Abraham Choi, Christina Berardi, Edgar Delgado, Frieden Stone
Introduction to Data Science - Assignment 4

MoneyBall Using Multiple Linear Regression and K-Means

For this group project, we decided to go with the moneyball dataset and figuring out how to predict the target wins, for no reason other than that it seemed to be the most interesting. The raw dataset has a lot of missing values and attributes that may be irrelevant for this purpose, so we will be using RStudios to clean up the dataset. Upon completing the dataset cleanup, we will use Weka to run multiple linear regressions to satisfy the use of our supervised algorithm and K-means for our unsupervised algorithm.

The original data set used includes statistics on 2277 baseball teams, split into offensive and defensive attributes. It includes offensive attributes such as base hits by batters, doubles by batters, triples by batters, home runs by batters, walks by batters, batters hit by pitch (get a free base), strikeouts by batters, caught stealing, and stolen bases. There are less defensive attributes, including strikeouts by pitchers, walks allowed, hits allowed, errors, double plays and home runs allowed. Each of these attributes can either have a negative or positive impact on wins.

The data set used for the analysis did not have all the attributes and teams as the original one. Through the exploratory data analysis, which will be explained further below, the final data set ended with 1889 teams. Also, attributes with overwhelmingly empty information were taken out.

The goal of the two models was to predict the number of wins for each team. This is determined through one unsupervised technique, k-means clustering, and one supervised, multiple linear regression. For the data cleaning, we used the R programming language, and for the algorithms we used Weka, a tool for machine learning practices. Through the analysis of both negative and positive winning attributes and defensive and offensive attributes, we can determine when a team will win or lose.

The raw moneyball dataset has many missing values and some questionable attributes, so we had to brainstorm on a whiteboard, and also look a few steps ahead, because we didn’t want to waste any time and effort on trying to do something that would be out of our scope for this project. For instance, one member in our group put a bit of thought into creating an entropy model for this dataset by categorizing the attribute values to low, medium, or high. However, another member in our group feared that we may have to evaluate the our data on a very detailed level which would take more time and we may run into an overfitting issue.

Our group cleaned up the raw dataset using RStudios and since we have used it before to clean up the rolling sales of New York, only a little more research was required to be able to utilize it for our moneyball dataset. By using the “sum(is.na(ATTRIBUTE NAME))” method, RStudios gives us the number of missing values for each attribute. Upon evaluating our results, we discovered that only 6 of the 15 attributes had missing values. The 6 attributes with missing values are; bat.hbp, baserun.cs, baserun.sb, bat.so, pitch.so, and field.dp.

The attribute with the most missing values was “bat.hbp” which indicates how many times a batter on the team walked to base due to being hit by a pitcher with 2085 missing values. We decided to remove this attribute because 2085 missing values out of 2277 rows was simply too great of a data loss. With 772 missing values, “baserun.cs” which represents the number of times a player was caught trying to steal a base, was next in terms of most missing values, and we removed it as well since it is about a third of rows missing. The last attribute we decided to remove was “baserun.sb”, the attribute that represents the number of stolen bases. Although it was only missing 131 rows, we grouped baserun.sb and baserun.cs together since they directly contrasted each other. Also, ScienceDaily states that, “A new analysis found hitting accounts for more than 45 percent of Major League Baseball teams' winning records, fielding for 25 percent and pitching for 25 percent. And, the impact of stolen bases is greatly overestimated.”[1] This information reinforces our decision to classify the attribute as an outlier. We remove the attribute and now we are left with 13 attributes, 3 of which still have some missing values.

The attribute bat.so and pitch.so each have 102 missing values, while field.dp has 286 missing values. The amount of the remaining missing values indicates that we have at least 286 rows with missing values and at most 490. The “na.omit(DATASET)” method deletes any row with missing values, since we have 2277 rows, we can assume that we will still end up with plenty of data entries even if we were to eliminate the maximum amount of rows, which would be 490 rows. Upon executing the omit method we end with 1889 data entries which is still a very large portion of the dataset.

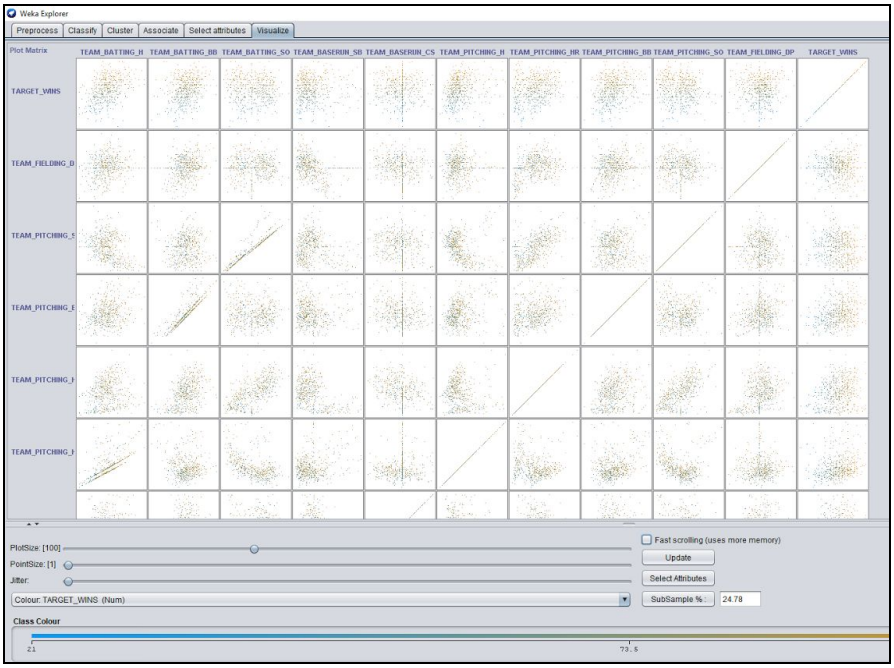
Building Our Models

After the data had been cleaned by eliminating empty columns, transforming missing data, ridding of uncorrelated attributes, and processing out outliers, we began to look at scatter plots generated in Weka. We found that (with the exception of TEAM_PITCHING_SO and TEAM_FIELDING_E which oddly the opposite effect) most of the attributes’ relation with TARGET_WINS aligned with the theoretical outcomes

[1] University of Delaware. "Baseball's winning formula: Statistical analysis debunks the old adage 'Pitching is 75 percent of the game'." ScienceDaily. www.sciencedaily.com/releases/2011/09/110929122932.htm (accessed December 7, 2017).

“positive impact” / “negative impact” outlined in the data set dictionary. Being that all of our decided attributes did appear to have some level of linear correlation with wins, we began to split the data into smaller categories in order to determine an order of importance towards winning a game. TEAM_BATTING attributes (H, BB, and SO) formed our Batting category, TEAM_PITCHING (H, HR, BB, and SO) formed our “Pitching” category, and TEAM_FIELDING (E and DP) formed our “Fielding” category. We split the data in an attempt to answer the question of which aspect of play is most associated with winning.

Linear Regression

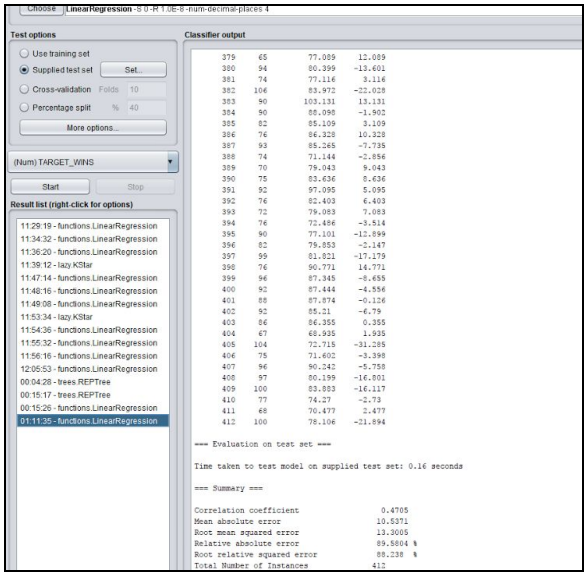


To build our linear regression models, the new data sets contained within each category were split into 3 smaller sets for analysis. Training sets contains 60 percent of the data, while the validation and test sets each contained 20 percent. The built-in Weka function for linear regression was then applied to each of the sets for each category using cross validation with a minimum of ten folds. We found that the training, and test sets yielded virtually identical results which indicates that the models ran properly. (More will be discussed on the actual results in a later section).

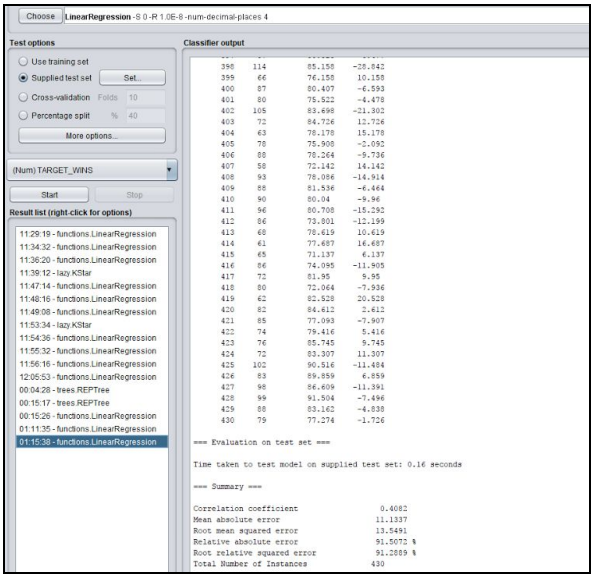
Upon completion, it was observed that the category “Batting” had our highest correlation coefficient while fielding had nearly no correlation whatsoever. Since TEAM_BATTING_H yielded the highest weight and consists of TEAM_BATTING_2B, TEAM_BATTING_3B and TEAM_BATTING_HR, we choose to further look into which base hits correlate most with TARGET_WINS. To account for first base hits, we simply subtracted 2B, 3B, and HR from TEAM_BATTING_H. The remainder represents hits

that were made with a player on first base. After running this new category through regression analysis, we discovered that 2B and 3B have the highest association with wins (3B with a slightly higher weight value).

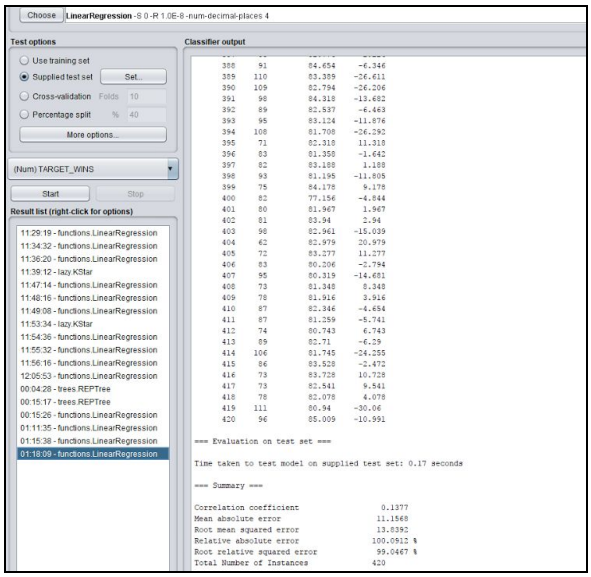
Batting



Pitching



Fielding



Admittedly, our correlation coefficients did not strike a lot of confidence with our models. Our goal was to achieve a 75 percent or greater correlation coefficient but were only able to observe (at most) about 50 percent with the data provided. In our further attempts, we created and added new attributes based on the following:

- TEAM_BATTING_H / TEAM_PITCHING_H
- TEAM_BATTING_BB / TEAM_PITCHING_BB
- TEAM_PITCHING_SO / TEAM_BATTING_SO

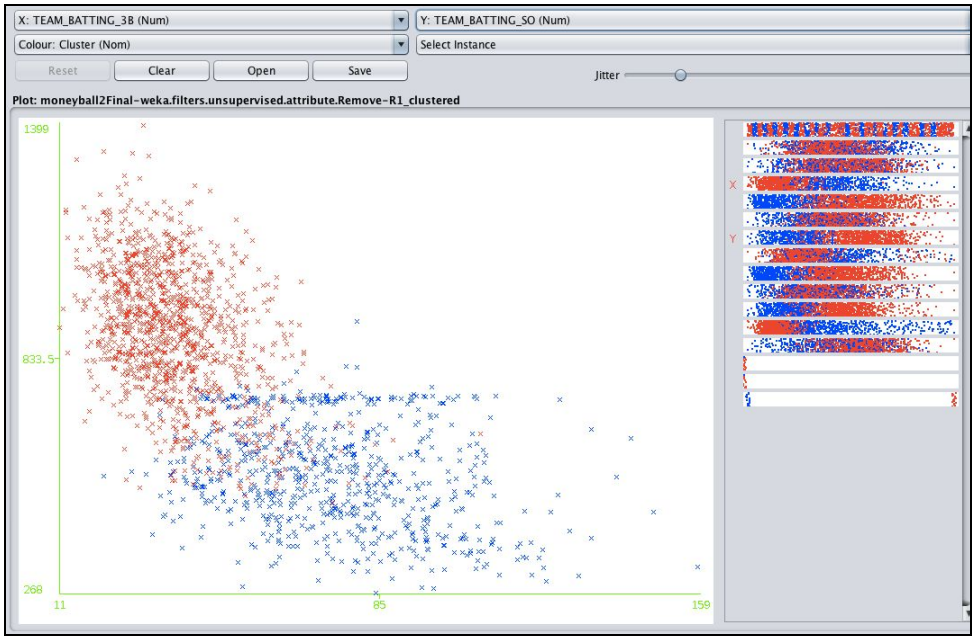
Still, correlation did not rise to any significant levels. Although we were able to mostly confirm the accuracy of the theoretical effects of each attribute on wins, we would have liked to have more statistical data on each team.

K-Means Clustering

For the unsupervised technique, K-means clustering was used. By definition of an unsupervised algorithm, there is no way to attain a “right answer,” like in supervised learning. Because there is no truth data, clustering is not used for classification. Instead, this technique is used to find hidden groupings in the data and analyze their meanings. In our model, we ran the k-means clustering algorithm with Weka with only two clusters. Although this algorithm would not be able to predict TARGET_WINS, it showed us patterns in certain attributes. The unsupervised model was an important addition to the analysis because, although we already had predicted TARGET_WINS, the results must be analyzed. With the clustering, the teams could see the successful strategies and their most common results (like achieving homeruns, strikeouts, etc.).

After running the model through Weka, we found noticeable clusters. In making the models, we only compared attributes of the same team and position. As stated above, the clustering model showed the best results with the TEAM_BATTING attributes.

TEAM_BATTING_SO vs. TEAM_BATTING_3B



TEAM_BATTING_HR vs. TEAM_BATTING_3B

These clusters show that there are different categories of teams. In the first model, Cluster 1 shows a team that has a high amount of strikeouts, with a low amount of third base hits. Cluster 0 shows a team with a more neutral profile: a medium amount of third based hits and a medium-to-low amount of strikeouts.

The second image shows a similar shape as the image above. In this one, Cluster 1 shows a team with a high number of homeruns, and a much lower number of third base hits. Cluster 0 exemplifies a team in which they hit a large amount of triples, instead of home runs. This can be analyzed to see what kind of players are on each team. Maybe the team hitting more triples does not have as strong of players, but they manage to score with Runs Batted In (RBI). This is a different strategy than teams who have incredible strong hitters, and hit home runs whenever they go up to bat. Both are different techniques to win a game, although a domain expert is needed to fully explore those different types of teams.

Performance on Our Models

Our strategy was to divide our dataset into three categories and apply multiple linear regression. We decided to divide our data into a batting, pitching, and fielding categories. Once we got those fields we broke each resulting dataset into training, validation, and test sets to which we ran multiple linear regression on each. For the Batting category on our Training Set we received 42% for correlation coefficient and 18% for our R^2. On our Cross Validation x10 folds we got 43% for correlation

coefficient and 20% for our R^2. For our Test Set we received 47% for correlation coefficient and 22% for our R^2. For the Fielding Category on our Training Set we received 0% on correlation coefficient and 0% for our R^2. For our Cross Validation we received 1% on correlation coefficient and 2% for our R^2. For our Test Set we received 1 % on correlation coefficient and 2% for our R^2. And for our Pitching Category on our Training Set we got 35% for correlation coefficient and 12% for our R^2. On our Cross Validation we got 32% for correlation coefficient and 13% for our R^2. Lastly, we got 38% for correlation coefficient and 13% for our R^2.

Conclusion

Based on our results our correlation tells us that the Batting category allows us to predict what team is most likely to win. We also noticed that Team Batting Hits had the highest weight in the Batting Category. Therefore, we decided to further look into hits broken down into singles(1B), doubles(2B), triples(3B), and home runs(HR). And triples(3B) had the highest weight which is of significance to our findings. Our results tells us that the team with the most hits is most likely to predict a win. Especially if most of the hits are triples (3B).