013582
482	Spencer Schwellenbach	ATL	NL	0	0.013286
432	Jose Quintana	NYM	NL	0	0.011114
471	Cristopher Sánchez	PHI	NL	0	0.011063
488	Luis Severino	NYM	NL	0	0.010961
282	Michael King	SDP	NL	0	0.008839
469	Chris Sale	ATL	NL	1	0.008185
268	Mitch Keller	PIT	NL	0	0.008044

Results - First Base

ROC AUC Scores

- Random Forest: 0.8854
- Logistic Regression: 0.9583
- Gradient Boosting: 0.8646
- **k-Nearest Neighbors (kNN): 0.9792**
- Support Vector Machine (SVM): 0.8958

The actual NL League Gold Glove winner (Christian Walker) for 2024 **ranked second** among the top predictions made by the kNN model

Top Players in AL League for KNN:

	Name	Tm	Lg	Win	KNN_Probability_Class_1
35	Carlos Santana	MIN	AL	1	0.232409
24	Nathaniel Lowe	TEX	AL	0	0.187653
11	Yainer Diaz	HOU	AL	0	0.000000
32	Salvador Perez	KCR	AL	0	0.000000
44	Justin Turner	SEA	AL	0	0.000000
43	Spencer Torkelson	DET	AL	0	0.000000
39	Tyler Soderstrom	OAK	AL	0	0.000000
38	Dominic Smith	BOS	AL	0	0.000000
37	Jon Singleton	HOU	AL	0	0.000000
36	Nolan Schanuel	LAA	AL	0	0.000000

Top Players in NL League for KNN:

	Name	Tm	Lg	Win	KNN_Probability_Class_1
31	Matt Olson	ATL	NL	0	0.263149
47	Christian Walker	ARI	NL	1	0.193005
20	Bryce Harper	PHI	NL	0	0.162176
0	Pete Alonso	NYM	NL	0	0.000000
1	Luis Arráez	SDP	NL	0	0.000000
48	Patrick Wisdom	CHC	NL	0	0.000000
46	LaMonte Wade Jr.	SFG	NL	0	0.000000
42	Michael Toglia	COL	NL	0	0.000000
41	Rowdy Tellez	PIT	NL	0	0.000000
40	Spencer Steer	CIN	NL	0	0.000000

Results - Second Base

ROC AUC Scores

- Random Forest: 0.9966
- Logistic Regression: 0.9832
- Gradient Boosting: 0.9966
- **k-Nearest Neighbors (kNN): 1.0000**
- Support Vector Machine (SVM): 0.9732

	Name	Tm	Lg	Win	KNN_Probability_Class_1
47	Andrés Giménez	CLE	AL	1	0.251801
120	Marcus Semien	TEX	AL	0	0.122301
46	Zack Gelof	OAK	AL	0	0.094885
110	Josh Rojas	SEA	AL	0	0.000000
107	Pablo Reyes	BOS	AL	0	0.000000
106	Luis Rengifo	LAA	AL	0	0.000000
105	Zach Remillard	CHW	AL	0	0.000000
104	Ceddanne Rafaela	BOS	AL	0	0.000000
103	Jorge Polanco	SEA	AL	0	0.000000
101	Kyren Paris	LAA	AL	0	0.000000

Top Players in NL League for KNN:

	Name	Tm	Lg	Win	KNN_Probability_Class_1
135	Brice Turang	MIL	NL	1	0.456014
128	Bryson Stott	PHI	NL	0	0.113996
115	Casey Schmitt	SFG	NL	0	0.000000
113	Thomas Saggese	STL	NL	0	0.000000
111	Miguel Rojas	LAD	NL	0	0.000000
109	Brendan Rodgers	COL	NL	0	0.000000
102	Jace Peterson	ARI	NL	0	0.000000
99	Joey Ortiz	MIL	NL	0	0.000000
96	Kevin Newman	ARI	NL	0	0.000000
1	Ozzie Albies	ATL	NL	0	0.000000

Results - Shortstop

ROC AUC Scores

- Random Forest: 0.9775
- Logistic Regression: 0.9640
- Gradient Boosting: 0.9685
- k-Nearest Neighbors (kNN): 0.7095
- **Support Vector Machine (SVM): 0.9820**

The actual AL and NL League Gold Glove winners (Bobby Witt Jr. and Ezquiel Tovar) for 2024 **ranked second** among the top predictions made by the SVM model

Top Players in AL League for SVM:

	Name	Tm	Lg	Win	SVM_Probability_Class_1
103	Anthony Volpe	NYY	AL	0	0.173609
112	Bobby Witt Jr.	KCR	AL	1	0.109599
65	Zach Neto	LAA	AL	0	0.080663
42	Gunnar Henderson	BAL	AL	0	0.076873
72	Jeremy Peña	HOU	AL	0	0.052406
79	Brayan Rocchio	CLE	AL	0	0.041397
47	Andy Ibáñez	DET	AL	0	0.041009
23	J.P. Crawford	SEA	AL	0	0.036603
88	Corey Seager	TEX	AL	0	0.032799
21	Carlos Correa	MIN	AL	0	0.026849

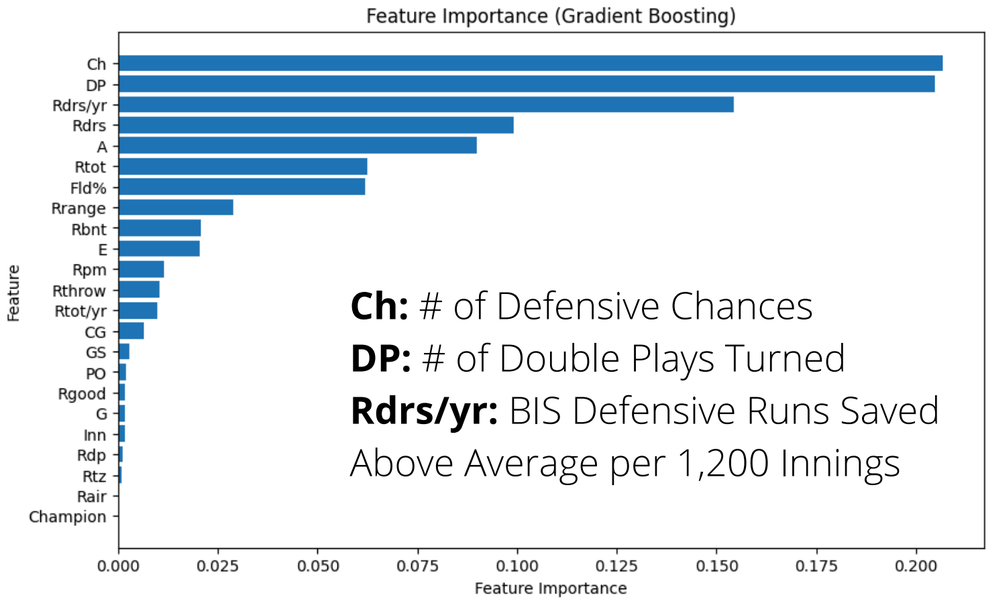
Top Players in NL League for SVM:

	Name	Tm	Lg	Win	SVM_Probability_Class_1
25	Elly De La Cruz	CIN	NL	0	0.175424
98	Ezequiel Tovar	COL	NL	1	0.119876
110	Masyn Winn	STL	NL	0	0.103984
96	Dansby Swanson	CHC	NL	0	0.061695
9	Orlando Arcia	ATL	NL	0	0.061160
2	Willy Adames	MIL	NL	0	0.058838
55	Francisco Lindor	NYM	NL	0	0.053940
0	CJ Abrams	WSN	NL	0	0.034960
80	Miguel Rojas	LAD	NL	0	0.033053
51	Ha-Seong Kim	SDP	NL	0	0.023282

Results - Third Base

ROC AUC Scores

- Random Forest: 0.9889
- Logistic Regression: 0.9333
- **Gradient Boosting: 0.9944**
- k-Nearest Neighbors (kNN): 0.7361
- Support Vector Machine (SVM): 0.9333



Top Players in AL League for Gradient Boosting:

	Name	Tm	Lg	Win	GB_Probability_Class_1
7	Alex Bregman	HOU	AL	1	1.282807e-07
10	Oswaldo Cabrera	NYY	AL	0	1.210077e-08
15	Ernie Clement	TOR	AL	0	9.330267e-09
56	Isaac Paredes	TBR	AL	0	2.614010e-09
58	José Ramírez	CLE	AL	0	2.614009e-09
9	José Caballero	TBR	AL	0	2.588726e-09
88	Davis Wendzel	TEX	AL	0	2.588726e-09
87	Eric Wagaman	LAA	AL	0	2.588726e-09
70	Nick Sogard	BOS	AL	0	2.588726e-09
20	Brandon Drury	LAA	AL	0	2.242320e-09

Top Players in NL League for Gradient Boosting:

	Name	Tm	Lg	Win	GB_Probability_Class_1
13	Matt Chapman	SFG	NL	1	8.312319e-02
46	Ryan McMahon	COL	NL	0	2.350518e-03
74	Eugenio Suárez	ARI	NL	0	6.198240e-08
8	Jake Burger	MIA	NL	0	1.690315e-08
42	Noelvi Marte	CIN	NL	0	1.186124e-08
3	Nolan Arenado	STL	NL	0	1.091719e-08
68	Nick Senzel	WSN	NL	0	1.090148e-08
84	Mark Vientos	NYM	NL	0	4.112279e-09
5	Alec Bohm	PHI	NL	0	2.614010e-09
55	Joey Ortiz	MIL	NL	0	2.614009e-09

Results - Left Field

ROC AUC Scores

- Random Forest: 0.9831
- Logistic Regression: 0.9976
- Gradient Boosting: 0.9758
- k-Nearest Neighbors (kNN): 0.4952
- **Support Vector Machine (SVM): 1.0000**

Top Players in AL League for SVM:

	Name	Tm	Lg	Win	SVM_Probability_Class_1
103	Steven Kwan	CLE	AL	1	0.129107
104	Wyatt Langford	TEX	AL	0	0.123482
192	Alex Verdugo	NYY	AL	0	0.085509
73	Riley Greene	DET	AL	0	0.081853
190	Daulton Varsho	TOR	AL	0	0.044894
1	Armando Alvarez	OAK	AL	0	0.037654
3	Miguel Andujar	OAK	AL	0	0.035164
25	Gustavo Campero	LAA	AL	0	0.031825
32	Evan Carter	TEX	AL	0	0.026154
53	Jarren Duran	BOS	AL	0	0.021595

Top Players in NL League for SVM:

	Name	Tm	Lg	Win	SVM_Probability_Class_1
79	Ian Happ	CHC	NL	1	0.179728
175	Cal Stevenson	PHI	NL	0	0.051744
29	Conner Capel	CIN	NL	0	0.037026
20	Vidal Bruján	MIA	NL	0	0.036479
12	Cody Bellinger	CHC	NL	0	0.036185
42	Garrett Cooper	CHC	NL	0	0.031225
186	Taylor Trammell	LAD	NL	0	0.031127
9	Jorge Barrosa	ARI	NL	0	0.025651
129	Andruw Monasterio	MIL	NL	0	0.025116
67	Joey Gallo	WSN	NL	0	0.022818

Results - Center Field

ROC AUC Scores

- Random Forest: 0.9483
- Logistic Regression: 0.8966
- Gradient Boosting: 0.9379
- **k-Nearest Neighbors (kNN): 0.9828**
- Support Vector Machine (SVM): 0.9276

The actual NL League Gold Glove winner (Brenton Doyle) for 2024 **ranked third** among the top predictions made by the kNN model

Top Players in AL League for KNN:

	Name	Tm	Lg	Win	KNN_Probability_Class_1
137	Daulton Varsho	TOR	AL	1	0.142857
110	Julio Rodríguez	SEA	AL	0	0.142857
128	Leody Taveras	TEX	AL	0	0.142857
0	Wilyer Abreu	BOS	AL	0	0.000000
85	Ryan McKenna	BAL	AL	0	0.000000
96	Rafael Ortega	CHW	AL	0	0.000000
92	Cedric Mullins	BAL	AL	0	0.000000
91	Mickey Moniak	LAA	AL	0	0.000000
89	Kameron Misner	TBR	AL	0	0.000000
88	Jake Meyers	HOU	AL	0	0.000000

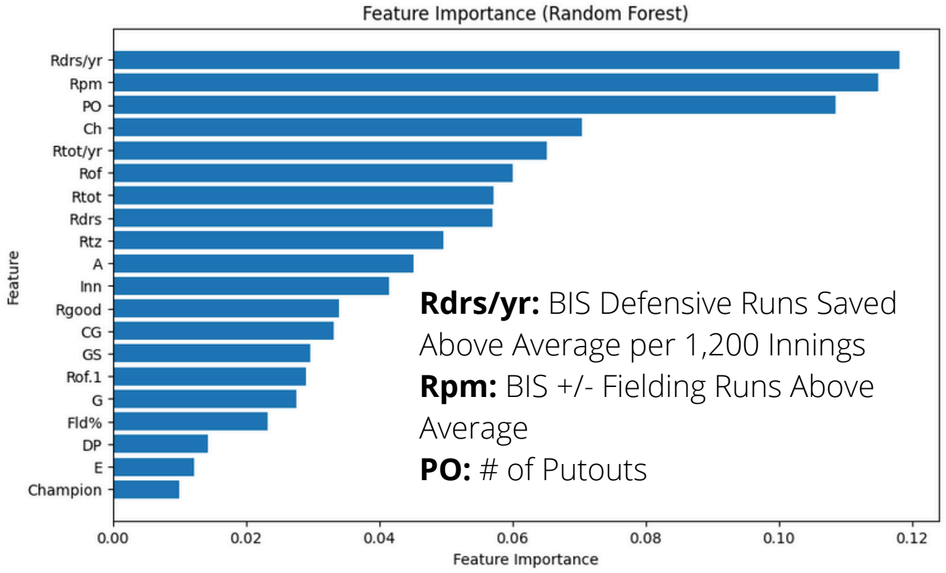
Top Players in NL League for KNN:

	Name	Tm	Lg	Win	KNN_Probability_Class_1
130	Michael A. Taylor	PIT	NL	0	0.285714
146	Jacob Young	WSN	NL	0	0.142857
36	Brenton Doyle	COL	NL	1	0.142857
113	Eddie Rosario	WSN	NL	0	0.000000
112	Chris Roller	MIL	NL	0	0.000000
111	Johan Rojas	PHI	NL	0	0.000000
107	Heliot Ramos	SFG	NL	0	0.000000
102	Blake Perkins	MIL	NL	0	0.000000
100	Joshua Palacios	PIT	NL	0	0.000000
99	Andy Pages	LAD	NL	0	0.000000

Results - Right Field

ROC AUC Scores

- **Random Forest: 0.9924**
- Logistic Regression: 0.9646
- Gradient Boosting: 0.9836
- k-Nearest Neighbors (kNN): 0.7475
- Support Vector Machine (SVM): 0.9520



Rdrs/yr: BIS Defensive Runs Saved Above Average per 1,200 Innings
Rpm: BIS +/- Fielding Runs Above Average
PO: # of Putouts

Top Players in AL League for Random Forest:

	Name	Tm	Lg	Win	RF_Probability_Class_1
0	Wilyer Abreu	BOS	AL	1	0.297635
165	Juan Soto	NYY	AL	0	0.129958
67	Adolis García	TEX	AL	0	0.022667
180	Kyle Tucker	HOU	AL	0	0.019750
3	Jo Adell	LAA	AL	0	0.015000
26	Lawrence Butler	OAK	AL	0	0.014310
156	Anthony Santander	BAL	AL	0	0.009048
148	Hunter Renfroe	KCR	AL	0	0.003929
140	Wenceel Pérez	DET	AL	0	0.003500
107	Josh Lowe	TBR	AL	0	0.000833

Top Players in NL League for Random Forest:

	Name	Tm	Lg	Win	RF_Probability_Class_1
41	Jackson Chourio	MIL	NL	0	0.077750
39	Nick Castellanos	PHI	NL	0	0.062179
63	Sal Frelick	MIL	NL	1	0.040833
155	Jesús Sánchez	MIA	NL	0	0.009476
176	Lane Thomas	WSN	NL	0	0.006667
82	Jason Heyward	LAD	NL	0	0.006167
173	Fernando Tatis Jr.	SDP	NL	0	0.005000
103	Ramón Laureano	ATL	NL	0	0.003500
164	Jorge Soler	ATL	NL	0	0.003333
113	Starling Marte	NYM	NL	0	0.002500

Discussion Pt. 1

Overall, models showed **strong ranking ability across positions, with high AUC scores**, but failed to predict winners due to the lack of sufficient training examples for class 1

- Models struggled to learn patterns that accurately classify the minority class (winners)
- Across all models, accuracy was high, reflecting the dominance of non-winners in the datasets

Challenges

- Class Imbalance
  - With winners making up only ~1% of the dataset, all models prioritized predicting the majority class (non-winners), leading to perfect recall for class 0 but zero recall for class 1
- Subjective Voting Criteria
  - Gold Glove decisions are influenced by subjective factors as the manager and six coaches for each MLB team vote not based solely on fielding metrics
  - Even with high probabilistic scores, some players may not win due to external factors unrelated to fielding performance

Discussion Pt. 2

Opportunities for Improvement

- Incorporate Data Before 2013
  - Extend the datasets to include seasons prior to 2013 by discarding metrics unavailable in earlier years
  - This would increase the number of observations, especially for the minority class (winners)
- Clean Data Based on Player Qualifications for the Gold Glove Award
  - Refine the datasets by ensuring only players who meet the qualifications for the award are included
  - Example: Minimum innings played or games started
  - This excludes high-performing players who do not qualify, aligning the datasets with real-world award criteria and improving model relevance
  - May also reduce noise caused by players with insufficient playing time, where short-term performance inflates metrics (e.g. fewer chances to make errors)

References

https://github.com/lucaskelly49/Machine-Learning-Model-Predicting-MLB-Gold-Glove-Award-Winners

- A guidance resource for a machine learning project developed during a course
- Developed as part of the Module 3 Final Project for a Flatiron School curriculum