Predicting and Forecasting MLB WAR

Stephen Fujiwara Mary McSweeney Soohyun Min Ryan Quach
Aashman Rastogi Sizhuo Tian

May 9, 2025

Abstract

Wins Above Replacement (WAR) is a key metric in Major League Baseball (MLB) used to evaluate a player's overall contribution to their team relative to a replacement-level player. This study focuses on predicting and forecasting pitcher WAR using detailed statistical data from past MLB seasons. By linking pitch-by-pitch and game-level statistics to recorded WAR values, we developed predictive models to estimate a pitcher's performance within a given season and project their future WAR. Using a panel data approach and regression modeling, we identify key performance indicators that correlate with WAR. The findings offer valuable insights for player evaluation, team decision-making, and talent scouting. Additionally, we discuss the limitations of our approach and suggest areas for further research in baseball analytics.

Contents

1 Introduction

In major league baseball, teams must frequently make decisions about rosters, contracts, and strategy. To make these decisions teams use data to decide a player's contribution to the team. Wins Above Replacement (WAR) is a statistic used in baseball that measures how many wins a player is worth than a replacement player at the same position (*Wins above replacement (War) — glossary* 2025). Pitchers can be evaluated in a number of ways, such as arm strength, and current pitch types (Eddie 2013). However, teams may opt to use WAR instead as it is comprehensive, using multiple performance factors to determine each player's value to the team.

The goal of this report is to develop predictive models that estimate a pitcher's WAR based on statistical performance indicators and historical trends. By analyzing a dataset containing pitch-by-pitch and game-level statistics from MLB seasons, we aim to identify key factors that influence WAR and build forecasting models for future seasons. This study leverages machine learning techniques and panel regression models to enhance the accuracy of WAR predictions. The insights gained from this analysis can aid front-office executives, coaches, and analysts in making data-driven decisions about pitcher performance and team composition.

2 Data Discussion

2.1 Data Overview

The dataset used in this study consists of multiple datasets merged together. The first dataset consists of detailed MLB pitcher statistics from 2010 to 2024. The MLB pitcher statistics were compiled from pitch-by-pitch and game-level data across multiple seasons. Each observation within this dataset represents a pitcher's game appearance, including features such as pitch type, velocity, strikeout rate, walks, and innings pitched. The next set of datasets contain WAR statistics for all MLB players active in a given year, covering the years 2010 to 2024. For each of these WAR datasets, one observation represents a single player, with features such as the number of plate appearances, innings pitched, pitching WAR, batting WAR, and total WAR.

2.2 Data Integration

To combine these datasets into one, single dataset, we systematically merged the pitcher statistics with the WAR data for each year. Specifically, we iterated through the WAR datasets year by year and extracted a subset of the pitcher statistics dataset that corresponded to that very same year, and merged the two together via an inner join according to the player name. This would ensure that only players with data available in both sources were the only ones retained. Finally, we concatenated these merged datasets into a single comprehensive dataset, integrating both game-by-game pitching statistics and WAR statistics across multiple years.

3 Methods

3.1 Feature Engineering

3.1.1 Feature Selection

For our analysis, we conducted a feature selection process, where we selected variables most intuitively aligned with our task. The chosen features include ageYrs (pitcher's age), seniorityYrs (years of experience), RHP (a binary indicator of right- or left-handedness), K (strikeouts), W (walks), HR (home runs allowed), outs (total outs recorded), and x (total events). We deemed that these variables capture key aspects of a pitcher's performance, while maintaining a reasonable level of granularity. The features excluded were either too granular, introducing excessive complexity, or deemed simply irrelevant.

3.1.2 Feature Transformation

To enhance the utility of our features, we performed a transformation on our game-level statistics. Specifically, we normalized the variables - K (strikeouts), W (walks), HR (home runs allowed), and outs by dividing by x, the total number of events. This allowed us to express these metrics as rates, as having raw counts were not significant pieces of information on their own. Following this, we aggregated the dataset by player and year (as a pair), computing the mean for all features, with the exception being RHP where we computed the mode. This aggregation would align us with our modeling approach, as our predictive models were designed to take summary statistics over a series of games for a given player as input.

3.2 Predicting Annual Pitcher WAR

3.2.1 Linear Regression

We used linear regression to predict annual Pitcher WAR (Wins Above Replacement) based on average statistics for each player across 2010–2024. Linear regression was chosen for its simplicity and interpretability, making it a suitable baseline model to understand the relationships between pitcher stats and WAR. Rather than a traditional train-test split, we utilized all available years' data for each player to predict WAR independently for each year. This approach allowed the model to leverage year-to-year variations in performance and capture the temporal dynamics of pitcher statistics. Predictor variables included average age, seniority, right-hand pitching, and average values of strikeouts, walks, home runs allowed, and outs. Aggregating these stats into yearly averages helped reduce noise and provided cleaner input for the model. Linear regression offered clear interpretability, with coefficients directly indicating how each stat influences WAR. Its efficiency was also beneficial for handling the large dataset. However, limitations included the assumption of linearity, sensitivity to outliers, and potential multicollinearity, which could affect the reliability of the results. Using averaged statistics minimized variability from game-to-game data, allowing the model to focus on consistent aspects of performance. While effective as a baseline, the model's limitations suggest that exploring non-linear models could further improve accuracy.

3.2.2 Decision Tree

To predict a pitcher's WAR based on our specific set of features, we also trained a decision tree model. Decision trees are well-suited for this task due to their ability to capture nonlinear relationships (as well as linear) and interactions between variables without any explicit set of assumptions. Even more importantly, they provide an interpretable modeling structure that can allow us to assess which variables contribute the most to the prediction of WAR.

Decision Trees offer advantages and disadvantages. They are flexible, capable of modeling complex interactions between features (as mentioned before), and do not require pre-processing such as feature scaling. Furthermore, they utilize a very interpretable decision making framework. However, a significant drawback is their tendency to overfit, particularly when the tree is very deep, and they also are sensitive to small variations in the data, producing high variance.

For training the decision tree model, we performed a simple train-test-split, allocating 70% of the dataset for model fitting, and the remaining portion for evaluation. To optimize the model, we conducted cross-validation, searching for the best parameters to balance complexity and generalization before applying it to the test set.

3.2.3 XGBoost

The next model we trained for this task was an XGBoost model. XGBoost is a gradient-boosting algorithm that builds a collection of decision trees sequentially, where predictions are refined by creating new trees that correct errors from previous trees. This allows it to model complex, nonlinear relationships much like a single decision tree, but potentially to a much higher degree.

XGBoost has several advantages. It tends to outperform simpler models in accuracy terms, and can capture complex feature interactions without any manual specification. It is also designed for computational efficiency, making it suited for even very large datasets. Another important advantage is that it incorporates regularization, including L1 regularization and L2 regularization, to combat overfitting. A disadvantage

however is that boosting methods, including XGBoost, are more computationally intensive than simpler models. Furthermore, while decision trees provide clear interpretability, the ensemble nature of XGBoost makes it less interpretable.

As for splitting the data for training, we used the same train-test-split as was used with the decision tree model, where 70% of the dataset was allocated for training, with the rest for model evaluation. Hyperparameter tuning was then conducted through cross-validation with the training set to optimize model performance before a final evaluation on the test set.

3.3 Forecasting Next-Year Annual Pitcher WAR

For the second task, we decided to utilize two different models: linear regression and panel regression. To do so, we added a column to the dataset that measures the next year WAR of a given player and year combination, which are recorded as rows. Given that the next year WAR of a player can be considered as linearly dependent on the current year WAR and other predictors in the dataset, we saw linear regression as a good choice for forecasting pitcher WAR in the following year. Such a decision makes sense because of the ease of interpretability and simplicity associated with the model, which should not be surprising in light of the fact that linear regression was also used for the first task. However, given that one cannot assume the errors are uncorrelated, a panel regression approach may be more optimal for predicting pitcher WAR of the following year. Hence, we decided to utilize panel regression using the plm() package, which specializes in producing and analyzing these kinds of models. More precisely, we decided to use a model that splits the data by pitcher and year, and one that treats the effects as fixed. Note that both of these models used the same predictors as the models of Task 1 with the single addition of current year WAR.

4 Results

4.1 Results of Predicting Annual War

4.1.1 Linear Regression

The linear regression model achieved a Mean Squared Error (MSE) of 1.010338, indicating moderate accuracy but also leaving room for improvement. An MSE above 1 suggests that some predictions significantly deviated from actual WAR values, likely due to the model's linear assumption failing to capture more complex relationships among the variables. The sensitivity to outliers might have also elevated the error, as extreme WAR values could disproportionately influence the results.

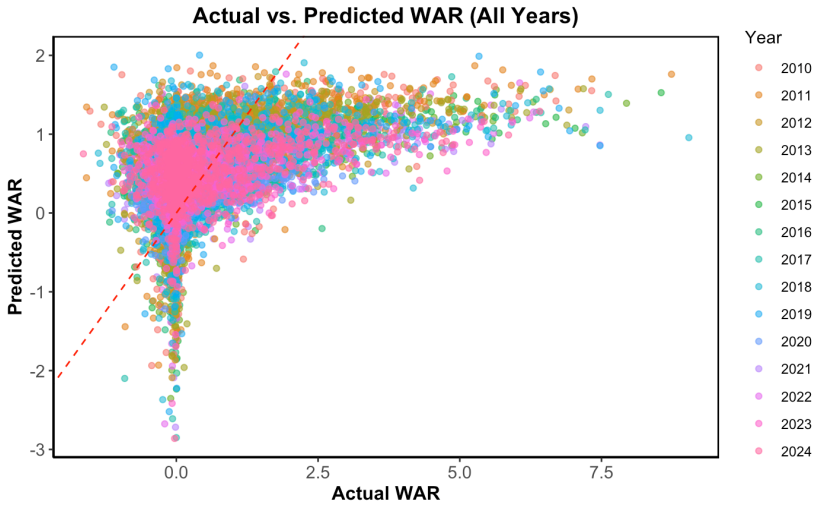

Figure 1: Scatterplot of Actual vs Predicted WAR for Linear Regression

The "Actual vs. Predicted WAR" plot reveals that while the model's predictions align moderately well for average WAR values, they significantly deviate for higher or lower WAR ranges. The dispersion of data points around the red dashed line suggests that the linear model struggles to capture non-linear patterns, leading to inaccuracies in predictions. The clustering of points near zero indicates that the model performs relatively better for players with average WAR but less so for others.

The "MSE Trend Over Years" plot further highlights the variability in the model's performance across different seasons. The fluctuations in MSE suggest that certain years posed greater challenges for the model, possibly due to anomalies or shifts in player performance trends. Notably, the sharp drop in MSE around 2020 implies an unusual year where the model fit the data more effectively, potentially due to less variability in player stats.

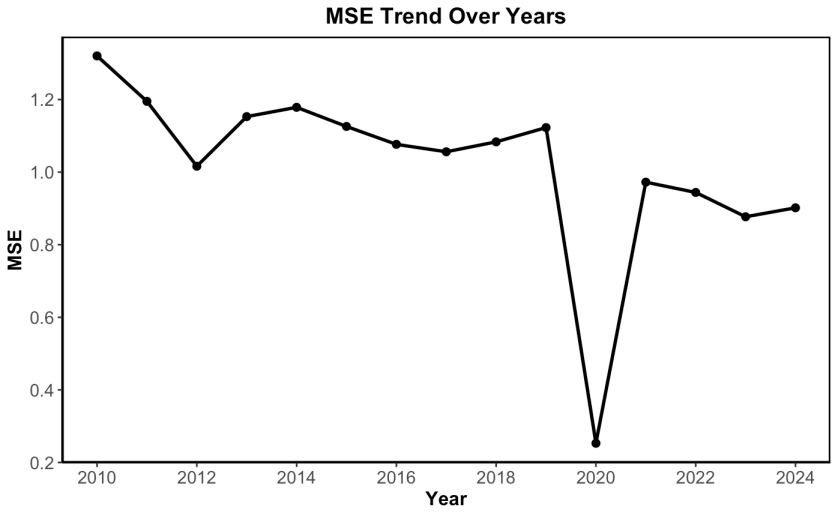

Figure 2: Line Graph of Linear Regression MSE Over Time

The findings imply that while linear regression offers a straightforward and interpretable baseline, it may not fully account for non-linear influences and interactions between pitcher stats. The use of average statistics helped reduce noise, but the model's limitations suggest that exploring more sophisticated techniques could enhance accuracy. In particular, non-linear models such as Decision Trees or XGBoost might better capture the complexities of pitcher performance, providing a promising direction for further analysis.

4.1.2 Decision Tree

Tuning the hyperparameters through cross-validation, we selected the best-performing decision tree model for evaluation on the test set. The best decision tree model yielded a Mean Squared Error (MSE) of 0.833, providing more satisfactory results than that of linear regression. However, one thing to note from what can be seen from the Figure, the model systematically underestimates WAR for some odd reason.

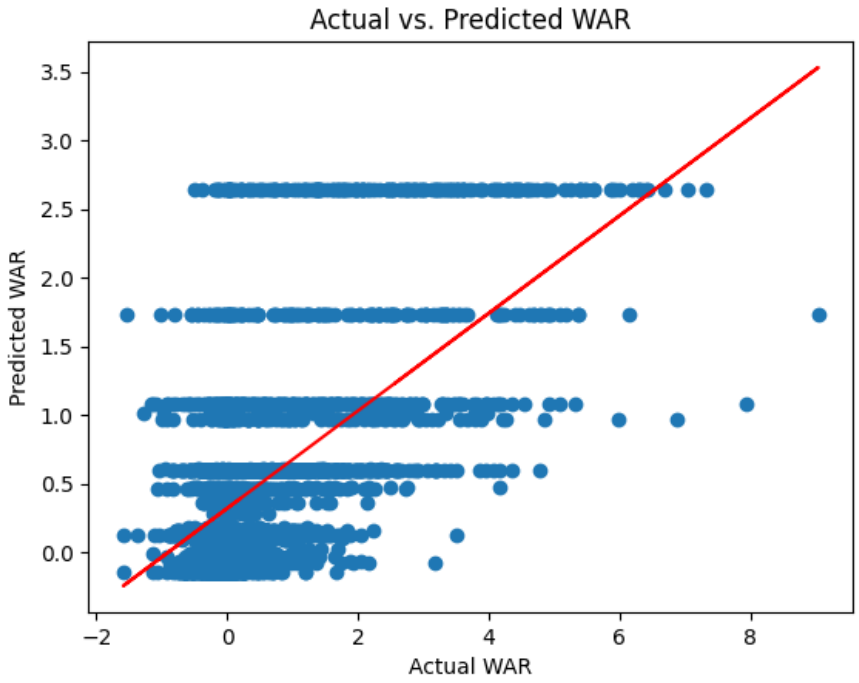

Figure 3: Scatterplot of Actual vs Predicted WAR for Decision Tree

To gain further insight into this model's decision-making process, we examined the feature importances of the model, which helps quantify the contribution of each feature towards making predictions. As the Figure indicates, avg_outs was the most influential feature, receiving the highest importance score of 0.425. In stark contrast, avg_ageYrs and RHP had importance scores of 0, indicating that the model simply did not use these features at all for prediction. The rest of the features moderately contributed to decision making.

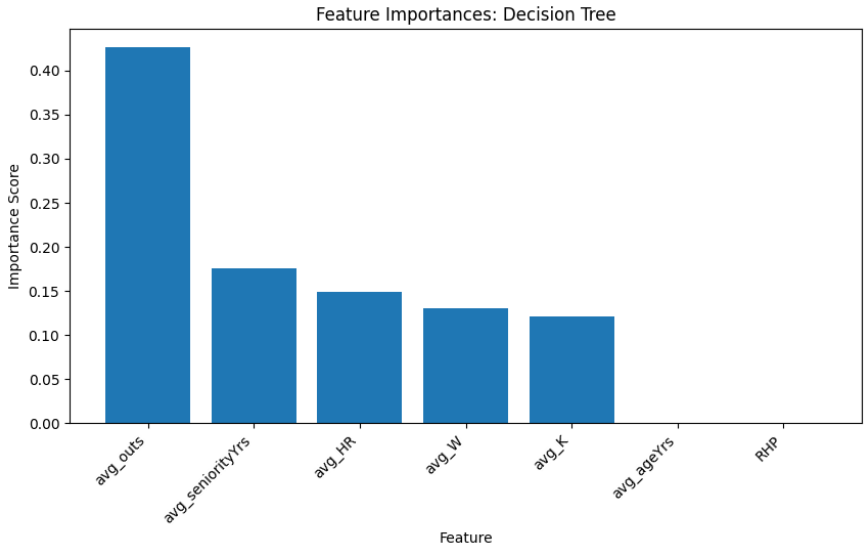

Figure 4: Barplot of Feature Importance for Decision Tree

4.1.3 XGBoost

Using cross-validation, we fine-tuned the hyperparameters to identify a best XGBoost model. This model was then evaluated on the test set, where it provided us with a Mean Squared Error (MSE) of 0.711, indicating an improvement over the decision tree model. Similar to the decision tree model, it seems to systematically underestimate WAR, however to a slightly less extreme degree than the decision tree model.

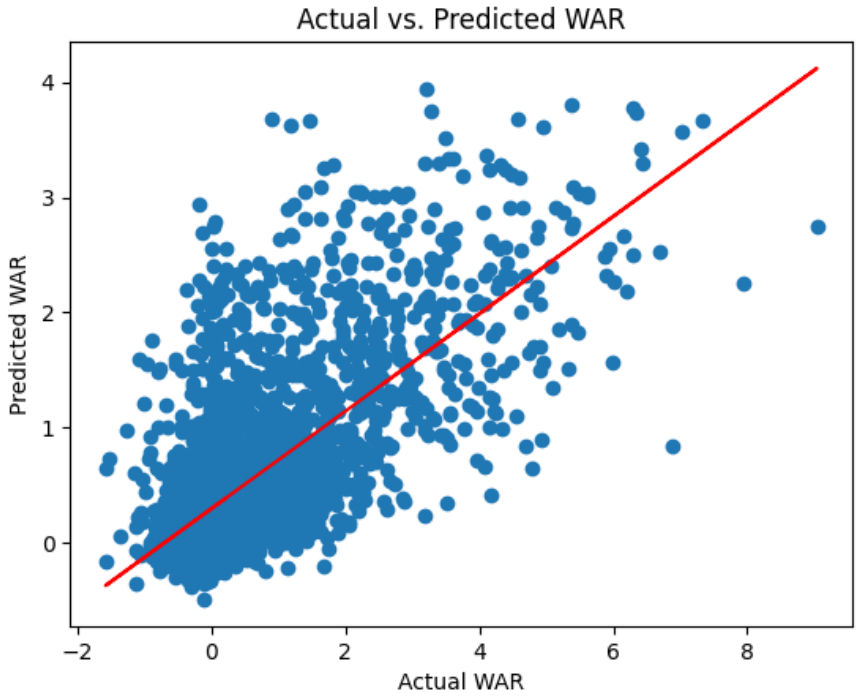

Figure 5: Scatterplot of Actual vs Predicted WAR for Decision Tree

We also analyzed the feature importances to understand the key features relevant to the decision-making process of the model. Unlike the decision tree model, which had a single dominant feature, XGBoost distributed the importance more evenly across the feature space. While avg_outs still remained as the feature with the highest importance score, it was by a very smaller margin from the feature in second place, avg_HR. In addition, avg_ageYrs, and RHP - which the decision tree model completely ignored - had small, but non-zero importance scores, suggesting that XGBoost had incorporated these features rather than choosing to discard them entirely, highlighting how boosting methods can capture slightly more nuanced relationships in data better than a single decision tree model.

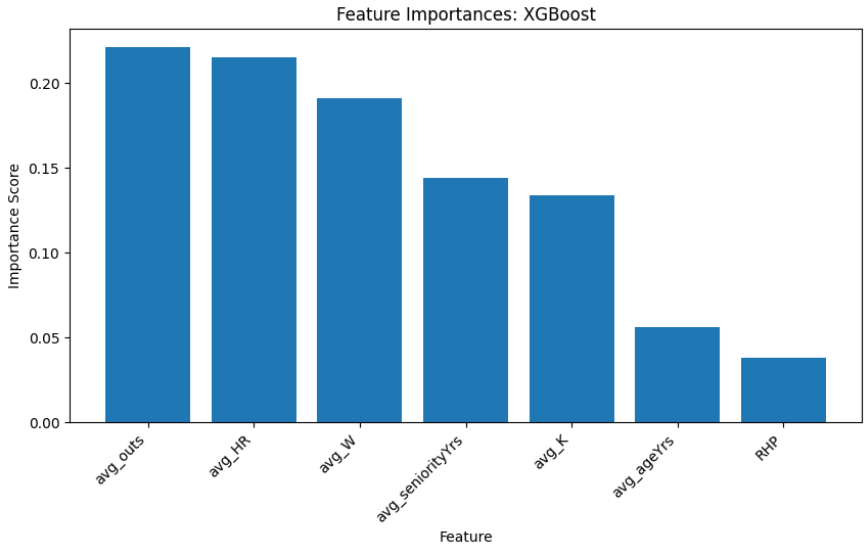

Figure 6: Barplot of Feature Importance for XGBoost

4.2 Results of Forecasting Next-Year Annual WAR

For predicting next year's WAR based on the previous WAR, which was the second task requested for this project, we employed a multiple linear regression model and a panel regression model. Having two models would ensure that we would not be limited to a single option when it comes to attempting the prediction of next year's WAR for each player.

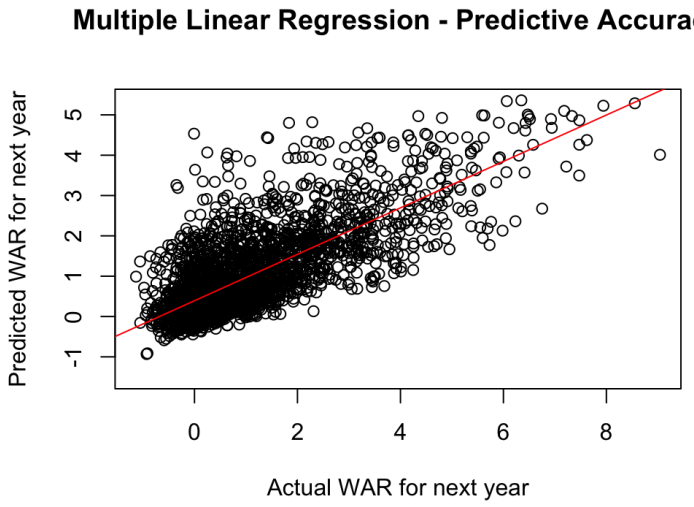

Figure 7: Scatterplot of Actual vs Predicted WAR for Linear Regression

For the linear regression model, it achieved an MSE of about 0.794 units (or an RMSE of about 0.891 units). Given that the standard deviation of the WAR variable is about 1 unit in the dataset, which makes sense given that WAR is standardized, the model is therefore quite accurate when it comes to predicting WAR of a given player for the next year. Such a high predictive ability was unanticipated because we decided to utilize fewer variables than we had initially expected to use, so the fact that this simple model did quite well was a pleasant surprise, to say the least.

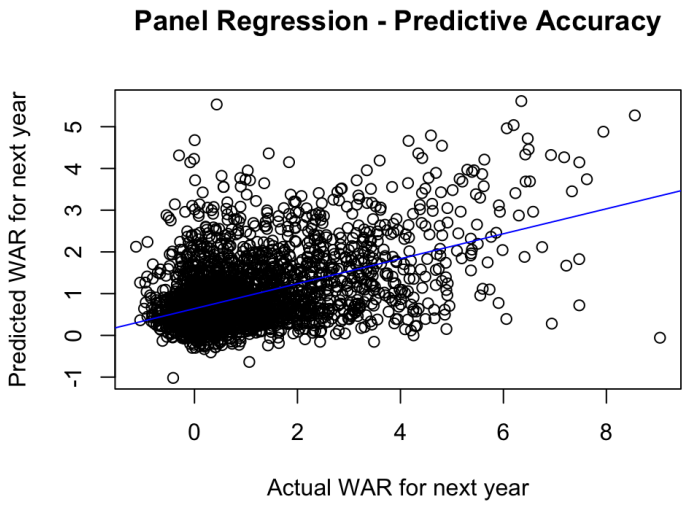

Figure 8: Scatterplot of Actual vs Predicted WAR for Panel Regression

As for the random-effects panel regression model, it achieved an MSE of about 1.179 units (or an RMSE of about 1.085 units). Initially, we expected the panel regression model to do much better than the relatively simple linear regression approach, so this drop in accuracy is surprising. However, the reason for this decline may be that panel regression is either simply too complex or overfitted for the data, although this increase in predictive error is somewhat strange and unexpected considering that we are working with panel data. Of course, it must be noted that based on the player-specific visualizations, the panel model appears to be better at predicting the direction of WAR changes over time.

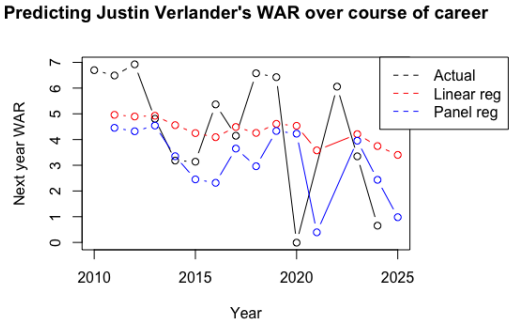

Figure 9: Line Graph of Predicted WAR for Justin Verlander

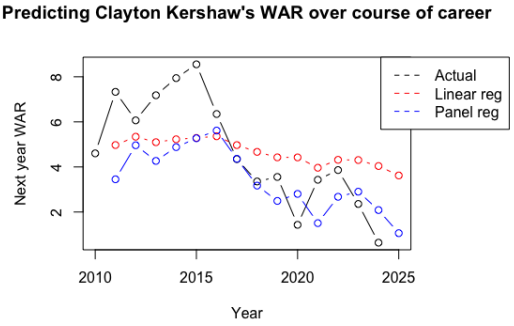

Figure 10: Line Graph of Predicted WAR for Clayton Kershaw

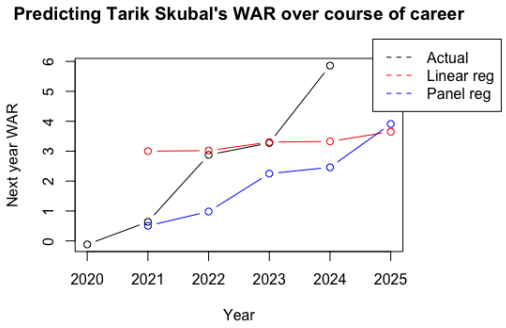

Figure 11: Line Graph of Predicted WAR for Tarik Skubal

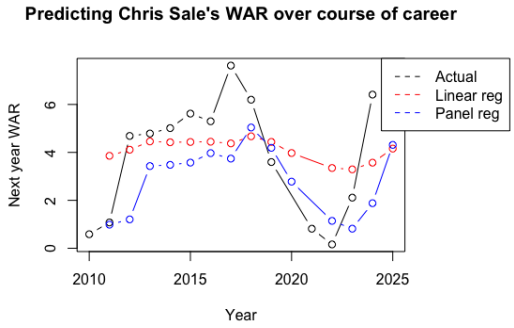

Figure 12: Line Graph of Predicted WAR for Chris Sale

5 Conclusions & Limitations

For predicting the WAR of the current year, all of our models were decently accurate in terms of MSE, with our decision tree, XGBoost, linear regression models all being effective. As for forecasting, it was more difficult, but our modeling was still decently accurate, especially when it comes to the linear regression forecaster. And even the less accurate model of the two forecasting algorithms (random effects panel regression) was still effective (and arguably more so than linear regression) at predicting direction of WAR change over time.

However, this study has several limitations that should be addressed in future research. First, some important factors might have been excluded. Variables such as psychological attributes, environmental conditions, and defensive support from teammates can significantly impact a pitcher's WAR but were not incorporated in this study.

Second, the assumption of linearity may oversimplify the relationship between pitcher performance metrics and WAR. While linear regression provides interpretability, the true relationship between these variables may be nonlinear. Given the complexity of WAR calculation, advanced models such as neural networks or random forests could capture nonlinear interactions more effectively.

Third, generalizability issues may arise due to the dataset's time frame. Our study utilized data from 2010 to 2024, but changes in MLB rules, advancements in player training methods, and evolving pitching strategies could impact the model's predictive power.

References

Eddie, Comeaux (2013). *Quick pitch: mlb talent evaluators*. en. URL: https://www.abca.org/magazine/magazine/2013-1-Winter/Quick_Pitch_MLB_Talent_Evaluators.aspx (visited on 03/13/2025).

Fangraphs baseball — baseball statistics and analysis (2025). en. URL: https://www.fangraphs.com (visited on 03/13/2025).

Mlb. Com — the official site of major league baseball (2025). en. URL: https://www.mlb.com/ (visited on 03/13/2025).

Wins above replacement (War) — glossary (2025). en. URL: https://www.mlb.com/glossary/advanced-stats/wins-above-replacement (visited on 03/13/2025).