Executive Summary: Predicting MLB All-Star Selections Using Machine Learning

This project applied a multitude of machine learning techniques to a structured Major League Baseball (MLB) dataset in order to predict whether an MLB player would be selected as an All-Star in the following season. Our dataset consisted of season statistics for hitters across 10 full MLB seasons (excluding 2019 and 2020 due to COVID). The binary classification target was "AS_next_year" with levels "Yes" or "No" indicating whether a player became an All-Star in the following season. The data included 4,514 rows and 31 features, including traditional stats (e.g., AVG, HR, PA), advanced metrics (OPS+), team and position data, and categorical variables for current All-Star status. Pitchers and players with low sample sizes (PA < 100) were excluded.

Because All-Star selections are rare events (~9% prevalence) in this dataset, the class imbalance presented a recurring challenge across models. Extensive data cleaning was applied: categorical variables were converted to factors, and redundant or derived stats were removed (OPS, TB, H, AB, G, and Player_ID ). Our dataset contained variables with very different ranges, such as OBP (around 0.2-0.5), and PA (100-800), so we normalized each variable using min-max normalization to address this issue.

Our primary evaluation metric was the F1 score, which balances both precision and sensitivity, a necessity given the class imbalance in our dataset. Accuracy was largely ignored, as it would reward models that simply predicted "No" in most cases. Sensitivity, precision, and Kappa were also recorded. All models were evaluated using 10-fold cross-validation, with threshold tuning applied to each probability-based model to further optimize F1 scores.

The following algorithms were trained, tuned, and evaluated:

K-NN, Logistic Regression, Naive Bayes, Decision Trees, Rules (JRip), Support Vector Machine, Bagging, Boosting, Random Forest, XGBoost, Neural Network

The process for each model was vaguely the same. We ran each model through caret, allowing for us to set up cross validation, as well as a tuning grid. We started each model by testing it with 10-fold CV, and no tuning grid, to get some baseline statistics. We then created a tuning grid with the relevant hyperparameters for each model, and re-ran each model however many times we needed to identify the combination that optimized the model's performance. After we found the optimal tuning grid, we set up a threshold test (for probability based models) to find the probability threshold that maximized our F1 score (our primary evaluation metric).

After doing this for each model, we were able to rank all our models based on the F1 score.

Model Rankings:

1. XGBoost – 0.4844

2. Random Forest – 0.4832
3. Logistic Regression – 0.4681
4. Decision Tree – 0.4294
5. Boosting – 0.4593
6. Neural Networks – 0.411
7. Bagging – 0.4103
8. Support Vector Machine – 0.3852
9. Naive Bayes – 0.3733
10. K-Nearest Neighbors (KNN) – 0.3487
11. Rules (JRip) – 0.336

While XGBoost and Random Forest delivered the top F1 scores, Logistic Regression somewhat surprisingly offered interpretability and consistent sensitivity. Our rule-based model, although easily interpretable, proved to be our least accurate model, and KNN continued to struggle with class imbalance and high-dimensionality.

Based on our F1 scores, we likely would not be ready to fully apply any of our models to a real life application setting, as our highest performing model didn't even break 0.5. Although determining what is a "good" F1 score can be virtually impossible considering it is all relative to the data being used, it is safe to assume that our high end F1 scores around 0.48 are not high enough to confidently use outside this project.

There are a few things we know could have been done better, and would likely result in better results from our models. Using WAR in our dataset is something we certainly should have done. WAR is generally recognized as the best stat for measuring a player's true value. WAR, short for Wins Above Replacement, is an impossibly complicated formula that takes factors like basic stats, to expected stats (based on different ballparks, altitude, and more), and even weather to create a simple number output. Using this would certainly increase our models' performances.

The major outstanding issue with using a dataset revolving around all-star status is the process that is used for selecting all stars. The all-star selection process is long and complicated, and each year provokes angry fans that don't see their favorite players selected for the game. A combination of fan votes, manager votes, and player votes are used to pick the final all-star rosters. A common narrative surrounding the all-star game is that players from popular teams like the Yankees, who have 4 million followers on Instagram, will naturally get more fan votes than players from teams like the Rockies, who only have 550K Instagram followers.

Overall, despite not getting great results in the grand scheme of things, we were able to use our knowledge of machine learning to maximize our results. We saw significant improvement from the basic version, to the optimized and fine tuned version of almost all of our models. A future project using the same dataset would have a great baseline, with plenty of ideas for even better results.

This semester was an eye opening experience for me in understanding what machine learning actually is, and the applications that it is actually used in. Before taking this class, I associated machine learning with highly technical, impossible algorithms, creating machines like ChatGPT, that were completely inaccessible to me. I didn't realize how versatile, genuinely usable, and practical using these machine learning techniques really were. More than anything, I learned that success in machine learning isn't just about choosing the right model, but it's about noticing and fixing all of the small things that silently trip up the models. Things like formatting variables, normalizing data, tuning steps, setting up grids, and binning data are all so important, and can easily be ignored, are what makes these models real learners.

Throughout my Augsburg education, one consistent theme I've noticed is that data isn't neutral. In many classes, including this one, I've learned how data can inherently incorporate social injustices, often unintentionally. This goes from biased algorithms, or incomplete data, or information that reinforces racism, sexism, and homophobia. I've come to understand that one of, if not the most important quality of a data scientist is to notice these patterns, and actively try to disrupt and change them. Instead of blindly taking data, and using it, it is just as important to inspect the data for these injustices, and make changes so that data being used represents everybody. Augsburg has done a great job of teaching me how data can be used for good, and teaching me what it looks like when it's not.

Outside of the classroom, being apart of the Augsburg baseball team has helped me grow as a team member, especially in a group project like this one. Collaboration, support, and stepping into my role are all important things that I have developed in my baseball career, and I look forward to continuing to apply these attributes in the future.

Looking at Augsburg's mission, to foster informed citizens, thoughtful stewards, critical thinkers, and responsible leaders, I most closely identify with the idea of being an informed citizen. I grew up in Salt Lake City, and to put it frankly, in a privileged neighborhood, and sheltered environment. Growing up, I took pride in understanding the injustice that shaped the world, and had parents that helped me understand that what I had was not what everyone had, however that alone cannot even touch what I've learned and experienced here at Augsburg. Augsburg broadened my view of the world, especially through the lens of data. Taking both Critical Race Theory and Intro to Data Science (taught alongside data injustice studies), really transformed the way I saw the world. Learning both history, and the impacts it has had on our current world's data was very powerful, and I have Augsburg to thank for that.

In terms of vocation, I would love to contribute to a company that reflects strong values and genuinely works to create a better world, by working in their data science department. I now see that machine learning is not just a technical tool used by tech powerhouses, but a tool that people like me can use to create genuine impact on my local community.