MLB Archetype Clustering and Classification

Jake Giguere
giguere@bu.edu

I. Abstract

This paper presents a data-driven approach to classifying and predicting player seasonal hits in Major League Baseball using K-Means clustering and logistic regression. Key performance metrics such as batting average (BA), on-base percentage (OBP), slugging percentage (SLG), and hits (H) are used to cluster players into distinct archetypes, revealing patterns among power hitters, contact hitters, and other roles. A logistic regression model is then applied to predict whether a player will achieve 100 or more hits in a season(180 games). The model's performance, evaluated on a dataset of over 12,000 data points, demonstrates strong predictive capabilities, with a precision of 0.965, recall of 0.965, F1-score of 0.945, and an area under the ROC curve (AUC) of 0.98.

This approach significantly outperforms a prior Parlay-Leg Optimizing Neural Network(PLONN) multi-layer perceptron (MLP) model that was trained on roughly 600 data points, underscoring the importance of data volume in achieving robust model performance. While the PLONN showed promising results with limited data, the larger dataset in this study allows for more accurate player archetyping and prediction of outcomes. Although this paper does not use MLPs the main takeaway is leveraging pySpark to achieve higher precision when larger datasets are utilized.

II. Introduction

In terms of sports analytics, the ability to accurately predict player performance has become an asset for teams, their analysts, and coaches. Major League Baseball (MLB) generates large amounts of data, which can be leveraged to discover meaningful patterns and improve decision-making processes.

This paper explores two machine learning techniques—K-Means clustering and logistic regression—to classify baseball players into distinct archetypes and predict their performance over a season. Specifically, the study aims to predict whether a player will achieve 100 or more hits in a season using a relatively large dataset of over 10,000+ player instances. Key performance metrics, including batting average (BA), on-base percentage (OBP), and slugging percentage (SLG), serve as inputs to the model, allowing for the creation of player archetypes such as power hitters and contact hitters.

A critical aspect of this work involves contrasting the results of this larger dataset approach with a prior effort that used a much smaller dataset of 600 data points in a Parlay-Leg Optimizing Neural Network (PLONN) multi-layer perceptron (MLP). While the PLONN MLP showed promise in predicting player daily outcomes, the limited data significantly constrained its accuracy. This study demonstrates that increasing the volume of data, along with leveraging clustering and logistic regression, leads to more accurate and reliable results.

By employing K-Means clustering to group players based on their statistical attributes and logistic regression to classify whether they will reach certain performance milestones. The model's effectiveness is evaluated using precision, recall, F1-score, and Receiver Operating Characteristic (ROC) curve analysis, with results showing a high area under the curve (AUC) of 0.98, indicating strong predictive capabilities. This work underscores the importance of larger datasets and machine learning techniques in sports analytics, highlighting their potential to generate actionable insights and enhance decision-making in MLB.

III. KMeans

To better understand player performance and identify distinct player archetypes, we applied the KMeans clustering algorithm to the dataset. KMeans is a widely used unsupervised learning technique that partitions data into a fixed number of clusters by minimizing the within-cluster variance. In this study, we used key performance metrics such as on-base percentage (OBP) and slugging percentage (SLG) to group players into similar archetypes.

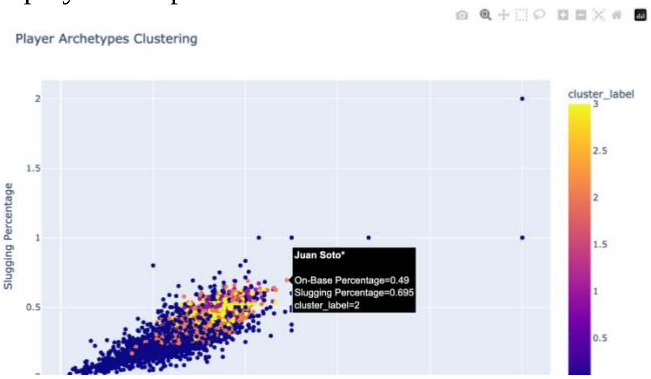
Methodology

Using KMeans, we divided players into four distinct clusters based on their offensive statistics. The algorithm iteratively assigns each player to the cluster whose centroid (mean point of the cluster) is closest to the player's performance data. Once all players have been assigned, the centroids are recalculated, and the process repeats until the clusters stabilize, minimizing the distance between each player and their cluster’s centroid.

Results and Interpretation

The clustering revealed clear groupings of players based on their performance metrics. Players with high slugging percentages and high on-base percentages were grouped together, while those with lower statistics formed separate clusters. The resulting clusters provide insight into player archetypes, such as power hitters and contact hitters, Non-performers, On-base Specialists, allowing us to better understand each player’s contribution to their team.

The attached figure (Fig. 1) visualizes the results of the KMeans clustering, where players are plotted based on their on-base percentage and slugging percentage, and the color represents the cluster they belong to. Notably, high-performing players such as **Juan Soto** appear in clusters characterized by higher OBP and SLG, which falls within contact hitters. Each cluster represents a group of players with similar statistical profiles, making it easier to identify patterns in performance.



Juan Soto*

On-Base Percentage=0.49
Slugging Percentage=0.695
cluster_label=2

Figure 1

IV. Logistic Regression

After clustering the players into distinct archetypes using KMeans, we proceeded with a supervised learning approach to predict whether a player would achieve a significant performance milestone: accumulating more than 100 hits in a season. For this task, we employed logistic regression, a classification algorithm commonly used to model binary outcomes.

To facilitate player performance classification, we created a binary target variable, **hit**, which classifies players based on whether they accumulated more than 100 hits in a season. Players with more than 100 hits are labeled as "1" (successful), while those with fewer hits are labeled as "0" (unsuccessful). This classification serves as the target variable for our logistic regression model, which predicts a player’s likelihood of reaching the 100-hit milestone.

Features

For the logistic regression model, we selected several key performance metrics as independent variables (features). These included:

- **Batting Average (BA)**: Hits per At Bat.
- **On-Base Percentage (OBP)**: Reflects a player’s ability to get on base.
- **Slugging Percentage (SLG)**: Indicates a player's power-hitting capability.
- **Home Runs (HR)**: A metric for power hitting.
- **Strikeouts (SO)**: Indicates how often a player strikes out.
- **Runs Batted In (RBI)**: Measures a player’s ability to drive in runs.

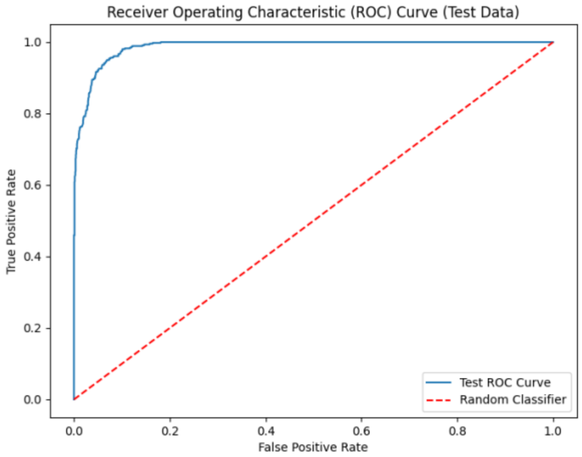
These features were combined into a feature vector using PySpark’s Vector Assembler to prepare the data for logistic regression.

We applied logistic regression to predict the likelihood that a player would achieve 100 hits, using the features mentioned. Logistic regression is well-suited to this problem because it outputs probabilities, making it an effective tool for binary classification tasks. The model predicts the probability that the target variable (hit) will be 1 (the player will reach 100 hits).

The results of the logistic regression model were as follows:

- **Precision**: 0.965
- **Recall**: 0.965
- **F1-Score**: 0.945
- **AUC**: 0.98

-



The ROC curve for the test data shows that the logistic regression model is highly effective at classifying whether players will exceed 100 hits. The high AUC value and the curve’s proximity to the top-left corner indicate that the model achieves a strong balance between True Positive Rate and False Positive Rate, making it a reliable classifier for this task.

V. Conclusion

In this study, we explored the application of machine learning techniques—KMeans clustering and logistic regression—to classify and predict player performance in Major League Baseball. By leveraging a dataset of over 10,000 player instances over 10 seasons, we identified distinct player archetypes and successfully predicted whether a player would accumulate more than 100 hits in a season.

The use of KMeans clustering allowed us to group players into archetypes based on metrics such as batting average (BA), on-base percentage (OBP), and slugging percentage (SLG). This unsupervised learning technique revealed patterns in player performance, highlighting distinct groups such as power hitters and contact hitters. These clusters provided valuable insights into player roles and their offensive contributions to the game.

Building on the archetypes, we applied logistic regression to classify players based on their likelihood of exceeding 100 hits in a season. The model achieved impressive results, with a **precision of 0.965**, **recall of 0.965**, **F1-score of 0.945**, and an **AUC of 0.98**. These metrics indicate that the logistic regression model is highly accurate in predicting player performance and demonstrates a strong ability to generalize to unseen data, as evidenced by the ROC curve analysis on the test set.

This study also highlighted the impact of dataset size on model performance. When compared to the PLONN model, which utilized only 600 data points, the logistic regression model benefited significantly from the larger dataset, capturing more intricate patterns in player performance.

References

Baseball Reference,
https://www.baseball-reference.com/leagues/majors/2020-standard-batting.shtml