Using Forecasted Statistics and Classification to Predict the MLB Silver Slugger Award

Kenny McAvoy, Daniel Quinn, Edward Yuan, Rayyan Karim

1 Abstract

This is the final paper for our Data Science course project. The ultimate goal of this project is to predict future individual statistical projections through regression methods, and then using those statistical projections to classify winners of the MLB Silver Slugger award by playing position.

2 Introduction

Every year, the MLB gives out an award to one player in each position for each league, the American League (AL) and the National League (NL), to recognize the best overall hitter at that position. These awards are called the Silver Slugger awards, and we seek to develop a model that can accurately project individual players' stats going into a season, and then use those projected stats to predict the winners of the Silver Slugger award. In order to predict the winners of the Silver Slugger award for an upcoming season, there are two major steps that we need to take. First, we have identified five major statistics that are generally used by the media to evaluate good hitting: runs, home runs, runs batted in, batting average, and OPS (on-base plus slugging percentage). Using data from past seasons, we ran correlation analyses on statistics to determine which other sub-statistics are good predictors of runs, home runs, etc. Once important sub-statistics were identified, we developed regression models using Linear Regression, Linear Support Vector Regression (SVR), and Support Vector Regression to predict values for the five major statistics for each player. Finally, we used our predicted statistics to develop classification models that predicts winners of the Silver Slugger award.

3 Related Work

Creating statistical projections has been a task many other companies do. The methods to do so are varied, from manually compiled projections based on the "eye" test to projections yielded from various algorithms. Predicting MLB player performance for the upcoming season is a common task for objectives such as creating pre-season fantasy baseball rankings or overall team performance projections. ESPN creates fantasy baseball projections by using a weighted average of players' three most recent seasons while factoring in batted ball data, age, park factors, etc. We take the same kind of data-centric model approach to create our projections by analyzing past data and using advanced statistics. There are other projection models, but it is generally agreed upon in all models that the past three seasons are the baseline for future projections. As far as we know, there is no existing classifier to predicting Silver Slugger winners specifically. Experts cast their own predictions based on their knowledge, but there are no known models that specifically predict the Silver Slugger.

4 Problem Definition

Who will win the 2019 Silver Slugger awards at each position in the MLB, based on our statistical predictions for the 2019 season? How well does our model predict this result based on previous years' winners?

5 Data Preprocessing

Data from the 2014-2019 MLB seasons were extracted from FanGraphs and used in the regression model. This data had 49 different columns of hitting statistics for each player who had at least 200 plate appearances in that given MLB season. One challenge with the data was that players who played on multiple teams in a season were assigned to no teams in the data set. This was manually filled by the team based on which team the player played the most games for. This was important because we thought some player statistics might be dependent on the team's overall performance or the home ballpark characteristics for a player. Another preprocessing issue was that some statistics had to be scaled using sklearn so that our regression models created reasonable projections. This means that an input sub-statistic, such as OBP, is no longer an on-base percentage but instead is a measure of the number of standard deviations above or below the mean OBP across the MLB. We also needed to account for retired players who did not play in the season we would be predicting, but this was easily accounted for after the predictions were made because we only made predictions for those players who were included in the dataset used for predictions; they were, therefore, given empty predictions and could easily be removed.

6 Methodology

Our first step was to obtain the hitting statistics data for both teams and individual players for the past three years. We obtained all our data from Fangraphs. Our data included many sub-statistics that are not in the five major statistics, such as ground ball percentage, speed, strikeouts, etc. For each of the five statistics, we ran a correlation analysis with the 2014-19 data to determine which sub-statistics are best predictors of the major statistics. Below is a table of the correlation analysis results.

Table 1. Correlation Analysis Results for Sub-Statistics

Major Statistic	**Sub-Statistics that factor into the major statistic regression**	**Correlation**
Runs	wRC	0.936
	PA	0.914
	H	0.903
	AB	0.893
	RBI	0.818
	G	0.794
	2B	0.795
Batting Average	BABIP	0.755
	OBP	0.737
	wOBA	0.734
	OPS	0.724
	wRC+	0.713
	wRAA	0.676
	1B	0.673
Home Runs	RBI	0.881
	ISO	0.843
	wRC	0.802
	SLG	0.795
	R	0.741
	HR/FB	0.738
	OPS	0.711
Runs Batted In	wRC	0.888
	HR	0.881
	PA	0.839

	AB	0.825
	R	0.818
	H	0.818
	2B	0.755
On Base Percentage + Slugging (OPS)	wOBA	0.992
	wRC+	0.973
	SLG	0.957
	wRAA	0.935
	OBP	0.827
	ISO	0.786
	HR	0.711

From these correlation statistics, we used the top seven highest correlated sub-statistics to build our regression models for each major statistic. The training data for the prediction is the statistics from the three seasons prior to the year being predicted/tested. Using a weighted average of the last three years of data, we used the regression model to predict 2017, 2018, 2019, and 2020 MLB statistics for those players. The ideal weighted average methodology is found in the table below (if a player only played in year 1 and year 2, we weight year 1 with 0.3 and year 2 with 0.7, and if a player only played one year, they receive a weight of 1. This is not an issue for the final results, since our predictions filter out players who did not play in the year being predicted).

Table 2. Season Weighting Methodologies

	Year - 3	**Year - 2**	**Year - 1**
Played all 3 years	0.1	0.3	0.6
Two years	-	0.3	0.7
One year	-	-	1

Once regression models were created for the predicted statistics for the 2017-20 MLB seasons. We tested give different methods to forecast the 2019/2020 seasons: Linear, Linear SVR, SVR, Huber, and Ridge. Each of these models employ different methods to make predictions. Linear Regression attempts to create a linear fit to which future data points are predicted against. Linear SVR works in a similar method using support vector regression with a linear kernel. SVR is the same as linear SVR without the support vector kernel. The Huber Regressor optimizes the squared loss for the samples where |(y - X'w) / sigma| < epsilon and the absolute loss for the samples where |(y - X'w) / sigma| > epsilon, where w and sigma are parameters to be optimized. The Ridge model is a regression model that uses the loss function as the linear least squares function and regularization given by the l2-norm.

The next step we took was to use two different methods of classification to predict the winners of the silver sluggers in each of those 4 seasons using our predicted statistics. We trained both the K-Nearest Neighbors and the Naive Bayes classifiers on the 3 years prior to the year we were predicting for to develop a classification model for how Silver Slugger winners are decided. Before being able to run the models, we had to do some preprocessing. We divided the player data first by league, and then within each league divided by each position.

For K-Nearest Neighbors, we used the scikit-learn KNeighborsClassifier to develop the model. As an additional step, we scale the data using the same sklearn method used in the regression models. As seen below, scaling the data before training the models yielded better classification results.

Table 3. KNN Average Metrics Over 2017-2019

	Scaled	Unscaled
Ranking	11.31	8.78
Recall(5)	12.33	11
Recall(4)	12.33	10.67
Recall(3)	11	9.33
Recall(2)	7.33	6.33
Recall(1)	3.67	3.67

Then, for each position, the average profile of a winner is calculated by averaging the statistics of the winners of the past three years at that position across both leagues. Because the position data will have very few winners, we append the other league's winners at the given position over the past three years. We then sort the position data by calculating the Euclidean distance of each player's stats to the average profile of a winner at their position. Now, we train a KNN model on each subset of our data, setting the k value to 1. Then, we test on the projected statistics for the upcoming year. If our model does not predict a minimum of five winners, we simply append players with a profile nearest the average winner profile until we have five winners at each position.

For Naive Bayes, we used the scikit-learn GaussianNB classification method to develop the model. The training data was obtained from each of the specific league-position data files using the three years prior to the year being tested. A column labeled "SS" was added to the files with a simple "Y" or "N" distinction to indicate if that player won the Silver Slugger award that year. The GaussianNB classifier was then trained on the values according to this column, and a prediction of the top candidates for Silver Sluggers was made using which players had the highest probability of classifying as "Y." A list of these players sorted by probability (most likely to win) was then created, and we could use this list to get the top *n* finishers for later evaluation.

We used the 5 main statistical categories listed above--R, HR, RBI, AVG, and OPS--to train our classification models, and we then ran the models on the 2017-20 seasons to get predictions for each of those 4 seasons. Once we had our predictions for each of those seasons, we used the actual results for the Silver Slugger awards for 2017-19 to evaluate how accurate our classification model was; since the 2020 season has not started yet, that was just for our own interest and could not be evaluated yet.

7 Results and Evaluation

After training our linear regression model on the testing data, we were able to apply the model to make predictions on both the sub-statistics and the major statistics. An example for runs scored is shown below.

Table 4. Predictive Output for Runs Scored, Top 3 Finishers

Player Name	Linear Regression Predicted Runs Scored	Linear SVR Predicted Runs Scored	SVR Predicted Runs Scored
Charlie Blackmon	111.07	106.48	97.26
Paul Goldschmidt	105.81	104.28	98.38
Francisco Lindor	105.60	102.34	94.68

We did the same process to obtain predicted statistics in all five categories. Next, we downloaded actual 2019 statistics from FanGraphs to do comparisons and evaluations. We calculated mean absolute error (MAE) and root mean squared error (RMSE) for each of the five major statistics (R, HR, RBI, AVG, OPS) between the actual values each player accrued in 2019 and the predicted values for each player, using sklearn's metrics module for the computations. The error rates can be seen below in the table below:

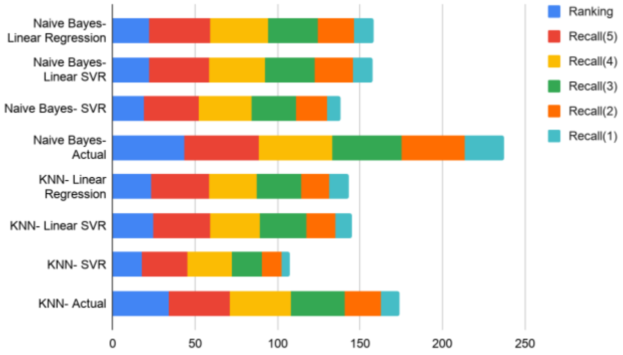
Table 5. MAE and RMSE for Linear SVR

	R	HR	RBI	AVG	OPS
MAE	18.00	6.51	18.08	0.027	0.076
RMSE	22.79	8.41	22.76	0.035	0.097

We see that AVG and OPS, which are not volume based stats, are more accurate, being very close to 0. The volume based stats (R, HR, RBI) have higher error. Part of the reason these stats have a higher error rate is that players might have drastically more or less plate appearances than they have had in the past three years. It is virtually impossible to accurately predict the number of plate appearances a batter will have in a season because of things such as injury or trade.

To evaluate our classification models, we ran the classification on each of our regressed statistics and measured the accuracy by counting the number of correct predictions we made (recall)--but also expanding that to the number of winners in our top *n* predictions--and by creating a ranking measure that is calculated by summing up each player's 1/*n* placement in our predictions, where *n* is where that winner placed in our rankings. Since it is difficult to get the exact winner for each position based on the number of players and how close some of them are in production, having these measures gives us an idea of how close we are to predicting the winners as well even if we don't predict the exact winners correctly every time. Figure 6 shows the breakdown for how each of our classification-regression model combinations performed according to these evaluation metrics.

Figure 6. Classification Evaluation



Based on the results, Naive Bayes is clearly the better classifier across the board, and it has significantly better results when using the actual season statistics than when using our projections, even compared to KNN. A sample of Naive Bayes classification on the Linear SVR regressions for 2019 can be seen below.

Table 7. Naive Bayes Classification Winners based on 2019 Linear SVR projections

	AL	NL
C	Gary Sanchez	J.T. Realmuto*
1B	Jose Abreu	Max Muncy
2B	Jose Altuve	Javier Baez
3B	Jose Ramirez	Nolan Arenado
SS	Francisco Lindor	Manny Machado
OF	Mookie Betts*	Bryce Harper

OF	Mike Trout*	Charlie Blackmon
OF	J.D. Martinez	Christian* Yelich
DH	J.D. Martinez	

** indicates actual 2019 award winner*

8 Conclusion

Many of the regression models used produce similar results, but we see that Linear Regression and Linear SVR are the regression models that produce the best classification results. In addition, it is shown in Figure 6 that Naive Bayes significantly outperforms the KNN classifier. Thus, using Linear Regression or Linear SVR to project statistics combined with using Naive Bayes to classify the Silver Slugger award winners will yield the best predictions for the upcoming season. Ultimately, it is difficult to come up with exact predictions for the Silver Slugger award even with our best models, but our top 3 predictions at each position do include the eventual Silver Slugger winner 64.5% (11/17) of the time, as seen in the results in Table 3. In general, this is consistent with how we see sports awards handed out in real life. Elite players consistently play well and win their fair share of awards, but there is variation in which one actually wins the award each year. In addition, there can be breakout stars that experts could not account for at the beginning of the season. This makes the task of predicting Silver Slugger winners a difficult one, even with the use of data science.