***Introduction to Data Mining***

***DATS 6103***

***Final Project***

*Milestone -3-*

**Data Exploration and Preparation**

## Author

## Michael Wegan G45192209

## Zheying Chen G43061235

## Weiwei Li G27494462

## Xueli Zhou G29955935

## Introduction

Milestones 3 is a stepping stone to complete your semester Final Project. The goals of this millstone deliverable is it to help you progressively achieve your end goal while you understand, practice and master each building block of your final product.

## Explore and Prepare Data:

Please answer the questions below and provide the following Project Proposal:

1. **Describe the Datasets used**

We used 4 data sets. We started with the batting.csv file from SeanLahman.com, which has annual batting statistics for every major league baseball player since 1911. We added some variables that we calculated from the data present in the batting.csv. Our analysis requires 3 other data sets that we were able to find from SeanLahman.com and baseball-reference.com. Those data included annual team performance for each player’s team, MVP winners, and wins over replacement (WAR) data, which is a measure of how many wins a player is worth for his team as compared to a replacement player.

1. **List your Data exploration results**

Figures are included as attached files. We elaborate on these results below.

1. **Describe in details you Data Cleaning Process**

The data cleaning process was arduous. We used dplyr to join the base data set, batting.csv, to the other team and WAR data. This required us to build new columns upon which to base these joins. The ID’s were consistent between the batting, MVP, and team data, as they all came from the same source – SeanLahman.com. However, we identified 402 records in the batting data through anti\_join() with differing player IDs from the WAR data. We were able to use adist() to match most of these records from the batting data to the appropriate records in the WAR data. Still, for 10 players, we had to manually change the ID’s in the WAR data set to accomplish the join in full.

The joins also had an unexpected consequence. There are many players that played for 2 teams within the same year. For these individuals, there are 2 records in the batting data and 2 records in the WAR data. When joining these data, the results have 4 records for these players within the same year if the join is based on a column we built consisting of the player ID pasted together with the year. We solved this issue by including the team ID in the pasted ID value, following this format “TEAM\_PLAYERID\_YEAR”. The join then worked as expected, returning 2 records for these players in the years that they played for 2 teams.

After cleaning, our “games\_played” column had a bimodal distribution. We could see that MVP winning pitchers clustered around 30 games played whereas other non-pitcher MVPs clustered around 145 games played. The minimum of the higher mode is around 100. Constraining for 100 games played made sense because this allowed for us to remove pitchers from our MVP data set and retained all non-pitchers. This was important because we are attempting to predict MVP award winners based on offensive metrics. Pitchers that receive the MVP award are not evaluated on their offensive metrics, but their pitching metrics. There is little offensive data for pitchers and no pitching data for non-pitchers. Coupled with the fact that only 13% of MVP winners are pitchers, our approach to prediction is that it is more likely for non-pitchers to win based on offensive metrics, meaning that our prediction of a non-pitcher is valid for 87% of MVP awards.

Below are questions that you should investigate, answer and report back in as many details as possible(written summaries, pictures, code, snapshots, ….:)

1. **You have already gathered your data and explored it , no what steps have you followed from the lectures notes to mine and prepare your data ( Data cleaning, integration, reduction, transformation,…) ? for instance, have you found missing values and what are they?, what descriptive statistics have you done on the data set ex: changing data types,…? Elaborate.**

The full data set contained 135 columns, many of which are of little value in predicting MVP award winners. We significantly pared down the number of variables to those that we believe will have some impact on voters. Generally, these variables measure offensive output in some way. For instance, slugging percentage measures the ability of a player to hit for power, not just singles.

Some of our variables were derived from other variables in the data set. One of the simpler examples is batting average, which is the number of hits (H) divided by at bats (AB). All our derived metrics were based on commonly accepted baseball statistics and we found the calculations at SeanLahman.com, baseball-reference.com, fangraphs.com, and retrosheet.org. One of the derived variables, weighted on-base average (wOBA), required a significant number of the other variables to calculate. A problem that we encountered was that some of these data points were not collected for all the players throughout the data set. Sacrifice flies (SF) and intentional base on-balls (IBB) were not recorded in early major league baseball, but are required for wOBA calculation. We therefore had to model the using a LASSO regression in the glmnet package for R. We used the half of the records with SF and IBB data to train the model and half to test it. From this process, we predicted the missing SF and IBB which we then used to calculate the wOBA for these players.

As we are attempting to mine the data to predict MVPs, we are comparing the actual known MVP winners to their non-MVP winning counterparts. Major League Baseball has two separate leagues, the National League (NL) and the American League (AL). Each league has an MVP for every year. This means that we need to compare the NL MVP of a certain year only to other NL players from that same year and vice versa for AL MVPs. We decided to scale the majority of the variables by the max value for that variable for each year and each league. For example, if the max number of hits for a the NL in 1985 was 215, we would have divided all number of hits values for that year and league by 215. This gives us a relative scale of how that player performed against all his counterparts for that year and league.

Finally, there are infrequent occurrences in which a pitcher wins the MVP. This occurred at a rate of approximately 13%. We decided to drop these events from the analysis, as we are focusing on offensive output data. MVP pitchers are evaluated on the strength of their pitching performance not offensive metrics. Coupled with the fact that pitchers generally have very poor offensive output these events in the data set would have clouded our conclusions.

We also changed the data types of some of the columns to logical where it made sense to do so. For example, we converted the MVP column from a factor length of 1 (Most Valuable Player) with most of the values as “NA” to a TRUE/FALSE logical data type. This will make coding our analysis easier.

1. **What features have you built into the data set?**

We have added slugging percentage, batting average, weighted on-base percentage (wOBA), and on-base percentage. These are commonly used baseball statistics measuring the offensive output of an individual player.

The calculation for wOBA can be found here: <http://www.fangraphs.com/library/offense/woba/>. The formula uses linear weights adjusted by year. The reason for adjusting by year is to scale the formula to the offensive/hitting environment for that year so that it is comparable across years. There have been eras of very different offensive production in baseball and failure to scale this metric by year would not allow for valid comparison across years.

The slugging percentage calculation can be found here: <http://www.baseball-reference.com/bullpen/slugging_percentage>. The weights are constant and represent the number of bases awarded for each type of hit (e.g., a double is worth 2 because a player got a hit with a value of 2 bases) divided by the number of at bats (AB) or opportunities to hit. This measures the ability of a player to hit for power as it values extra-base hits over singles. It is an value of the number of bases a player is expected to receive from each at bat.

The OPS is a measure of a player’s ability to both get on base by any means of which there are three: get a hit, get a walk, or get hit by a pitch. It is calculated as: (hit by pitch + walks + hits)/(at-bats + sacrifice flies + hit by pitch + walks).

Batting average is simply the rate at which a player gets a hit. It is calculated by dividing the number of hits by the number of at-bats.

Adding these metrics was important because we believe they offer more complete information about a batter’s performance and offensive output than do the basic metrics such as hits or runs-batted-in. The “hits” column, for example, is a count, and is highly correlated with at-bats. It is not valid to make an inference about the relative superiority of two players just based on hits, especially if they have very different numbers of at-bats. Because hits is a function, to a degree, of the number of at-bats, the player with the greater number of opportunities to get a hit will have the most hits. This is not a measure of which is the superior hitter, but of which player had the most opportunities to hit. The derived metrics are scaled in a way that give an expected value per at-bat, so the absolute number of at-bats is totally uncorrelated with the measurement, allowing us to compare the players under fair criteria not based on the number of opportunities, but their success relative to those opportunities. We also believe that the voters who "elect” an MVP value these metrics and that they will be significant in predicting MVPs in the future.

1. **What insights have you found working with the data?**

After scaling, our data have multiple distributions by variable. Interestingly, the MVP winners do not always rank the highest in these metrics. There are possibly interactions that are important between variables that can explain why someone who has the best offensive output in some categories does not win. Also, as baseball writers vote on the MVP, there is inherent bias toward some players with “good” personalities and away from others that are considered to have a “bad” attitude. There is not a reliable means to model this perception of players, but I believe it does play into voters’ decisions. This may complicate the ability of our model to predict MVP winners. Additionally, there are many metrics that are correlated which will affect the final decisions we make regarding which variables to include in the model building process.

1. **Have you changed or refocused your initial question after exploring the data? How so?**

Yes. Initially, we were going to look at both MVP and Cy Young winners. However, these would be two totally separate processes as MVP is more likely to be driven by offensive output and the Cy Young award is solely based on pitching metrics. This would require us to essentially double our effort in data accrual, cleaning and research. With the amount of time this process took just for the offensive data for the MVP data mining, we did not feel it was appropriate to continue with the Cy Young award modeling. Also, in terms of an academic exercise, data mining the Cy Young data would have been exactly the same process as the MVP data, so there would have been little to gain in terms of technical skills building.

1. **What problems have you run into as you have explored the data?**

Most of the data sets were from the same source and so they had the same player ID field. The wins over replacement (WAR) data came from a different source, though, and the player ID field was slightly different. We performed an anti\_join() with the dplyr package in R to figure out which player IDs did not match between the data sets. Once we found the 402 records in the WAR data that did not match any records in the SeanLahman.com data, we ran a fuzzy matching process with adist(). This adequately matched the majority of the unmatched records. For 10 players, we had to manually change player IDs to achieve a successful join between the data sets.

Also complicating the join was the fact that many players had played for multiple teams in a single year. When joining, the result was that these players would have twice the number of records than they should have. For example, if a player played on team A and B, the join would have A to A, A to B, B to B, and B to A – all possible combinations of the two teams. We had to establish an additional key field which specified the player ID, year, and team pasted together as a character string. This allowed for the proper joining process. We also found 3 duplicated records in the SeanLahman.com data which we removed.

Through exploration after cleaning, we realized that our initial censoring criteria were inadequate. We started by including all individuals with >= 50 at-bats in a season. Upon investigating the boxplots it was clear that MVPs play more games and have more at-bats than the average non-MVP player. It is not appropriate to compare the MVP winners to the players with less opportunity who are not really competitive for the MVP award. A non-pitcher with less than 100 games played has never won the MVP. With that discovery, we decided to make our censoring criteria more stringent, requiring players to have played in at least 101 games in a season. This gives us a more realistic picture of who the MVPs competed against for the award.

Finally, we realized that we were accidentally including more pitcher MVPs that we had previously thought. We want to exclude this group of MVP winners because they won due to their pitching performance, not their offensive performance. Our goal is to predict the winner of the MVP when it is awarded based on offensive output, which occurs 87% of the time. Through the exploration of the games played distribution from the MVP winners, we identified this issue. It was a bimodal distribution with one mode around 35 games played and one around 145 games played. Pitchers generally don’t appear in more than 50 games and no non-pitcher will win the MVP with so few games played. We therefore knew that all MVP winners clustered around the first mode were pitchers. By setting our criteria at 101 games played, we not only select for a better representation of the everyday players competing for the MVP, but also only for position players (i.e., non-pitcher) who won the MVP. We dropped all records from the year/league when pitchers won the MVP because we are comparing MVP winners to their same year/league cohort. Since we removed the MVP for that year (because they were a pitcher), we had no MVP to compare to the rest of the players for that year/league, making those records useless.

1. **Does the available data limit your ability to answer your question? Do you need to gather additional data ?**

I don’t believe so. We still have >14400 records over 105 years with 167 MVP winners. This is greatly reduced from the 103,000 records we started with, but I think the variables we have are sufficient and the number of records ample.

1. **Has the scope of the question changed ?**

Yes. We will no longer be mining the Cy Young award winner data. There is little to gain from it academically, as the process would be analogous to the MVP data mining process and it would take an inordinate amount of time.

1. **What tools did you use and for what specific purpose ?**

We used dplyr and glmnet packages in R. dplyr was useful for the process of joining the various data sets and summarizing by variable as we explored the data. glmnet provided the lasso, ridge, and glmnet functions for modeling the SF and IBB missing data. We used ggplot for visualization. The rest of our work was done in the base R packages.

We also tapped into the treasure trove of baseball statistics information online. Baseball-reference.com and fangraphs.com were especially helpful in choosing which variables, such as wOBA and on-base percentage, we should derive and why. These sites also provided rationale for which means of calculation were superior for those metrics which have multiple ways to calculate and the annual linear weights for wOBA. We also contacted the data managers at these sights to find additional data and ask questions about the statistics and data.

1. **Provide the code that you used to clean your data set. Comment the code to explain how it transforms the original data set.**

This is included as a separate file.

1. **Are you using more than one data set?**

We have integrated multiple data sets (batting, award winners, team data, WAR). The process for joining them is described above and the data files and outputs are included in the uploaded zip file.

1. **If so how did you bring the datasets together?**

Using different join functions in dplyr. This is explained at length above.

1. **What models do you think you might use based on initial experience with the data.**

We aim to figure out a model that can help us to predict the. We have a significant number of variables, and the target is a 1-D value. The model will be classification-based.

**Avoid the curse of dimensionality**

We have collected and created as much data as are possible to build our model. Then the most important thing we need to do is avoid the curse of dimensionality.

1.Since we have normalized our data, we can use the Low Variance Filter to find that is there some variables has a small variance that make no sense for our prediction.

2. Some of these data have a high correlation to each other. We can use High Correlation Filter, to reduce part of them.

3. PCA is a good method. However, if we use this method, we will lose the explanations of data. Because, the new variables are not derived from the factual world.

4. We can also use the random forest to reduce dimensions. And this method can also be used to do classifications.

**Models that can be used for classification:**

* Use the decision tree (combine the random forest) to predict the answer.
* KNN (K-Nearest Neighbor)
* BBN (Bayesian Belief Network)
* SVM (Support Vector Machine)
* ANN (Artificial Neural Network)

1. **Elaborate and report on your data exploration process using R.**

We ran summaries, built histograms, and boxplots. There are some data that will need to be normalized if we use models that assume normal distributions. Most of our findings were intuitive – the MVPs tend to rank much higher in the derived categories that are scaled by at-bats. Looking at a non-standardized metric such as total hits is not a valid means for evaluating players relative to their counterparts, even when scaled by the max value for that year. This is because the mean and range of the number of at-bats for MVP winners is significantly higher than the rest of the population. Hits are naturally a function, to a degree, of the number of opportunities to hit. So to compare the number of hits between an MVP winner with 500-600 at-bats and another player with 50 at-bats is naturally biased in favor of the MVP winner and more a measure of opportunity rather than ability to hit. We therefore required a minimum of 101 games played in a season for inclusion in the analysis. The truth is that no non-pitcher that is not a regular, everyday player will win the MVP. No one that plays half of the season will win. The only position (i.e., non-pitcher) players that have a chance are the ones who play most games. Imposing the 101 games played restriction gives us a more realistic set of candidates to compare to the MVP winner.

We will most likely need to transform some of our variables, but how we do so is dependent upon the methods we ultimately use for mining these data. As we move ahead in the project to milestones 4 & 5 and begin to actually apply the models we will transform the variables as appropriate.

Some initial findings are that MVP winners have higher average values in the derived metrics, which is to be expected. These are the people considered to be the elite offensive players in the game. I was surprised at how much home runs are different between MVPs and non-MVPs. It appears that hitting for power is more highly valued by the voters than I had initially though. This means that an exceptional player that has a high on-base percentage and produces a lot of runs may lose to a player with inferior statistics but more home runs. It will be interesting to see how the data mining process weighs each of the variables and which will be considered important.

**From boxplots**

MVPs have better scores in :

l Played\_games, but have many outliers

l at\_bats

l runs\_scores, obvious distinct

l hits

l doubles

l home\_runs

l runs-batted

l IBB

l Strikeouts, but nonMVP have many outliers in this variable

l Total\_bases

l Slugging\_percent

l Team-rank &team wins , but have many outliers

l WAR

non-MVPs have close scores in:

l triples

l walks

l sacrifice\_flies

l onbase\_percentage

l Batting\_average

1. **What are some constraints R tool presented to you during data exploration?**

The only real constraint that we found was with the performance of adist(). We needed adist() to fuzzy match ID’s between two data sets so that we could join them. For the most part this worked. However, there were 10 individuals for which the matching process provided spurious results. We used the anti\_join() function in dplyr to figure out which records were not properly joining and then had to manually change their IDs to complete the join.

Otherwise, we were not constrained by any tools in R. We found dplyr especially helpful in joining the disparate data sets and managing data frames.

## Evaluation and Summary

Make your milestone summary main points here. Evaluate your findings so far. Draw some early potential conclusions.

The data cleaning process was more difficult than we thought it would be. We had fairly clean data (i.e., well-structured with definitions for each variable) so we thought the process would be straightforward. When we started to piece the different data sets together, we found that there was much more work than we originally anticipated. The joining of the data sets presented the most time-consuming problem in the cleaning process. This is most likely due to our inexperience with dplyr and learning to navigate through it. Once we became familiar with it, we were more quickly able to problems within the joins that needed to be addressed.

The fact that MVPs have higher summary values in most categories associated with offensive output is not surprising. We are surprised by the magnitude of the difference. One of the basic tenets of WAR is that baseball talent is normally distributed across the general population and that professionals exist in the upper tail. If that is the case, then the MVP winners are deep into that tail, representing an exceedingly small proportion of the professional baseball player population. Based on the magnitude of the differences we see, we expect that predicting the MVP will not be difficult because their performance is so far beyond that of the average player.

A temporal aspect to the model may be interesting to add. We hypothesize that earlier in the data set, the MVP may be selected by slightly different criteria than more recent selections. This is because the derived metrics such as WAR and OBP are relatively recent developments. Baseball statistics have become infinitely more complicated since the early 1990s, meaning that voters have had more and different information on which to base their MVP selections. Because these data were not available prior to the 1990s, we may find that the model needs to be adjusted depending upon the year. That would be an interesting finding and give some insight about changes in the evaluation of baseball players’ performance through time.

Another interesting finding from our exploration is that the number of games won by MVPs’ teams seems to be much higher than we would expect from random selection. This indicates that not only does a player need to excel in offensive output but also their team needs to do well. Simply having a great offensive year on a bad team is not enough to win, generally. Players generally have little control over the team for which they play so there is some stochasticity in the system. We found that MVPs were almost evenly split between teams that won their division and those that did not, but there are far more teams that lose the division than win. We would have expected many more MVPs to come from teams that did not win the division if occurring at the same rate as teams which won the division. Also, several teams were more likely than others to produce MVPs. Some of this may be because those teams generally win more than would be expected and therefore have a higher probability of producing an MVP. From these findings, we believe team performance is going to be a valuable factor in the analysis.