Final Paper

Kin Tong Ng, Euijun Kim, Taylor Vassar

Feature	Response
Name	Kin Tong Ng, Euijun Kim, Taylor Vassar
Kaggle Nickname	Kin Tong Ng-Lec-2
Kaggle Rank	Between 37 and 38(Since one of the groupmates left so the current ranking on Kaggle is not using our final model.)
Kaggle R^2	0.79416
Total Number of Predicators	10
Total number of Betas	16
BIC	5345.813
Complexity Grade	94

Abstract

This project utilizes the statistical practice of multiple linear regression, specifically to predict the winning proportion of all NCAA (National Collegiate Athletic Association) Division I basketball teams, given a sample of their games and 22 predictor variables. This game data is compiled from the years of 2013-2021, all played in the United States. We built this model based off of a training dataset and resulted in an R square value of 0.79416, resulting in ranking between 37 and 38.

Introduction

Regarding NCAA Division I college basketball teams, our goal was to predict the winning proportion of all NCAA Division I basketball teams given a sample of their games and using 22 predictors, either categorical or quantitative. Since SEED is not used here so the dataset basically has 21 predicators, which is listed below:

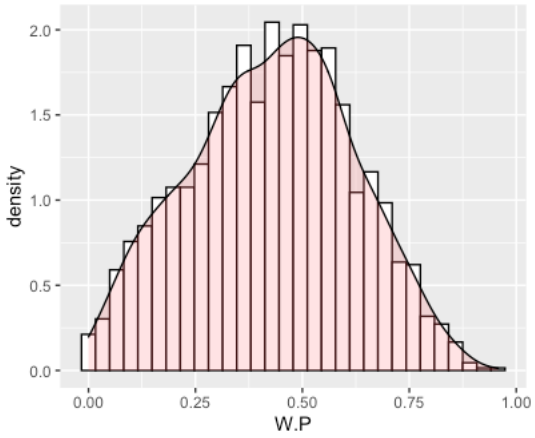
	Variables	Type
1	X500.Level	Categorical
2	ADJOE	Numerical
3	ADJDE	Numerical
4	EFG_O	Numerical
5	EFG_D	Numerical
6	TOR	Numerical
7	TORD	Numerical
8	ORB	Numerical
9	DRB	Numerical
10	FTR	Numerical
11	FTRD	Numerical
12	X2P_O	Numerical
13	X2P_D	Numerical
14	X3P_O	Numerical
15	X3P_D	Numerical
16	WAB	Numerical
17	YEAR	Numerical
18	NCAA	Categorical
19	Power.Rating	Categorical
20	Adjusted.Tempo	Numerical
21	W.P	Numerical

Methodology

The Response Variable: The Wining Proportion

Before constructing our model, we first checked the data distribution first. For our response variable, W.P, we expect to see it follows the normal distribution. If it doesn't follow the normal distribution, then we might need to apply transformations on the response variable. For the Wining Proportion in the training data set, we have created a histogram. According to the plot, we believe that it approximately follows normal distribution so no transformation for the response variable is needed at this point.

Histogram for Winning Proportion:



The Numerical Predictors

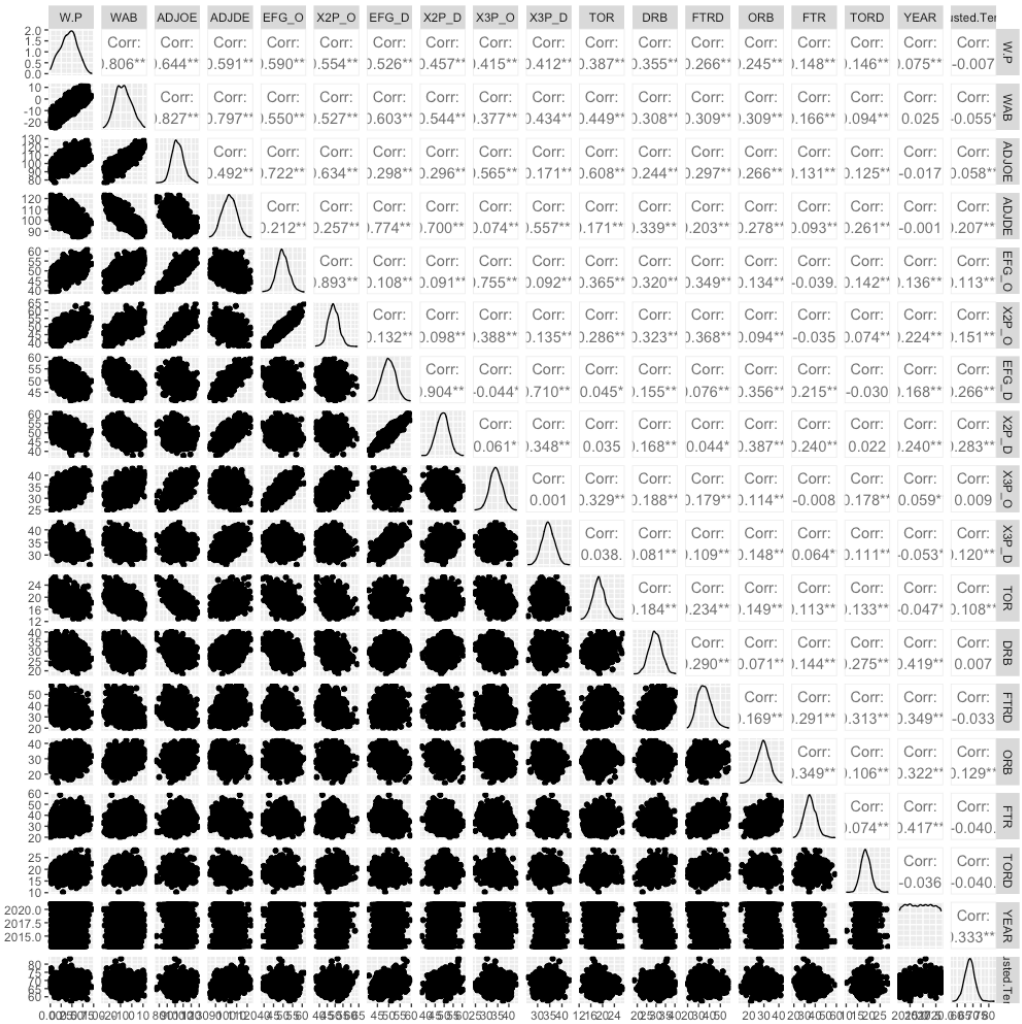
After we had checked the response variable, we looked into the numerical variables. There are 17 numeric variables for this data set. If we want to choose the predictors for our model, we need to first look at their correlation among these predictors and the response variable. Below is the correlation of all numeric variables with respect to the response variable.

Table2: Correlation among numeric variables

	Numeric.predictors	Correlation
1	WAB	0.806
2	ADJOE	0.644
3	ADJDE	-0.591
4	EFG_O	0.590
5	X2P_O	0.554
6	EFG_D	-0.526
7	X2P_D	-0.457
8	X3P_O	0.415
9	X3P_D	-0.412
10	TOR	-0.387
11	DRB	-0.355
12	FTRD	-0.266
13	ORB	0.245
14	FTR	0.148
15	TORD	0.146
16	YEAR	0.075
17	Adjusted.Tempo	-0.007

Furthermore, here is the correlation matrix among all numerical variables and the response variable.

Scatterplot Matrix



Based on the correlation plots above, we can see that the variable, WAB, has the highest correlation with the response variable, W.P. Moreover, ADJOE, ADJDE, EFG_O, X2P_O, EFG_D have the next highest correlation with the response variable. However, the variable ADJOE has a very high correlation, about 0.827, with WAB. Moreover, X2P_O is also greatly correlated to EFG_O, given the fact that the correlation with them is 0.893. Thus, if we create a model that includes all the numeric predictors, collinearity could be a problem. Nevertheless, we still create a multiple linear model based on all the numeric predictors. Below is the summary of the model: (the zero's in all tables are actually very small values)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.8381	2.2831	-0.3671	0.7136
WAB	0.0227	9e-04	25.0438	0
ADJOE	-0.0113	0.001	-11.3494	0
ADJDE	0.0118	0.001	11.7366	0
EFG_O	0.0414	0.0076	5.4373	0
X2P_O	-0.0123	0.0048	-2.5373	0.0112
EFG_D	-0.0161	0.0102	-1.5836	0.1134
X2P_D	-0.0022	0.0065	-0.335	0.7377
X3P_O	-0.0071	0.004	-1.761	0.0784
X3P_D	-0.0038	0.0054	-0.6976	0.4855
TOR	-0.0186	0.0017	-11.2477	0
DRB	-0.0116	9e-04	-12.2563	0
FTRD	-0.0025	4e-04	-6.259	0
ORB	0.0087	8e-04	10.8802	0
FTR	0.002	4e-04	4.3566	0
TORD	0.0221	0.0015	14.8887	0
YEAR	6e-04	0.0011	0.4931	0.622
djusted.Tempo	0.0019	7e-04	2.6393	0.0084

	Observation	Residual.Std..Error	R.square	Adjusted.R.sauare
1	2000	0.08595	0.799942331746891	0.798226398164498

The model of all numeric predictors has the R square of 0.799 and the adjusted R square of 0.798. The summary of the model shows that the intercept and the variables, EFG_D, X2P_D, X2P_O, X3P_O, X3P_D, YEAR, are not significant because their p-values are not less than 0.05. Therefore, we considered removing them.

After Removing some variables:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6024	0.0989	6.0907	0
WAB	0.0249	9e-04	27.0498	0
ADJOE	-0.0188	9e-04	-21.8349	0
ADJDE	0.0024	7e-04	3.3441	8e-04
EFG_O	0.0377	0.002	19.3125	0
X2P_O	-0.0046	0.0014	-3.3279	9e-04
TOR	-0.025	0.0016	-15.7356	0
DRB	-0.0067	8e-04	-8.5995	0
FTRD	-0.0015	4e-04	-3.6893	2e-04
ORB	0.0127	7e-04	17.6852	0
FTR	0.0028	4e-04	6.4714	0
TORD	0.0106	0.0012	8.9503	0
Adjusted.Tempo	0.0015	7e-04	2.1348	0.0329

	Observation	Residual.Std..Error	R.square	Adjusted.R.sauare
1	2000	0.0902116730521755	0.779032746980701	0.777698269358038

After taking out some predictors, now we see that all the predictors in the new model are significant, and the R square of this model is about 0.779. However, we still need to check the VIF of this model to avoid collinearity which would make our model invalid. Below is the VIF table of this model:

	vif(p2)
WAB	9.7975
ADJOE	9.8651
ADJDE	5.2944
EFG_O	9.0663
X2P_O	5.476
TOR	2.6855
DRB	1.5935
FTRD	1.5886
ORB	2.233
FTR	1.3481
TORD	1.6828
Adjusted.Tempo	1.1501

According to the table, we can tell that the variable, WAB, ADJOE, ADJDE, EFG_O, X2P_O, have VIF over 5, which means we might need to remove some of them. Therefore, for the next step, we started to remove the variable with the highest VIF one by one while checking the VIF every time we remove one. As the predictor WAB has the highest correlation with the response variable, we tried to keep it and delete other predictors. Below is the model that all predictors have no VIF over 5:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0115	0.0777	-0.1483	0.8821
WAB	0.0154	6e-04	27.2124	0
X2P_O	0.011	9e-04	12.385	0
TOR	-0.008	0.0014	-5.8012	0
DRB	-0.0089	8e-04	-10.9591	0
FTRD	-0.0013	5e-04	-2.8842	0.004
ORB	0.0039	7e-04	5.6482	0
FTR	0.0024	5e-04	4.773	0
TORD	0.0141	0.0012	11.5949	0
Adjusted.Tempo	3e-04	8e-04	0.3429	0.7317

	Observation	Residual.Std..Error	R.square	Adjusted.R.sauare
1	2000	0.103862433347292	0.706657736983478	0.705331063432147

Thus, we got a model with no VIF over 5; however, the R square decreased by 0.07 from the previous model. Since we wanted to improve the model, we decided to shrink the number of predictors first and add back some of the predictors that are helpful with predicting the winning proportion.

Creating new variable

The one way we can think of to reduce the number of predictors is to combine some of the predictors. After exploring the data, we found out that some predictors are similar to others. For example, EFG_O is the effective field Goal percentage shot, and EFG_D is the effective field goal percentage allowed; moreover, TOR is the turnover percentage allowed and TORD is the turnover percentage committed. Therefore, we combined these variables by subtracting each two of them (we also investigated other methods like adding, multiplying, and dividing, but we discovered that subtracting works the best for our model). Therefore, we create new variables like new_EFG from subtracting EFG_O by EFG_D. After we created new variables from the original data set, we decided to replace the variable of the previous model with the new variable. Below is the new model after the replacement of the variable. Moreover, among the variables, X2P_O, X2P_D, X3P_O, X3P_D, we discovered that X2P_O is the most helpful one for our model, and if we add the rest of these variables to our model, the VIF will exceed five.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0402	0.0605	-0.6647	0.5063
WAB	0.0148	6e-04	26.4995	0
X2P_O	0.0118	8e-04	14.0077	0
new_TOR	-0.0115	0.001	-11.4121	0
new_RB	0.0061	5e-04	11.0918	0
new_FTR	0.0016	4e-04	4.0893	0
Adjusted.Tempo	0	8e-04	0.0216	0.9828

	vif(p6)
WAB	2.673
X2P_O	1.51
new_TOR	1.4865
new_RB	1.4385
new_FTR	1.3335
Adjusted.Tempo	1.0624

	Observation	Residual.Std..Error	R.square	Adjusted.R.sauare
1	2000	0.1046	0.7018	0.7009

Above is the model that we obtained with the new variables. Our new model has six predictors with approximate 0.7 adjusted R square while our previous model has nine predictors and similar value of adjusted R square. Therefore, for simplicity, we would prefer this model. To improve the model, we considered to add back some variables that have high correlation with the response variable.

Adding new_EFG_O back to the model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1824	0.058	3.1446	0.0017
WAB	0.0074	7e-04	10.8493	0
X2P_O	0.0022	0.001	2.2655	0.0236
new_EFG	0.0172	0.001	17.1421	0
new_TOR	-0.0167	0.001	-16.8543	0
new_RB	0.0068	5e-04	13.33	0
new_FTR	0.0022	4e-04	6.0526	0
Adjusted.Tempo	0.0029	7e-04	3.9515	1e-04

	vif(p7)
WAB	4.524
X2P_O	2.2773
new_EFG	4.2022
new_TOR	1.6378
new_RB	1.4495
new_FTR	1.3467
Adjusted.Tempo	1.1213

	Observation	Residual.Std..Error	R.square	Adjusted.R.sauare
1	2000	0.0977	0.7401	0.7392

Adding the variable new_ADJ(ADJOE – ADJDE + 60):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3373	0.0521	6.4702	0
WAB	0.0225	9e-04	25.0477	0
new_ADJ	-0.0116	5e-04	-22.7758	0
X2P_O	0.0024	9e-04	2.7381	0.0062
new_EFG	0.0222	9e-04	24.16	0
new_TOR	-0.0205	9e-04	-22.7828	0
new_RB	0.0092	5e-04	19.7222	0
new_FTR	0.0021	3e-04	6.4344	0
Adjusted.Tempo	0.0022	7e-04	3.3869	7e-04

	vif(p8)
WAB	10.0094
new_ADJ	9.8996
X2P_O	2.2775
new_EFG	4.461
new_TOR	1.6951
new_RB	1.5275
new_FTR	1.347
Adjusted.Tempo	1.1237

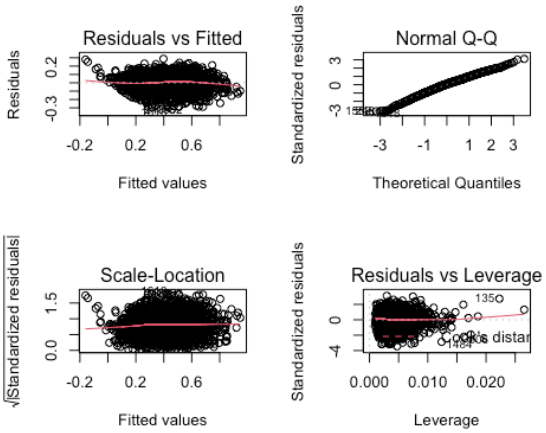
	Observation	Residual.Std..Error	R.square	Adjusted.R.sauare
1	2000	0.0871	0.7938	0.793

We noticed that after adding the variable new_ADJ, the R square increased significantly but the VIF also became invalid. Therefore, we started to think that is there any way to combine this variable with other variable? After trying couples of combinations, we discover that if we add WAB and ADJDE together and create a new variable WAB_AD, we will get higher R square and with no VIF over five. Moreover, we also found out that if we just replace new_EFG and new_TOR by the original variables, we can obtain higher R square. In addition, we also add the variable YEAR, and we also remove X2P_O to avoid collinearity.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15.6414	1.6791	-9.3154	0
WAB_AD	0.0126	7e-04	18.0829	0
EFG_D	-0.0293	9e-04	-32.9628	0
EFG_O	0.0191	9e-04	22.2541	0
TOR	-0.0138	0.0012	-12.027	0
TORD	0.0271	0.001	27.4578	0
new_RB	0.0082	4e-04	18.3817	0
new_FTR	0.0026	3e-04	8.0097	0
YEAR	0.0075	8e-04	8.9772	0

	vif(p9)
WAB_AD	2.0848
EFG_D	1.5529
EFG_O	1.6777
TOR	1.3412
TORD	1.1053
new_RB	1.2379
new_FTR	1.2438
YEAR	1.0806

	Observation	Residual.Std..Error	R.square	Adjusted.R.sauare
1	2000	0.0924	0.7679	0.7669



Above is the summary of the model and the diagnostic plots of it. The model that has no VIF over 5, we had 0.739 adjusted R square with seven predictors. After we adjusted the predictors, now we obtained 0.767 adjusted R square with 8 predicators. We all think that's a significant improvement. Based on the diagnostic plots, the normality of the model looks good, and the errors also look independently distributed even though it slightly showed a double bow shape. Moreover, based on the graph, the variance is also constant here, and there might be a few good and bad leverage points for the model. We will look back to this again after making changes to the model

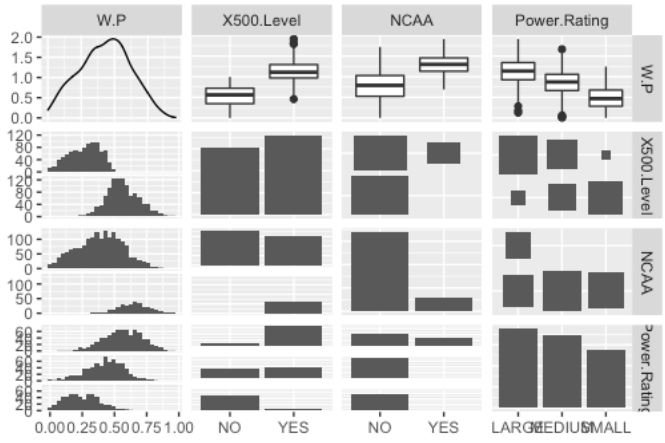
For the next step, we started to look at categorical predictors and interaction terms.

Categorical Variable

For this data set, we have three categorical predictors, and they are X500.Level, NCAA, and Power. Rating. Below we show the boxplot and density matrix for those three categorical variables with response to winning proportion. According to the matrix, we consider that the original categories are already good enough for predicting the winning price; thus, we decided not to do transformation on the categorical variables.

In addition, based on the matrix, we also discovered that NCAA has the similar effect on the response variable as X500.Level does. Therefore, we will just include one of them in our model.

Scatterplot matrix for categorical variables:



In the model below, the R square increased from 0.767 to 0.7964, compared to the previous model. We also check the VIF, and there are no VIF over five, which means collinearity is also not significant for this model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.1146	1.5872	-7.6326	0
WAB_AD	0.0099	7e-04	14.3136	0
EFG_D	-0.0226	0.001	-22.1233	0
EFG_O	0.0153	9e-04	16.9856	0
TOR	-0.0117	0.0012	-9.9	0
TORD	0.0218	0.001	21.6648	0
new_RB	0.0068	4e-04	15.2566	0
new_FTR	0.0021	3e-04	6.6299	0
YEAR	0.0058	8e-04	7.3636	0
X500.LevelYES	0.0976	0.0062	15.8207	0
Power.RatingMEDIUM	0.0212	0.0055	3.8296	1e-04
Power.RatingSMALL	0.0078	0.008	0.967	0.3337

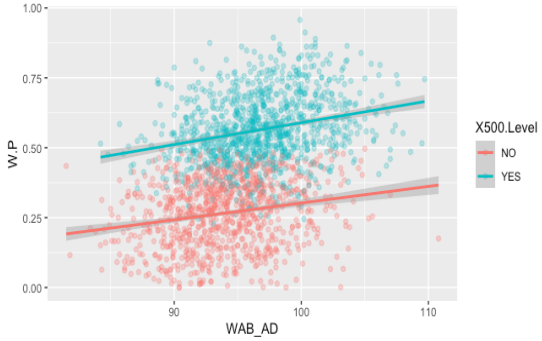
	GVIF	Df	GVIF^(1/(2*Df))
WAB_AD	2.3328	1	1.5274
EFG_D	2.3592	1	1.536
EFG_O	2.0967	1	1.448
TOR	1.6161	1	1.2713
TORD	1.3213	1	1.1495
new_RB	1.4093	1	1.1871
new_FTR	1.3016	1	1.1409
YEAR	1.1036	1	1.0505
X500.Level	2.5382	1	1.5932
Power.Rating	2.8917	2	1.304

	Observation	Residual.Std..Error	R.square	Adjusted.R.sauare
1	2000	0.0863	0.7976	0.7964

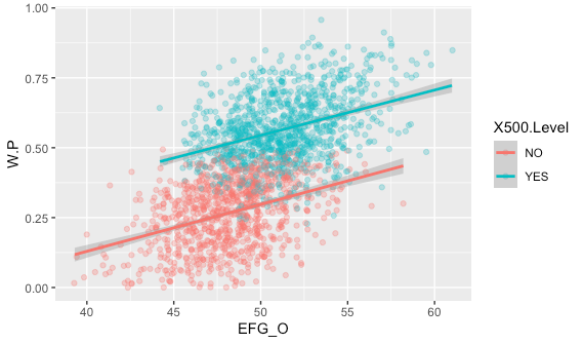
Exploring Interaction

For interaction terms, we first constructed interaction plot for some of the numeric variables with the X500.Level and Power. Rating. In the plots below, we did not see a strong or significant interaction with these two categorical variables. For example, in the interaction plot of X500.Level and the variable WAB_AD, the slops of different lines do not seem to have significant difference, and it is similar for the other plots. Nevertheless, we still tried to add the interaction terms to our model, and it turned out that only the interaction terms are not significant, the X500.Level itself also turned to insignificant. Thus, we decide not to add interaction terms due to simplicity.

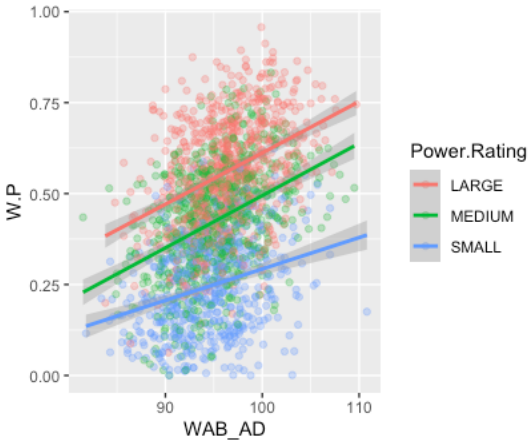
Interaction plot between X500.Level and WAB_AD.



Interaction plot between X500.Level and EFG_O



Interaction plot between X500.Level and WAB_AD

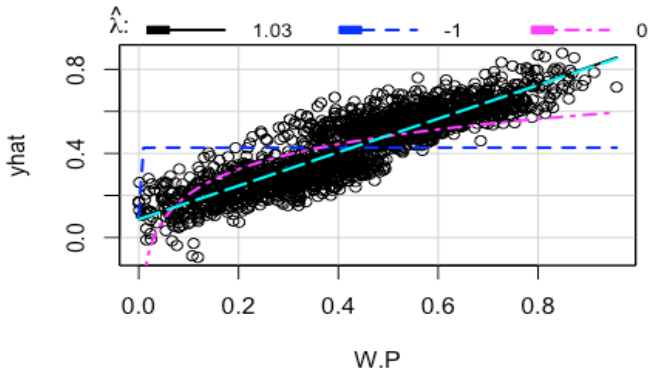


	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.472	1.6034	-7.7785	0
WAB_AD	0.0113	0.0016	7.216	0
EFG_D	-0.0221	0.0014	-15.9958	0
EFG_O	0.0152	9e-04	16.7465	0
TOR	-0.0116	0.0012	-9.7695	0
TORD	0.0219	0.001	21.6785	0
new_RB	0.0067	4e-04	15.0566	0
new_FTR	0.0021	3e-04	6.5962	0
YEAR	0.0059	8e-04	7.454	0
X500.LevelYES	0.2606	0.1274	2.0449	0.041
Power.RatingMEDIUM	0.0393	0.1143	0.3434	0.7313
Power.RatingSMALL	0.2011	0.1486	1.3536	0.176
WAB_AD:X500.LevelYES	-0.0011	0.0015	-0.7367	0.4614
EFG_D:X500.LevelYES	-0.0012	0.0018	-0.6796	0.4969
_AD:Power.RatingMEDIUM	-2e-04	0.0012	-0.1504	0.8805
AB_AD:Power.RatingSMALL	-0.002	0.0016	-1.3086	0.1908

Transformation

Besides interaction, we also wanted to investigate the possible transformation that improve our model. Therefore, we first try the inverse response plot transformation. According to the plot below, the response variable does not seem to need any transformation, which matches

what we discussed before, that the response variable follows an approximate normal distribution.



	LRT	df	pval
LR test, lambda = (0 0 0 0 0 0 0 0 0)	2859.56492	9	< 2.22e-16
LR test, lambda = (1 1 1 1 1 1 1 1 1)	65.11818	9	1.3683e-10

Furthermore, we are also interested in the transformation of the numeric predictors. Thus, we checked if there are any power transformations needed for our model. As a result, we found out that not applying any transformation could be the best for our current model. Below is the result of testing inverse response transformation and power transformations:

	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
W.P	0.968	1	0.917	1.018
WAB_AD	0.309	0	-0.338	0.955
EFG_D	0.813	1	0.281	1.344
EFG_O	0.774	1	0.311	1.238
TOR	0.181	0	-0.113	0.476
TORD	0.466	0.5	0.205	0.727
new_RB	1.341	1	0.97	1.713
new_FTR	1.418	1.42	1.132	1.704
YEAR	-2.142	-2.14	-2.218	-2.066

Polynomial

As our last step, we would like to explore the polynomial degree of the numeric predictors. Hence, we decided to apply polynomial degree 3 to all the numeric predictors in our model and delete those insignificant polynomial degrees one by one until all the betas are significant. The charts below are the first and eighth attempts and the final attempt. In our final attempt, we can see some betas are not exactly significant, but we still kept it. For example, the degree 2 of YEAR is not significant, but we still kept it since the degree 3 of

YEAR is significant. The adjusted R^2 increased from 0.7976 to about 0.8 after applying polynomial degrees.

First attempt:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3658	0.0056	65.6214	0
poly(WAB_AD, 3)1	1.9726	0.1345	14.6611	0
poly(WAB_AD, 3)2	-0.1014	0.0928	-1.0933	0.2744
poly(WAB_AD, 3)3	-0.1038	0.0877	-1.1837	0.2367
poly(EFG_D, 3)1	-3.0465	0.1347	-22.623	0
poly(EFG_D, 3)2	0.1039	0.0881	1.1804	0.238
poly(EFG_D, 3)3	0.1638	0.0873	1.8768	0.0607
poly(EFG_O, 3)1	2.0474	0.1269	16.1398	0
poly(EFG_O, 3)2	0.2068	0.0926	2.2337	0.0256
poly(EFG_O, 3)3	0.0442	0.0887	0.4977	0.6187
poly(TOR, 3)1	-1.125	0.1111	-10.1288	0
poly(TOR, 3)2	0.0811	0.0904	0.8972	0.3697
poly(TOR, 3)3	0.2132	0.0872	2.4442	0.0146
poly(TORD, 3)1	2.1548	0.0998	21.5945	0
poly(TORD, 3)2	0.2291	0.0871	2.6308	0.0086
poly(TORD, 3)3	-0.1011	0.0867	-1.1663	0.2436
poly(new_RB, 3)1	1.5638	0.1027	15.2232	0
poly(new_RB, 3)2	0.1427	0.0879	1.6228	0.1048
poly(new_RB, 3)3	-0.0178	0.086	-0.2063	0.8365
poly(new_FTR, 3)1	0.6528	0.0979	6.669	0
poly(new_FTR, 3)2	-0.1502	0.0861	-1.7443	0.0813
poly(new_FTR, 3)3	0.0323	0.0858	0.3767	0.7064
poly(YEAR, 3)1	0.7036	0.0903	7.7903	0
poly(YEAR, 3)2	-0.0461	0.0897	-0.5141	0.6072
poly(YEAR, 3)3	-0.452	0.0885	-5.1078	0
X500.LevelYES	0.0938	0.0062	15.1029	0
Power.RatingMEDIUM	0.0256	0.0056	4.5662	0
Power.RatingSMALL	0.008	0.0081	0.9864	0.3241

Eighth attempt:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3659	0.0056	65.7453	0
poly(WAB_AD, 1)	1.983	0.1341	14.7848	0
poly(EFG_D, 2)1	-3.0353	0.1339	-22.6735	0
poly(EFG_D, 2)2	0.1033	0.0877	1.1786	0.2387
poly(EFG_O, 2)1	2.0318	0.1255	16.1877	0
poly(EFG_O, 2)2	0.1749	0.0891	1.9633	0.0498
poly(TOR, 3)1	-1.122	0.1102	-10.184	0
poly(TOR, 3)2	0.0694	0.0885	0.7844	0.4329
poly(TOR, 3)3	0.2263	0.0861	2.628	0.0087
poly(TORD, 2)1	2.1561	0.0991	21.7634	0
poly(TORD, 2)2	0.2095	0.0867	2.4172	0.0157
poly(new_RB, 2)1	1.5712	0.1019	15.4199	0
poly(new_RB, 2)2	0.1428	0.0878	1.6271	0.1039
poly(new_FTR, 2)1	0.6607	0.0978	6.7585	0
poly(new_FTR, 2)2	-0.1474	0.086	-1.7131	0.0869
poly(YEAR, 3)1	0.706	0.0903	7.8224	0
poly(YEAR, 3)2	-0.0362	0.0895	-0.4048	0.6857
poly(YEAR, 3)3	-0.4477	0.0883	-5.0717	0
X500.LevelYES	0.0945	0.0061	15.4026	0
Power.RatingMEDIUM	0.0248	0.0056	4.4572	0
Power.RatingSMALL	0.0072	0.008	0.9034	0.3664

Final attempt:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.7503	0.0851	-8.8161	0
X500.LevelYES	0.1366	0.0935	1.4604	0.1443
EFG_D	-0.0234	0.001	-22.656	0
EFG_O	0.0147	9e-04	16.2775	0
poly(TOR, 3)1	-1.1212	0.11	-10.1927	0
poly(TOR, 3)2	0.1042	0.089	1.1712	0.2417
poly(TOR, 3)3	0.2472	0.086	2.8726	0.0041
poly(TORD, 2)1	2.1437	0.0992	21.6175	0
poly(TORD, 2)2	0.2221	0.0862	2.5756	0.0101
WAB_AD	0.0107	9e-04	12.5113	0
new_RB	0.0068	4e-04	15.3778	0
new_FTR	0.0021	3e-04	6.7372	0
poly(YEAR, 3)1	0.6983	0.0902	7.7407	0
poly(YEAR, 3)2	-0.0312	0.0897	-0.348	0.7279
poly(YEAR, 3)3	-0.4496	0.0883	-5.0906	0
Power.RatingMEDIUM	0.0231	0.0055	4.183	0
Power.RatingSMALL	0.0089	0.008	1.1163	0.2644
X500.LevelYES:WAB_AD	-4e-04	0.001	-0.4519	0.6514

Results and Discussion

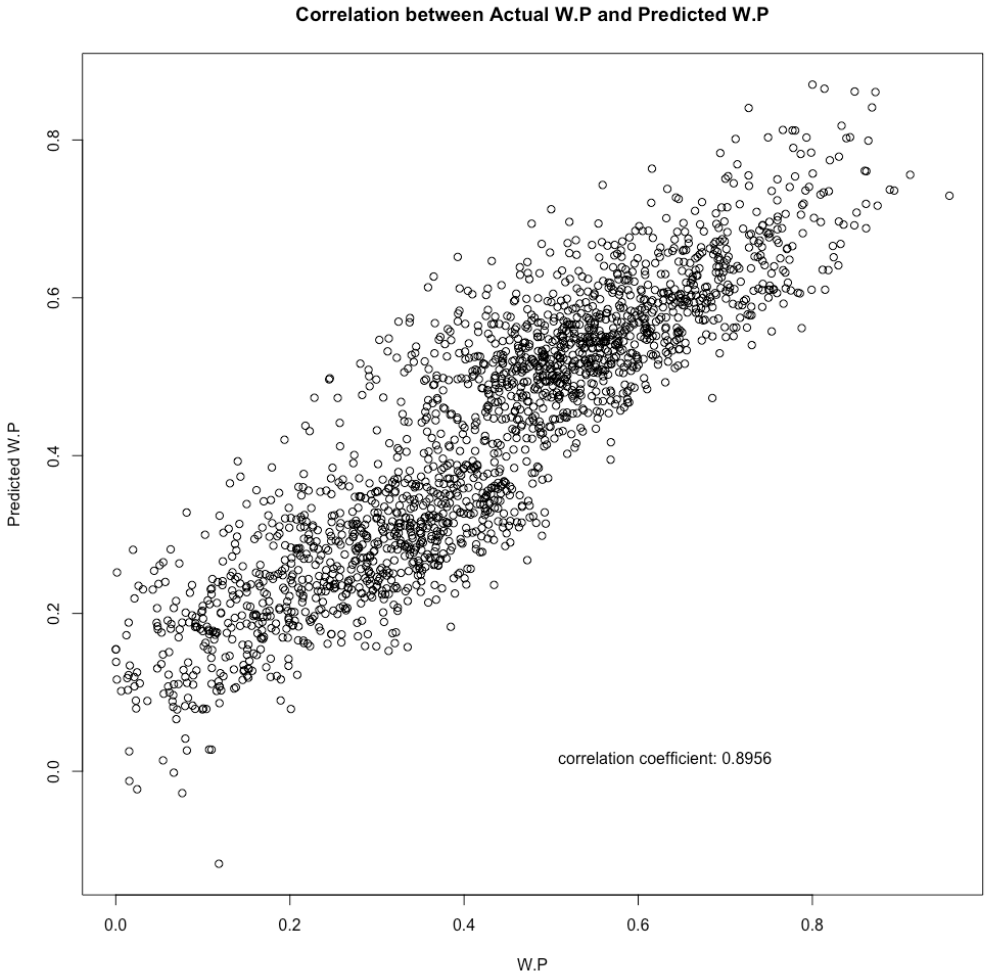
Final Model

After all the attempts above, we conclude our final model as:

$$\widehat{W.P} = -0.64 + 0.091(X500.Level) - 0.24(EFG\_D) + 0.0147(EFG\_O) - 1.004(TOR) + 0.117(TOR)^2 + 0.218(TOR)^3 + 1.938(TORD) + 0.18(TORD)^2 + 0.01(WAB\_AD) + 0.0064(new\_RB) + 0.0022(new\_FTR)$$

After constructing our final model, we wanted to test the accuracy of our model by splitting the data into two datasets based on the seed to "1000". We split the data in a way that there are 500 observation in the testing data and 1500 observation in the training data. Thus, we can see the adjusted R square is about 0.7989 and even the 2nd polynomial degree of TOR is insignificant, the 3rd polynomial degree of TOR still shows that the predictor is significant. In the testing data, we also obtained similar R square, about 0.8065; however, the 2nd polynomial degree of TORD and the 3rd polynomial degree become insignificant. Even there might a problem with these two betas, we still use this model due to its accuracy shown in both datasets. Moreover, the correlation between the predicted winning proportion and the

actual winning proportion is 0.8957, so we are confident that our model will be at least accurate to some degree.



Training Model (adj $R^2 = 0.7989$).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.6404	0.0791	-8.0983	0
X500.LevelYES	0.0908	0.0072	12.6585	0
EFG_D	-0.0239	0.0012	-19.6953	0
EFG_O	0.0147	0.001	14.2118	0
poly(TOR, 3)1	-1.0039	0.1105	-9.0831	0
poly(TOR, 3)2	0.1175	0.0881	1.3338	0.1825
poly(TOR, 3)3	0.218	0.0867	2.5151	0.012
poly(TORD, 2)1	1.9381	0.0993	19.527	0
poly(TORD, 2)2	0.1801	0.0869	2.0728	0.0384
WAB_AD	0.01	8e-04	12.1492	0
new_RB	0.0064	5e-04	12.1725	0
new_FTR	0.0022	4e-04	6.1888	0
poly(YEAR, 3)1	0.6808	0.091	7.4823	0
poly(YEAR, 3)2	-0.0769	0.0902	-0.8524	0.3941
poly(YEAR, 3)3	-0.359	0.0892	-4.0266	1e-04
Power.RatingMEDIUM	0.0181	0.0065	2.7951	0.0053
Power.RatingSMALL	0.0011	0.0094	0.1143	0.909

Testing Model (adj $R^2 = 0.8065$).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.9885	0.1408	-7.0218	0
X500.LevelYES	0.1007	0.0121	8.3521	0
EFG_D	-0.0219	0.002	-11.0769	0
EFG_O	0.0145	0.0019	7.7365	0
poly(TOR, 3)1	-0.4834	0.1115	-4.3334	0
poly(TOR, 3)2	0.0057	0.0857	0.0662	0.9473
poly(TOR, 3)3	0.16	0.0848	1.8854	0.06
poly(TORD, 2)1	0.9414	0.0999	9.4245	0
poly(TORD, 2)2	0.137	0.0846	1.6198	0.1059
WAB_AD	0.0118	0.0014	8.6839	0
new_RB	0.0082	8e-04	9.7244	0
new_FTR	0.0016	6e-04	2.5987	0.0096
poly(YEAR, 3)1	0.22	0.0882	2.4928	0.013
poly(YEAR, 3)2	0.0876	0.0885	0.9901	0.3226
poly(YEAR, 3)3	-0.288	0.0863	-3.3389	9e-04
Power.RatingMEDIUM	0.0368	0.0106	3.4524	6e-04
Power.RatingSMALL	0.0261	0.0153	1.7037	0.0891

Additional Diagnostic Plots

Leverage Points

There appear to be only two bad leverage points:

Leverage	outliers	
	No	Yes
No	1834	100
Yes	64	2

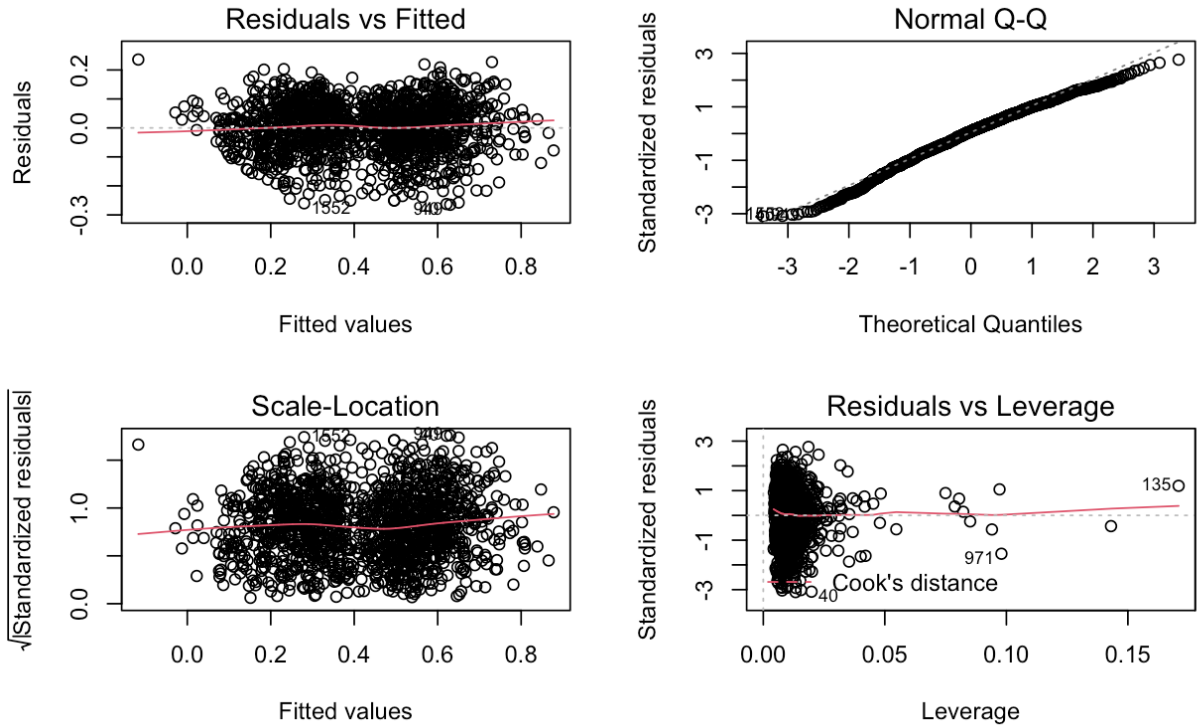
Final Diagnostics

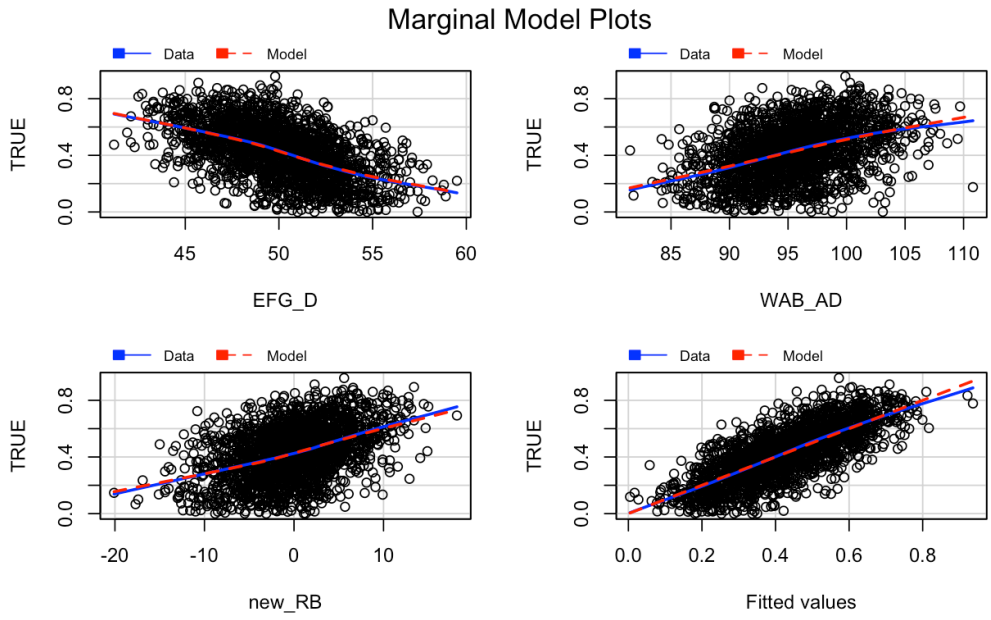
Here is the summary of the final model, the diagnostic plots, marginal density plot, and leverage plots. The diagnostic plots indicate that the assumptions are met, and that the predictors are relatively accurate in predicting the W.P. However, we still consider how the predictor “poly (YEAR, 3)2”, “poly (TOR, 3)2”, and “Power.Rating” are not statistically significant, but we still keep them because other betas from these predictors are statistically significant.

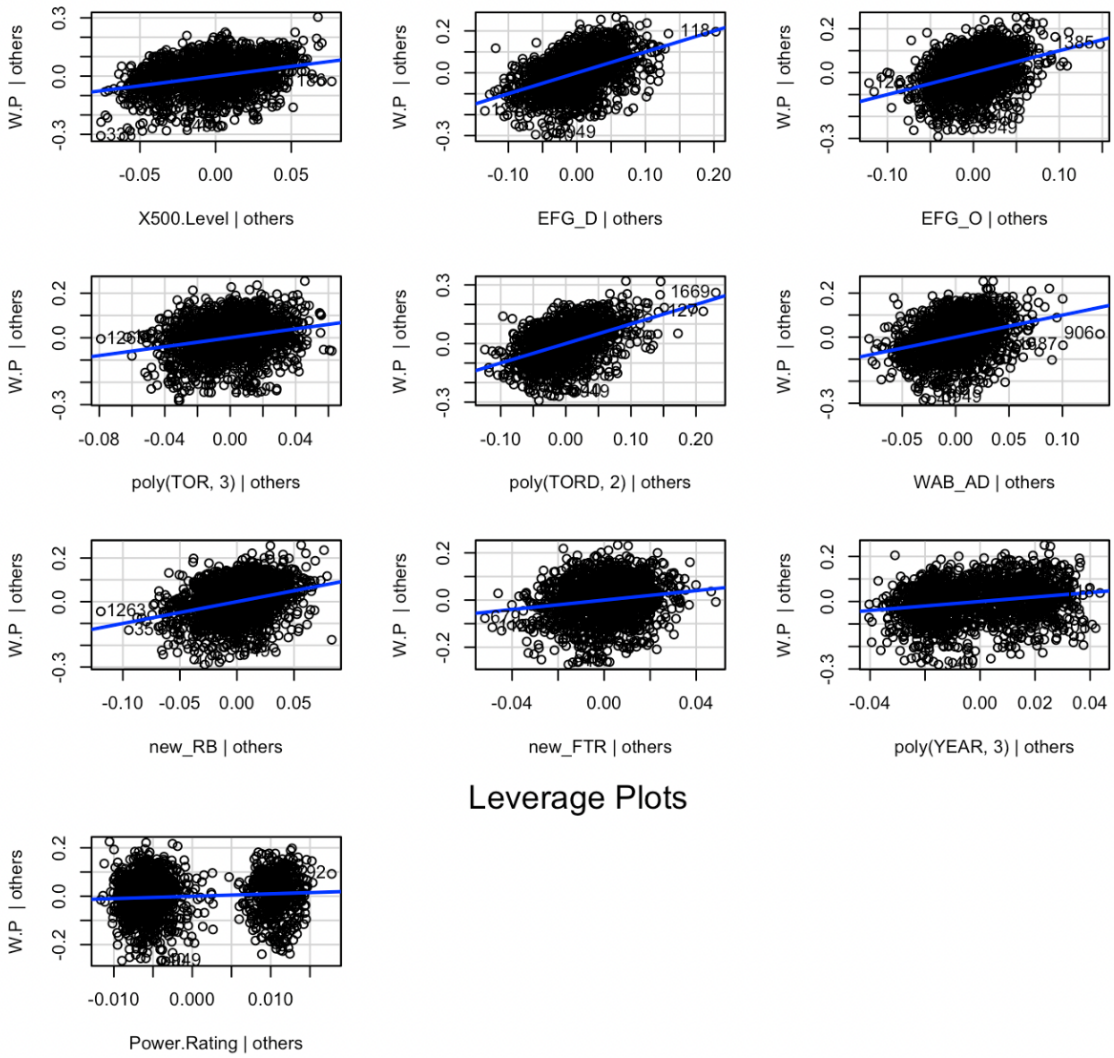
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.7276	0.0687	-10.5944	0
X500.LevelYES	0.0944	0.0061	15.3881	0
EFG_D	-0.0234	0.001	-22.6596	0
EFG_O	0.0147	9e-04	16.2973	0
poly(TOR, 3)1	-1.1203	0.11	-10.1879	0
poly(TOR, 3)2	0.0957	0.087	1.1008	0.2711
poly(TOR, 3)3	0.2481	0.086	2.8851	0.004
poly(TORD, 2)1	2.1447	0.0991	21.6368	0
poly(TORD, 2)2	0.224	0.0861	2.6019	0.0093
WAB_AD	0.0104	7e-04	14.8686	0
new_RB	0.0068	4e-04	15.3841	0
new_FTR	0.0021	3e-04	6.743	0
poly(YEAR, 3)1	0.6999	0.0901	7.7658	0
poly(YEAR, 3)2	-0.0288	0.0895	-0.3215	0.7479
poly(YEAR, 3)3	-0.4487	0.0883	-5.0827	0
Power.RatingMEDIUM	0.023	0.0055	4.1771	0
Power.RatingSMALL	0.0089	0.008	1.1155	0.2648

Final Model

Observations	Residual	R^2	Adjusted R^2
2000	0.08546	0.8021	0.8005







Limitations and Conclusion

According to the diagnostic plots above, we would say that there are slight violations on constant variance and normality, but they are not significant; thus, we would still consider our model valid.

However, one of the drawbacks of this model is its low R square. As we split the dataset into two, we discovered that some variables are not significant anymore, and we assume that

could be the reason that our model has a lower adjusted R square with the given testing data compared to the training data.

In addition, I do admit that the model could be better since other students could obtain higher R square even with fewer predictors than us. After considering our mistakes, I assume that other groups have done a better job of creating new variables than us. In other words, they might discover some patterns among variables and combine them together with more complicated methods than multiplication or subs traction. We also think that we should have looked into our model more closely and maybe delete some bad leverage points too.

Frankly, we did not do as good as other groups in this competition, but we all believe that this is a great opportunity to apply the methods and concepts from class to actual competition, and the competition is also a fruitful lesson for us. Thus, everyone in our group looks forward to doing better in the future with more advanced skills and experience.

References

Almohalwas, Akram. *2021. Chapter 3 scanned notes and Examples updated Jun 27 2020*

Almohalwas, Akram. *2021. STAT 101A Chapter-5 Winter-2020*

Almohalwas, Akram. *2021. STAT 101A Chapter 6 W2020*

———. 2021. STAT 101 a *Winter 2021 Kaggle Competition: Predicting Winning Proportion Using College Basketball Data*

College Basketball Data 2013-2021

Sheather, Simon J. 2009. *A Modern Approach to Regression with R. New York: Springer.*