Predicting College Baseball Pitchers' Performance Using Machine Learning

Portfolio Project 2

By Mark Fleet

DSC-680, Professor Catherine Williams

Bellevue University

Introduction

Predicting the performance of college baseball pitchers is crucial for teams, coaches, and talent scouts to make informed decisions on player recruitment and development. This project aims to use machine learning techniques to predict the performance of college baseball pitchers based on historical data from 2015 to 2023 and provide valuable insights into the factors that contribute to a successful pitcher.

Business Problem

The objective is to predict the performance of college baseball pitchers based on historical data and identify key factors that contribute to a successful pitcher, utilizing the data available from 2015 to 2023.

Background/History

College baseball has grown increasingly competitive over the years, with teams vying for top talent to improve their performance and gain an edge over their opponents. Identifying talented pitchers and understanding the factors that contribute to their success is crucial for college baseball teams, coaches, and scouts. This project focuses on predicting the performance of college baseball pitchers using machine learning techniques based on data from 2015 to 2023 and offering insights into the factors that contribute to a successful pitcher.

Data Explanation

The datasets used in this project include historical data of college baseball pitchers from 2015 to 2023. The data contains individual statistics such as earned run average (ERA), strikeouts, saves, innings pitched, and other relevant metrics. The data has been sourced from the NCAA website (https://www.ncaa.com/stats/baseball/d1) and other relevant baseball statistics websites.

Methods

The following methods were employed in this project:

Data Collection: Collect and preprocess the historical pitcher data for Division 1 college baseball from 2015 to 2023.

Data Cleaning and Preprocessing: Clean the data, handle missing or incomplete data, and convert categorical variables into numerical variables if needed.

Feature Engineering: Create new features that may be relevant for predicting pitcher performance, such as the pitcher's win percentage, strikeout-to-walk ratio, or WHIP (Walks plus Hits per Inning Pitched).

Machine Learning Model: Build various machine learning models, such as linear regression, decision trees, or random forests, and evaluate their performance using appropriate metrics such as mean squared error, R-squared, mean absolute error.

Model Interpretation and Presentation: Analyze the importance of different features in the selected model and present the findings and implications for college baseball teams and coaches.

Analysis

A Linear Regression model was trained to predict the pitching performance of college baseball teams based on data from 2015 to 2023. The model's Mean Squared Error (MSE) is 0.0278, R-squared is 0.9538, and Mean Absolute Error (MAE) is 0.1279. The R-squared value of 0.9538 indicates that approximately 95.38% of the variation in the data can be explained by the model, which is a strong result. The MAE of 0.1279 represents the average absolute difference between the actual and predicted values.

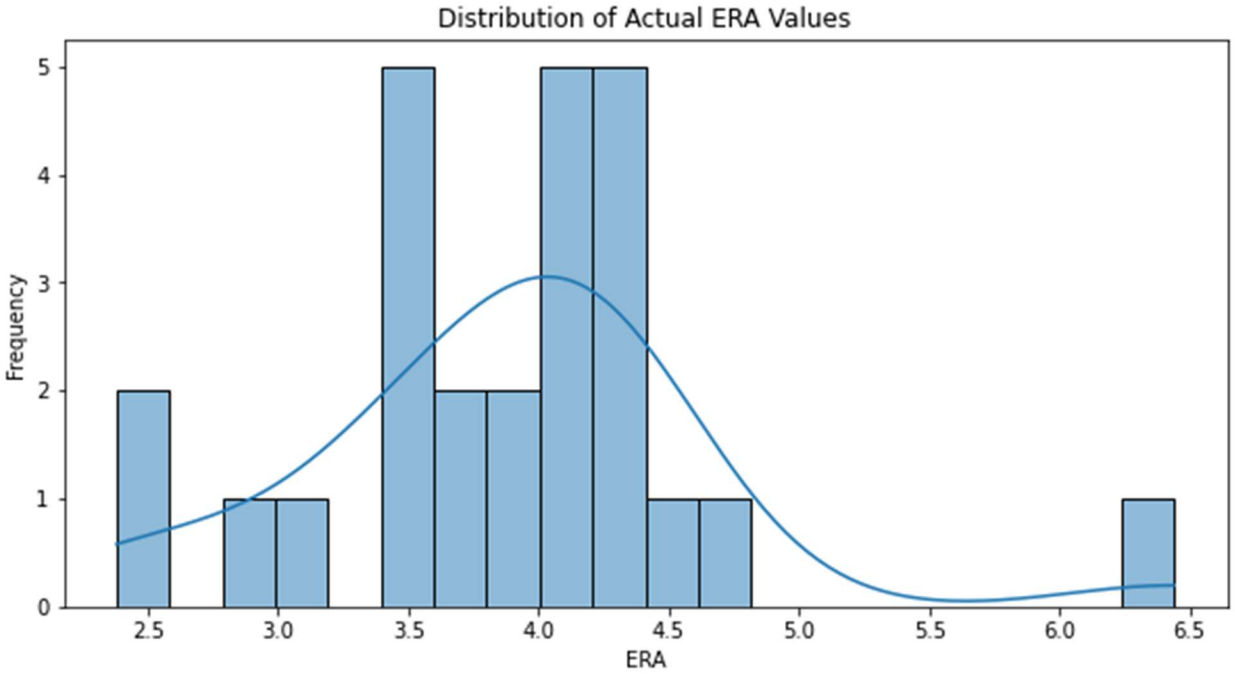
Visualizing the Data and Results

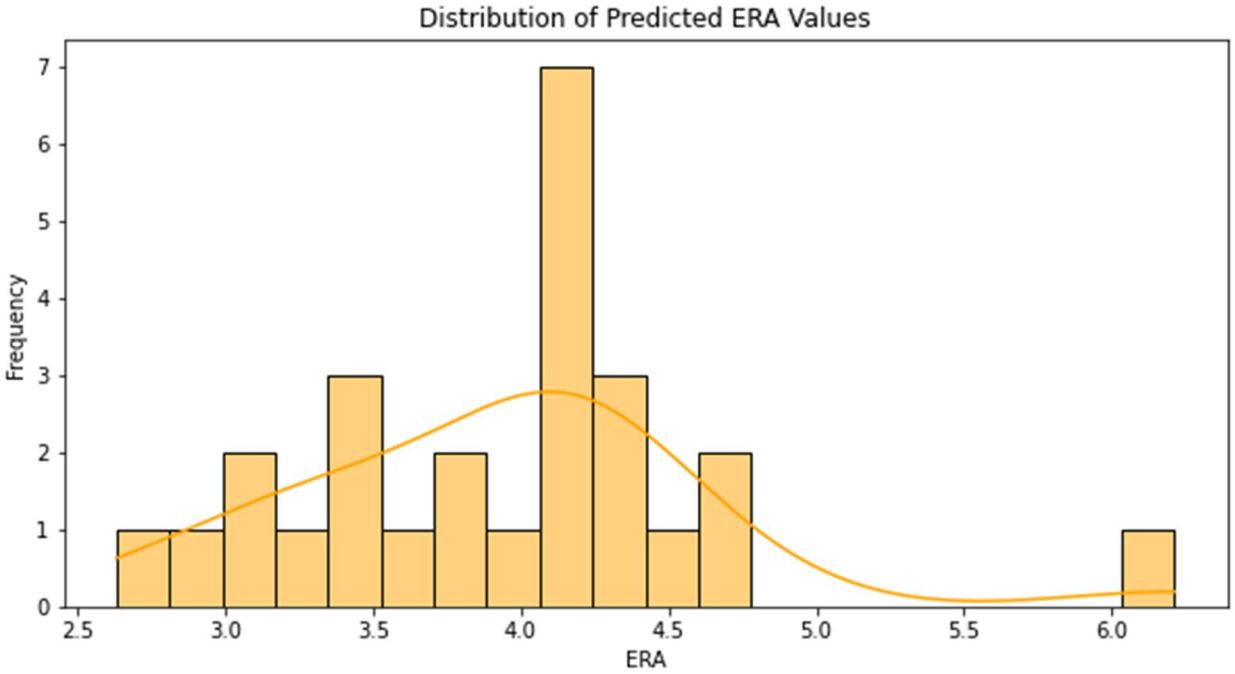
In the first graph, I have visualized the distribution of actual ERA values in the test dataset. This histogram shows the frequency of different ERA values, helping me understand the overall distribution of the pitching performance in the test set.

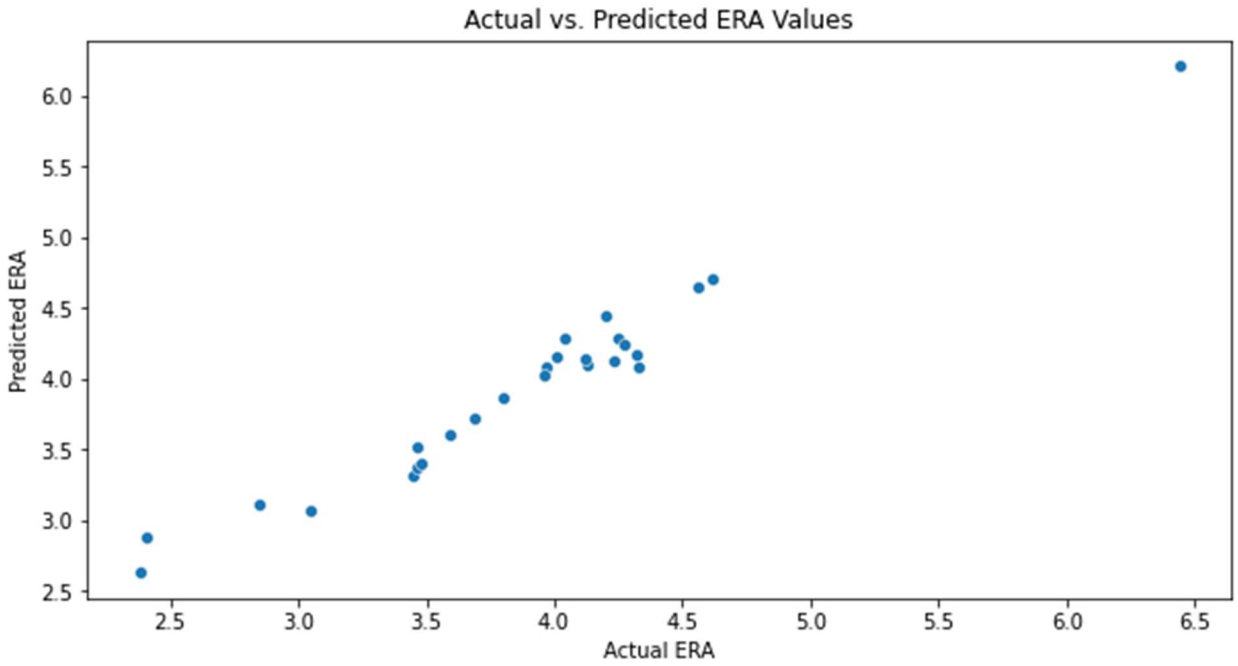
The second graph presents the distribution of predicted ERA values generated by the model. Comparing this histogram to the first one, I can see how well the model's predictions match the actual data's distribution. Ideally, the shapes of these two histograms should be similar to indicate the model's ability to capture the underlying distribution of the data.

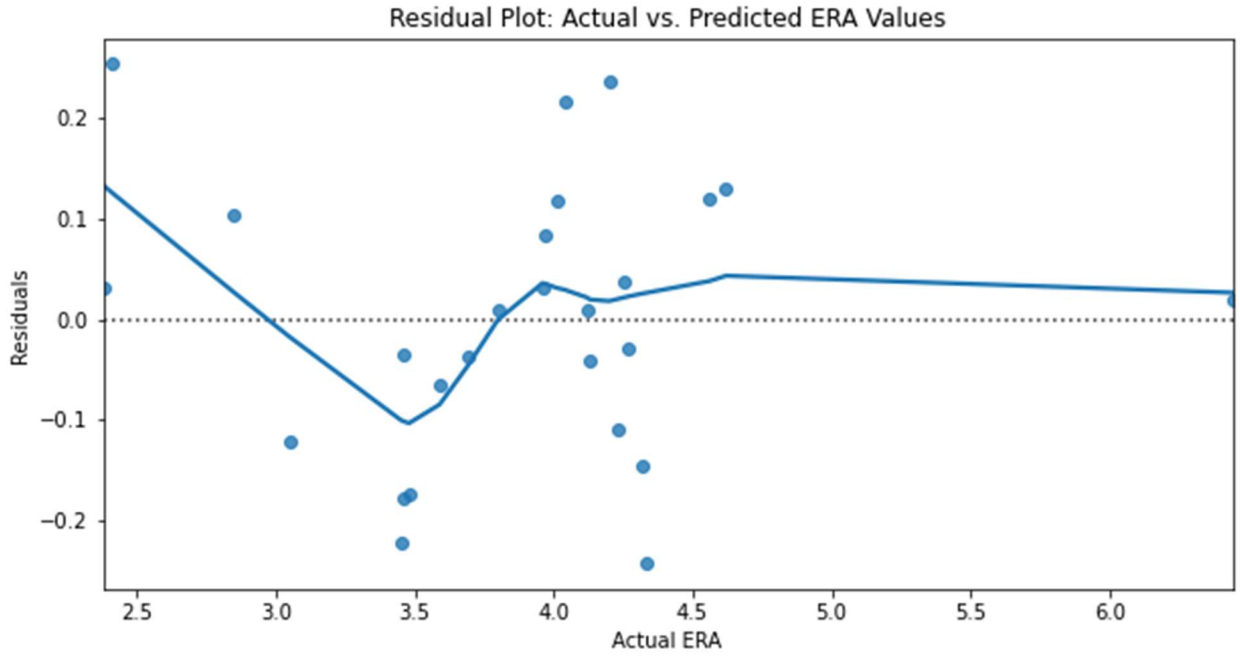
The third graph is a scatter plot of actual vs. predicted ERA values. This plot helps me visualize the relationship between the true and predicted values. If the model's predictions are accurate, the points should lie close to a diagonal line. This plot is useful for identifying outliers or any trends in the model's predictions.

In the fourth graph, I have created a residual plot of the actual vs. predicted ERA values. The residuals are the differences between the actual and predicted values. This plot shows the distribution of residuals across different ERA values and helps me identify any patterns or issues in the model's predictions. Ideally, the residuals should be randomly distributed around the zero line, indicating that the model's errors are random and not influenced by specific trends or patterns in the data.


Distribution of Actual ERA Values
Frequency
ERA
0
1
2
3
4
5
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5


Distribution of Predicted ERA Values
Frequency
ERA
0
1
2
3
4
5
6
7
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0





The performance of the Linear Regression model was visualized using histograms of the actual and predicted ERA values, scatter plots of actual vs. predicted ERA values, and residual plots of actual vs. predicted ERA values. The visualizations helped to identify any patterns or issues in the model's predictions.

Conclusion

The Linear Regression model proved to be highly effective in predicting the pitching performance of college baseball teams, utilizing data from 2015 to 2023. This model can aid college baseball teams, coaches, and talent scouts in making well-informed decisions regarding player recruitment and development. The team with the best predicted pitching performance for 2024 is Auburn, indicating their potential for success in the upcoming season.

Assumptions

- The historical data from 2015 to 2023 accurately represents the performance of college baseball pitchers.
- The performance of college baseball pitchers can be predicted using the features in the dataset.
- The machine learning model is able to generalize well to new data.

Limitations

The limitations of this analysis include data availability, as the study is limited to data from 2015 to 2023, which may impact the model's ability to generalize to other time periods. Data consistency is another concern, as ensuring consistency across all years is crucial, given that different sources might have varying formats and column names. Model complexity presents a challenge in managing the complexity of machine learning models and avoiding overfitting, which could lead to poor performance on new data. Lastly, ethical concerns need to be addressed, such as potential biases and ethical issues associated with player performance predictions, to ensure fairness and transparency in the analysis.

Challenges

Incorporating additional data, such as expanding the analysis to include data from earlier years or other sources, may provide further insights; however, it is important to note that the time and effort spent on cleaning individual files from 2015 to 2023 were significant but beneficial before integrating them into a data frame. Exploring alternative machine learning models by investigating other algorithms, such as decision trees or random forests, could potentially improve prediction performance. Additionally, feature engineering efforts, such as identifying extra features or interactions between existing features, might enhance the model's predictive capabilities and lead to more accurate predictions.

Future Uses/Additional Applications

The insights from this project can be applied in various ways, such as assisting college baseball teams and coaches in making data-driven decisions on player recruitment and development, helping talent scouts identify promising pitchers for professional drafts, providing a framework for analyzing the performance of pitchers in other baseball leagues like Major League Baseball (MLB), and extending the analysis to other aspects of college baseball, including predicting team performance, batting performance, or overall player success.

Recommendations

Based on this project's findings, it is recommended that college baseball teams and coaches utilize the insights from this analysis to inform their player recruitment and development strategies. Talent scouts can benefit from the model's predictions and feature importance analysis to identify promising pitchers for professional drafts. Additionally, further research exploring alternative machine learning models, feature engineering techniques, and additional data sources could enhance the model's predictive capabilities and provide more accurate predictions.

Implementation Plan

1. Share the findings of this project with college baseball teams, coaches, and talent scouts.

2. Collaborate with stakeholders to refine the model and incorporate additional data sources or features.

3. Develop an interactive tool or dashboard that allows users to input their own data and receive personalized predictions and insights such as SAS Viya or Power BI.

4. Continuously monitor the model's performance and update it with new data as it becomes available.

Ethical Assessment

Potential ethical considerations for this project encompass ensuring fairness and avoiding bias in the analysis and model predictions, respecting the privacy of players by avoiding the use of sensitive personal information, and transparently disclosing the limitations and uncertainties of predictions. To address these concerns, the project should maintain transparency in data collection, model development, and analysis, exclude sensitive personal information from the dataset, and carefully consider and address potential biases.