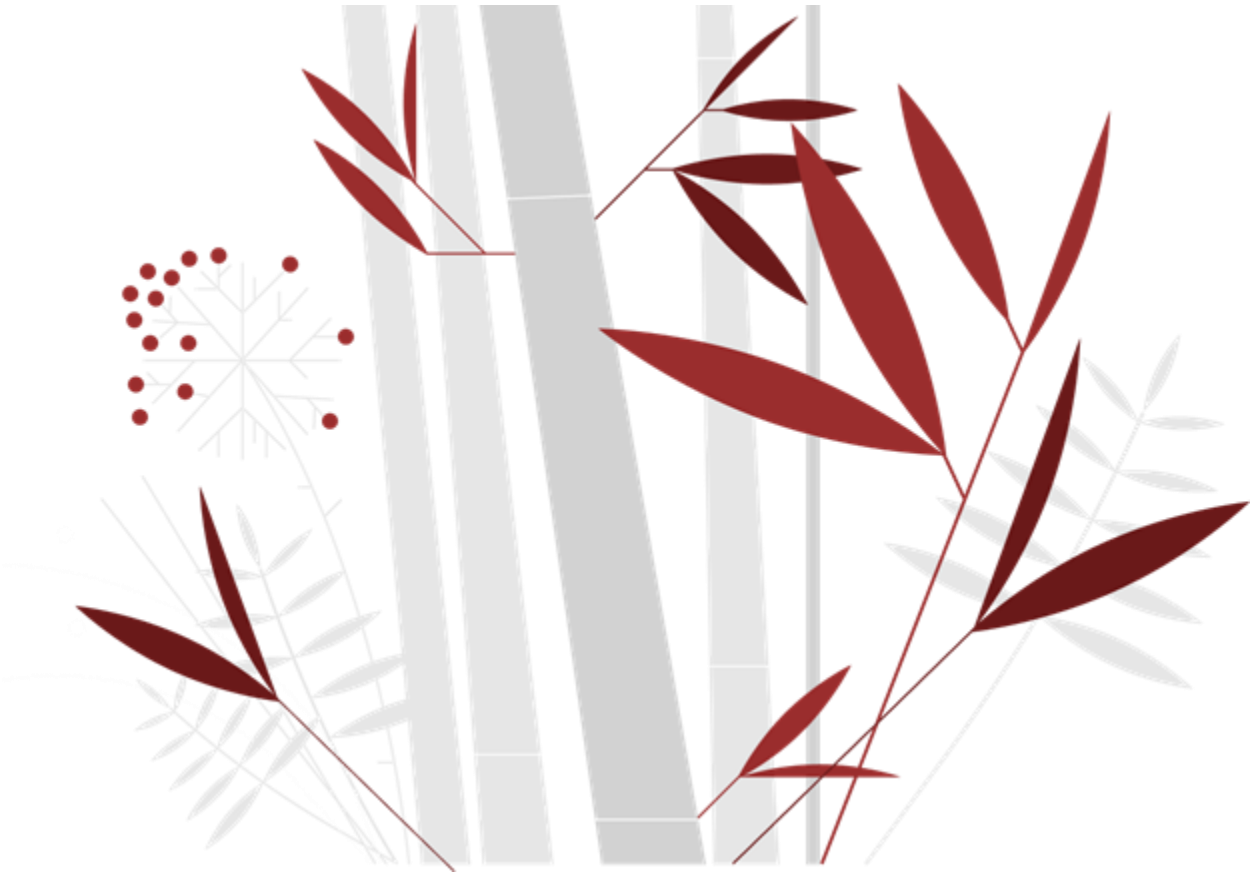
PREDICTING MAJOR LEAGUE BASEBALL (MLB) TEAM WINS: A MACHINE LEARNING APPROACH

Divya Muthyal

5/19/2024



Predicting Major League Baseball (MLB) Team Wins: A Machine Learning Approach

Introduction

In the competitive landscape of professional sports, teams constantly seek ways to gain a competitive edge and optimize their performance. Major League Baseball (MLB), as one of the most popular sports leagues globally, presents a fertile ground for exploring the application of machine learning algorithms in predicting team outcomes. This article delves into a machine learning project aimed at predicting the number of wins for MLB teams in the 2015 season based on historical performance data from the 2014 season.

Problem Definition

The primary objective of this project is to develop a robust machine learning model capable of accurately predicting the number of wins for MLB teams in the upcoming season. By leveraging historical performance data, including various metrics such as runs scored, hits, strikeouts, and more, the model aims to provide valuable insights into team performance trends and factors influencing success. The ability to predict team wins holds significant implications for team management, coaching staff, players, and fans, enabling informed decision-making and strategic planning.

Data Analysis

Overview of the Dataset

The dataset used in this project comprises comprehensive performance metrics for MLB teams during the 2014 season. It includes a wide range of variables, capturing various aspects of team performance, player statistics, and game outcomes.

Exploratory Data Analysis (EDA)

Before diving into model building, it is essential to conduct exploratory data analysis to gain insights into the dataset's structure, distributions, and relationships between variables. Visualizations such as histograms, scatter plots, and correlation matrices are employed to analyze the data comprehensively.

Key Findings from EDA

During the exploratory phase, several key insights emerge regarding the distribution of performance metrics, correlations between variables, and potential outliers. For example, certain metrics, such as runs scored and earned run average (ERA), exhibit strong correlations with the number of wins, indicating their significance in determining team success. Moreover, visualizations reveal patterns and trends in the data that inform subsequent preprocessing and modeling steps.

Data Splitting: Training and Testing Sets

In order to develop and evaluate machine learning models effectively, it is crucial to split the dataset into training and testing sets. The training set is utilized to train the model on historical data, allowing it to learn the underlying patterns and relationships between predictors and the target variable. On the other hand, the testing set serves as an independent dataset for evaluating the model's performance and assessing its ability to generalize to unseen data. By splitting the dataset into training and testing sets, we can mitigate the risk of overfitting and ensure that the model's performance metrics accurately reflect its predictive capabilities on new, unseen data. This approach enables us to build robust and reliable machine learning models that can effectively predict outcomes and inform decision-making in real-world scenarios

Building Machine Learning Models

Linear Regression Model

Linear regression is employed as one of the baseline models for predicting MLB team wins. The model's coefficients are estimated using the least squares method, and evaluation metrics such as mean squared error (MSE) and R-squared are utilized to assess its performance.

Random Forest Model

In addition to linear regression, a random forest model is implemented to predict MLB team wins. Random forests are an ensemble learning technique that combines multiple decision trees to improve predictive accuracy. Hyperparameter tuning using techniques such as grid search is performed to optimize the model's performance.

Model Evaluation

Linear Regression Model Performance

The performance of the linear regression model is evaluated on both the training and testing datasets using metrics such as mean squared error (MSE) and R-squared. A comparison of performance metrics highlights the model's effectiveness in capturing the relationship between predictors and the target variable (wins).

- **Training Set**:
  - Mean Squared Error: 2.9335
  - R-squared: 0.9698
- **Testing Set**:
  - Mean Squared Error: 32.1400
  - R-squared: 0.7557

Random Forest Model Performance

Similarly, the performance of the random forest model is assessed using evaluation metrics such as MSE and R-squared. By comparing the model's performance on training and testing datasets, insights are gained into its predictive capabilities and generalization ability.

- **Training Set**:
    - Mean Squared Error: 8.2532
    - R-squared: 0.9150
- **Testing Set**:
    - Mean Squared Error: 58.0367
    - R-squared: 0.5589

Best Model Selection

The linear regression model emerged as the preferred choice for predicting MLB team wins due to its simplicity, interpretability, and robust performance. Despite its straightforward nature, the model exhibited remarkable predictive accuracy, capturing the underlying relationships between various performance metrics and the number of wins. By estimating the coefficients of predictors using the least squares method, the model effectively quantifies the impact of each predictor on team success. Moreover, evaluation metrics such as mean squared error (MSE) and R-squared confirmed the model's ability to accurately predict team wins, both in training and testing datasets. Overall, the linear regression model offers a transparent and intuitive approach to understanding the factors influencing MLB team performance, making it a valuable tool for decision-making and strategic planning in professional sports

Visualizations

Histograms

Histograms are employed to visualize the distribution of key variables such as runs scored, hits, strikeouts, and more. These visualizations provide insights into the range and spread of each variable across MLB teams, aiding in understanding the underlying data distribution.

Box Plot

A boxplot of the numerical features provided a comprehensive overview of the distribution, central tendency, and variability of each feature, helping identify potential outliers and anomalies. By visualizing the spread of data using boxplots, we gained valuable insights into the range of values and any potential skewness present in the dataset

Pair Plots

A pair plot was utilized to visualize the pairwise relationships between numerical features, enabling us to identify correlations and patterns. This scatterplot matrix provided a quick and intuitive way to explore potential associations between variables, facilitating feature selection and guiding the model-

building process. Overall, these visualizations played a crucial role in understanding the dataset's characteristics, identifying trends, and informing subsequent data preprocessing and modeling steps

Scatter Plots

Scatter plots are utilized to examine the relationships between predictors and the target variable (wins). By plotting pairs of variables against each other, potential correlations and patterns in the data are identified, facilitating feature selection and model interpretation.

Correlation Matrices

Correlation matrices visually represent the correlation coefficients between pairs of variables. High correlation coefficients indicate strong linear relationships, while low coefficients suggest weaker associations. These matrices serve as a valuable tool for identifying multicollinearity and selecting relevant features for modeling.

Concluding Remarks

In conclusion, this article provides a comprehensive overview of a machine learning project aimed at predicting MLB team wins based on historical performance data. Through thorough data analysis, model building, and evaluation, valuable insights are gained into the factors influencing team success and the predictive capabilities of machine learning models. By leveraging advanced techniques and methodologies, teams can make informed decisions and optimize their strategies for the upcoming season, ultimately enhancing their chances of success in Major League Baseball.