Joe Stella

Data Science Project Proposal

Bat Oracle

Introduction:

The mission of Bat Oracle is to revolutionize MLB batting predictions. Traditional betting odds often favor casinos and profit, while coaches frequently rely on subjective player knowledge and raw stats rather than a set algorithm. By harnessing unbiased statistical analysis, Bat Oracle aims to provide a more accurate assessment of a batter's performance against specific pitchers.

Historically, batters exhibit consistent patterns when facing pitchers with similar handedness, favored pitch types, velocity ranges, K\9, ERA, and exHR. once rivalries and streaks are removed from consideration. By leveraging comprehensive data from pitchers and batters spanning the last 10 years, my goal is to develop a predictive model capable of accurately projecting a batter's performance against a given pitcher and their tendencies. I aspire for the model to achieve a success rate of over 70%, which will be judged by the precision of the stats when compared to the true result.

Drawing inspiration from successful ventures like 'Moneyball' and 'Statcast,' Bat Oracle strives to bring cutting-edge analytics to the forefront of baseball strategy, offering teams and enthusiasts alike a powerful tool for informed decision-making.

Proposed work:

* The project would start by first refining data sets.
  + First, I would need batters and pitchers stats from the last 10 years
    - This can be found on pybaseball
  + Then, I would need a database of every baseball game in the last 10 years
    - Also found in pybaseball
  + I would then process the data by including the performance of batters against given pitcher stats.
  + I would then train the model with the new data set.
* After this I would create the program that would take the given pitcher and given batter and find the batters expected performance based on the trained model

Timeline:

Data Collection: I plan to gather two main types of data:

Batter and pitcher stats from the last 10 years, which I can obtain from sources like pybaseball.

A database of every baseball game in the last 10 years, also available in pybaseball.

Data Processing: After collecting the data, I would need to process it. This would involve:

Combining the batter and pitcher stats with the game data to create a comprehensive dataset.

Possibly cleaning and organizing the data to make it suitable for analysis.

Model Training: With the processed dataset, I would train a model. This model would learn the relationships between various factors such as pitcher stats, batter stats, game conditions, etc., and the performance of the batter.

Model Evaluation and Refinement: After training, I'll evaluate the model's performance using appropriate metrics. If necessary, I might refine the model by adjusting parameters or trying different algorithms.

Development of Prediction Program: Once I have a trained and validated model, I would develop a program that takes inputs such as a pitcher's stats and a batter's stats, and outputs the expected performance of the batter based on the trained model.

Testing and Deployment: Finally, I'll test the prediction program thoroughly to ensure it performs as expected. Once satisfied with its performance, I could deploy it for actual use.