USING SURVIVAL ANALYSIS FOR BETTER MLB PLAYER VALUATION

Williams, Kaitlyn & Woods, Kimmi & Gottlieb, Eliana & Hawkins, Will & Zabriskie, Ben
Department of Computer Science
Georgia Institute of Technology
Atlanta, GA 30332, USA
{kwilliams380, kwoods36, egottlieb6, whawkins32, bzabriskie3}@gatech.edu

ABSTRACT

Major League Baseball (MLB) teams are constantly looking for new ways to gain a competitive advantage. One such way is using analytics to improve player valuation and acquisition. Inspired by the pioneering Moneyball approach, this paper introduces a novel approach combining survival analysis and machine learning to better predict player decline and incorporate this decline information into valuation. Using data from the Lahman Baseball Database (1871-2023), we developed a Random Survival Forest model to predict individual player's survival and hazard functions. From these functions, we then predicted performance decline events, achieving 88.72% accuracy for position players and 84.06% accuracy for pitchers within a one-year margin. Building on these insights, we implemented two Random Forest Regressors for salary prediction - one using traditional statistics and another incorporating survival analysis features. Our results show that current salaries are not properly accounting for player decline and that older players are being overvalued. Overall, we show that using a survival analysis-centric approach can give MLB decision makers more information on potential player acquisitions and current roster members.

1 INTRODUCTION

In 2002, the Oakland Athletics revolutionized baseball by implementing an analytics-driven approach to player evaluation, known as 'Moneyball.' Faced with budget constraints, they leveraged data analysis to identify undervalued players, ultimately achieving unexpected success. This analytical approach remains relevant today in Major League Baseball, a $12B industry where improved on-field performance directly impacts fan engagement and revenue. Building on this foundation, we explore how machine learning can enhance player evaluation, particularly focusing on a critical challenge facing MLB general managers: accurately valuing veteran free agents. Consider a GM with a competitive roster seeking to maximize immediate championship potential through free agency. While these seasoned players offer proven MLB performance, their age-related decline poses a significant risk. Our research addresses this challenge by developing a predictive model that incorporates traditional baseball statistics to forecast player performance and determine appropriate salary valuations, with special attention to age-related performance degradation.

2 RELATED WORK

The original Moneyball approach (Hakes & Sauer, 2006) used a linear regression model to identify which player statistics were most important for team success. For example, on-base percentage was identified as a key statistic that most traditional scouts were undervaluing. More recent work (Heaton & Mitra, 2022) has built on this original Moneyball approach, using advancements in Natural Language Processing and Computer Vision to better measure player performance and value. Separately, other work (Melling, 2017) has used a survival analysis approach to predict the length of a player's career. However, we found no prior work that used survival analysis as a way to valuate players. Thus, our work seeks to fill this gap by using an approach founded on survival analysis to make player valuations and roster decisions.

3 DATA COLLECTION AND PREPARATION

3.1 DATASET

Our data came from the Lahman Baseball Database (Lahman, 2024). More specifically we used the following datasets from the database:

- Batting: 22 features x 113,800 entries; batting statistics for each player by season
- Fielding: 18 features x 151,507 entries; fielding statistics for each player by season
- Pitching: 30 features x 51,368 entries; fielding statistics for each player by season
- Teams: 47 features x 3045 entries; summed statistics for each team by season
- Salary: 5 features x 26,428 entries (data from 1985 to 2016); salary data for each player by season
- People: 24 features x 21,010 entries; player information linked to player codes in other datasets

The data contains statistics from the MLB from 1871 through 2023.

We combined the Batting and Fielding datasets to get the stats for position players, and separately, we used the Pitching dataset for pitchers. Additionally, we combined both the position player and pitcher datasets with the People dataset to get a player's age during each season.

3.2 Data Cleaning/Preparation

We removed players who were only in the MLB for one season. Some players played for multiple teams within a season, and those players had a different row of data for each season. To get a player's total stats for a season year, we used the sum of the columns of each row for a player a season year (excluding age). Additionally, we took out players who played fewer than 115 games in a season to account for injured players. We imputed missing data points for players using mean, rolling average (taking the average values for a player using the values from the last 3 seasons), and forward fill. Some missing statistics could be calculated directly from other non-missing fields.

4 Exploratory Data Analysis (EDA)

4.1 Correlation Analysis

By the nature of baseball statistics, many features are very correlated, so we wanted to note which features were the most correlated and be cautious of multi-collinearity in our models when using these features. We created correlation matrices for both position players and field players to confirm our thoughts about the correlated features, finding the most correlated features for pitchers to be Walks, Hits, and Innings Pitched, and the most correlated features for field players to be: Hits, Home Runs, and Earned Runs. Appendix Section A.1 contains the correlation matrices as visuals.

5 Methodology

5.1 Survival Analysis

5.1.1 Survival Analysis Overview

The core of our approach is survival analysis, a method that we have not discussed in class. This approach aims to find a connection between input features and the time of an event. While this type of analysis is common in the medical industry and for machinery, where the event may be the death of a patient or the failure of a machine part, we can apply similar principles to baseball players. The key challenge with this type of data is the use of censored data, which results when someone or something has not yet experienced the event of interest. In contrast, uncensored data are data points where we know the exact time until the event, or survival time, for that case. Because censored cases haven't experienced the event, we do not yet know their exact survival time. However, this censored data should not be omitted as it gives valuable insights by telling us the minimum time until an event will occur for that observation. By combining censored and uncensored data, we can build survival curves, which show the probability of ”survival” over time, and hazard functions, which represent the instantaneous risk of the event occurring at any given time. These tools allow us to analyze patterns of player longevity, identify periods of heightened risk for performance decline, and make more informed decisions about player valuation and contracts.

Appendix Section A.2 walks through an example of how censored data is leveraged to build a survival function. This basic example shows how censored data can be used to build survival curves and hazard functions for an overall group. In the sections that follow, we will describe how this approach can be extended to create survival curves on a a player-by-player basis.

5.1.2 Random Survival Forest

To actually develop the survival curves and hazard functions for the players, we decided to use a Random Survival Forest. Random Survival Forests (RSF) are an extension of Random Forests specifically designed to handle time-to-event data and survival analysis with censored observations. Specifically, we used the sksurv.ensemble.RandomSurvivalForest package (Pölsterl, 2023). The algorithm works by constructing multiple survival trees through bootstrap sampling, where each tree is trained on a randomly selected subset of the original data while preserving crucial time-to-event structure and censoring information. During tree construction, each node splits the data using a log-rank test statistic calculated from a random subset of features, creating branches that maximize the difference in survival patterns between resulting groups. For prediction, when a new case is introduced, it traverses down each tree until reaching terminal nodes, where survival functions are estimated non-parametrically using the Kaplan-Meier and Nelson-Aalen estimators based on the training samples that reached those nodes. The final prediction is an ensemble estimate that combines survival curves from all trees in the forest. RSFs are particularly beneficial because they offer several advantages: they naturally handle censored data, capture complex variable relationships, reduce overfitting through ensemble averaging, and provide more stable and accurate survival predictions than single tree models. The ensemble approach, combining predictions from multiple trees, creates a robust model that can generate individualized survival curve predictions while maintaining high predictive accuracy. For these reasons, we selected RSF as the preferred approach for the survival analysis element of our project.

5.1.3 RSF Implementation

To develop a survival model for predicting player performance decline, we first established a clear definition of what constitutes a "decline event." For each performance metric—such as batting average, home runs, and strikeouts—we identified the year in which each player achieved their peak performance. Using annual league averages as a benchmark, we determined whether a player's performance in a given metric dropped below the league average in the years following their peak. A decline event was defined as occurring when a player's performance fell below the league average in at least 50% of the tracked metrics after their respective peak years. This approach ensured that decline events represented meaningful and sustained decreases in performance rather than short-term variability.

Feature selection was a critical step to ensure the predictive validity of the model. Since we were using these same performance metrics to define decline, we focused on choosing features that were likely to be early indicators of a player's decline, rather than those influenced by the decline itself. For example, we avoided features such as reduced playtime or changes in role, as they are more likely consequences of decline. Instead, we prioritized metrics like batting average, home runs, and strikeouts—indicators that have historically shown patterns leading to a decline. This careful selection ensured that the model used meaningful, predictive signals to forecast player decline, rather than post-decline effects. By systematically identifying decline events across the dataset, we established a robust framework for conducting survival analysis and modeling player performance trajectories.

For pitchers, the features used to identify decline events were selected based on their relevance to performance. The chosen metrics provided a comprehensive view of a pitcher's effectiveness and were integral to determining decline events. For position players, the features used for identifying decline events emphasized both offensive and defensive performance. By incorporating both offensive and defensive statistics, these features offered a balanced approach to capturing a position player's overall contributions and identifying meaningful declines. A full list of features used can be found in A.4.

Figure 3 in A.3 visualizes the performance trajectory of Reed Johnson, illustrating his metrics relative to league averages. Each feature, such as batting average, home runs, and runs batted in

(RBI), is plotted over time, with the player's performance compared to the league averages for each respective year. The graph highlights key events in the player's career: the peak year, marked by a vertical green line, represents the year in which the player achieved their highest performance for a specific metric. The decline year, shown with a red vertical line, marks the point when the player's performance fell below the league average in 50% or more of the tracked metrics, signaling the occurrence of a decline event.

As shown in the graph, Reed Johnson experienced peaks in certain metrics even after his decline event. However, these later improvements in individual metrics occurred after the player had already declined in 50% of the tracked metrics, which aligns with our definition of a decline event. This pattern is expected, as the decline event is based on a sustained drop in performance across multiple metrics, and improvements in some areas do not necessarily offset the broader decline. The graph reinforces the idea that a decline event represents a sustained downturn in performance, even if isolated metrics show occasional peaks after the event.

After identifying the decline events, we prepared the data for survival analysis by creating key variables that would drive the model. We added a binary "decline event" column, where a value of 1 indicated a player experienced a decline event, and 0 indicated they had not. For those players who did experience a decline, we set the "time to event" column to reflect the number of years from their debut to the year they reached their decline event. For players who had not yet declined by the end of the observation period, we treated them as "censored." These censored players were assigned a "time to event" value based on the duration from their first season to the end of their career or the last season available in the dataset. This step ensured that we were accounting for both players who had experienced a decline event, players who retired before experiencing a decline event, and players whose careers were still ongoing at the time of analysis.

To capture the nuances of player performance over time, we engineered a set of features that would help the survival model predict decline. Key to this was the creation of trend-based variables, such as 3-year rolling averages and year-to-year differences. The rolling averages smoothed out short-term fluctuations in performance, giving a clearer view of a player's broader trend. The year-to-year differences highlighted any sudden changes in performance, which could signal an emerging decline. These trend features allowed us to focus on leading indicators—patterns in a player's performance that were predictive of future decline events—rather than simply using the raw values of the performance metrics themselves.

Once the data was prepared and trends were engineered, we applied the survival model to analyze the relationship between the selected features and the likelihood of a decline event. The model focused on individual players' performance trajectories, using time-to-event data to predict when a player might experience a decline. By examining the trends in performance metrics leading up to observed declines, the model identified patterns that could forecast future declines. This methodology allowed us to build a robust survival model capable of predicting player performance decline based on observable trends in their data.

5.2 Valuation

5.2.1 Early Attempts

The second primary element of our project was to assign players a value. We first tried to correlate individual player statistics with both run differential and win percentage using regression approaches. This was done to try to understand how a player's individual performance contributes to team success and create our own metric to measure player performance and value. However, results were poor using this approach. VIF analysis showed very high multicollinearity even when trying composite variables, non-linear terms, transformations, etc. Principal Component Analysis (PCA) was also tried, but results were still poor. Overall, we struggled to find any meaningful ways to exactly determine an individual player's contributions to his team's success.

5.2.2 Random Forest Regressor

After our initial attempts at creating our own method to valuate players, we instead turned to creating a model to predict a player's future salary given their performance statistics in prior years.

To do this, we used a Random Forest Regressor (RFR). Random Forest Regression combines many decision trees, each trained on different random subsets of data, and averages their predictions together. This ensemble approach produces more accurate and stable predictions than a single decision tree could achieve. We chose this model as it helps avoid overfitting and can capture complex relationships, which players salaries are based on. Specifically, we used sklearn.ensemble.RandomForestRegressor (learn Developers, 2024b). First, we trained the RFR using only the player statistics found in our dataset. Then, we introduced an additional feature derived from our survival analysis work to determine how this information would impact the predictions for a player's salary.

5.2.3 RFR Implementation

Building on this methodology, we implemented two RFR models to predict player salaries. Both models utilized performance statistics as feautres, but one model incorporated an additional feature derived from our survival analysis. This approach allowed us to assess the impact of survival-derived insights on salary prediction accuracy.

Our initial feature selection focused on comprehensive performance metrics that could meaningfully represent a player's value. For position players, we incorporated both offensive and defensive statistics. This multifaceted approach ensured that the model captured a player's holistic contribution rather than relying on isolated performance metrics. For pitchers, we also adopted a comprehensive feature set. The statistics chosen provided a nuanced view of pitcher effectiveness. A full list of features used can be found in A.4.

A critical innovation in our modeling approach was the incorporation of survival analysis features into the salary prediction model. We introduced two additional features derived from our earlier survival analysis work: a binary "decline event" column and a "time to event" column. These features allowed us to explore how a player's projected performance decline might influence their market value. By integrating these survival analysis metrics, we sought to analyze the predictive power of long-term performance trajectories.

We implemented two parallel RFR models to rigorously evaluate the impact of these additional features. The first model utilized only traditional performance statistics outline above, while the second incorporated the survival analysis metrics. This comparative approach enabled us to assess whether the decline event information provided meaningful predictive improvements.

Our model development process involved several key steps. We first prepared the dataset by creating paired observations where each row represented a player's previous season's statistics and the corresponding current season's salary. This approach allowed us to train a model that could predict future salary based on past performance. We employed standard scaling to normalize the feature spaces, ensuring that no single metric would disproportionately influence the model's predictions.

To optimize model performance, we utilized GridSearchCV (learn Developers, 2024a) to perform hyperparameter tuning, specifically exploring different numbers of estimators (trees) in the random forest. This process helped us identify the optimal model configuration that balanced model complexity with predictive accuracy. The grid search ranged from 20 to 500 estimators, allowing for a comprehensive exploration of model complexity.

The models were rigorously evaluated using standard regression metrics, include Mean Squared Error (MSE) and R-squared. By comparing the performance of the original model with the augmented model containing survival analysis features, we could quantitatively assess the incremental predictive value of these additional metrics.

Feature importance analysis played a crucial role in understanding the model's internal mechanics. By examining the relative importance of different features, we gained insights into which performance metrics most strongly influence salary predictions. This analysis not only validated our feature selection approach but also provided a nuanced understanding of how different performance indicators contribute to player valuation.

Our methodology represented a sophisticated approach to player valuation, moving beyond traditional statistical methods. By integrating survival analysis insights with a robust machine learning framework, we developed a more dynamic and predictive model of player salary determination.

6 Experimental Results

6.1 Survival Analysis Results

From the RSF model, the survival and cumulative hazard function for each player that had not yet experienced our defined decline event was predicted. To test the accuracy of these predictions against our test data, we had to convert these survival functions into a single time-to-event number. This was done by determining the point at which the survival function dropped below 0.50 and using that time to compare to the defined decline event in our test data. The results of the survival analysis highlight the ability of the model to predict events of player decline with high precision. For position players, the model achieved an accuracy of 88.72% within a one-year margin, with training and test concordance indices of 0.949 and 0.937, respectively. For pitchers, the accuracy within one year was 84.06%, with higher concordance indices of 0.984 (training) and 0.977 (test). The concordance index measures the rank correlation between predicted decline events and actual observations, providing an evaluation of the model's predictive reliability. The one-year margin was chosen because of the high cost of retaining declined players and the risks associated with prematurely releasing players.

Permutation feature importance was used to evaluate the importance of each feature in the Random Forest survival analysis. For position players, the most critical predictors included hits, at-bats, doubles, walks, RBIs, "time to decline event", and whether or not the decline event has occurred. For pitchers, key features were ERA, walks, innings pitched, total batters faced, cumulative batting average against, total earned runs, and total strikeouts. This method of assessing feature importance ensures a more accurate understanding of the contribution of each statistic to the model's performance by quantifying the impact of randomly shuffling each feature on prediction accuracy. These insights help identify the statistics most indicative of performance decline.

The survival analysis for pitchers is visualized in Figure 4 found in A.3, which illustrates the survival functions for a random subset of 100 pitchers in the test set. Each line represents the survival probability of an individual pitcher, with stepwise drops indicating decline events over time. The graph demonstrates that survival probability decreases as time progresses, reflecting the increasing likelihood of performance decline as pitchers age or accumulate seasons in their careers. This visualization was generated using Python rsf.predict_survival_function method, part of the RSF implementation. This function estimates the survival probabilities over time for each player in the test set by aggregating predictions from multiple survival trees within the forest. Each tree calculates survival probabilities based on training data, handling censored observations by including players who had not yet experienced a decline event. The final survival function for each player is an ensemble prediction that combines estimates from all trees to provide an individualized survival curve. By leveraging the RSF model and the predict_survival_function method, the analysis captures the risk of decline for each player, offering actionable insights into individual career trajectories and enhancing the ability to predict and manage potential performance drops.

The cumulative hazard analysis for pitchers is depicted in Figure 5 in A.3, which showcases the cumulative hazard functions for a random subset of 100 pitchers in the test set. Each line represents a pitcher's accumulated risk of decline over time, with steeper slopes indicating higher cumulative risks. This visualization was generated using the Python rsf.predict_cumulative_hazard_function method, which computes cumulative hazard estimates for each player by aggregating predictions from all survival trees in the Random Survival Forest (RSF) model. The ensemble approach ensures an accurate risk assessment by combining individual tree-level hazard estimates. Cumulative hazard analysis provides valuable information on the progressive risk of performance decline, allowing teams to anticipate long-term trends and make data-informed decisions about player management. [1]

[1] see Figures 6&7 in A.3 for similar figures for position players

6.2 Valuation Results

The RFR models for salary prediction revealed nuanced insights into player valuation, highlighting the complex relationship between performance metrics and compensation. We developed two parallel models for both position players and pitchers: one utilizing traditional performance statistics and another incorporating survival analysis features.

For position players, the initial model using only performance statistics demonstrated moderate predictive capability, with an R-squared value of 0.387 and a mean squared error (MSE) of $1.034\text{e}13$. Upon introducing survival analysis features, the model's performance deteriorated, with the MSE increasing to $1.823\text{e}13$ and the R-squared value dropping to -0.081. This negative R-squared indicates that the model with additional features performed worse than a horizontal line, suggesting that the survival analysis metrics introduced significant noise into the prediction process.

The pitchers' model exhibited a similar pattern. The initial model achieved an R-squared of 0.248 with an MSE of $8.556\text{e}12$. After incorporating survival analysis features, the model's performance degraded, with the MSE increasing to $1.206\text{e}13$ and the R-squared value declining to -0.0598.

Feature importance analysis provided additional context to these observations. For both position players and pitchers, the "decline event" feature ranked as the least important among the 26 total features considered. In the pitchers' model, the "time to event" feature was the second least important variable. This consistent finding suggests that the survival analysis features did not contribute meaningful predictive power to the salary estimation models.

Our hyperparameter tuning process explored the number of trees in the random forest, ranging from 20 to 500. While 500 trees marginally improved model performance compared to 160 trees - offering and increase of $9\text{e}{-4}$ in train R-squared and $1.7\text{e}{-3}$ in test R-squared for position players, and $1.8\text{e}{-3}$ in train R-squared and $3.2\text{e}{-3}$ in test R-squared for pitchers - we ultimately selected 160 trees. This decision balanced model complexity with computational efficiency, recognizing that the marginal performance gains did not justify the additional computation cost and model complexity.

7 Discussion and Interpretation

7.1 Strengths and Weaknesses

The survival analysis model can handle large datasets with numerous features and predict player decline events effectively. It does this with minimal runtime, enabling it to be scalable with low computational costs. Results of the model allow teams to save money by identifying players with consistent performance and building rosters that avoid severe under-performance. In the future, access to additional data - such as injury reports - would improve the model accuracy and prediction ability.

As we expected, the valuation model had poor predictive power and severe overfitting. The results confirm that real salaries are based on much more than on-field performance. The results also demonstrate that many players, especially elderly players, are being over-valued compared to their on-field performance.

7.2 Result Interpretation

Our findings align with our initial hypothesis that player compensation extends beyond on-field performance metrics. The models' inability to improve predictive accuracy with survival analysis features suggests that player valuation involves complex, intangible factors. Notably, older players often receive larger contracts despite being closer to potential performance decline, indicating that factors such as reputation, past achievements, and market dynamics play significant roles in salary determination.

The divergence between performance metrics and salary predictions underscores the multifaceted nature of player valuation in profession sports. While quantitative performance measures provide valuable insights, they represent only one dimension of a player's market value. Factors such as leadership qualities, marketability, fan appeal, and historical significance likely contribute to compensation in ways not captured by traditional statistical analysis.

8 Error Analysis

Potential sources of error in the model's predictions stem from several factors. One significant issue is the imputation of missing data, as filling gaps with averages can introduce bias, especially when outliers—such as exceptionally high-performing or under-performing players—skew the league averages. This can distort the accuracy of the model's predictions. Unexpected player injuries also pose a challenge, as they can accelerate the time to a decline event in ways the survival model cannot predict. Additionally, analyzing different types of pitchers, such as starters and closers, together may lead to inaccuracies due to the distinct metrics and roles associated with each position. Lastly, when players experience declines in certain statistics, such as at-bats or plate appearances, it can affect other metrics and reduce the availability of reliable data points. These factors underline the need for refined preprocessing and modeling techniques to improve predictive accuracy.

Figures 8&9 in A.3 compare the predicted versus actual decline years for test set players who have already experienced a decline event, with one graph representing pitchers and the other position players. The x-axis denotes the predicted decline year, while the y-axis shows the actual decline year. The diagonal red dashed line represents perfect agreement between predictions and actual outcomes. Points clustered near this line indicate accurate predictions, while deviations reflect prediction errors. The intensity of the shading corresponds to the number of players at each data point, with darker shades indicating higher player counts. These visualizations highlight the model's performance in accurately predicting decline years for players across different roles.

9 Conclusion

This study provides valuable insights into player performance decline and salary valuation in professional sports. The survival analysis model effectively predicts player decline events, offering teams a data-driven approach to roster management. By identifying potential performance declines early, teams can optimize their investments, reduce risks of under-performance, and build sustainable strategies for long-term success.

Our findings highlight the importance of considering age and the likelihood of performance decline in salary negotiations. MLB general managers can leverage survival analysis models to better assess the risk of offering long-term, high-value contracts to aging players. While these contracts may often reflect past achievements and fan appeal, integrating data-driven insights into these decisions can help teams strike a balance between rewarding performance and managing financial risk.

However, our results also reveal the limitations of performance metrics in fully capturing player valuation. The observed discrepancies between on-field performance and compensation suggest that intangible factors—such as leadership, marketability, and historical significance—play significant roles in determining salaries. These factors often result in higher salaries for older players, even as they approach potential performance decline.

The integration of survival analysis features into the salary prediction model did not enhance accuracy, further underscoring the multifaceted nature of player valuation. This finding reinforces the importance of incorporating qualitative aspects alongside quantitative measures in future valuation models. Looking forward, expanding the dataset to include additional variables such as training data, psychological attributes, or team dynamics could refine predictions and enhance the model's utility.

Ultimately, this research demonstrates the potential of survival analysis as a tool for understanding player dynamics while acknowledging the complexities of player valuation. MLB general managers can benefit from these insights to make more informed, data-driven decisions, not only improving team performance but also ensuring financial sustainability in a highly competitive market.

References

Jahn K. Hakes and Raymond D. Sauer. An economic evaluation of the moneyball hypothesis. *Journal of Economic Perspectives*, 20(3):173–186, 2006.

Connor Heaton and Prasenjit Mitra. Using machine learning to describe how players impact the game in the mlb, 2022. URL https://www.sloansportsconference.com/research-papers/using-machine-learning-to-describe-how-players-impact-the-game-in-the-mlb. Presented at the MIT Sloan Sports Analytics Conference.

Sean Lahman. Sean lahman baseball database, 2024. URL https://sabr.org/lahman-database/. Accessed: 2024-12-08.

Scikit learn Developers. *GridSearchCV*, 2024a. URL https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html. Accessed: 2024-12-08.

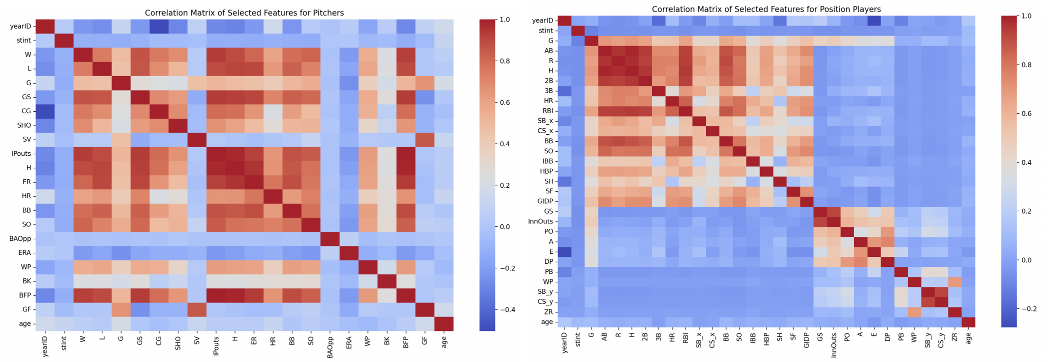
Scikit learn Developers. *RandomForestRegressor*, 2024b. URL https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Accessed: 2024-12-08.

Micah Melling. Survival analysis: How long do careers last?, 2017. URL https://www.baseballdatascience.com/survival-analysis-how-long-do-careers-last/. Accessed: 2024-12-08.

Sebastian Pölsterl. *scikit-survival Documentation*, 2023. URL https://scikit-survival.readthedocs.io/en/v0.23.0/index.html. Version 0.23.0.

A Appendix

A.1 Correlation Matrices



(a) Correlation Matrix: Pitchers

(b) Correlation Matrix: Position Players

Figure 1: Correlation Matrices for Pitchers and Position Players

A.2 Survival Analysis Applied Example

To demonstrate how survival analysis works in practice, let's examine a simple hypothetical medical study that tracked patient survival after a treatment. The study followed four patients until their deaths, providing us with the following data points. Patient A survived for 2 months, Patients B and C both survived for 5 months, and Patient D survived for 8 months after treatment. The Kaplan-Meier method calculates survival probabilities at each time point where an event (death) occurs. Initially, all four patients were alive, giving us a survival probability of 100%. At the 2-month mark, Patient A passed away. With one death out of four patients at risk, the survival probability became (4-1)/4 = 0.75 or 75%. At the 5-month mark, two more deaths occurred (Patients B and C). At this point, there were three patients at risk (the original four minus Patient A who died at 2 months). The survival probability is calculated by multiplying the previous survival probability (0.75) by the probability of surviving this time period (3-2)/3, resulting in $0.75 \times 0.33 = 0.25$ or 25%. Finally, at 8 months, the last patient (Patient D) passed away. With only one patient at risk and one death, the survival probability dropped to 0.

The Kaplan-Meier curve visually represents this step-wise decrease in survival probability over time. The curve starts at 1.0 (100% survival) and drops at each time point where deaths occur, creating a distinctive stepped pattern. The steeper drops at 5 months, where two deaths occurred simultaneously, illustrates periods of higher risk.

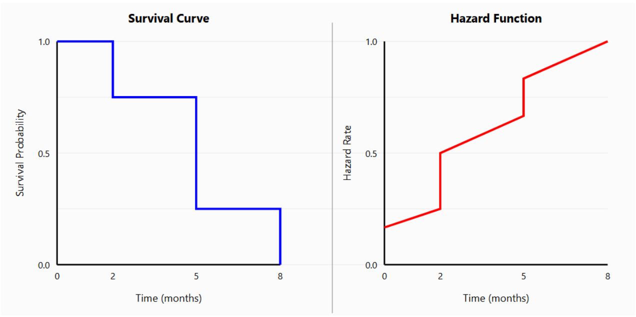


Figure 2: Medical Study Example Survival Function and Hazard Curve

The corresponding hazard function shows the instantaneous risk of death at each time point. Unlike the survival curve, which shows the cumulative probability of survival, the hazard function reveals the immediate risk of death. In this example, we see three distinct periods of hazard:

Initial period (0-2 months): Low hazard rate with only one death
Middle period (2-5 months): Increased hazard rate with two deaths occurring simultaneously
Final period (5-8 months): Highest hazard rate, as the last remaining patient died

The hazard function helps identify periods of heightened risk, which can be particularly valuable for medical interventions or patient monitoring. The sharp increase in hazard at the 5-month mark suggests this might be a critical period requiring additional attention.

A.3 FIGURES

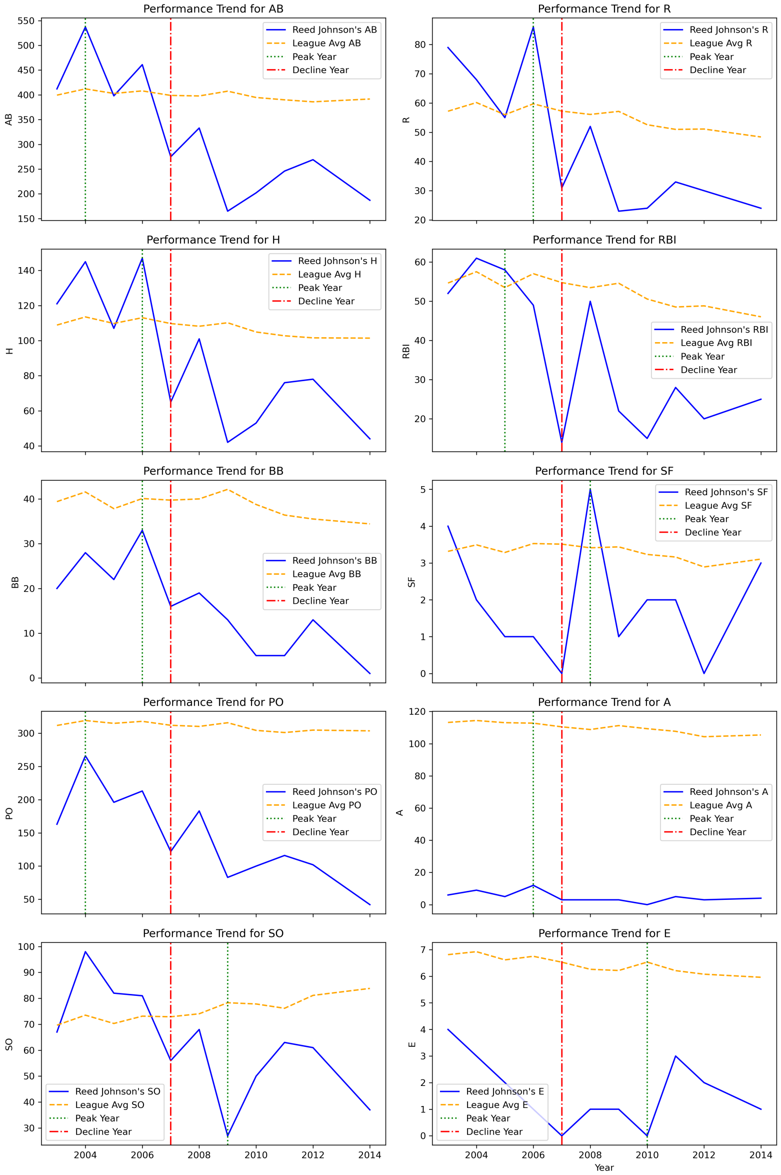


Figure 3: Performance Trends and Decline Event for Player johnsre02 (Reed Johnson)

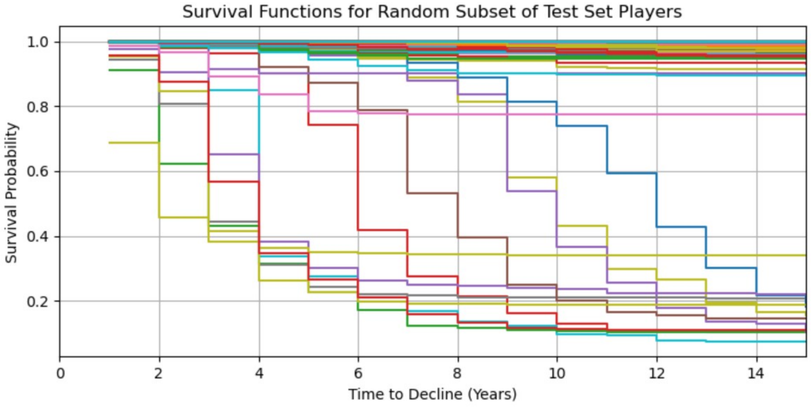


Figure 4: Survival functions for a random 100 pitchers from test data.

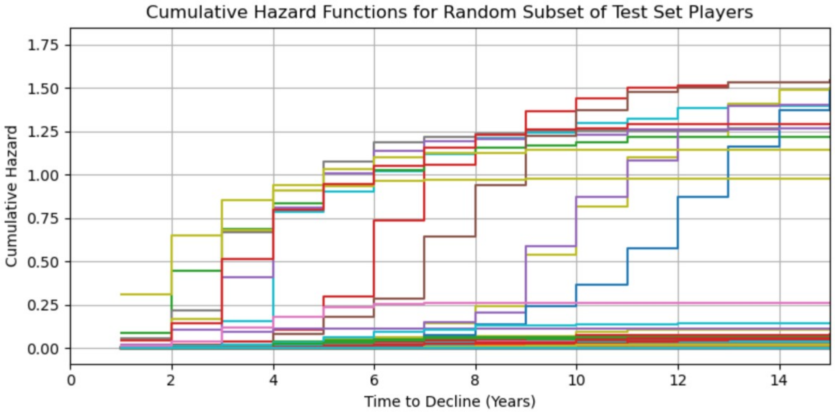


Figure 5: Cumulative hazard function for a random 100 pitchers from test data.

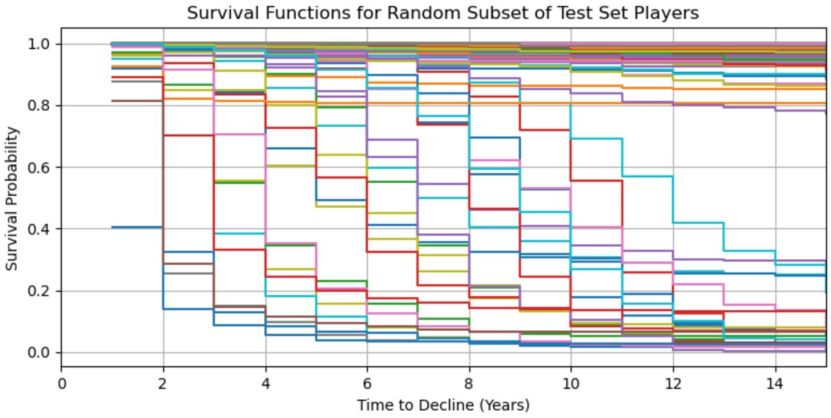


Figure 6: Survival functions for a random 100 position players from test data.

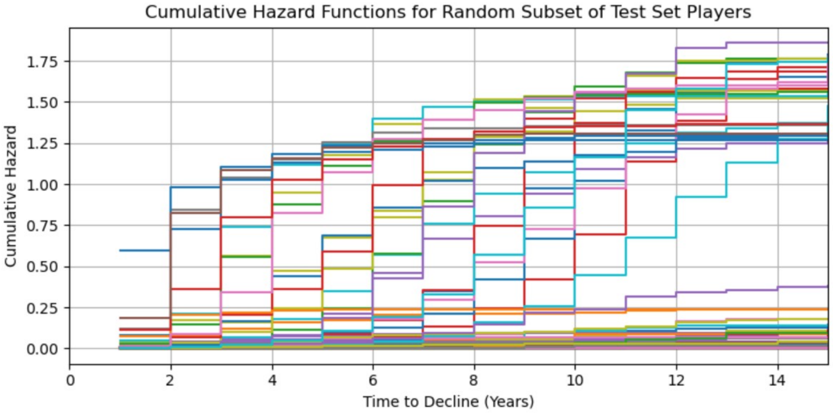


Figure 7: Cumulative hazard function for a random 100 position players from test data.

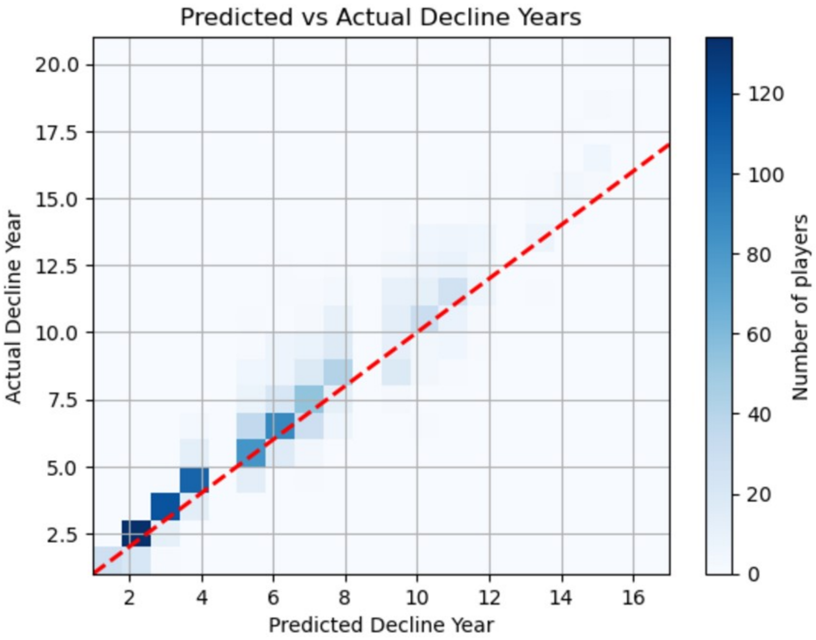


Figure 8: Predicted vs Actual of Position Players

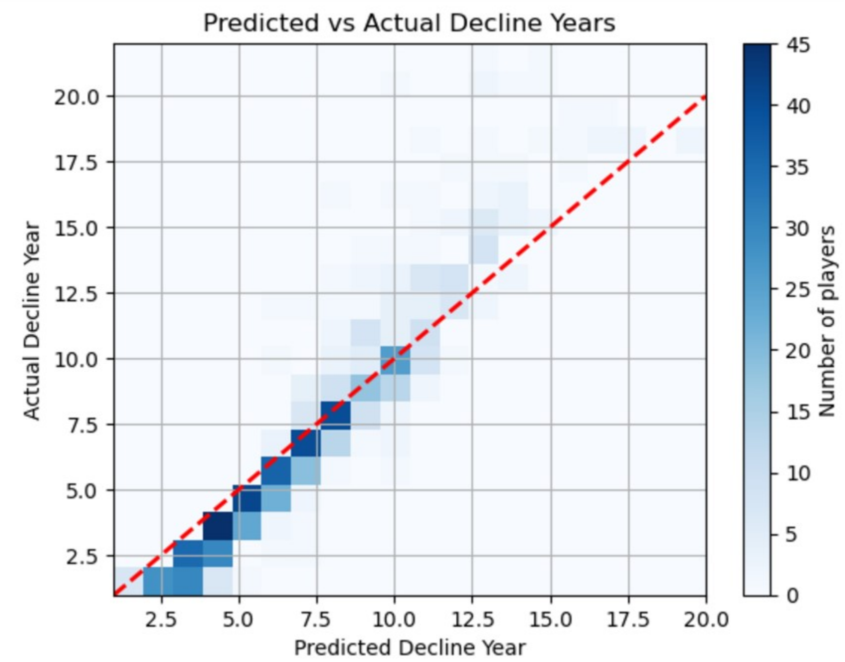


Figure 9: Predicted vs Actual of Pitchers

A.4 Features

Below is a tabulated summary of the features used in our data. The middle three columns indicate which dataset(s) the feature belongs to, and the last two columns indicate which model(s) used the feature.

Feature Name	Abbreviation	Pitchers	Batters	Fielders	Survival Analysis	Valuation
Wins	W	X	-	-	X	X
Losses	L	X	-	-	X	X
Games	G	X	X	X	-	X
Games Started	GS	X	-	X	-	X
Complete Games	CG	X	-	-	-	X
Shutouts	SHO	X	-	-	X	X
Saves	SV	X	-	-	X	X
Outs Pitched	IPOuts	X	-	-	X	X
Hits	H	X	X	-	X	X
Earned Runs	ER	X	-	-	X	X
Home Runs	HR	X	X	-	*	X
Walks	BB	X	X	-	X	X
Strikeouts	SO	X	X	-	X	X
Opponent's Batting Average	BAOpp	X	-	-	X	X
Earned Run Average	ERA	X	-	-	X	X
Wild Pitches	WP	X	-	-	X	X
Balks	BK	X	-	-	X	X
Batters Faced	BFP	X	-	-	-	X
Games Finished	GF	X	-	-	-	X
At Bats	AB	-	X	-	X	X
Runs	R	-	X	-	X	X
Runs Batted In	RBI	-	X	-	X	X
Singles	1B	-	X	-	-	X
Doubles	2B	-	X	-	-	X
Triples	3B	-	X	-	-	X
Stolen Bases	SB	-	X	-	-	X
Caught Stealing	CS	-	X	-	-	X
Intentional Walks	IBB	-	X	-	-	X
Hit by Pitch	HBP	-	X	-	-	X
Sacrifice Hits	SH	-	X	-	-	X
Sacrifice Flies	SF	-	X	-	X	X
Grounded into Double Plays	GIDP	-	X	-	-	X
Outs Played	InnOuts	-	-	X	-	X
Putouts	PO	-	-	X	X	X
Assists	A	-	-	X	X	X
Errors	E	-	-	X	X	X
Double Plays	DP	-	-	X	-	X

*Used in Survival Analysis for Pitchers, but not Batters