Bryan Rogers
6/17/2024
Bellevue University
DSC680: Applied Data Science

The First Project – Milestone 2

Draft White Paper: Predicting Major League Baseball Game Outcomes with Machine Learning

Business Problem

Major League Baseball (MLB) has historically relied on expert opinions, data analysts and historical statistics to help determine the outcomes of games and attempt . This white paper hopes to explore the development of a machine learning model to predict game winners, offering a data-driven approach for informed decision-making and enhanced game analysis.

Introduction

Accurately predicting game winners for Major League Baseball (MLB) can provide a significant benefit to many groups of people including fans, team analysts, and sports gamblers among others. This project hopes to investigate the application of machine learning, a powerful tool within artificial intelligence, to predict the outcomes of professional baseball games. By analyzing historical data and identifying key attributes that influence the outcomes of baseball games, the model can potentially provide valuable insights that complement traditional analysis methods and add another level of advantages for teams looking to gain a competitive edge.

Background/History

Baseball has long embraced statistical analysis to understand player performance and team strategies, analytics have been a key part of the game for over 100 years and are more heavily relied on than any other major professional sport. Sabermetrics, a term coined by Bill James, has revolutionized the game by applying statistical methods to gain a competitive advantage. It is time to take baseball analytics to the next level by using machine learning algorithms and AI to develop future strategies and ways to gain a competitive advantage.

Data

This project aims to use a robust dataset encompassing historical MLB games.

Data Source

- Baseball-Reference.com **-** https://www.baseball-reference.com/
- Retrosheet - https://www.retrosheet.org/
- Kaggle - https://www.kaggle.com/datasets/saurabhshahane/major-league-baseball-dataset

Data Preparation

The raw data will go through rigorous data cleaning and preprocessing to address:

- **Missing Values:** Imputation or data deletion can be used to handle missing data
- **Inconsistencies:** Data formatting errors or terminology issues can be fixed for better analysis
- **Errors:** Data errors will be found and fixed to ensure data accuracy

Data Dictionary

The prepared data will include a diverse range of features potentially relevant to game outcomes, some of these data points and categories can include:

- Team Statistics
  - Batting Average (BA / AVG.)
  - On-base percentage (OBP)
  - Slugging percentage (SLG)
  - Earned run average (ERA)
  - Fielding percentage (FLD)
- Player Statistics
  - Batting Average (BS / AVG.)
  - Home Runs
  - Hits
  - Runs Batted In (RBI)
  - Plate Appearances (PA)
- Matchup Statistics
  - Head-to-Head performance data
- Game Factors
  - Day and time of game
  - Weather Conditions

Methodology

Machine Learning is the fundamental method that will be used in this project and the development and training of a model will be the most important.

Feature Engineering

Beyond the raw data points that are already in the data, new features will potentially be created to enhance the model's predictive power. Examples could include:

- Team batting average against left-handed pitchers
- Recent Win/Loss streaks for teams
- Pitching Matchups

Exploratory Data Analysis (EDA)

EDA is a crucial step in the understanding of data and identifying relationships between data points and the target variable to help determine game outcomes. Techniques like correlation analysis such as matrices and visualization will be used to explore these relationships.

Machine Learning Model Selection and Training

I will use a few different machine learning algorithms to determine the best possible method for determining game outcomes will be evaluated for their suitability in predicting game winners, these algorithms could include:

- Logistic Regression
- Random Forrest
- Decision Trees

The chosen model from these will be trained on a significant portion of the historical data, allowing it to learn from past trends and identify patterns that influence game outcomes.

Model Evaluation

The model's performance will be rigorously evaluated using metrics like:

- Accuracy: Amount of correctly predicted game winners
- Precision: The ratio of true positives vs. all predicted outcomes
- Recall: Ratio of true positive to actual wins
- F1-score: Mean of precision and recall

These metrics will provide insights into the model's effectiveness and identify areas for improvement.

Analysis

The analysis will focus on:

- **Model Performance:** Evaluating the model's ability to predict game winners accurately.
- **Feature Importance:** Identifying the features that contribute most significantly to the model's predictions.
- **Insights and Interpretations:** Understanding the factors that influence the model's predictions and their implications for game outcomes.

Assumptions:

- Historical data is accurate and reflects future trends to for game outcome prediction
- Chosen features are helpful and can significantly impact game outcomes.

Limitations:

- Data quality issues may affect model performance, i.e. not enough years of data to use.
- Unforeseen circumstances (e.g., injuries, weather conditions, external factors) can influence outcomes.
- The model is a tool and reference, not a guarantee, and should be used in conjunction with other analysis methods and not be a catch-all solution.

Challenges:

- Selecting the most impactful features for prediction and not using irrelevant data, wasting time and accuracy of the model.
- Ensuring data quality and accuracy, as well as addressing potential biases.
- Interpreting the model's predictions and results for human comprehension.

Future Uses/Additional Applications:

- Developing real-time game outcome prediction models to optimize future player performance.
- Analyzing player performance and team strategies.
- Optimizing fantasy baseball and sports gambling decisions.

Recommendations:

- I will attempt to use a combination of historical and real-time data for enhanced accuracy.
- I will continuously monitor and improve the model as new data becomes available over time.
- It makes sense to integrate the model with other analytical tools for comprehensive game analysis and not just use one model to answer all problems.

Implementation Plan:

- Secure necessary datasets and establish data collection procedures using accurate and reliable data.
- Develop and train the machine learning model using the methods listed above.
- Monitor model performance and adjust the model as needed.
- Transform the model into a user-friendly interface for accessibility and ease of access.

Ethical Assessment:

- **Data Bias:** It is important to be aware of potential biases in historical data and explore methods to mitigate their impact by ensuring data accuracy.
- **Transparency:** The model's limitations and assumptions will be clearly communicated to users.
- **Responsible Use:** The model is intended for entertainment and informational purposes, it is important to highlight responsible gambling practices based on the model results.

Questions from the audience:

- **How accurate is your model at predicting game winners?**
- **What are the most important factors that your model considers when predicting a winner?**
- **How does your model handle unexpected events, like injuries or sudden player slumps?**
- **Can your model be used for purposes beyond predicting winners, like predicting game scores or margins of victory?**
- **How often will you need to update the model with new data to maintain accuracy?**
- **Does your model consider ballpark factors, like dimensions or weather patterns, that might influence the game?**
- **How does your model compare to existing methods of baseball game prediction, like expert analysis or betting odds?**
- **Are there any ethical concerns surrounding the use of machine learning in predicting sporting events?**
- **Could this technology be used by teams to gain a competitive advantage?**
- **What are the limitations of your model, and what are some areas for future development?**

Conclusion:

This white paper proposes a machine learning approach to accurately predict Major League Baseball (MLB) game winners. By analyzing historical data and utilizing advanced algorithms, the model can potentially provide valuable insights for fans, analysts, and teams.

References:

- Moneyball: The Art of Winning an Unfair Game by Michael Lewis ([Book] Moneyball: The Art of Winning an Unfair Game)
- "Baseball and Machine Learning: A Data Science Approach to 2021 Hitting Projections" by S. Nakagawa (https://www.linkedin.com/pulse/baseball-machine-learning-how-helps-predict-future-mesk%C3%B3-md-phd)
- Resources from Kaggle competitions on MLB game prediction (https://www.kaggle.com/datasets/saurabhshahane/major-league-baseball-dataset)
- Baseball-Reference.com - Major League Statistics and Information. https://www.baseball-reference.com/
- Retrosheet. Retrieved from https://retrosheet.org/

Note: This is a draft white paper, and the specific methods and implementation plan may evolve as the project progresses.