DATA 621 Homework #1 - Moneyball

Daniel Hong, Mauricio Alarcon, Maxwell Wagner

September 11, 2016

Contents

Data Exploration

The Moneyball dataset includes roughly 2200 records dating from 1871 to 2006. Each record includes statistics regarding the performance of the team adjusted for a 162 game season. Table 1 below provides the mean, standard deviation, median, max, min, and completeness of each variable. Looking at the completeness column the most noticable outliers are the `Batters hit by pitch` and `Caught stealing` variables. These will likely be removed in the models due to lack of information.

Table 1 :

	mean	sd	median	max	min	completeness
TARGET_WINS	80.79086	15.75215	82.0	146	0	1.00
TEAM_BATTING_H	1469.26977	144.59120	1454.0	2554	891	1.00
TEAM_BATTING_2B	241.24692	46.80141	238.0	458	69	1.00
TEAM_BATTING_3B	55.25000	27.93856	47.0	223	0	1.00
TEAM_BATTING_HR	99.61204	60.54687	102.0	264	0	1.00
TEAM_BATTING_BB	501.55888	122.67086	512.0	878	0	1.00
TEAM_BATTING_SO	735.60534	248.52642	750.0	1399	0	0.96
TEAM_BASERUN_SB	124.76177	87.79117	101.0	697	0	0.94
TEAM_BASERUN_CS	52.80386	22.95634	49.0	201	0	0.66
TEAM_BATTING_HBP	59.35602	12.96712	58.0	95	29	0.08
TEAM_PITCHING_H	1779.21046	1406.84293	1518.0	30132	1137	1.00
TEAM_PITCHING_HR	105.69859	61.29875	107.0	343	0	1.00
TEAM_PITCHING_BB	553.00791	166.35736	536.5	3645	0	1.00
TEAM_PITCHING_SO	817.73045	553.08503	813.5	19278	0	0.96
TEAM_FIELDING_E	246.48067	227.77097	159.0	1898	65	1.00
TEAM_FIELDING_DP	146.38794	26.22639	149.0	228	52	0.87

In order to view the distribution of variables, a scaling method was used. The scaling is necessary to correct the values in the `Hit's Allowed` and `Strikeout by Pitchers` variable. It also provides evidence for the later use of transformations to balance the variables.

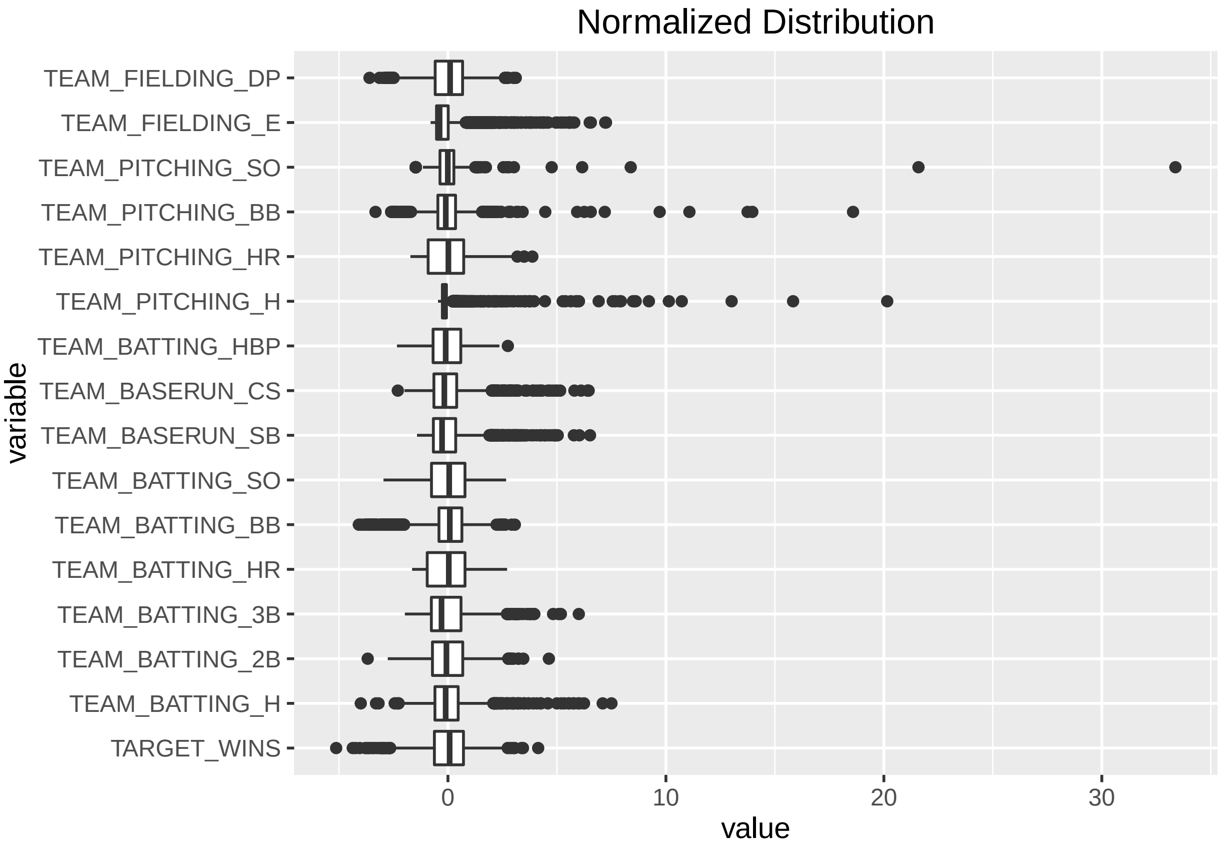

Table 2 below shows the correlation between each variable and the `Target_wins` (number of wins) variable. Predictably, some variables are near zero correlation, which means they have little impact on the wins variable.

Table 2 :

	r
TARGET__WINS	1.0000000
TEAM__BATTING__H	0.2578130
TEAM__BATTING__2B	0.1624482
TEAM__BATTING__3B	0.0811299
TEAM__BATTING__HR	0.1102447
TEAM__BATTING__BB	0.1661288
TEAM__BATTING__SO	-0.0515269
TEAM__BASERUN__SB	0.0775086
TEAM__BASERUN__CS	-0.0052072
TEAM__BATTING__HBP	0.0195018
TEAM__PITCHING__H	0.1514105
TEAM__PITCHING__HR	0.1155198
TEAM__PITCHING__BB	0.1505342
TEAM__PITCHING__SO	-0.0633447

	r
TEAM_FIELDING_E	-0.0884986
TEAM_FIELDING_DP	-0.0352128

Data Preparation

Build Models

Select Models

Appendix