October 16, 2022

Final Team Project: MLB Player Valuations Using Salary Predictions

**Business Understanding:**

* Baseball is a business just as much as it is a game; organizations constantly weigh the value of on-field performance with their desire to maximize profits and increase revenues. Each of the thirty MLB organizations have different priorities each season – some prefer profits, some prefer building organizational talent depth, and others will exhaust all resources (including prospects) season after season to remain in contention for a World Series Championship. These goals and preferences for each team are largely imposed by ownership, however, they typically correlate with the team’s location and market size. For example, it is foolish to expect the Milwaukee Brewers — who operate in the second smallest television market — to allocate resources the same way the San Francisco Giants do. The revenue disparity between these two organizations (and many others) is in the hundreds of millions. For that reason, teams like the Brewers and Giants must approach their roster construction process in very different ways. The Giants can afford to spend big (into the 300M+ range) whereas the Brewers operate on the margins and heavily rely on their development pipeline to progress players through their minor league system to the major leagues.
* Data capturing technologies and decade’s worth of statistical data has afforded MLB organizations the luxury of projecting player performance with relatively high confidence. While each team has different roster needs each year, all teams can place a monetary value on a player based on given performance metrics they value most. Our project is designed to build a model that can replicate the player valuation process teams use in free agency and arbitration salary negations using publicly accessible data. The data used is results based so the predictions provided indicate how much a player should be worth next season based on his previous season numbers. The modeling process must consider many nuances specific to the business of baseball and the way numbers might indicate a level of performance unrealistic from a visual evaluation. We considered many of these nuances, such as the growing difference between maximum and minimum annual salaries, throughout our data mining and model construction. Through EDA we discovered that there were many constraints we would need to apply to the data for our model to predict salaries with Chart, histogram

  Description automatically generatedmoderate to high accuracy. The large number of rookies, or players who played in the majors for only a brief stint, would reduce the variability in the train data and anchor the model on the low end of predictions. These players each earn league minimum; despite that minimum salary increasing over time, there were too many players with too few games played to keep them in the training data. Additionally, we discovered several interesting insights in looking at the changes in some important performance metrics from 1986-2022, despite only testing our data on 2002-2016. The latter range of years are the only ones that we were able to use salary information and team valuation information, so we trained on these ~5.5 thousand observations. The team valuation information was an important part of our data mining, as it included a variable in our model that could somewhat adjust for the sheer growth in annual salaries offered as baseball itself has grown into the 21st century. Overall, our model can be used by organizations to determine if a player is over or undervalued as a product of his performance, discover the best players per dollar, the diamonds in the rough that should be signed, the blooming stars, the maximum salary to offer a player, and many more actionable insights to aid in roster construction across the league despite differing organizational goals.

**Figure 1**

**Data Understanding:**

* Our data came from the Sean Lahman Baseball Database, an open-source collection of Major League Baseball statistics from 1871 to 2021.  The database contains separate tables for annual batting, pitching, and fielding statistics for MLB players along with salary information, award winners, team standings, team stats, managerial and player records, post-season stats, and more. We chose to utilize several of these tables, joining several times to get to our final data set that included the batting, fielding, awards, and salary tables from the Lahman database.
* The salary information for each player each year was accurate, however we needed to adjust for inflation to standardize salaries that have inherently grown over time. A million-dollar player in the 1980’s is far superior results-wise to a million-dollar player today. To make this adjustment in salaries to the present dollar value, we sourced inflationary data from the Federal Reserve Bank of Minneapolis and used the average consumer price index from 2021 [(Minneapolis Federal Reserve).](https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1913-) The 2022 CPI is rapidly changing as a result of the current global economy so we did not want to add further bias and inflate salaries above the normal range.
* An area of bias briefly addressed in the business understanding is the growing difference between minimum and maximum salaries over time, especially in the last decade. Figure 1 shows the nonlinear growth that has occurred, especially in maximum salaries since the early 2000’s. This positive bias reflects the impact of different ways teams view players; brand value, sponsorship potential, and advertising revenues can grow for a team simply by signing high profile players — their contracts often reflect this inherent value. With this in mind, we would expect our model to overestimate player salaries — especially in the early years of their career.
* To reduce this bias, we included team valuation data gathered using Statista and Forbes records [(MLB Valuation Data).](https://www.statista.com/statistics/193637/franchise-value-of-major-league-baseball-teams-in-2010/) The data accessible ranged from 2002 to 2022, slightly minimizing the range of years our model could be trained using. Figure 2 represents the average team valuation from Chart, histogram

  Description automatically generated2002 to 2022, demonstrating the exponential growth of baseball in the United States over that time frame.

**Figure 2**

**Data Preparation:**

* Based on an initial evaluation of the database, we chose to look at players who debuted after 1985, since the database only had salary data from 1985 to 2016. Another consideration was the introduction of a standardized strike zone in 1969. We then filtered our analysis to only hitters and used their batting and fielding statistics along with awards received. The decision to only build this model to predict hitter salaries was made to combat the difference in meaning several stats have for pitchers and hitters. For example, a high number of strikeouts for a pitcher typically correlates with a massive payday whereas a high number of strikeouts for a hitter typically results in a pay cut or release from a team’s roster.
* Our training data included players from 2002 to 2016, and our final model would be deployed on active players from a single team (San Francisco Giants) using results from the 2021 season. We conducted EDA with the full 1986-2021 data but limited training to 2002-2016 because it contained salary and team valuation data that would be of value when deploying our model on a specific team.
* We chose to look at only regular season data and filtered out players based on a number of criteria including the number of games played, total at-bats in a season, number of walks taken, and removed pitchers. Each player is identifiable in each table by a unique playerID, along with information on their age, and dates for their debut game and final game.
* Another reason we removed select hitters who played less than five games or very few batting stats is because often there are extenuating circumstances such as injuries or service time manipulation that limits the number of appearances a player could have in a year. MLB determines the salary bracket and ability to earn arbitration or free agency contracts based on this service time clock; organizations often attempt at manipulating this clock to maximize the years they will have control of a player and their teams “championship window.” Keeping these players in the train data would convince the model that playing below 50 games is normal when the true number should be between 130 and 160.
* For fielding metrics, we identified that certain players could have played in multiple positions in a single season, so we combined the metrics for them and created a new variable for the number of positions he played in to capture the added value a versatile player brings. Additionally, if a player was traded mid-season, we combined the metrics for his stints at each team to ensure that we only had one record per player per season.
* As previously mentioned, salaries we adjusted for inflation using the 2021 CPI. We took only the maximum salary a player earned in a season for those who were traded as the team receiving a player via trade rarely pays the full amount owed with the majority covered by the team trading a player.
* We also narrowed down the awards a player received based on their performance in that season, anticipating that it could impact their salaries. We chose to look at certain pertinent awards based on our knowledge of the sport and which awards were most valued or carried the most prestige, such as the Most Valuable Player award. Subsequently, we joined the individual tables into a single data frame and the data unit was all the information about a player, his batting and fielding performance, salary, and awards in a particular season.
* To capture detailed insights based on domain knowledge of the sport, we created calculated variables for advanced metrics such as Strikeout and Walk Percentages, double plays per at bat; both are widely used metrics that can be used to judge player performance over the course of a season.
* Finally, we added the team valuations for the teams in our dataset as an additional column, which resulted in a final dataset of 57 variables, which are explained in the data dictionary in Appendix A.

**Exploratory Data Analysis:**

* To identify any linear relationships in our data, we computed correlation matrices to narrow the number of parameters that would be included in our modeling. The first correlation matrix captured the relationship between the advanced stats, followed by an additional matrix on all regular stats (Refer to Appendix B for these intermediate correlation matrices). Any collinear variables were removed from our modeling. After testing the correlation between variables independently, we then combined the remaining variables from both the matrices to check the correlation among them, removing any remaining collinear variables (Figure 3).

**Figure 3**

Chart

Description automatically generated

* Certain variables had a strong linear relationship with each other, such as batting average with batting-average-on-balls-in-play, runs with at-bats, etc. Including these pairs in our models would introduce the issue of multicollinearity in our models, reducing the accuracy of our estimates. Much of the variable selection and row reduction processes were done through visually analyzing the tables of various queries and implementing the insights.

**Modeling:**

* We approached the core data mining task as a regression problem since our target variable and most of our features were numerical variables. Additionally, a regression task would not require standardizing our features and would retain the value of some of our advanced stats. We evaluated multiple models as outlined below.
* For our first choice, we used a **Linear Regression** model to understand the underlying relationships in the data and the impact of certain variables on the player’s salary. We attempted to do a LASSO analysis which did not prove to be helpful, as outlined in Appendix C. For our next model choice, we implemented a **Regression Tree** to account for the size and noisiness of our dataset. Compared to linear regression, which attempts to find the best fitted line for the data, while a tree-based approach uses if-else rules based on thresholds that explain the variance in the data, instead of relying on linearity in the data. To build upon the results from the regression tree we decided to utilize an ensemble learning method, **Random Forest**. As mentioned above, we were not able to identify the key features in our data through a LASSO analysis, which was resolved using Random Forest as it combines results from multiple trees, each of which are trained on a different set of features. Interpreting the variable importance plot was key to arriving at the best feature set.
* For each of the above models, we initially trained them on all our features to set a baseline for performance. Subsequently, based on domain knowledge of the most pertinent stats to quantify a player’s performance, we implemented the models again to estimate the improvement in prediction. Finally, we built another model which included features derived from the EDA done above, along with the valuation data for a team over the years to compensate for the large positive bias introduced due to a nonlinear increase in player salaries in the last decade.
* The best model predicts salaries that we can use to determine the value of a player relative to the salary he earned in that season. Teams can use this information for several different business use cases, including decisions on whether an extension should be offered, the maximum salary they will offer, how different performance metrics lead to higher salaries, and more. While we ultimately deployed the model on one team’s hitters in the 2021 season, the model can easily be applied to any team’s roster and afford them the same value and insights.

**Evaluation:**

* To evaluate all the models outlined above, we employed a **k-fold cross-validation** to ensure that we made the best use of all our training data. We trained the models on nine folds and had one test fold that acted as a holdout sample on which we appraised the model performance.
* The metric we arrived on was the **root mean square error**, an appropriate choice for a regression task. As we were predicting salaries, it made the most business sense to compute the deviations from actual salaries in the same unit (dollars). To see the RMSE of each fold, please refer to Figure 6 in Appendix D.
* Based on the average RMSE for each model over the 10 folds, we found that the Random Forest model built based on the EDA which accounted for the team valuations had the best improvement from the baseline model. Please refer to the variable importance plot and the changing errors against the number of trees for this model seen in Figure 7 and Figure 8 respectively in Appendix D.
* Using our model to determine player value and the appropriate salary to offer can benefit an organization in several situations. Teams can rely on their intuitive valuation based on the organizations perceived value — on-field or otherwise — or utilize historic data and predictive tools to ensure they do not overpay or wrongly derive expected value. While the latter approach fails to account for brand equity and potential revenues a player can bring, it takes all player contracts and the corresponding performance to make an accurate determination of a players financial worth. Each season, the front office and several players engage in negotiations over perceived worth; our model affords an indisputable dollar amount that players and their agents will struggle to defend against. The model is effective for players who have accrued six years of service time and are entering free agency just as it is effective for players early in their careers who have performed well and could receive an extension before hitting the open market. It is also effective in arbitration hearings, where players and the team put forth their salary demands, and an independent arbiter determines which side has the more convincing case. Overall, the model can help teams make more informed decisions, leading to saving money and effective roster building habits.

**Deployment:**

* We selected a single team to deploy our model on — the 2021 San Francisco Giants — and predicted salaries for hitters active at the end of the 2021 season. Using these predictions based on the players’ stats for the 2021 season, we can estimate his projected value to help the team in contract negotiations, i.e., if a player is highly valued, they could offer a more lucrative deal than his present contract to lock him down or release players who are overpaid but underperforming to manage the team roster and continue building for future success.
* We concluded that the change in a team’s franchise value was able to act as a proxy for the nonlinear growth in salaries over the last decade, but not entirely. There was still a positive bias when predicting salaries that we were unable to eliminate. A possible solution would be to include variables that capture the brand value of a player such as his popularity or existing sponsorships.
* We are ignoring the fundamental aspect of a team sport by looking at player performances as individual and self-contained, not accounting for the impact of team dynamics and player relationships on performance. Further, ballpark factors can influence the stats a player is able to accumulate over the course of a season. Coors Field in Colorado is notorious for being a hitter-friendly ballpark due to the elevation. Coors Field is consistently ranked as the most-hitter friendly ballpark based on the number of home runs and other offensive measures compared to the rest of the MLB ballparks.
* Some ethical factors that should be considered is the impact that a predicted salary should have in negotiations. Each team is obligated to offer a fair wage to players as there is more to baseball than winning games. Often families of MLB players are from low-income classes and the earnings a player makes over the course of his career are shared among many family members. It is also important to consider that baseball is a business and player production directly determines the wage a player can and will earn. There are countless stories of teams who have “no sense of loyalty” and will release a player quickly because they failed to produce. Teams must manage these concerns to avoid being branded as an organization that does not care about their players as this will lead to fewer free agents interested in signing with that organization.
* The risk associated with this kind of valuation comes from quantifying a player’s value based solely on statistics rather than a more holistic approach, since we could be undervaluing a player that could perform for the team in the future. Also, there are much more informative information available to MLB teams that can be used to predict performance on a more individual scale. For example, Trackman is a system that captures ball-flight metrics on each pitch and hit that happens over the course of a game. The implementation and development of these technologies has changed the way teams develop their players. Teams can discover in above average spin rate (the number of revolutions a ball has on its way through the strike zone) for a specific pitch, for example, and design everything about the way that player executes around utilizing that strength. These ball flight metrics would allow teams a more effective way to forecast future ability rather than basing salary numbers off of historic performance.
* In our deployment case with the 2021 Giants, we interestingly discovered that the model accounted for aging star players as having salaries above their perceived worth. Evan Longoria, Brandon Belt, and Buster Posey have all been in the league since the early 2010’s and received their current contracts several years ago. While at one point their performance warranted that salary, they are now overvalued by several millions of dollars. For some teams, this can be the difference in operating at a profit or deficit. The Giants are a larger market team so this does not impact them as much; still our model shows other players, like Wilmer Flores, who are undervalued and well deserving of the contract he currently plays on and will continue to positively impact the Giants probability of winning games.

**Bibliography:**

Federal Reserve Bank of Minneapolis. “Consumer Price Index, 1913.” *Consumer Price Index, 1913-*, Federal Reserve Bank of Minneapolis, https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1913-.

Forbes, Editors. “The Business of Baseball.” *Forbes*, Forbes Magazine, https://www.forbes.com/mlb-valuations/list/.

Gough, Christina. “MLB Franchise Values US 2022.” *Statista*, 27 May 2022, https://www.statista.com/statistics/193637/franchise-value-of-major-league-baseball-teams-in-2010/.

Lahman, Sean. “Sean 'Lahman' Baseball Database.” *The Comprehensive R Archive Network*, Comprehensive R Archive Network (CRAN), 26 Apr. 2022, https://cran.r-project.org/web/packages/Lahman/index.html.

**Appendix**

**Team Member Contribution:**

Ethan Shear: Business Understanding, Domain Expertise, Data Cleaning, EDA, Report & Presentation

Aayush Chordia: Data Cleaning, Modelling, Evaluation, Report & Presentation

Deepika Karan: EDA, Modelling, Evaluation, Report & Presentation

Haoyang Dong: Business Understanding, Preliminary Data Cleaning

Anh Tuong: Data understanding

**Appendix A: Data Dictionary**

|  |  |
| --- | --- |
| **yearID** | Year/Season |
| **teamID** | 3 letter code for MLB team |
| **playerID** | Unique identifier for a player |
| **Games** | Number of games played in the season |
| **AtBats** | Number of plate appearances minus sacrifice hits, walks, or hit by pitches |
| **Runs** | Number of runs the player hit that season |
| **Hits** | Number of times a hitter hits the ball and reaches base safely |
| **Doubles** | Number of times the hitter has safely reached second base |
| **Triples** | Number of times the hitter has safely reached third base |
| **HomeRun** | Number of home runs the player hit that season |
| **RunsBattedIn** | Number of run-scoring hits the hitter makes |
| **StolenBases** | Number of times baserunner takes a base to which he isn't entitled |
| **CaughtStealing** | Number of times a runner attempts to steal but is tagged out before reaching base |
| **Walk** | The number of base-on-balls; four balls before a batter is retired equates to a walk |
| **StrikeOut** | Number of strike outs the player had that season |
| **IntentionalWalk** | A strategic move made by the pitcher to award a walk without attempting to retire a hitter |
| **HitByPitch** | Number of times a hitter is struck by a pitched ball without swinging at it |
| **SacrificeHit** | Number of times a player successfully advances a runner at least one base with a hit |
| **SacrificeFly** | Number of times a batter hits a fly-ball out to the outfield/foul territory that allows a runner to score |
| **GroundIntoDoublePlay** | Number of times a player hits a ground ball that results in multiple outs on the bases |
| **BattingAvg** | Player's hits divided by his total at-bats |
| **PlateApps** | Number of completed batting turns |
| **TotalBases** | Number of bases gained by a batter through his hits |
| **SluggingPct** | Total number of bases a player records per at-bat |
| **OnBasePerc** | How frequently a hitter reaches base per plate appearance |
| **OPS** | Addition of on-base percentage and slugging percentage |
| **BABIP** | Measures a player's batting average exclusively on balls hit into the field of play |
| **GamesStarted** | Number of games the player has started in the season |
| **InnOuts** | The total number of outs by which a defender is in the game on defense |
| **PutOuts** | Number of times a fielder physically completes an out himself |
| **Assists** | Number of times a fielder touches the ball before a putout is recorded by another fielder |
| **Error** | Number of times a fielder fails to convert an out on a play that an average fielder should have made |
| **DoublePlay** | Number of times two offensive players are ruled out within the same play |
| **nPOS** | Number of positions the player has fielded in the season |
| **Salary** | Salary earned (not adjusted for inflation) |
| **GG** | If the player won the Golden Glove award that season |
| **MVP** | If the player won the Most Valuable Player award that season |
| **SilvSlug** | If the player won the Sliver Slugger award that season |
| **HankAaron** | If the player won the Hank Aaron award that season |
| **RookOfYear** | The total number of awards the player won that season |
| **totAwards** | Total awards won by the player in the season |
| **AdjustedSal** | Present dollar value of salary earned |
| **debutYear** | Year of the player's debut season |
| **Age** | Age of the player in years |
| **ExtraBaseHits** | Number of hits that are not singles |
| **Singles** | The number of time the hitter has reached the first base safely |
| **xbhPerhits** | Extra base hits per hit |
| **WalkPerc** | Walks by the total number of plate appearances |
| **StrikeOutperc** | Strikeouts by the total number of plate appearances |
| **KtoBB** | How many strikeouts a pitcher records for each walk he allows |
| **IsolatedPower** | Measures the raw power of a hitter by taking only extra-base hits and the type of extra-base hit into account |
| **FieldingPerc** | Total number of putouts and assists by a defender, divided by the total number of chances (putouts, assists and errors) |
| **HomerPerAB** | Total homeruns divided by at-bats |
| **DoublePlayPerAB** | Total double plays divided by at-bats |
| **sacPerAB** | Total number of sacrifice hits divided by number of at bats |
| **SqAge** | Square of the player's age |
| **YearSinceDebut** | Years since the player's debut |
| **Valuation** | Franchise value of the team in millions (USD) |

**Appendix B: Correlation Matrices**

**A picture containing treemap chart

Description automatically generatedFigure 4: Correlation Matrix for Advanced Stats**

**Chart

Description automatically generatedFigure 5: Correlation Matrix for Standard Stats**

**Appendix C: Additional Modeling Tasks**

* **KNN:** Our initial approach to evaluate K Nearest Neighbors as a potential model choice was to gauge whether we could assign players with similar statistics a similar salary, through naturally forming groups in our dataset, based on player ability. However, as we started working on the modeling we discovered that our data was not suitable for this model as our variables had varying scales. Standardizing some of the advanced stats of a player led to the variable losing its meaning and hence its predictive power. Therefore, we decided KNN was not an appropriate choice for our data mining task.
* To narrow down our features to the most significant ones, we employed a **LASSO** analysis on a linear regression model. Using cross validation to test various values of lambda, our results were not meaningful since our lambda values were very large and had a very large range, which meant that all our variables were being penalized heavily. We then tried our own values for lambda which resulted in almost all our variables having a nonzero coefficient, rendering the analysis not helpful to our data.

**Appendix D: Model Evaluation**

**Chart

Description automatically generatedFigure 6: RMSE for Various Models Across Folds**

**Table

Description automatically generatedFigure 7: Variable Importance Based on Best Random Forest Model**

**Chart, histogram

Description automatically generatedFigure 8: Changing Errors Against the Number of Trees:**