Predicting the Role of Softball Players Based on Objective Performance Metrics

Rachel Shurberg, Data Science Capstone Project '23 with College Sports Evaluation

Background and Research Question

Sports analytics are used to analyze all aspects of sport which can help teams and individuals evaluate their performance. With the increase in technology, there has been advances in the ability to analyze player performance to gain insight into the weaknesses and strengths of players. For softball specifically, many statistics regarding speed, strength, hitting, and pitching are considered.

This capstone project looks to answer the question: Can performance statistics be used to classify the role of a college softball player on their team using machine learning techniques? This project analyzes speed, hitting, and throwing data collected by College Sports Evaluation.

Data

Description

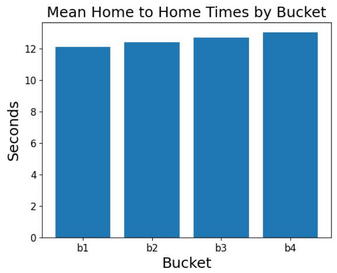
The data was collected by College Sports Evaluation. The dataset includes 13 statistics on player performance, the role of the player, and what division they play. The 13 statistics include the time it takes the player to run from home plate back to home plate, the exit velocity of the ball when thrown, the bat speed when hitting, and more. The hitting statistics were collected using a Blast sensor attached to the bat. There are three types of players: Impact player, starter, and role player. The player type is subjective and determined by the coach of each team while the performance statistics are objective measurements. For all buckets, the least common player type is impact player. The most common type of player is starter.

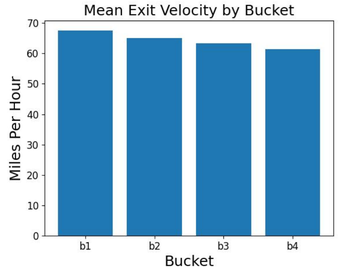
The data is split into 4 buckets that are subsets based on the division since the level of play greatly varies.

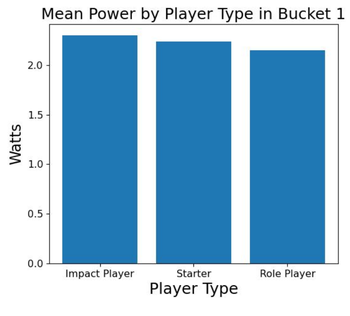
Bucket	Divisions	Sample Size
1	Power 5, High D1, Top JUCO 1	591
2	Mid & Low D1, High & Mid JUCO 1, High D2	823
3	Mid & Low D2, Mid & Low JUCO 1, Top NAIA, High D3	1040
4	Bottom NAIA, Mid & Low D3, Bottom JUCO 1, JUCO 2, JUCO 3	794

Data Cleaning

There was some missing data to handle. The non-hitting data had multiple trials, but most players only completed one trial for each statistic. For those four statistics, the average of the trials they completed was averaged. Players with missing data from the five non hitting statistics were dropped. Missing data from the Blast hitting statistics was imputed using the mean of that statistic within that player's bucket. The mean was used since the distribution of the variables were normal.







Models

Four multiclass classifiers were ran to classify the player role from the 13 objective statistics. They were ran using sklearn packages in python. The four classifiers are **K-nearest neighbors, random forest classifier, decision tree classifier, and gaussian naive bayes classifier**. Each classifier was ran on each bucket's dataset separately.

The K-nearest neighbors model was fitted with the hyperparameters of a k value of 13 and the euclidean distance metric. The random forest classifier was fit with the 600 as the number of estimators. The decision tree classifier and gaussian naive bayes classifier was fit with their default parameters.

Results

The classifiers overall did not show successful classification of player type from objective statistics. The results ranged in accuracy from 34.5% to 47.3% across all buckets and all classifiers. The classifier with the highest accuracy on average is the random forest classifier. Buckets two and three had the highest accuracy across all classifiers. This makes sense since buckets two and three have the largest sample size, so there is more data for training.

Accuracy of Each Classifier by Bucket

Bucket	K-Nearest Neighbors	Random Forest	Decision Trees	Gaussian Naive Bayes
1	.462	.445	.345	.361
2	.424	.467	.358	.473
3	.418	.471	.389	.438
4	.403	.447	.396	.390
Average	.427	.458	.372	.415

The accuracy of classification for each player type within each bucket for each classifier was also calculated. The accuracy of impact players was consistently by far the lowest. The accuracy of each type of player is more reflective of the proportion of each type of player in each bucket. The starter and role player accurate classification rate is higher, and usually over 50%.

Accuracy by Position for KNN and Random Forest

Bucket	K-Nearest Neighbors			Random Forest		
	Impact	Starter	Role	Impact	Starter	Role
1	.130	.511	.571	.130	.532	.510
2	.038	.486	.507	.115	.457	.609
3	.073	.578	.429	.171	.530	.500
4	.139	.540	.412	.222	.587	.433

Conclusion

These results indicate that there is good reason to believe that there are other factors that contribute to the effectiveness of a player. This analysis doesn't conclude that there is no relation, but only 13 statistics were included in this analysis.

Limitations and Future Work

There are many more components that were not included that could contribute to player success. The sample sizes of each bucket were not very large, so the accuracy could improve with larger sample sizes and buckets that don't combine as many levels of play. An ideal future study would have larger sample sizes and more data that assess a wider range of abilities. The player type metric is subjective, and making predictions based on a subjective measurement is not ideal. Another possible related project would be trying to categorize a player into the correct division.

CSE COLLEGE SPORTS EVALUATION

WELLESLEY