Unit 1 Assignment: “MONEYBALL OLS REGRESSION PROJECT”

MSDS 411 Section 56, spring 2019

Prabhat Thakur 04/20/2019

Introduction

The purpose of this assignment is to build a multiple linear regression model on the moneyball baseball training data to predict the number of wins for a team in a game session. Each record in training data set represents a professional baseball team from the years 1871 to 2006 inclusive. For arriving to a best predictive model, different multiple linear regression models are evaluated, and a final model is chosen based on model fit statistics and diagnostic plots.

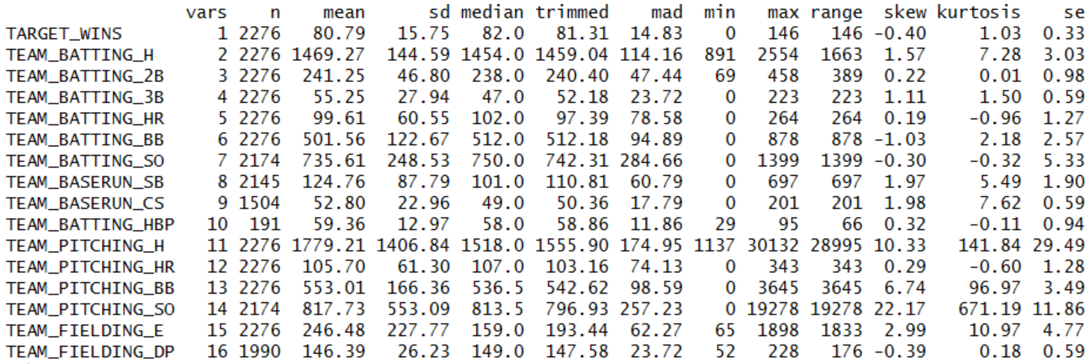
Bonus Problem

I have attempted two bonus problem for total of 40 points in this assignment. 1. Create model using glm function to compare its results with lm results for the default glm options. 2. I used mice package to impute missing values using decision trees. Also used random forest model for variable selection for one of the model.

Data Exploration

The data for this assignment contains summary information about individual team’s baseball statistics. There are 2276 records in the training dataset and 259 in the test dataset. The Training data set has 17 columns and out of those 17 columns, INDEX is simply an index value for observation identification while TARGET_WINS represents the response variable. The remaining 15 elements are all potential predictor variables for building the linear models. The test dataset contains same columns except the TARGET_WINS. The trained model will be used to predict TARGET_WINS for test dataset observations. Below is a summary table for the training dataset excluding INDEX column. All variables are given as continuous numeric which is ideal for any regression model building.

From the below summary of training data, it appears that several of the variables has missing values. Also some variables are skewed and has outliers.



```
                  vars    n    mean      sd median trimmed    mad  min   max range  skew kurtosis    se
TARGET_WINS          1 2276   80.79   15.75   82.0   81.31  14.83    0   146   146 -0.40     1.03  0.33
TEAM_BATTING_H       2 2276 1469.27  144.59 1454.0 1459.04 114.16  891  2554  1663  1.57     7.28  3.03
TEAM_BATTING_2B      3 2276  241.25   46.80  238.0  240.40  47.44   69   458   389  0.22     0.01  0.98
TEAM_BATTING_3B      4 2276   55.25   27.94   47.0   52.18  23.72    0   223   223  1.11     1.50  0.59
TEAM_BATTING_HR      5 2276   99.61   60.55  102.0   97.39  78.58    0   264   264  0.19    -0.96  1.27
TEAM_BATTING_BB      6 2276  501.56  122.67  512.0  512.18  94.89    0   878   878 -1.03     2.18  2.57
TEAM_BATTING_SO      7 2174  735.61  248.53  750.0  742.31 284.66    0  1399  1399 -0.30    -0.32  5.33
TEAM_BASERUN_SB      8 2145  124.76   87.79  101.0  110.81  60.79    0   697   697  1.97     5.49  1.90
TEAM_BASERUN_CS      9 1504   52.80   22.96   49.0   50.36  17.79    0   201   201  1.98     7.62  0.59
TEAM_BATTING_HBP    10  191   59.36   12.97   58.0   58.86  11.86   29    95    66  0.32    -0.11  0.94
TEAM_PITCHING_H     11 2276 1779.21 1406.84 1518.0 1555.90 174.95 1137 30132 28995 10.33   141.84 29.49
TEAM_PITCHING_HR    12 2276  105.70   61.30  107.0  103.16  74.13    0   343   343  0.29    -0.60  1.28
TEAM_PITCHING_BB    13 2276  553.01  166.36  536.5  542.62  98.59    0  3645  3645  6.74    96.97  3.49
TEAM_PITCHING_SO    14 2174  817.73  553.09  813.5  796.93 257.23    0 19278 19278 22.17   671.19 11.86
TEAM_FIELDING_E     15 2276  246.48  227.77  159.0  193.44  62.27   65  1898  1833  2.99    10.97  4.77
TEAM_FIELDING_DP    16 1990  146.39   26.23  149.0  147.58  23.72   52   228   176 -0.39     0.18  0.59
```

To explore predictor variables from training and test data set together, both datasets were merged for data exploration purpose. Below are the boxplot and histogram of all predictor variables. Figure 1, shows the distribution of the predictor variables. Based on the box plots, it is clear that the testing data likely comes from the same population as the training data and that several of the variables are skewed. In addition to needing to adjust for skewness, the data presents outliers for which flooring or capping should be considered.

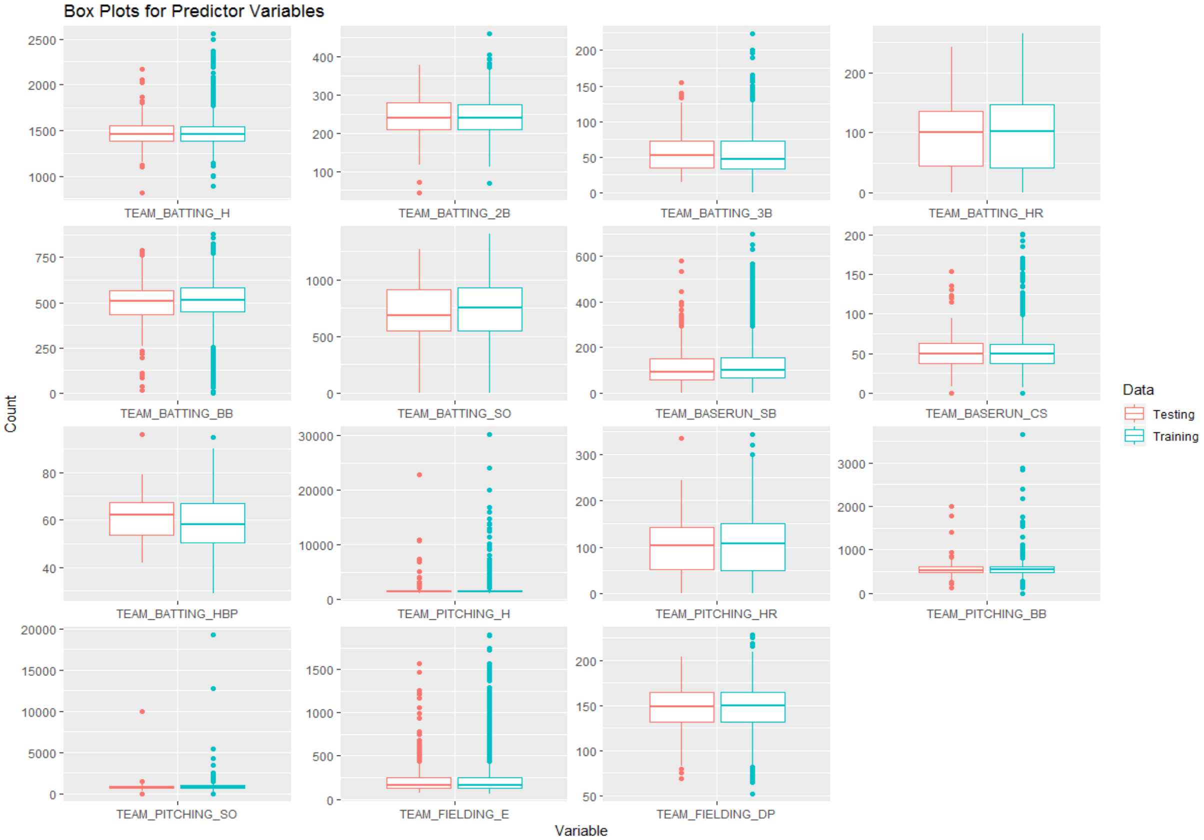


Figure 1 – Box Plots of Training and Testing Predictor Variables

Form the series of histograms charts, Figure 2, it appears that pitching data is skewed by outliers. Some of the predictor distributions also appear bimodal. This indicates that important information such as the number of games played in a particular season or era-specific characteristics are missing. If the year of the season was available, that context could adjust for historical baseball trends.

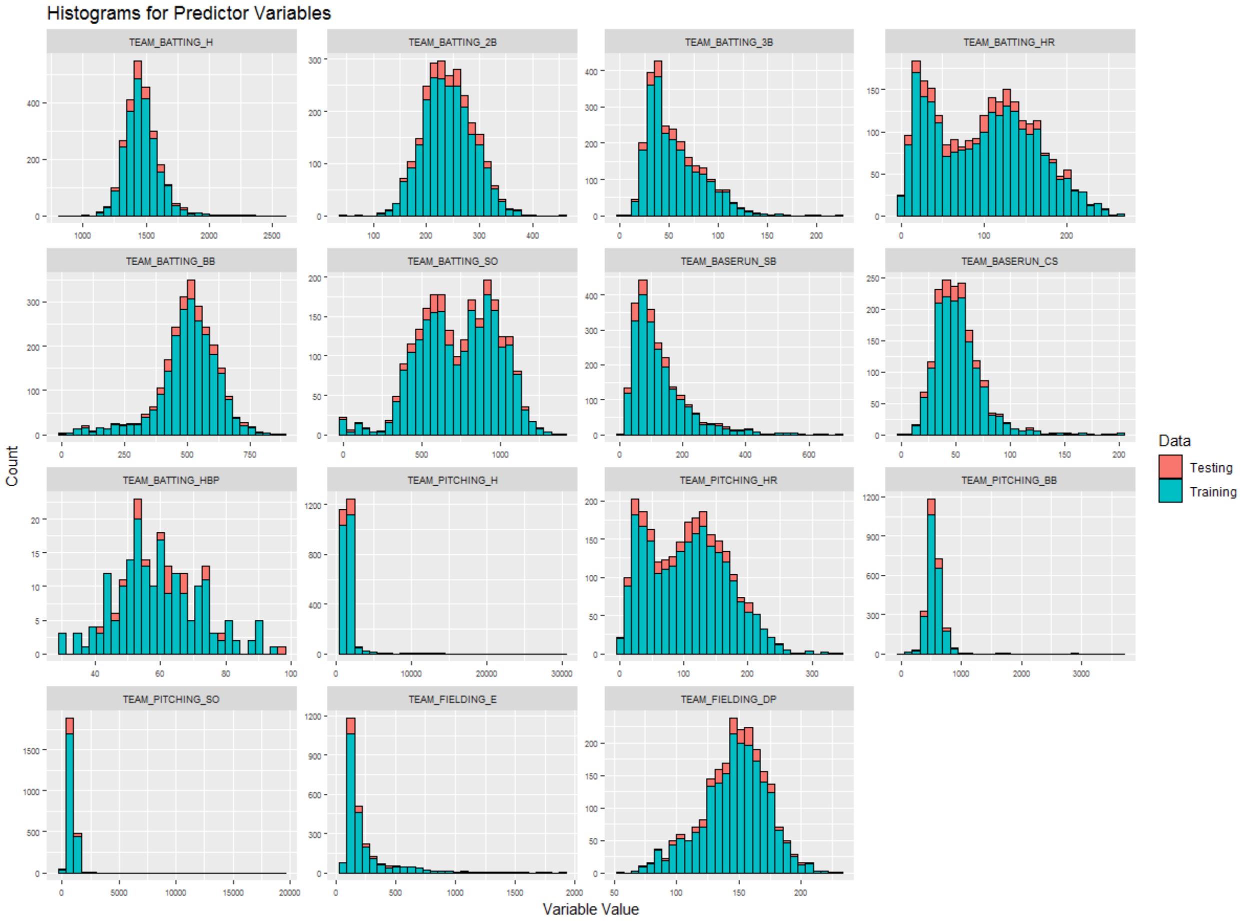


Figure 2 – Histograms of Training and Testing Data

The correlation matrix heat map (Figure 3) of all 16 variables on training dataset shows that some of the predictors are moderately correlated with target and will be useful in creating predictive model. From the matrix, most predictors are not strongly correlated with each other with the exception of the pitching and batting variables (hits, home runs, walks, and strikeouts) which are perfectly correlated. Including both type of variables in the model could result in multicollinearity issue. Based on the relationship with the target, it would appear that the pitching data is inaccurate and should be dropped from consideration. For example, it shows that teams with more pitching walks result in more wins which is unintuitive. Also some strong correlations between TEAM_BATTING_H and

TEAM_BATTING_2B, TEAM_BATTING_3B and TEAM_BATTING_HR since these individual variables are each a subset of hits.

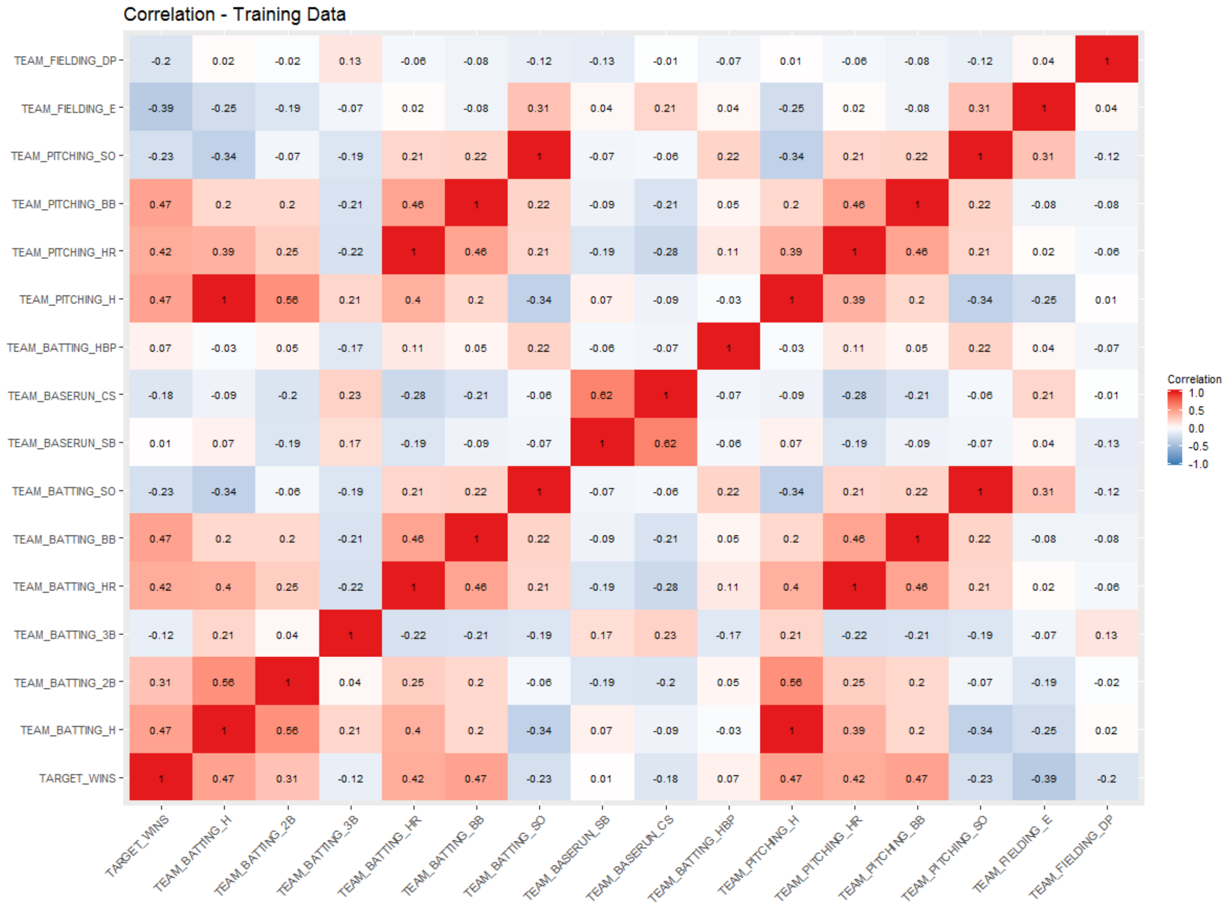


Figure 3 – Correlation in Training Data

Figure 4 shows the portion of values missing in the training data for variables with at least one missing value. These missing values are imputed in the Data Preparation section. The large percentage of missing TEAM_BATTING_HBP values indicate that even with imputation, this variable should not be considered in a model.

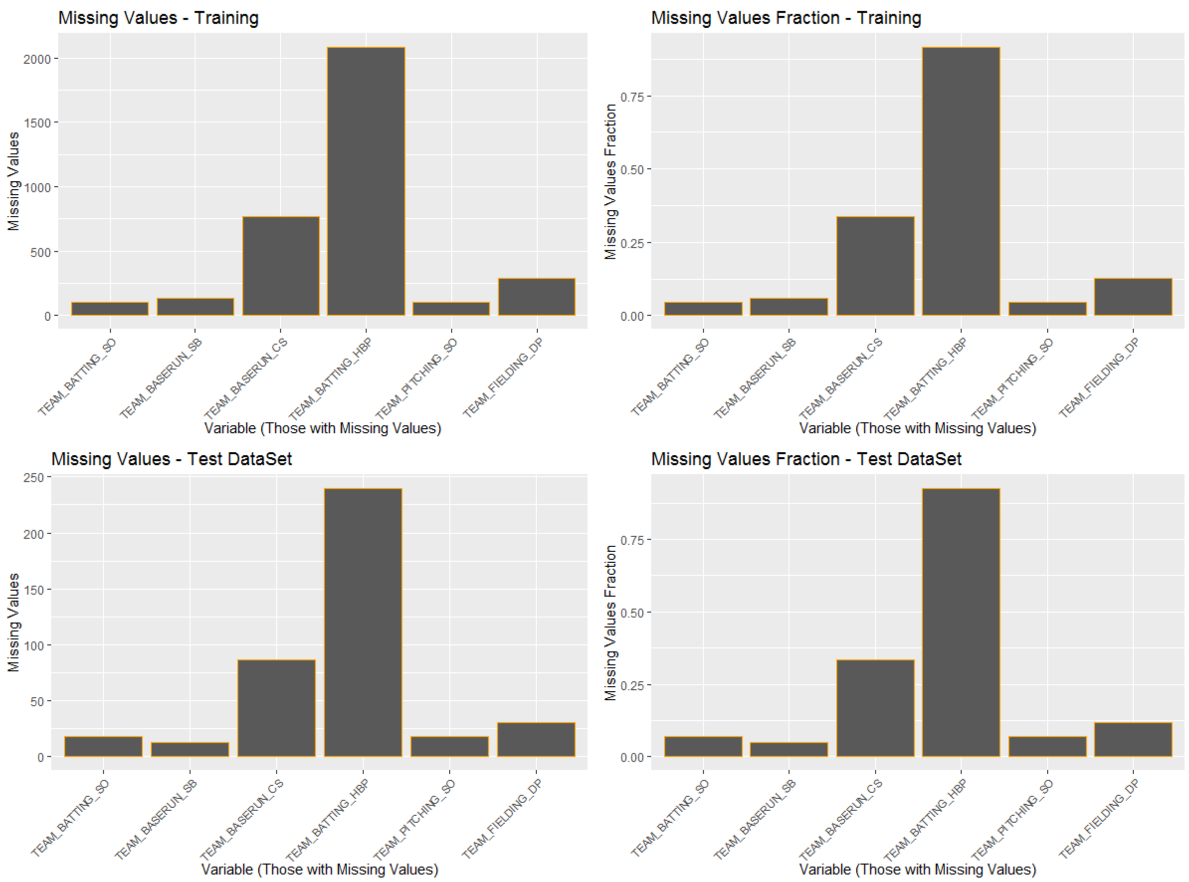


Figure 4 – Missing Value in Training and Test dataset

Data Preparation

From the data exploration exercise it is stablished that pitching variables TEAM_PITCHING_H, TEAM_PITCHING_HR, TEAM_PITCHING_BB, and TEAM_PITCHING_SO are perfectly correlated to batting variables. These variables are excluded from further model building exercise to avoid potential multicollinearity issue. Next, win totals less than 20 are caped to 20 and similarly win total greater than 116 as caped to 116 as those are major league records and observations outside that range may influence the model undesirably.

An additional variable for batting singles is created so that Base Hits by batters (1B,2B,3B,HR) variable so that all batting variable can be treated independently in the model, as a result it is removed from the model. 5 new missing value indicator variables are also created for the variables with missing values. These variables are used in the modeling process to determine if the fact that a value is missing adds predictive value itself. These same steps are also performed on test dataset.

In next data preparation step, the missing values are imputed using decisions trees within the MICE R package. Due to the way that the MICE package works, the imputation formula is not able to be saved for future use. That means that testing data and training data are combined together for imputation and then separated again for further model building. This works for the small size of data in this example but would not be practical for large data sets.

Next, the target and missing value indicator variables are added back to the now imputed data set. For skewed variable singles, triples, stolen bases, and errors, the log of the variable is used instead. To account for large outliers in the predictors, singles, doubles, triples, stolen bases, errors, double plays, and strikeouts variable's upper limit are capped at the 99th percentile and lower limit is capped at the 1th percentile of the training data.

So far no observation are removed from the training data base, however after reviewing residual vs fitted value and residual QQ plot, two observations stand out as extreme outliers causing models to violate OLS linear regression modeling assumptions. Observation number 1211 and 1342 are removed before fitting the model.

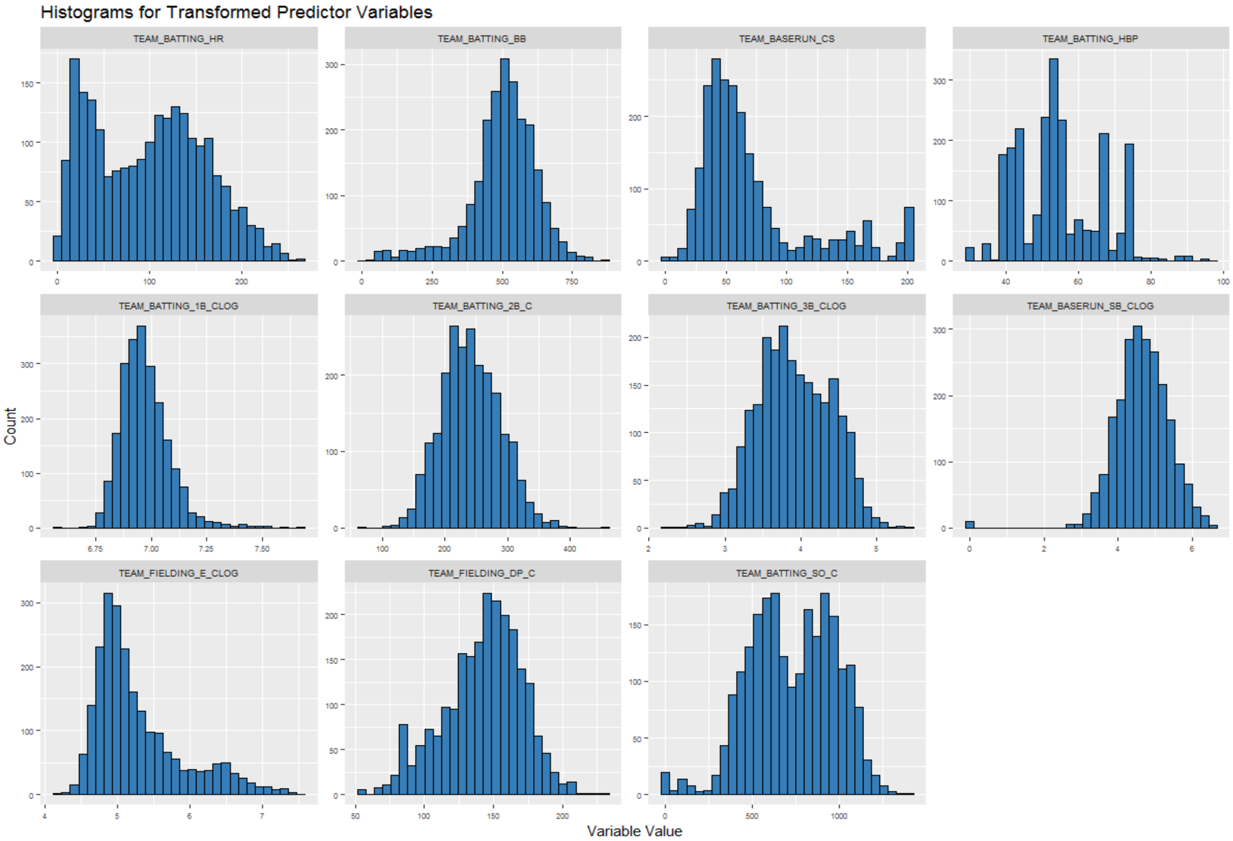


Figure 5 – Histograms of Transformed Training Data

The distributions of these final continuous variables are included in Figure 5. The charts show that the transformations do reduce the skewness in the data but the bimodal nature of some of the variables persist. Figure 6 shows the correlation between these final variables. This check shows that variables highly correlated with each other have been successfully removed, so multicollinearity should be avoided. The correlation with the target also shows promise for predictive models.

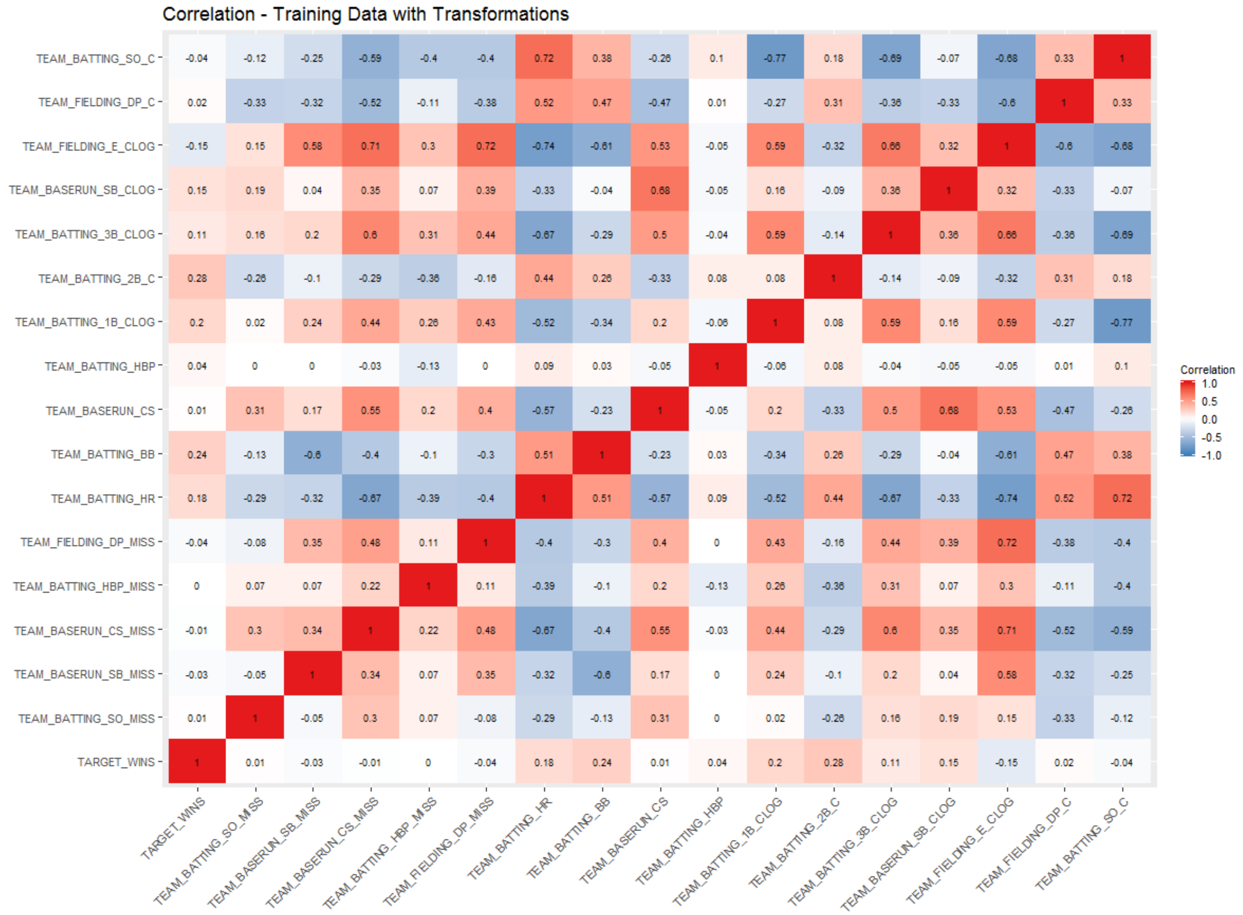


Figure 6 – Correlation in Transformed Training Data

Figure 7 shows correlation of each predictor variables with the target in training dataset. Single, double and walks by batters variables seems to be good predictors.

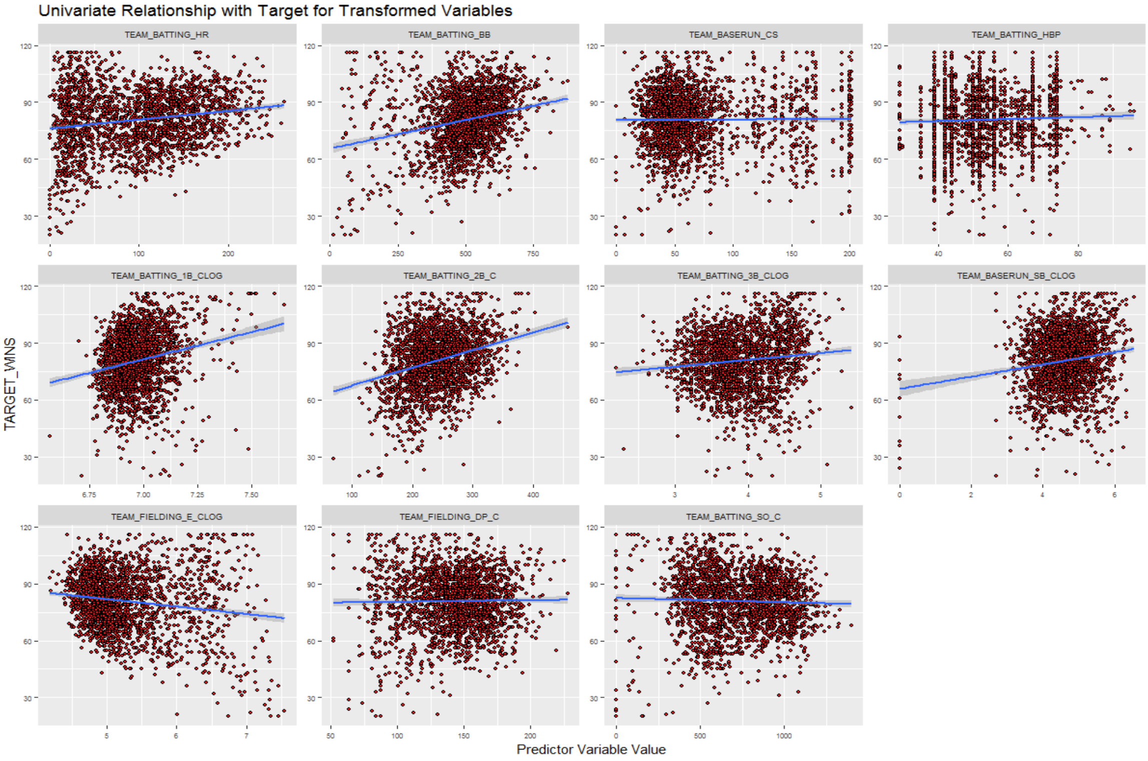


Figure 7 – Univariate Relationship with Target for Transformed Variables

Build Models

To test the model accuracy on unseen data by model, training data set is spited in 70/30 ratio. 70% of the training dataset is then used for fitting the model (model training) and rest 30% is used to assess model performance and estimate how a model would perform on new data.

Model1: This model is regress on all batting variable, as player's ability to get on base was stressed the best predictor of future performance. Except TEAM_BATTING_HBP variable, all other batting variables are statistically significant, also all predictor coefficients are positive and has an intuitive relationship with the target.

Model 2: This model contains all of the variables identified in the Data Preparation section and shown in Figure 7. Models predictive accuracy is relatively high compared to model1 however some of the variables are statistically insignificant and have unintuitive relationships with the target. Variance inflation factors for different variables are under the suggested threshold limit of 10, there is a variable with factors greater than six which raises concerns regarding the stability of the model's coefficients.

Model 3: This model applies stepwise variable selection to the variables included in Model 2. This model performance is very similar to model2 and do not remove any predictors.

Bonus Model 4: This is a same model as model 2 using glm function. The default options for that function specify the identity link function and a gaussian error distribution. Model performance and coefficients are exactly same as model2. The default options of glm function matches with linear regression exactly. It can be concluded that linear regression is a special case of a generalized linear model.

Bonus Model 5: For this model random forest model built to determine top eight variables and these variables are then used in a linear regression. This model performs poorly compared to model 2 and 3. Also top 8 variables also includes less statistically significant predictor variables.

Bonus Model 6: This model includes all the variables same as model 2 with additional interaction terms between home runs and stolen bases, and errors and strike outs. These interactions were selected by looking at the correlation between these variables and testing the impact of adding them to the model. This is the best model out of all 6 model. Figure 9 shows the model diagnostic plots for Model 6. These plots show that assumptions of linear regression are met, however the model does do a relatively poor job of fitting to extremely low win seasons.

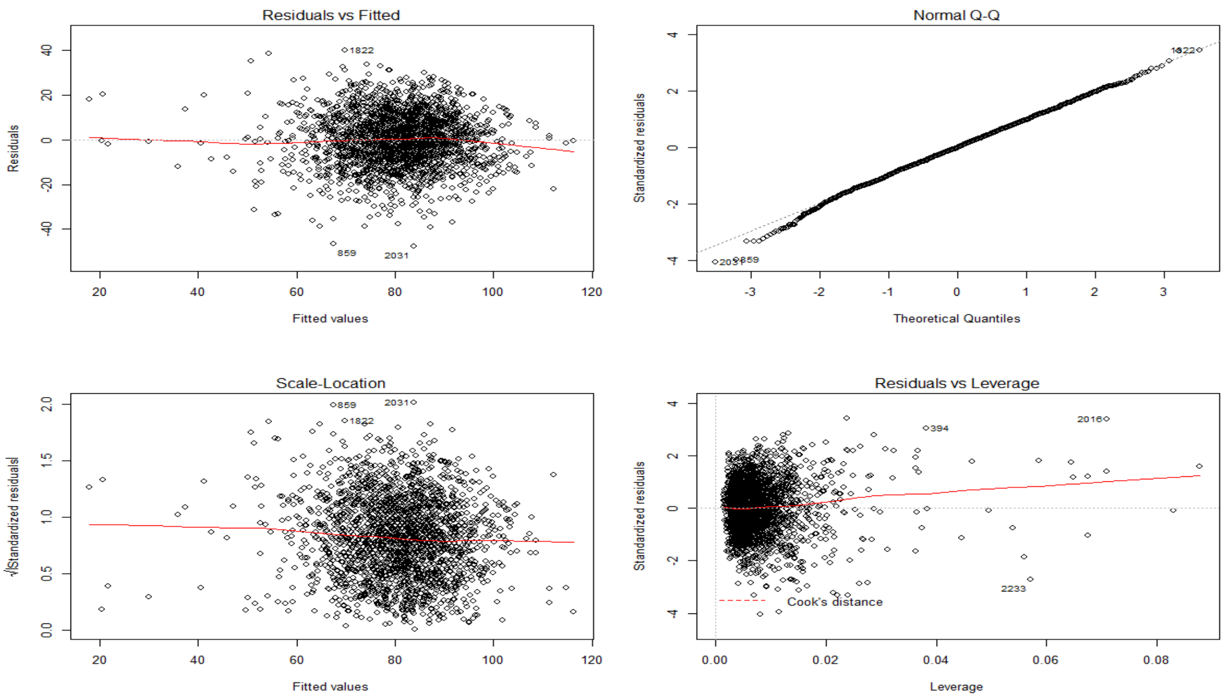


Figure 8 – Final Model Diagnostic Plots

Select Model

Below table lists different model performance matrices for all models built. Model 6 is selected as best model based on highest Adj R squared value and lowest Training and Validation prediction error. Once Model 6 is selected, the model is fit to the entire training data. Due to the large training sample it provides more accurate coefficients.

Model	Adj R-Squared	AIC	Training MSE	Validation MSE
Model1	0.2256	12850.78	184.7695	180.9046
Model2	0.3985	12458.34	142.6229	135.1502
Model3	0.3986	12457.06	142.6867	135.3734
Model4	0.3986	12457.06	142.6867	135.3734
Model5	0.2847	12726.25	170.4474	165.1728
Model6	0.4019	12450.09	141.7076	133.3392
Final Model	0.4139	17709.29	138.8144	131.1281

Following is the regression equation of final model which is used to generate scores (predict target wins) on test data set.

```
P_TARGET_WINS = round(-276.0711 +
                0.2491296 * TEAM_BATTING_HR  +
                0.02751336 * TEAM_BATTING_BB +
                0.04213978 * TEAM_BASERUN_CS +
                0.02432542 * TEAM_BATTING_HBP +
                57.04187 * TEAM_BATTING_1B_CLOG +
                0.01864378 * TEAM_BATTING_2B_C  +
                8.182519 * TEAM_BATTING_3B_CLOG +
                5.449530 * TEAM_BASERUN_SB_CLOG -
                21.19661 * TEAM_FIELDING_E_CLOG -
                0.1330793 * TEAM_FIELDING_DP_C +
                6.287021 * TEAM_BATTING_SO_MISS +
                26.46384 * TEAM_BASERUN_SB_MISS +
                2.572081 * TEAM_BASERUN_CS_MISS +
                5.292837 * TEAM_BATTING_HBP_MISS +
                4.294441 * TEAM_FIELDING_DP_MISS -
                0.03003720 * TEAM_BATTING_HR * TEAM_BASERUN_SB_CLOG -
                0.002560585 * TEAM_FIELDING_E_CLOG * TEAM_BATTING_SO_C ,
               2)
```

Coefficient value and there sings are in sync with the predictor variable impact on win except Double Plays. It is supposed to have positive impact on win but as per the regression equation based on provided training data coefficient sign is negative which suggest that its impact is negative on win.

Conclusion

Data preparation is an important step of predictive model building process. Proper handling of missing, outliers and skewed values in training and testing dataset is very important. Subject knowledge can be very helpful when doing data exportation and preparation. In this assignment, in order to predict the number of wins a team will earn in a season, 6 different multiple linear regression models were trained. While the final model fits the data well, its coefficients can also illustrate the factors that make an individual team successful. Teams that succeed in common offensive categories such as home runs will likely win more games on average. The model can be greatly improved if other commonly used baseball statistics are also included in the analysis.