Predicting 2025-2026 MLB Free Agent Contracts Using R, Clustering, and Random Forests

Executive Summary:

The purpose of this project is to build based on the foundation of part 1 by implementing random forest (RF) models to improve the prediction of MLB free agent contract values and lengths for both hitters and pitchers. The primary objective was to assess whether nonlinear modeling techniques could outperform the original linear regression models when it came to capturing the complex relationships between player performance metrics and market outcomes.

Multiple RF models were developed and iteratively refined for hitters. The most accurate model incorporated 20 performance variables. The model explained 72% of variation in AAV, and 73% in contract length for hitters. Pitchers were a bit harder to predict, as the model explained 76% of variation in AAV; however, just 35% for contract length. This highlights the greater variance in pitcher contract structures. Visualizations and residual analysis confirmed strong predictive alignment, especially among high-value elite free agents.

The most significant finding was the importance of the variable “Walk.Year.WAR”. This was the most significant variable in 3 of the 4 random forest models. The adjusted r-squared values dropped notably when this variable was removed. Originally this variable was omitted from the linear model, due to multicollinearity concerns. However, it became apparent how valuable and impactful this variable became in RF models.

Finally, the trained random forest models were applied to the 2026 free agent class to forecast projected contracts. The “Walk.Year.WAR” values were scaled from current season-to-date stats, and estimated over a full 162 game season. Missing data for PCA, and HR.FB values were imputed utilizing medians, or model-derived estimates. The resulting forecasts provide realistic AAV and contract length expectations by player.

Overall, this phase demonstrated that random forest modeling offers a more flexible and accurate approach when it comes to free agent prediction. These models offer a clear blueprint for applying machine learning to future contract valuations.

Hitters_Fa_Join Random Forest Model:

The first random forest model created for the “hitters_fa_join” consisted of the following variables; WAR, SLG, HR, Age, BsR, K., and Cluster. These variables were chosen as a preliminary

test run. When predicting AAV, the adjusted r-squared of this model came out to 0.65. While this was a slight improvement over the linear model, continued iteration was necessary.

Since multicollinearity doesn't negatively affect random forest models the same way it affects linear models, the next iteration consisted of all 20 variables that made up our "hitters_fa_join" data frame. The following variables that were included in this RF model are; Walk.Year.WAR, WAR, PA, AVG, OBP, SLG, HR, RBI, R, BB., K., SB, ISO, wOBA, Off, Def, Cluster, PC1, and PC2. This model produced an adjusted r-squared score of 0.72 when predicting AAV, noticeably better than our linear model.

So far, models have consisted of 7 and 18 variables. A third model consisting of 12 variables was produced, to see if that was the optimal number of variables. However, this model scored much lower than the other two, with an adjusted r-squared score of 0.60.

With one last effort to produce the highest adjusted r-squared score possible, the last model consisted of multiple feature engineered variables such as; War per PA, SLG to OBP ratio, an Athleticism score, Age buckets (higher values for younger players), and Premier position grouping (C, SS, CF). Multiple different iterations of this model were run, with different combinations between all 5 feature engineered groups. Results with all combinations were similar across the board. The adjusted r-squared when predicting AAV ranged anywhere from 0.66-0.71.

The final model utilized for future predictions was the 20 variable model, which scored 0.72. The final decision to use this model was based upon the adjusted r-squared score, along with easier interpretability when compared to the feature engineered model. The same variables and framework were utilized for the predicted years models. What follows is a deeper summary of the finalized Predicted_AAV model.

1. **Predicted_AAV Model:**

The first image represents the code ran, along with the adjusted r-squared for the predicted_aav model (0.72).

```
> # RANDOMFOREST MODEL_1_AAV:
> 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)
>
>
> # RANDOMFOREST MODEL_1_AAV ELAVUATION:
> predictions_1 <- predict(rf_model_aav, newdata = test)
> mse_1 <- mean((predictions_1 - test$AAV)^2)
> print(mse_1)
[1] 12.96211
> rss_1 <- sum((predictions_1 - test$AAV)^2)
> tss_1 <- sum((test$AAV - mean(test$AAV))^2)
> r_squared_1 <- 1 - rss_1/tss_1
> print(r_squared_1)
[1] 0.8090542
>
>
> # RANDOMFOREST MODEL_1_AAV: ADJUSTED R-SQUARED: (0.72)
> n_1 <- nrow(test)
> k_1 <- 18
> adj_r_squared_1 <- 1 - ((1 - r_squared_1) * (n_1 - 1) / (n_1 - k_1 - 1))
> print(adj_r_squared_1)
[1] 0.7231285
```

The second image helps assess predictor importance in the random forest model. The importance() function provides

two key metrics: %IncMSE and IncNodePurity. %IncMSE represents how much the model's prediction error increases when each variable is altered. A higher value indicated a greater contribution to general prediction accuracy. IncNodePurity captures the total reduction in variance contributed by each variable across all trees. This measures how often and effectively a variable is used for splitting. Just like %IncMSE, a higher IncNodePurity value contributes to general prediction accuracy.

```
> importance(rf_model_aav)
                %IncMSE IncNodePurity
Walk.Year.WAR 29.534519     3682.3190
WAR           11.284835     2035.2289
PA             8.048225      394.6118
AVG            4.118815      252.7606
OBP            3.282707      289.2898
SLG            8.800367      550.2506
HR            10.465199      747.4530
RBI            8.698895      561.8865
R              9.528830      807.9770
BB.            2.696058      190.3621
K.             4.171440      243.5014
SB             6.310603      212.5316
ISO            6.419752      226.4671
wOBA           7.478679      443.1961
Off           10.014959     1613.9419
Def            6.993103      257.9011
BsR            3.445796      181.9204
Cluster        6.137246      547.8089
PC1           11.782620     1666.0486
PC2            4.428331      170.3852
```

Walk.Year.WAR was by far the most important predictor across both metrics. Essentially, this variable is the most influential variable when it comes to determining AAV. Other variables such as; PC1, WAR, HR, and Off, the other highest %IncMSE scores. This is confirmation that recent, high-level power and run creating performance drives player value more than anything else.

2. Predicted_Years Model:

The first image represents the code run, along with the adjusted r-squared for the predicted_years model (0.74).

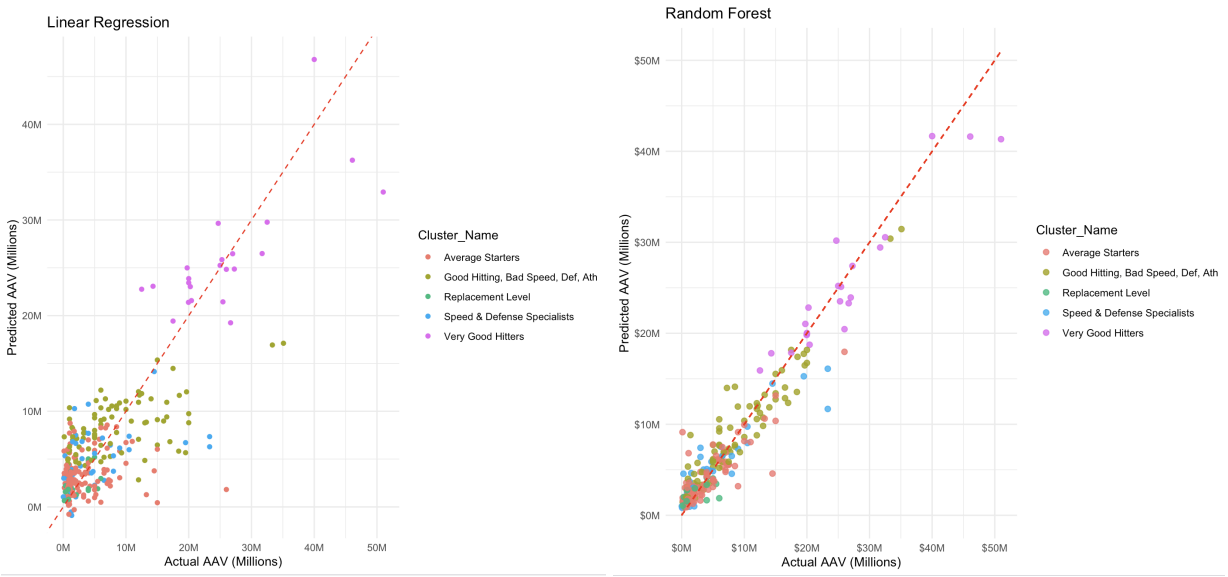
```
> # RANDOMFOREST MODEL_1_YEARS:
> set.seed(123)
> rf_model_years <- randomForest(Years ~ Walk.Year.WAR + WAR + PA + AVG + OBP + SLG + HR + RBI +
+   R + BB. + K. + SB + ISO + wOBA + Off + Def + BsR + Cluster + PC1 + PC2,
+     data = train,
+       ntree = 500,
+         mtry = 5,
+           importance = TRUE)
>
>
> # RANDOMFOREST MODEL_1_YEARS: ELAVUATION:
> predictions_1_years <- predict(rf_model_years, newdata = test)
> mse_1_years <- mean((predictions_1_years - test$Years)^2)
> print(mse_1_years)
[1] 0.5831408
> rss_1_years <- sum((predictions_1_years - test$Years)^2)
> tss_1_years <- sum((test$Years - mean(test$Years))^2)
> r_squared_1_years <- 1 - rss_1_years/tss_1_years
> print(r_squared_1_years)
[1] 0.8185634
>
>
> # RANDOMFOREST MODEL_1_YEARS: ADJUSTED R-SQUARED: (0.74)
> n_1_years <- nrow(test)
> k_1_years <- 18
> adj_r_squared_1_years <- 1 - ((1 - r_squared_1_years) * (n_1_years - 1) / (n_1_years - k_1_years - 1))
> print(adj_r_squared_1_years)
[1] 0.7369169
```

The second image once again helps us assess predictor importance in the model. Walk.Year.WAR was once against the most important variable followed by; WAR, PA, R, and Off. HR and RBI were near the top as well, which were in the top 5 for AAV. One interesting finding is the inclusion of PA (plate appearances). This makes logical sense, as plate appearance acts as a standard for durability and health. Obviously, the most plate appearances a player has, the less games they've missed due to injury. Teams value PA when determining a potential player's contract length. Just like the AAV model, this is confirmation that recent, high-level power and run creating performance drives player contract length more than other variables.

```
> importance(rf_model_years)
                %IncMSE IncNodePurity
Walk.Year.WAR 22.7360302     169.90756
WAR           10.3072253     125.86837
PA             7.0839658      29.97720
AVG            1.5705137      28.33936
OBP            2.2510794      19.51188
SLG            5.9604711      19.92185
HR             5.8609355      18.63544
RBI            4.8348996      13.36383
R              5.0447560      49.82163
BB.            0.7098892      14.10091
K.             1.2107750      13.24175
SB             2.0681156      11.48237
ISO            5.3009479      15.31402
wOBA           5.1770030      23.61076
Off            6.0876902      67.31353
Def            4.8694686      11.82119
BsR            2.8407090      21.20468
Cluster        3.5410297      26.64071
PC1            5.8730320      43.63859
PC2            3.5489419      16.27122
```

Hitters_Fa_Join: Random Forest vs Linear Model:

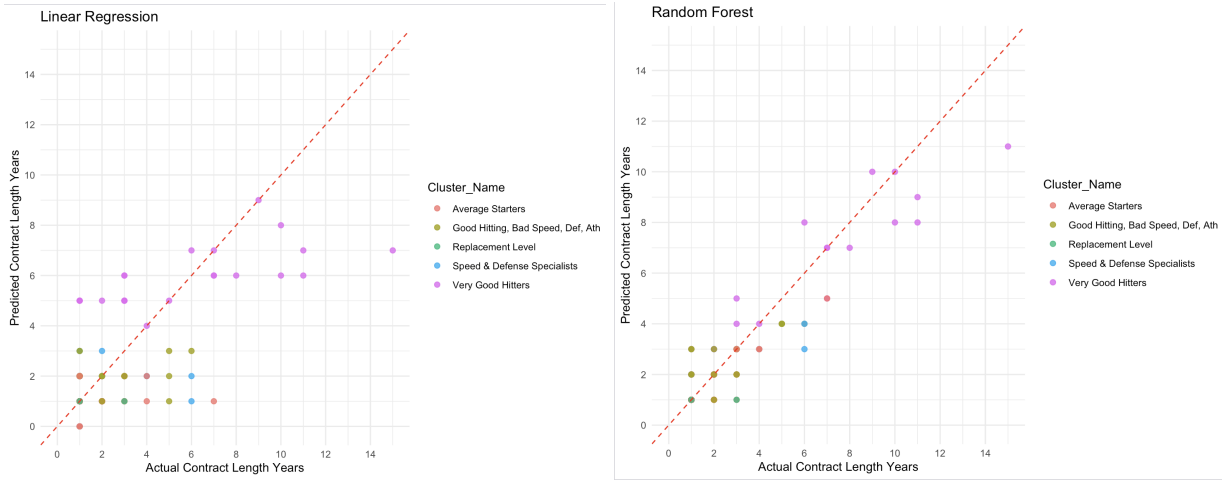
When comparing the random forest to the linear model, the first item of comparison is the model accuracy plot. On the left is the linear regression plot, while the random forest plot is on the right. Once again, the x-axis represents actual AAV values, while the y-axis is what each model predicted.



Visually, the random forest model shows a tighter clustering of data points around the line of best fit, especially in the higher salary ranges. In contrast, the linear model exhibits more dispersion and several clear outliers, particularly for elite-level free agents. It's important to clarify that the red line in both graphs is not the actual equation of the model, but simply a best-fit line for visual reference. While both models show some prediction error especially at the high end, the random forest model appears to reduce the magnitude of those errors and avoid the extreme mispredictions seen in the linear model.

To quantitatively evaluate model performance, RMSE (Root Mean Squared Error) was calculated for both the linear and random forest models when predicting AAV. The linear regression model yielded an RMSE of 9.36, while the random forest model produced an RMSE of 3.60. This suggests predictions were off by $9.36 million on average for the linear model, but just $3.60 million on average for the random forest model (AAV). The random forest model was more than twice as accurate for this comparison. This paired with the adjusted r-squared values for both models (0.64 linear vs. 0.72 rf) prove the random forest model captured underlying relationships that the linear model failed to catch.

Below is the same concept however; for years. The linear model is on the left, while the random forest model is to the right.



Unlike the AAV plots, this comparison is less clear due to the small range of integer values on both axes and overlapping points. Since contract years were rounded in both models, the distribution appears more discrete and clustered. Still, the random forest model shows tighter grouping along the line of best fit, indicating more consistent performance. The linear model has broader dispersion, especially for mid-range contract values, suggesting weaker fit. Despite the limitations of the rounded response variable, the random forest model again displays an advantage in predictive accuracy.

The RMSE once again was calculated for both models when predicting years. The linear regression model yielded an RMSE of 1.25, while the random forest model produced an RMSE of just 0.76. The linear model was an average of 1.25 years off when predicting contract length, while the random forest model was just 0.76 years off. This combined with the adjusted r-squared values (0.52 linear vs. 0.74 rf) prove the random forest models were much more accurate when it came to predicting years for a free agent's contract.

Pitchers_Fa_Join Random Forest Model:

The "pitchers_fa_join" data frame followed the same format as "hitters_fa_join". The first random forest model consisted of 8 variables which were; WAR, ERA, xFIP, K.9, BB.9, IP, vFA..pi, and Age. As a test run, this model produced an adjusted r-squared score of 0.62 when predicting for AAV.

The next random forest model consisted of the following 19 variables; 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. When predicting for AAV, this model yielded an adjusted r-squared value of 0.76. This was considerably better than our first random forest model, along with the linear model.

1. **Predicted_AAV Model:**

The first image represents the code run, along with the adjusted r-squared for the predicted_aav model (0.76).

```
> # PITCHERS_FA_JOIN: RANDOMFOREST MODEL AAV: (19 VARIABLES)
> 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)
>
>
> # PITCHERS_FA_JOIN: RANDOMFOREST MODEL AAV: (19 VARIABLES): EVALUATION
> predictions_aav_p19 <- predict(rf_model_aav_p19, newdata = test_p)
> mse_aav_p19 <- mean((predictions_aav_p19 - test_p$AAV)^2)
> print(mse_aav_p19)
[1] 7.728745
> rss_aav_p19 <- sum((predictions_aav_p19 - test_p$AAV)^2)
> tss_aav_p19 <- sum((test_p$AAV - mean(test_p$AAV))^2)
> r_squared_aav_p19 <- 1 - rss_aav_p19/tss_aav_p19
> print(r_squared_aav_p19)
[1] 0.826875
>
>
> # PITCHERS_FA_JOIN: RANDOMFOREST MODEL AAV: (19 VARIABLES): (0.76)
> n_aav_p19 <- nrow(test_p)
> k_aav_p19 <- 19  # number of predictors used
> adj_r_squared_aav_p19 <- 1 - ((1 - r_squared_aav_p19) * (n_aav_p19 - 1) / (n_aav_p19 - k_aav_p19 - 1))
> print(adj_r_squared_aav_p19)
[1] 0.7583464
```

The second image helps assess predictor importance in the random forest model. PC1 was the most important variable when it came to predicting AAV. This variable is originally from our PCA analysis, and captured overall pitching performances by combining key statistical indicators into a single score. The individual variables that had the strongest influence on PC1 were; WAR, xFIP, ERA, IP, and AAV. Pitchers that scored high PC1 values generally had stronger run prevention metrics. This resulted in more accumulated innings, as they commanded higher annual values in free agency. As a result, PC1 served as a comprehensive performance index and was used in both the clustering process, and as a predictor in the random forest model. Its high variable importance score confirmed its value in summarizing pitcher quality across multiple dimensions.

```
> importance(rf_model_aav_p19)
                %IncMSE IncNodePurity
Walk.Year.WAR 16.4995660     2510.8681
WAR           10.2999123     1526.6739
W              8.8633748      934.3536
L              9.5681414      642.1285
SV             7.2971979      247.5687
IP             8.6553343      702.3103
K.9            6.5577513      465.2782
BB.9          -0.2536372      277.5605
BABIP          3.8905624      247.1124
LOB.           4.2399292      265.5582
GB.            3.9998022      270.6307
HR.FB          5.8156757      274.9787
vFA..pi.       7.9657674      432.4187
ERA            5.6408164      348.6502
FIP            6.2807961      374.8303
xFIP           7.8670533      509.4719
PC1           22.7373463     2464.6525
PC2            6.2509544      463.1610
Cluster        9.8139872     1044.2877
```

The other variables that rounded out the top 5 were; Walk.Year.WAR, WAR, Cluster, and notably L (losses). While losses are traditionally viewed as a team stat, and are outdated in terms of judging pitchers, the model suggests that they may proxy for factors like pitcher usage, team context, or consistency over a full season. All of these variables captured recent performance, overall value, and pitcher archetypes.

2. **Predicted_Years Model:**

The first image represents the code run, along with the adjusted r-squared score for the predicted_years model (0.35).

While the other three random forest models scored very well (0.72-0.76), it's clear this model does a poor job of estimating pitcher contract length. Despite testing for both 8 and 19 variables, the adjusted r-squared score remains well below 0.50. This isn't due to a lack or abundance of variables, but rather the low signal quality in the current variable set. Most pitchers receive 1-2 year deals, creating a skewed low-variance target that performance metrics along can't effectively explain. Variables such as; injury history, velocity trends, and pitch sequence / mix are likely critical to front-office decision making when it comes to determining the contract years for a pitcher. Without these variables, it remains difficult to predict an accurate contract length for pitchers. It's clear the random forest models enhanced the other three predictive values; however, contract years for pitchers needs further engineering.

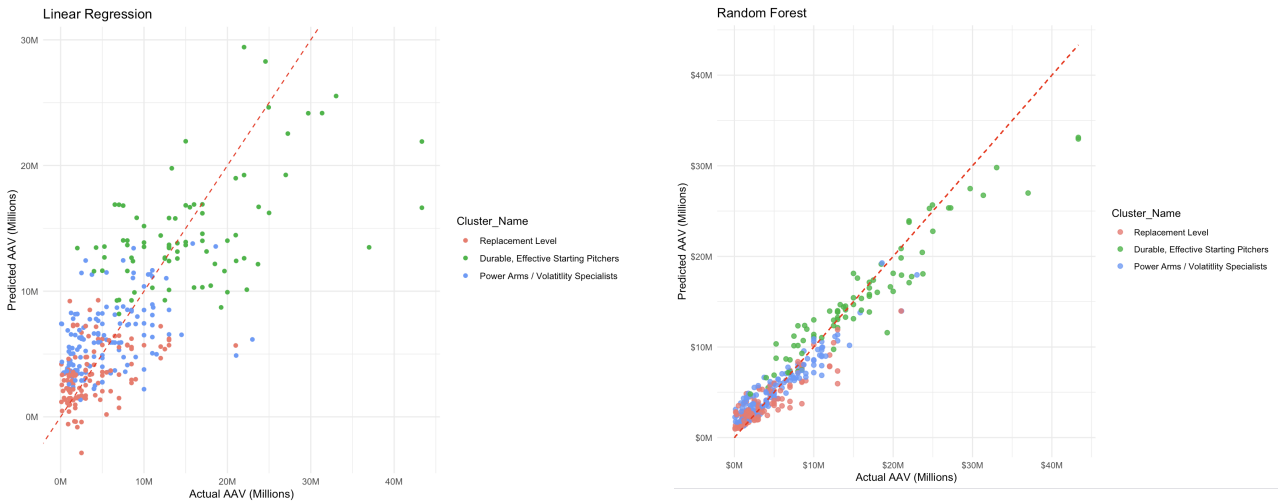
```
> # PITCHERS_FA_JOIN: RANDOMFOREST MODEL YEARS: (19 VARIABLES)
> set.seed(123)
> rf_model_years_p19 <- randomForest(Years ~ Walk.Year.WAR + WAR + W + L + SV + IP +
+   K.9 + BB.9 + BABIP + LOB. + GB. + HR.FB + vFA..pi. + ERA + FIP + xFIP + PC1 + PC2 + Cluster,
+     data = train_p,
+       ntree = 500,
+         mtry = 3,
+           importance = TRUE)
>
>
> # PITCHERS_FA_JOIN: RANDOMFOREST MODEL YEARS: (19 VARIABLES): EVALUATION
> predictions_years_p19 <- predict(rf_model_years_p19, newdata = test_p)
> mse_years_p19 <- mean((predictions_years_p19 - test_p$Years)^2)
> print(mse_years_p19)
[1] 0.8553456
> rss_years_p19 <- sum((predictions_years_p19 - test_p$Years)^2)
> tss_years_p19 <- sum((test_p$Years - mean(test_p$Years))^2)
> r_squared_years_p19 <- 1 - rss_years_p19/tss_years_p19
> print(r_squared_years_p19)
[1] 0.4410517
>
>
> # PITCHERS_FA_JOIN: RANDOMFOREST MODEL YEARS: (19 VARIABLES): (0.35)
> n_years_p19 <- nrow(test_p)
> k_years_p19 <- 9  # number of predictors used
> adj_r_squared_years_p19 <- 1 - ((1 - r_squared_years_p19) * (n_years_p19 - 1) / (n_years_p19 - k_years_p19 - 1))
> print(adj_r_squared_years_p19)
[1] 0.3543184
```

PC1 was the most important predictor in the model, narrowly outperforming Walk.Year.WAR. These were followed by WAR, vFA..pi., and FIP to round out the top five. While the model underperformed in predicting contract length, analyzing which variables contributed most provides insight into its structure and limitations. The dominance of performance related metrics like WAR, FIP, and velocity highlights that the model relies heavily on statistical output, yet lacks critical context such as injury history or long-term risk factors. To improve accuracy in future iterations, more contextual variables not present in the current data frame such as; IL stints, pitch mix trends, and durability indicators should be incorporated.

```
> importance(rf_model_years_p19)
                %IncMSE IncNodePurity
Walk.Year.WAR 14.453884     34.552122
WAR           11.554795     24.834961
W              5.354633     16.884498
L              5.012857     12.550975
SV             5.668229      4.923350
IP             8.308807     16.954928
K.9            7.236862     11.634289
BB.9           4.153525      7.169061
BABIP          2.460269      5.737945
LOB.           4.520355      6.997389
GB.            1.532765      6.876633
HR.FB          5.467438      8.597971
vFA..pi.       9.245442     11.223153
ERA            4.357381      7.073962
FIP            8.896229     10.475560
xFIP           7.012655     11.795741
PC1           15.112070     42.317804
PC2            5.414391     11.077390
Cluster        5.331826      7.817304
```

Pitchers_Fa_Join: Random Forest vs Linear Model:

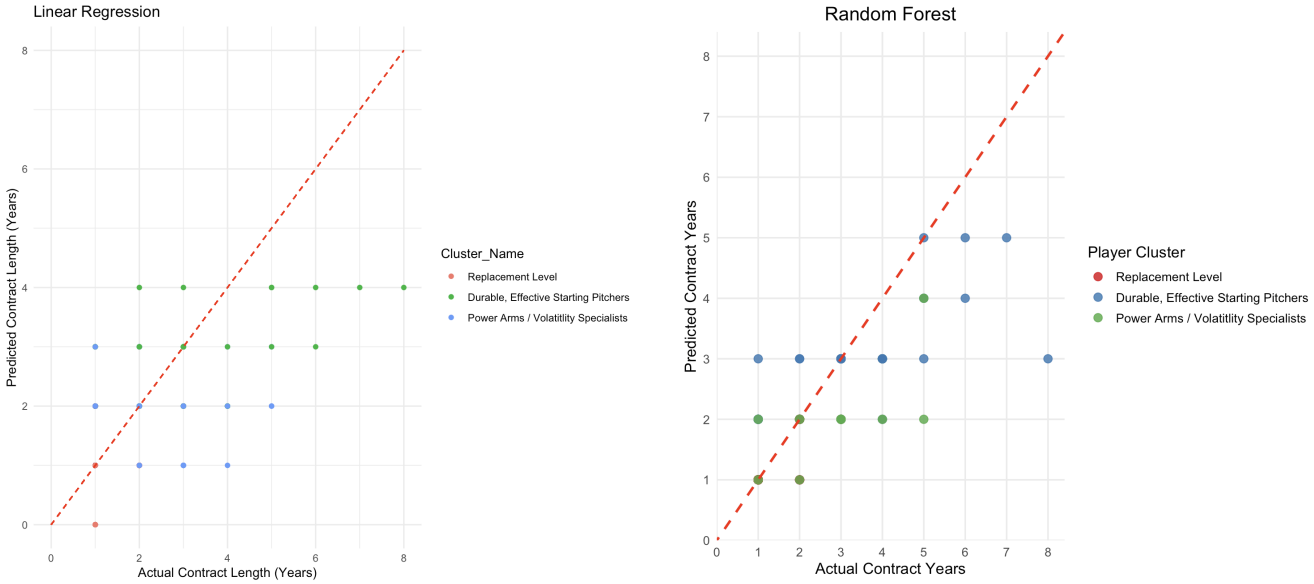
The first item of comparison between these two is the model accuracy plot. The linear regression plot is to the left, while the random forest plot is on the right. The x-axis represents actual AAV values, while the y-axis is what each model predicted.



Visually, the random forest model displays a much tighter concentration of predictions around the line of best fit. This is especially noticeable at the upper end of the AAV range, particularly for pitchers clustered in the elite or high-variance groups. The random forest model better handles non-linear patterns, reducing the magnitude of extreme mispredictions and aligning more closely with actual AAV values across clusters. It's important to note that the red line in both graphs is a best-fit reference line, not the actual prediction equation. While both models show error at the margins, the random forest's ability to minimize outliers and adjust to non-linear interactions results in a visibly more accurate and stable predictive distribution.

To quantitatively evaluate model performance, RMSE was calculated for both the linear and random forest models when predicting AAV. The linear regression model yielded an RMSE of 4.86, while the random forest model produced an RMSE of 2.78. Predictions were off by $4.86 million on average for the linear; however, the random forest model was off by $2.78 million (AAV). This paired with the adjusted r-squared values for both models (0.52 linear vs. 0.76 rf) prove the random forest model captured underlying relationships much better that the linear model failed to catch.

Below is the same concept for years. The linear model is on the left, while the random forest model is to the right.



Unlike the AAV plots, the comparison for predicted contract length is inherently less distinct due to the small integer range of years and the rounding applied in both models. This results in overlapping points and a more discreet appearance across both plots. Even so, the random forest model displays a tighter concentration of predictions near the identity line, particularly for players receiving 2 to 5-year deals. In contrast, the linear regression model shows flatter predicted values and wider vertical dispersion, especially in the mid-range indicating underfitting. It's also worth noting that some cluster groups, such as Replacement Level pitchers, may be present but not visually distinguishable due to overlapping coordinates. Despite these visual constraints, the random forest model again demonstrates a performance advantage, capturing more nuanced relationships in contract length predictions.

The RMSE was calculated for both models when predicting years. The linear regression model yielded an RMSE of 0.83, while the random forest model produced an RMSE of 0.92. This along with the adjusted r-squared scores (0.38 linear vs. 0.35 rf) prove the linear model slightly outperformed the random forest model.

This outcome suggests that contract length may follow a more linear relationship with the chosen predictors compared to AAV. Teams often anchor contract length to easily interpretable variables like age or recent WAR, which may exhibit more consistent, additive effects, favoring linear modeling. Additionally, the smaller variance in contract years (compared to AAV) may reduce the benefit of non-linear models like random forest, which tend to excel in capturing complex patterns or interactions that may not be as present in this outcome variable. It’s directionally sound, but vague. Sharpen the wording to reflect the low explanatory power

and propose a more specific next step. While the linear model slightly outperformed the random forest both models performed poorly overall, as the adjusted R-squared scores were both below 0.40. This suggests that the current predictors fail to capture the key drivers of contract length. Future models should incorporate external variables such as; injury history, team needs, or market scarcity, to improve predictive accuracy.

2026 Free Agent Predictions Hitters:

Moving onto the 2026 free agent class of hitters, most of the infrastructure of the data frame was already set up due to part 1 of the project. The only column that needed work was the Walk.Year.WAR column. Since Walk.Year.WAR wasn't a variable utilized in the linear model, this column wasn't an issue in part 1. Since the MLB season is currently happening, these upcoming free agent's Walk.Year.WAR available numbers aren't available. Since this variable was the most important in terms of the %IncMSE, it had to be included in the model. To estimate each upcoming free agent's Walk.Year.WAR, a new current data frame had to be uploaded to R, consisting of each free agent's 2025 stats to date. This new data frame "war_2025_clean" consisted of each player's 2025 War and games played. This was merged into the existing data frame "fa_2026_hitters_stats", by utilizing "Name_clean". After column renaming, and eliminating players who didn't meet the 2026 free agent criteria, the current 2025 war to date column (War_to_date), along with the amount of games played so far in 2025 were updated. To estimate their current pace and trajectory, their current war to date was divided by how many games they've played so far in 2025. This essentially produces how much WAR a player has produced per game so far. This number was then multiplied by 162 (162 game season), to create their final estimated "Walk.Year.WAR".

While a lot can change over the second half of the season, this was the most effective way to estimate these player's Walk.Year.WAR. It will be very interesting to see how much these predictions can change for a player if they have an excellent, or negative second half of the season. After this process, both random forest models were added to this dataframe, effectively predicting each player's AAV and contract length.

Two additional variables, PC1 and PC2, also required imputation in the "fa_2026_hitters_stats" data frame. These values originally came from "hitters_fa_join", so a merge was performed to incorporate them. Because not all upcoming free agents had signed contracts between 2022 and 2025, some players were missing PC1 and PC2 values. Initially, these missing values were filled using the overall median for each variable. However, this approach disproportionately impacted elite-level free agents by assigning them average values that didn't reflect their performance. To improve accuracy, missing values were instead imputed

using the median PC1 and PC2 values within each player's respective cluster group. This method better preserved group characteristics and avoided penalizing elite players with missing data.

The data frame shown is the upcoming 2026 class of free agent hitters, filtered by the random forest AAV values from highest to lowest. This view highlights the top 33 projected free agent deals based on predicted AAV. The linear model's predictions don't have any abbreviation at the end of them (Predicted_AAV, Predicted_Years, Total_Salary) . However, the random forest model's predictions all end in "rf" (Predicted_AAV_rf, Predicted_Years_rf, Total_Salary_rf"). This screenshot focuses on the differences between models for the top 33 predicted free agent hitters.

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

Showing 1 to 33 of 87 entries, 29 total columns

The random forest model produces more accurate AAV predictions than the linear regression model, especially for elite free agents. Although differences in AAV are not extreme, the random forest model assigned higher AAV values in 29 of the 33 cases shown, indicating a stronger alignment with actual market behavior. Unlike the linear model, which compresses predictions around the mean, the random forest captures non-linear relationships between player statistics and salary outcomes. This is particularly evident in the valuation of players like

Alex Bregman, Kyle Tucker, Pete Alonso, and Cody Bellinger, whose predicted AAVs better reflect market expectations.

The model also excels at distinguishing between tiers of players, producing a wider and more realistic spread in salary estimates. High-performing players with strong WAR, favorable age curves, and solid offensive metrics are appropriately valued without inflating the salaries of replacement-level players. For example, low-tier players are assigned modest total salary projections, avoiding the upward distortion common in the linear model.

By assuming constant marginal effects, the linear regression model systematically underestimates elite players and overestimates fringe contributors. In contrast, the random forest model's flexibility and variable interaction handling allows for more responsive and realistic valuation.

While both models offer insight, the random forest clearly outperforms in projecting AAV for top-tier hitters. To further improve predictive accuracy, future models should incorporate external market dynamics such as; team payroll constraints, agent influence, and player marketability to capture value drivers beyond performance metrics.

2026 Free Agent Predictions Pitchers:

Now onto the 2026 free agent class of pitchers, again most of the infrastructure and data was already in place. However, estimating Walk.Year.WAR took on a different approach compared to before. Initially, Walk.Year.WAR was estimated the same exact way as the "fa_2026_hitters_stats" data frame. However, since some relievers had performed very well in limited opportunities, this framework estimated some player's WAR to be as high as 35.0. Since this is extremely unrealistic and impossible, another approach was taken. At the time of calculation for this segment, the duration of the MLB season had reached the 45% mark. Walk.Year.WAR was calculated by simply dividing each player's "2025.WAR.td" (their current WAR), by 0.45. This gave a projection of what each player was on pace for regarding the 2025 season. This approach resulted in logical 2025 WAR values, and is how "Walk.Year.WAR" was ultimately estimated.

Just like the "fa_2026_hitters_stats" data frame, a few more columns had to be imputed and estimated. PC1, PC2, and HR.FB values had to be merged from "pitchers_fa_join". Those who didn't sign any contracts from 2022-2025, this resulted in NA values for those columns. Just like the "fa_2026_hitters_stats" data frame, these values were estimated based on the medians for each respective cluster group.

The data frame shown is the upcoming 2026 class of free agent pitchers, filtered by the random forest AAV values from highest to lowest. This view highlights the top 33 projected free agent deals based on predicted AAV. The linear model's predictions don't have any abbreviation at the end of them (Predicted_AAV, Predicted_Years, Total_Salary). However, the random forest model's predictions all end in "rf" (Predicted_AAV_rf, Predicted_Years_rf, Total_Salary_rf"). This screenshot focuses on the differences between models for the top 33 predicted free agent pitchers.

	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

The random forest model improved accuracy over the linear model in predicting AAV for free agent pitchers. In 21 of the 33 pitcher cases shown, the random forest prediction exceeded and more accurately matched the expected contract figures based on past deals and player performance tiers. While 21/33 was lower than "fa_2026_hitters_stats" 29/33, this was the

most accurate model in the entire project, with an adjusted r-squared score of 0.76 in the "pitchers_fa_join" data frame.

The linear model assumes constant marginal effects across all variables, leading to underestimation for pitchers with elite outlier metrics and overestimation for aging or injury-prone arms. By incorporating variable interactions and conditional splits, the random forest model allows for more accurate and differentiated predictions.

Ultimately, while both models provide baseline insights, the random forest model clearly excels in valuation realism for the 2026 free agent pitcher class. Future improvements should include variables and data such as; injury history, velocity trends, pitch sequence / mix, team-specific demands, and player marketability.

Findings on Past Free Agent Contracts:

	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	avisaíl 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

	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

The tables above display the 15 largest overpays from 2022–2025 for hitters, comparing predictions from the linear regression model (left) and the random forest model (right). It's immediately clear that the random forest model produced more accurate AAV predictions. For example, the 15th largest overpay in the linear model had a difference of $10.18M, while the 15th in the random forest model was only $3.83M off, a significant improvement.

Shohei Ohtani offers a compelling case study. Ohtani signed a 10-year, $700M deal with the Los Angeles Dodgers, theoretically equating to $70M/year. Fangraphs, however, split this into $46.08M as a hitter and $23.92M as a pitcher. Since our models focus exclusively on hitter contracts, only the $46.08M figure was used. Notably, Ohtani didn't appear in the top 15 overpays for the linear model and ranked 16th overall. The linear model predicted his hitter AAV at $36.25M, much closer to the actual than others, yet still less accurate than the random forest model, which predicted $41.04M. Despite a smaller margin of error, Ohtani still appeared on

the random forest table. This illustrates how the random forest model better captured high-end market behavior and delivered stronger predictive accuracy, especially for elite hitters.

	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

	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

These tables also display the 15 largest overpays from 2022-2025, comparing models but with pitchers this time. Once again, it's clear the random forest model (right) produced more accurate predictions than the linear model (left). The 15th largest overpay in the linear model had a difference of $7.97M, while the 15th in the random forest model was only $4.32M.

The linear model severely underestimated top-end contracts, most specifically with; Max Scherzer, Jacob DeGrom, and Justin Verlander, all of which were off by at least $21M. In contrast, the random forest model's worst difference was not even half of the linear model at $10.37M. This displays how much better the random forest model captured market dynamics, particularly at the top of the free agent pitcher market.

Key Takeaways:

This phase of the project highlights the advantage of incorporating machine learning techniques, particularly random forest modeling, to improve the prediction of free agent contracts. Compared to linear regression, the random forest model captured complex, non-linear relationships between player performance and contract outcomes, especially for elite-tier players who were consistently undervalued by the linear model.

Ultimately, this project underscores the importance of using non-linear, flexible models in sports contract forecasting. Accounting for interactions, segmentation, and imputation strategies can significantly enhance the realism and utility of predictive outputs in high-stakes markets like MLB free agency.

While the random forest models performed much better than the linear models, there's always room for improvement. Next time around, external variables such as; market scarcity,

agent influence, team dynamics / goals, player marketability, and pitch trends / velocity should be incorporated for improved accuracy.

Future Work:

This project was fueled by a mix of interest in data analytics, and a personal passion for the game of baseball. Throughout this extensive project, I enhanced my data analytics skills in numerous ways. This is a project I may revisit in the future; however, I'd like to start writing scouting reports for potential breakout candidates in the future. If I do revisit this project, I'll be mainly interested in improving the adjusted r-squared score for the "pitcher years" model, along with overall enhanced accuracy. Incorporating external variables would also be a priority in the future.

Appendix:

Data Sources:

- Fangraphs.com - Updated player statistics through June 12, 2025.
- Spotrac.com - Free agent contract details, and upcoming free agent lists.
- Datasets created or updated in R:
    - hitters_fa_join.csv
    - pitchers_fa_join.csv
    - fa_2026_hitters_stats.csv
    - fa_2026_pitchers_stats.csv
    - war_2025_clean
    - war_2025_clean_p

Data Dictionary:

- WAR: Wins Above Replacement.
- OBP: On Base Percentage.
- SLG: Slugging Percentage.
- HR: Home Runs.
- BB.: Walk Percentage.
- K.: Strikeout Percentage.
- SB: Stolen Bases.
- BsR: Baserunning Runs.
- Age: Player Age.
- ERA: Earned Run Average.
- xFIP: Expected Fielding Independent Pitching.
- FIP: Fielding Independent Pitching.
- K.9: Strikeouts per 9 Innings.
- BB.9: Walks per 9 Innings.
- IP: Innings Pitched.
- vFA..pi.: Average Fastball Velocity.
- AAV: Average Annual Value of Contract.
- Years: Length of Contract in Years.
- Cluster_Name: Categorical Label from K-means Output.
- PC1, PC2: Principal Components from PCA.
- Walk.Year.WAR: Projected 2025 WAR using Per-Game Extrapolation.
- HR.FB: Homerun to Fly Ball Ratio.

All Modeling specifications, variable imputations, and assumptions are fully detailed in the main body under each respective section.