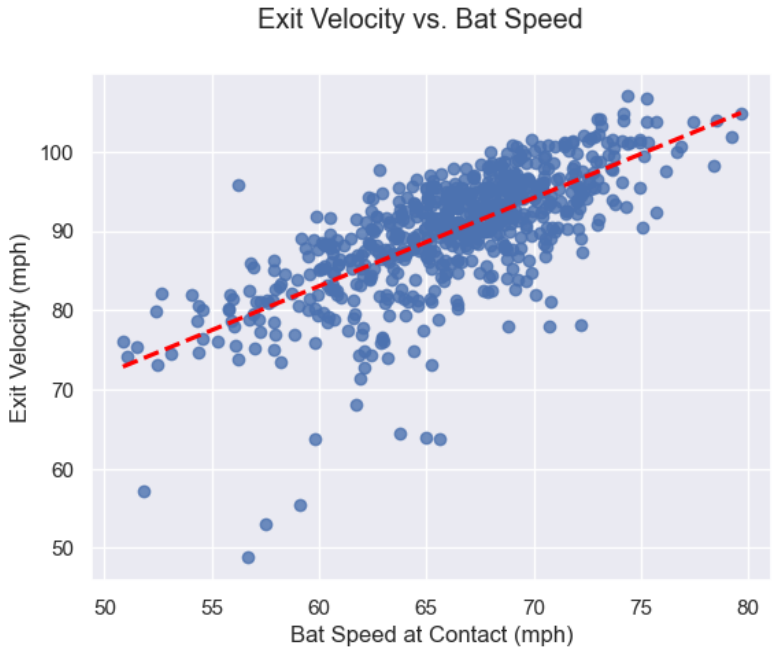
Predicting Bat Speed Using Biomechanical Data

Why Bat Speed Matters

A hitter's approach in any given at bat involves a hit-hit ball in play. Whether it's behind a runner or a deep fly ball, nobody would object to the idea that, excluding bunts, hitting a baseball hard is a good outcome. Yet, a league-average hitter manages to succeed only 2 to 3 times out of 10 plate appearances. With so many uncontrollable variables, the hitter must maximize what can be controlled. In a bat-ball collision, there are three primary factors affecting the post-collision velocity of a struck ball, or exit velocity: the initial ball/bat velocities and collision efficiency, also known as Driveline Baseball's metric called Smash Factor. The latter depends on where the ball is impacted in relation to the "sweet spot," the area on the bat experiencing the least vibration. Barreling a baseball is hard enough, though a hitter's swing speed and approach are fully within their control. Therefore, in a vacuum, if two hitters hit the same pitch at the same spot on the barrel, then the player who can move the bat the fastest at impact will produce a higher exit velocity.

Physics aside, statistical evidence points to a significant positive linear relationship (0.71 Pearson correlation) between bat speed and exit velocity, and higher average bat speeds per hitter also correlate with lower average time to contact and a lower rate of mishits, supporting the notion that increasing bat speed raises the floor for productive mishits. From a player evaluation standpoint, Major League Baseball scouts consider bat speed an aspect of the nebulous "hit tool" and a prerequisite for playing in The Show. Essentially, bat speed is to hitters as velocity is to pitchers – swinging the bat faster generally produces more positive outcomes.



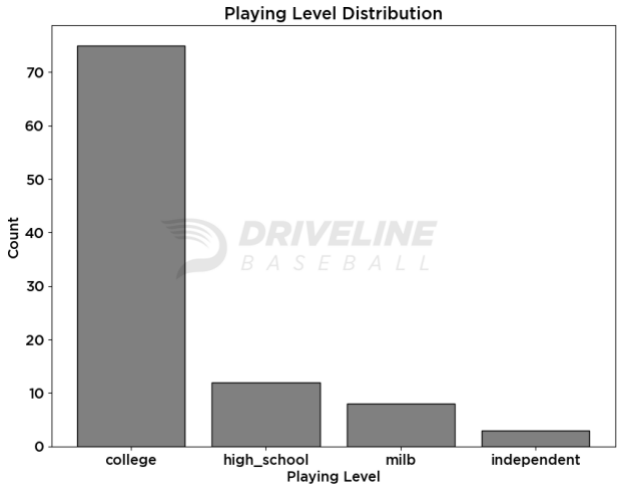
How to Increase Bat Speed

There have been multiple studies revealing significant increases in bat speed from various training methods. Progressive, velocity-based resistance training enhances raw strength, power, and lean body mass, thereby increasing the total amount of force applied into the ground. Generating this ground force, however, is only a piece of the puzzle. Although many players meet the requisite strength benchmark, the ability to *quickly* and *efficiently* transfer this energy into the swing is what separates elite hitters from the rest. Hitting, after all, is a sequential and rotational movement, and proper sequencing and movement patterns efficiently transfer energy through the kinetic chain. Specific resistance implements, such as medicine balls and overweighted/underweighted bats, train fast-twitch muscle groups and explosiveness while reinforcing movement efficiency. Although a rarity among hitting programs, these evidence-based methods produce greater force outputs and establish better rotational movement patterns, unlocking an athlete's untapped bat speed.

As mentioned earlier, elite hitters are elite movers. Not every ballplayer who can deadlift 400 pounds can hit a 95 mile-per-hour heater 400 feet; nor are all power hitters physical specimens built like Aaron Judge or Matt Olson. Otherwise, how would players like Mookie Betts or Jose Altuve put up multiple 30-plus home-run seasons?

The OpenBiomechanics Project

Motion capture (sometimes referred as "mocap") technology is permeating throughout all levels of baseball in an attempt to gather and study kinematic movement data, offering insight into the biomechaical features of elite-level movers. Driveline Baseball's OpenBiomechanics Project (OBP) is the "largest high-fidelity, open-source set of raw and processed motion capture files on elite baseball players in the world." Most athletes in this dataset play in college, with some professional and high-school age athletes included as well. Utilizing this information, this project aims to identify the most important biomechanical features contributing to bat speed.

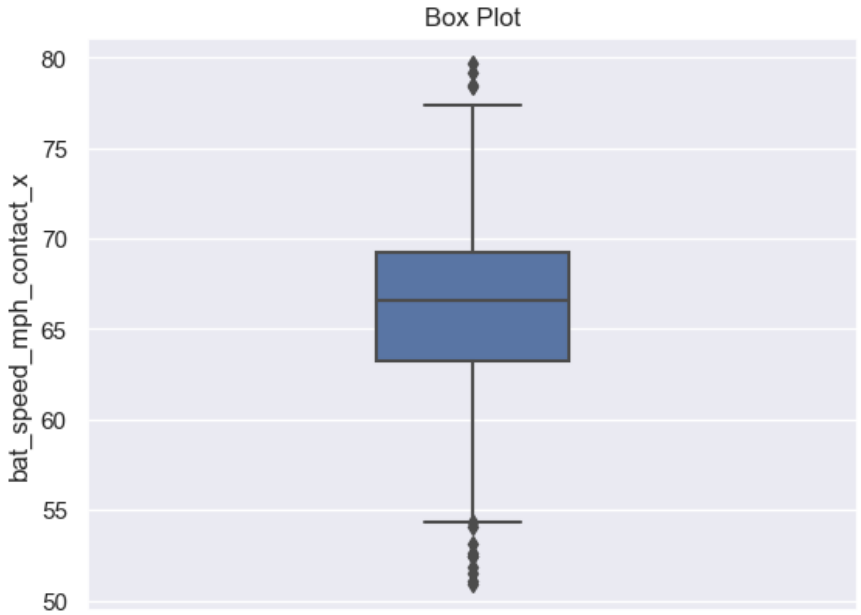


Exploratory Data Analysis

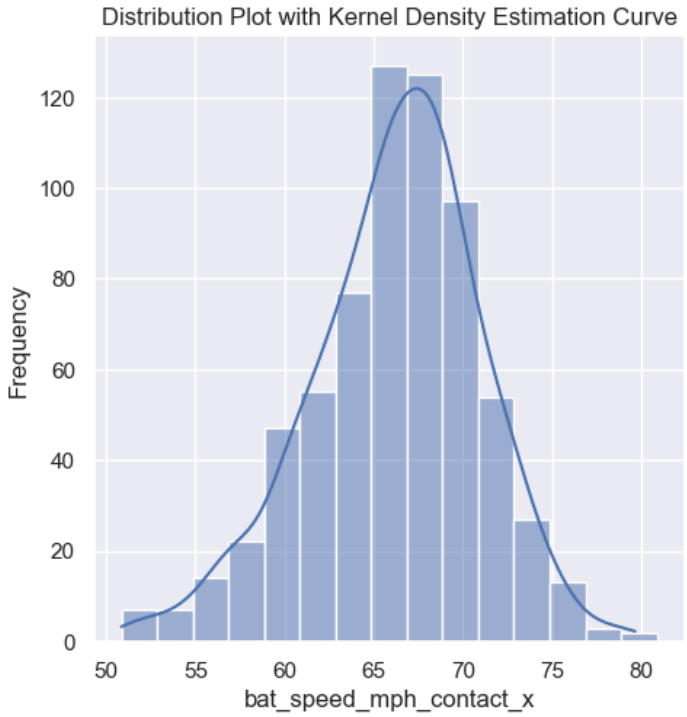
The OBP hitting database consists of biomechanical point-of-interest (POI) metrics (poi_metrics.csv), pitch-level HitTrax data (hittrax.csv), and athlete metadata (metadata.csv) for 98 hitters. The POI dataset contains swing metrics measured by motion capture technology and kinematic metrics for different body parts at the various phases of the swing. The HitTrax data is not used in this analysis since none of the columns relate to bat speed. The metadata contains relevant biological information about each athlete, such as age, height, and weight, that may be associated with bat speed – the thought being that older, taller, and/or heavier athletes may tend to swing the bat faster than less developed or undersized athletes. The POI and metadata datasets are joined to make the final dataset consisting of 677 total observations and 129 columns.

Dependent Variable

The dependent variable to be modeled is bat_speed_mph_contact_x, i.e. bat speed at contact measured in miles per hour (mph). While blast_bat_speed_mph_x measures bat speed using a Blast Motion sensor placed on the knob of the bat, bat_speed_mph_contact_x is measured using motion capture cameras which provide more accurate and reliable readings. It also has no missing data points and less variation, while the Blast data is missing 30 out of 677 entries and has more extreme outliers. The average swing speed is 66.2 mph with a standard deviation of 4.8 mph. The slowest recorded bat speed is 50.9 mph while the fastest is a whopping 79.6 mph, which is considered elite at the professional level. There also appears to be multiple outliers on both sides indicated by the box plot below.



The histogram and kernel density estimation curve show a unimodal, bell-shaped distribution of the dependent variable, and thus is assumed to have an approximately normal distribution.



Feature Variables

Selection and Missingness

For simplicity and interpretability purposes, not all 129 variables can be included in the model. Therefore, a subset of columns in the dataset will be used as feature/predictor variables. First, certain columns associated with bat speed such as exit velocity, sweet-spot velocity, and hand speed are dropped from the original dataset. Additionally, bat_torso_angle_ds_y is dropped since it is the same as bat_torso_angle_connection_x, and session is also excluded since it is a unique identifier, narrowing the number of predictors to 112. Lastly, there are no missing values in the feature data.

Multicollinearity

The variance inflation factor (VIF) quantifies the amount of multicollinearity among the feature variables in the dataset. A VIF greater than 10 indicates that a particular variable is highly correlated with another variable.

Feature	VIF
x_factor_fp_z	inf
upper_arm_speed_mag_max_x	inf
torso_launchpos_y	inf
upper_arm_speed_mag_swing_max_velo_x	inf
upper_arm_speed_mag_seq_max_x	inf
...	...
athlete_age	2.505029
attack_angle_contact_x	2.236508
upper_arm_speed_mag_fm_x	2.179170
bat_torso_angle_ds_z	1.679973
bat_max_x	1.656145

According to the table above, multiple features (35 in total) have infinite VIFs, suggesting perfect multicollinearity. Mathematically, this equates to an R^2 value of either 1 or -1, which intuitively means that one or more variables are exact linear combinations of other variables. To address this multicollinearity issue, the set of features is iteratively reduced until all VIFs fall below 10. A VIF of 10 corresponds to a 0.9 R^2 value, so the final set of features should not have any variables that are highly correlated. Many of these highly correlated variables are recorded at nearly simultaneous phases of the swing or body parts that tend to move in similar fashions, such as the torso and pelvis. The reduced set of features contains 25 variables.

Model Building

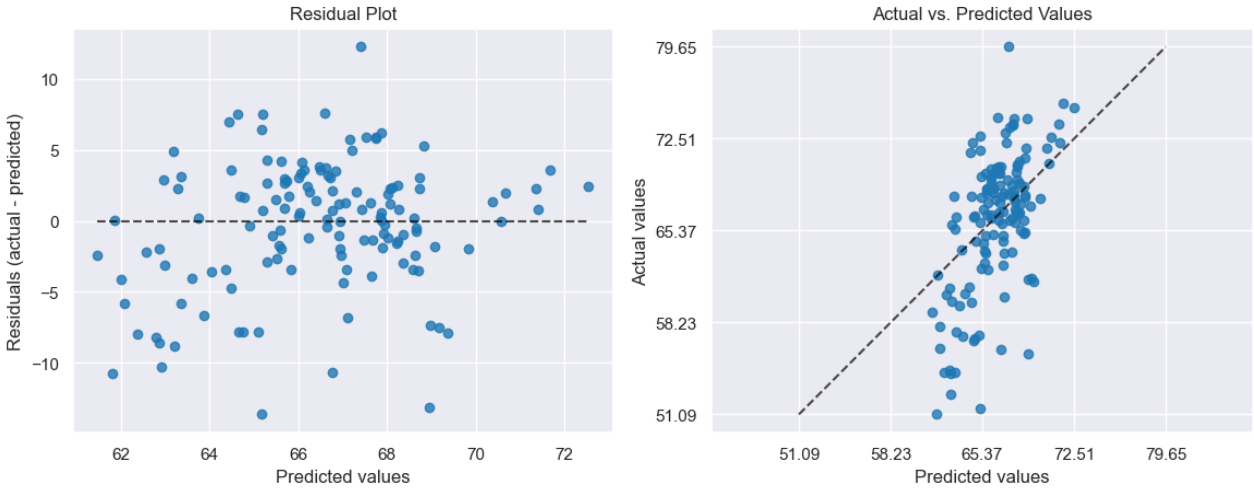
Multiple Linear Regression

The first model will use multiple linear regression (MLR) to estimate bat_speed_mph_contact_x using an 80-20 split of training and test sets containing 541 and 136 observations, respectively. After fitting the model on the training data and making predictions on the test set, it is evaluated using the following performance metrics:

- Mean Squared Error (MSE): 20.21

- Root Mean Squared Error (RMSE): 4.50
- R-Squared (R^2): 0.31

The R^2 value of 0.31 suggests that the model explains approximately 31% of the variance in the bat speeds within the test dataset. The RMSE value shows that the average difference between the model's predicted values and the actual values in the test dataset is about 4.5 mph.



The low R^2 value suggests that the model does not fit the data well, and the actual versus predicted values plot shows that the range of predicted values is much less than the range of actual values. The residual plot shows no clear patterns as the residuals are relatively symmetrically distributed about the horizontal axis. However, there are some significant prediction errors, with some deviating by more than 10 mph, indicating the poor model performance and possibly some influence by the outliers in the dataset.

Outliers

The presence of outliers tends to affect the calculation of regression coefficients, skewing the slope and intercept of the line. Therefore, these outliers will be identified using the interquartile range (IQR) approach. IQR is the difference between the third quartile (Q3) and the first quartile (Q1). Values that are above or below the following upper and lower bounds are labeled outliers and removed from the dataset:

- Upper Bound: Q3 + 1.5 * IQR
- Lower Bound: Q1 - 1.5 * IQR

According to these thresholds, 14 observations are considered outliers and removed from the dataset. This new dataset is again split into training and test sets before it is fitted to another multiple linear regression model. While the MSE and RMSE slightly decreased, the R^2 decreased to about 0.10. Since the overall performance of this MLR model without outliers is worse than the previous model, it may be beneficial to try a different modeling technique to improve these metrics.

Elastic Net

Elastic net regularization is particularly useful for regression models where multicollinearity exists among the features and/or when you want to perform feature selection. The regularization term in elastic net is a linear combination of L1 and L2 penalties, which essentially control the complexity and capacity of the model. The L1 (Lasso) penalty helps in reducing the number of features by shrinking coefficients to zero (effectively performing feature selection) while the L2 (Ridge) penalty shrinks the coefficients towards zero but does not set them to zero, which helps in dealing with multicollinearity.

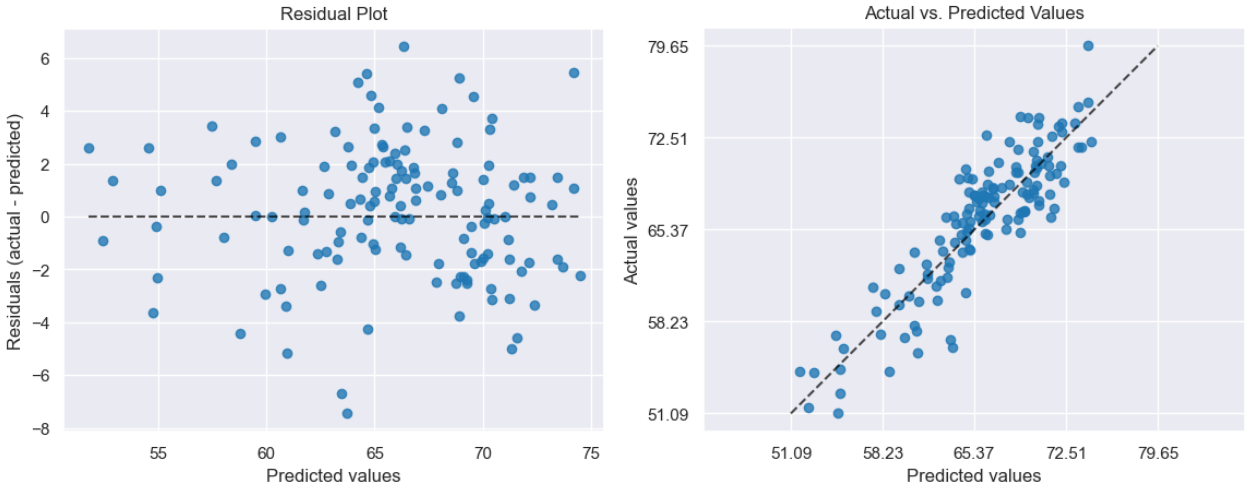
An elastic net function from Python's scikit-learn library is designed specifically for regression models incorporating this type of regularization. This function automatically performs 5-fold cross validation to determine the best parameters for the model, specifically focusing on the regularization strength (alpha) and the mix between L1 and L2 penalties (lambda or l1_ratio). It tests 100 different alpha values with each l1_ratio by default and evaluates the model's performance for each combination.

Most of the default parameters are used for this model, but a list of values for l1_ratio is used instead of the 0.5 default value. A higher l1_ratio pushes more coefficients to zero, which is useful for feature reduction and enhancing model simplicity. Conversely, a lower l1_ratio integrates more L2 penalty and handles features with high collinearity. By passing a list of these values, the function tests each value with each alpha and ultimately selects the one yielding the best prediction score.

Lastly, feature scaling ensures that each feature contributes equally to the regularization process, allowing the model to converge more quickly and accurately. Unscaled or unnormalized features can disproportionately influence the penalty applied to different coefficients, leading to biased estimates and an inefficient learning process.

The full dataset (including outliers) is used for the train-test split with 541 observations in the training set and 136 observations in the test set. The model incorporates all 111 features as it is designed to automatically address the multicollinearity issue. The fitted model used an alpha value of about 0.008 and l1_ratio of 0.5, indicating an equal mix between L1 and L2 penalties. The following performance metrics and plots exhibit a significant increase in the elastic net model's performance:

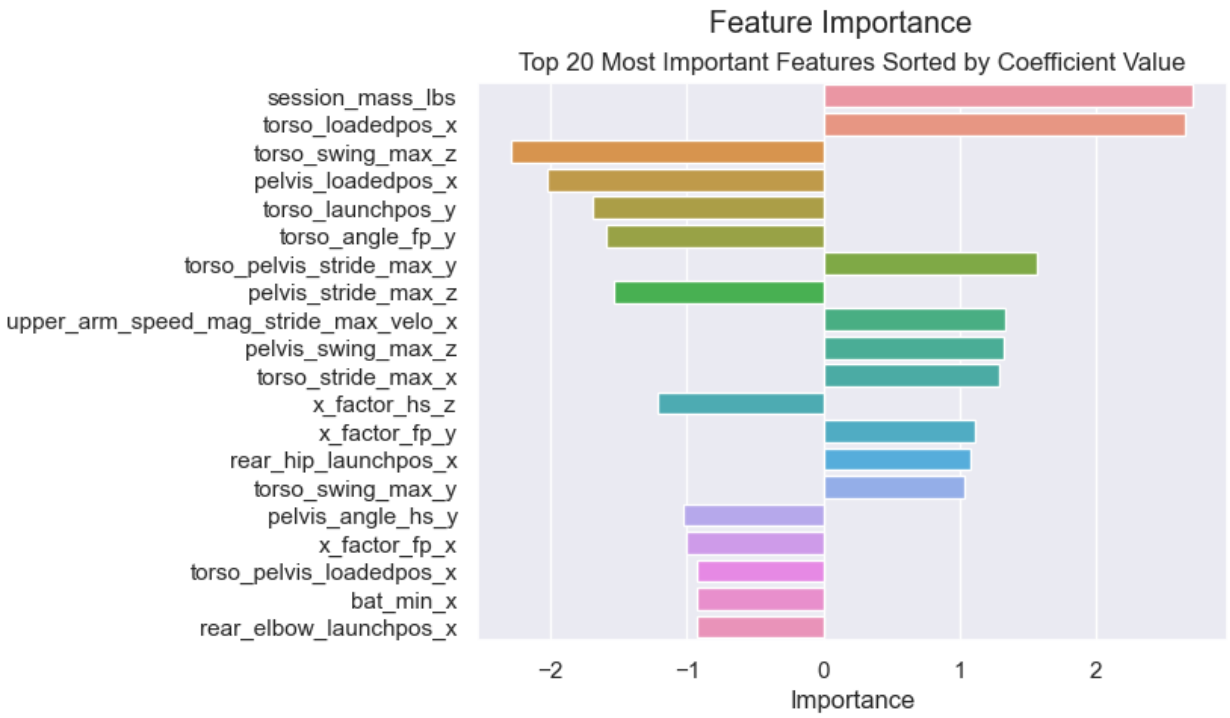
- MSE: 6.41
- RMSE: 2.53
- R^2: 0.78



Again, the residual plot shows no clear pattern, although some of the errors tend to increase with a predicted value greater than 60 mph. However, the actual versus predicted values plot suggests improved model accuracy since the data points cluster around the identity line where the predicted values equal the actual values. Additionally, the range of predicted values closely resembles the range of actual values.

Feature Importance and Insights

A bar chart of the elastic net model's coefficient values in descending order is shown below. Athlete body mass (session_mass_lbs) is the most influential variable contributing to bat speed, supporting the notion that strength training and an increase in muscle mass leads to higher bat speeds. Torso angle at the loaded position (torso_loadedpos_x) is the most influential biomechanical variable, with the positive value suggesting that more counter rotation (away from the pitcher) could help a hitter swing the bat faster. Unsurprisingly, maximum torso-pelvis angle (i.e. hip-shoulder separation) has a positive impact on bat speed since this aspect usually signals a mechanically efficient rotary movement. Maximum torso angle between the load and contact (torso_swing_max_z) has the largest negative influence on bat speed, though the negative sign is likely because the torso is rotating toward the pitcher during this time. One counterintuitive finding from the feature importance chart is that no variables related to angular velocities related to the torso or pelvis are included. One would think that faster trunk rotation would correlate to a higher bat speed, but the model seemed to not take this into consideration.



Model Improvements

Although the results of the elastic net model are encouraging, there are several other aspects to consider that could further improve its performance. First, other manual variable selection methods, such as forward or backward stepwise selection, could lead to similar or better results since these methods do not require feature scaling, therefore boosting the model's interpretability. However, this technique would be quite time consuming given the large number of columns in the data. Principal component analysis is another dimensionality reduction technique that was considered but not used due to time constraints and lack of interpretability.

Only 98 athletes are included in this dataset, so increasing the data collection to a wider range of ages and playing levels could provide more convincing results. Additional information pertaining to an athlete's strength metrics or force plate data could potentially provide other possible factors relating to bat speed. Consulting with a biomechanist or sport scientist might also be beneficial to learn more about the POI dataset and the interaction between its variables.

Finally, no advanced machine learning techniques, like ensemble methods or deep learning, were utilized to enhance the prediction accuracy. While these methods may be beneficial for handling larger and more complex datasets, the performance of the elastic net regression was deemed satisfactory.

References

Nathan, Alan M. "Characterizing the performance of baseball bats." *American Journal of Physics* 71.2 (2003): 134-143.

Szymanski, David J., Coop DeRenne, and Frank J. Spaniol. "Contributing factors for increased bat swing velocity." *The Journal of Strength & Conditioning Research* 23.4 (2009): 1338-1352.

Wasserberger KW, Brady AC, Besky DM, Jones BR, Boddy KJ. The OpenBiomechanics Project: The open source initiative for anonymized, elite-level athletic motion capture data. (2022).

https://www.drivelinebaseball.com/2021/02/smash-factor-a-data-driven-approach-to-assessing-the-hit-tool/

https://www.drivelinebaseball.com/2019/05/debunking-bat-speed-myths/

https://blogs.fangraphs.com/scouting-the-hit-tool-pt-1/