Nicholas Sullivan

Md Mainul Islam Mamun

CSCI 4930 Big Data Analytics

11 December 2022

A Machine Learning Approach to MLB Player Performance and Salary Evaluation

**Introduction**

The modern statistical approach to the game of baseball was revolutionized by Bill James, the “godfather” of sabermetrics and popularized through the book and movie *Moneyball* about Billy Beane and the Athletics. As teams adopted these advanced analytical methods over the past twenty years, teams now face a new challenge to finding a competitive edge when everyone is now relying on advanced sabermetrics. Attempting to model some of these teams’ internal analytical systems, I take one approach to evaluating current MLB hitters using several basic machine learning models. My goal was to use a combination of traditional and advanced statistics aside salary data to discover under- and over-valued players, which could help guide decision making for a front office. While the analysis and data used could be altered to change the focus on free agents or minor leaguers, my specific results would be optimal recommendations for trade targets. A competing team would be inclined to find these value players that other teams may not see the full value of, or a selling team may want to trade players they view as over-valued but other teams think has more true value.

**Background**

MLB provides a wealth of data to the public about player’s statistics and pitch by pitch game logs on the Baseball Savant website. For this evaluation, the standard statistics included for batters range from traditional batting average to average release extension and more. I removed a few statistics at the beginning that I was unsure of, such as release extension and fielding position against of all fielders. My analysis only factored in offensive statistics, and therefore does not account for the defensive value of a player. The statistics I used were from the complete 2021 season, which I paired with salary data from the beginning of the 2022 season, according to USA Today’s database. For documentation of each statistic, refer to Baseball Savant. This was the most direct way to analyze players current value at the start of the 2022 season. Due to many different players’ contract status as rookies, newly signed, or veterans several years into a deal, this should not be treated as an analysis of how well contracts evaluate the previous year. Rookies can way outperform their minimum contract that they may be locked into, and veteran players can depreciate over later years of their contract, while still getting paid a pre-negotiated amount based on stats before 2021. One change to the data that that was important was to change the count variables of swings, whiffs, and takes into swinging strike percentage. This is calculated as:

The primary inspiration for this report stemmed from a systematic literature review by two Miami University students on machine learning applications in baseball (Koseler). They discuss the different ways people have analyzed aspects of the game with corresponding machine learning methods and offer many proposals of ways more machine learning can be used to predict baseball analytics. Another specific project report by Tatsuya Ishii completed a similar analysis, using clustering methods to identify undervalued players (Ishii). With limited detail summarized in Ishii’s results presentation on pitchers, I aimed to follow a similar path and discover the results for the 2022 season for batters specifically.

**Project Specification**

This background information did not represent a full analysis of players, but rather I was able to fine tune my methods and goals. To put the focus of this project on the education of machine learning models and understanding the power those may have in the future of baseball analytics, there are insufficient factors included to make a full evaluation of a player. The models are built only using offensive statistics with an equal consideration for all salaried and healthy baseball players in 2021 and 2022. Additionally, the focus was put on classification methods although other methods such as regression are potentially just as or more valuable for certain circumstances. By focusing on clustering, I could compare results with Ishii’s project.

**Problem Analysis and Solution Design**

The three clustering algorithms modeled were K-Means, Hierarchical, and 3 types of Vector Machine with Support Vector Machine being the most complicated and likely having the highest accuracy. For k-means clustering I chose a k of three clusters based off the elbow method as seen in Figure 1. The elbow method shows the within clusters sum of squares, which is ideally minimized, as that means the data within each cluster is highly similar, and therefore predictable. However, as the number of clusters keeps increasing the sum of squares marginally decreases to a point that overfitting the data with more clusters is less effective. Therefore, the largest elbow in the plot, often shown by the largest change in slope represents the ideal number of clusters. While it appears to be 2 in Figure 1, 3 clusters were preferred in this situation since only two clusters of salary range were not descriptive or detailed enough. Additionally, I scaled my data so that all values were measured on a scale of -1 to 1, since some attributes were count variables and others were percentages. This balances out the weights given to the variables.

Hierarchical clustering has the added benefit of not needing to specify a certain number of clusters beforehand. Different clustering methods such as complete linkage, mean linkage, and Ward’s minimum variance work better with certain types of data distributions, as they cluster according to different methods. For this project, Ward’s minimum variance had the highest agglomerative coefficient, a measure of the distance between clusters. After running the model and producing a dendrogram tree, the gap stat method is used, as seen in Figure 2, which calculates a gap stat for different numbers of clusters. The highest gap stat is usually preferred, but a cluster of one does not make sense, so the optimal cutting of the dendrogram is into four clusters. Figure 3 is the created dendrogram split into four clusters.

Finally, there are three vector machine models to cluster data: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and support vector machine (SVM). LDA usually assumes that the data used is normally distributed and the covariance matrices are identical. To test for normal distributions by attribute the Shapiro-Wilkes Normality test was used. Although only LDA assumes normality, results are compared in Figure 4 of the accuracy of each model using only normal variables and all variables for each model. The accuracy values were calculated from confusion matrices, which required preemptive groupings of players by salary into five salary ranges. Those ranges were from 0-1 million, 1-5 million, 5-15 million, 15-25 million, and more than 25 million annual salary in 2022. This represented arbitrary groupings of players of a similar salary range, with small enough ranges to distinguish between each group’s assumed talent range. The most undervalued players offensively according to this model were those who were predicted to a high salary group, but true salary range was much lower.

Between all three models, there were three separate lists of potentially undervalued players based off offensive performance in 2021. Analyzing all three lists, the players with the most cross-over were starting points for deeper research. However, each player’s situation had to be addressed case by case, as there are many different factors that go into a player’s previous success, current contract status, and future value.

**Validation**

For each clustering approach, a different validation technique verified the validity of such a model1. For k-means clustering, the visualization of the first two primary components verified that the clusters had at least some degree of separation from each other with no overlap (see Figure 5). This model had the most notable separation than any other combination of prediction variables. The only unincluded offensive statistic was ISO, which made no difference on the model, as the statistics it is built on are already included in the analysis. For hierarchical clustering, the primary validation was the visualization of the four clusters as a box and whisker plot, which showed noticeable differences between each cluster (see Figure 6). Therefore, the clusters created based on performance generally follow a similar trend with salary, the highest salaried players also having the best offensive performance. However, this is an awfully generalized trend and assumption, which would not be preferred to a numerical validation technique1. Finally, the vector machine models had the strongest validation technique – accuracy values. These are easy to understand and measurable numbers to choose and validate which model is the best and how successful it truly is at modeling the data.

**Evaluation**

For the hierarchical clustering model, three players I researched more into that had

1 K-means and hierarchical clustering models still can be strengthened by superior validation techniques to which I was unable to measure or unaware of. In the future I would hope to include a stronger focus on validation, which may have also allowed me to choose which model was the best of the three, instead of relying on the combined results of the three. Ishii, who focused on clustering of pitcher analysis used Root Mean Square Error to evaluate his predictive models, which included K-means and Hierarchical Clustering.

potential upside for different reasons were CJ Cron, Luis Robert, and Jed Lowrie. Cron’s contract is $7.25 for each of the next two years (as of the start of 2022). He could be a solid starting 1B option for two years to help a team bridge a gap to a prospect or 3B’s move to first. Luis Robert will have an increasing contract up to $15 million by 2025 with $20 million club options in 2026 and ’27. He is in his prime years, provides solid defense out in centerfield, and the control is in the hands of the club to retain his contract if he trends upward to outperform his contract. Lastly, Lowrie would be a one year $850k utility veteran. He still hit league average in 2021, despite being 37, has positional versatility, and could be an important veteran presence for a club. These thought processes are examples of further reasons a club may add or remove value from a player beyond the model itself.

For the SVM model the two players that were valued as being in the $15-25 million salary group but are in the lowest group ($0-1 million) were Will Smith and Ryan Mountcastle. They were both still under team control pre-arbitration, which makes sense why they still have the lowest contracts, but nonetheless teams may still want to consider them as trade targets or for early contract extensions to save on future years. This was the largest value difference produced by the SVM model. Cross-checking all three models for undervalued players, some of the most common names included Jed Lowrie, Willy Adames, Bryan Reynolds, Austin Riley, and Ryan Mountcastle.

One area left unaddressed that could still be valuable from the models are overvalued players – names to avoid or trade away depending on other teams’ evaluations of the player. A few future ways to enhance these models and create more predictive possibilities include classifying future players into this model to measure offensive performance against salary range, predicting actual salary values rather than groups, which requires regression analysis not classification, and to consider a situation without salary values. In this final scenario, one should predict based on raw metrics and then find the performers in each group that are abnormally low in traditional statistics, as this signifies their raw capabilities for power, discipline, speed, etc. are not fully seen by their results on the field. Finally, these models only include offensive statistics, however, there are many more opportunities to capture a player’s value and alter the weight of consideration applied to different parts of a player’s game. The possibilities are countless; however, it heavily relies on a creative imagination to ask questions and not being afraid of failing through trial and error.

**Conclusion**

This process is rooted in having a strong understanding of the use of machine learning models, their effectiveness, and purpose to analyze data for effective results. Over time, experience working with these models leads to a stronger methodical approach to hypothesizing, building, testing, and analyzing. As this comfortability and approach improves, more factors can be included to the analysis to truly get that competitive edge. By finding players who give teams that competitive edge, teams can ultimately reach that final goal of winning and increasing organization revenue and value, as well as building a strong team and city culture. The trickle-down effects start with each little front office decision that machine learning models can help support and enhance across the organization.

**References**

Ishii, Tatsuya. 2016. Using Machine Learning Algorithms to Identify Undervalued Baseball Players. http://cs229.stanford.edu/proj2016/poster/Ishii-UsingMachineLearningAlgorithmsToIdentifyUndervaluedBaseballPlayers-poster.pdf.

Koseler, Kaan, and Matthew Stephan. “Machine Learning Applications in Baseball: A Systematic Literature Review.” *Applied Artificial Intelligence*, vol. 31, no. 9-10, 26 Feb. 2018, pp. 745–763., https://doi.org/10.1080/08839514.2018.1442991.

Data Sources:

Baseball Savant: [Statcast Search | baseballsavant.com (mlb.com)](https://baseballsavant.mlb.com/statcast_search)

Savant Documentation: [Statcast Search CSV Documentation | baseballsavant.com (mlb.com)](https://baseballsavant.mlb.com/csv-docs)

USA Today: [Major League Baseball Salaries 2022 | USA TODAY Databases | USA TODAY NETWORK](https://databases.usatoday.com/mlb-salaries-2022/)

Chart, line chart

Description automatically generated

Figure 1: Elbow Method of K-Means Clustering

Chart, line chart, box and whisker chart

Description automatically generated

Figure 2: Gap-stat plot for hierarchical clustering, with 4 clusters chosen

Diagram

Description automatically generated with low confidence

Figure 3: Dendrogram of hierarchical clustering, cut into 4 clusters

Chart, bar chart

Description automatically generated

Figure 4: Accuracy values of all vector machine models

Chart, radar chart

Description automatically generated

Figure 5: K-Means Clustering clusters plotted by first two principal components

Chart, box and whisker chart

Description automatically generated

Figure 6: Hierarchical Clustering Box-and Wisker of four cluster by salary