Predicting MLB Regular Season and Postseason Success

Based on Roster Construction and Payroll Allocation

Parker Gibbons

May 7, 2024

Professor Brad Hartlaub

STAT 306

Introduction and Literature Review

Ever since the 2002 Oakland Athletics made history with their great success despite the departures of key players and a constricting budget, many smaller market Major League Baseball franchises have attempted to employ similar strategies in attempts to create similar success. Even with teams like the New York Yankees and New York Mets that have payrolls that exceed 300 million dollars, teams such as the Oakland Athletics and Cleveland Guardians with payrolls under 100 million dollars are forced to compete for the same players and on-field success despite an obvious financial disadvantage. However, this disadvantage has not stopped small market teams from achieving high levels of success. The 2015 Kansas City Royals (payroll of 126 million dollars) and 2016 Cleveland Guardians (payroll of 105 million dollars) are two examples of highly successful teams despite having a below average league payroll for that season.

Seeing the success of small market teams, especially against the large market teams, inspired me to explore how these small market teams construct their rosters and allocate their limited payroll. In my research I found work done by Lopez et al (2011) that used data from 1995 to 2009 to predict winning, runs scored, and playoff appearance using offensive, defensive, and pitching statistics. Through creating linear regression models, it was concluded that ERA and OPS, two widely used statistics to evaluate pitchers and hitters respectively, can be used to predict wins and playoff appearance. Somberg et al (2012) and Moses (2019) used data from multiple seasons to examine the relationship between team payroll and the probability of playing in the postseason. Through the use of probit and linear regression models, it was concluded by both works that teams with higher payrolls have a higher likelihood of regular season success and advancing to the postseason. Tao et al (2015) looked at how the dispersion of a team's payroll impacts team success and concluded that the ways in which teams allocate their payroll does not impact team success as much as total payroll. This suggests that the specific ways in which teams allocate their payroll does not matter as much as simply having a high payroll.

Despite these past findings, none of these works took into account the difference in market sizes between teams. All of the works explored how success and postseason appearance can be predicted in Major League Baseball but all of the models included the entirety of the teams in the league without factoring in the possibility that a team like the New York Yankees employs methods for team success that differ from those of a team like the Oakland Athletics. Thus, I wanted to explore whether teams of different market sizes allocate their payroll and construct their roster to put forth the best possible chance of success. This leads to my main research question: What is the ideal roster construction and payroll allocation to win games, in the regular season and in the playoffs? I also want to ask whether teams of different market sizes spend money on different parts of their roster (specific position, infield, outfield, starting pitching rotation, bullpen, bench) and are there certain methods of this that work better for teams of larger markets versus smaller markets? Also do teams of similar market size rely more heavily on certain parts of their roster than others for value and success?

To begin this process, I will classify all 30 Major League Baseball teams into classes based on their market size via clustering. I will then further examine these clusters to see how teams in each market size construct their roster and allocate their payroll. Using these clusters, I will attempt to predict regular season success and postseason success and determine whether there are certain methods of roster construction that work better for teams of larger markets versus smaller markets. Lastly, I will examine some case studies of teams more in depth to dive deeper into their specific roster construction.

Data

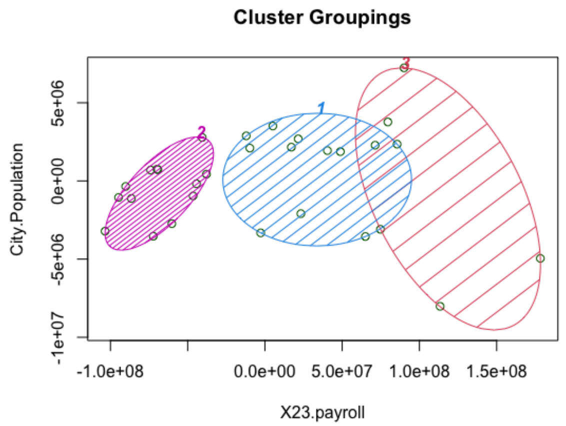
The data used for this analysis was obtained mostly from Spotrac (team financial data) and Baseball Reference (team statistics). In addition, the scope of the data is the last four seasons of Major League Baseball from 2020 to 2023. For the clustering part of this analysis, I will consider the variables total payroll from the past four seasons, city population, media market size, team valuation, and attendance per game from the past three seasons (there were no fans allowed during the 2020 COVID shortened season). However, for the analysis I will consider the response variables of regular season win percentage and an indicator variable to specify when the team was eliminated in the postseason (0 = team did not make playoffs, 1 = team made the Wild Card Round, 2 = team made the Divisional Series, 3 = team made the League Championship Series, 4 = team made the World Series, 5 = team won the World Series). The explanatory variables I will consider include total payroll, payroll by position, competitive balance tax threshold (CBT), number of players on arbitration, acquired as free agents, acquired by trade, or re-signed or extended, number of rookies, money spent in free agency, money spent re-signing or extending players, average age of batters, and average age of pitchers.

Methodology

K-means Clustering

In order to classify each of the teams in Major League Baseball into classes corresponding to their market size, I will use k-means clustering. K-means clustering is an unsupervised non-linear algorithm that clusters data based on their similarity to one another. It uses a pre-specified number of clusters that I determined using the elbow method which determined that three clusters would be most appropriate. These clusters will help later see what variables matter more to each market class and identify similarities/differences between the different market sizes.

Figure 1 - Cluster Groupings of MLB Team Market Sizes



Notes: Cluster 2 indicates teams in the small market class, cluster 1 indicates teams in the medium market class, cluster 3 indicates teams in the large market class. City population is on the Y-axis, 2023 team payroll is on the X-axis.

Figure 2 - MLB Team Market Size Cluster Means

	X23.payroll	X22.payroll	X21.payroll	X20.payroll	City.Population	media.market.size	Team.valuation	X23.attendence.g	X22.attendence.g	X21.attendence.g
1	195238927	170387988	150722702	68943348	6575736	2911750	2636.667	32797.83	31270.17	21761.08
2	97365321	92387753	83894109	42645849	3620110	1686462	1445.231	21713.08	18353.31	13474.31
3	272799899	252125385	209895987	93088715	12325086	5079000	3825.000	39515.40	36608.80	24652.80

Cluster means:

Notes: Cluster 2 indicates teams in the small market class, cluster 1 indicates teams in the medium market class, cluster 3 indicates teams in the large market class.

Table 1 - Teams in each MLB Team Market Size

Market Size	# of teams	Teams
Large	5	LAD, NYM, NYY, PHI, SDP
Medium	12	ATL, BOS, CHC, CHW, COL, HOU, LAA, MIN, SFG, STL, TEX, TOR
Small	13	ARI, BAL, CIN, CLE, DET, KCR, MIA, MIL, OAK, PIT, SEA, TBR, WSH

Notes: Team abbreviation guide https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Baseball/Team_abbreviations

The clustering process resulted in fairly good accuracy. In addition to passing the eye test, three clear classes of market sizes were distinguished with clear differences in cluster means and no overlap. This obviously supports the past findings that differences in market size in Major League Baseball exist and that both large and small market sizes are very distinguishable. A potential limitation of this process could be that specific team financials, including team revenue, are unavailable. However, I believe this process was successful and sets up the next step of creating models to determine which financial metrics impact success for each market classification.

Creating Models

With the three clusters of market sizes, I will now attempt to predict both regular season winning percentage and overall season success, measured by how far the team went (ranges from not making the playoffs to winning the World Series).

Multiple Linear Regression

To begin the modeling process, I started with Multiple Linear Regression which extends linear regression to incorporate multiple independent variables which allows for the modeling of complex relationships between the dependent variable and multiple predictors variables. The process assumes a linear relationship between the dependent variable and each predictor variable, with coefficients representing the strength and direction of these relationships. The coefficients are estimated using techniques such as least squares estimation which aims to minimize the difference between observed and predicted values across all variables. I believe using MLR will illuminate the significant variables for each market size and remove any multicollinearity. In the case of each market size, both the methods of backwards elimination and best subsets were used to find models with the most variability explained.

Large Market MLR Models

- Model to predict winning percentage (Adjusted R-squared = 0.6159)
- Model to predict playoff results (Adjusted R-squared = 0.7936)

Table 2 - Large market model to predict winning percentage

Variable	Sign	Significant at 99%?
Catcher payroll	-	No
Pitcher payroll	+	Yes
CBT space	-	Yes
# players resigned/extended	+	No
# free agents signed	-	No
Money spent on free agents	-	No
Average pitcher age	-	Yes

Table 3 - Large market model to predict playoff results

Variable	Sign	Significant at 99%?
Infield payroll	-	Yes

Outfield payroll	+	No
Pitcher payroll	+	Yes
CBT space	-	Yes
# players resigned/extended	+	Yes
Money spent on free agents	-	Yes
Average batter age	+	No
# of rookies	+	No

These models suggest that successful large market teams (in the regular season and postseason) noticeably spend more money on pitchers and spend over the CBT threshold. This makes sense because large market teams have the payroll to spend large amounts of money on players and in many cases, outbid teams with lower payrolls. Both models explained the majority of variability in the dependent variable with the model predicting playoff results explaining almost 80% of the variability. Thus, both of these models performed fairly well and set a high bar to be improved upon later.

Medium Market MLR Models

- Model to predict winning percentage (Adjusted R-squared = 0.4862)
- Model to predict playoff results (Adjusted R-squared = 0.3698)

Table 4 - Medium market model to predict winning percentage

Variable	Sign	Significant at 99%?
Total payroll	-	No
Catcher payroll	+	No
Pitcher payroll	+	Yes
# players resigned/extended	-	No
Money spent on players resigned/extended	+	No
Average batter age	+	No

# of rookies	-	No

Table 5 - Medium market model to predict playoff results

Variable	Sign	Significant at 99%?
Infield payroll	+	No
Outfield payroll	-	Yes
Pitcher payroll	+	No
CBT space	-	No
Average pitcher age	-	No
# of rookies	-	No

These models suggest that successful medium market teams (in the regular season and postseason) noticeably spend more money on pitchers and less money on outfielders. Another noticeable similarity is that both models show that successful medium market teams have less rookies. This could mean that these teams with more constricted payrolls believe that pitchers bring enough value to the team to warrant being spent on more rather than outfielders. In addition, these teams could be relying on more experienced players rather than their young talent. Both of these models explained a decently less amount of variability than the large market models which could be attributed to the medium market cluster being the most murky of the three. The spectrum of market size is noticeably heavier at the tails (large and small), leaving the medium market cluster to be the least distinctive of the three.

Small Market MLR Models

- Model to predict winning percentage (Adjusted R-squared = 0.4719)
- Model to predict playoff results (Adjusted R-squared = 0.2783)

Table 6 - Small market model to predict winning percentage

Variable	Sign	Significant at 99%?
Total payroll	-	Yes
Infield payroll	+	Yes
Outfield payroll	+	Yes
Pitcher payroll	+	No
# players acquired	+	No

via trade		
# of rookies	-	Yes

Table 7 - Small market model to predict playoff results

Variable	Sign	Significant at 99%?
Total payroll	-	Yes
Infield payroll	+	No
Outfield payroll	+	No
# players acquired via trade	+	No
Money spent on players resigned/extended	+	No
# of rookies	-	Yes

These models suggest that successful small market teams (in the regular season and postseason) noticeably spend less money and rely less on their young talent. This is interesting because these teams do not spend much money to spend on new players anyway with their highly constricted payroll thus leaving them with the players that they draft and trade for. The successful small market teams are the ones in which the players acquired without needing to spend money succeed rather than players that are given expensive contracts. The model to predict regular season success does a decent job at explaining variability which could speak to methods of roster construction becoming more commonly used by small market teams to succeed in the regular season. However, the model to predict playoff results does not explain a ton of variability which could speak to the unlikely chance of small market teams even having success in the postseason, much less making it.

MARS (Multivariate Adaptive Regression Splines)

While the multiple linear regression models did an overall decent job at explaining variability in both regular season and postseason success, I wanted to go more in-depth with modeling and explore some possible non-linear relationships and interaction among the predictor variables. This leads to MARS being the subsequent method to be used. Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression technique that builds upon multiple linear regression by allowing for more flexible and complex relationships between predictors and the dependent variable. MARS does this by partitioning the feature space into smaller regions and fits linear models within each region which allows for piecewise linear relationships. It iteratively adds basis functions, called "splines," to capture non-linearities and interactions between predictors, which offers improved predictive accuracy compared to traditional linear regression. MARS also selects the most relevant predictors and their interactions automatically, thus providing a more interpretable and efficient model for capturing complex data patterns.

For the large market teams, both MARS models to predict winning percentage and postseason success explained a decently less amount of variability (Generalized R-squared = 0.37 and 0.49 respectively). This could have resulted from overfitting or the lack of data since only four MLB seasons were considered. However, for medium market teams both MARS models improved upon the variability explained by the predictors (Generalized R-squared = 0.54 and 0.50 respectively). These models emphasized the increased spending on pitchers and infielders, and having more rookies on the team. For small market teams, the MARS model to explain regular season win percentage explained noticeably less variability (Generalized R-squared = 0.33) but the model to explain playoff results explained noticeably more variability (Generalized R-squared = 0.50). This model emphasized the presence of rookies on the team which contrasts with the findings from MLR and suggest that the small market teams that perform well in the postseason have many rookies on the roster. The variable importance plots for each MARS model are in Appendix 1. Moreover, while the MARS models did not improve the models for every market, it did capture nonlinear relationships that improved models for medium and small market teams. This could suggest that while large market teams have more distinct methods of roster construction and payroll allocation, there are methods and tendencies shared by medium and small market teams which could speak to teams with lower payrolls employing similar strategies.

Lasso Regression

Lasso regression (Least Absolute Shrinkage and Selection Operator) extends from multiple linear regression and MARS by introducing a regularization term that penalizes the absolute size of regression coefficients. This penalty term encourages sparsity which effectively shrinks coefficients towards zero and selects only the most relevant predictors. Thus Lasso regression serves as a feature selection technique and offers a more interpretable model compared to multiple linear regression and MARS, which may include all predictors despite how relevant each of them are. Lasso regression addresses overfitting concerns and provides a flexible tool by providing model simplicity and predictive accuracy as well as its ability for handling high-dimensional data with potentially collinear predictors. For all the models, Alpha is equal to one due to the tuning of alpha not causing much difference in the variability explained by the models. Thus, the penalty term is maximized.

Large Market Lasso Regression Models

- Model to predict winning percentage (Adjusted R-squared = 0.5968206, MSE =

0.001448442)
- Model to predict playoff results (Adjusted R-squared = 0.7935581, MSE = 0.2805708)

Figure 3 - Large Market Lasso model estimates to predict win percentage

```
9 x 1 sparse Matrix of class "dgCMatrix"
                                  s0
(Intercept)            2.060569e+00
(Intercept)            .
Catcher..             -2.042740e-09
Pitcher..              1.507723e-09
CBT.Space             -1.641846e-09
X..resigned.extended   9.208100e-03
X..free.agents.signed -3.514664e-03
free.agent...spent    -2.106910e-10
Pitcher.average.age   -5.651487e-02
```

Figure 4 - Large Market Lasso model estimates to predict playoff results

```
10 x 1 sparse Matrix of class "dgCMatrix"
                                 s0
(Intercept)          -2.648478e+01
(Intercept)           .
Infield..            -4.078325e-08
Outfield..            1.503476e-08
Pitcher..             2.166252e-08
CBT.Space            -2.439006e-08
X..resigned.extended  4.465613e-01
free.agent...spent   -5.758465e-09
Batter.Average.Age    7.995957e-01
X..Rookies            1.290884e-01
```

The Lasso regression models performed better than MARS models but slightly worse than the MLR models in both cases. However the results did corroborate with the previous models by emphasizing the spending of large market teams over the CBT threshold and the retention of players. Interestingly, both models suggest that successful large market teams do not spend much money on free agents which in conjunction with the retention of players could suggest that successful teams have a foundation of players that do not change greatly over time.

Medium Market Lasso Regression Models
- Model to predict winning percentage (Adjusted R-squared = 0.4800229, MSE = 0.002569442)
- Model to predict playoff results (Adjusted R-squared = 0.3697544, MSE = 1.190254)

Figure 5 - Medium Market Lasso model estimates to predict win percentage

```
9 x 1 sparse Matrix of class "dgCMatrix"
                                       s0
(Intercept)                  2.183039e-01
(Intercept)                  .
Total.Payroll               -3.780750e-10
Catcher..                    1.654607e-09
Pitcher..                    1.751823e-09
X..resigned.extended        -2.369588e-03
resigned.extended...spent    4.118146e-10
Batter.Average.Age           1.147561e-02
X..Rookies                  -4.216140e-03
```

Figure 6 - Medium Market Lasso model estimates to predict playoff results

```
8 x 1 sparse Matrix of class "dgCMatrix"
                                 s0
(Intercept)           1.365142e+01
(Intercept)           .
Infield..             1.231462e-08
Outfield..           -3.410783e-08
Pitcher..             2.748706e-08
CBT.Space            -1.111786e-08
Pitcher.average.age  -4.270898e-01
X..Rookies           -7.373363e-02
```

The lasso regression models performed very similarly to the MLR models for medium market teams and worse than the MARS models. These results did, however, corroborate with the previous findings that successful medium market teams spend more money on pitchers and less on outfielders. Both models also suggest that successful medium market teams do not have many rookies on the roster. The lesser performance of the Lasso models could result from sensitivity to the scale of the features since Lasso regression penalizes the absolute size of regression coefficients. It could also result from the relationships in the data being more non-linear in which MARS would work better.

Small Market Lasso Regression Models

- Model to predict winning percentage (Adjusted R-squared = 0.471865, MSE = 0.003646168)
- Model to predict playoff results (Adjusted R-squared = 0.2783112, MSE = 0.6224189)

Figure 7 - Small Market Lasso model estimates to predict win percentage

```
8 x 1 sparse Matrix of class "dgCMatrix"
                                         s0
(Intercept)                    5.065636e-01
(Intercept)                    .
Total.Payroll                 -2.015350e-09
Infield..                      5.186027e-09
Outfield..                     3.563054e-09
Pitcher..                      2.666940e-09
X..players.aquired.via.trade   4.080143e-03
X..Rookies                    -4.867100e-03
```

Figure 8 - Small Market Lasso model estimates to predict playoff results

```
8 x 1 sparse Matrix of class "dgCMatrix"
                                         s0
(Intercept)                    1.346161e+00
(Intercept)                    .
Total.Payroll                 -1.340847e-08
Infield..                      2.134629e-08
Outfield..                     2.394206e-08
X..players.aquired.via.trade   5.125372e-02
resigned.extended...spent      3.899888e-09
X..Rookies                    -6.358581e-02
```

As the medium market lasso models did, the small market lasso regression models performed very similarly to the MLR models as the adjusted r-squared were very similar. As the MLR models did, the model to predict regular season success explained more variability than the model predicting playoff success. However, both models emphasized the lack of spending, payroll, and interestingly enough the lack of rookies which is in line with the MLR results. Both models also emphasize spending more heavily on infielders and outfielder than other positions and relying on players acquired through trade.

Discussion and Conclusions

Through the use of K-means clustering, it is clear that the teams in Major League Baseball can be classified in different clusters according to market size. The large market teams noticeably have higher payrolls and reside in areas with higher population and larger media markets. In contrast, the smaller market teams noticeably have lower payrolls and reside in areas with lower population and smaller media markets. While not having a ton of similarities, the medium market teams clearly do not fall on either end and thus distinguish themselves as being in the middle. Having these clusters allowed for each market to be explored separately as to how they construct their rosters and allocate their payroll.

For large market teams, the Lasso and Multiple Linear Regression models worked very well and explained the majority of variability with the winning percentage model almost having 80% of the variability explained in both cases. The MLR model helped determine which variables were significant and the Lasso Regression model built off this to further determine how significant each of the variables were. Through this, it could be concluded that successful large markets teams take advantage of their large payroll to spend. However, this money is usually spent on retaining current players rather than new players. When they spend their money, these large market teams spend heavily on pitchers and extend their players long-term.

For medium market teams, the MARS models clearly worked the best by explaining the most variability in both winning percentage and playoff success which suggests that the data for medium market teams involved non-linear relationships that were more easily captured by MARS models. These models conveyed that successful small market teams spend their payroll mainly on pitchers and infielders while also valuing their young talent to contribute. This could illuminate as to where these medium market teams feel their money is best spent as they deal with payrolls that are lower than their large market counterparts.

Small market teams had the most interesting case because the MLR and Lasso Regression models worked the best in terms of predicting regular season success but the MARS model noticeably worked better for predicting playoff success. With regards to regular season success, it is shown that small market teams do not rely on their young talent for success and instead invest in infielders and outfielders while also using trades to build their roster. This makes intuitive sense because many non-successful small market teams have many rookies simply because they have no one else to play. Thus, the more successful small market teams invest what money they have into infielders and outfielders while also using trades to acquire players that they want.

Future Work

This work showed that MLB teams of different market sizes rely on different elements of roster construction to win and have success in the playoffs. Obviously, an easy way to improve on this work as a whole is to have data from more MLB seasons. Specifically, the clustering process could be improved with access to data on variables such as team revenue, contract lengths, merchandise sales and ticket prices. These clustering results can also be compared to alternative clustering methods such as hierarchical clustering. The analysis of this work could be improved by using Elastic Net regression which is a regularization technique that combines the penalties of Lasso and Ridge regression. This would allow the limitations of the Lasso regression models to be built off of. By incorporating both lasso) and ridge penalties, elastic net regression can deal with collinearity more effectively while maintaining the feature selection capabilities of Lasso regression.

In addition, this analysis could be intensified by looking at each position specifically rather than positional groups to identify what specific positions are valued and spent on by teams of different market sizes. This could also include the value produced and previously produced by these players to see if teams are adjusting their roster construction strategies over time based on the value per dollar that they are getting from certain positions. This analysis could possibly employ the use of Bayesian regression modeling which is a statistical approach that incorporates prior knowledge or beliefs about the parameters of a regression model. This would allow for the estimation of uncertainty in the model parameters by providing posterior distributions instead of point estimates. This approach could allow for probabilistic predictions to be made which could be used to make predictions as to how MLB teams of different market sizes will approach large transactional events (offseason, winter meetings, trade deadline).

References

2023 MLB team position performance by wins above replacement. Baseball. (n.d.). https://www.baseball-reference.com/leagues/team_compare.cgi?request=1&year=2023&lg=MLB&stat=WAR

Author links open overlay panelYu-Li Tao a, a, b, 1, 2, AbstractThis study examines the relation between compensation and performance in Major League Baseball (MLB), Arellano, M., Blundell, R., Depken, C. A., Jane, W. J., Levine, D. I., Mohan, N., Wang, M. C., Anderson, T. W., Avrutin, B. M., Berri, D. J., Bloom, M., & Bond, S. R. (2015, October 23). Compensation and performance in Major League Baseball: Evidence from salary dispersion and team performance. International Review of Economics & Finance. https://www.sciencedirect.com/science/article/pii/S1059056015001872

Lewis, J. (2024, February 23). Major pro and college sports teams ranked by Market Size. Sports Media Watch. https://www.sportsmediawatch.com/nba-market-size-nfl-mlb-nhl-nielsen-ratings/

Lopez, J. (n.d.). Multiple linear regression viewpoints volume 37 • number 2 ... http://www.glmj.org/archives/MLRV_2011_37_2.pdf

Somberg, A. K., & Sommers, P. M. (2012, May 17). Payrolls and playoff probabilities in Major League Baseball - Atlantic Economic Journal. SpringerLink. https://link.springer.com/article/10.1007/s11293-012-9317-3

Spotrac. (n.d.). 2024 MLB team salary Cap Tracker. spotrac.com. https://www.spotrac.com/mlb/payroll/_/year/2024/sort/cap_total2

Team, D. (2022, March 25). Lasso and Ridge regression in python tutorial. https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression

Top 40 metro areas - MLB teams. (n.d.).

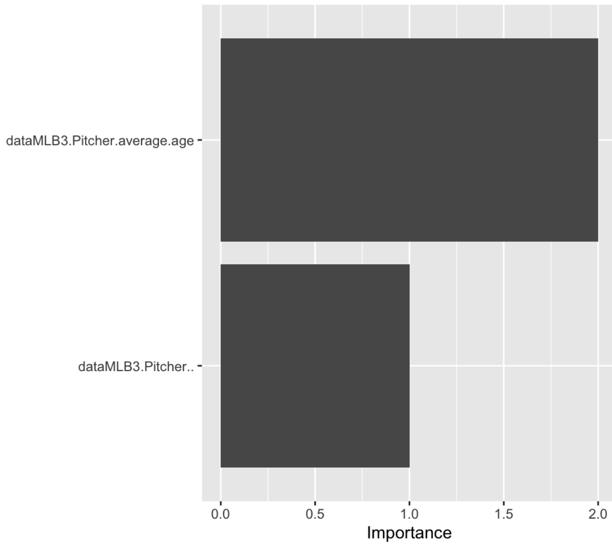
https://www.census.gov/content/dam/Census/newsroom/stories/baseball/stories-baseball.pdf

Wikimedia Foundation. (2024, April 17). Forbes list of the most valuable MLB clubs. Wikipedia.

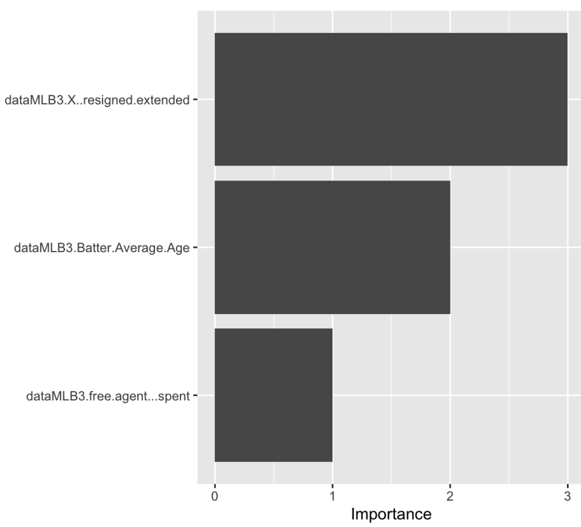
https://en.wikipedia.org/wiki/Forbes_list_of_the_most_valuable_MLB_clubs

Appendix 1: MARS Variable Importance Plots

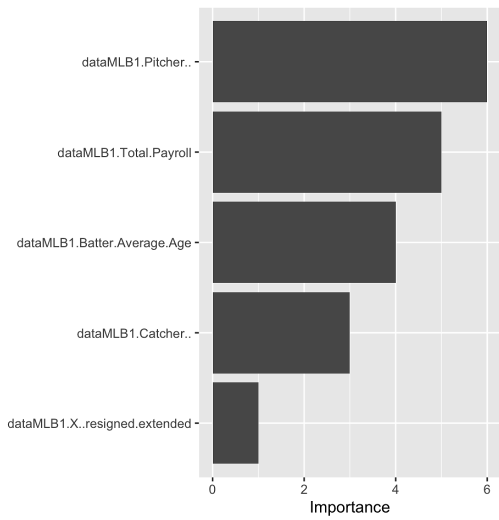
Model 1: Predicting Win Percentage for Large Market Teams



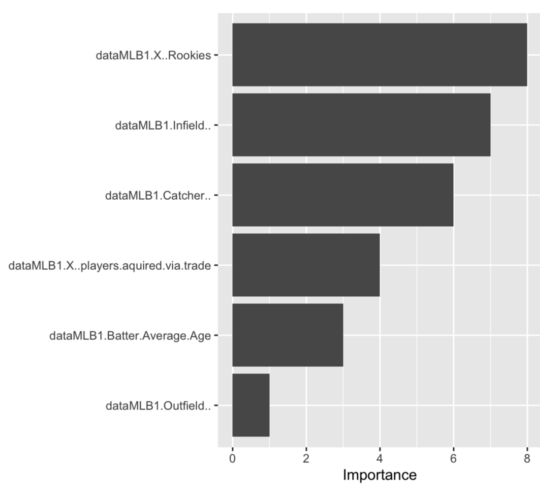
Model 2: Predicting Playoff Success for Large Market Teams



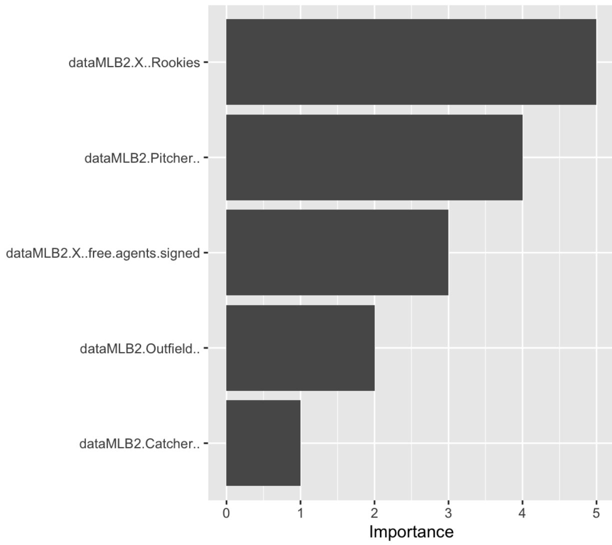
Model 3: Predicting Win Percentage for Medium Market Teams



Model 4: Predicting Playoff Success for Medium Market Teams



Model 5: Predicting Win Percentage for Small Market Teams


dataMLB2.X..Rookies
dataMLB2.Pitcher..
dataMLB2.X..free.agents.signed
dataMLB2.Outfield..
dataMLB2.Catcher..
0
1
2
3
4
5
Importance