Strikeout Prediction

Building a Neural Network Model to predict Starting Pitcher Performance in Baseball Games

By: The Strikeout Squad
Aditya Chakravarthi (achakravart4@wisc.edu)
Zachary Majca (majca@wisc.edu) †

1. Overview: 1-2 sentences describing this project
    - Our goal for this project is to use a neural network to predict MLB strikeouts per inning. We will use data from the top 300 pitchers and 32 teams to train our model on predicting the expected strikeouts per inning and use that to predict the strikeouts per game for a starting pitcher.

2. Background: Explain a bit background of this project, what have been done before, what is the state of the art outcome (A good place to cite a couple of most relevant works, including URLs, citation to references, etc.)

- Pitcher strikeout percentage has been one of, if not the most important statistics for determining a pitcher's success. Many models have been created determining the outcome and pitching performance of specific games, however we are making it more specific by focusing on strikeouts. A couple of advanced projects have been done regarding the outcome of a specific game, however, they are outside the scope of this class.
- MLB.com has a predicted list of who will lead the league in strikeouts at the end of the season. This link is beneficial as they mention key stats that we will use in our project.
    - https://www.mlb.com/news/2023-mlb-strikeout-leader-predictions

3. Statement of Work

3-1. Datasets: where it comes from? What are in this dataset? Your preliminary inspection of the datasets and what others have said about this dataset.

- Our primary dataset comes from statcast and Fangraphs. MLB statcast allows a user to select the data sets they want to look at and download them into a csv file. For our top 300 starting pitchers, we will look at 9 different categories.
    - k%
    - Out of zone swing and miss %
    - In zone contact %
    - Whiff %
    - First strike %
    - GB, FB, LD, Popup %
- For hitting we will use the stat cast hitting to find the strikeout percentage per team. Using strikeouts/plate appearances
- The data set appears to be elite. With no missing categories it was overwhelming to decide. We have handpicked the stats that we think are the most important.
- Other people rely heavily on statcast MLB for their existing neural network models and have only said good things about it.

3-2. Method: What you plan to do? What are existing similar works you find so far? What have been accomplished so far (prior work). What you plan to do? Why you want to do this?

Both of us are interested in baseball and all the advanced stats that tell the story of the sport. On top of that, we have also taken an interest in the betting side of sports, particularly with pitcher strikeout lines. As stated before, projects that can predict and model the outcome of certain games, given the statistics of a particular lineup and pitcher have been done before, however we are solely focusing on starting pitcher strikeouts. We plan to find the average strikeouts per inning for the top 300 pitchers in the league.

Similar projects are located in the reference section.

3-3. Outcome and Performance evaluation: What outcome you anticipate in doing this project? what criteria (performance, costs) will be used to gauge quantitatively whether the project is successful?

There are a few potential outcomes that we are anticipating and hope to see by the end of the project. Firstly, The neural network model aims to provide more accurate predictions of the number of strikeouts compared to other traditional statistical models or baseline approaches. The anticipated outcome is a model that demonstrates improved performance in terms of mean squared error (MSE), mean absolute error (MAE), or other relevant performance metrics. Secondly, By training the neural network on a dataset containing various pitcher and game-related features, the project can uncover insights into the factors that significantly influence the number of strikeouts. This analysis can help identify the key variables that contribute to successful pitching performances. However, The ultimate outcome of the project could be the practical deployment and utilization of the developed model. If the model proves accurate and useful, it can be integrated into baseball analytics systems, scouting reports, or used by coaches and analysts to assess pitcher performance and make informed decisions.

As briefly mentioned before, the main criteria that will be used to gauge the success of the project are some performance metrics to be decided upon in the future. Metrics such as Mean Squared Error (MSE) to calculate the average squared difference between the predicted and actual number of strikeouts, where a lower MSE indicates better performance. Also, another metric such as Mean Absolute Error (MAE), which is determined by calculating the average absolute difference between the predicted and actual number of strikeouts, where a lower MAE would indicate a better performance. As there are a lot of independent variables within our dataset that can and will impact

the outcome, a metric such as R^2 score which takes the proportion of the variance between the dependent variable, strikeouts, and the independent variables, the various statistics we are using. In this case a higher R^2 score would indicate better performance. Other evaluators such as various cross validation techniques that determine the generalizability of the model and reduce the amount of overfitting we do, while assessing our models performance over various subsets of data, would be important as well. We also need to be mindful of the cost considerations of our project, such as things like training time and the amount of computational resources we are using.

4. Project Plan

Break down the project into a set of tasks and provide a Gantt chart that schedule the beginning and end of each task within a duration of 7 weeks. Potential tasks may include (but not limited to) setting up environment (e.g. GitHub account), data pre-processing, programming (specifying specific programming tasks as detail as possible), progress report and final report preparation, project presentation preparation, etc.
Each project will be conducted using a GitHub project. The project plan section should include the URL link to the GitHub project and will provide read access by the instructor and the TA.

Week	1	2	3	4	5	6	7
Setup	Github account						
	Project proposal						
Data	Collect Data	Data collection					
		Preprocessing Data					
		Splitting Data Set					
Architecture			Choose Architecture	Program Neural Network	Create test/validation methods		
Model						train/test model	

						Validate / Evaluate	
Report							Deploy Model
							Write Report

5. References

Give a numbered list of research papers, reports, blogs that you cite in your proposal. Output is strikeouts per inning.

1. Fangraphs - For Data Collection
2. Statcast - For Data Collection
3. An Advanced Neural Network For predicting MLB Outcomes
4. This project is a prediction of strikeout rates
5. Pitching Projections in 2021 Baseball

Github Repository

https://github.com/zmajca/StrikeoutPredict.git