Homework 1 – PREDICT 411

Nate Bitting

Introduction and Data Exploration

The first step in any model building exercise is to first explore the data in order to understand any nuances in the data that need to be addressed during the data preparation phase. In this assignment, we will be using the Moneyball dataset which contains 2,276 observations of baseball statistics with a corresponding number of wins as the response variable. We will leverage this data to build an OLS regression model to predict the number of wins, given a set of statistics provided in the dataset for each observation. The primary objective is to deploy a predictive model that can be leveraged to predict new observations.

Table 1 – Moneyball Dataset Variable List

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
1	INDEX	Num	8	
2	TARGET_WINS	Num	8	
10	TEAM_BASERUN_CS	Num	8	Caught stealing
9	TEAM_BASERUN_SB	Num	8	Stolen bases
4	TEAM_BATTING_2B	Num	8	Doubles by batters
5	TEAM_BATTING_3B	Num	8	Triples by batters
7	TEAM_BATTING_BB	Num	8	Walks by batters
3	TEAM_BATTING_H	Num	8	Base Hits by batters
11	TEAM_BATTING_HBP	Num	8	Batters hit by pitch
6	TEAM_BATTING_HR	Num	8	Homeruns by batters
8	TEAM_BATTING_SO	Num	8	Strikeouts by batters
17	TEAM_FIELDING_DP	Num	8	Double Plays
16	TEAM_FIELDING_E	Num	8	Errors
14	TEAM_PITCHING_BB	Num	8	Walks allowed
12	TEAM_PITCHING_H	Num	8	Hits allowed
13	TEAM_PITCHING_HR	Num	8	Homeruns allowed
15	TEAM_PITCHING_SO	Num	8	Strikeouts by pitchers

As can be seen in Table 1, there are fifteen variables in the dataset. The dependent variable we are interested in predicting is TARGET_WINS. The INDEX variable is just a unique identifier for each observation. The remaining thirteen variables are numeric variables we will consider for use in the OLS regression model to predict TARGET_WINS for each observation. A sample view of the first 10 observations of the dataset can be seen in Table 2 below.

Table 2 – 10 Observations of Moneyball Dataset (does not include all variables to save space)

Obs	INDEX	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS
1	1	39	1445	194	39	13	143	842	.	.
2	2	70	1339	219	22	190	685	1075	37	28
3	3	86	1377	232	35	137	602	917	46	27
4	4	70	1387	209	38	96	451	922	43	30
5	5	82	1297	186	27	102	472	920	49	39
6	6	75	1279	200	36	92	443	973	107	59
7	7	80	1244	179	54	122	525	1062	80	54
8	8	85	1273	171	37	115	456	1027	40	36
9	11	86	1391	197	40	114	447	922	69	27
10	12	76	1271	213	18	96	441	827	72	34

In order to get to know the data better before understanding what data processing might be required, we will investigate each variable for missing values and outliers in the next section.

Missing Values

To assess what data we are working with, we will explore summary statistics and identify the magnitude of any missing values for each predictor variable.

Table 3 – Summary from SAS Means Procedure

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Minimum	5th Pctl	50th Pctl	90th Pctl	95th Pctl	99th Pctl	Maximum
TARGET_WINS		2276	0	80.7908612	82.0000000	0	54.0000000	82.0000000	100.0000000	104.0000000	114.0000000	146.0000000
TEAM_BATTING_H	Base Hits by batters	2276	0	1469.27	1454.00	891.0000000	1280.00	1454.00	1636.00	1696.00	1950.00	2554.00
TEAM_BATTING_2B	Doubles by batters	2276	0	241.2469244	238.0000000	69.0000000	167.0000000	238.0000000	303.0000000	320.0000000	352.0000000	458.0000000
TEAM_BATTING_3B	Triples by batters	2276	0	55.2500000	47.0000000	0	23.0000000	47.0000000	96.0000000	108.0000000	134.0000000	223.0000000
TEAM_BATTING_HR	Homeruns by batters	2276	0	99.6120387	102.0000000	0	14.0000000	102.0000000	180.0000000	199.0000000	235.0000000	264.0000000
TEAM_BATTING_BB	Walks by batters	2276	0	501.5588752	512.0000000	0	246.0000000	512.0000000	635.0000000	671.0000000	755.0000000	878.0000000
TEAM_BATTING_SO	Strikeouts by batters	2174	102	735.6053358	750.0000000	0	359.0000000	750.0000000	1049.00	1104.00	1193.00	1399.00
TEAM_BASERUN_SB	Stolen bases	2145	131	124.7617716	101.0000000	0	35.0000000	101.0000000	231.0000000	302.0000000	439.0000000	697.0000000
TEAM_BASERUN_CS	Caught stealing	1504	772	52.8038564	49.0000000	0	24.0000000	49.0000000	77.0000000	91.0000000	143.0000000	201.0000000
TEAM_BATTING_HBP	Batters hit by pitch	191	2085	59.3560209	58.0000000	29.0000000	40.0000000	58.0000000	76.0000000	83.0000000	90.0000000	95.0000000
TEAM_PITCHING_H	Hits allowed	2276	0	1779.21	1518.00	1137.00	1316.00	1518.00	2059.00	2563.00	7093.00	30132.00
TEAM_PITCHING_HR	Homeruns allowed	2276	0	105.6985940	107.0000000	0	18.0000000	107.0000000	187.0000000	210.0000000	244.0000000	343.0000000
TEAM_PITCHING_BB	Walks allowed	2276	0	553.0079086	536.5000000	0	377.0000000	536.5000000	694.0000000	757.0000000	924.0000000	3645.00
TEAM_PITCHING_SO	Strikeouts by pitchers	2174	102	817.7304508	813.5000000	0	420.0000000	813.5000000	1095.00	1173.00	1474.00	19278.00
TEAM_FIELDING_E	Errors	2276	0	246.4806678	159.0000000	65.0000000	100.0000000	159.0000000	542.0000000	716.0000000	1237.00	1898.00
TEAM_FIELDING_DP	Double Plays	1990	286	146.3879397	149.0000000	52.0000000	98.0000000	149.0000000	178.0000000	186.0000000	204.0000000	228.0000000

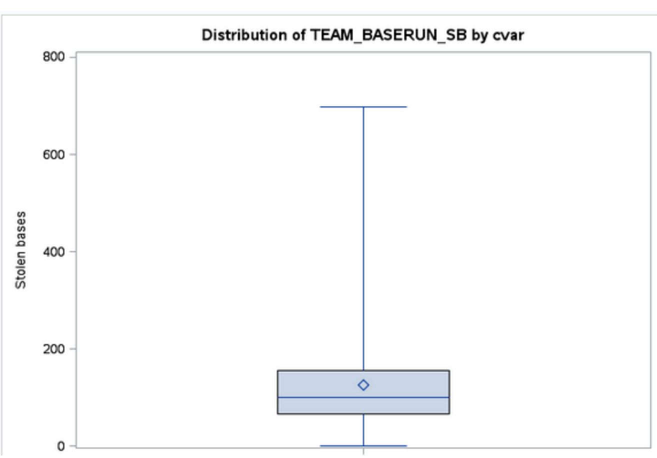
N = number of observations;
N Miss = number of missing observations
Mean = mean value for each numeric variable

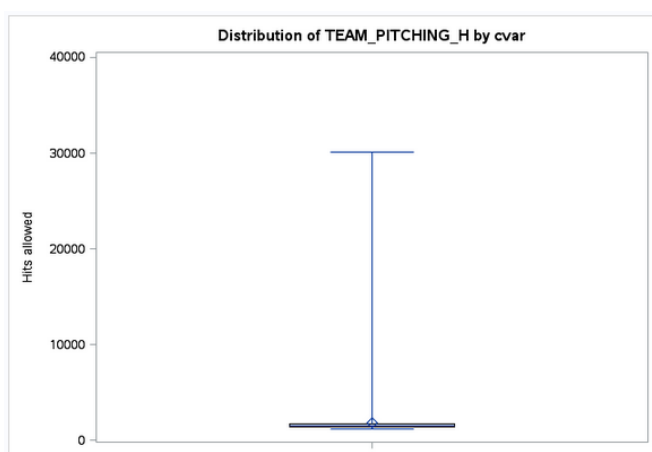
Table 3 provides us the visibility to know what numeric variables we will need to address in the data preparation phase in terms of handling missing values and any potential outliers. As you can see the variables with missing variables include Strikeouts by batters, Stolen bases, Caught stealing, Batters hit by pitch, Strikeouts by pitchers, and Double Plays. You should also notice the Batters hit by pitch variable is extremely sparse as 92% of the observations have a missing value. As a result of this fact, we will exclude the Batters hit by pitch variable from our model. For all other variables with missing values listed above, we will impute the *mean* value in place of each missing value.

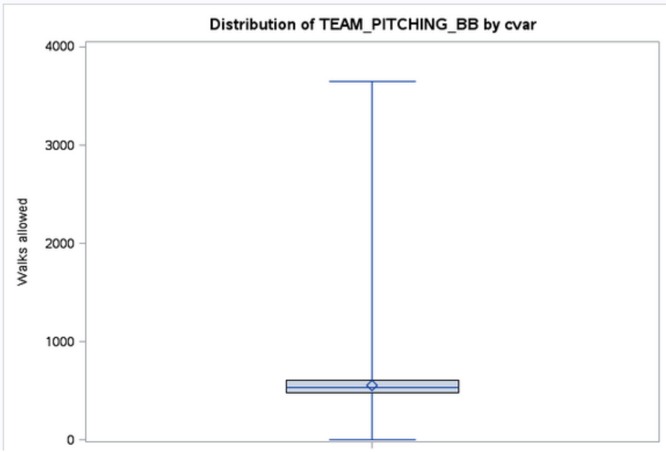
Assessment of Outliers

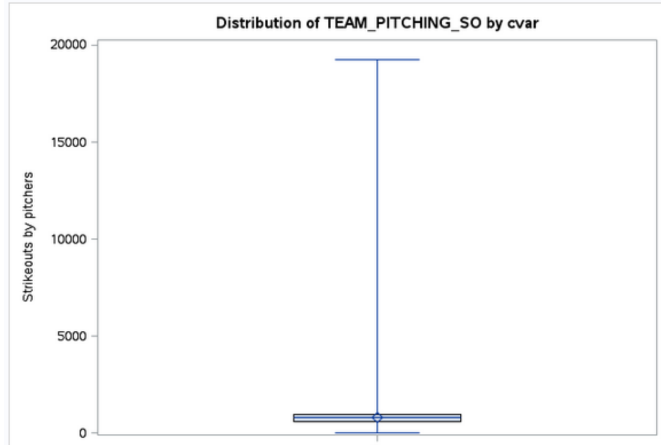
In order to determine what variables may contain extreme observations, we assessed the boxplots of each numeric variable. In Table 4 shown below, we selected the variables that will need to be addressed for outliers in the data preparation phase by means of variable transformation or trimming for each of the extreme observations.

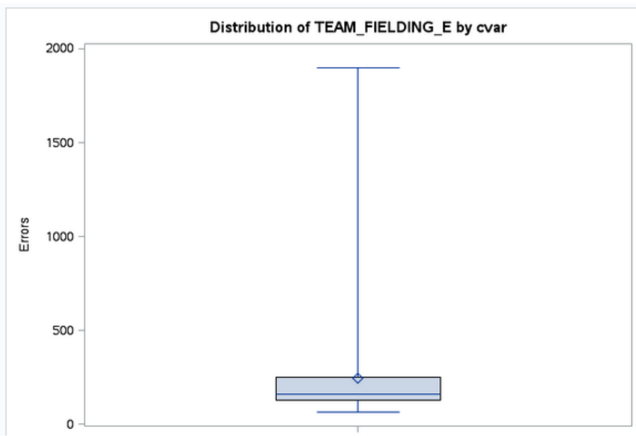
Table 4 – Boxplots for numeric variable with possible











When investigating each of these variables show in Table 4, we want to better understand what is driving such a wide range of values in each of these variables. To confirm our suspicions for the extreme observations we leverage the UNIVARIATE procedure to explore further into the detail.

Table 5 – UNIVARIATE Extreme Observations Examples

TEAM_BASERUN_SB

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1584	562	2023
0	1211	567	643
14	1825	632	642
18	2079	654	279
18	942	697	2022

TEAM_PITCHING_HR

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2239	297	426
0	2233	301	1810
0	2136	320	964
0	2016	320	1882
0	2015	343	832

TEAM_PITCHING_BB

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1211	2169	1340
119	1350	2396	1083
124	1824	2840	282
131	299	2876	2136
140	861	3645	1342

TEAM_PITCHING_SO

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2239	3450	282
0	2233	4224	1826
0	2016	5456	1
0	2015	12758	1342
0	1824	19278	2136

TEAM_FIELDING_E

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
65	1891	1567	391
66	390	1728	1584
68	1386	1740	1825
72	837	1890	1211
74	1335	1898	415

TEAM_PITCHING_H

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1137	1456	16038	1342
1168	1353	16871	415
1184	1001	20088	2136
1187	232	24057	1211
1202	1354	30132	1584

We noticed there were extreme observations in each variable (leveraging the UNIARIATE output in Table 5) that must be addressed in the data preparation phase. We will explore multiple techniques for dealing with these outliers through the use of trimming and/or variable transformation.

Correlation Analysis

Before we move into the data preparation phase, we should first explore if any relationships exist between the predictor variables and the response variable or any relationships among the predictor variables. By leveraging the CORR procedure (refer to Table 6 to the right) we assessed the correlation between the predictor and response variable, TARGET_WINS.

Most of the correlations made intuitive sense; Base hits by batters having a positive correlation with WINS and Errors having a negative correlation with WINS. However, there were a few predictor variables that were counterintuitive to the theoretical effect one would expect in terms of correlation to wins. For example, TEAM_BASERUN_CS should have a negative impact on WINS, but is showing a positive correlation to wins in Table 6. The same phenomena is present with TEAM_PITCHING_BB and TEAM_PITCHING_HR. We must keep a close eye on this during the model building phase to see if this phenomena is present in the predictor coefficients.

Table 6 – Correlation Matrix Predictors vs Response

		TARGET_WINS		
TARGET_WINS		1.00000		2276
TEAM_BATTING_H	Base Hits by batters	0.38877	<.0001	2276
TEAM_BATTING_2B	Doubles by batters	0.28910	<.0001	2276
TEAM_BATTING_3B	Triples by batters	0.14261	<.0001	2276
TEAM_BATTING_HR	Homeruns by batters	0.17615	<.0001	2276
TEAM_BATTING_BB	Walks by batters	0.23256	<.0001	2276
TEAM_BATTING_SO	Strikeouts by batters	-0.03175	0.1389	2174
TEAM_BASERUN_SB	Stolen bases	0.13514	<.0001	2145
TEAM_BASERUN_CS	Caught stealing	0.02240	0.3853	1504
TEAM_BATTING_HBP	Batters hit by pitch	0.07350	0.3122	191
TEAM_PITCHING_H	Hits allowed	-0.10994	<.0001	2276
TEAM_PITCHING_HR	Homeruns allowed	0.18901	<.0001	2276
TEAM_PITCHING_BB	Walks allowed	0.12417	<.0001	2276
TEAM_PITCHING_SO	Strikeouts by pitchers	-0.07844	0.0003	2174
TEAM_FIELDING_E	Errors	-0.17648	<.0001	2276
TEAM_FIELDING_DP	Double Plays	-0.03485	0.1201	1990

Data Preparation

Address Missing Values

As mentioned in the previous section, the variables with missing values include Strikeouts by batters, Stolen bases, Caught stealing, Batters hit by pitch, Strikeouts by pitchers, and Double Plays.

<u>Results of assessing predictors with missing values</u>

- TEAM_BATTING_SO: Median is 750 and 102 values are missing
- TEAM_BASERUN_SB: Median is 101 and 131 values are missing
- TEAM_BASERUN_CS: Median is 49 and 772 (34%) values are missing
- TEAM_BATTING_HBP: Median is 58 and 2085 (92%) values are missing
- TEAM_PITCHING_SO: Median 813.5 and 102 values are missing
- TEAM_FIELDING_DP: Median is 149 and 286 values are missing

Batters hit by pitch will be dropped completely from the dataset given 92% of the observations have missing values. For the remaining variable listed, we will create new imputed variables leveraging the median value for each variable. We chose to use the median to account for any extreme observations that may have strong influence or leverage on the mean. We will create new variables for the imputed values designated with the "IMP" prefix (e.g. IMP_TEAM_BATTING_SO). We will also create flags to indicate which observations have imputed values using the "M" prefix (e.g. M_TEAM_BATTING_SO).

Variable Transformations

In the Data Exploration phase, we highlighted five predictor variables that contained extreme observations including TEAM_BASERUN_SB, TEAM_PITCHING_HR, TEAM_PITCHING_BB, TEAM_PITCHING_SO, TEAM_PITCHING_H, and TEAM_FIELDING_E. To address these extreme observations, we will cover multiple techniques in this section including trimming, standardization, and logarithmic transformations. We will also explore the use of studentized residuals in the next section.

From a trimming perspective, we will use the 95^{th} and 99^{th} percentiles as the basis for trimming out the extreme observations. The 99^{th} percentile for each listed predictor variable will have the "T99" prefix and similarly, the 95^{th} percentile will have the "T95" prefix. We will also standardize each variable by calculating the z-scores for each observation. The standardized variables will contain the "STD" prefix and the trimmed standardized variables will have the "T_STD" prefix (trims values to only fall between -3 to 3). Lastly, we will transform each variable leveraging logarithmic transformations including natural and base 10 logarithm which will be contain the prefix "LN" and "LOG10", respectively. We will utilize each of these transformation approaches in order to see which method results in the best performing model. The third consideration was to bin the values of each predictor variable into specific bins. We chose the approach of leveraging the MEANS procedure to determine the 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles as our guidelines for the boundaries of each bin.

Handling of Extreme Observations Leveraging Studentized Residuals

In order to deal with some of the extreme observations in the dataset, we first built an OLS regression model to include all variables in order to output the studentized residuals for each observation. We then built in the logic to delete any observation that had a studentized residual value that exceeded the absolute value of 2. This process removed 99 observations from the dataset, leaving us with a total of 2,177 observations to train our predictive models.

Model Building

For this section, we will review several models and perform basic model validation techniques to assess the overall model fit. We will leverage key statistical measures such as Adjusted R-square, MSE, and p-values and perform residual analysis to determine what actions we might need to take to improve the overall model fit.

Model A: Simple model using all input variables and the imputed variables we created to handle missing values

The first model we will start with is to leverage all of the original predictor variables as well as the imputed variables we created to handle the missing values. We chose not to leverage any of the transformed variables in the first model so that we can perform the residual analysis to guide us in what type of transformation might be necessary. The resulting SAS output from our first OLS regression model is shown in Table 7 on the following page.

Table 7: Model A (*RMSE*: 11.59 *ADJRSQ*: 0.36 *CP*: 15 *AIC*: 10683.71)

The REG Procedure
Model: A
Dependent Variable: TARGET_WINS

Number of Observations Read	2177
Number of Observations Used	2177

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	163408	11672	86.86	<.0001
Error	2162	290530	134.38039		
Corrected Total	2176	453938			

Root MSE	11.59226	R-Square	0.3600
Dependent Mean	81.13000	Adj R-Sq	0.3558
Coeff Var	14.28850		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	28.34183	5.12926	5.53	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.04582	0.00363	12.61	<.0001	4.13360
TEAM_BATTING_2B	Doubles by batters	1	-0.02453	0.00840	-2.92	0.0035	2.46144
TEAM_BATTING_3B	Triples by batters	1	0.07652	0.01580	4.84	<.0001	2.99829
TEAM_BATTING_HR	Homeruns by batters	1	0.04577	0.02625	1.74	0.0814	40.66191
TEAM_BATTING_BB	Walks by batters	1	0.01928	0.00632	3.05	0.0023	8.99027
IMP_TEAM_BATTING_SO		1	-0.01180	0.00237	-4.98	<.0001	5.18797
IMP_TEAM_BASERUN_SB		1	0.02821	0.00409	6.89	<.0001	1.90365
IMP_TEAM_BASERUN_CS		1	-0.01326	0.01421	-0.93	0.3507	1.17668
TEAM_FIELDING_E	Errors	1	-0.01843	0.00240	-7.68	<.0001	3.99937
IMP_TEAM_FIELDING_DP		1	-0.12974	0.01177	-11.02	<.0001	1.34966
TEAM_PITCHING_BB	Walks allowed	1	-0.00446	0.00493	-0.91	0.3651	8.17383
TEAM_PITCHING_H	Hits allowed	1	-0.00044336	0.00041988	-1.06	0.2911	4.02498
TEAM_PITCHING_HR	Homeruns allowed	1	0.03484	0.02345	1.49	0.1375	33.52689
IMP_TEAM_PITCHING_SO		1	0.00223	0.00091438	2.43	0.0150	3.05470

As you can see in Table 7, the Adjusted R-square for Model A is 0.3558 and the MSE is 134.38. You will also notice that many of the predictor variables' p-values are higher than .05, indicating that those variables are not statistically significant for this model in their current form. We can also see the variance inflation factors for TEAM_BATTING_HR and TEAM_PITCHING_HR exceed 9, indicating multicollinearity exists among some of the variables. This may indicate that a new metric must be created or a variable must be removed in order to address the multicollinearity.

When assessing the residuals of the predictors against the dependent variable, we also notice that the errors are not constant and have a distinct pattern including TEAM_BATTING_HR, IMP_TEAM_BASERUN_SB, IMP_TEAM_BASERUN_CS, IMP_TEAM_FIELDING_E and TEAM_PITCHING_HR. We can perform transformations on these variables to see if the residuals improve and become constant for all observations. It is also clear that there are several outliers that exist in TEAM_PITCHING_BB, TEAM_PITCHING_H, and IMP_TEAM_PITCHING_SO. These issues can also be addressed through some of the variable transformations we discussed in the previous section.

Model B: Addition of new calculated variables and binning techniques

Given the significant issues covered with Model A, we must look for other methods to improve the model fit and deal with the multicollinearity that exists among the predictor variables. Some techniques to address multicollinearity include the use of new combination variables. For example, we can calculate the total number of bases earned by leveraging some of the other variables in the dataset including: TEAM_BATTING_HR (4 bases), TEAM_BATTING_3B (3 bases), TEAM_BATTING_2B (2 bases), TEAM_BATTING_1B (1 base), TEAM_BATTING_BB (1 base), TEAM_BATTING_SB (1 base), and TEAM_BATTING_CS (-1 base). We used these variables to create a new variable called TEAM_BASES_EARNED. As we mentioned in an earlier section, we also created bins for the numeric variables to see if that improves the overall model fit or not. We also chose to use forward selection to aid us with variable selection for Model B (please refer to Table 8 below for the output of Model B).

Table 8: Model B (*RMSE*: 10.51 *ADJRSQ*: 0.47 *CP*: 18.74 *AIC*: 10262.36)

The REG Procedure
Model: B
Dependent Variable: TARGET_WINS

Number of Observations Read	2177
Number of Observations Used	2177

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	216067	10289	93.21	<.0001
Error	2155	237871	110.38090		
Corrected Total	2176	453938			

Root MSE	10.50623	R-Square	0.4760
Dependent Mean	81.13000	Adj R-Sq	0.4709
Coeff Var	12.94987		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	33.12662	5.42614	6.11	<.0001	0
IMP_TEAM_BASERUN_SB		1	0.03401	0.00487	6.99	<.0001	3.27583
IMP_TEAM_BATTING_SO		1	-0.01522	0.00226	-6.74	<.0001	5.73242
IMP_TEAM_FIELDING_DP		1	-0.09887	0.01223	-8.08	<.0001	1.77367
IMP_TEAM_PITCHING_SO		1	-0.00134	0.00080848	-1.66	0.0973	2.90732
M_TEAM_BASERUN_CS		1	1.35766	0.79674	1.70	0.0885	2.73125
M_TEAM_BASERUN_SB		1	39.18903	1.88703	20.77	<.0001	3.36930
M_TEAM_BATTING_SO		1	8.01470	1.38674	5.78	<.0001	1.63049
M_TEAM_FIELDING_DP		1	2.54211	1.42112	1.79	0.0738	3.83552
TEAM_BASES_EARNED		1	0.02695	0.00303	8.89	<.0001	17.68344
TEAM_BATTING_1B		1	0.01308	0.00435	3.01	0.0026	5.70961
TEAM_BATTING_2B	Doubles by batters	1	-0.05030	0.00954	-5.27	<.0001	3.86987
TEAM_BATTING_3B	Triples by batters	1	0.04458	0.01586	2.81	0.0050	3.67619
TEAM_BATTING_BB	Walks by batters	1	0.01294	0.00644	2.01	0.0447	11.38575
TEAM_FIELDING_E	Errors	1	-0.06350	0.00351	-18.08	<.0001	10.41754
TEAM_PITCHING_BB	Walks allowed	1	-0.00970	0.00380	-2.55	0.0107	5.91410
TEAM_PITCHING_H	Hits allowed	1	0.00329	0.00043274	7.61	<.0001	5.20505
TEAM_BATTING_HR_2		1	1.03792	1.31061	0.79	0.4285	5.27491
TEAM_BATTING_HR_3		1	-2.19014	1.22784	-1.78	0.0746	5.49747
TEAM_BATTING_HR_4		1	-0.10049	0.84733	-0.12	0.9056	2.72750
TEAM_PITCHING_HR_1		1	-4.14703	1.90131	-2.18	0.0293	3.44989
TEAM_PITCHING_HR_2		1	-3.70386	1.30355	-2.84	0.0045	5.28408

Since we saw high VIFs in Model A for both TEAM_BATTING_HR and TEAM_PITCHING_HR, we chose to remove the two variables and replace them with the bins we created in the data preparation section. As we can see, that did end up correcting the VIFs for the bin indicator variables, however we see extremely high VIFs for TEAM_BASES_EARNED, TEAM_BATTING_1B, TEAM_BATTING_H and a few others that exceed 9. While the Adjusted R-square improved significantly compared to Model A, we must address the high multicollinearity that still exists among the predictor variables. One will also notice the signs on a few of the coefficients are counterintuitive including IMP_TEAM_FIELDING_DP, IMP_TEAM_PITCHING_SO, TEAM_BATTING_2B, TEAM_FIELDING_E, and TEAM_PITCHING_H. We will explore one more model to see if these issues can be addressed.

Model C: A combination of multiple variable transformation techniques

To address the issues we saw in both models A and B, we will explore the use of multiple variable transformations on in Model C including binning, trimming, and logarithmic transformations. Please refer to Table 9 to see the output of Model C.

Table 9: Model C (*RMSE*: 10.58 *ADJRSQ*: 0.46 *CP*: 13.61 *AIC*: 10297.35)

The REG Procedure
Model: C
Dependent Variable: TARGET_WINS

Number of Observations Read	2177
Number of Observations Used	2177

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	213100	8523.99069	76.13	<.0001
Error	2151	240838	111.96580		
Corrected Total	2176	453938			

Root MSE	10.58139	R-Square	0.4694
Dependent Mean	81.13000	Adj R-Sq	0.4633
Coeff Var	13.04251		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.72234	4.50414	0.16	0.8726	0
LN_TEAM_BASERUN_SB	1	6.46467	0.59526	10.86	<.0001	2.62218
IMP_TEAM_BATTING_SO	1	-0.01856	0.00177	-10.50	<.0001	3.46147
M_TEAM_BASERUN_SB	1	37.01918	2.20764	16.77	<.0001	4.54618
M_TEAM_BATTING_SO	1	11.43922	1.29912	8.81	<.0001	1.41070
TEAM_BASES_EARNED	1	0.03295	0.00127	25.84	<.0001	3.08486
T95_TEAM_FIELDING_E	1	-0.06754	0.00348	-19.40	<.0001	6.09336
T95_TEAM_PITCHING_H	1	-0.00576	0.00168	-3.43	0.0006	4.69944
TEAM_BASES_EARNED_1	1	-2.96731	1.13250	-2.62	0.0089	1.81456
TEAM_BATTING_1B_1	1	-2.63365	1.24004	-2.12	0.0338	1.44670
TEAM_BATTING_1B_2	1	-1.68035	0.69059	-2.43	0.0150	1.51554
TEAM_BATTING_2B_2	1	3.99684	0.69280	5.77	<.0001	1.48699
TEAM_BATTING_2B_3	1	4.29464	0.59164	7.26	<.0001	1.26944
TEAM_BATTING_3B_1	1	-7.76377	1.29179	-6.01	<.0001	1.86113
TEAM_BATTING_3B_2	1	-5.44978	1.00360	-5.43	<.0001	3.18484
TEAM_BATTING_3B_3	1	-5.57251	0.91518	-6.09	<.0001	3.03368
TEAM_BATTING_3B_4	1	-3.23268	0.75541	-4.28	<.0001	2.07462
TEAM_BATTING_BB_1	1	12.82368	2.11292	6.07	<.0001	3.51330
TEAM_BATTING_BB_2	1	2.55878	0.68209	3.75	0.0002	1.42881
TEAM_FIELDING_E_4	1	-2.19646	0.63844	-3.44	0.0006	1.48551
TEAM_PITCHING_H_1	1	4.52152	1.18663	3.81	0.0001	1.38110
TEAM_PITCHING_H_4	1	-1.91656	0.58039	-3.30	0.0010	1.23067
team_baserun_sb_1	1	5.40011	1.24409	4.34	<.0001	1.50572
team_baserun_sb_3	1	-2.02191	0.62420	-3.24	0.0012	1.53403
team_baserun_sb_4	1	-3.20413	0.65914	-4.86	<.0001	1.44755
team_baserun_cs_3	1	1.61579	0.61106	2.64	0.0082	1.74451

Our criteria for variable selection in model three include the following:

1. Had a p-value less than .05
2. Had a VIF less than 9
3. Had an appropriate sign for the betas that aligned with the theoretical effect provided in the moneyball data dictionary

In order to arrive at a final listing of predictor variables that met these three criteria, we had to perform several variable transformations. These transformations also aided us in reducing the variance shown in the residual plots for each predictor variables. For this final model, our QQ plot indicated normality with regards to the residuals. For the TEAM_BASERUN_SB variable, we had to perform a natural logarithmic transformation to reduce the variance of the residuals. For TEAM_FIELDING_E we had to apply a 95th percentile trimming transformation in order to reduce some of the outliers in the dataset. The same trimming transformation was also applied to the TEAM_PITCHING_H variable as well to handle the extreme observations. The remaining variables included in the model are subsets of the bins we created for each numeric predictor variable. We removed several bin indicator variables as they did not meet the three criteria listed above.

Model Selection

After reviewing each model in depth in the prior section, we will review the model fit statistics in order to finalize on a model to select for deployment. The primary measures we will use to assess the models include the following:

1. Adjusted R-Square
2. RMSE
3. Mallow's Cp
4. AIC
5. All predictors with p-values less than 0.05
6. All predictors with VIFs less than 9
7. All predictors with appropriate sign for betas

Table 10: Model Selection Criteria

	Model A	Model B	Model C
Adjusted R-Square	0.3558	0.4709	0.4633
RMSE	11.5923	10.5062	10.5814
Mallow's Cp	15.0000	18.7499	13.6120
AIC	10683.7166	10262.3591	10297.3507
P-values < .05	No	No	Yes
VIFs < 9	No	No	Yes
Appropriate Signs (+/-)	No	No	Yes

Given the criteria we have outlined above, only Model C meets all criteria and has the lowest Mallow's Cp value for all models. Therefore, we selected Model C as our choice to be deployed into production.

Conclusion

Now that we have arrived at a final model for deployment, we will create the necessary data step to predict the number of wins with a test dataset provided by the instructor. The key takeaway from this assignment is that you should not assume you are working with clean data before jumping into the model building phase. You first must learn about the data. If you are dealing with data in a new industry or domain outside of your sphere of knowledge, then you must perform the necessary due diligence to become familiar with the data. This can be achieved by research online or by reaching out to an expert in the particular field you are exploring for the model building exercise. Luckily, I did play baseball for several years and am very familiar with the game. I did reference online materials to see what insight anyone has documented with regards to baseball statistics. The process of exploring the data and preparing the data were critical steps before I could move into the model building phases. Overall, this assignment was an excellent learning opportunity and I look forward to the challenges that lay ahead.