UNIVERSITY OF
CALGARY

Predicting MLB Pitcher Salaries

Alex Havers
Srikant Joshi
Femi Onafalujo
Andrey Sivyakov

4/11/2018

Introduction

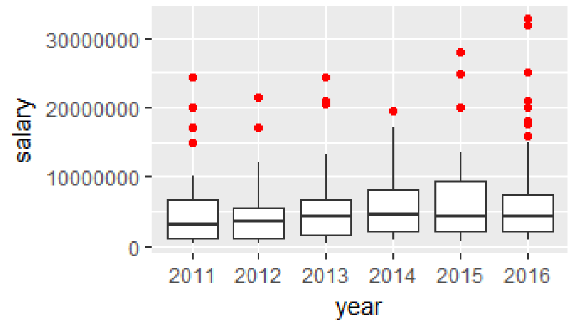
Major League Baseball (MLB) salaries have been extensively researched over the last few decades. In 2003, Michael Lewis authored "Moneyball: The Art of Winning an Unfair Game", which exposed the inner workings of team economics in baseball to the public. It revealed a movement towards using advanced statistical analysis to determine the value of players to maximize the win/cost relationship. Player salaries are typically the largest cost for Major League Baseball teams. Players often get paid millions of dollars because of their unique skills, as well as the profitability of the teams that employ them. The most important factor that determines how much a player's salary is his on-field production. The better one plays, the more he will be paid. The goal of this project is to develop models that explain and predict the relationship between skill and pay.

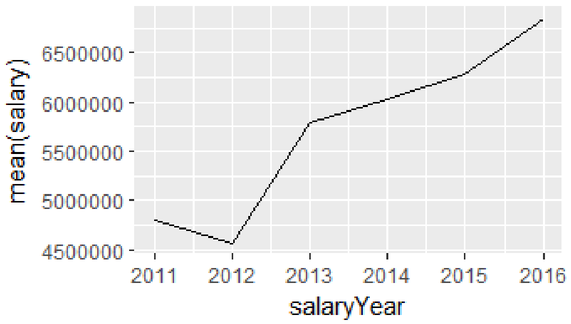
The data utilized was mainly gleaned from the Lahman database[1], a public repository of baseball statistics and other valuable information about players and teams. These statistics are largely results-based metrics that attempt to describe player performance. This database was merged with more granular contract information gathered from Cot's Contracts[2]. For further explanation of the variables within, please visit the reference urls.

The project initially focused on the broad question: can MLB salary be effectively predicted from our existing dataset? Through the analysis process, this question was pared down to predicting salary of **free agent pitchers post-2010.** This paring was due to the original question being too broad; it included attempting to predict position players' salary, which was a separate dataset. The analysis necessary would have therefore doubled. Next, we decided to focus on free agent contracts, which is one of several available contract types. This focus was necessary because teams have some control over players on other contract types, while a free agent can sign with whoever he wants for whatever price he wants. This is the most valuable contract to predict for MLB teams, as they are the costliest of all contract types. Allowing for a team to accurately predict their largest cost is incredibly valuable for understanding their projected bottom-line. Last, contract years were filtered to post-2010. In the MLB, the Competitive Bargaining Agreements (CBA) that govern labour/management relations are negotiated every 5 years. Often, these agreements adjust rules involving free agent compensation. So, limiting our data set to only 2 collective bargaining agreements[3] should minimize noise attributable to any free agent value affected by the CBA.

Exploratory data analysis

The graphs below show distribution of pitcher FA salaries (left) and the trend of mean pitcher FA salaries (right) in the MLB in 2011-16.

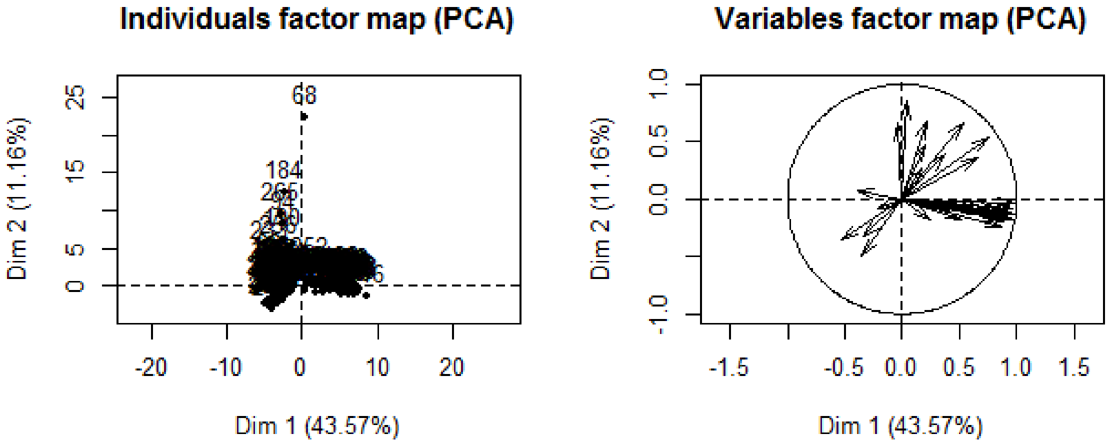




[1] http://www.seanlahman.com/baseball-archive/statistics/
[2] https://legacy.baseballprospectus.com/compensation/cots/
[3] http://legacy.baseballprospectus.com/compensation/cots/league-info/cba-history/

Since there is a trend in the pitcher FA mean salaries in the observed period, an independent 2-group t-test was utilized to deduce if the true mean of the train data, pitcher FA salaries in 2011-15, is not statistically different from the true mean of the test data, pitcher FA salaries in 2016. The p-value of the test was >0.05, which supports the null hypothesis that the true means of the two subsets of the data are not different from 0. It is also noticeable on the left graph that there are several outliers in each year, which may influence performance of the predictive models. This makes some intuitive sense, as the 'best' FA in any given year are typically subject to a bidding war, creating upward pressure on those players' contracts.

Principal Component Analysis (PCA) was then used to further explore the data. An important insight provided by PCA was that variables pull data points in three opposite directions within the first two primary components. Data points are grouped in two clusters almost symmetrically around 0-load of the first PC.

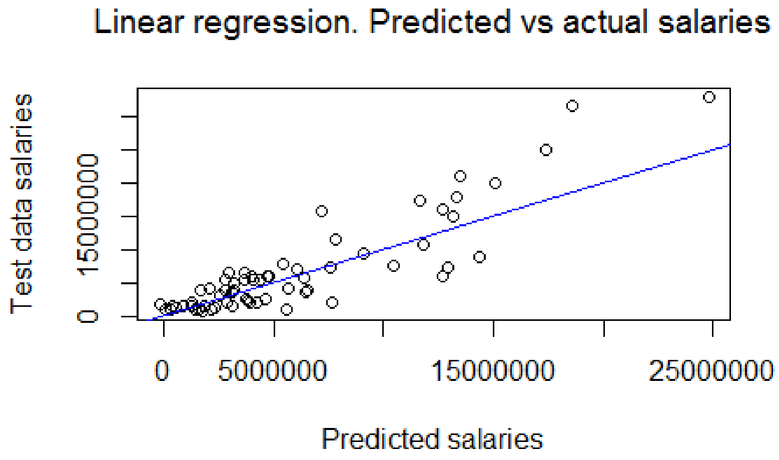


Predictive models

In this project, the data is high-dimensional. This can be a problem, because a high number of variables and relatively low number of observations may lead to models that fit well on test data but perform poorly on an independent data set. The following statistical learning approaches allow to overcome this issue: linear regression with stepwise model selection, principal components regression (PCR), and random forest. We also used boosting, another tree-based method, to see if its performance will be better than performance of random forest. The train data set included data for 2011-15, and data for 2016 was used as the test data set.

Linear regression with stepwise model selection

First, pitcher FA salaries were regressed on the rest of the variables (player performance) in the train data set. As expected, the model was over-fit – very few predictors showed statistical significance in predicting the salaries. Stepwise model selection allowed us to create a regression model that included only 11 variables (out of a possible 36), but still indicated relatively good fit; the drop in adjusted R-squared was only 2%. Then model diagnostics were analysed to determine if the selected model satisfied the standard model assumptions. Diagnostics included: normality of error terms, correlation of residuals over time, heteroscedasticity and multicollineraity. One predictor (GS) that showed multicollineraity with other predictors was removed. This resulted in the final regression model of 10 predictors. The graph below shows how the predicted values compare to the actual test data. There seems to be a decrease in prediction accuracy with a requisite increase in salary value. This phenomenon may be explained by the presence of outliers as previously mentioned. Root mean squared error (RMSE) of the linear regression model was ~$3.7 million.



From a birds-eye view, this model makes sense. The final variables are obviously linked to player skill. Strikeouts (SO) are the most ideal outcome for a pitching event and are only controlled by the pitcher's skill. This is because a strikeout is an event between only a pitcher and a hitter, unlike other types of outs that involve fielders' skill. Therefore, teams should be looking to maximize this stat. Walks (BB) and Hit by Pitch (HBP) are similar - they are also controlled by pitcher/hitter skill but are a negative value to the pitcher. However, they are a cumulative statistic; worse pitchers will not accumulate as many as better pitchers because it is preferable to pitch the better pitchers more. So, there is a positive relationship. Games (G) shows the split between starters and relievers; good starters get paid more but pitch few games (though face more batters total). There is a large increase in salary around ~30 G, the typical amount a full-time starter pitches. Past that point, salary drops then slowly slopes upward, showing the increase in salary amongst more trusted relievers. Wins (W) and Saves (SV) are statistics that are positively correlated to pitcher skill and are some of the most commonly cited statistics when analysing a player. Therefore, the data selected makes some sense.

PCR

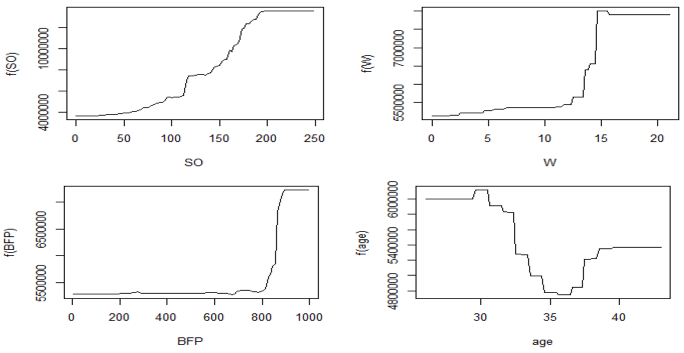
Another approach suitable for analysis of high dimensional data is PCR. According to the validation plot (in our R script), the PCR model with the first 22 primary components will have the lowest RMSE. The RMSE of the PCR model on the test data was ~$3.61 million, which is like the linear regression model. The main disadvantage of PCR is that the results are hard to interpret. Since the primary components were used in regression, it is unclear how predictors influenced the model output.

Random forest

Random forest (RF) is a tree-based method that de-correlates the trees by considering only a limited and random sample of predictors at every split. It allows for a decrease in influence of a single predictor, or a group of predictors, highly correlated with the response variable. The latter improves predictive power, but unlike trees, which are easy to interpret, the results of RF are not very intuitive. According to the RF model, two predictors – strikeouts (SO) and saves (SV) have the highest importance in terms of % increase in mean squared error (MSE). This is consistent with the selected linear regression model, which also included SO and SV. The RMSE of the RF model on the test data was ~$3.99 million, noticeably higher than that of linear regression and PCR.

Boosting

Boosting was our final method. Unlike the other models, it does not reduce dimensionality. Boosting is a tree-like method which develops trees using information from previously grown trees, which improves model accuracy and fit. As expected, boosting had a better prediction accuracy compared to RF. The RMSE of the model was ~$3.57 million, which is also slightly better than linear regression and PCR. Another advantage of boosting is better interpretability. The graph below shows salary as a function of four variables, SO, BFP, W, and IPouts, with the highest relative influence. In the BFP (Batters Faced by Pitchers) graph, there is a clear threshold of ~850 that leads to a salary increase. This makes intuitive sense, as starting pitchers face more batters as a function of pitching more innings, and typically get paid more as a result. This threshold, then, is likely the threshold that separates starters and relievers. A similar threshold is present within the W graph, as starters are more likely to garner wins than relivers. SO has a smoother slope, which is interesting because we could expect a similar threshold as W and BFP. However, there do exist relivers that have a greater number of SO per BFP. These relievers, who are therefore displaying a higher level of skill than their peers, may earn more as a result, and smooth the relationship. Last, we can see that younger free agents typically make a higher salary than their older compatriots, though there is an increase in salary as age moves past 36 years. This is likely because players that can last to that point in their career are at a higher base skill level than their peers, who often retire in their mid thirties.



Model Outcome Analysis

The RMSE, though seemingly high, is within an understandable margin. Free agent contracts in baseball have no cap, and as a result are often very lucrative for players. Wins Above Replacement (WAR) is a metric that attempts to assign an overall (pitching, hitting and pitching) value to a player, standardized through era, position, and ballpark. It is the best way we currently have to compare players against each other within an individual metric. 0 is a 'replacement' player, or the next player a team would be able to obtain externally to the major league roster. 2 is an average player, 3 above average, and so on. From 2011-2015, teams paid between $7.4 million and $9.6 million per WAR[4]. So, an average player could expect to get a contract worth $14.8 million/y and $19.2 million/y. The RMSE in this analysis is between $3.5 and $4 million, or roughly 0.5 WAR. Though this RMSE may seem large at first, it is a marginal amount when it comes to paying for real-world skill. This is not to say that baseball management is willing to spend millions of dollars without thought, but rather that an incremental $4 million dollars will not allow for a large pitching upgrade.

[4] https://www.fangraphs.com/blogs/the-recent-history-of-free-agent-pricing/

Limitations

The main limitation in this analysis is the data. A public dataset was utilized, with many performance variables, but there are performance variables that this dataset does not include. Specifically, physical skill data is now being gathered and analysed in the MLB but is not included in the Lahman database. There are also qualitative factors that weren't considered as they are not gathered or curated into an easy to access dataset. The market situation, for example, is a huge factor in free agent decisions from year to year. If a player has 3 teams that value his skill set, he will be more relatively valuable than a player with 2 teams that value his skill set. Therefore, in different years, the same skill set may be worth a different salary. This may help explain the outliers with high salary values in our dataset. Indeed, when removing outliers, the models return a 30% better RMSE; around $2.5 million. However, the outliers should not be removed, as they are valid data points, and a result of supply and demand for elite players. Pragmatically, teams would not be incentivized to remove top-tier players to improve their prediction model; these are the most important players to predict. Instead, it would be wise to devise a market 'factor'; a categorical variable that can be applied to each player season to attempt to adjust for his market. With this method, the model may be trained to understand market factors rather than ignore them.

The data is also limited to only the salary in the first year of a free agent's contract. Players can sign single or multi-year contracts. The latter increases risk for the team. To compensate, the annual average value (AAV) of a contract will decrease given an increase in years. For example, a player may be able to earn 20 million dollars on a 1-year contract, but if he was to sign a 5-year contract, he may only make $15 million per year. Of course, the guaranteed $75 million is well above the $20 million of the 1-year contract, so the player should always take the longer deal. In this analysis, only the first year is considered, and contract length is ignored. This may understate longer term contract signee value, or overstate short term contract signee value.

Conclusion

Management can glean various insights from this analysis. Each model predicted salary to a reasonable degree, and by adding one year of data to the training set there was an improvement of the RMSE in each model of ~$300k. Each subsequent year of data should therefore improve the models. Management can use these models as a baseline in conjunction with market context and other qualitative factors to determine the salary their preferred free agent may command. This will help them estimate their bottom line and plan more efficiently. EDA revealed that higher-value free agents are more difficult to model with performance data, which is very relevant to management decisions. Boosting displayed the effect of various factors to salary, specifically bounding the difference between starters and relievers. Overall, it was surprising to note that public data estimated salary so well. As teams utilize proprietary and other data, this model can be supplemented to create stronger predictive capabilities.