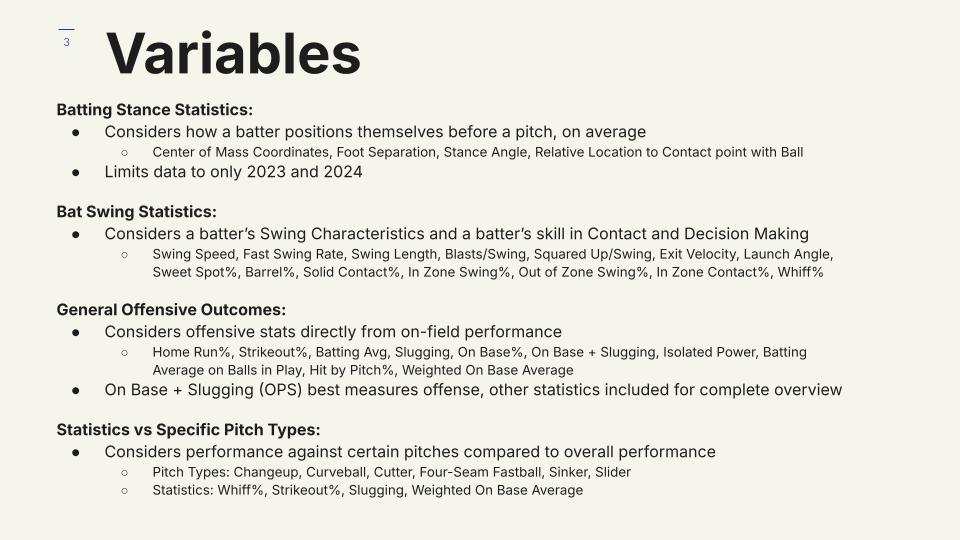


Major League Baseball has some of the best data in all of professional sports. This is helped massively due to the isolated, discrete plays that occur between individual batters and pitchers, but can also be attributed to the MLB’s consistent investment into the field of Data Analytics. New sensors, tracking equipment, and other novel data-acquisition techniques are being introduced into the league all the time. Last year, MLB Statcast and BaseballSavant released to the public Bat Tracking Data, an example of it shown in the image in the top right, giving brand new, quantifiable insights into how the best players swing their bat. And just last month, they also released Batting Stance data, showing where and how the league’s batters choose to stand in the batters box.

My project seeks to look at this new set of data, and find an analytical lens to view baseball offense through. In this analysis, I look to answer 3 main questions.

1. Does Batting Stance, an unoptimized factor in hitting left mostly to player preference, actually meaningfully affect hitting performance?
2. Can we use a batter’s base characteristics, including Stance, Swing, and Smarts (or Decision Making), to predict offensive outcomes, such as bases, strikeouts, or runs?
3. Can we identify which hitting factors are most impactful towards increasing offensive value, giving new insights into what hitters should train around?

Through these questions, I’ll obtain a holistic view of how statistics can best be used within the framing of Baseball offense.



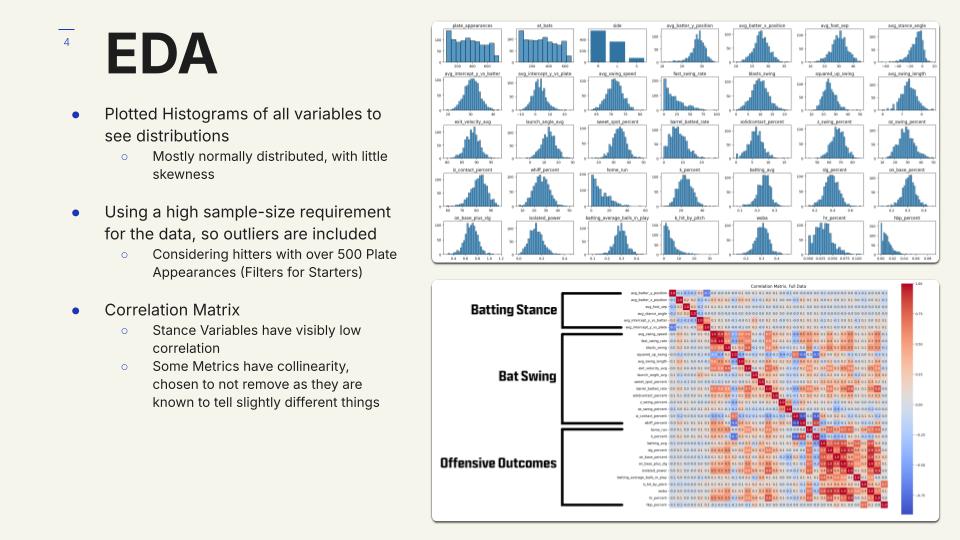
Because of Baseball’s aforementioned investment into data, a large number of variables are included in this analysis. I’ve separated them into groups of 4 for easier discussion.

The first group is Batting Stance Statistics. These predictors consider how a batter positions themselves before a pitch, on average through a season. Specifically, this covers the x and y coordinates of the batter’s center of mass, foot separation, stance angle, and the location where the batter on average makes contact with the ball, relative to the plate and to the center of mass. Unfortunately, due to the novelty of this tracking, the data is only available for the 2023 and 2024 seasons, limiting the amount of usable data for this project.

Next is Bat Swing Statistics. These consider a batter’s Swing Characteristics, such as Swing Speed and Swing Length. It also includes a batter’s skill in Contact (measured with Whiff%, rate at which the bat is swung and completely misses the ball, among others) and Decision Making (In Zone Swing%, the rate at which a strike is swung at, among others).

We also consider general offensive outcomes as a target, acknowledging the stats that come directly from performance and production on the field. On Base Plus Slugging, or OPS, is considered a gold standard in measuring offense, but other stats are included for a wider picture.

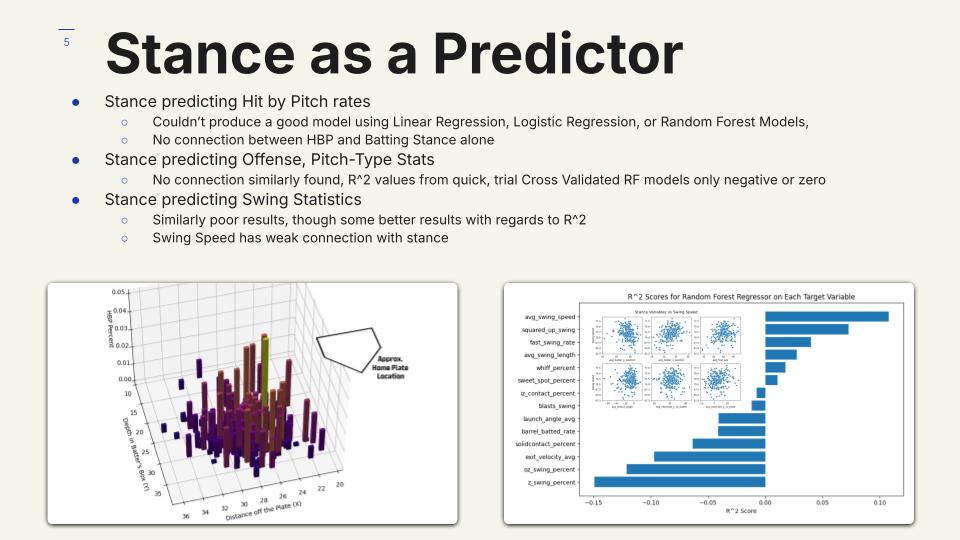
Finally, we also look at batting performance facing specific pitch types, for the purpose of seeing whether we can predict if certain batters will perform better vs certain pitches based on their batting characteristics. There’s limitations is stat and pitch-type availability here, but the variables chosen here cover a good deal.



In Exploratory Data Analysis, I plotted histograms of all the variables in use, to see how they are distributed. The results of this are shown in the top right. Most of the variables, with few exceptions, are normally distributed, making later analysis a bit easier.

For this project, only hitters with over 500 plate appearances in a season were considered. This number is fairly standard among baseball stats and record keeping, known as making a player “qualified”, and roughly filters for only an MLB team’s starting batters. Because of this high sample-size requirement for all the data points, all outliers would be included, as over this large number of plate appearances, extreme stats wouldn’t be flukes or misrepresentation.

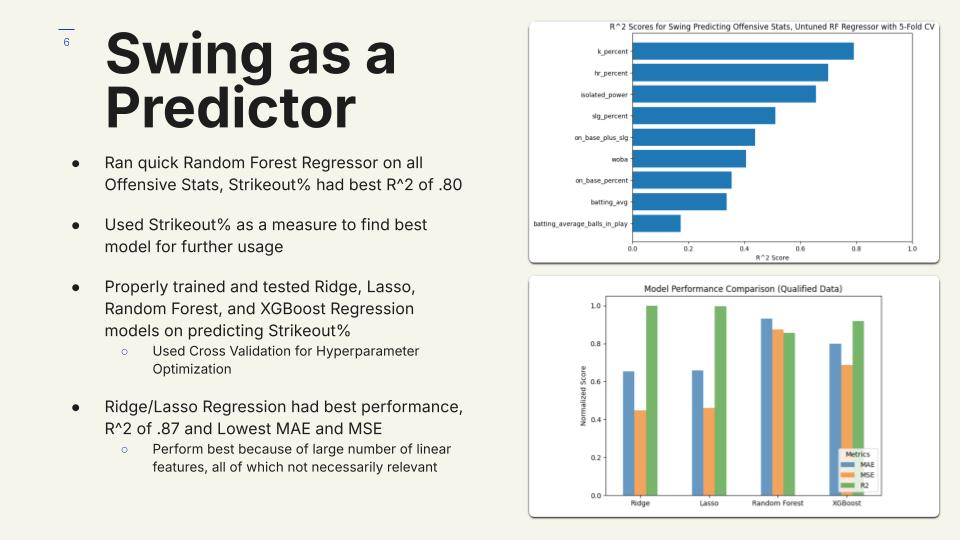
To cap off EDA, a correlation matrix was made from all the variables, shown in the bottom right. Note that pitch specific variables are cut from the visualization shown, for both brevity, and which will be seen later, these pitch specific stats fail to give interesting conclusions. Notable on this heatmap is that batting stance has visibly low correlation with the rest of the variables, which will be explained further in the next slide. Also, while some metrics show collinearity in the correlation matrix, I chose not to remove or modify them as they are known to tell slightly different stories.



I began my statistical analysis by using the stance data to predict the rate at which a batter is hit by a pitch, a question I assumed would be low stakes and give an obvious conclusion. I had assumed that by having a batting stance close to the plate, and maybe close to the pitcher as well, the rate at which you get hit by pitch would dramatically increase. Surprisingly, after experimenting with Linear Regression, Logistic Regression Classification, and Random Forests, I could not build a good model using stance as a predictor for hit-by-pitch rates, with the best R^2 value I obtained being around 0.03. As it turns out, there is much less connection between the two than I had thought, as is especially apparent in the plot in the bottom left between location of a batter and the rate of being hit by pitch as the Z Axis.

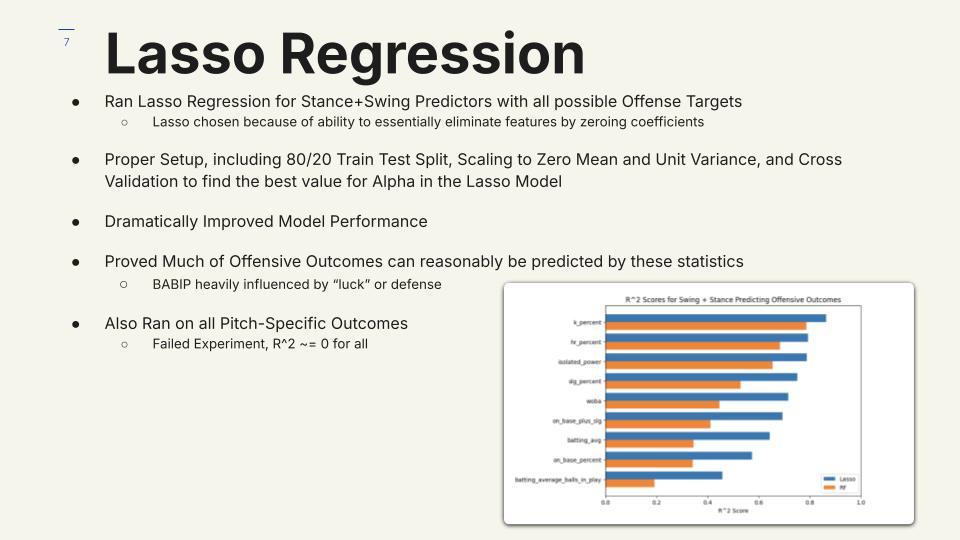
There was but more failure found when using stance to predict general offensive performance, or performance against specific pitch-types. The batting stance variables were used as predictors in creating quick, trial 5-Fold Cross Validated Random Forest Regression Models, and obtained R^2 values that were only negative or close to zero.

Similar results to the aforementioned test were obtained when using stance to predict swing characteristics, though the results were slightly better. The R^2 values found are plotted in the bottom right. Also plotted is the scatterplot between swing speed and each of the batting stance predictors. Swing speed showed the best predictability from stance, with an R^2 of 0.11, and the very loose correlations shown in the scatterplots demonstrates this.



Swing characteristics, on the other hand, performed much better as predictors towards offensive outcomes. The same quick, cross-validated Random Forest Regression test was run using swing stats to predict all of the offensive outcomes, with R^2 scores shown in the top right. Notably, Strikeout percentage, the rate at which a batter gets struck out, was the easiest to predict, with an R^2 of 0.80.

With this, Strikeout Percentage was used as a measure for qualitatively finding the best type of model for further usage. Using the proper methods of the Machine Learning Pipeline, including hyperparameter optimization, Ridge, Lasso, Random Forest, and XGBoost Regression models were all trained and tested on predicting strikeout percentages. The results of this analysis are in the bottom right, with scores of Mean Absolute Error, Mean Squared Error, and R^2 values plotted, normalized from 0 to 1. Ridge and Lasso Regression had the best performance, with an R^2 value for both around 0.87 and low error values. This can be expected due to the linear nature of the stats at play, and the large number of linear features, all of which not being strictly relevant.

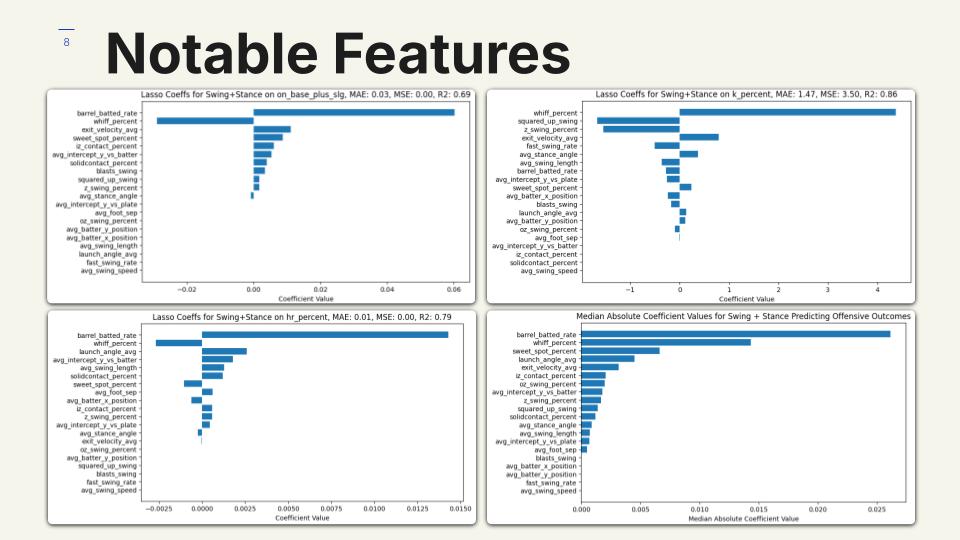


Lasso Regression was then used with Stance and Swing Predictors on all possible Offense Targets. Lasso was chosen over Ridge because of its ability to essentially eliminate features by zeroing coefficients, which is likely to perform best with the large number of combined predictors at play in this scenario, all of which may not be relevant at the same time.

A proper setup was undergone for this examination, including an 80/20 train-test split, scaling to zero mean and unit variance, and cross validation for getting the best value for alpha in the Lasso model. This proper run dramatically improved model performance, shown in the bottom right, where the R^2 values for both the untuned RF regressor from the previous slide in orange and the Lasso model in blue are plotted.

From this, we can see that much of Offensive Outcomes can reasonably be predicted by these statistics, with high R^2 scores showing that much of the variance in the offense targets is explained by Swing and Stance data. Of course, batting characteristics cannot fully predict outcomes. A good indication of this is Batting Average on Balls in Play, or BABIP, that tracks outcomes only when the opposing defensive field players are involved. BABIP can be increased with strong, well placed hits, but much of the stat is determined by luck, or the strength of the opposing defense, so it makes sense that swing and stance characteristics fail to explain a majority of the variance in this target.

This method of Lasso Regression was also run on all the pitch-specific outcomes, though the experiment failed. R^2 scores were around 0 for all of these outcomes, indicating there is more at play.



This final slide of the presentation shows the feature importances, or coefficient values, taken from the Lasso Regression models trained on Offensive outcomes, as described on the previous slide. OPS, the aforementioned offensive gold standard, is in the top left. The top right shows strikeout percentage, the bottom left shows home run percentage, and the bottom right shows the median absolute coefficient value among all targets, looking to show the general usefulness of each predictor.

Notably, Barrel Rate, or the rate at which a batter gets a specific, high-quality launch angle and exit velocity, tracks very well with statistics favoring run output, like OPS and home run percentage. Launch angle, sweet-spot percent, and exit velocity on their own also track with these types of targets, which makes intuitive sense given that these stats inform batting power. On the other side, a target like strikeout percentage favors predictors that inform decision making, such as whiff percentage and swing percentage when the pitch is in the strike zone.

Overall, barrel rate and whiff percentage are shown to be strong predictors of offensive performance, and can be indicative that accuracy and power are important factors towards increasing performance in the batters box.