Will Gibbs and Drake Lewis

Predicting MLB Hitters’ Salaries

This project aims to find the most important variables in determining player salary based on their hitting statistics from the 2011-2022 seasons. Seeing all of the big MLB contracts given out in the past two years, our goal is to see the reasoning behind handing out so much money. Are the contracts driven by the desire to win or by the desire to sell tickets? The data was taken from a GitHub repository called MLB-salary-prediction (created by cj0121) and includes all of the players who received contracts during the 2011-2022 time period with their corresponding statistics from the past 5 seasons (must have had at least 400 ABs).

Our analysis started with a multiple linear regression model (R^2=.669) founded by the best subsets technique which included Age, X2B_pos, C_pos, PA, TB, ISO, wRAA, and WAR as significant predictors for Salary (besides X2B_pos). This model indicated that X2B_pos, Age, PA, and ISO were all negatively correlated with salary, slightly supporting the motivation of winning. The LASSO regression model (R^2=.6657), included 15 predictors that had variables such as Age, X1B_pos, X2B_pos, BB, HBP, and OBP with negative coefficients. These findings suggest that monetary gains more influence the motivation.

To find the most important predictors for salary, we used the random forests and XGBoosting techniques. The random forests model had an R^2=.664 and found that WAR was by far the most important predictor in determining salary, followed by wRAA. For the XGBoost model, it resulted in an R^2=.696 and also clearly saw WAR as the most important predictor, again followed by wRAA. We were expecting OPS to have more importance, as much of the literature we read stated that OPS was very important, but for our models WAR consistently dominated.

Our findings in this project tell us that WAR is the most important predictor in determining salary, which makes sense considering the variable is made from hitting, fielding, and baserunning statistics. In addition, XGBoosting produced the most accurate model compared to the other methods, showing its effectiveness in predicting values.