Baseball Case Study

Introduction

In the world of professional sports, predicting team performance is a crucial aspect for coaches, managers, and fans alike. In this case study, we will explore how machine learning techniques can be applied to predict the number of wins for a team in the 2015 Major League Baseball (MLB) season based on various offensive and pitching statistics from the 2014 season.

Problem Statement

The goal of this project is to develop an algorithm that can accurately predict the number of wins for a given MLB team in the 2015 season. The dataset contains 16 input features, including offensive statistics such as runs scored, hits, and home runs, as well as pitching statistics like earned run average (ERA), shutouts, and saves. The output variable is the number of predicted wins (W) for each team.

Data Exploration and Preprocessing

The dataset consists of 30 records, each representing a team in the MLB. The input features are a mix of integer and float data types, while the output variable (Wins) is an integer. After a thorough analysis, we found no missing values in the dataset.To gain a better understanding of the data, we performed exploratory data analysis (EDA). The EDA revealed that the target variable (Wins) follows a normal distribution, indicating that linear regression models might be suitable for this problem.Next, we analyzed the correlation between the input features and the target variable. The analysis showed that some features, such as Runs (R), Home Runs (HR), Doubles (2B), Walks (BB), Shutouts (SHO), and Saves (SV), had a positive correlation with Wins. On the other hand, features like At Bats (AB), Hits (H), Triples (3B), Strikeouts (SO), Stolen Bases (SB), Complete Games (CG), and Errors (E) had a low correlation with Wins.Interestingly, Runs Allowed (RA), Earned Runs (ER), and Earned Run Average (ERA) had a high negative correlation with Wins. This suggests that teams with a lower number of runs allowed and a lower ERA tend to win more games.Further analysis revealed that RA, ER, and ERA were highly correlated with each other, indicating the presence of multicollinearity. To address this issue, we decided to remove one of these features (ER) from the dataset.Additionally, we found that features like AB and H had a strong correlation of 74%, which could also lead to multicollinearity. However, since these features had a low correlation with the target variable, we decided to keep them in the dataset for now.

Feature Engineering and Selection

Based on the insights gained from the EDA, we decided to keep the following features for further analysis:

* Runs (R)
* At Bats (AB)
* Hits (H)
* Doubles (2B)
* Triples (3B)
* Home Runs (HR)
* Walks (BB)
* Strikeouts (SO)
* Stolen Bases (SB)
* Runs Allowed (RA)
* Earned Run Average (ERA)
* Shutouts (SHO)
* Saves (SV)
* Complete Games (CG)
* Errors (E)

We also identified the presence of outliers in some features, such as R, ERA, SHO, SV, and E. To handle these outliers, we applied appropriate techniques like winsorization or removal, depending on the severity of the outliers and their impact on the overall distribution.After handling the outliers, we observed that the skewness in the feature distributions had been reduced, indicating that the data was more normally distributed.

Machine Learning Model Development

Since this is a regression problem, where the goal is to predict a continuous target variable (Wins), we decided to use various regression algorithms to build our predictive model.We started by splitting the dataset into training and testing sets, using a random state of 175 to ensure reproducibility. The training set was used to train the models, while the testing set was held out for final evaluation.Next, we performed 5-fold cross-validation on the training set to evaluate the performance of different regression models and tune their hyperparameters. The models we considered include:

1. Linear Regression
2. Random Forest Regressor
3. XGBoost Regressor

The cross-validation results helped us identify the best-performing model, which in this case was the Random Forest Regressor. The model achieved an average R-squared score of 0.76 across the 5 folds, indicating that it was able to explain 76% of the variance in the target variable (Wins).To further improve the model's performance, we applied feature transformation techniques to handle the remaining skewness in the feature distributions. This step helped in normalizing the data and improving the model's ability to capture the underlying patterns.Finally, we evaluated the best-performing model on the held-out testing set. The Random Forest Regressor achieved an R-squared score of 0.76 on the test set, confirming its ability to generalize well to unseen data.

Results and Insights

The key insights gained from this project are:

1. **Offensive statistics** like Runs (R), Home Runs (HR), and Walks (BB) are positively correlated with Wins, while **defensive statistics** like Runs Allowed (RA) and Earned Run Average (ERA) are negatively correlated.
2. **Handling multicollinearity** and **outliers**, as well as **removing skewness**, were crucial steps in improving the model's performance.
3. The use of **cross-validation** helped in selecting the best-performing model and tuning the hyperparameters.

These insights can be valuable for MLB teams and analysts in understanding the key factors that contribute to a team's success and developing more accurate predictive models for forecasting team performance.

Conclusion

In this case study, we successfully developed a machine learning model that can predict the number of wins for an MLB team based on various offensive and pitching statistics. By applying techniques like feature engineering, outlier handling, and cross-validation, we were able to build a robust Random Forest Regressor model that achieved an R-squared score of 0.76 on the test set.The insights gained from this project can be used by MLB teams to make informed decisions regarding player acquisitions, lineup optimization, and in-game strategy. Additionally, this approach can be extended to other sports leagues and domains where predicting team performance is crucial for success.