Evolution of an MLB Pitcher's Performance: A Five-Year Analysis of Corbin Patrick

STA 160 - Final Project

Johnson Tian, Yibo Ren - May. 2024

Introduction

Major League Baseball (MLB) is one of the most popular and competitive professional sports leagues in the United States, boasting a rich history and tradition. Since its inception in the late 19th century, MLB has undergone significant evolution. Currently, MLB comprises 30 teams divided into the American League (AL) and the National League (NL). Each season, teams compete in a series of games, culminating in the World Series, where the champions of the AL and NL face off. As the oldest of the four major professional sports leagues in the United States, MLB's charm has spread from America to the rest of the world.

Baseball is known for its high complexity and unpredictability. The outcome of a single play can vary dramatically, sometimes with no warning, adding to the excitement and suspense of the game. Unlike other team sports, the influence of star players in baseball is relatively limited. In baseball, every player has a fixed turn to bat, regardless of their superstar status, emphasizing the importance of teamwork and strategy. Baseball's unique nature lies in its unpredictability and the equal opportunity given to every player during the game.

MLB is not only the starting point of professional sports in the United States but also the origin of professional sports worldwide. Other sports leagues like the NFL, NBA, NHL, Premier League, and Bundesliga emerged decades after MLB. In the U.S., 85% of young people choose baseball as their first team sport at the age of four, making it an introductory sport for most American children. MLB runs from April to October, attracting over 80 million spectators to the ballparks each year. Recently, Shohei Ohtani signed a record-breaking $700 million contract, highlighting baseball's significant role in the global sports arena.

Research Subject and Objectives

This study focuses on MLB's renowned pitcher, Corbin Patrick. Patrick made his MLB debut in 2012 and has become known for his exceptional pitching skills. He has played for several teams throughout his career, but our study concentrates on his performance from 2017 to 2021. This period was selected because it encompasses a significant phase of his career during which he demonstrated notable performance and development.

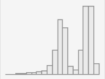
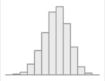
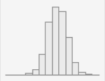
Why Corbin Patrick?

Corbin Patrick was chosen for this study due to his consistent and impactful presence in MLB during the specified period. His diverse pitching style and strategic release points make him an intriguing subject for in-depth analysis. By studying Patrick's performance, we aim to uncover insights that can be generalized to other pitchers and contribute to the broader understanding of pitching dynamics in baseball.

Data Collection

Due to baseball's high complexity, the use of big data analysis has become increasingly important. We collected five years of data on Corbin Patrick from 2017 to 2021 from Baseball Savant. These data include detailed pitch-by-pitch information, player statistics, and game outcomes. By leveraging this

rich dataset, we aim to explore relationships between pitching characteristics and game performance and to predict his future performance.

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
pitch_type [character]	1. (Empty string) 2. CH 3. CU 4. FF 5. SI 6. SL	13 (0.1%) 701 (5.1%) 466 (3.4%) 3170 (23.1%) 4070 (29.6%) 5316 (38.7%)		0 (0.0%)
release_speed [numeric]	Mean (sd) : 86.4 (6.3) min ≤ med ≤ max: 61.3 ≤ 88.9 ≤ 95.9 IQR (CV) : 10.5 (0.1)	328 distinct values		12 (0.1%)
release_pos_x [numeric]	Mean (sd) : 2.3 (0.2) min ≤ med ≤ max: 1.6 ≤ 2.3 ≤ 2.9 IQR (CV) : 0.2 (0.1)	118 distinct values		12 (0.1%)
release_pos_z [numeric]	Mean (sd) : 6.3 (0.1) min ≤ med ≤ max: 5.6 ≤ 6.3 ≤ 6.9 IQR (CV) : 0.2 (0)	112 distinct values		12 (0.1%)
events [character]	1. (Empty string) 2. field_out 3. strikeout 4. single 5. walk 6. double 7. home_run 8. grounded_into_double_play 9. force_out 10. field_error [11 others]	10144 (73.8%) 1368 (10.0%) 895 (6.5%) 531 (3.9%) 251 (1.8%) 179 (1.3%) 112 (0.8%) 66 (0.5%) 65 (0.5%) 24 (0.2%) 101 (0.7%)		0 (0.0%)
description [character]	1. ball 2. hit_into_play 3. foul 4. called_strike 5. swinging_strike 6. blocked_ball 7. swinging_strike_blocked 8. foul_tip 9. foul_bunt 10. hit_by_pitch [2 others]	4367 (31.8%) 2425 (17.7%) 2186 (15.9%) 2172 (15.8%) 1408 (10.3%) 622 (4.5%) 379 (2.8%) 98 (0.7%) 47 (0.3%) 15 (0.1%) 17 (0.1%)		0 (0.0%)

The data contain several important variables, Here's some visualization of important variables.

Data Cleaning

1. Filling Missing Values

In MLB season time series data, it is common to encounter discontinuous or missing data, such as during off-season periods or due to large gaps caused by the pandemic or player injuries. The methods for filling in missing values include forward fill and backward fill. These two methods use the previous or next valid data point to fill in the missing values, thus converting the time series data into uniform intervals.

Forward Fill: Fill in missing values using the previous valid data point.

Backward Fill: Fill in missing values using the next valid data point.

2. Interpolation

For continuous variables, interpolation methods are used to fill in missing data points. Since games do not occur every day or at regular intervals, using interpolation to connect the points helps in analysis. Linear interpolation is a common method that estimates missing values based on the linear relationship between known data points. Interpolation methods are suitable when there are few missing values and the data changes smoothly.Linear Interpolation: Estimate missing values by performing linear interpolation between known data points.

3. Using Frequency Conversion

The frequency of season time series data may be uneven. By resampling, data can be converted to a fixed frequency (such as daily, weekly, or monthly), making it easier to analyze and model.

Daily Resampling: Suitable for daily data, facilitating fine-grained analysis.

Weekly Resampling: Suitable for weekly data, helping to capture periodic changes.

Monthly Resampling: Suitable for monthly data, helping to observe long-term trends.

4. Deleting Missing Values

For some data with many missing values that are not important, deletion can be chosen. This simplifies the dataset and reduces interference with analysis and modeling. Before deleting missing values, ensure that these data do not significantly impact the analysis results.

Delete Missing Values: Remove rows or columns with many missing values that are not important.

Variable Processing

The variables in the dataset can be categorized into different types, including binary variables,

categorical variables, and numerical variables. Different types of variables need to be processed separately.

Binary Variables: Encode binary variables, such as converting Yes/No to 1/0.

Categorical Variables: Label categorical variables.

Numerical Variables: Standardize numerical variables for better performance in modeling.

5. Creating New Integrated Variables

By combining variables, new integrated variables can be created. These variables can capture more complex relationships and patterns, helping to improve model performance.

Creating Matrices by Combining Coordinates: Combine multiple coordinate variables into a matrix form to capture spatial relationships.

Multiplying Variables: Create interaction variables by multiplying two or more variables.

By following these data cleaning steps, discontinuous game dates and seasonal data can be effectively handled, ensuring data completeness and consistency, and laying a solid foundation for subsequent analysis and modeling.

Research Goals

This study has multiple objectives, each aimed at providing a comprehensive analysis of Corbin Patrick's performance. The primary goals are as follows:

1. Predicting Release Speed and Release Point:
    1. Objective: To develop models that can accurately predict the release speed and release point of Patrick's pitches.
    2. Significance: Understanding and predicting these factors can help in game strategy and training, allowing pitchers to optimize their performance.
2. Predicting Hitting Results:
    1. Objective: To predict the outcomes of at-bats based on various pitch characteristics.
    2. Significance: This can provide insights into how different pitches affect batter performance and help in developing effective pitching strategies.
3. Analyzing Player Performance Over Time:
    1. Objective: To conduct a time series analysis of Patrick's performance metrics across different seasons.
    2. Significance: This analysis helps in understanding trends and variations in player performance over time, which can inform decisions about training and career management.

Methodology

This study employs advanced machine learning techniques and statistical models to achieve the aforementioned objectives. The research data are sourced from publicly available MLB databases, which include detailed pitch-by-pitch information, player statistics, and game outcomes. We use these data to build various predictive models, including linear regression, random forests, and gradient boosting trees, to address different research goals.

1. Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \qquad i = 1,\ldots,n$$

Linear regression models the relationship between a dependent variable Y and multiple independent variables X. The coefficients β represent the impact of each independent variable on Y, and ε is the error term.

2. Random Forest Model

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Random forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of individual trees to improve predictive accuracy and control overfitting.

3. Support Vector Regression (SVR) Model

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)K(x_i, x) + b$$

Support Vector Machine (SVM) regression aims to find a function that deviates from the actual observed values by a value no greater than ϵ and is as flat as possible. The kernel function K transforms the input data into a higher-dimensional space.

4. Gradient Boosting (XGBoost) Model

$$\min\left(\frac{1}{n}\sum_{i}(\hat{y}_i - y_i)^2\right)$$

Gradient Boosting is a machine learning technique for regression and classification that builds a model in a stage-wise fashion from weak learners (typically decision trees). It minimizes the mean squared error of the predictions.

5. Decision Tree Model

$$I_E(i) = -\sum_{j=1}^{m} f(i,j)\log_2 f(i,j)$$

Decision trees use information entropy to determine the best splits at each node. The entropy I_E measures the impurity or uncertainty in the data, guiding the tree's structure.

6. Time Series (ARIMA)

$$X_t - \alpha_1 X_{t-1} - \ldots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

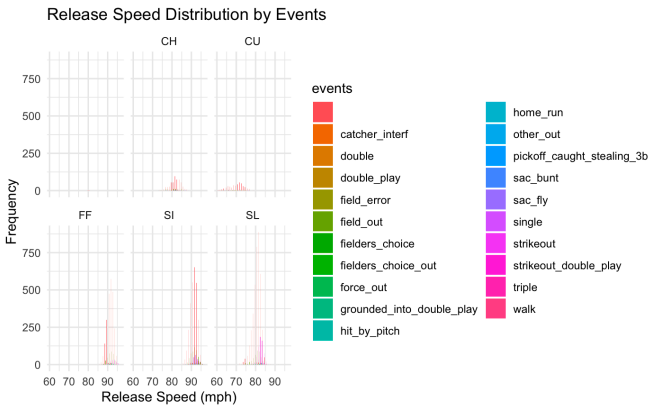
This ARIMA equation represents the relationship between the current value of the time series Xt and its past values and errors, incorporating autoregressive parameters α and moving average parameters θ.

7. Ridge Regression

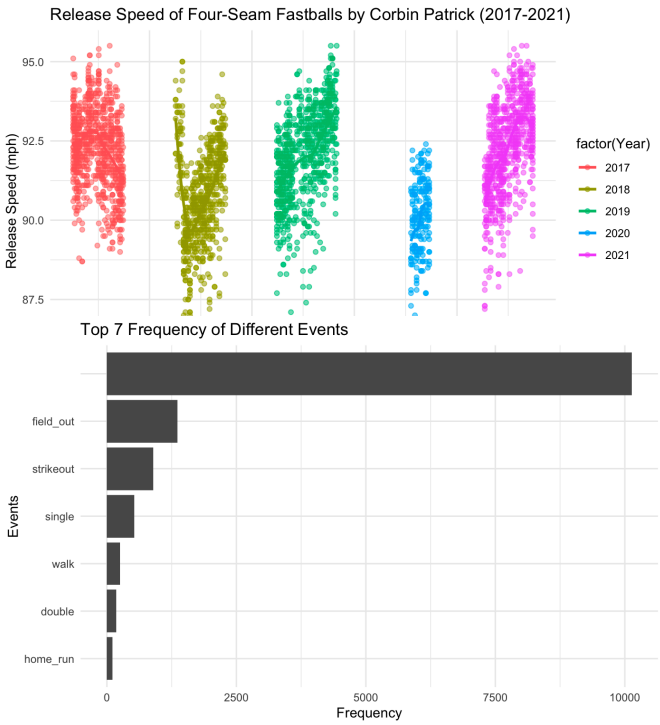
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Base Data Operation

"Release Speed Distribution by Events," shows the distribution of release speeds for different pitch



types (e.g., CH, CU, FF, SI, SL) and color-codes



them based on various events. It is evident that the distribution of release speeds varies across different pitch types, and different events are represented by different colors. This indicates that different pitch types and speeds can lead to different game outcomes.

"Top 7 Frequency of Different Events," displays the frequency of the top 7 events. The image shows that the field_out event has the highest frequency, followed by strikeout, single, walk, double, and home_run. This indicates that field_out is the most common event type among all the game events, which is significant for understanding the most frequent outcomes in the games.

As we can see, researching release speed and events is crucial for analyzing a pitcher's performance. First, release speed is a key indicator of pitching effectiveness. Higher pitch speeds generally make it more challenging for the batter to hit, thereby enhancing the pitcher's game performance. Analyzing the variation in release speeds over time helps understand the pitcher's performance stability and potential physical or technical changes over different seasons. Second, events directly reflect game outcomes. Events such as strikeout, single, and home_run provide direct insights into the result of each pitch. Analyzing the frequency and distribution of different events helps assess the pitcher's performance under various conditions. Understanding the distribution of events for different pitch types and speeds aids pitchers and coaches in formulating more effective game strategies. For instance, a particular pitch type at a certain speed might result in more strikeouts, making it a preferred choice in critical situations. By analyzing release speed and events, this study aims to gain an in-depth understanding of Corbin Patrick's pitching performance over different seasons and the impact of various pitch types and speeds on game outcomes. This not only helps evaluate the overall performance of the pitcher but also provides scientific insights for developing game strategies.

Prediction of Ball Releasing Speed

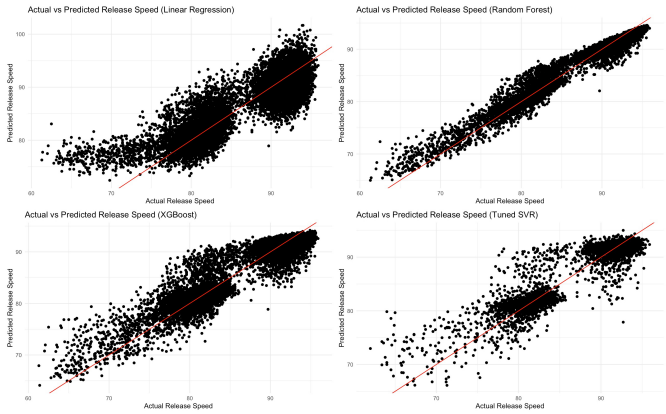
In this study, we selected the following variables to predict the pitching speed of Corbin Patrick: release_pos, pfx_interact, vx_vy_interact, ax_ay_az_interact, and vz0. These variables were chosen based on their potential impact on the pitch speed during the pitching process. Specifically, release_pos (release position) represents the hand position of the pitcher when releasing the ball, directly influencing the pitch speed. pfx_interact (pitch spin interaction) and vx_vy_interact (velocity interaction) describe the interactions of the ball's spin and velocity during the pitch. ax_ay_az_interact (acceleration interaction) combines the effects of the

ball's acceleration in three directions. vz0 (initial velocity) is the initial speed of the ball at release, which is crucial for predicting the final pitch speed.

We applied several machine learning methods to model and predict the pitch speed, including Linear Regression, Random Forest, Support Vector Regression (SVR), and Gradient Boosting (XGBoost). Each of these methods has unique characteristics that capture the complex relationships within the data from different angles. The specific results are as follows:

1. Linear Regression Model

The results indicate that all variables significantly impact pitch speed (all p-values are less than 0.001). The model's multiple R^2 is 0.7568, and the adjusted R^2 is 0.7567, indicating that the model



explains approximately 76% of the variance in the data. The residual standard error is 3.083, the Mean Squared Error (MSE) is 6.291449, and the Root Mean Squared Error (RMSE) is 2.508276.

2. Random Forest Model

The Random Forest model, constructed with 500 decision trees, explains 84.14% of the variance, with an MSE of 1.433921 and an RMSE of 1.197464. Feature importance analysis shows that vx_vy_interact and release_pos are the most important features.

3. Support Vector Regression (SVR) Model

The initial SVR model had an MSE of 6.291449, an RMSE of 2.508276, and an adjusted R^2 of 0.8410958. After tuning, the optimized SVR model's MSE was 6.22629, RMSE was 2.495254, and R^2 improved to 0.8425676.

4. Gradient Boosting (XGBoost) Model

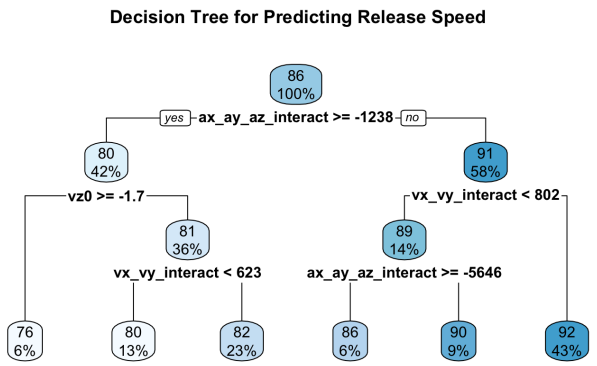
The XGBoost model demonstrated the best predictive performance, with an MSE of 4.568579 and an RMSE of 2.137424. The XGBoost model further confirmed the importance of ax_ay_az_interact and pfx_interact as key features.

These results highlight the differences in performance among various models for predicting pitch speed. The XGBoost model performed the best, followed by the Random Forest model. The SVR model showed decent performance, with improvements after parameter tuning. Although simple, the Linear Regression model also provided valuable explanatory power.

5. Decision Tree Model

Additionally, we chose to use a decision tree model to predict Corbin Patrick's release speed. The selection of the decision tree model was mainly based on its simplicity and strong visualization capabilities.

Through constructing the decision tree model, we discovered several key findings. In the process of decision tree splitting, ax_ay_az_interact (acceleration interaction variable) emerged as the first and most important splitting variable, indicating that acceleration interaction has a significant impact on release speed. Following this, vz0 (initial speed) and vx_vy_interact (velocity interaction variable)



were also identified as important variables. Although the decision tree model provided a clearer path of variable influence, its predictive performance was relatively low. This is because the decision tree model is prone to overfitting and has limitations in capturing complex nonlinear relationships. The results of the decision tree model can be explained through simple rules.

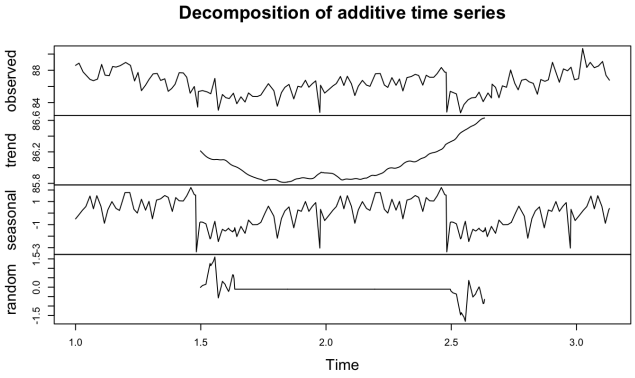
In summary, the decision tree model offers a simple and interpretable tool for understanding the impact of various variables on release speed. Despite its predictive performance being inferior to more complex models (such as random forest and XGBoost), the decision tree is highly valuable for initial analysis and identifying variable importance. It confirmed that ax_ay_az_interact is the most crucial variable for determining release speed.

6. Time Series (ARIMA)

Next, we attempted to fit the performance of the player Corbin Patrick using an ARIMA model. However, we faced the challenge that the player does not play every day due to injuries, off-seasons, and the pandemic, leading to gaps in the time series data. To address these issues, I chose to use linear interpolation to fill the gaps between games and defined each year's off-season to remove those periods.

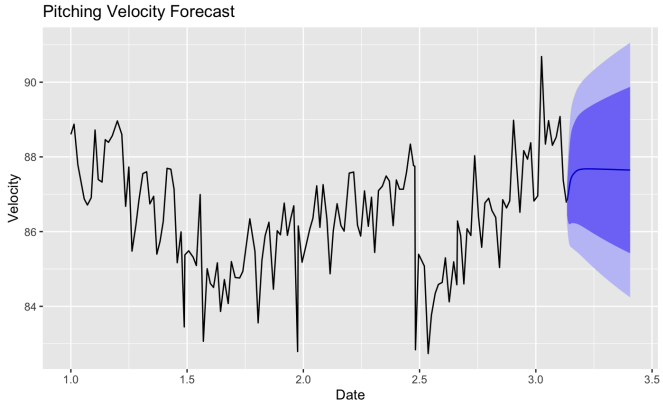
The specific methods are as follows: First, I used the na.approx function from the zoo package to linearly interpolate the missing data between games. Then, based on the defined off-season periods for each year (e.g., 2017-10-01 to 2018-02-28), I removed the data for those dates to ensure model accuracy. The steps included decomposing the time series, conducting ADF tests on the original data and the first-differenced data, fitting an ARIMA model, diagnosing the model residuals, and forecasting future release speeds.

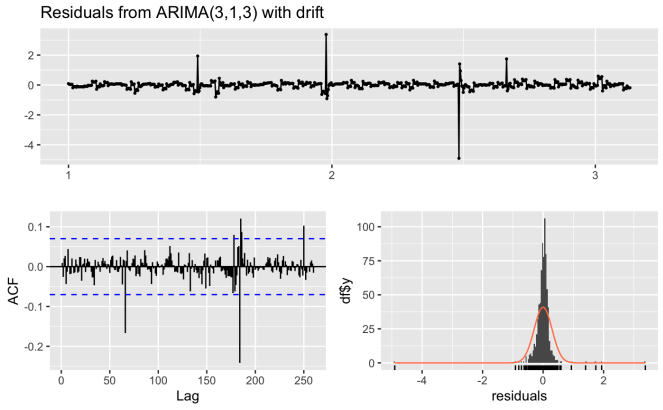
The residual diagnostics plot for the ARIMA(3,1,3) with drift model reveals three key aspects. Firstly, the residual time series plot shows that the residuals fluctuate around zero without any apparent pattern, indicating that the model has successfully captured the underlying structure of the time series data, leaving the residuals as white noise. Secondly, the autocorrelation function (ACF) plot of the residuals indicates that the autocorrelation



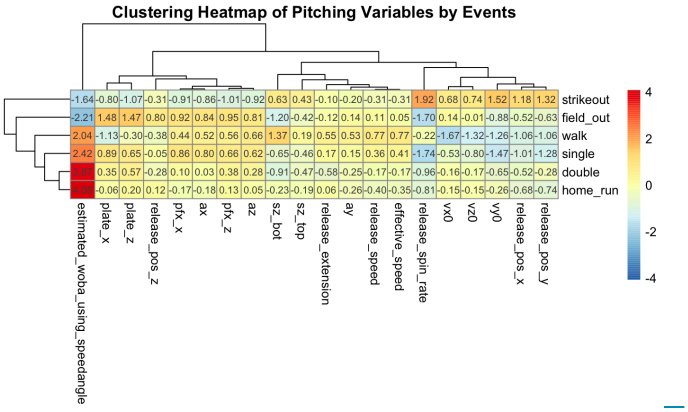
coefficients are mostly within the 95% confidence interval, suggesting that there is no significant autocorrelation in the residuals and that the model has adequately captured the dependencies in the time series. Lastly, the histogram of the residuals, which approximates a normal distribution centered around zero, supports the assumption of normally distributed errors. Overall, these results indicate that the ARIMA(3,1,3) model fits the data well and is suitable for forecasting future pitching speeds.

From the results, the time series decomposition graph showed the fluctuations in the original time series. The trend component exhibited





a downward and then upward trend, the seasonal component displayed periodic variations, and the random component showed some irregular fluctuations. The ADF test results indicated that both the original data and the first-differenced data were stationary, further confirming the predictability of the data. The ARIMA model's residual diagnostics showed no significant autocorrelation in the residuals, indicating a good model fit. The forecast graph for future release speeds indicated that future pitching speeds would fluctuate between 86 and 89, suggesting relatively stable performance in upcoming games.



Prediction of Event in Games

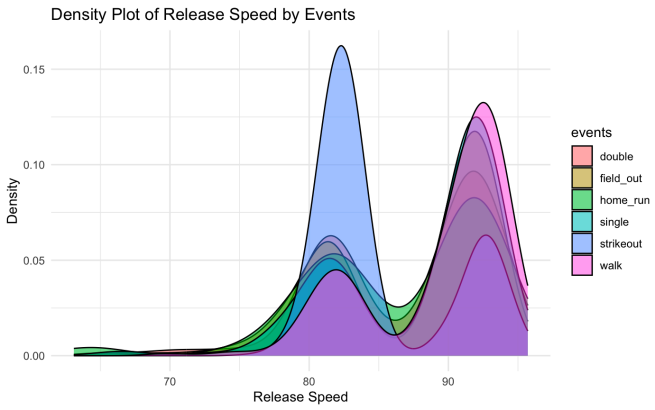
In this study, we selected multiple pitching variables including release_speed, release_pos_x, release_pos_z, pfx_x, pfx_z, plate_x, plate_z, vx0, vy0, vz0, ax, ay, az, sz_top, sz_bot, effective_speed, release_spin_rate, release_extension, release_pos_y, and estimated_woba_using_speedangle. These variables are considered to reflect pitching characteristics and outcomes, thereby influencing the game results.

1. Clustering Heatmap

The purpose of creating the clustering heatmap was to visualize the relationships and clustering patterns between different pitching variables and event types. We aimed to identify which variables were closely associated with specific events (such as strikeouts, walks, and home runs).

The estimated_woba_using_speedangle shows high values in home_run events, indicating its significant role in these events. The release_speed shows higher values in strikeout and field_out events, suggesting that higher pitch speeds might be associated with these events. There are significant differences in the performance of different event types across variables, providing a basis for

analyzing how pitching characteristics affect game



outcomes.

2. Density Plot of Pitching Speed

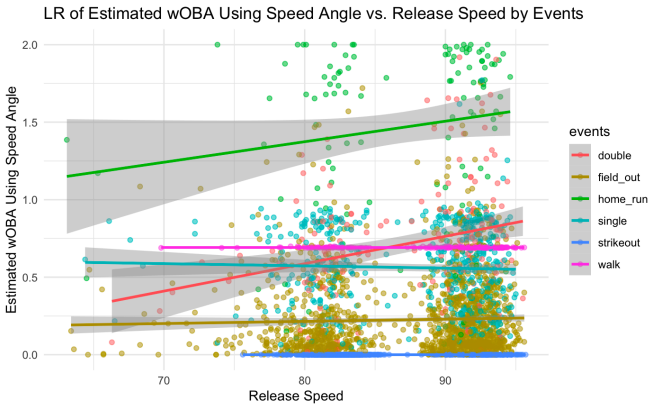
The density plot shows the distribution of pitching speeds for different events. By observing the density plot, we can understand the frequency of different events at different pitching speeds. Strikeout events have higher density at higher pitching speeds (close to 90 mph), indicating that higher pitch speeds might increase the likelihood of strikeouts. Walk events have higher density at lower pitching speeds, suggesting that lower pitch speeds might be associated with walks.

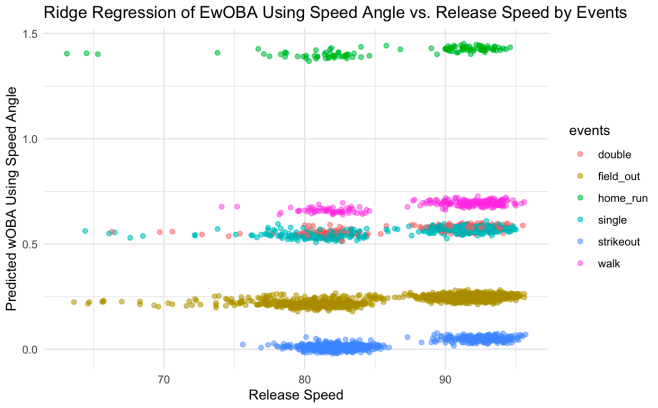
The density plot successfully displayed the distribution of pitching speeds for different events, confirming our hypothesis that pitching speed significantly affects the occurrence of specific events.

3. Regression Plot: Pitching Speed vs. Estimated wOBA

The linear regression plot shows the relationship between pitching speed and estimated wOBA using speed and angle for different event types. The regression line shows the linear relationship between pitching speed and estimated wOBA. The home_run events' regression line has a positive slope, indicating that the estimated wOBA increases with pitching speed. The field_out events' regression line is almost flat, suggesting that pitching speed has little effect on the estimated wOBA.

The ridge regression plot shows the relationship between pitching speed and estimated wOBA using ridge regression. The plot includes predicted values for different event types. The predicted wOBA for home_run events shows higher values at higher pitching speeds, indicating that faster pitch speeds might more easily lead to home runs. The predicted wOBA for field_out and strikeout events shows little variation across different pitching speeds, overall remaining at lower levels.





By performing ridge regression, we aimed to address these challenges and enhance the reliability of our analysis of the relationship between pitching speed and estimated weighted on-base average (wOBA) across different events. The results from ridge regression provide a more stable and

interpretable model, confirming the trends and relationships observed in the initial linear regression analysis.

Conclusion

From 2017 to 2021, Corbin Patrick experienced significant moments in his career, including outstanding performances and challenging periods. In 2017 and 2018, his pitching speed and performance reached their peak. However, in the following years, he faced several issues, including injuries, season interruptions, and the impact of the COVID-19 pandemic.

Performance Decline During the Pandemic: The COVID-19 pandemic in 2020 had a profound impact on the performance of many professional athletes. The interruptions and uncertainties of the season not only disrupted players' regular training and game routines but also put considerable stress on their mental and physical states. For Corbin Patrick, the long off-season during the pandemic likely led to inconsistencies in training and game conditions, partially explaining his performance decline in 2020. This period's performance fluctuations highlight the importance of a stable training and game environment for professional athletes.

Impact of Injuries: Injuries are an unavoidable issue for professional athletes. Throughout his career, Corbin Patrick experienced several significant injuries that directly affected his attendance and pitching speed. The recovery process from injuries is often long and uncertain; even after physical recovery, athletes may need time to return to optimal performance levels. These injuries not only affected his physical condition but also posed substantial psychological challenges.

Psychological Factors: In addition to physical challenges, psychological factors also play a crucial role in the performance of professional athletes. Facing long interruptions, recovery periods, and the pressure of returning to the field, a player's mental state can be significantly impacted. For Corbin Patrick, maintaining a positive attitude and strong mental resilience was essential for his ability to continue competing after multiple setbacks. His mental toughness helped him overcome many obstacles, allowing him to perform well even in the face of adversity.

In summary, Corbin Patrick's career not only demonstrates his technical and physical prowess but also showcases his resilience and indomitable spirit in the face of challenges. Through in-depth analysis of his pitching data and performance, we have gained insights into various factors influencing his game performance and derived important recommendations for future pitching strategies. These findings can help coaches and pitchers develop more effective strategies to improve game performance. Further research and analysis can deepen our understanding of the complex relationships between pitching characteristics and game outcomes, providing scientific evidence for the development of game strategies. Corbin Patrick's experiences offer valuable lessons and insights for future athletes in managing the ups and downs and challenges of their careers.

Results & Discussion

1. Prediction of Release Speed

   Linear Regression Model:

- The Linear Regression model indicated that all selected variables (release_pos, pfx_interact, vx_vy_interact, ax_ay_az_interact, and vz0) significantly impacted pitch speed. The model's multiple R^2 was 0.7568, and the adjusted R^2 was 0.7567, explaining approximately 76% of the variance in the data. This suggests a strong linear relationship between these variables and the release speed of Corbin Patrick's pitches.

Random Forest Model:

- The Random Forest model, constructed with 500 decision trees, explained 84.14% of the variance, with an MSE of 1.433921 and an RMSE of 1.197464. Feature importance analysis revealed that vx_vy_interact and release_pos were the most influential features. This model demonstrated superior performance in capturing complex interactions between variables.

Support Vector Regression (SVR) Model:

- The initial SVR model had an MSE of 6.291449, RMSE of 2.508276, and an adjusted R^2 of 0.8410958. After parameter tuning, the optimized SVR model's performance improved slightly, with an MSE of 6.22629, RMSE of 2.495254, and R^2 of 0.8425676. SVR provided a good balance between simplicity and predictive power.

Gradient Boosting (XGBoost) Model:

- The XGBoost model showed the best predictive performance, with an MSE of 4.568579 and an RMSE of 2.137424. It further highlighted the importance of ax_ay_az_interact and pfx_interact as key features. This model's ability to handle non-linear relationships and interactions made it the most effective for predicting pitch speed.

Decision Tree Model:

- Although the Decision Tree model provided clear insights into variable importance, its predictive performance was lower than the ensemble methods. The model identified ax_ay_az_interact as the most crucial variable, followed by vz0 and vx_vy_interact.

Discussion:

- The consistent identification of ax_ay_az_interact, release_pos, and vx_vy_interact across various models underscores their significance in determining pitch speed. The superior performance of ensemble methods like Random Forest and XGBoost highlights the complex, non-linear nature of pitching dynamics. These findings can inform training strategies to optimize pitch release speed.

2. Prediction of Game Events

Clustering Heatmap:

- The clustering heatmap revealed distinct patterns between pitching variables and event types. For instance, estimated_woba_using_speedangle showed high values in home_run events, while release_speed was higher in strikeout and field_out events. This visualization aids in understanding how different pitching characteristics influence game outcomes.

Density Plot of Pitching Speed:

- The density plot showed that higher pitching speeds were associated with strikeouts, while lower speeds were linked to walks. This confirms the hypothesis that pitching speed significantly affects the likelihood of specific events. This information

is crucial for pitchers and coaches in developing game strategies.

Regression Plot: Pitching Speed vs. Estimated wOBA:

- The regression analysis indicated a positive relationship between pitching speed and estimated wOBA for home runs, suggesting that faster pitches might lead to more home runs. Conversely, the relationship was flat for field_out events, implying little impact of pitch speed on these outcomes.

Ridge Regression:

- Ridge regression provided a more stable model, confirming the trends observed in linear regression. It demonstrated that higher pitch speeds increase the likelihood of home runs while having minimal impact on field_out and strikeout events.

Discussion:

- These findings highlight the trade-offs in pitching strategies. While higher speeds can increase strikeouts, they also pose a risk of more home runs. Understanding these dynamics can help pitchers balance their approach, optimizing for both effectiveness and minimizing risks.

3. Time Series Analysis of Performance

ARIMA Model:

- The ARIMA(3,1,3) model effectively captured the underlying structure of the time series data, with residuals fluctuating around zero and showing no significant autocorrelation. The forecast suggested relatively stable future pitching speeds between 86 and 89 mph.

Performance Trends:

- The time series decomposition revealed that Corbin Patrick's performance exhibited a downward trend followed by an upward trend over the five-year period. Seasonal variations and irregular fluctuations were also observed, likely due to injuries, off-seasons, and the COVID-19 pandemic.

Impact of External Factors:

- The COVID-19 pandemic in 2020 significantly disrupted regular training and game routines, leading to performance inconsistencies. Injuries further impacted his pitching speed and consistency. Psychological factors also played a role, affecting his mental state and resilience.

Discussion:

- Understanding the temporal dynamics of a pitcher's performance provides valuable insights for career management and training optimization. The analysis highlights the importance of maintaining a stable training environment and managing injuries effectively to ensure consistent performance.

Reference

1. Reddington, P. (2024, March 25). Washington Nationals news & notes: Patrick Corbin enters final year of 6-year deal in D.C. Federal Baseball. https://federalbaseball.com
2. Molis, J. (2023). Time series analysis of Major League Baseball organizations' fan attendance. Bridgewater State University.
3. Milano, V. (2009). Investigation of MLB data with multivariate statistics. Statistics Department, Cal Poly State University, San Luis Obispo, CA.

4. Major League Baseball. (n.d.). MLB history - people. http://mlb.mlb.com/mlb/history/mlb_history_people.jsp?story=com

5. Major League Baseball. (n.d.). About MLB. https://www.mlb.com/official-information/about-mlb

6. Cho, E. (2021, September 28). Opinion: Shohei Ohtani is the face and future of baseball. The Occidental News. https://theoccidentalnews.com/opinions/2021/09/28/opinion-shohei-ohtani-is-the-face-and-future-of-baseball/2903653

Appendix

STA_160FP_part1.pdf
STA_160FP_part2.pdf