NCAA Softball Data Analysis Final Report

1. Introduction

I started playing softball when I was 5 years old and grew up on the dirt. When I was in high school, I played for a top-ten nationally ranked travel softball team and used to be on leadership for the Notre Dame club softball team through my junior year. College softball has grown so much from when I was a kid and would watch the postseason and Women College World Series on TV. College softball had 339.6% revenue growth in the last year and is currently the fastest growing NCAA sport (*The Athletic*). Despite its increasing popularity, the development of data analytics in softball is much further behind than its counterpart in baseball. As opposed to baseball where the MLB is a well-established and highly visible league, the biggest stage for softball is college softball. Yet, the difference between the data available and the level of analysis performed on it at the highest levels of the respective sports are astounding. There is more analysis performed on NCAA baseball statistics than NCAA softball even though it is not as popular.

My goal is to identify the trends in the NCAA softball offensive data from the 2023 season and utilize my findings to make suggestions for what areas the Notre Dame softball team should focus on improving going into the 2024 season. I examined batting trends in the NCAA softball data from the 2023 season to identify what factors most strongly contributed to the success of the top individual players and teams. Then, I determined which metrics are the most closely related to generating runs and which ones are the strongest predictors of player performance. I applied these metrics to analyze the Notre Dame offense and to make actionable suggestions for areas of improvement to work on in the offseason. Notre Dame had an average season last year, finishing 30-19-1 at 7th place in the ACC. Notre Dame had solid offensive statistics, with a team batting average of .315, but it was not enough to differentiate them from their opponents and to provide enough runs for them to win an increased number of games. The goal of my analysis is to provide insights into what will help push them over the edge to be on par with the elite offenses in NCAA softball.

2. Related Work

There was not any previous analysis that had been performed on NCAA softball data over the course of a season that I could find. The data I obtained was from a publicly available R package, so other people may have utilized it to perform analysis without publishing a written report about their findings. The NCAA softball website offers individual and team statistics aggregated over the course of the season that a user can go on and search for. A few other third party sites, such as D1Softball.com also offer similar aggregated statistics, but beyond very basic information the rest of the data is behind a paywall. While this raw data is available, neither NCAA softball nor the third party sites have performed analysis on it in any form that is publicly available. Many individual teams have their batting and pitching statistics over the course of a season available on their websites as well, but these are just on a pdf document and not in a csv data table format that can be downloaded and analyzed with methods that we have learned about in class. However, these datasets are mainly used as a reference tool if a person is trying to look up an individual players' statistics, the statistics for a team they enjoy, or who the leaders are in certain statistical categories. From my research I was not able to find any work that performed anywhere near the same level of data analysis on softball data as we perform in class on the data from other sports. My goal was to perform a comprehensive analysis of these NCAA statistics because very little has been done before that applies the sabermetrics used in baseball with the complexity of data analysis that we learn in our class on NCAA softball data so I feel like there is a lot of room to find unique insights in the trends of the data as a whole.

3. Data Description

The dataset I am using comes from the softballR R package that I uploaded from GitHub. I am analyzing a function that provides the batting box scores for the 2023 season where each row in the dataset provides the statistics for an athlete in one game. The dataset contains 209,242 samples (rows) and 32 columns that contain explanatory variables of the offensive statistics of each player by game including number of RBI's, number of each type of hit (singles, doubles, triples, and homeruns), total bases, number of walks, number of strikeouts, number of stolen bases, and number of times the batter hit into a double or triple play. One limitation is that the datasets provided by this R package only provide basic statistics and do not have as specific and in-depth of data as the statcast data we used in our MLB analysis. Therefore, my analysis will focus more on overall season trends rather than specific matchups, pitch type/location breakdowns, and individual player strengths and weaknesses. I created calculated fields for my batting dataset so that I would be able to work with more advanced metrics in my analysis. First of all, I calculated the number of singles and the number of plate appearances and added them into the dataset as new columns. Then, I created calculated fields of metrics that measure player quality using the data available such as batting average, on base percentage, slugging percentage, OPS, isolated power, runs created, weighted on base average, strikeout rate, and walk rate. The data is broken down by game, so I created a loop that calculated and aggregated the totals for each of these advanced statistics, in addition to the basic ones given in the original dataset, for the entirety of the season both by player and by team. Next, I filtered the data to only consider batters who had greater than or equal to 30 plate appearances over the course of the season. This focused my analysis on players who were consistently in the starting batting lineup and prevented the data from being skewed by players who only had a few at-bats throughout the course of the season.

4. Methods

After performing my initial exploratory analysis, my goal was to examine what metrics are most closely related to winning. Since my dataset did not include data on wins and losses, I used the number of runs scored as the response variable because scoring runs is the ultimate goal of an offense and therefore a good metric to measure its success. The two key predictive models that I utilized in my analysis are linear regression and XGBoost. Prior to creating both of these models, I created a training dataset using a random sample of 60% of the team-level dataset. The validation sets (40% of each set) were created using the rows that were not included in the training sets. I chose to run these models on team-level data due to the fact that there were so many samples in the player-level data.

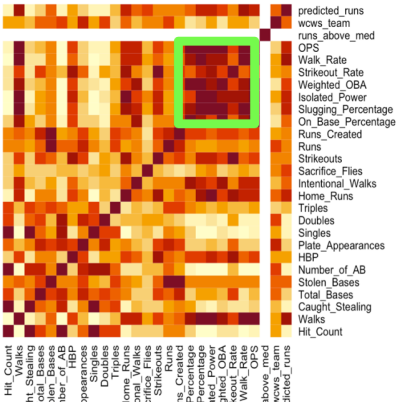
I then proceeded to run a linear regression model. The reduction in the number of samples at the team level decreased the variance and the number of outliers. When running summary statistics on my response variable, the number of runs scored, the mean was less than the median and they were very close in value which indicates that there are likely not any significant outliers in this dataset. Next, I created a density plot of runs which revealed a distribution that was pretty close to normal, so there was no need to perform a log transformation for the linear regression. I ran both a backward elimination method and a forward selection algorithm to select the variables that were most important in predicting the number of runs. The linear regression was effective for inference because it provided a straightforward interpretation of the coefficients and the statistical significance of each variable. However, there were limitations because the relationship between the predictors and the response variable was non-linear. A weakness of linear regression models is that they often perform poorly when attempting to model non-linear relationships. Also, there were a lot of predictors that were selected for the model, even after running variable selection, which is a type of problem linear regression does not tend to handle as well.

Finally, I used XGBoost to create a model that predicts the number of runs that each team would score, and applied the model to individual players to determine their offensive value to the team. Before running the model, I converted the training and test data into a matrix format. Then, I trained the XGBoost model and ran parameter tuning. At each step, I interpreted the output graph to choose the combination of parameters with the lowest RMSE. The optimal values were a minimum child weight of 1, a max depth of 3, a gamma of 0.10, a subsample of 1, a column sample by tree equal to 1, a learning rate of 1, and 250 trees. One strength of the XGBoost model is that it has stronger

predictive power than the linear regression model, which allows it to identify characteristics of more difficult to classify samples. It is also more resilient to overfitting, which was a problem in my linear regression model despite partitioning the data. For these reasons, it is a stronger model choice than linear regression to create a more accurate predictive measure for the number of runs scored.

5. Discussion

First of all, I filtered out the 10 players with the highest number of runs created and ran a correlation analysis to see what factors that they all shared in common. This revealed that there was high correlation between the metrics OPS (on-base percentage + slugging percentage), walk rate, isolated power, slugging percentage, and on-base percentage. The high correlation with on-base percentage and walks demonstrated that it is important to consistently get on base in order to manufacture more runs. The high correlation between power statistics such as isolated power and slugging percentage is logical because extra base hits are likely to score runners that are already on base when they are batting and put them in scoring position if it was not a home run where they scored themselves on the hit. My main takeaway from this chart was that a combination between consistently getting on-base and power to drive in more base runners is key for the players who were the most effective in driving in runs in the 2023 season.

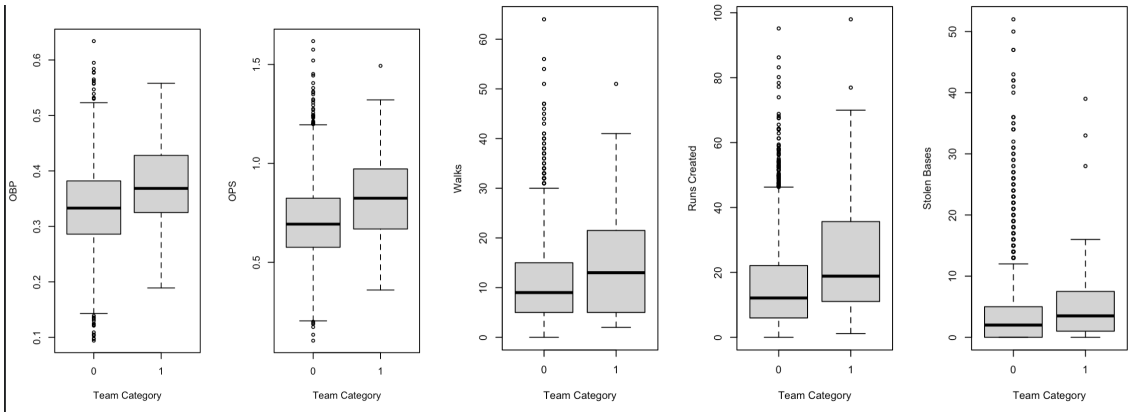


Next, I progressed into creating a linear regression model with the goal of identifying predictors that were statistically significant in predicting the number of runs. I ran this regression on the team-level statistics because there were so many samples in the total player data. Since there were 31 explanatory variables in the dataset, I utilized backward elimination and forward selection to select the variables that were the most significant in making the prediction. There was not a large difference in the accuracy between the backward and forward methods. However, the backward elimination method yielded less variables than the forward selection algorithm, and most of the variables in the backward model overlapped with the ones yielded in the forward selection. Therefore, I chose to use the variables from the backward elimination model in my linear regression model: walks, total bases, stolen bases, intentional walks, runs created, OBP, walk rate, and batting average. When I ran the regression, it identified walks, stolen bases, runs created, batting average, and OBP as statistically significant predictors. One interesting insight is that intentional walks were the variable with the largest positive correlation with runs scored. This makes sense because if players are strong enough offensively that they are being intentionally walked they are probably responsible for a significant portion of their respective team's scoring. Across all of the batters in the NCAA and not just the top players that I examined in my correlation chart, the linear regression model identified that it is more important to consistently get on base rather than rely on power to generate runs. This makes sense because the top players likely have more consistent power, and it requires a rare skill level to generate these results that the majority of players in the NCAA do not have. While the Notre Dame softball team can focus on increasing their power, a more effective area to devote their efforts to will be in finding effective strategies for each of their players to get on base and advance bases. However, one drawback to this model was that the adjusted R squared and multiple R squared were both 98% even after running variable selection. This indicates that the model was likely overfit, due to

the fact that even after using backward elimination there were still a lot of predictors in the model, which caused the regression to fit the data too well so it captured the general trend and a lot of random noise. As a result, I used the output from this model to identify the statistically significant variables but used the XGBoost to make the actual predictions for the number of runs scored.

Team	OPS (on base + slugging)	Walks	Stolen bases	Runs Created	Isolated Power
Average of all Division 1 Teams	.72	129	50	169	.12
Notre Dame	.89	144	45	260	.18
Florida State (ACC Champions)	.89	249	134	354	.19
Oklahoma (National Champions)	1.12	242	53	491	.3

The linear regression and correlation chart indicated that metrics related to getting on base, advancing runners, and power were important so I wanted to delve deeper into these insights. I compared Notre Dame's statistics in walks, stolen bases, runs created, OPS, and isolated power to the average of all Division I programs, the ACC Champions and World-Series runner ups Florida State, and the National Champions Oklahoma. While Notre Dame softball is on par with or better than the national average for all of these statistics, they fell significantly behind both Oklahoma and Florida State in runs created and walks. Florida State has higher metrics in walks and stolen bases which shows that they are generating their runs through getting lots of runners on base and finding ways to advance them while Oklahoma dominates the power statistics which shows that power is key in generating their offensive production. In order to become an elite level offense and boost their runs created, Notre Dame must find a way to either increase the consistency of their power or increase the number of walks and stolen bases as a way to advance runners and therefore generate more runs. Delving deeper into World Series level offenses, I created boxplots for the same statistics and filtered them by teams who had qualified for the World Series in the 2023 season (1 on graph below) and those who did not (0 on graph below). This analysis was performed on the player level statistics. Unsurprisingly, teams that qualified for the World Series performed higher on all of these metrics; however, the World Series teams had almost no outliers as opposed to the rest of the NCAA where there were many outliers on both ends. This reveals a big differentiator between elite level offenses and the rest of the NCAA: consistent performance of all of the players in the batting lineup. Teams that did not qualify for the World Series had one or a couple superstars who they relied on but were not consistent through their lineup. On the other hand, players on the teams that made the World Series performed slightly better on average which shows they were stronger offensively but the minimal outliers shows that all of the players on these teams performed well throughout the season.



I created an XGBoost model to predict the number of runs that a team would score and then applied the model to individual players from the Notre Dame softball team to determine their value. For my output, I got an MAE of 18

and a RMSE of 23. For the Notre Dame softball team, the actual number of runs they scored was 294 while the number of runs that the XGBoost model predicted was 270, so the differential was 24 runs. The differential was between 25-30 runs for each of the ones that I predicted, which when analyzing season total number of runs for teams that fell in the 300s range is a reasonable error in prediction that demonstrates the model is imperfect but also eliminates concern of overfitting. In line with previous analysis, the XGBoost model identified metrics related to getting on base (sacrifice flies, OBP, walks, stolen bases) and power (home runs, doubles, and slugging percentage) as the most important variables in predicting the number of runs. Sacrifice flies is a variable that the linear regression did not identify but is a key way to score runs if there is a runner on third tagging up or to advance runners in order to ultimately generate scoring. This reiterates the importance of finding ways to get on base and move runners in combination with consistent power as the key ways to boost the number of predicted runs.

Team	Runs	Predicted Runs by XGBoost model	Differential
Virginia Tech	350	321	29
Oregon	318	286	32
Notre Dame	294	270	24

I then applied the predicted runs metric on a player level and more specifically to the Notre Dame softball team. Lexi Orozco (34.4), Karina Gaskins (36.3), and Joley Mitchell (33.0) were the 3 players on the Notre Dame softball team with the highest number of predicted runs. This prediction makes sense because they were clearly the top 3 offensive performers for Notre Dame according to their season statistics. These 3 players had the highest batting averages, on-base percentages, number of walks, home runs, and sacrifice flies for the Notre Dame softball team, which were all predictors that the XGBoost model identified as significant in generating runs. One observation is that these players were all girls who hit for power but also consistently found a way to get on base. Unfortunately, Orozco graduated and Mitchell transferred for this season, leaving only one of our top offensive performers in the lineup. There was a big drop off in predicted runs between these players and the rest of the team, as demonstrated by the average predicted number of runs for the Notre Dame softball team being 18 (only including the 13 players on the roster with more than 30 at-bats). The key takeaway from this model has reiterated the importance of finding ways to boost statistics to get on base more often and increase power. It will be important for the Notre Dame softball team to find ways to more effectively get on base and move their runners, which may come in the form of increasing sacrifice plays, finding ways to steal extra bases, or being more selective at the plate to increase the number of walks for players who do not have as much power.

6. Conclusions and Future Work

In order to elevate their offense to an elite level, Notre Dame must find a way to get on base more often and advance runners through increasing consistency in their extra base hits, sacrifice plays, or stealing extra bases. The most important factor in creating runs is consistently getting on base, and what differentiates the upper echelon of players is the fact that they combine this with power. My models both determined that walks, stolen bases, runs created, batting average, and OBP are the most critical variables in predicting the number of runs scored.

One downfall of this dataset is that I was limited in the analysis that I could perform because I only had access to basic statistics. I used the number of runs scored as the response variable because I figured that this was a good measure of offensive productivity and success. However, if I had a dataset where the outcome of each game was given as a 0 or 1 variable, I would have been able to more directly predict which offensive statistics led to winning which would have been interesting. If given more time, I would love to delve deeper into the pitching statistics and perform the opposite of this analysis and predict the number of runs given up by a pitcher in a season and what variables have the most significant impact in determining that value.

7. Contributions

I worked alone so therefore I did this whole project.

8. Bibliography

Auerbach, Nicole, and Richard Deitsch. “WCWS Championship Series Averages 1.6 Million Viewers.” *The Athletic*, The Athletic, 9 June 2023, theathletic.com/4598276/2023/06/09/womens-college-world-series-2023-viewership/.

“Softball Overall Statistics.” *Notre Dame Fighting Irish - Official Athletics Website*, 20 May 2023, fightingirish.com/sports/softball/.