Predicting Winning Percentages of College Basketball Teams

Info

1. Names and Kaggle nicknames
    a. Ethan Moy, Kaggle: ethmoy
    b. Cara Drake, Kaggle: caradrake
    c. Stacy Deng, Kaggle: stacydeng
    d. Dylan O'Leary, Kaggle: dylanoleary
2. Latest Kaggle rank
    a. 14 as of 3/18 at 3pm
3. Kaggle R2 score
    a. 0.80039
4. The total number of predictors used.
    a. 10
5. The total number of Betas including B0
    b. 11
6. Latest BIC score of the final MLR model. "Use ExtractAIC(model,k=log(n))"
    a. -9837.745

Abstract

The purpose of this project is to generate a multiple linear regression model that best predicts the winning proportion of college basketball teams based on a stats sheet of games. Using regression techniques, we were able to build a multiple linear regression model for the dataset in order to predict the winning proportions for each team. In total, we have 10 predictors in our final model. Our model was developed using the training dataset provided under the "Predicted Winning Proportions" class Kaggle competition, where our R^2 value was approximately 0.80039. In the final Kaggle competition that used the testing data, we placed 14th with an R^2 value of 0.81741, under the team name "Lec 1 Stacy Deng, Ethan Moy, Cara Drake, Dylan O'Leary."

Introduction

College basketball is often confusing and disorganized with so many teams of different skill levels and different matchups with varying skill sets. To address these differences, historical trends are often analyzed to determine game results, such as the fact that a 1-seed in the NCAA tournament wins 78.7% of the time, or the fact that these teams win 99.3% of first-round matchups (Sergent, 2021). There are also several indicators to determine overall winning percentage; for example, Broome notes that adjusted offensive efficiency is "the amount of points a team scores per 100 possessions," and strength of schedule is "total efficiency of the opponents a team has faced" (2019). Using these indicators as our predictors, we had to act as a statistician that wanted to calculate the response variable W.P., which is the winning proportion of a college basketball team in the U.S.

The testing data consists of 2000 observations with 20 predictor variables -- 16 numerical and 4 categorical.

Table 1: Types of Predictors

Variable Name	Variable Description	Variable Type
X500. Level	Whether or not a team has more wins than losses	Categorical
ADJOE	Adjusted offensive efficiency	Numerical
ADJDE	Adjusted defensive efficiency	Numerical
EFG_O	Effective field goal percentage shot	Numerical
EFG_D	Effective field goal percentage allowed	Numerical
TOR	Turnover rate	Numerical
TORD	Steal rate	Numerical
ORB	Offensive rebound rate	Numerical
DRB	Offensive rebound rate allowed	Numerical
FTR	Free throw rate	Numerical
FTRD	Free throw rate allowed	Numerical
X2P_O	Two-point shooting percentage	Numerical
X2P_D	Two-point shooting percentage allowed	Numerical
X3P_O	Three-point shooting percentage	Numerical
X3P_D	Three-point shooting percentage allowed	Numerical
WAB	Wins above bubble	Numerical
YEAR	Seasons 2013-2021	Categorical

NCAA	Whether team made it into the NCAA	Categorical
Power.Rating	Level of chance of beating an average Division 1 team	Categorical
Adjusted.Tempo	Adjusted tempo against an average Division 1 team	Numerical

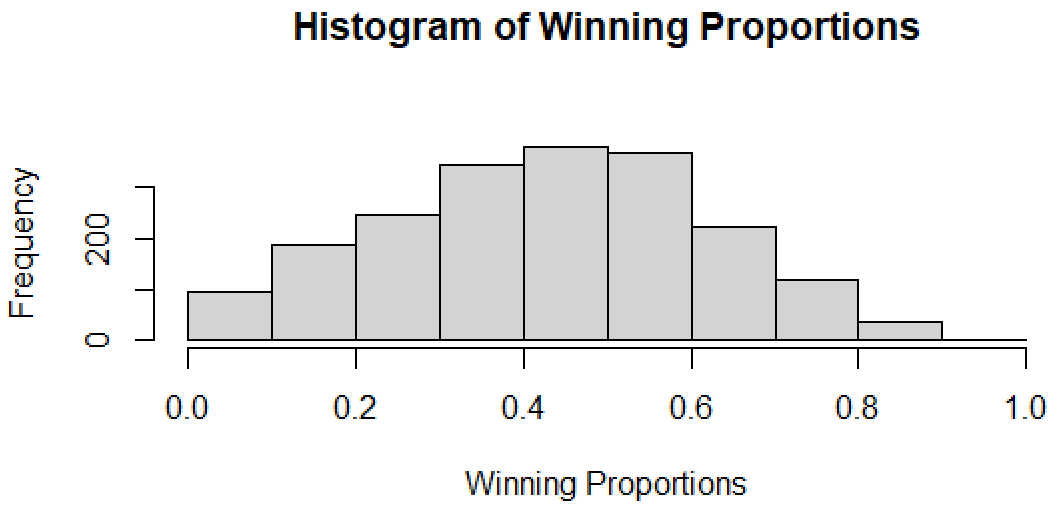
Methodology

The Response Variable

The first thing we did was to plot the response variable, winning proportion (W.P), to determine its distribution. Below are its summary statistics and is a histogram of its distribution.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000233	0.2909135	0.4357239	0.4275516	0.5634299	0.9574060

Table 2: Summary Statistics of Winning Proportion



Based on this graph, W.P appears somewhat symmetrical, so it does not violate the assumption of normality. Some variation from a normal trend is acceptable when dealing with raw data in the real world. We can verify this by implementing a Box-Cox transformation of the response variable.

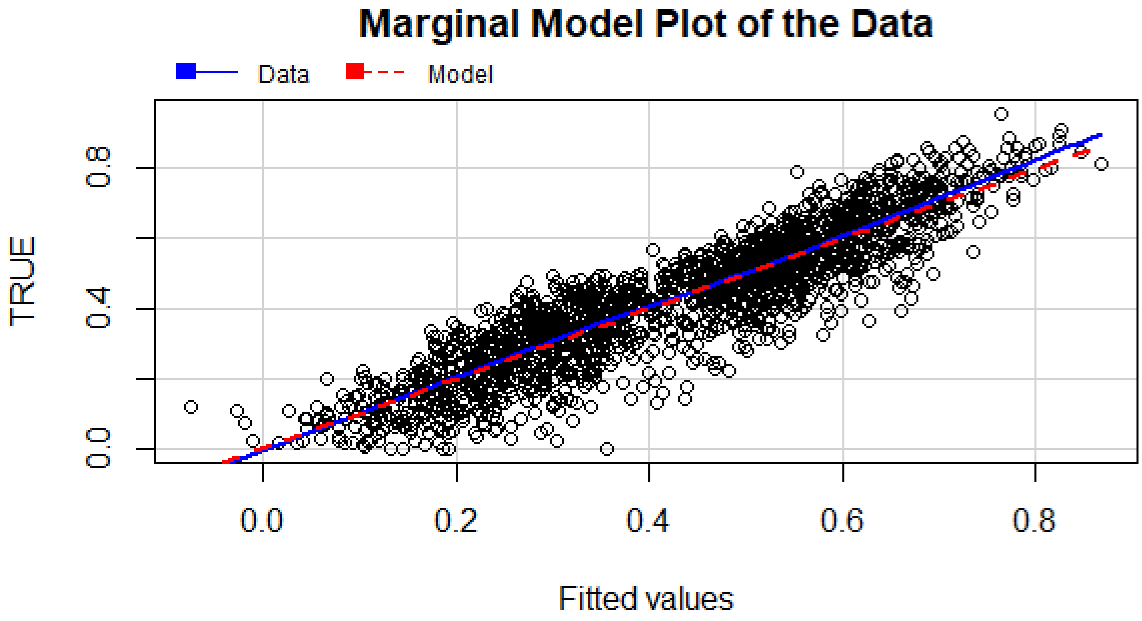
Table 3: Box-Cox Transformation Results

Estimated transformation parameter
Y1
0.8871871

This plot confirms that the best fit for W.P is a lambda of 1, meaning that it does not need to be transformed. Box Cox tends to overfit when run in RStudio, which is why the value of one also fits the lambda.

Degree of Regression

Next, we created a marginal model plot to determine if a linear regression was the best fit for the data.



Because the best-fit curve for the data points appears linear, a linear model appears to be best for the data. The linearity of this model is evident in the overlap of the trend line for the data and the line given by the model.

Building the Model

To begin with, we created a linear model with all 20 predictor variables that were simply added together without any transformation; in other words, W.P. = $\beta_1 x_1 + \beta_2 x_2 + ... + \beta_{20} x_{20}$.

Where each of the betas was a predictor variable from our data set.

We then conducted a backwards stepwise test to eliminate variables that did not contribute to the model, with AIC as the criteria.

Table 4: AIC Results

Start: AIC = -9979.33				
W.P ~ X500.Level + ADJOE + ADJDE + EFG_O + EFG_D + TOR + TORD + ORB + DRB + FTR + FTRD + X2P_O + X2P_D + X3P_O + X3P_D + WAB + YEAR + NCAA + Adjusted.Tempo + Power.Rating				
	Df	Sum of Sq	RSS	AIC
X2P_D	1	0.00009	13.320	-9981.3
YEAR	1	0.00092	13.320	-9981.2
X3P_D	1	0.00114	13.321	-9981.2
		.		
		.		
		.		
Step: AIC = -9981.32				
W.P ~ X500.Level + ADJOE + ADJDE + EFG_O + EFG_D + TOR + TORD + ORB + DRB + FTR + FTRD + X2P_O + X3P_O + X3P_D + WAB + YEAR + NCAA + Adjusted.Tempo + Power.Rating				
	Df	Sum of Sq	RSS	AIC
YEAR	1	0.00099	13.321	-9983.2
NCAA	1	0.00486	13.325	-9982.6
X3P_D	1	0.01248	13.332	-9981.4
		.		
		.		
		.		
Step: AIC = -9983.17				
W.P ~ X500.Level + ADJOE + ADJDE + EFG_O + EFG_D + TOR + TORD + ORB + DRB + FTR + FTRD + X2P_O + X3P_O + X3P_D + WAB + NCAA + Adjusted.Tempo + Power.Rating				
	Df	Sum of Sq	RSS	AIC
NCAA	1	0.00467	13.325	-9984.5

```
<none>                                    13.321        -9983.2

X3P_D           1          0.01425        13.335        -9983.0

                              .
                              .
                              .
```

```
                      Step:  AIC = -9984.47
```

```
W.P ~ X500.Level + ADJOE + ADJDE + EFG_O + EFG_D + TOR + TORD +  ORB + DRB + FTR +
      FTRD + X2P_O + X3P_O + X3P_D + WAB + Adjusted.Tempo + Power.Rating
```

```
                Df         Sum of Sq         RSS           AIC

<none>                                    13.325        -9984.5

X3P_D           1          0.01404        13.339        -9984.4

X3P_O           1          0.01412        13.339        -9984.4

                              .
                              .
                              .
```

```
                              Call:
   lm(formula = W.P ~ X500.Level + ADJOE + ADJDE + EFG_O + EFG_D +
   TOR + TORD + ORB + DRB + FTR + FTRD + X2P_O + X3P_O + X3P_D +
          WAB + Adjusted.Tempo + Power.Rating, data = train)
```

```
Coefficients:
 (Intercept)    X500.LevelYES       ADJOE             ADJDE             EFG_O
  0.216043        0.074893        -0.009398          0.010351          0.033274

   EFG_D            TOR             TORD              ORB               DRB
 -0.015915       -0.015214         0.019043          0.007247         -0.010019

    FTR             FTRD            X2P_O             X3P_O             X3P_D
  0.001417       -0.002199        -0.009754         -0.005586         -0.001609

    WAB        Adjusted.Tempo   Power.RatingMEDIUM  Power.RatingSMALL
  0.018920        0.001728         0.010776          -0.010742
```

Based on the test, we retained seventeen of the variables and removed X2P_D, YEAR, and NCAA. We then conducted a summary of the current model and eliminated variables that were not statistically significant at a 1% level.

Table 5: Summary Results of 17 Predictor Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2160434	0.0935554	2.309	0.021032 *
X500.LevelYES	0.0748925	.0060707	12.337	< 2e-16 ***
ADJOE	-0.0093978	0.0009941	-9.453	< 2e-16 ***
ADJDE	0.0103511	0.0010096	10.252	< 2e-16 ***
EFG_O	0.0332736	0.0072968	4.560	5.43e-06 ***
EFG_D	-0.0159146	0.0016600	-9.587	< 2e-16 ***
TOR	-0.0152142	0.0015405	-9.876	< 2e-16 ***
TORD	0.0190433	0.0013954	13.647	< 2e-16 ***
ORB	0.0072473	0.0007353	9.856	< 2e-16 ***
DRB	-0.0100189	0.0008054	-12.440	< 2e-16 ***
FTR	0.0014167	0.0004043	3.504	0.000469 ***
FTRD	-0.0021988	0.0003777	-5.822	6.79e-09 ***
X2P_O	-0.0097540	0.0046174	-2.112	0.034773 *
X3P_O	-0.0055861	0.0038562	-1.449	0.147610
X3P_D	-0.0016085	0.0011133	-1.445	0.148668
WAB	0.0189204	0.0009222	20.517	< 2e-16 ***
Adjusted.Tempo	0.0017277	0.0006319	2.734	0.006314 **
Power.RatingMEDIUM	0.0107758	0.0068627	1.570	0.116531
Power.RatingSMALL	-0.0107421	0.0104096	-1.032	0.302224

RSE	DF	R^2	Adjusted R^2	p-value
0.08202	1981	0.8179	0.8163	< 2.2e-16

These were X3P_O, X3P_D, and Power.Rating. Note that all categories of Power.Rating were insignificant at both the 1% and 5% levels, so we removed the variable altogether. After removing these predictors, our model was down to fourteen variables.

Table 6: Summary Results of 14 Predictor Model

RSE	DF	R^2	Adjusted R^2	p-value
0.0824	1985	0.8158	0.8145	< 2.2e-16

Candidate Model 1

We ran an ANOVA test and noticed that variables FTR, X2P_O, and Adjusted.Tempo had very low sum of squares values: 0.122, 0.024, and 0.052, respectively. Other variables had sum of squares values ranging from 0.420 to 4.581. However, the p-values were statistically significant at the 5% level, so we ran a partial F-test. The test was also statistically significant at the 5% level, but decided to remove them anyway to reduce any collinearity problems.

Table 7: ANOVA test for Candidate Model 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X500.Level	1	44.280	44.280	6521.4624	< 2.2e-16 ***
ADJOE	1	4.581	4.581	674.7203	< 2.2e-16 ***
ADJDE	1	1.407	1.407	207.2318	< 2.2e-16 ***
EFG_O	1	1.701	1.701	250.5091	< 2.2e-16 ***
EFG_D	1	0.509	0.509	75.0033	< 2.2e-16 ***
TOR	1	0.420	0.420	61.8015	6.179e-15***
TORD	1	1.135	1.135	167.1319	< 2.2e-16 ***
ORB	1	1.196	1.196	176.1743	< 2.2e-16 ***
DRB	1	0.914	0.914	134.6549	< 2.2e-16 ***
FTR	1	0.122	0.122	17.9732	2.343e-05***
FTRD	1	0.560	0.560	82.4755	< 2.2e-16 ***
X2P_O	1	0.024	0.024	3.5812	0.058582 .
WAB	1	2.801	2.801	412.5749	< 2.2e-16 ***
Adjusted.Tempo	1	0.052	0.052	7.6886	0.005609 **
Residuals	1985	13.478	0.007		

Then, we looked at the variance inflation factor (VIF). Variables ADJOE, ADJDE, EFG_O, EFG_D, and WAB all had VIFs of over five. This is a violation to collinearity since the variance inflation factor should not be over five for any value in a model because that is an indicator of overfitting. We then created a correlation matrix to determine what variables were most closely related to each other. This yielded pairs of ADJOE and ADJDE, as well as EFG_O and EFG_D.

Table 8: VIF test for Candidate Model 1

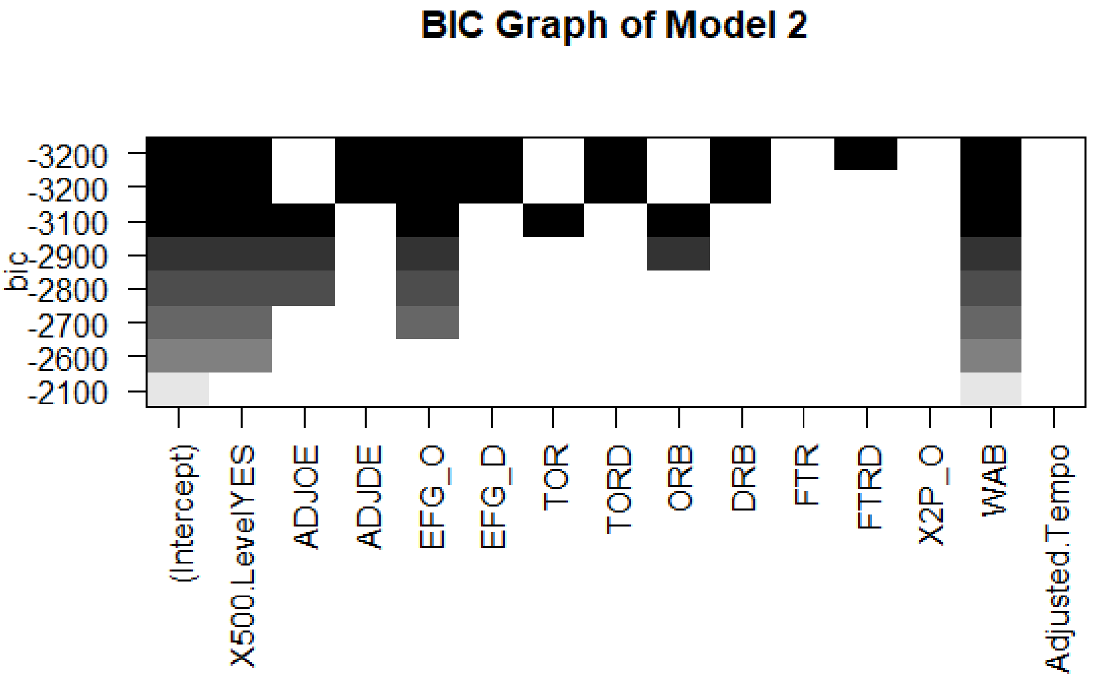
X500.Level	ADJOE	ADJDE	EFG_O	EFG_D	TOR	TORD
2.588679	14.499939	11.067124	10.611523	5.277519	3.013787	2.742649
ORB	DRB	FTR	FTRD	X2P_O	WAB	Adjusted.Tempo
2.744477	2.046735	1.396618	1.675655	5.490531	11.573594	1.156814

To fix this, we created a new numerical variable ADJOE/ADJDE as “efficiency” and a new numerical variable EFG_O/EFG_D as “efg.” We then noticed that the VIF for efficiency and WAB were still high, so we tried to add an interaction term between the two. The interaction term was not statistically significant at both the 1% and 5% levels. Furthermore, ANOVA analysis found that it contributed a low sum of squares to the overall model, and a partial F-test found no statistically significant difference without the interaction term. So, we removed it.

At this point, we weren’t sure how the self-created categories would correlate with the actual test data, so we reverted back to the original predictors ADJOE, ADJDE, EFG_O, and EFG_D. We also felt that the justification behind this model was shaky, especially with regards to how we removed variables at the beginning of this section. Thus, we returned to the model we obtained after conducting the backwards stepwise test; in other words, the one immediately before the start of this section.

Candidate Model 2

After that, we conducted another best subsets selection test using the regsubsets function in R. From the previous test, we noticed that removing more variables had a very small tradeoff in AIC; for example, the AIC would worsen by 0.1 to 2 compared to the current score of around -9984. However, the computer did not remove those variables because it technically leads to a worse AIC. Therefore, we conducted this next test using BIC because our objective was to simplify the model, and BIC is designed to give us a greater penalty for complexity.



Based on this test, we simplified our model to only 8 predictor variables: X500.Level, ADJDE, EFG_O, EFG_D, TORD, DRB, FTRD, and WAB.

Table 9: Summary of 8 Predictor Model

RSE	DF	R^2	Adjusted R^2	p-value
0.08514	1991	0.8028	0.802	< 2.2e-16

Before proceeding further, we checked the diagnostic plots for our current model and found no major violations. However, we wanted to transform the predictors to see if we could “transform two or more variables towards joint normality” (Sheather, 2009). To do this, we used the powerTransform function in R. Note that the code for diagPlot was taken from Chapter 6 notes (Almohalwas, 2020).

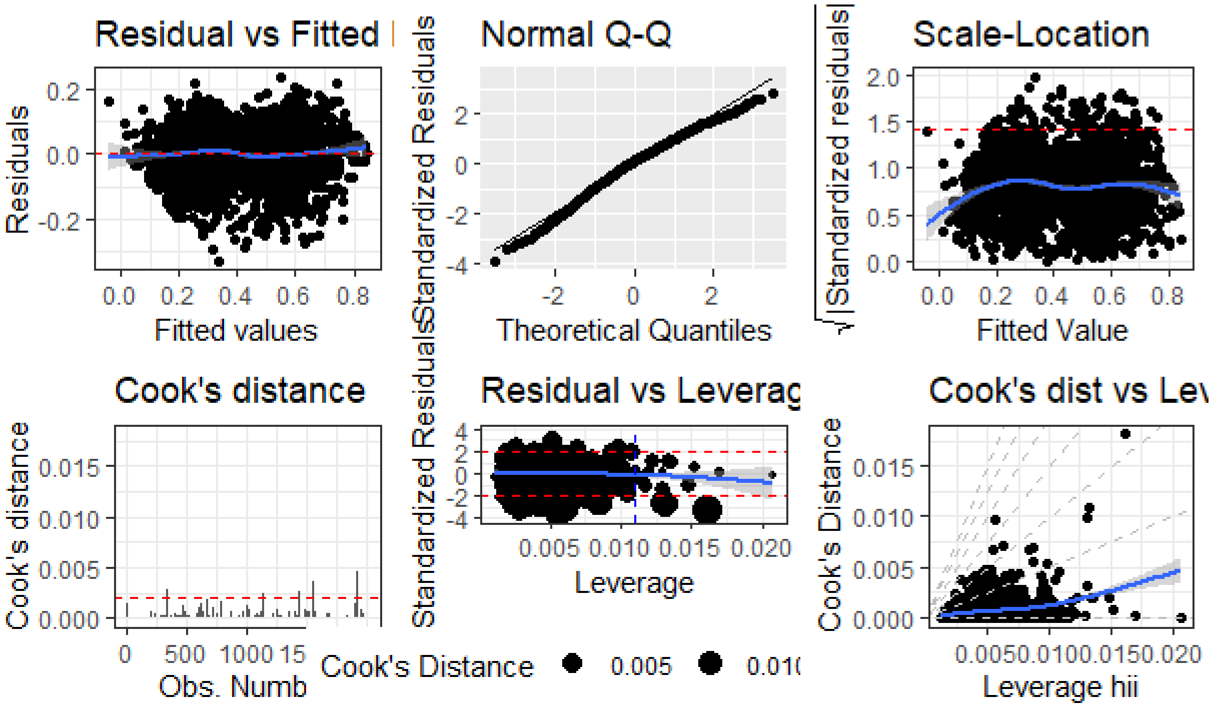


Table 10: bcPower Transformations to Multinormality (Candidate Model 2)

	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
ADJDE	1.2482	1.0	0.8021	1.6943
EFG_O	0.6419	1.0	0.1157	1.1681
EFG_D	0.6372	1.0	0.1310	1.1435
TORD	0.3871	0.5	0.1359	0.6383
DRB	1.0972	1.0	0.7959	1.3985
FTRD	0.1407	0.0	-0.0534	0.3349

Likelihood ratio test that transformation parameters are equal to 0 (all log transformations)

	LRT	df	pval
LR test, lambda = (0 0 0 0 0 0)	96.11857	6	< 2.22e-16

Likelihood ratio test that no transformations are needed

	LRT	df	pval
LR test, lambda = (1 1 1 1 1 1)	103.4895	6	< 2.22e-16

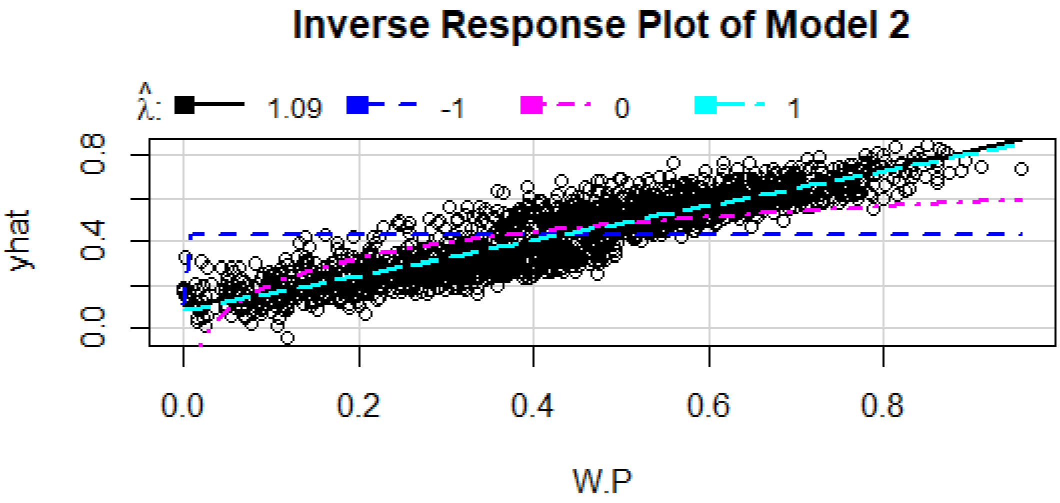
The test indicated that we should raise the following predictors to the following powers: ADJDE to 1.25, EFG_O to 0.5, EFG_D to 0.5, TORD to 0.33 (the cubic root), and take the natural log of FTRD. However, when we did so, we noticed a noticeable drop in our R^2 value of almost 0.04, which was significant considering the range of the Kaggle leaderboard was about 0.05.

Table 11: Summary After Transformation

RSE	DF	R^2	Adjusted R^2	p-value
0.09356	1991	0.7619	0.7609	< 2.2e-16

Therefore, we decided to keep our predictor variables untransformed. The drop in $R^2$was too significant to keep the model despite the indication from power transform.

After this, we wanted to conduct an inverse response plot on the response variable to determine if we could reduce the SSE for our model. We had already determined our response variable was linear but this would be another chance to verify its linearity. Below is our result:



This plot confirms that the best fit for W.P is a lambda of one, meaning that it does not need to be transformed. The given lambda by the program was 1.09, which is extremely close to one, thus giving us confirmation that our response variable should not be transformed.

Next, we checked the VIF of the predictors to see if they were independent and if there were any issues with multicollinearity. We noticed that ADJDE had a VIF of about 8, and WAB had a VIF of about 6.

Table 12: VIF of Candidate Model 2

X500.Level	ADJDE	EFG_O	EFG_D
2.522043	8.339393	2.067200	3.712546
TORD	DRB	FTRD	WAB
2.013757	1.714721	1.418960	6.189991

We then tried to remove WAB as a predictor, but noticed that our R^2 dropped by about 0.07, which was again a violation. Furthermore, a partial F-test on the full model with WAB and a reduced model without WAB showed that we needed to keep WAB in our model, as the p-value was less than 10^{-16}.

Table 13.1: Summary with no WAB

RSE	DF	R^2	Adjusted R^2	p-value
0.09859	1992	0.7354	0.7345	< 2.2e-16

Table 13.2: Analysis of Variance Table with no WAB

Model 1: W.P ~ X500.Level + ADJDE + EFG_O + EFG_D + TORD + DRB + FTRD
Model 2: W.P ~ X500.Level + ADJDE + EFG_O + EFG_D + TORD + DRB + FTRD + WAB

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1992	19.362				
2	1991	14.432	1	4.9301	680.14	< 2.2e-16***

We conducted a similar procedure on ADJDE had similar results: the R^2 dropped about 0.04, and the partial F-test showed that there was a statistically significant difference between a model with and without ADJDE.

Table 14.1: Summary with no ADJDE

RSE	DF	R^2	Adjusted R^2	p-value
0.09199	1992	0.7696	0.7688	< 2.2e-16

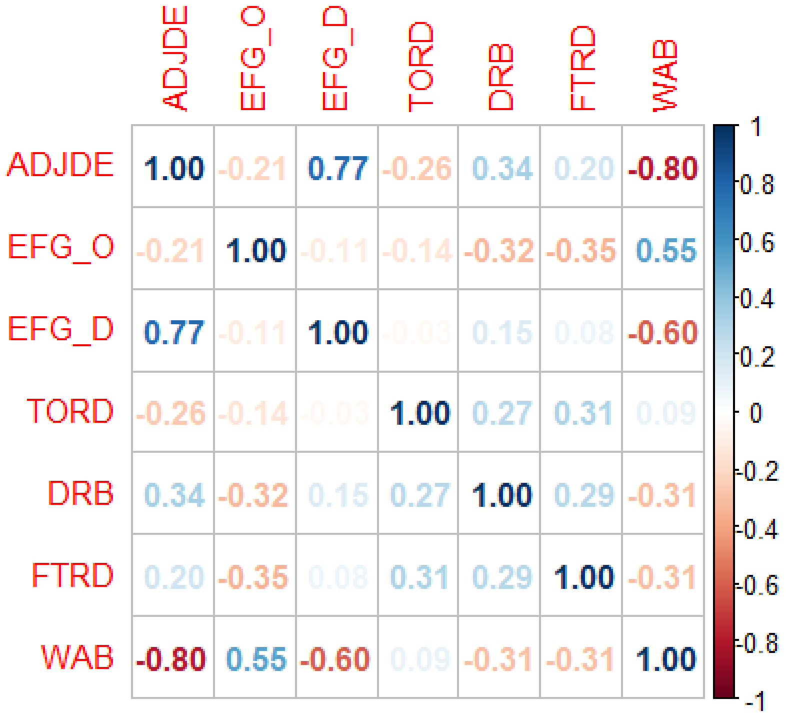
Table 14.2: Analysis of Variance Table with no ADJDE

Model 1: W.P ~ X500.Level + EFG_O + EFG_D + TORD + DRB + FTRD + WAB

Model 2: W.P ~ X500.Level + ADJDE + EFG_O + EFG_D + TORD + DRB + FTRD + WAB						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1992	16.858				
2	1991	14.432	1	2.4262	334.71	< 2.2e-16***

Therefore, we decided to keep both predictors despite their VIF scores. This was a violation of collinearity, but without it our R^2 dropped substantially, thus it made more sense to keep the variables.

Next, we created a correlation matrix to determine which predictors were most related to each other, and we could perhaps add interaction terms for variables with high linearity. Below is our matrix.



	ADJDE	EFG_O	EFG_D	TORD	DRB	FTRD	WAB
ADJDE	1.00	-0.21	0.77	-0.26	0.34	0.20	-0.80
EFG_O	-0.21	1.00	-0.11	-0.14	-0.32	-0.35	0.55
EFG_D	0.77	-0.11	1.00	-0.03	0.15	0.08	-0.60
TORD	-0.26	-0.14	-0.03	1.00	0.27	0.31	0.09
DRB	0.34	-0.32	0.15	0.27	1.00	0.29	-0.31
FTRD	0.20	-0.35	0.08	0.31	0.29	1.00	-0.31
WAB	-0.80	0.55	-0.60	0.09	-0.31	-0.31	1.00

Based on these results, we noticed similar results as the VIF test: ADJDE was closely correlated with most of the remaining predictors, and WAB was also closely correlated to most of the others. Because we wanted a simpler model and didn't want to overfit for the testing data, we decided not to add interaction terms for all the variables that had a correlation higher than about 0.6. Therefore, our current model remained the same.

After that, we tried to merge ADJDE and WAB into one new numerical predictor by adding, subtracting, or multiplying them. However, the difference in R^2 was only about 0.001 for adding these variables, and the others negatively affected it; combined with the fact that ADJDE and WAB are conceptually very different, we decided to leave them as they originally were.

Table 15.1: Summary with ADJDE+WAB

RSE	DF	R^2	Adjusted R^2	p-value
0.08546	1992	0.8012	0.8005	< 2.2e-16

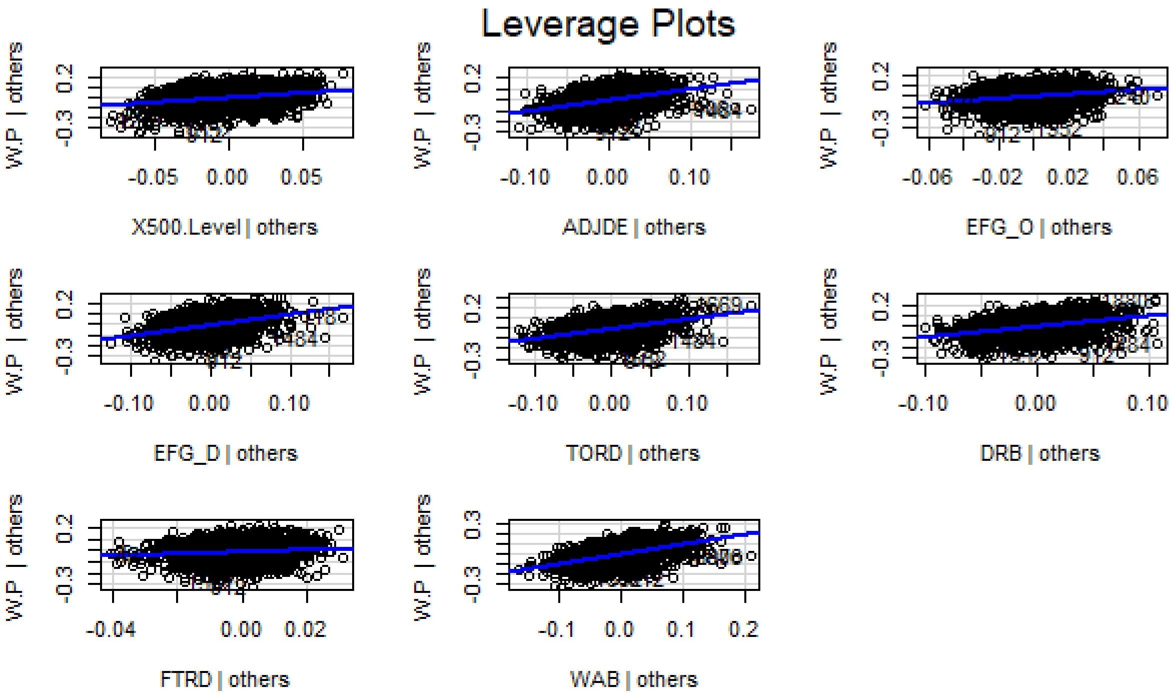
Table 15.2: Summary with ADJDE-WAB

RSE	DF	R^2	Adjusted R^2	p-value
0.09643	1992	0.7469	0.746	< 2.2e-16

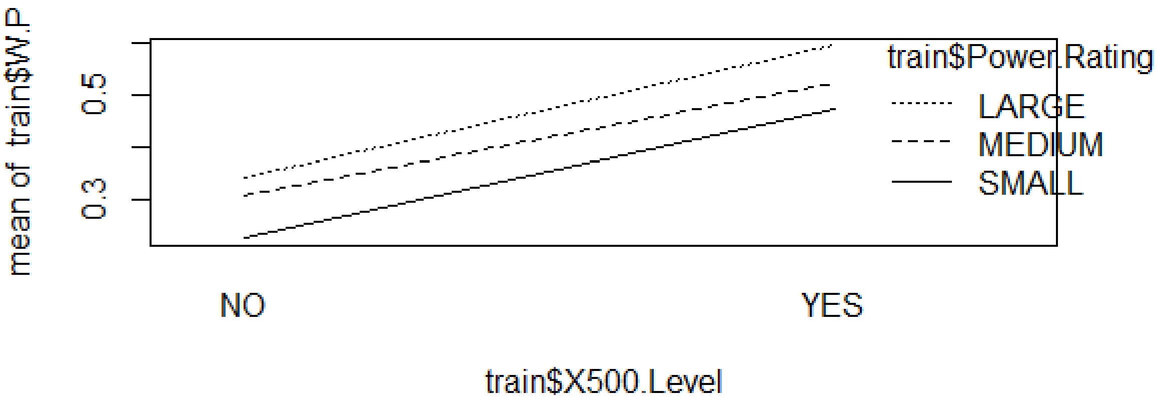
Table 15.3: Summary with ADJDE*WAB

RSE	DF	R^2	Adjusted R^2	p-value
0.09246	1992	0.7673	0.7665	< 2.2e-16

Finally, to confirm our current results, we created leverage plots of every predictor. None of the lines were flat, indicating that all predictors contributed to the model. Below are our leverage plots.

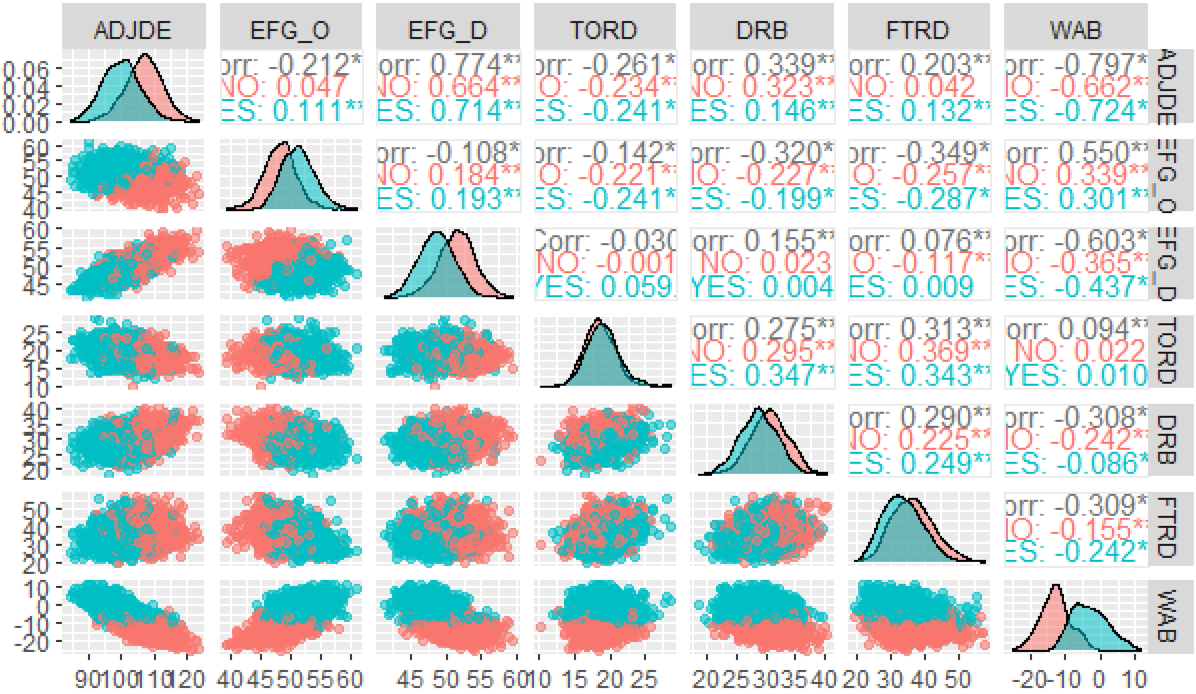


Finally, we checked interactions between numerical and categorical predictor variables. We first wanted to make sure that there were no interactions involving Power.Rating, which was one of the predictors we initially removed. This was because conceptually, Power.Rating was already designed as a predictive term so we expected it to be significant. Thus, we added Power.Rating into our current model and created an interaction plot.



The lack of interceptions in the lines indicate limited to no interaction between Power.Rating and X500.Level, which was the other categorical variable. Parallel lines would indicate no interaction at all, and these lines are close to parallel since they do not intersect each other at all and all have a similar positive trend. This therefore confirms that removing Power.Rating did not change the model, so we removed it and proceeded with the previous model.

We then wanted to explore possible interactions between the only remaining categorical predictor in our model, X500.Level, and the other numerical predictors. To do this, we created a ggpairs plot, as shown below.



This indicated that almost every numerical predictor had an interaction term with X500.Level. We thus added an interaction term between X500.Level and ADJDE, EFG_O, EFG_D, TORD, DRB, FTRD, and WAB.

Because we also wanted our model to not be unnecessarily complex, we conducted a summary and removed interaction terms with p-values statistically insignificant at the 1% level.

Therefore, we removed interaction terms between WAB, EFG_O, and FTRD. Then we did another summary with those remaining four values and removed the two which were now insignificant at the 1% level. This led us to the summary model 16.3 which gave us a final R^2 higher than the simple eight predictors model while avoiding too many betas.

Table 16.1: Summary with 7 Interaction Terms

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.2855888	0.0921510	-3.099	0.001968 **
X500.LevelYES	0.0541274	0.1256224	0.431	0.666607
ADJDE	0.0121681	0.0011506	10.576	< 2e-16 ***
EFG_O	0.0094952	0.0012638	7.513	8.69e-14 ***
EFG_D	-0.0197396	0.0017568	-11.236	< 2e-16 ***
TORD	0.0204531	0.0018463	11.078	< 2e-16 ***
DRB	-0.0092250	0.0011356	-8.123	7.88e-16 ***
FTRD	-0.0014193	0.0005127	-2.768	0.005684 **
WAB	0.0179264	0.0010219	17.542	< 2e-16 ***
X500.LevelYES:ADJDE	0.0068303	0.0016969	4.025	5.91e-05 ***
X500.LevelYES:EFG_O	-0.0030020	0.0017600	-1.706	0.088225 .
X500.LevelYES:EFG_D	-0.0084148	0.0025313	-3.324	0.000902 ***
X500.LevelYES:TORD	0.0084824	0.0024752	3.427	0.000623 ***
X500.LevelYES:DRB	-0.0069412	0.0015260	-4.549	5.73e-06 ***
X500.LevelYES:FTRD	-0.0012526	0.0007172	-1.747	0.080868 .
X500.LevelYES:WAB	0.0012115	0.0013992	0.866	0.386676

RSE	DF	R^2	Adjusted R^2	p-value
0.08463	1984	0.8058	0.8044	< 2.2e-16

Table 16.2: Summary with 4 Interaction Terms

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.2337781	0.0810080	-2.886	0.00395 **
X500.LevelYES	-0.0301876	0.0933820	-0.323	0.74653

ADJDE	0.0127906	0.0010156	12.594	< 2e-16 ***
EFG_O	0.0079915	0.0008793	9.089	< 2e-16 ***
EFG_D	-0.0201316	0.0016960	-11.870	< 2e-16 ***
TORD	0.0212104	0.0017411	12.182	< 2e-16 ***
DRB	-0.0095232	0.0011227	-8.482	< 2e-16 ***
FTRD	-0.0020435	0.0003586	-5.699	1.38e-08 ***
WAB	0.0185109	0.0006976	26.536	< 2e-16 ***
X500.LevelYES:ADJDE	0.0053482	0.0012013	4.452	8.99e-06 ***
X500.LevelYES:EFG_D	-0.0075093	0.0023454	-3.202	0.00139 **
X500.LevelYES:TORD	0.0069316	0.0021672	3.198	0.00140 **
X500.LevelYES:DRB	-0.0063512	0.0014714	-4.316	1.66e-05 ***

RSE	DF	R^2	Adjusted R^2	p-value
0.08467	1987	0.8053	0.8042	< 2.2e-16

Table 16.3: Summary with 2 Interaction Terms (Final Model)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.485e-01	7.470e-02	-3.327	0.000894 ***
X500.LevelYES	-5.439e-05	7.661e-02	-0.001	0.999434
ADJDE	1.441e-02	9.197e-04	15.668	< 2e-16 ***
DRB	-1.112e-02	1.042e-03	-10.673	< 2e-16 ***
TORD	2.494e-02	1.225e-03	20.367	< 2e-16 ***
EFG_O	8.099e-03	8.816e-04	9.186	< 2e-16 ***
EFG_D	-2.374e-02	1.267e-03	-18.743	< 2e-16 ***
FTRD	-2.073e-03	3.595e-04	-5.768	9.31e-09 ***
WAB	1.823e-02	6.958e-04	26.197	< 2e-16 ***

X500.LevelYES:ADJDE	1.861e-03	7.344e-04	2.534	0.011368 *
X500.LevelYES:DRB	-3.481e-03	1.259e-03	-2.766	0.005735 **

RSE	DF	R^2	Adjusted R^2	p-value
0.08494	1989	0.8039	0.8029	< 2.2e-16

Comparing Candidate Model 1 and 2

Just to be thorough in our analysis, we compared the BIC of this model (model 2) against our partially finished model from the previous section (model 1). Model 1 yielded a BIC of -9935 with 12 betas, whereas model 2 yielded a better BIC score with fewer betas. Therefore, we proceeded with model 2 as our final model.

Results

Result Model

As can be seen above, our final model was selected as

$$\begin{gathered}
\textit{Win\_Percentage} = -.2485 - 5.439*10^{-5}\ \textit{X500.LevelYes}\ (1.861*10^{-3} \\
\textit{Adjusted\_Defense\_Efficiency} - 3.481*10^{-3}\ \textit{Offensive\_Rebound\_Rate\_Allowed}) + 1.441*10^{-2} \\
\textit{Adjusted\_Defense\_Efficiency} - 1.112*10^{-2}\ \textit{Offensive\_Rebound\_Rate\_Allowed} + 2.454*10^{-2} \\
\textit{Turnover\_Percentage\_Committed} + 8.099*10^{-3}\ \textit{Effective\_Field\_Goal\_Percentage} - 2.374*10^{-2} \\
\textit{Effective\_Field\_Goal\_Percentage\_Allowed} - 2.07*10^{-3}\ \textit{Free\_Throw\_Rate\_Allowed} + 1.823* \\
10^{-2}\ \textit{Wins\_Above\_Bubble}
\end{gathered}$$

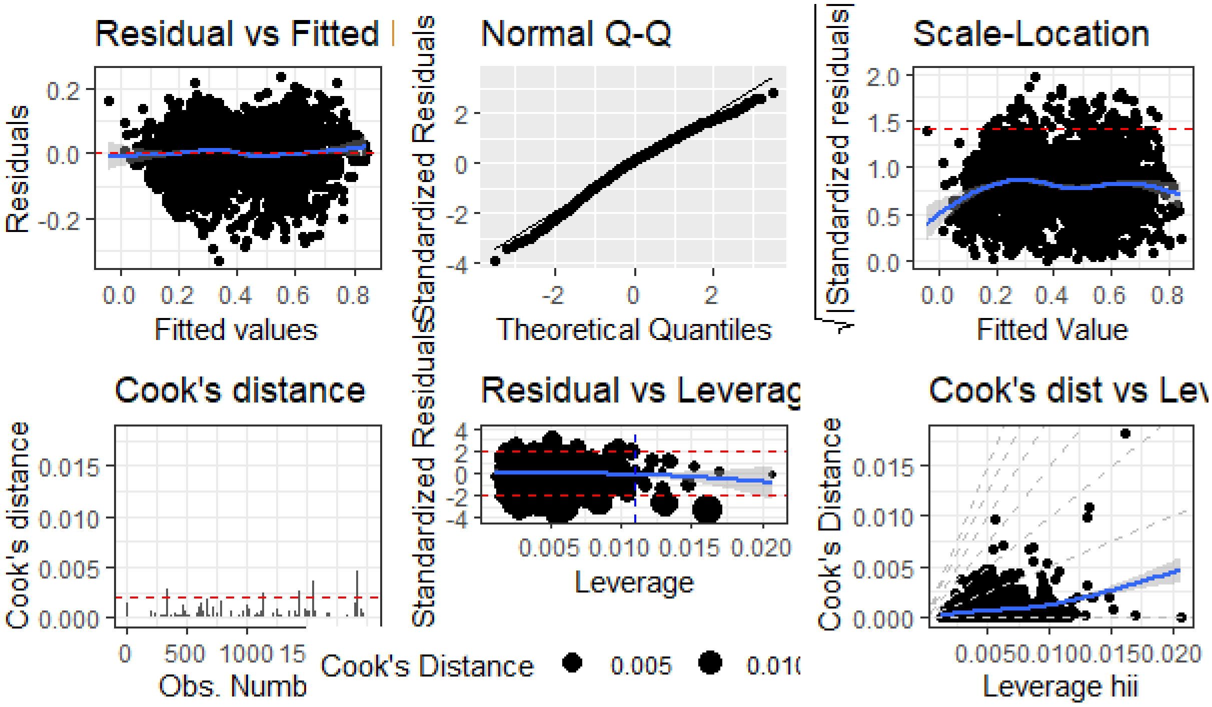
This model was chosen for the same reason explained above over our other model, to stay within the parameters of the challenge. These predictors were chosen based on the AIC, BIC, and p-value test which were run to help understand the values which affected the win percentage the most.

Results Diagnostics

For the model that we created, one of the major factors in consideration was the diagnostic plots created from it. A high R^2 on its own does not mean a good model if the diagnostics are bad. Avoiding as many violations as possible was a major concern while attempting to achieve a high R^2 value. Thus, there were many different diagnostic tests that were conducted on this model.

Six way diagnostic plot

When analyzing and determining this model, the primary concern was determining a valid model. To verify our model did not have any major violations, we checked our diagnostic plots. These graphs would verify the linearity of the model, the linear distribution, the spread of the residuals, and checks for influence points.



As can be seen in our diagnostics, although there are still some minor violations, the trend of the data has a minimal amount of violations. The first graph shows the residuals versus the fitted values, the trend of the red line overlaps largely with the blue. The second graph shows the distribution of standard residuals and should be a linear trend. There are some violations past two and under negative two, however the overall trend is fairly linear. This third graph checks for equal variance, which is why a flat line is good. Once again, there is a bit of drop off at the start, but overall the line is relatively flat, so the trend overall is decent. The fourth graph shows any points which exceed Cook's Distance, and thus are likely leverage points. There are very few in this model, and simply removing leverage points is not a reliable tactic, so having only a few violations in a model is acceptable with so many values. The fifth graph graphs the leverage points versus the square root of the standardized residual in order to locate bad leverage points. Here we can see only a few values are outliers, which means there are not many bad leverage points in this model, especially since only a few of those are past the mean value for h_{ii}. For the final model, the sixth, we can see that the leverage value is graphed against the Cook's Distance, and overall it has a flat relationship without an overwhelming amount of outliers.

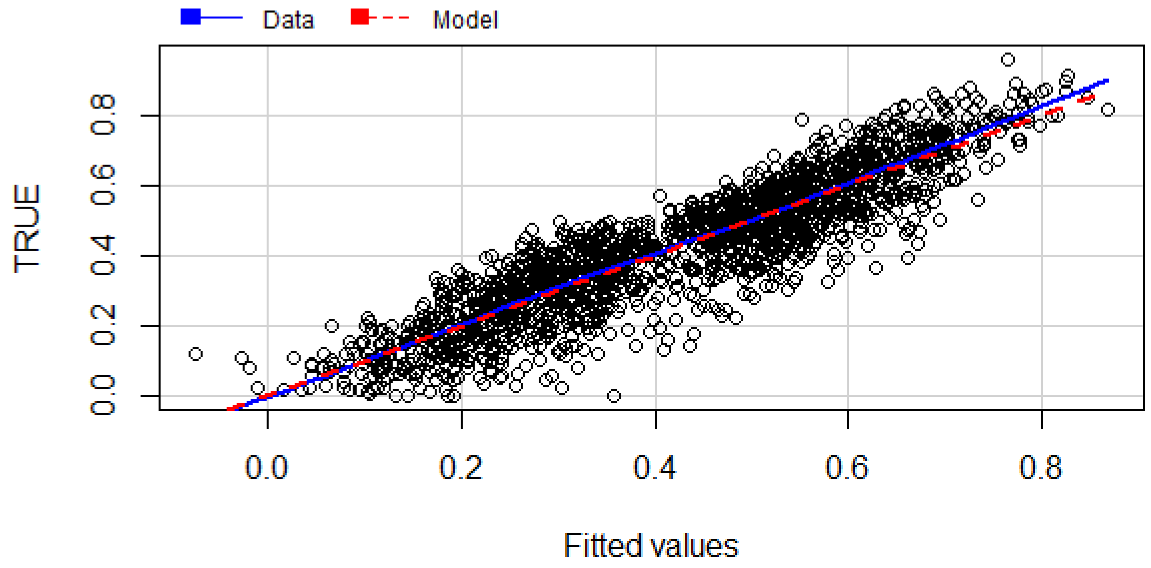
Leverage Points

Leverage / Outlier	No	Yes
No	1109	42
Yes	802	47

For the amount of data points collected, the number of leverage points that are concerning is not that many.

Mmp plots

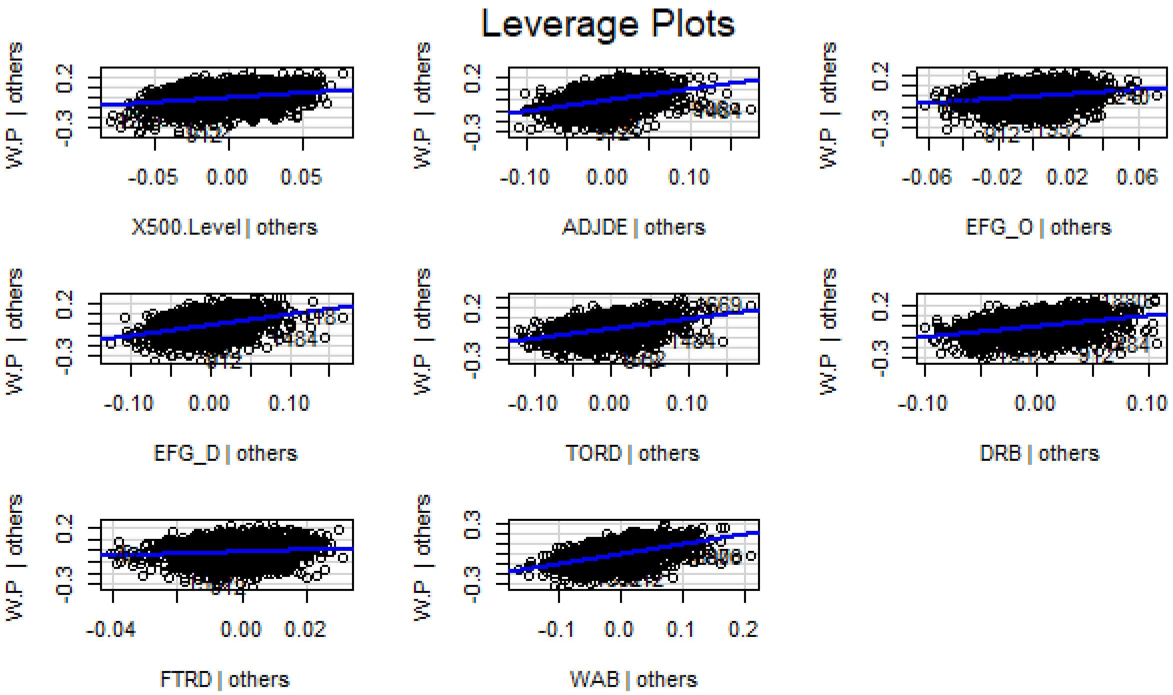
A further investigation into checking linearity can be seen by Marginal Model Plotting. Marginal Model Plots plot the terms against the fitted values. (Fox, 2021) The blue line is the trend of the data plots and the red line is our corresponding model. A strong overlap indicates the model is linear. If the data line is shaped differently than the model, then the model should be reworked or the data needs to be manipulated in some way.



As can be seen in the plot, for the most part, the blue and red lines are virtually inestinquesable, and thus the model fits the data's trend well. The starting value of the data's trend is slightly lower than our model's, and the end is slightly higher, once again showing some minor violation, but overall they are very strongly overlapping.

Leverage plots

Leverage plots are broken down one predictor at a time, and the outcome shows whether or not the predictor should be included in the model. A flat line indicates that that variable is not a good predictor for the model and those should be removed accordingly.



The trend of all these models has a slope higher than zero. Some more so than others, for instance, ADJDE has a much higher slope than FTRD. FTRD has a very small slope, it is not flat and thus does still warrant remaining in the model because they are significant.

The latest BIC of this model is -9837.745, which is a significantly low value. One of the major violations is the VIF, Variance inflation factor. The max vif should be five, but our model violates this. To solve this we attempted to remove variables which had high correlation with each other to reduce collinearity, however in doing so that caused the R^2 to drop significantly so we decided to maintain the variables despite the violations.

Thus, even though the model was not a perfect model, its violations are the minimal we could do so that the model would still maintain a high R^2.

Discussion

Limitations and Conclusions

As for the diagnostic plots shown above, there is a slight violation in normality, as the values towards the latter part of the data are not exactly on the normal line. Also, variance is not completely constant, as there is a slight increase after the first few data points. Cook's distance as well as the residuals vs. leverage plot show that there are some potential bad leverage points as well, which would in turn be violations for the model.

As for the model as a whole, we took into account the value of R^2 in comparison to the complexity of our model. We decided that a simpler model would be more beneficial at the cost of a few decimal points in our R^2 value. Of course, these decimal values play a role, but we concluded that anything under .001 is insignificant. We also faced some issues when it came to transforming certain numerical predictors into categorical variables. We stayed away from doing this mainly due to a fear of skewing the data if our calculations didn't fit the context of the problem at hand. For example, creating a cutoff value for Adj. Offensive Efficiency was unclear to us, so we stayed with the original numerical predictor. Attempts made to create new variables were largely unsuccessful because they either did not make sense at a higher level or didn't have an impact on our R^2.

Another issue was that we did not want to overfit, so we weren't certain as to how our model holds up to the real testing data. Transforming between both data sets created some issues with the R^2 and the overall fit of the model. The transformations failed to have any large correlation with R^2 and thus we were advised against doing impactful changes to our data when communicating with the teaching assistant for this course. Further analysis could involve conducting a logistic regression and further categorizing predictor variables into whether they occurred when a team was home or away, as was modeled by Lopez and Matthews (2015). One variable we specifically wanted to keep but could not find correlation was Power.Rating, thus perhaps through more logistic regression or reformatting tests that were not discussed in this course, we could find a better analysis of the data.

References

Almohalwas, Akram. (2020). *Chapter 5 Updated Winter 2020.*

Almohalwas, Akram. (2020). *Chapter 6 Updated Winter 2020.*

Broome, Anthony. (2019). *KenPom 101: What the college basketball metric system is and how it ranks Michigan*. SB Nation. Retrieved March 17, 2022, from https://www.maizenbrew.com/2019/10/23/20928669/kenpom-explained-what-it-means-michigan-basketball-ranking

John Fox [aut, cre]. (2021, November 6). Marginalmodelplot: Marginal model plotting in car: Companion to Applied Regression. marginalModelPlot: Marginal Model Plotting in car: Companion to Applied Regression. Retrieved March 19, 2022, from https://rdrr.io/cran/car/man/marginalModelPlot.html

Lopez, M. & Matthews, G. (2015). Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports, 11(1), 5-12.* https://doi.org/10.1515/jqas-2014-0058

Sergent, Jim. (2021). *NCAA Tournament bracket, by the numbers: Historical trends from March Madness*. USA Today. Retrieved March 18, 2022, from https://www.usatoday.com/in-depth/sports/ncaab/2021/03/16/march-madness-numbers-ncaa-tournaments-historical-trends/4647793001/

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