CSCI 183 Final Project Report

Strikeout predictions

Jonathan Wu jywu@scu.edu and Jared Maeyama jmaeyama@scu.edu

I. Abstract

We wanted to apply data science that we have learned in our class to sports. One way we thought that we could use data science was sports betting. Sports betting is very popular, and is still getting very popular due to many reasons such as technology making it more convenient and straightforward for people to get into sports betting. Out of all the sports out there, we were the most comfortable and knowledgeable in baseball, so we decided to narrow down the focus to baseball, since collecting all data from all the sports known to world wide would've been too much. Since we narrowed it down to baseball, we wanted to make a prediction algorithm for strikeouts, so people could have a general idea of where to bet their money. We would use the data set of all starting pitchers in the MLB for the 2022 season, all MLB team batting statistics and the strikeouts from every MLB start. The program would output the predicted number of strikeouts as a reassurance that they are making the correct bet, or a correction that they will make a wrong bet. In general, people could use this algorithm themselves for sports betting and it could cause less harm from sports betting.

II. Strikeout bets

First, what is a *strikeout bet*? Strikeout bets consist of a set over/under that the pitcher you bet on needs to hit. This means that you are betting on the number of strikeouts a pitcher throws in a game. For example, ifa pitcher had a line over 4.5 strikeouts, you would want that pitcher to throw 5 strikeouts for the bet to cash. Anything under 4 strikeouts would cash if you bet the under. Most of these bets are on half points, which means that there is always a winning side and losing side. There are other popular bets on baseball, which are moneyline, run line, and run total bets. These bets are the simplest since you are either betting on a team to win, to win or lose by a specific margin, or on a combined number of runs scored between two teams total.Where to place these bets? Well there are many online websites and apps where you can place these bets in, and of course the sportsbooks.

III. Sports betting

Sports betting is popular among everyone. There was a study done by Emma Seal, where they had survey responses from a sample of nearly 15,000 Australian sports fans that were used to study the determinants of gambling behavior, including if a person does gamble and the type of gambling engaged with, the number

of sports and non-sports bets made over a 12-month period, and attitudes towards betting on sports. In that study, it showed how the probability of betting on sports decreased with increasing age and was lower for women and people with a university education. Another result was that within friendship circles, “the views that sports betting is perceived as harmless, common and very much a part of enjoying sports were stronger among young men” (Seal).

IV. Data Collection

There are public data sets that exist for the stats of every MLB player for each season, individual player statistics, and each team batting statistics. There are challenges that we encountered in the data collection process. In the data set we had, the information showed the pitcher’s statistics for every single game of the whole season, but there wasn’t any data that showed the team they played against on that day, only the team that they played for. In order to solve this problem, we found the team’s pitching statistics for every game and also the batting statistics for every game and we combined the date of every game, Pull%, Cent%, and Oppo% into a string so every single game would have a unique id. So to speak, we could match up for every single game to see which team played against who. The K% and SO% are the same thing, which means a strikeout. There was also one other challenge that we faced. Sometimes there would be players that would play for two teams in the entire season, so when we combined the data sets, it would show that the team they played for was 2 Tms, meaning 2 teams. What we did to fix this problem, was instead of basing it off of the team column from the player’s year statistics, we based it off of the players team column for the individual starts as for each game that players would play for one team at a time. The data we collected was the pitcher’s statistics, the pitcher’s advanced statistics, team batting’s statistics for every game, team batting’s advanced statistics for every game, team’s pitching statistics for every game, and the statistics of every single start of each pitcher. We also used extra data such as the statistics of the batted balls of each team for every game for the first challenge that we encountered to figure out which team played which team.
And this doesn’t even vary in one sport, it contains all the sports. There are no sports events where you can’t find a betting type. The study shows how sports betting is very popular and the reasons behind the popularity in sports betting such as that players can now enjoy more exciting ways to place bets. Nowadays, technological innovations have also made an impact on sports betting, making it more diverse and providing more opportunities to bet on where that person's interest is. It makes sports betting more convenient and straightforward, making it simple and easy for people to make sports bets. For an algorithm to make for sports betting, it can be very useful nowadays since many people could use this algorithm for their own use. Still, there are people

V. Implementation

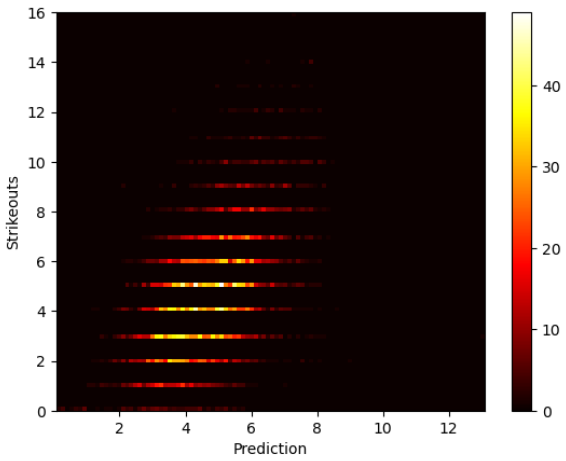
First we wanted to use all the data that we collected. To do this, we wanted to combine the spreadsheets of the data together. This would cause some errors from the ones we mentioned before, but we used the solutions to fix these errors. After mashing all the spreadsheets together, we used linear regression as our algorithm for the prediction of the strikeouts. We used three variables in our predictions. The three variables were K% of the pitcher, K% of the batter, and the total strikeouts that were thrown from the pitcher. As adding further variables the calculations used the games stat from the batting team which counted up how many unique players played in each game. We stopped at three to avoid the algorithm from learning the noise of the data. We used liner_model form sklearn for our linear regression

VI. Results

After using the linear regression model we got the K% from the hitter, total SO from the pitcher and K% from the pitcher were our three inputs to our algorithm. Our formula was

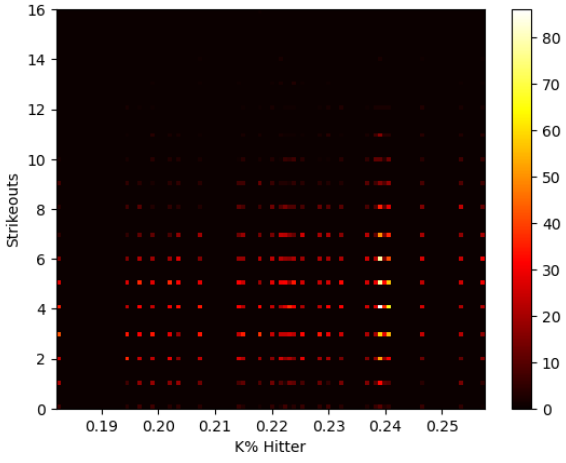
$$20.3552487 * K\%_{Pitcher} + 0.012251037 * SO + 12.6090668 * K\%_{Hitter}$$

Using our algorithm we graphed a heat map of our Prediction Vs. Strikeouts for each outing.

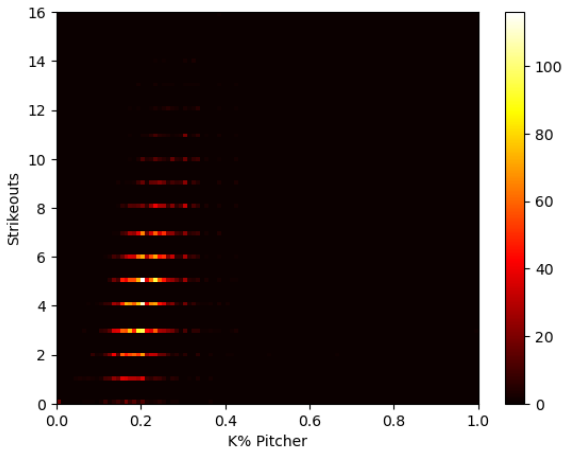


Graph 1

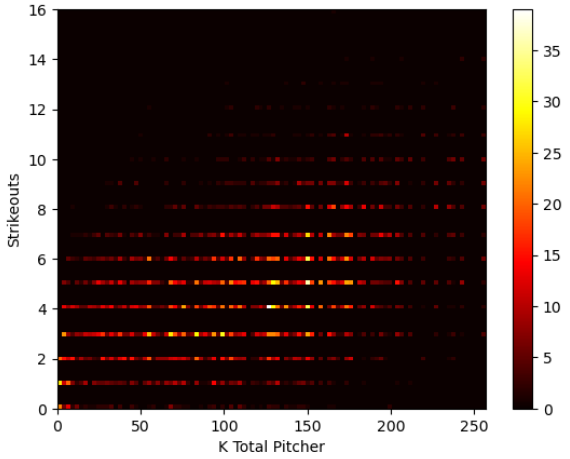
Ideally there is a liner 1:1 line between the prediction and the strikeouts but due to the data used and the inherent randomness of baseball there is a .529298 correlation between our prediction and strikeouts. The model has a squared error of 10.21. As you can see in the graphs below there is a loose liner correlation between the three input variables and the Strikeouts for the outing.



Graph 2



Graph 3



Graph 4

VII. Conclusion

In this paper we presented an algorithm that can determine the number of strikeouts that a certain pitcher would make, with the input of the stats of the pitcher and the batter they are facing from the user. People can use this output as reassurance that they are making a good or bad decision on the bet, or even make a correct bet if someone does not have the information from the pitchers and the batters. We collected a dataset that consists of all the starting pitchers in the MLB for the 2022 season, and all MLB team batting statistics and the strikeouts from every MLB start. Using this output, people could make better bets or predictions, causing more money gained than lost in bets. For future iterations we would like to look more into advanced statics and nonlinear regression models. Though inherently bound by the randomness of baseball, stats outside of just K% and total SO could help create more accurate predictions. Along with the ideas of garbage in garbage out, statistics that can better correlate with the data would lead to better predictions. The most accurate model might not be linear and so exploring more types of prediction algorithms would be more beneficial.

VIII. References

Davis, Kevin. “How to Bet on Baseball and Win: Beating MLB Pitcher Strikeout Total Props without a Model.” *American Betting Experts*, 17 Oct. 2022, https://americanbettingexperts.com/how-to-bet-on-baseball-and-win/.

“MLB Strikeout Prop Bet Picks Today: Explaining Baseball Strikeout Betting.” *OddsJam*, 5 Dec. 2022, https://oddsjam.com/mlb/mlb-strikeout-prop-bet-picks-today.

Seal, Emma, et al. “The Gambling Behaviour and Attitudes to Sports Betting of Sports Fans - Journal of Gambling Studies.” *SpringerLink*, Springer US, 1 Feb. 2022, https://link.springer.com/article/10.1007/s10899-021-10101-7.

“Splits Leaderboards.” *FanGraphs*, https://www.fangraphs.com/leaders/splits-leaderboards?splitArr=&splitArrPitch=&position=B&autoPt=false&splitTeams=false&statType=team&statgroup=1&startDate=2022-3-1&end

Date=2022-11-1&players=&filter=&groupBy=season&wxTemperature=&wxPressure=&wxAirDensity=&wxElevation=&wxWindSpeed=&sort=22%2C1&pageitems=10000000000000&pg=0.

"Splits Leaderboards." *FanGraphs*, https://www.fangraphs.com/leaders/splits-leaderboards?splitArr=&splitArrPitch=&position=B&autoPt=false&splitTeams=false&statType=team&statgroup=2&startDate=2022-3-1&endDate=2022-11-1&players=&filter=&groupBy=season&wxTemperature=&wxPressure=&wxAirDensity=&wxElevation=&wxWindSpeed=&sort=22%2C1&pageitems=10000000000000&pg=0.

"Splits Leaderboards." *FanGraphs*, https://www.fangraphs.com/leaders/splits-leaderboards?splitArr=&splitArrPitch=&position=B&autoPt=false&splitTeams=false&statType=team&statgroup=3&startDate=2022-03-01&endDate=2022-11-01&players=&filter=&groupBy=game&wxTemperature=&wxPressure=&wxAirDensity=&wxElevation=&wxWindSpeed=&sort=23%2C1.

"Splits Leaderboards." *FanGraphs*, https://www.fangraphs.com/leaders/splits-leaderboards?splitArr=&splitArrPitch=&position=P&autoPt=false&splitTeams=false&statType=team&statgroup=3&startDate=2022-03-01&endDate=2022-11-01&players=&filter=&groupBy=game&wxTemperature=&wxPressure=&wxAirDensity=&wxElevation=&wxWindSpeed=&sort=23%2C1.

"Splits Leaderboards." *FanGraphs*, https://www.fangraphs.com/leaders/splits-leaderboards?splitArr=42&splitArrPitch=&position=P&autoPt=false&splitTeams=false&statType=player&statgroup=1&startDate=2022-3-1&endDate=2022-11-1&players=&filter=IP%7Cgt%7C0&groupBy=season&wxTemperature=&wxPressure=&wxAirDensity=&wxElevation=&wxWindSpeed=&sort=22%2C1&pageitems=10000000000000&pg=0.

"Splits Leaderboards." *FanGraphs*, https://www.fangraphs.com/leaders/splits-leaderboards?splitArr=42&splitArrPitch=&position=P&autoPt=false&splitTeams=false&statType=player&statgroup=2&startDate=2022-3-1&endDate=2022-11-1&players=&filter=IP%7Cgt%7C0&groupBy=game&wxTemperature=&wxPressure=&wxAirDensity=&wxElevation=&wxWindSpeed=&sort=22%2C1&pageitems=10000000000000&pg=0.

"Splits Leaderboards." *FanGraphs*, https://www.fangraphs.com/leaders/splits-leaderboards?splitArr=42&splitArrPitch=&position=P&autoPt=false&splitTeams=false&statType=player&statgroup=2&startDate=2022-3-1&endDate=2022-11-1&players=&filter=IP%7Cgt%7C0&groupBy=season&wxTemperature=&wxPressure=&wxAirDensity=&wxElevation=&wxWindSpeed=&sort=22%2C1&pageitems=10000000000000&pg=0.

Staff, Props. "Best MLB Strikeout Props Today (Daily Picks)." *PROPS*, 9 Nov. 2022, https://props.com/best-mlb-strikeout-props-today/.