DATA 621

Homework 1, 9/27/20

Group 1

Members: Tony Mai, Justin Hsi, Lin Li, Santosh Cheruku, Michael ODonnell

1. **Data Exploration**

The “moneyball” training data set contains 2276 rows and 17 columns, including variables such as TARGET\_WINS, TEAM\_BATTING, TEAN\_BASERUN, etc. The variables are thought to have a positive or negative effect on the number of games the baseball team won during the season. Running a summary() function on the data set, we are able to get the mean, median, first and third quartile and the minimum and maximum values for each variable. We included a correlation plot and pairs plot to visualize the relationship among the variables. Histograms were created for each type of hits to observe the normality of the variables. We explored the structure of the variables for both the training and evaluation data sets and finally observed how TARGET\_WINS are affected by other factors. Interestingly, the number of wins seems positively correlated with all hits by batters except triples by batters, which the correlation plot shows as slightly negatively correlated. One potential explanation may be that getting triples, while good, is actually always worse than getting homeruns, so having a large number of triples may actually mean the team is just barely falling short. Nothing from the correlation plot can be used to conclude this, but it is something that can be investigated further in the future. Also surprising is that stolen bases barely has any positive correlation with wins, but that may just be due to the rarity of the event (stolen bases). TEAM\_PITCHING\_H, TEAM\_PITCHING\_BB, and TEAM\_PITCHING\_HR surprisingly shows a positive correlation with team wins, but maybe this alludes to having good batters and getting runs being more important to winning than stopping the opponent from getting runs. Similarly, TEAM\_PITCHING\_SO and TEAM\_PITCHING\_DP are events of denying the opponent runs, but they show a negative correlation with number of wins and may also point to getting runs for your team as the key to winning

1. **Data Preparation**

We addressed issues with imperfect data before building models or performing statistical analysis. We observed that several variables have high numbers of NA or missing values. TEAM\_BATTING\_HBP has the highest number of missing cases i.e., 2085 (~ 90%). Based on the variable definitions given in the assignment, it seemed reasonable that NA values meant that there were no occurrences of that event. So we chose to create additional columns flagging whether the original variable was NA or not (1 if NA, 0 if not NA), and then filled NAs with 0.

1. **Build Models**

First we built a model using the backward elimination process. In this process, we rejected predictors with p-value greater than 0.05 and stopped after all remaining model predictors had p-values of less than 0.05. For our second model we decided to use stepwise selection. Stepwise selection uses an automated process of building a model by adding or removing predictors repeatedly based on an improvement of a criterion (Akaike information criterion in our case). We noticed one of the variables, TEAM\_PITCHING\_SO, had a 1 p-value greater than 0.05 in the second model so we decided to build a third model using stepwise regression with the TEAM\_PITCHING\_SO predictor removed. The third model’s R squared dropped slightly, so we decided to stick with our second model.

1. **Select Models**

Out of the three models we created, the second model with stepwise selection was the best of the three. The Adjusted R squared is 0.4098 which translates to approximately 41% of variation in Target Wins can be explained by our model. The F statistic tells us if there is a relationship between the dependent and independent variables we are testing. Generally, a large F indicates a stronger relationship and we have 113.9. The normal quantile quantile plot for residuals displays an approximately straight line so the residuals are approximately normally distributed. However, there is slight deviation at the extreme values, meaning our model does have a bit of trouble predicting a very high or low number of wins accurately. The MSE is 743.6606. Using this model we were able to make predictions for the test dataset. Finally, we made a histogram of wins from the training and evaluation set to see if the prediction distribution looked fairly similar to the training distribution, which it does.

**5. Link to Code**

Link to code: <https://github.com/odonnell31/data621-HW1/blob/master/hw1_v2.Rmd>