Modeling MLB Free Agent Value

Predicting AAV and Contract Years Using Linear Regression & Random Forests.

Hunter Graham

Overview & Goals

- **Predict MLB free agent contracts** (AAV and Years) using player performance and past free agent data.
- Apply **K-means clustering** & **PCA** to develop **linear regression** & **random forest** learning models.
- **Separate models** by position group: hitters and pitchers
- Identify which **variables** have the greatest impact on contract outcomes.
- Evaluate performance using **Adjusted R^2** & **RMSE** to find the most accurate model.

Data → Clean → Cluster/PCA → Model (LR & RF) → Evaluate (R^2, RMSE)

Data Sources & Preparation

- Collected **MLB player performance data** (2022-2025) on every player from *Fangraphs*
- Collected **MLB free agent outcomes** (2022-2025) on every signing from *Spotrac*
- **Cleaned & merged** data frames to align player stats with contract outcomes
- **Standardized variables** for PCA input
- Created **player clusters** (k-means) to segment by performance type

Fangraphs + Spotrac → Clean & Merge → Standardize → PCA → Cluster

Linear Regression: Hitters

- **Formulas:**

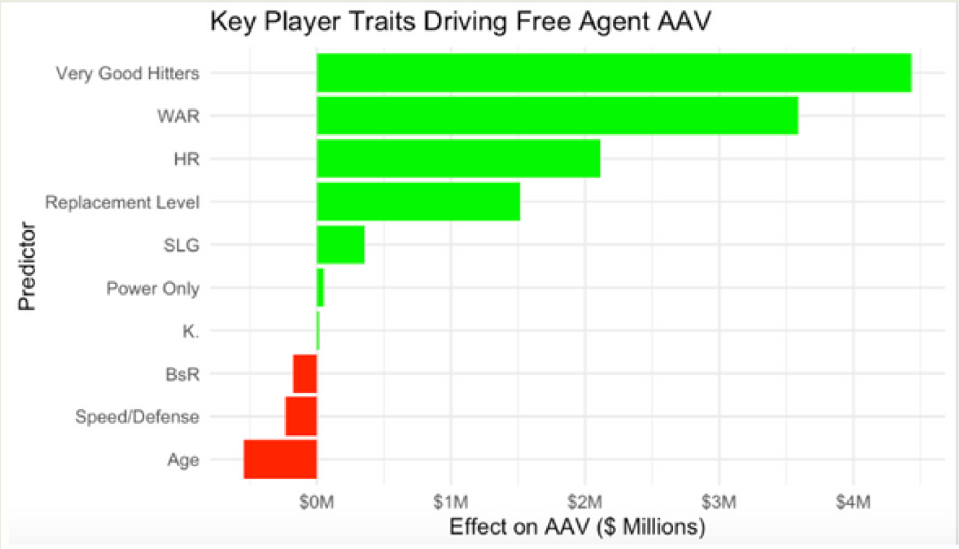
```
lm(formula = AAV ~ WAR + SLG + HR + Age + BsR + K. + Cluster_Name,
    data = hitters_standardized)
```

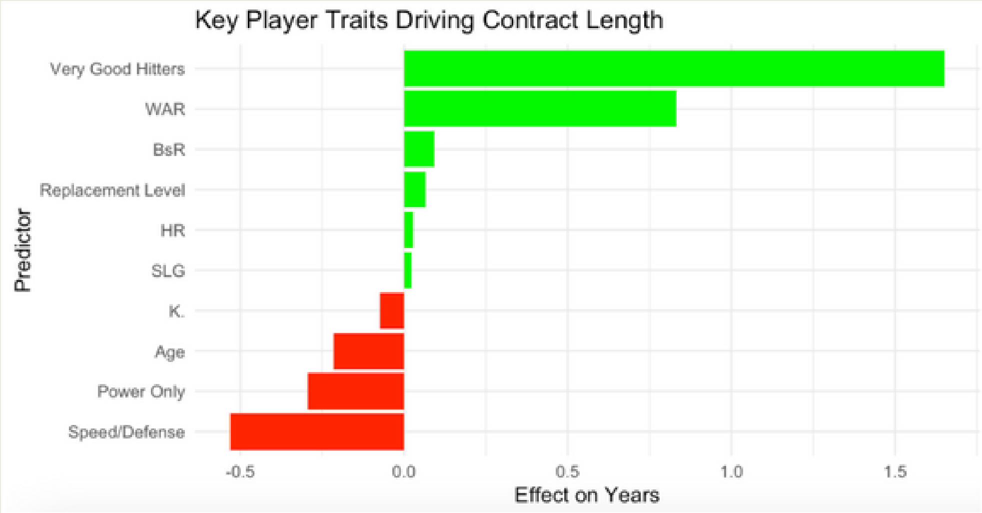
- **Adjusted R^2 AAV: 0.64**
- Model constantly underpredicted elite-level free agents.
- Likely due to linearity assumptions & regression to the mean.

```
lm(formula = Years ~ WAR + SLG + HR + Age + BsR + K. + Cluster_Name,
    data = hitters_standardized)
```

- **Adjusted R^2 Years: 0.52**
- Lower R^2 suggests contract length is harder to predict than AAV.
- Likely missing external factors (injury history, team need, marketability).

- **Significant Variables:**





Linear Regression: Pitchers

- **Formulas:**

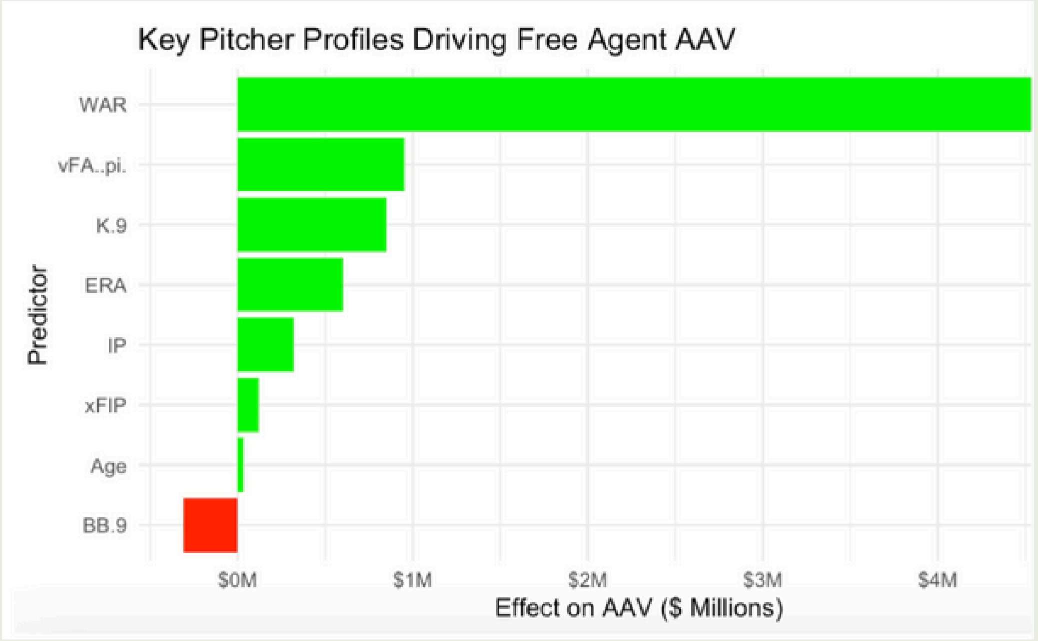
```
lm(formula = AAV ~ WAR + ERA + xFIP + K.9 + BB.9 + IP + vFA..pi. +
    Age, data = pitchers_standardized)
```

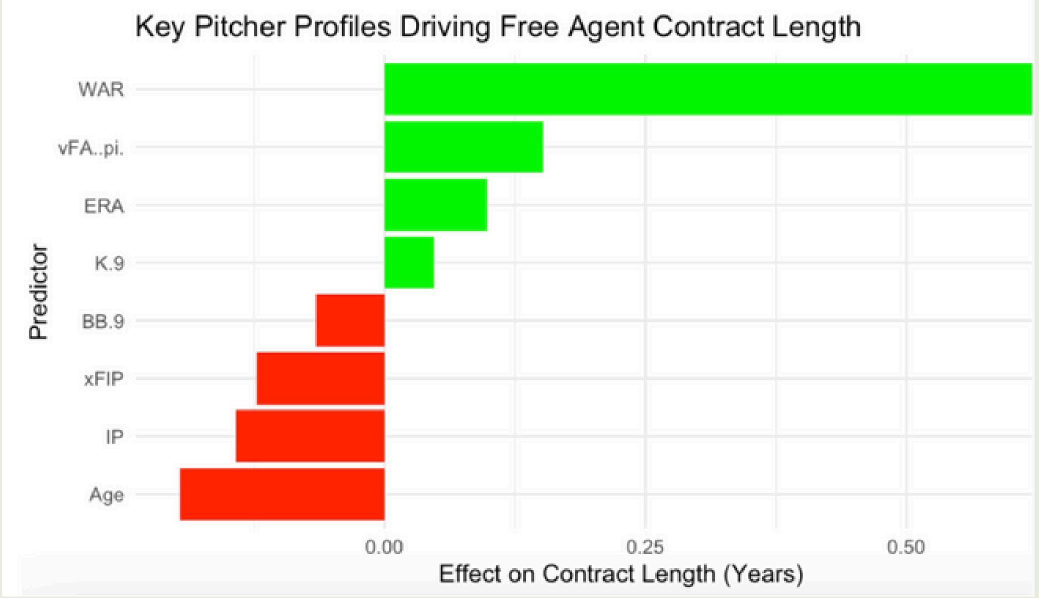
- **Adjusted R^2 AAV: 0.53**
- Model constantly underpredicted elite-level free agents.
- As with hitters, underprediction likely stems from linear assumptions and regression toward the mean.

```
lm(formula = Years ~ WAR + ERA + xFIP + K.9 + BB.9 + IP + vFA..pi. +
    Age, data = pitchers_standardized)
```

- **Adjusted R^2 Years: 0.38**
- Contract length's lower R^2 in both models reinforces that it's harder to predict than AAV.
- Future iterations to achieve a higher R^2 score are necessary.

- **Significant Variables**





Random Forest: Hitters

- **Formula:**

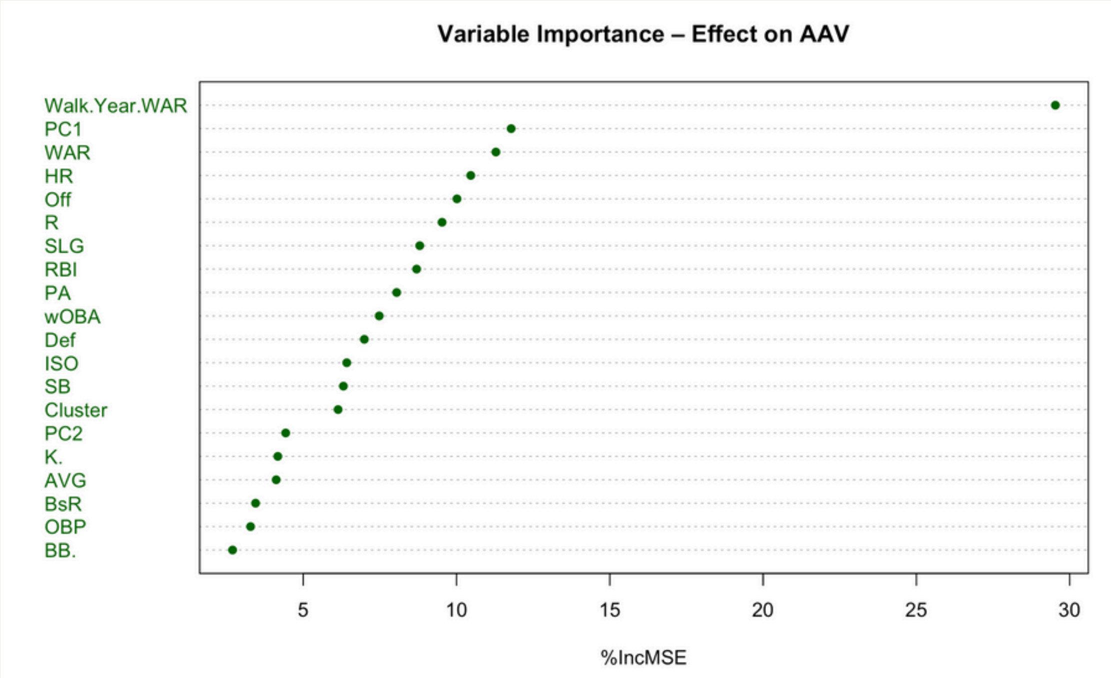
```
set.seed(123)
rf_model_aav <- randomForest(AAV ~ Walk.Year.WAR + WAR + PA + AVG + OBP + SLG + HR + RBI +
  R + BB. + K. + SB + ISO + wOBA + Off + Def + BsR + Cluster + PC1 + PC2,
    data = train,
      ntree = 500,
        mtry = 5,
          importance = TRUE)
```

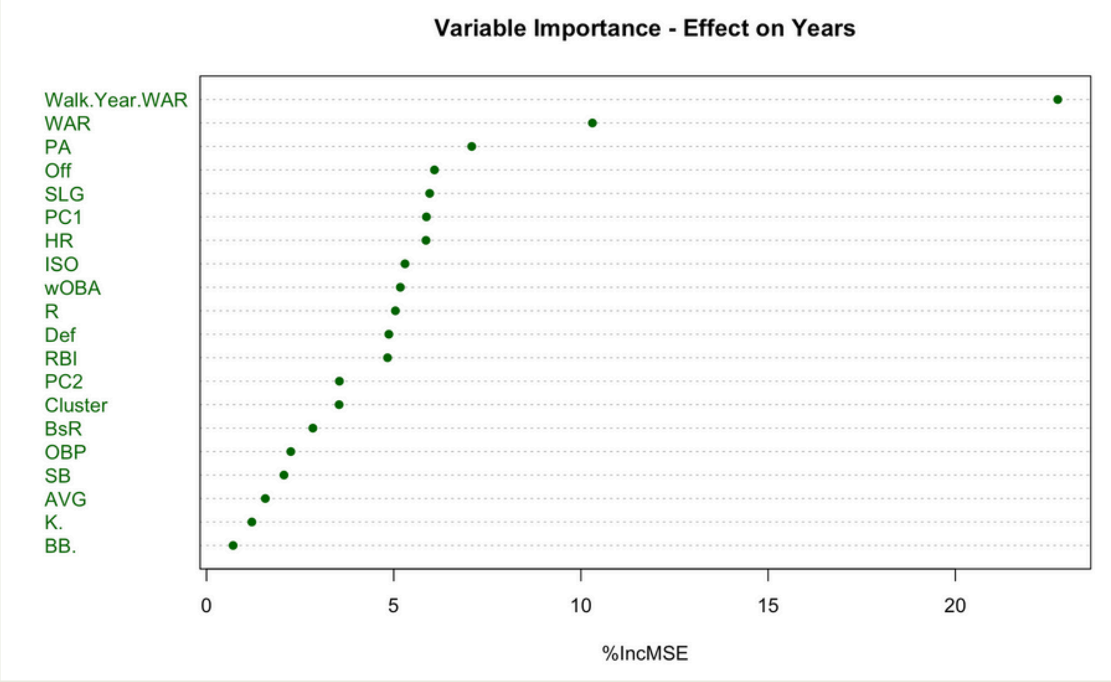
- **Adjusted R^2 AAV: 0.72**
- RF predicted elite-level free agents more accurately.

```
set.seed(123)
rf_model_years <- randomForest(Years ~ Walk.Year.WAR + WAR + PA + AVG + OBP + SLG + HR + RBI +
  R + BB. + K. + SB + ISO + wOBA + Off + Def + BsR + Cluster + PC1 + PC2,
    data = train,
      ntree = 500,
        mtry = 5,
          importance = TRUE)
```

- **Adjusted R^2 Years: 0.74**
- The inclusion of “Walk.Year.WAR” significantly improved predictive accuracy in both models.

- **Significant Variables:**





Random Forest: Pitchers

- **Formula:**

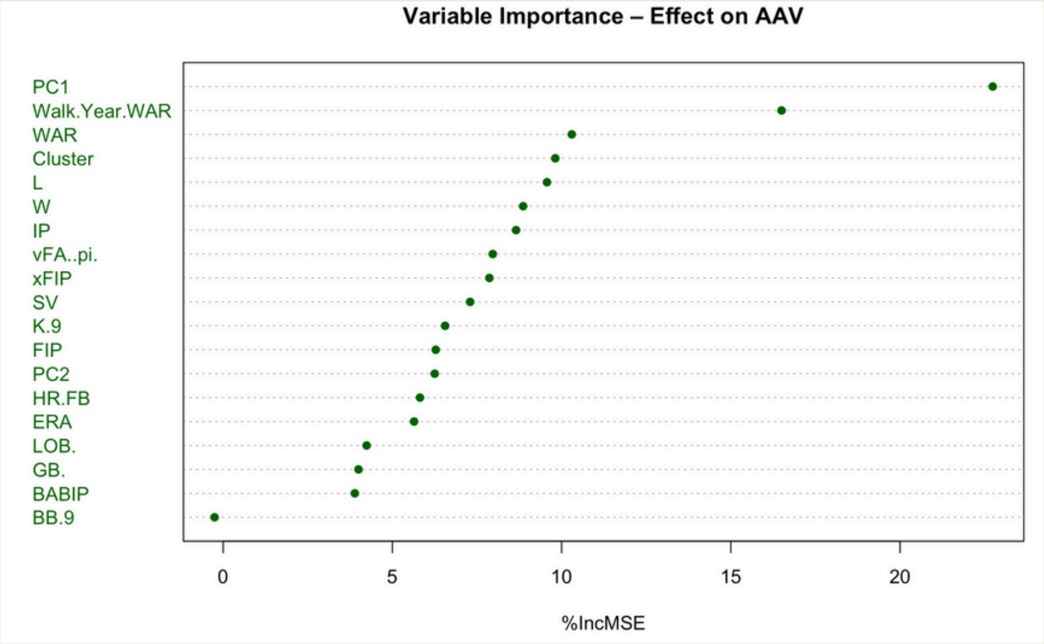
```
set.seed(123)
rf_model_aav_p19 <- randomForest(AAV ~ Walk.Year.WAR + WAR + W + L + SV + IP + K.9 + BB.9 +
    BABIP + LOB. + GB. + HR.FB + vFA..pi. + ERA + FIP + xFIP + PC1 + PC2 + Cluster,
      data = train_p,
        ntree = 500,
          mtry = 3,
            importance = TRUE)
```

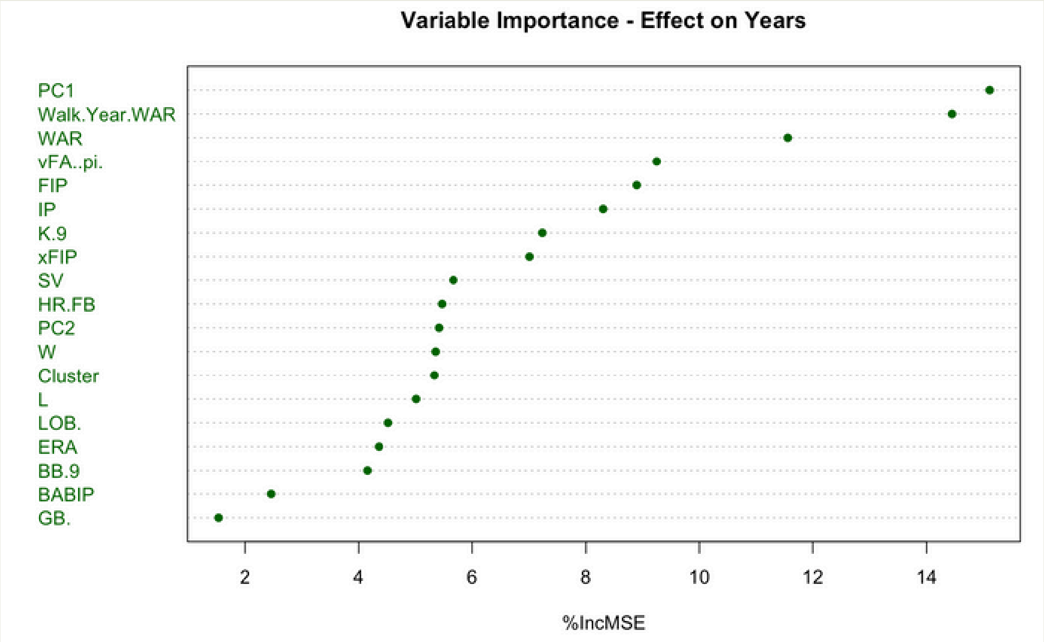
- **Adjusted R^2 AAV: 0.76**
- RF predicted all cluster groups much more accurately.

```
set.seed(123)
rf_model_years_p19 <- randomForest(Years ~ Walk.Year.WAR + WAR + W + L + SV + IP + K.9 + BB.9 +
  BABIP + LOB. + GB. + HR.FB + vFA..pi. + ERA + FIP + xFIP + PC1 + PC2 + Cluster,
    data = train_p,
      ntree = 500,
        mtry = 3,
          importance = TRUE)
```

- **Adjusted R^2 AAV: 0.35**
- RF failed to predicted Years more accurately. ‘Pitchers, Years’ consistently returned the lowest R^2 values.

- **Significant Variables**

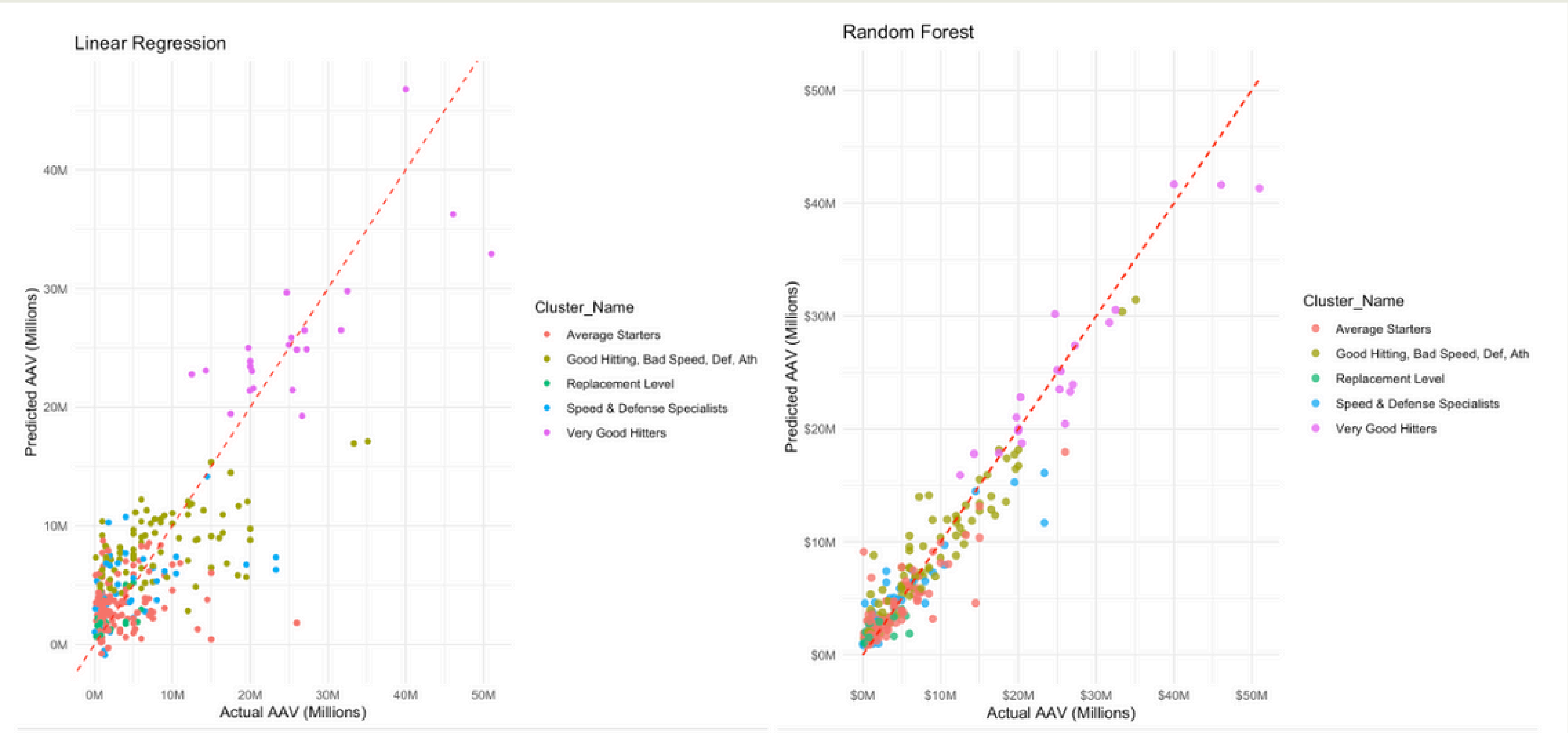




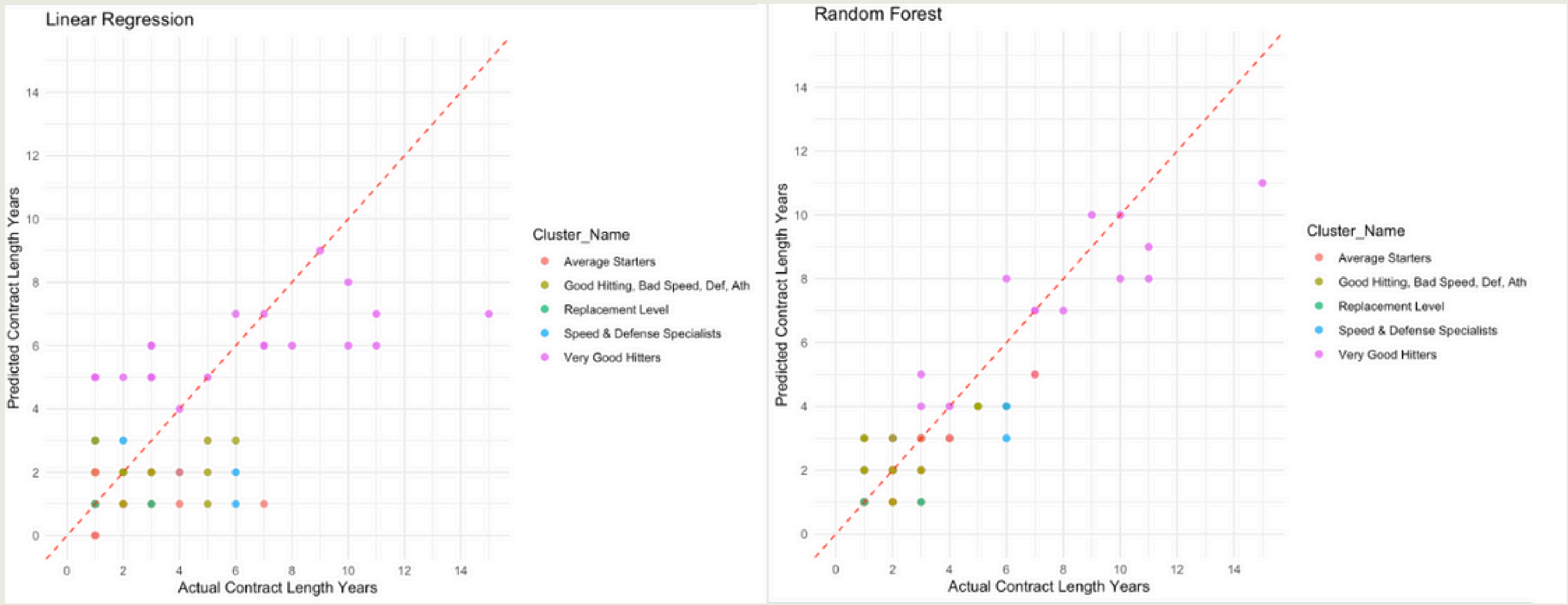
Model Comparison: Hitters

Metric:	Linear:	Random Forest:
• Adjusted R^2 (AAV):	• 0.64	• **0.72**
• Adjusted R^2 (Years):	• 0.52	• **0.74**
• RMSE (AAV in $M):	• 9.36	• **3.60**
• RMSE (Years):	• 1.25	• **0.76**

Model Accuracy Comparison (AAV)



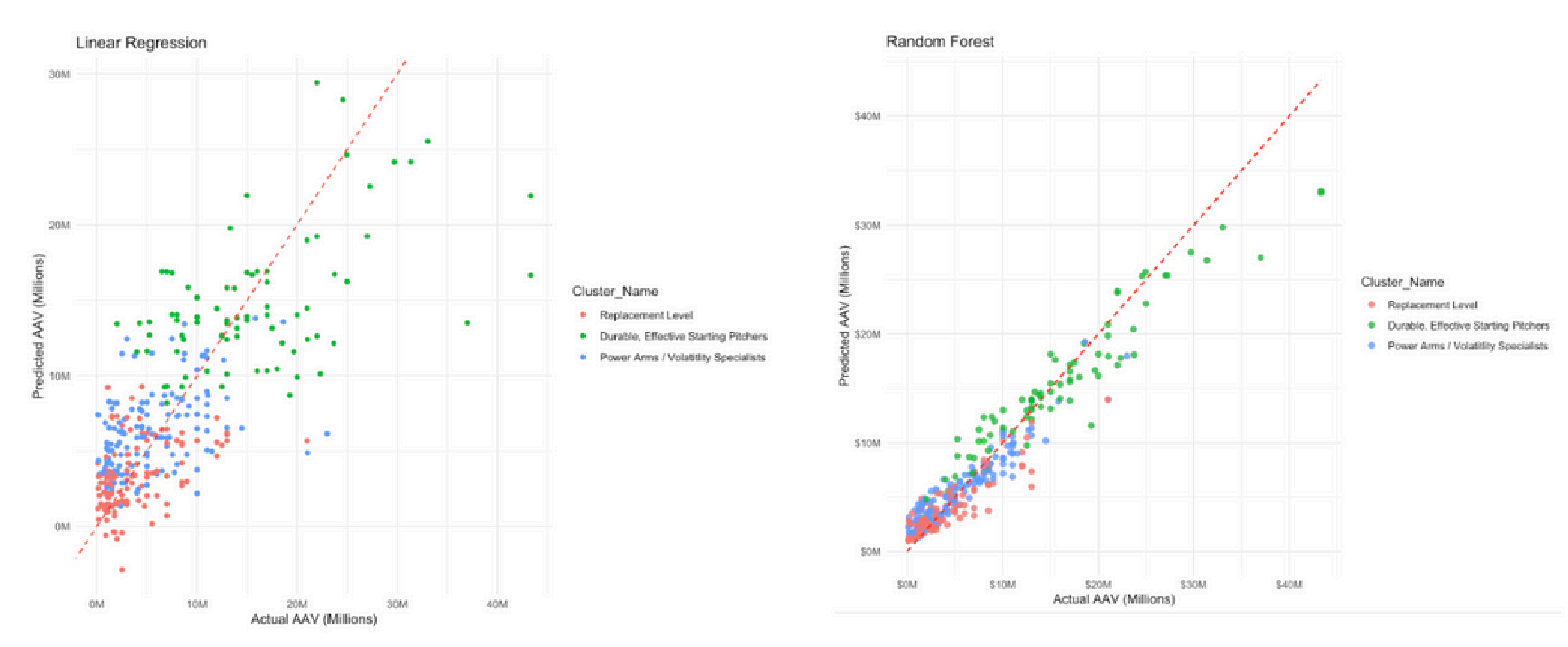
Model Accuracy Comparison (Years)



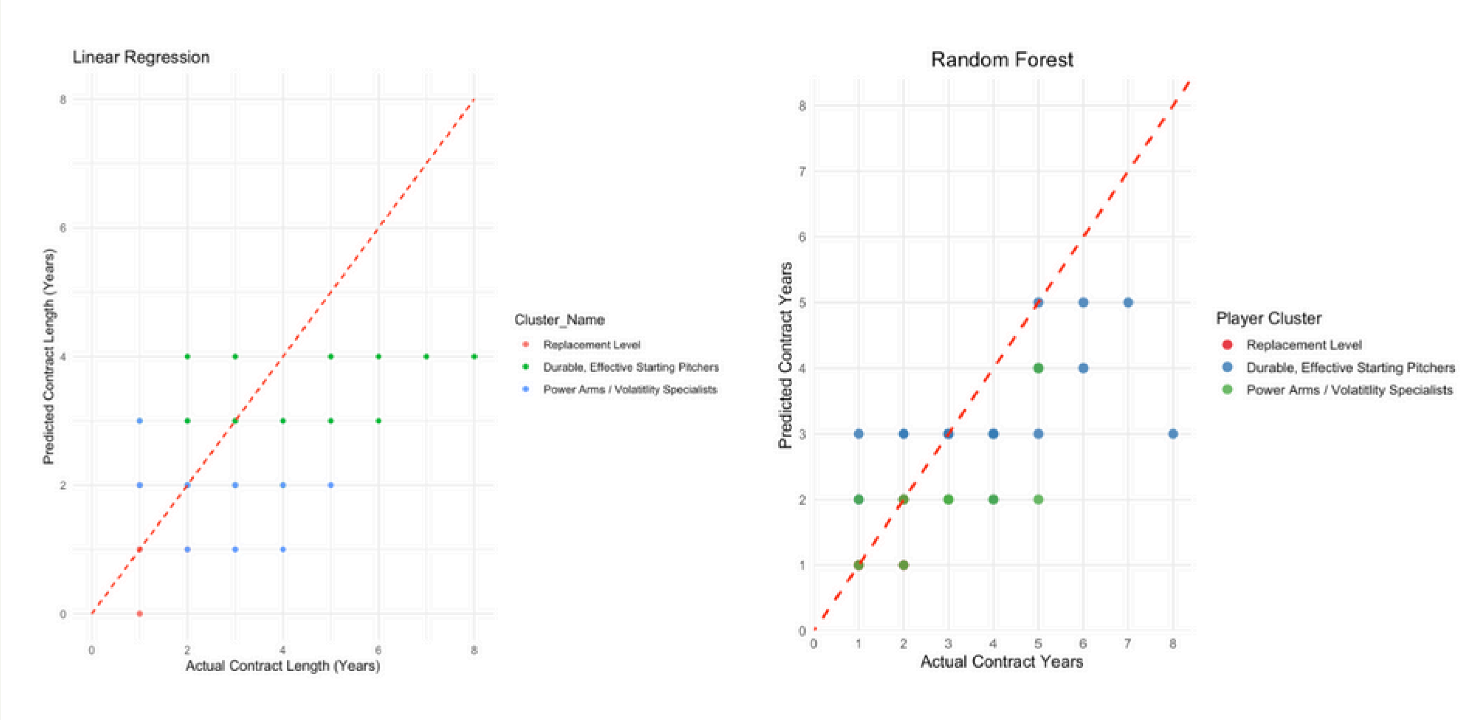
Model Comparison: Pitchers

Metric:	Linear:	Random Forest:
• Adjusted R^2 (AAV):	• 0.53	• **0.76**
• Adjusted R^2 (Years):	• **0.38**	• 0.35
• RMSE (AAV in $M):	• 4.86	• **2.78**
• RMSE (Years):	• **0.83**	• 0.92

Model Accuracy Comparison (AAV)



Model Accuracy Comparison (Years)



2026 Free Agent Predictions: Hitters

Variable Estimation:

- Walk.Year.WAR isn't available midseason. Each player's current WAR was extrapolated using pace-based projections from partial 2025 stats.

Key Takeaways:

- RF predicted higher AAV in 29/33 cases.
- This is significant since the linear model consistently undervalued elite level fa's.
- RF predicted higher AAVs and longer contracts for elite hitters by modeling nonlinear relationships and variable interactions, producing valuations that better matched recent free agent trends and market behavior.

Table displays linear and RF predictions for 2026 free agent hitters, ranked by RF's predicted AAV (top 33 shown).

	Name_clean	Age	Predicted_AAV	Predicted_AAV_rf	Predicted_Years	Predicted_Years_rf	Total_Salary	Total_Salary_rf
2	alex bregman	32.1	26.09	31.92	6	6	156.54	191.52
44	kyle tucker	29.2	27.58	28.77	7	8	193.06	230.16
67	pete alonso	31.3	25.49	27.78	5	5	127.45	138.90
86	william contreras	28.3	24.52	25.12	6	6	147.12	150.72
43	kyle schwarber	33.1	23.01	25.08	4	5	92.04	125.40
52	marcell ozuna	35.4	21.07	22.82	4	5	84.28	114.10
17	cody bellinger	30.8	13.20	22.39	3	3	39.60	67.17
66	paul goldschmidt	38.6	21.87	21.40	5	5	109.35	107.00
23	gleyber torres	29.3	13.82	21.06	3	3	41.46	63.18
38	josh naylor	28.8	13.92	18.85	3	3	41.76	56.55
10	bo bichette	28.1	13.96	17.62	3	3	41.88	52.86
29	jarren duran	29.6	11.21	16.52	3	3	33.63	49.56
15	cedric mullins	31.6	12.41	16.29	3	3	37.23	48.87
55	max muncy	35.7	11.90	15.23	2	2	23.80	30.46
72	ryan o'hearn	32.8	5.30	14.80	1	3	5.30	44.40
27	j.t. realmuto	35.1	15.09	14.29	3	2	45.27	28.58
7	austin hays	30.8	7.44	14.18	2	3	14.88	42.54
11	brandon lowe	31.8	9.74	12.91	2	2	19.48	25.82
70	rob refsnyder	35.1	2.78	12.44	1	3	2.78	37.32
79	trent grisham	29.5	7.42	12.20	2	3	14.84	36.60
24	harrison bader	31.9	5.04	12.14	1	3	5.04	36.42
48	luis arraez	29.0	9.10	11.98	3	3	27.30	35.94
64	ozzie albies	29.2	11.70	11.42	3	1	35.10	11.42
69	rhys hoskins	33.1	6.89	11.27	1	2	6.89	22.54
47	lourdes gurriel jr.	32.5	9.10	11.11	2	2	18.20	22.22
32	joc pederson	34.0	9.82	10.73	2	1	19.64	10.73
35	jorge polanco	32.8	7.03	10.42	1	1	7.03	10.42
8	austin hedges	33.7	1.81	10.39	1	3	1.81	31.17
13	carlos santana	40.0	9.00	10.31	1	1	9.00	10.31
81	ty france	31.8	6.66	10.27	1	2	6.66	20.54
50	luis robert jr.	28.8	12.78	9.49	2	1	25.56	9.49
18	danny jansen	31.0	8.26	9.34	2	1	16.52	9.34
60	mike yastrzemski	35.7	8.51	8.62	2	1	17.02	8.62

Showing 1 to 33 of 87 entries, 29 total columns

2026 Free Agent Predictions: Pitchers

Variable Estimation:

- As with hitter projections, Walk.Year.WAR isn't available midseason. Each player's current WAR was extrapolated using pace-based projections from partial 2025 stats.

Key Takeaways:

- RF predicted higher AAV in 21/33 cases.
- Both models tend to overvalue aging pitchers.
- RF's use of non-linear effects and variable interactions yielded more realistic AAV estimates that aligned more closely with actual market valuations.

Table displays linear and RF predictions for 2026 free agent pitchers, ranked by RF's predicted AAV (top 33 shown).

	Name_clean	Age	Predicted_AAV	Predicted_AAV_rf	Predicted_Years	Predicted_Years_rf	Total_Salary	Total_Salary_rf
47	justin verlander	43.2	20.27	25.08	2	2	40.54	50.15
34	framber valdez	32.4	24.65	22.11	4	4	98.60	88.46
30	dylan cease	30.3	27.27	21.79	4	4	109.08	87.14
18	chris bassitt	37.2	17.56	21.23	2	3	35.12	63.69
35	freddy peralta	29.8	16.84	19.88	3	3	50.52	59.65
66	michael king	30.9	17.32	18.94	3	3	51.96	56.83
21	chris sale	37.1	18.96	18.63	3	2	56.88	37.25
65	merrill kelly	37.5	14.03	18.19	2	2	28.06	36.38
64	max scherzer	41.8	15.28	17.72	2	2	30.56	35.44
71	miles mikolas	37.7	16.94	16.33	2	3	33.88	48.98
73	nick martinez	35.7	11.27	16.10	2	2	22.54	32.19
111	zac gallen	30.8	23.97	15.67	4	3	95.88	47.02
62	marcus stroman	35.0	11.14	14.27	2	2	22.28	28.54
17	charlie morton	42.4	12.68	14.26	1	2	12.68	28.51
83	robert suarez	35.2	7.10	14.26	2	2	14.20	28.51
40	jack flaherty	30.5	12.09	13.96	2	2	24.18	27.92
112	zach eflin	32.0	18.20	13.65	3	2	54.60	27.30
109	walker buehler	31.8	4.34	13.45	2	1	8.68	13.45
72	nestor cortes	31.3	14.54	13.07	2	2	29.08	26.14
105	tyler mahle	31.6	8.07	12.00	2	2	16.14	24.01
91	seth lugo	36.4	15.43	11.94	2	2	30.86	23.89
7	aroldis chapman	38.2	10.54	11.77	2	2	21.08	23.55
2	aaron civale	30.8	9.93	11.75	2	2	19.86	23.50
23	clayton kershaw	38.1	12.65	11.40	2	2	25.30	22.80
58	lucas giolito	31.8	8.02	11.24	2	2	16.04	22.47
75	paul blackburn	32.3	6.47	10.96	2	2	12.94	21.92
32	erick fedde	33.2	8.40	10.91	2	2	16.80	21.82
113	zack littell	30.5	7.60	10.71	2	2	15.20	21.43
53	kyle gibson	38.5	12.51	10.55	2	1	25.02	10.55
107	tyler rogers	35.3	0.36	9.99	1	2	0.36	19.98
8	austin gomber	32.4	6.74	9.78	1	1	6.74	9.78
103	tyler anderson	36.3	12.54	9.58	2	2	25.08	19.16
85	ryan helsley	31.8	14.61	9.45	3	2	43.83	18.90

Showing 1 to 33 of 113 entries, 24 total columns

Model Accuracy on Past Free Agents (2022-2025)

Tables represents the 15 largest AAV overpays from 2022-2025 for hitters & pitchers, comparing actual contract values to predictions from both linear and rf models.

- For hitters, the 15th largest overpay by the linear model was off by \$10.18M vs \$3.83M for the rf.
- For pitchers, the 15th largest overpay was off by \$7.97M (linear) vs. \$4.32M (rf).
- Rf model produced significantly more accurate AAV predictions than the linear model for both hitters and pitchers.

Linear

	Player	Year	Actual AAV (M)	Predicted AAV (M)	Diff (M)
1	kris bryant	2022	26.00	1.82	-24.18
2	juan soto	2025	51.00	32.92	-18.08
3	carlos correa	2022	35.10	17.11	-17.99
4	javier báez	2022	23.33	6.29	-17.04
5	carlos correa	2023	33.33	16.93	-16.40
6	trevor story	2022	23.33	7.35	-15.98
7	nelson cruz	2022	15.00	0.44	-14.56
8	josé abreu	2023	19.50	5.67	-13.83
9	starling marte	2022	19.50	6.72	-12.78
10	brandon belt	2022	18.40	5.83	-12.57
11	avisail garcía	2022	13.25	1.28	-11.97
12	anthony rizzo	2023	20.00	8.80	-11.20
13	mitch haniger	2023	14.50	3.77	-10.73
14	nick castellanos	2022	20.00	9.75	-10.25
15	michael conforto	2025	17.00	6.82	-10.18

Random Forest

	Player	Year	Actual AAV (M)	Predicted AAV (M)	Diff (M)
1	javier báez	2022	23.33	11.34	-11.99
2	mitch haniger	2023	14.50	4.75	-9.75
3	juan soto	2025	51.00	41.34	-9.66
4	kris bryant	2022	26.00	17.47	-8.53
5	trevor story	2022	23.33	16.53	-6.80
6	eddie rosario	2022	9.00	3.35	-5.65
7	starling marte	2022	19.50	14.04	-5.46
8	brandon belt	2022	18.40	13.28	-5.12
9	shohei ohtani	2024	46.08	41.04	-5.04
10	nelson cruz	2022	15.00	10.03	-4.97
11	willy adames	2025	26.00	21.10	-4.90
12	michael conforto	2025	17.00	12.16	-4.84
13	carlos correa	2022	35.10	30.61	-4.49
14	mike zunino	2023	6.00	2.13	-3.87
15	tyler o'neill	2025	16.50	12.67	-3.83

Linear

	Player	Year	Actual AAV (M)	Predicted AAV (M)	Difference (M)
1	max scherzer	2022	43.33	16.64	-26.69
2	jacob degrom	2023	37.00	13.49	-23.51
3	justin verlander	2023	43.33	21.91	-21.42
4	robbie ray	2022	23.00	6.16	-16.84
5	walker buehler	2025	21.05	4.88	-16.17
6	noah syndergaard	2022	21.00	5.68	-15.32
7	luis severino	2025	22.33	10.12	-12.21
8	marcus stroman	2022	23.67	12.15	-11.52
9	lucas giolito	2024	19.25	8.71	-10.54
10	eduardo rodriguez	2024	20.00	9.92	-10.08
11	sean manaea	2025	22.01	12.62	-9.39
12	nathan eovaldi	2025	25.00	16.23	-8.77
13	nick martinez	2025	21.05	12.41	-8.64
14	martín pérez	2023	19.65	11.60	-8.05
15	matthew boyd	2025	14.50	6.53	-7.97

Random Forest

	Player	Year	Actual AAV (M)	Predicted AAV (M)	Diff (M)
1	max scherzer	2022	43.33	32.96	-10.37
2	justin verlander	2023	43.33	33.13	-10.20
3	jacob degrom	2023	37.00	26.99	-10.01
4	lucas giolito	2024	19.25	11.58	-7.67
5	walker buehler	2025	21.05	13.97	-7.08
6	james paxton	2024	13.00	5.95	-7.05
7	noah syndergaard	2022	21.00	13.97	-7.03
8	jordan montgomery	2024	23.75	18.07	-5.68
9	zack greinke	2022	13.00	7.35	-5.65
10	robbie ray	2022	23.00	17.96	-5.04
11	sean manaea	2025	22.01	17.10	-4.91
12	aaron loup	2022	8.50	3.74	-4.76
13	blake snell	2025	31.36	26.74	-4.62
14	luis severino	2025	22.33	17.77	-4.56
15	matthew boyd	2025	14.50	10.18	-4.32

Model Limitations

Variable Estimation:

- Walk.Year.WAR was projected using partial 2025 stats, which may not reflect end-of-season performance.
- PC1 & PC2 values were imputed for some upcoming free agents using cluster group averages, possibly reducing individual accuracy.

Omitted Variables:

- Key predcictors like; injury history, market size, team needs, positional scarcity, player marketability, and team financial positioning were excluded.

Contract Structure Oversight:

- Models only predicted AAV, ignoring contract structure details such as front-/back-loading, opt-outs, bonuses, and incentives.

Future Improvements

Refine Pitcher “Years” Model:

- Random forest improved 3 of the 4 models, except for predicting pitcher contract length. While the other models achieved adjusted R^2 scores above 0.70, this model remained below 0.50, signaling a clear need for further development.

Integrate External Factors:

- Incorporate missing variables like injury history, market size, team needs, positional scarcity, player marketability, and team financial health to enhance prediction accuracy.

Continue Model Iteration:

- Test additional modeling techniques and explore alternative non-linear approaches for improved performance.

Key Takeaways & Conclusion

- Random forest models significantly outperformed linear regression in predicting free agent AAV & contract length, especially for elite players.
- Results reinforce the vlauye of flexible, machine learning-based methods for contract forecasting over rigid linear approaches.
- This project demonsrtaed the effectiveness of non-linear modeling in high-stakes markets like MLB free agency.