Forecasting and Analyzing Baseball

Double Play Analytics

Overview

This project analyzes the factors associated with forecasting baseball game outcomes and provides key recommendations for success to industry leaders.

Business understanding

In the highly competitive and financially driven landscape of professional baseball, accurately predicting game outcomes can significantly impact team performance, fan engagement, and revenue generation. By leveraging historical game data our predictive modeling approach aims to provide teams, broadcasters, and stakeholders with actionable insights into future game outcomes. This predictive analytics tool will empower teams to optimize their strategies, such as player selection, pitching rotations, and in-game tactics, leading to improved win rates and overall performance. Ultimately, a reliable predictive model for baseball game outcomes has the potential to drive fan excitement, increase ticket sales, and attract valuable sponsorships, thus contributing to the long-term success and profitability of baseball organizations.

Data understanding and analysis

The dataset was taken from https://www.retrosheet.org/. They came from the Game logs dataset for 2010 through 2023. The data dictionary can be found on the webpage and in the supplemental materials of this repository.

Description of data

Data shape

After concatenating the thirteen years of data there were 32,484 regular season games (rows of data) and 179 columns/features.

Data manipulation

The target 'win' was created using the (a) score of the visiting team, (b) score of the home team, (c) name of the visiting team, and (d) name of the home team to calculate the winning and losing team. The rows that represented each game were then divided into two row per game, one for each team and the winning team of the pair was assigned a win=1 and the losing team of the pair was assigned an win=0. Each pair was given the same ID and kept together through the train_test_split.

X features varied by model and included: 'at_bats', 'hits', 'double', 'triple', 'home_run', 'sacrifice_hit', 'sacrifice_fly', 'hit_by_pitch', 'walk', 'intent_walk', 'strikeout', 'stolen_base', 'caught_stealing', 'grounded_into_double_plays', 'left_on_base', 'pitchers_used', 'wild_pitches', 'assists', 'errors', 'double_def'. All features used in X were numeric and standard scaled.

Data and analysis limitations

This analysis spanned the COVID-19 pandemic. The pandemic introduced unique challenges for analyzing baseball data, requiring careful consideration of context, data quality, and potential biases when interpreting and modeling data from 2020 and beyond. In 2020 there were 60 regular season games per team, less than half of the standard 162. The shortened and interrupted season impacted player salaries in 2020 and may have affected player performance metrics such as batting averages, earned run averages (ERAs), and fielding percentages. Some players may have performed better or worse than expected due to factors like altered training routines, health concerns, or changes in game dynamics. COVID-19 outbreaks and health protocols may have impacted player availability due to positive cases, contact tracing, or precautionary measures. This could affect team rosters, playing time, and lineup strategies, leading to shifts in player statistics and team performance. Many sports events in 2020 were held without spectators or with limited attendance to comply with public health guidelines. The absence of fans in stadiums and arenas could have affected player morale, home field advantage, and game dynamics, potentially influencing game outcomes and performance metrics.

Baseball has one of the richest data caches in sports, offering a plethora of statistics and metrics for analysis. Nonetheless, navigating this vast sea of information poses challenges, demanding a focused approach, organization, and an understanding of big data modeling techniques to extract meaningful insights from the wealth of available data.

Data modeling, evaluation, and interpretation

Data are split into a training set and test set for modeling. Games are paired so that features for the winning team and losing team of a game are not split between the training and test set to prevent data leakage. Features are standard scale standardized. No additional preprocessing is used. There are no categorical features, no missing data, and the target is 50:50 balanced.

The primary evaluation metric used for these models is accuracy. With a balanced dataset and equal interest in classification of wins as much as losses, accuracy is an appropriate evaluation metric.

- The first model is a Dummy Classifier. As expected using the most frequent strategy the accuracy score of this balanced data set is 50%.
- The second model is a simple Decision Tree Classifier using parameters for criterion, maximum depth, and the full set of X features listed above. The accuracy score for this model is 71%. The model identifies: hits, at_bats, home_run, intent_walk, pitchers_used, sacrifice_fly, and grounded_into_double_plays as the most important features.
- Next a Random Forest Classifier uses a grid search for the estimator, and maxium depth and features. The best parameters are used in the model with the full set of X features listed above. The accuracy score for this model is 78%. The model identifies: hits, at_bats, home_run, intent_walk, walk, double, and errors as the seven most important features.
- The first Logistic Regression model uses the full set of X features listed above. The accuracy score for this model is 80%.
- The second Logistic Regression model uses the subset of X features identified as the most important features from the Random Forest: hits, at_bats, home_run, intent_walk, walk, double, and errors. The accuracy score for this model is 76%.

- The last model is a Neural Network. This model is built using extensive support from ChatGPT. The full set of X features listed above are used in the model. The accuracy score for this model is 83%; the best performing model. The seven most important feature identified in order of importance are: at_bats, hits, left_on_base, walk, pitchers_used, grounded_into_double_plays, and intent_walk.
- Add probabilities and/or coefficient interpretation
- Interpretation/recommendation
- Add visualizations