MLB Free Agent Evaluation

BY: ELIANA GOTTLIEB, WILL HAWKINS, KAITLYN WILLIAMS, KIMMI WOODS AND BEN ZABRISKIE

DECEMBER 2, 2024

Motivation

- **Moneyball**: Prior to the 2002 MLB season, the Oakland A's altered their player evaluation strategy. Having a much smaller budget than other teams, they turned to analytics to help them evaluate potential players and find the best value among the available players. Using this "moneyball" strategy, the team outperformed expectations and performed much better than outsiders anticipated.
- **Expanding Moneyball:** how we can use machine learning to help us evaluate baseball players?
- The MLB is a roughly $12B industry annually. Improved on-field performance can attract more fans and TV revenue, allowing teams to get more of this money.

Objective

We wanted to extend this Moneyball approach to 2024.

Imagine you are the **general manager** for an MLB team. You have several roster spots to fill in an offseason. Your roster has several very talented players on contract for the next few seasons. Therefore, you want to **maximize your chance of winning** now and get proven players to add to your team. With this, you are primarily looking at older free agents who have proven MLB performance. You want to **find the free agents that add the most value to your team**. However, you are concerned how these players will perform as they age.

Question

How can a GM properly valuate older players whose age may become a limiting factor?

Approach

- Using traditional baseball statistics, build a model that predicts a player's future performances and assigns them a value (salary).
- This model should account for decline in performance for aging players.

Data Overview

Our data comes from:

1. Lahman Baseball Database - link
   - Contains batting and pitching statistics from 1871-2023 seasons
   - The datasets we used:
     - Batting: 22 features x 113,800 entries
     - Fielding: 18 features x 151,507 entries
     - Pitching: 30 features x 51,368 entries
     - Teams: 47 features x 3045 entries
     - Salary: 5 features x 26,428 entries (data from 1985 to 2016)
     - People: 24 features x 21,010 entries

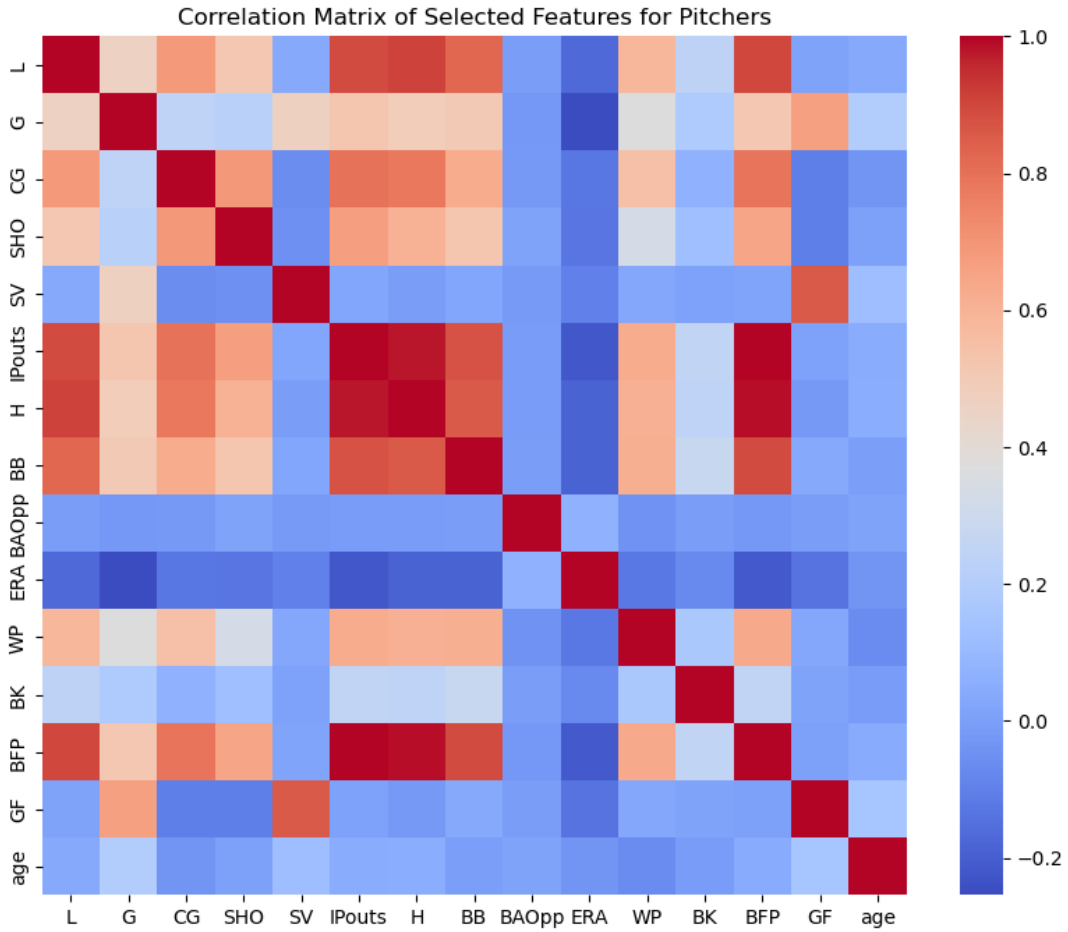
Data Preparation

Imputing Values into Data

- Imputed data points using player mean, rolling averages, or forward fill
- Removed players that were only in the MLB for one year
- For position players, removed players that played in a low number of games, to try and remove injured players from the data set.
- Will discuss the data pre-processing in more depth as we go through our methods

Exploratory Data Analysis

Correlation: Position Players



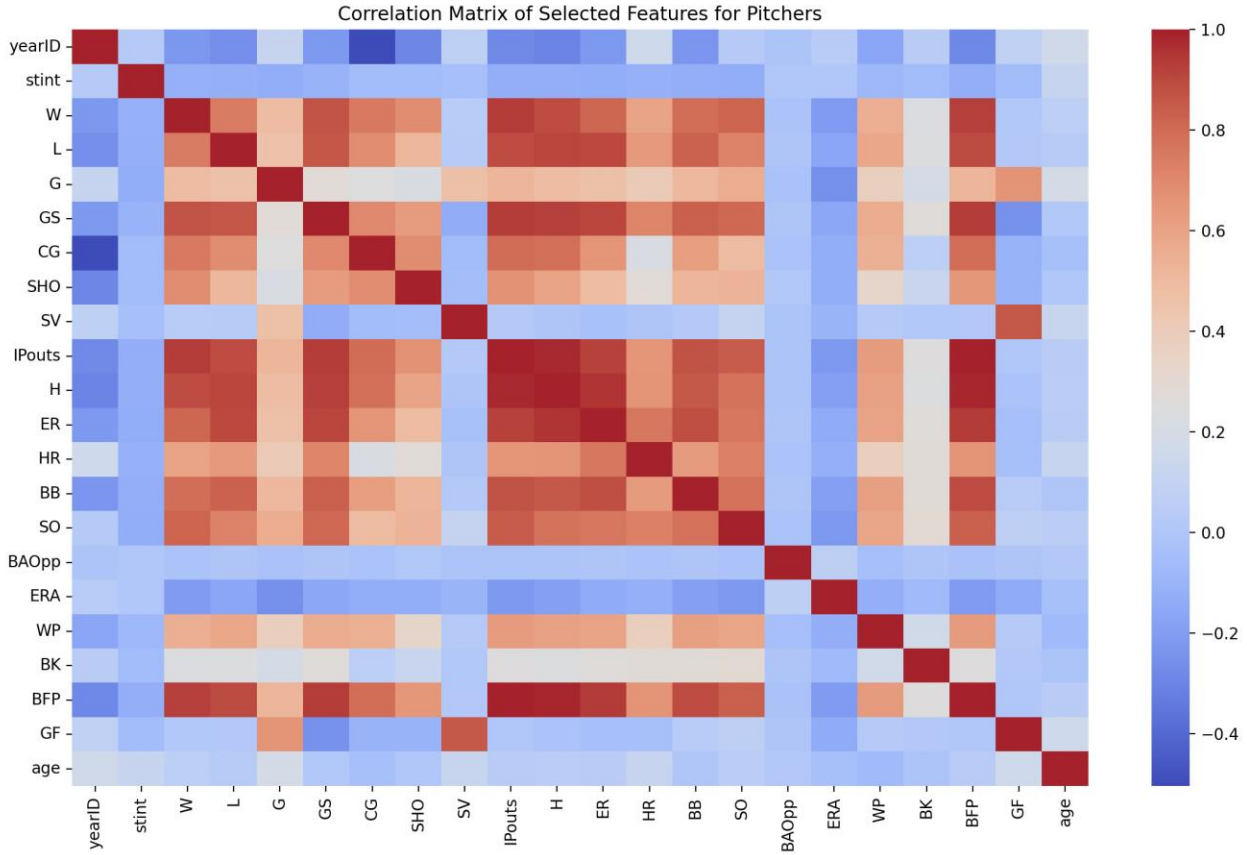
-By the nature of baseball stats, many features are very correlated

-Needed to be cautious of multi-collinearity in our models

-The most correlated features for position players include:

Games, At Bats, Runs, Hits, Doubles, Triples, Home Runs, RBI, Stolen Bases, Caught Stealing, Base on Balls, Strikeouts, Intentional walks, Hit by pitch, Sacrifice hits, Sacrifice flies, Grounded into double plays, Games started, Time played in the field

Correlation: Pitchers

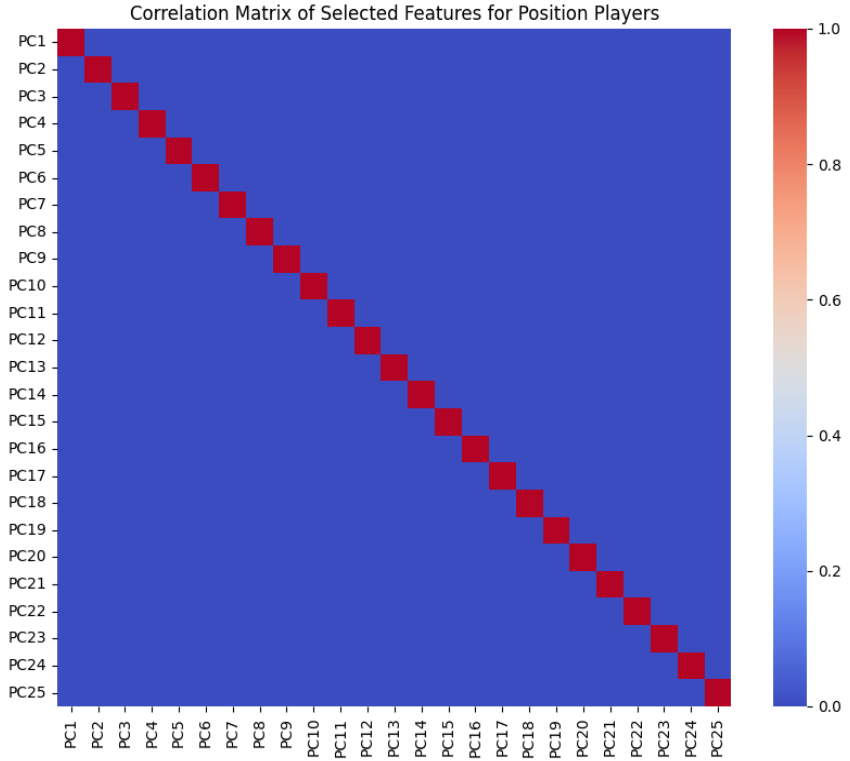


The most correlated features for pitchers included:

Wins, Losses, Games, Games started, Innings Pitched, Hits, Earned Runs, Home Runs, Walks, Strikeouts, Wild Pitches, Batters faced by pitchers

PCA: Position Players

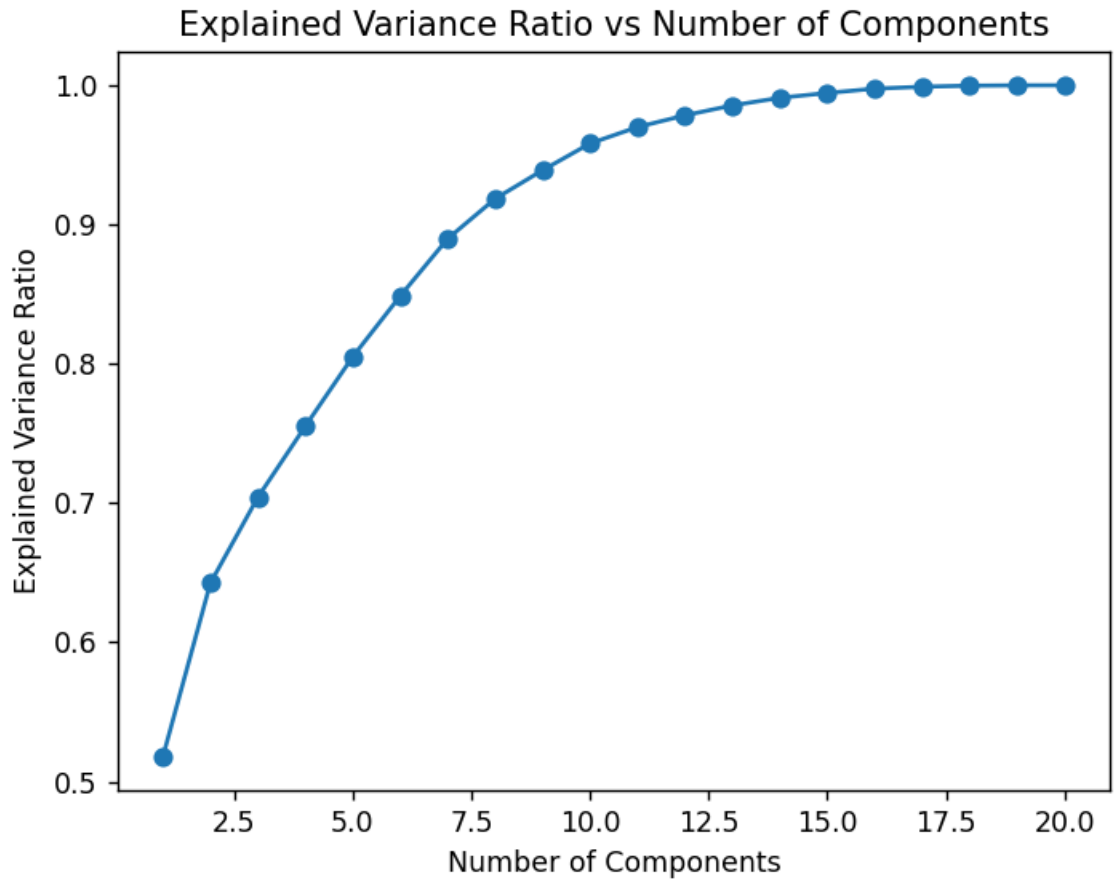
- Data Pre-Processing for Field Players:
  - Combine batting and fielding columns for fielding players
  - Select rows only after 1954 to remove large number of NaN values
  - Fill in 0s for players who batted but never fielded (or vice versa)
  - Remove catcher-specific columns
  - Standardize the data
  - Perform PCA
    - 4 components explain 80% of variance
    - 8 components explain 90% of variance
    - 12 components explain 95% of variance
    - 17 components explain 99% of variance



PCA: Pitchers

- Handle missing values:
  - Replace NaN with mean for BAOpp (Opponent's Batting Average) for that player across their career
  - Calculate ERA = 9*Earned Runs/Outs Pitched if ERA column is empty

  - 80% of variance explained by 5 components
  - 90% of variance explained by 8 components
  - 95% of variance explained by 10 components
  - 99% of variance explained by 14 components



Methodology and Models

Survival Models

- Survival analysis models aim to find a connection between features and the time of an event
- This type of analysis is commonly used in medical applications or for predicting when machinery might fail
- A key challenge with this type of modeling is censored data
  - If someone or something has not experienced the failure event yet, we do not know their exact time until failure
- Different models handle this censored data differently, but it provides valuable insights and should not be thrown out – we know the minimum time until failure

Using Censored Data – Basic Example

-Use the Kaplan-Meier method

-Consider a small study tracking patient survival after a treatment, where we have these uncensored data points (patients who passed away during the study):

-Patient A: 2 months, Patient B: 5 months, Patient C: 5 months, Patient D: 8 months

First, we look at how many patients were at risk at the start: 4 patients

At 2 months:

1. One death (Patient A)
2. Survival probability = (4-1)/4 = 0.75 or 75%

At 5 months:

1. Two deaths (Patients B and C)
2. Number at risk = 3 (original 4 minus the one who died at 2 months)
3. Survival probability = 0.75 × (3-2)/3 = 0.25 or 25%

At 8 months:

1. One death (Patient D)
2. Number at risk = 1
3. Survival probability = 0.25 × (1-1)/1 = 0

This gives us a step function that can help us predict patient's time to death

Random Survival Forest

-A Random Survival Forest is an extension of Random Forests that can handle time-to-event data with censoring.

-Builds multiple decision trees, where each tree estimates a survival function.

-At each node in the trees, separates patients into groups with different survival profiles.

-For terminal nodes in each tree, the survival and hazard function is calculated using the survival times of the training samples that reached that node

- Provides an estimate of the survival probability over time for that specific subgroup.

-The final prediction combines the survival estimates from all trees, creating an ensemble model that can predict individual patient survival curves

Defining Decline Event

Defining Player Decline:

- To create the survival model, we first needed a clear definition of what it means for a player to "decline."

Peak Performance Identification:

- For each performance metric (e.g., batting average, home runs), we identified the year in which each player achieved their peak value.

League Average Comparison:

- Using league averages for each metric, calculated annually, we determined whether a player's performance dropped below the league average after their peak.

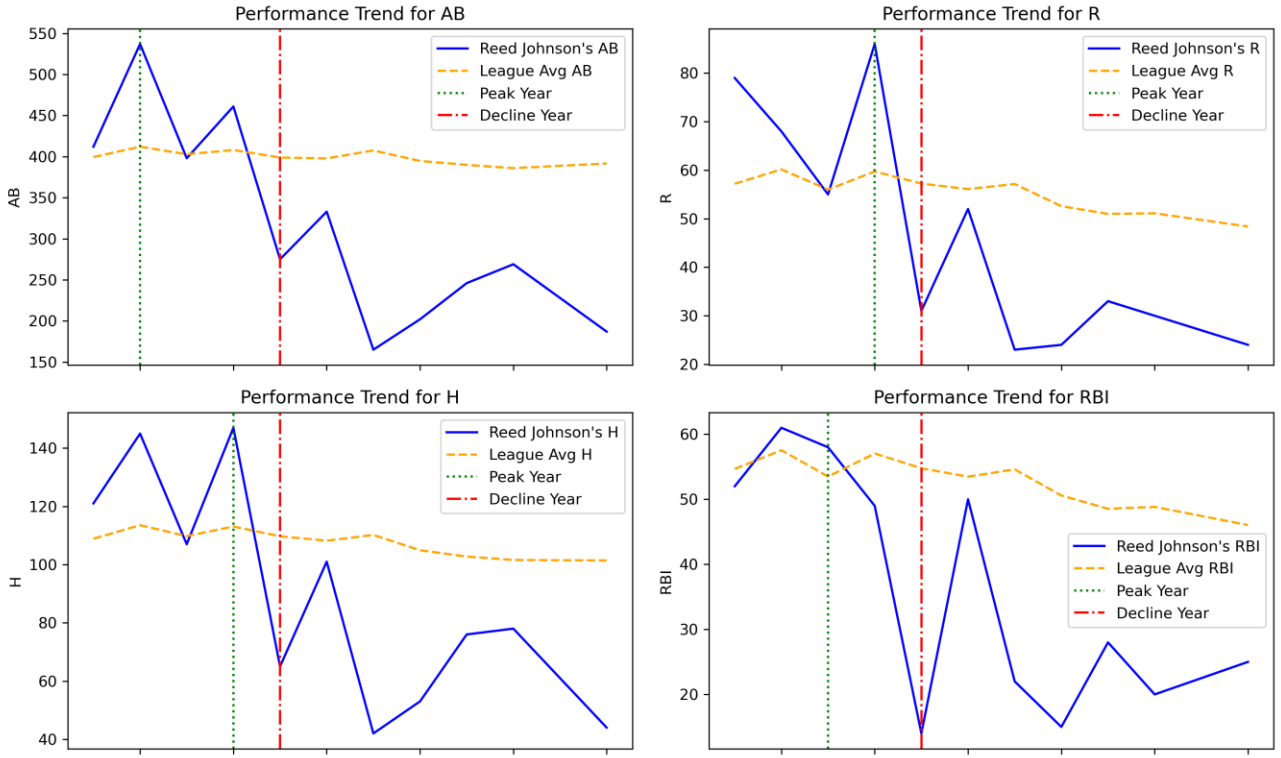
Decline Event Criteria:

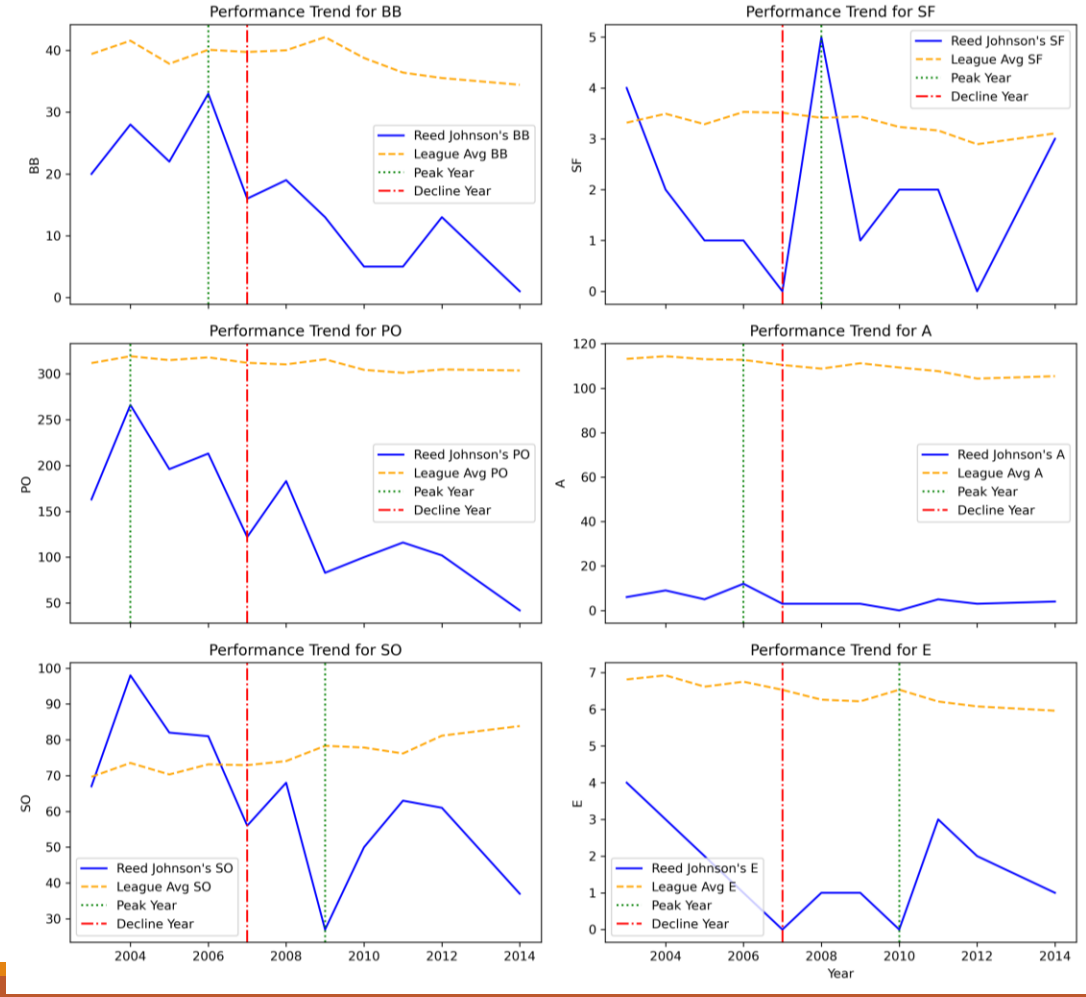
- A "decline event" was recorded if a player's performance fell below the league average in **50% or more of the tracked metrics** after their peak year in a specific metric.

Outcome:

- This method allowed us to systematically identify players who experienced a significant decline in performance and use these decline events as the basis for our survival analysis.

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





Implementing the Survival Model

Preparing Data for Survival Analysis:

- Each player was assigned a value in the **"decline event" column** (1 if they experienced a decline, 0 if not) and the **"time to event" column** (years until decline or until the player's last year in the dataset/end of their career).

Handling Censored Data:

- About 1/3 of the dataset included players who had already experienced a decline event.
- The remaining 2/3 consisted of **censored players**—those who had not yet declined during the observation period.
- Censored values for the "time to event" column were set as the time from the player's first season to the end of their career.

Implementing the Survival Model

- **Feature Engineering for Trend Analysis:**
  - Created **trend-based columns** for survival model predictors, including:
    - **3-year rolling averages** to smooth short-term fluctuations.
    - **Year-to-year differences** to capture changes in performance over time.
  - These trends allowed the model to focus on leading indicators of decline rather than directly predicting "time to event."

- **Feature Selection:**
  - Selected features based on **domain knowledge** to ensure they predicted decline rather than being outcomes of it.
  - Avoided using features that are directly impacted by a decline in player performance (e.g., reduced playtime or role changes) to maintain predictive validity.

- **Running the Model:**
  - The survival model analyzed **selected features and their trends**, focusing on individual players' trajectories to predict decline events.
  - Predictions were based on patterns in player performance leading up to a decline, using the **engineered trend data** for each feature.

Early Valuation Attempts

- First tried to correlate individual player statistics with both run differential and win percentage
- VIF analysis showed very high multicollinearity even when trying composite variables, non-linear terms, interaction terms, various transformations, etc.
- Trying to correlate to wins also produced high multicollinearity and had low explanatory power(R^2 of 0.41)

Valuation: Random Forest Regression

- Random Forest Regressor Models:
  - 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
- Model:
  - Train Random Forest Regressor to predict salaries given a player's statistics from previous years
  - Then, added survival analysis feature and retrained model

Valuation: Data Preparation

- Aggregate fielding data for players for each year
- Combine player statistics with salary information
- Filter for data only from 1985 to 2016
- Create time-series data set:
  - Take a player's statistics from one year (features) and pair with the salary from the following year (target)
- Pre-process:
  - Scale the features
  - Split into training and test sets

Results

Survival Analysis Results

Position player survival analysis:

Accuracy (±1 year) – 88.72%

Training Concordance Index: 0.9489721316729486

Test Concordance Index: 0.936922715431373 5

Pitcher survival analysis:

Accuracy (± 1 year) – 84.06%

Training Concordance Index: 0.983931139104 81

Test Concordance Index: 0.976700797534 0233

Concordance Index: It measures the rank correlation between predicted decline events and observed events. It is calculated as the ratio of concordant pairs (correctly ordered pairs) to comparable pairs.

Accuracy (±1 year): It measures how often the predicted "decline year" is within one year of actual event

- chose 1 year because keeping a declined player around too long impacts team performance and cost and cutting a player too soon risks losing potential value

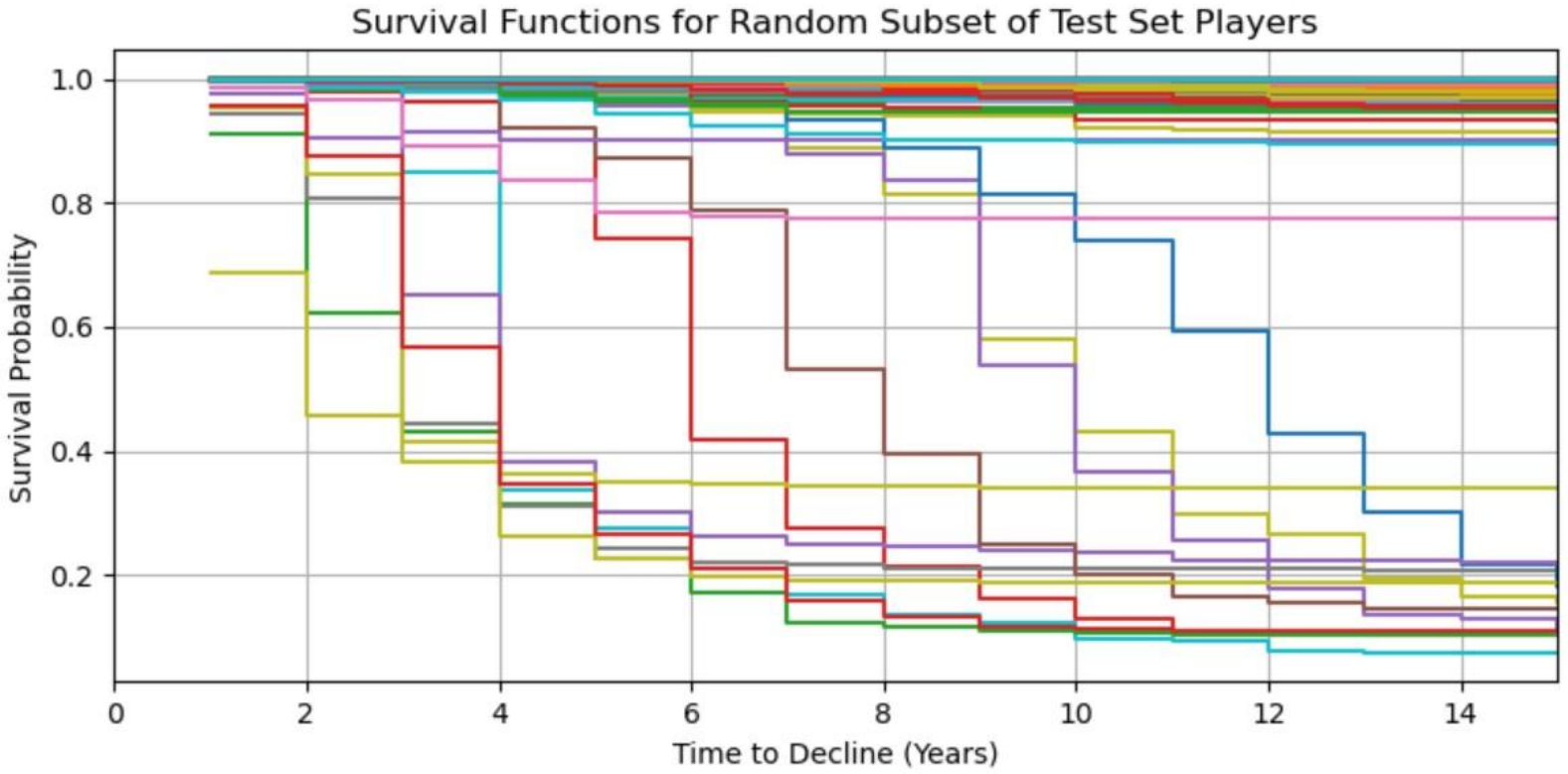
Survival Analysis Most Predictive Features

For position player survival analysis:

Decline Event, Time to the event, Hits, At Bats, Doubles, Walks, RBI

For pitcher survival analysis:

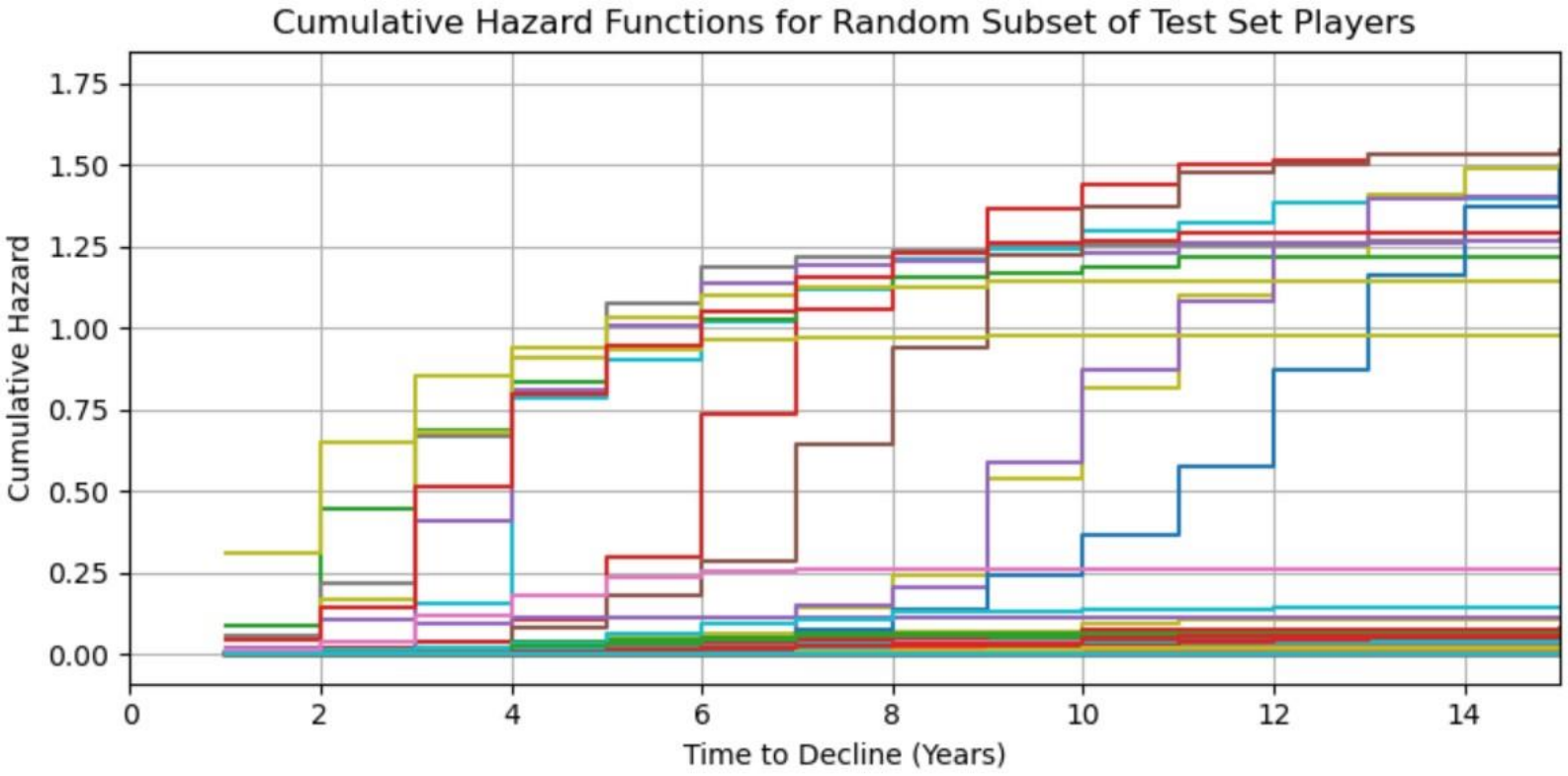
Decline Event, Time to the event, ERA, Walks, Innings Pitched, Total Batters Faced, Cumulative Batting Average Against, Total ER, Total Strikeouts



Survival Analysis - Pitchers

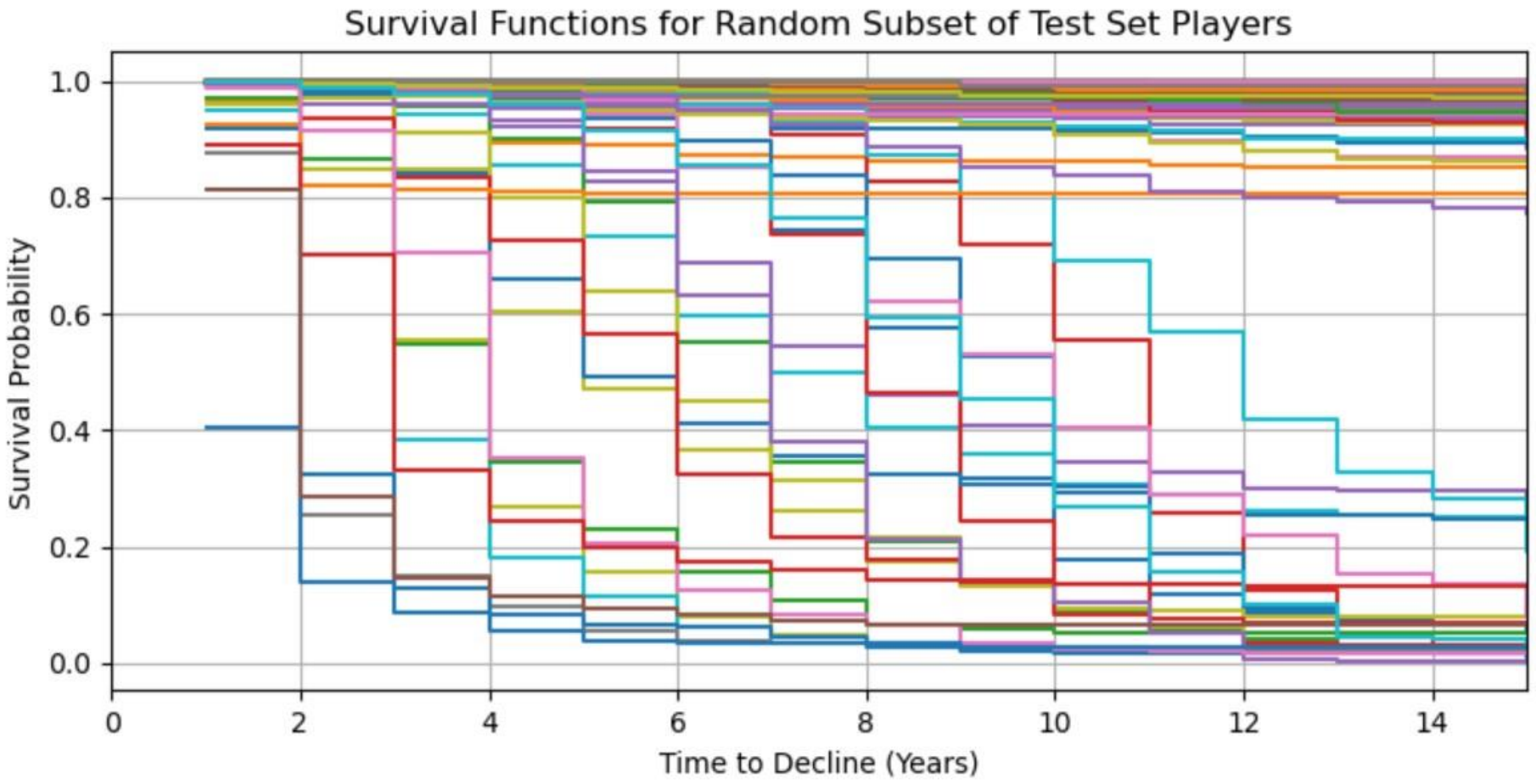
- Random subset of 100 pitchers of test set
- Each line represents a pitcher's survival probability, where stepwise drops indicate decline events over time
- Survival probability decreases over time as higher chances of decline exist as players age

Hazard Function - Pitchers

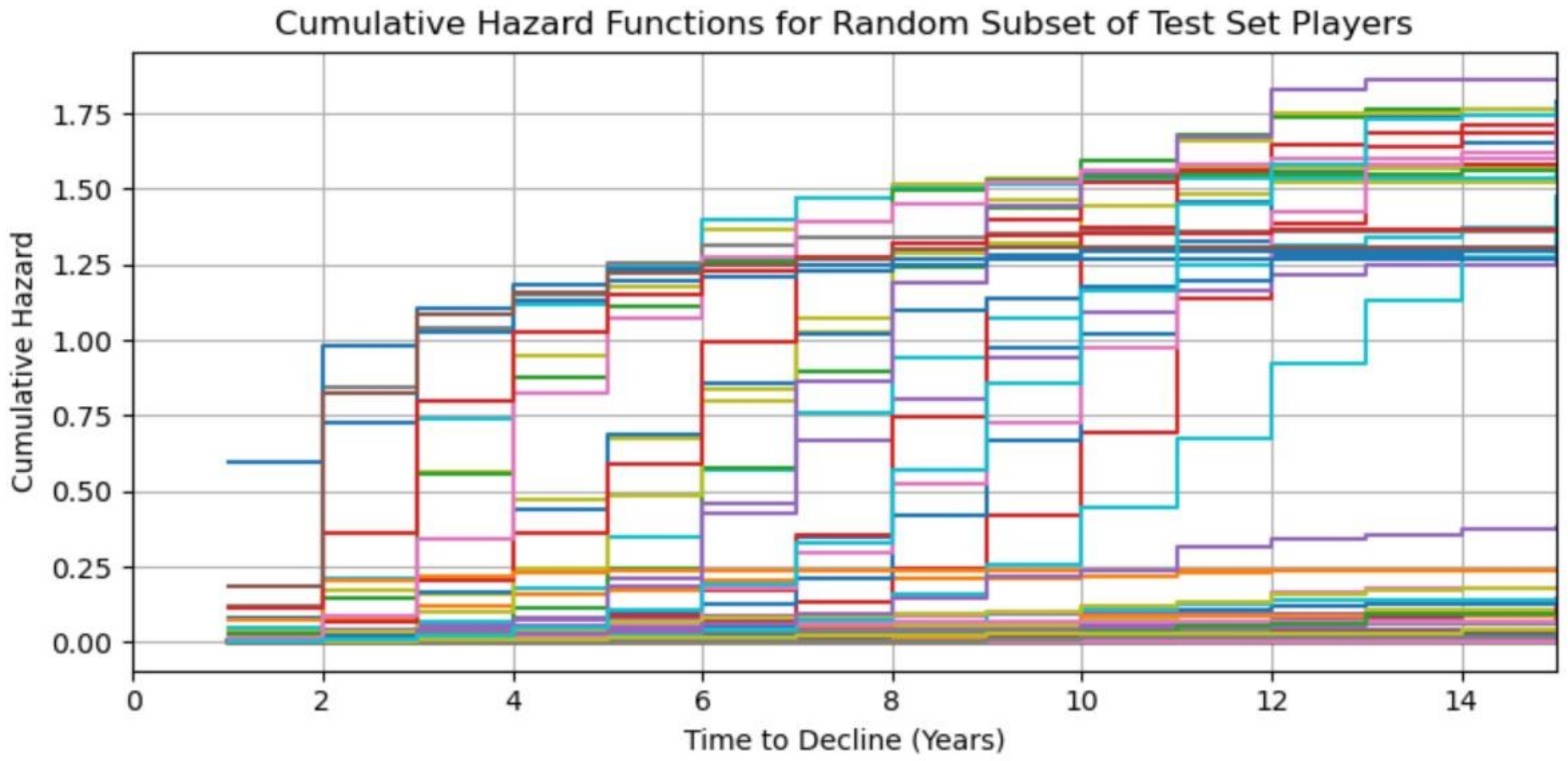


- Random subset of 1000 pitchers of test set
- Demonstrates pitcher's accumulated risk of decline over time, steeper slopes showing higher cumulative risks
- Individual cumulative hazard increases over time
- Created using rsf.predict_cumulative_hazard_function which estimates cumulative risk over time for each pitcher in RSF model

Survival Analysis - Position Players



Hazard Function - Position Players



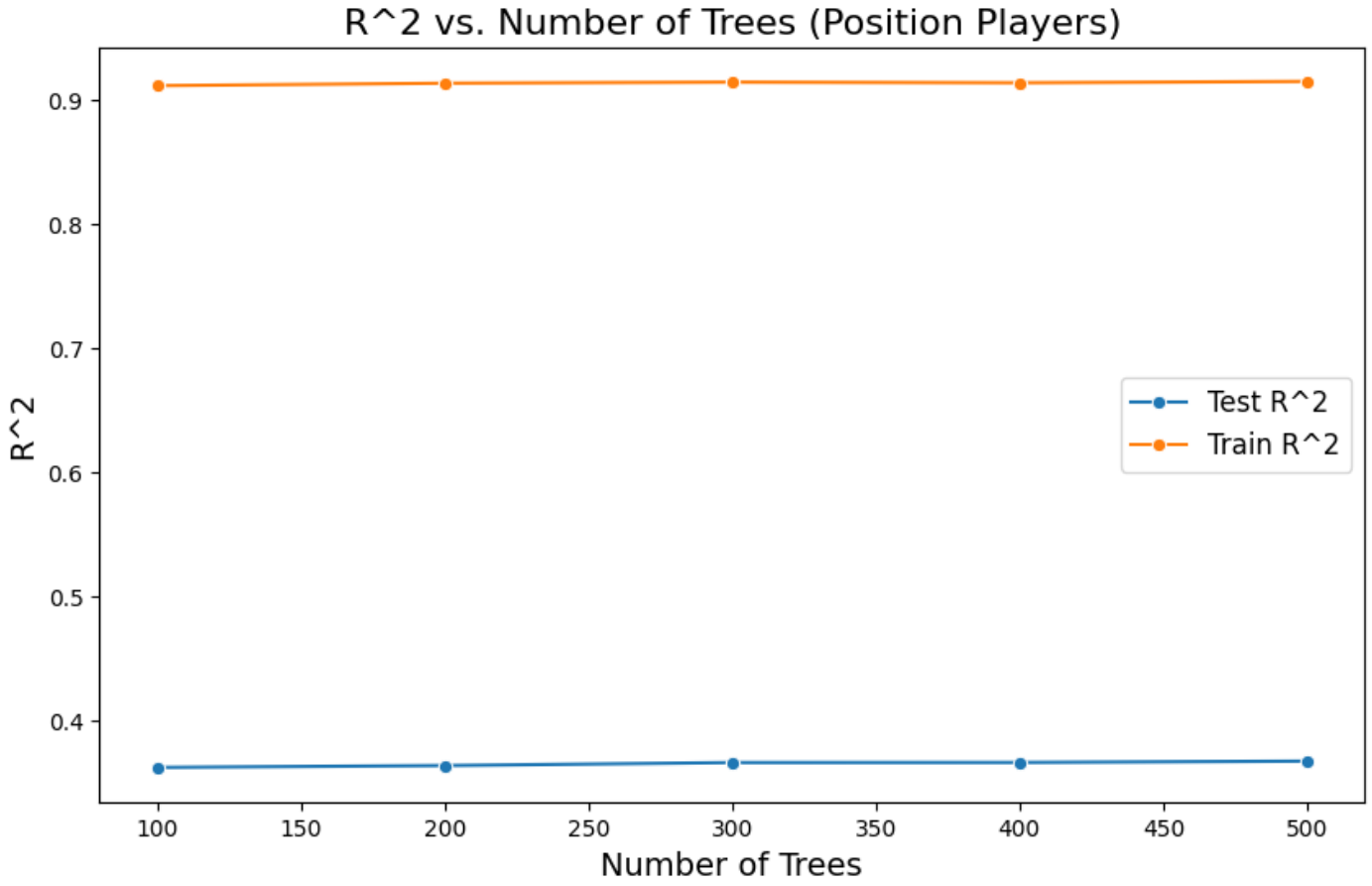
Valuation Results: No Survival Analysis

Hyperparameter search on the number of trees shows little effect on results

Test R^2 is approximately 0.36

Salaries were expected to be difficult to predict based on raw stats – much more goes into the real process

Train vs Test R^2 suggests Random Forest Regressor may not be generalizing well enough



Valuation with Survival Analysis Feature

Incorporated survival analysis results as another feature to include in model training

Original model (without survival analysis) had R^2 of 35.5%

New model was significantly worse – which we expected!

Most players make the most money near the ends of their careers, once they have accumulated experience

However, this often coincides with declining performance

This suggests our assumptions were correct – players at the end of their career are paid for their experience more so than their statistical output

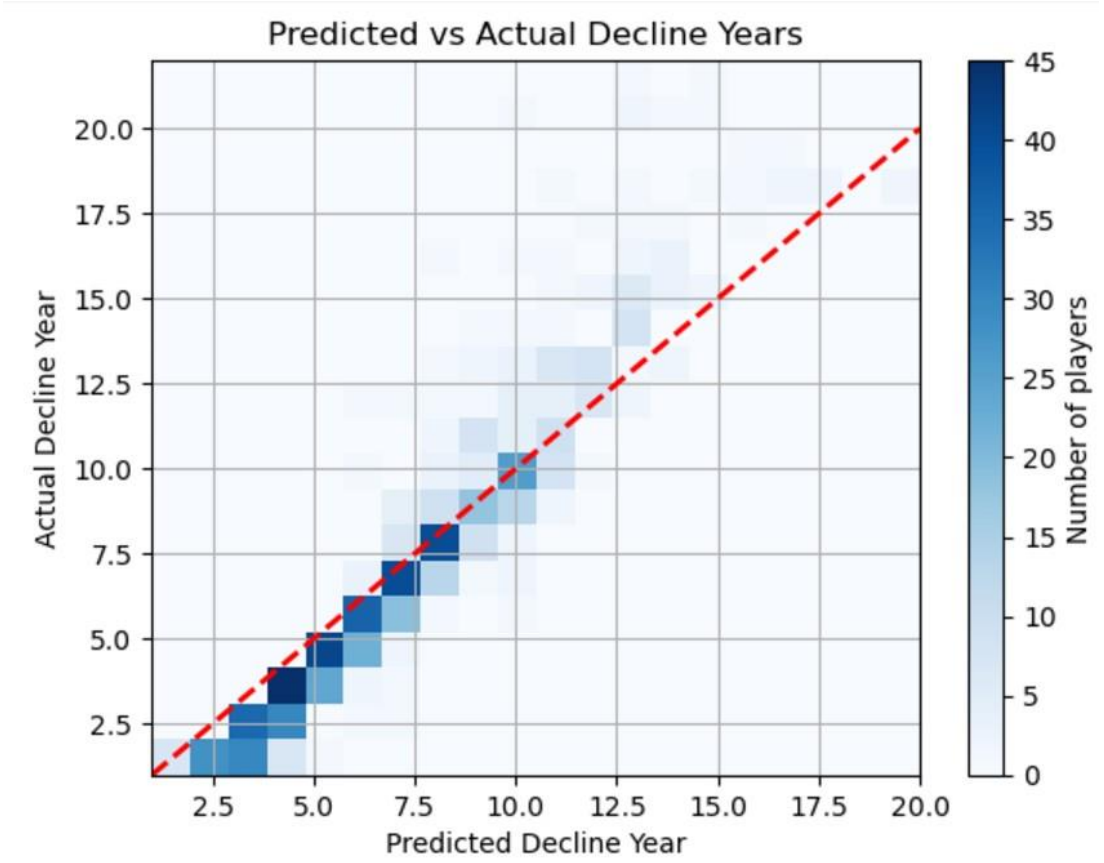
```
Original Model Metrics:
Mean Squared Error: 9421300094621.21
R-squared: 0.3551616349194193

New Model Metrics:
Mean Squared Error: 159690006208140.785
R-squared: -0.09299435871893436
```

Error Analysis

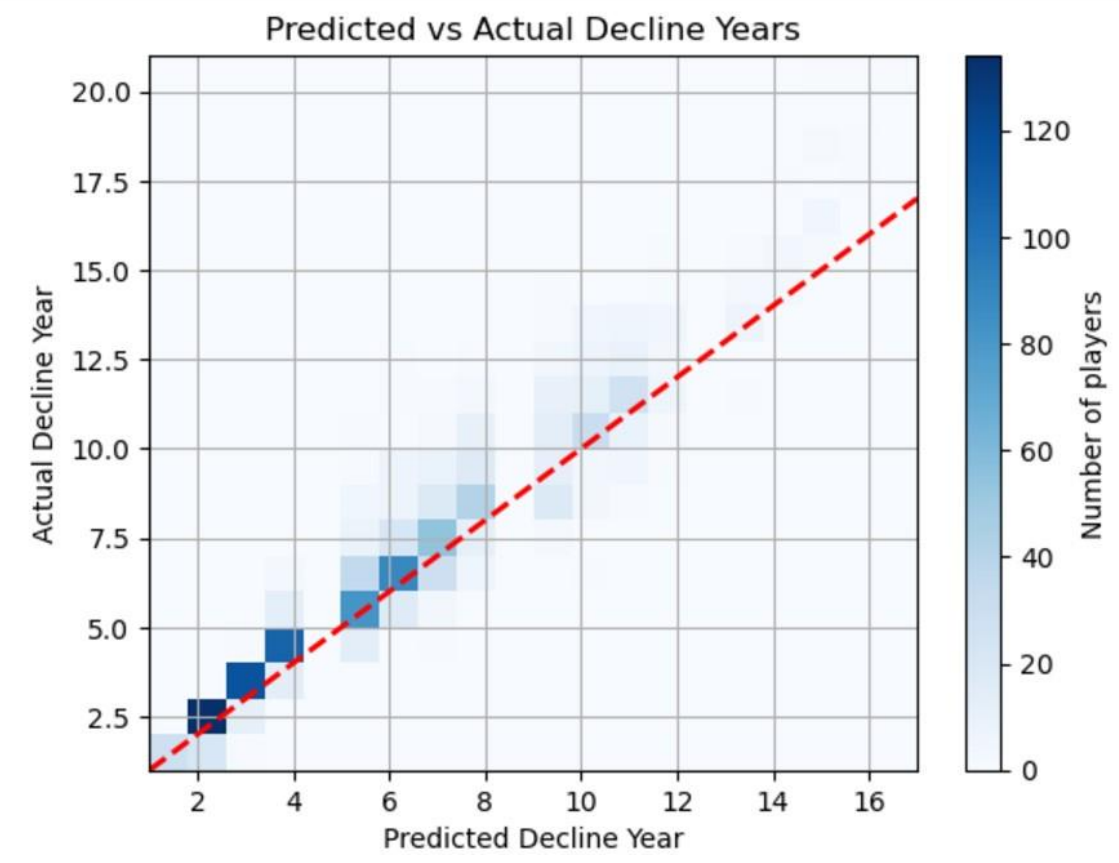
Possible Errors

- Imputation skews players' statistics, either improving or worsening them.
  - Imputing with mean introduces bias due to outliers having a significant impact on the overall league average.
- Unexpected injury may accelerate the time to a decline event for a player, which the survival model cannot predict
- Analyzing starting pitchers and closers together can lead to prediction errors due to differences in specific statistics and the contrasting roles these two types of pitchers play.
- Players showing decline in some statistics may experience a decline in others (ex: AB, PA, G, etc.) due to the nature of playing time and game decisions.
  - Data points for the player lessens

Density of Predicted vs Actual Decline Years

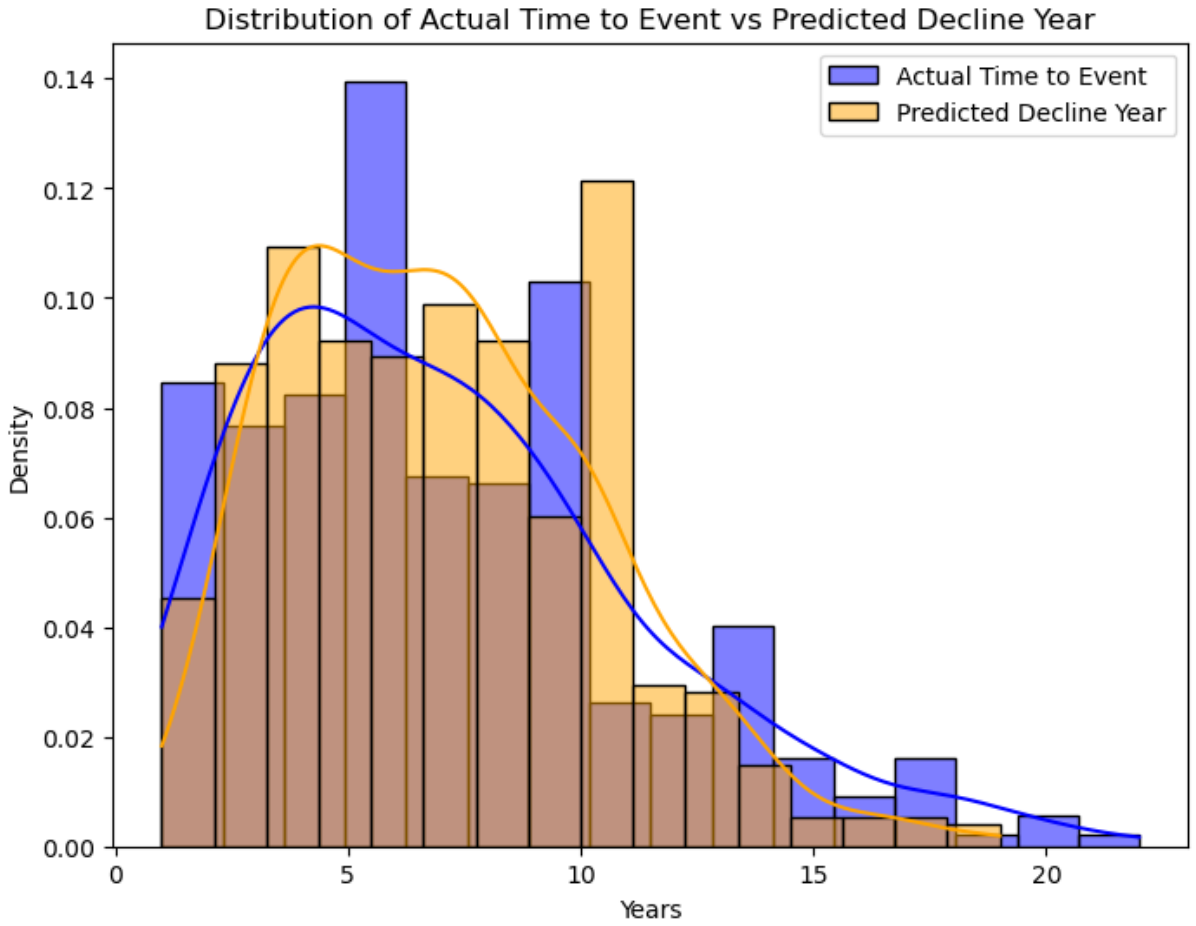


Pitchers

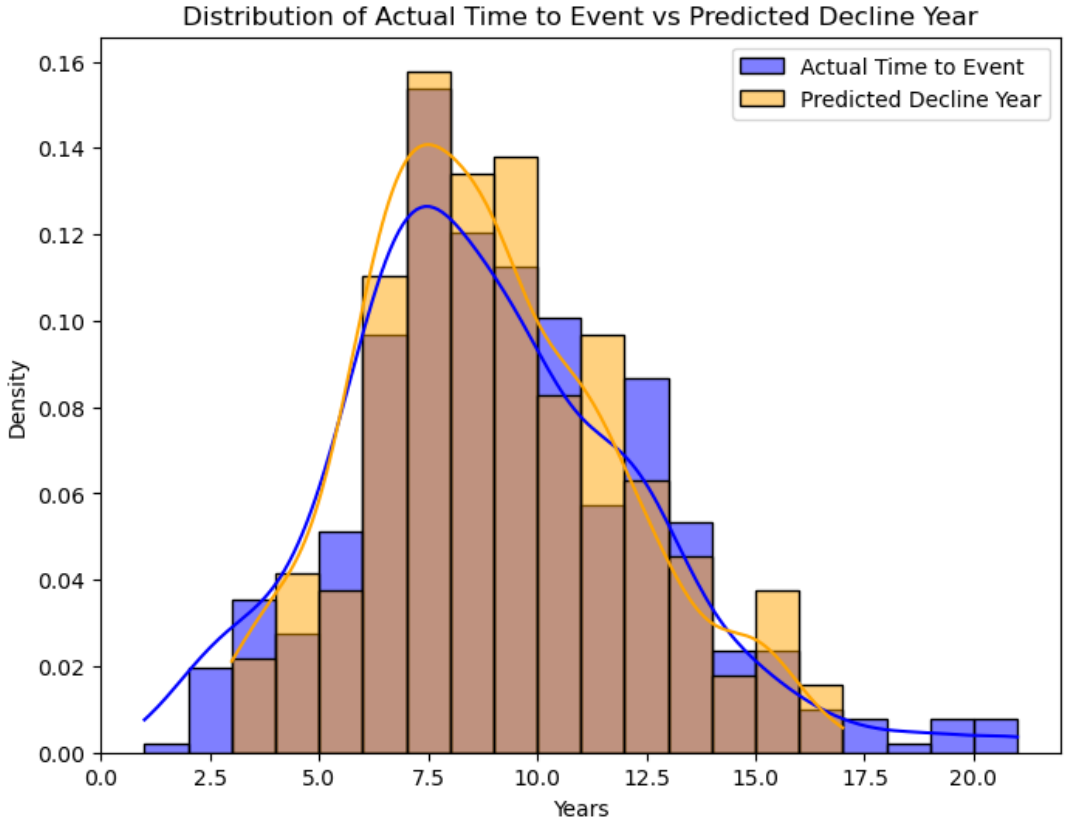


Position Players

Pitching: Distribution of Actual Time to Decline vs. Predicted Decline Year



Position Players: Distribution of Actual Time to Decline vs. Predicted Decline Year



Project Evaluation

Strengths + Weaknesses

Survival Analysis:

-Survival model can handle large datasets with numerous features and predict player decline events effectively.

-Minimal runtime → scalable with low computational costs.

-Results allowing teams to save money by identifying players with consistent performance and building rosters that avoid severe underperformance.

-Additional data (ie. player injury reports) would increase model performance

Valuation:

-The valuation model had poor predictive power and severe overfitting.

-Real salaries are based on factors beyond on-field player performance.

-Results demonstrate that many players are not being valuated properly, especially older players.

Conclusion

-Our goal: Find a way to determine if a player's salary is properly valued and consider aging players

-Survival analysis results showed that it is possible to predict accurately when a player will begin to decline

-Valuation results showed that players are not being valued properly, particularly those that are older and declining

-Overall, our results show that this kind of survival analysis combined with valuation could be useful for MLB decision makers to make better player acquisition decisions

Questions?

Thank you!

Appendix

Average MLB Career Length

According to the NYT, the average length of an MLB career for position players is 5.6 years and for pitchers it is hard to determine since they are so prone to injury and have a much more volatile career.

https://www.nytimes.com/2007/07/15/sports/baseball/15careers.html