1. Title

Analysis of the Winning Proportion in NCAA Basketball Tournament with Multiple Linear Regression

Project Summary	
Names	Ken Fukuyama Regina Thahir Tori Wang Rawi Baransy
Emails	kenfukuyama@ucla.edu reginathahir@ucla.edu toriluwang2020@gmail.com rawi.baransy@yahoo.com
Kaggle Nicknames	Regression Team Lec2
Kaggle Rank	32nd
Kaggle R Score	0.80935
Total predictors used	9 predictors
Total number of Betas including B0	10 β's
Latest BIC Score	-9813.395
Complexity Grade	100 (110 - 10)

2. Abstract

The purpose of this project was to build a multiple linear regression model that predicts the Winning Proportion of basketball teams in the National Collegiate Athletic Association or NCAA. In the annual NCAA tournament, predictions of the winning team and team ranks are often created by fans and the results are highly anticipated. Our job was to see which team statistics are useful in predicting the winning proportion and create a valid model using these predictors.

Using the training data, we applied regression techniques to build a multiple linear regression with selected variables to predict the winning proportion of each basketball team. The final model was constructed with 9 predictors and 10 betas. In the development stage, we tuned our model using the training dataset and achieved 0.8054 R-squared (the coefficient of determination). Ultimately, our team, Regression Team Lec2, submitted our model to the Kaggle competition for evaluation. With the testing dataset, we obtained an R-squared value of 0.80935 and ranked 32nd.

3. Introduction

The purpose of this project was to use the statistics of basketball teams in the NCAA in order to predict the winning proportions of each team. College basketball is governed by collegiate associations which include the NCAA. During conference play, the teams are ranked through the entire NCAA as well as in the tournament plays leading up to the NCAA tournament. We were provided with a dataset gathered from the 2013-2021 Division I college basketball seasons in the U.S. The data files were divided up into a training dataset and a testing dataset. The training dataset contains 20 predictors (not including the response variable W.P) and 2000 observations of different teams (Kaggle, 2022). The final regression model was then submitted to a class Kaggle competition, in which it was used to predict the winning proportions of the testing dataset that contains 1155 observations. Below are the set of initial variables and their class provided in the dataset:

Table 1: Types of variables

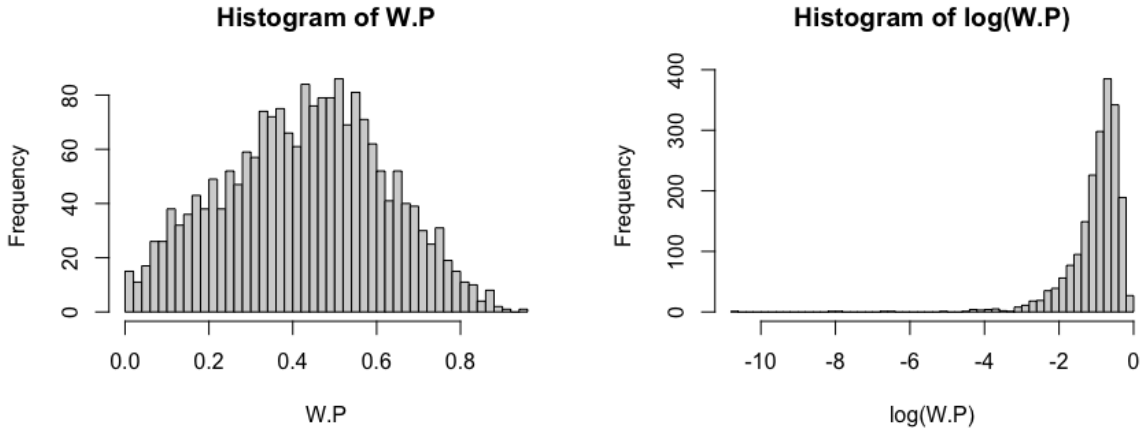
X500.Level	Categorical
ADJOE	Numerical
ADJDE	Numerical
EFG_O	Numerical
EFG_D	Numerical
TOR	Numerical
TORD	Numerical
ORB	Numerical

DRB	Numerical
FTR	Numerical
FTRD	Numerical
X2P_O	Numerical
X2P_D	Numerical
X3P_O	Numerical
X3P_D	Numerical
WAB	Numerical
YEAR	Categorical
NCAA	Categorical
Power.Rating	Categorical
Adjusted.Tempo	Numerical
W.P	Numerical

4. Methodology

The Response Variable: Winning Proportion (W.P)

The histogram of W.P shows that there is no need to do Box-Cox transformation on the response variable since it is already normally distributed. For example, the histogram of log(W.P) ruins the normality of the response variable and makes Winning Proportion left-skewed.

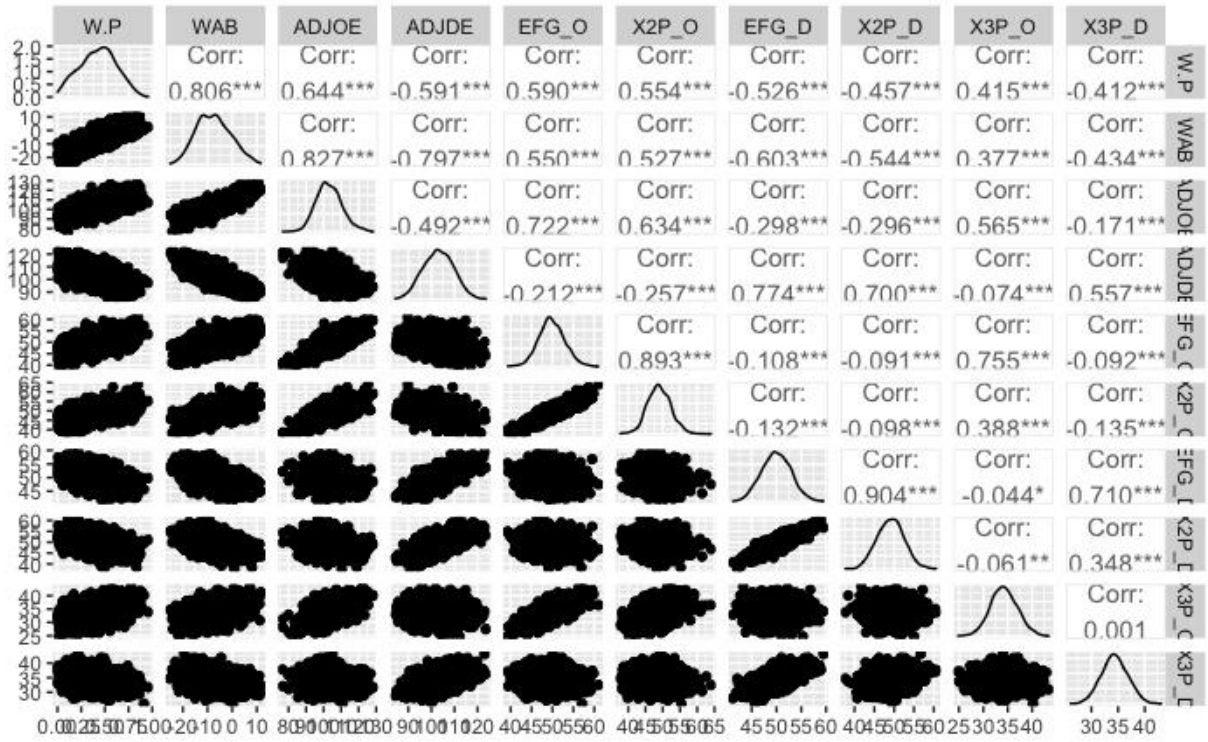


The Numerical Predictors

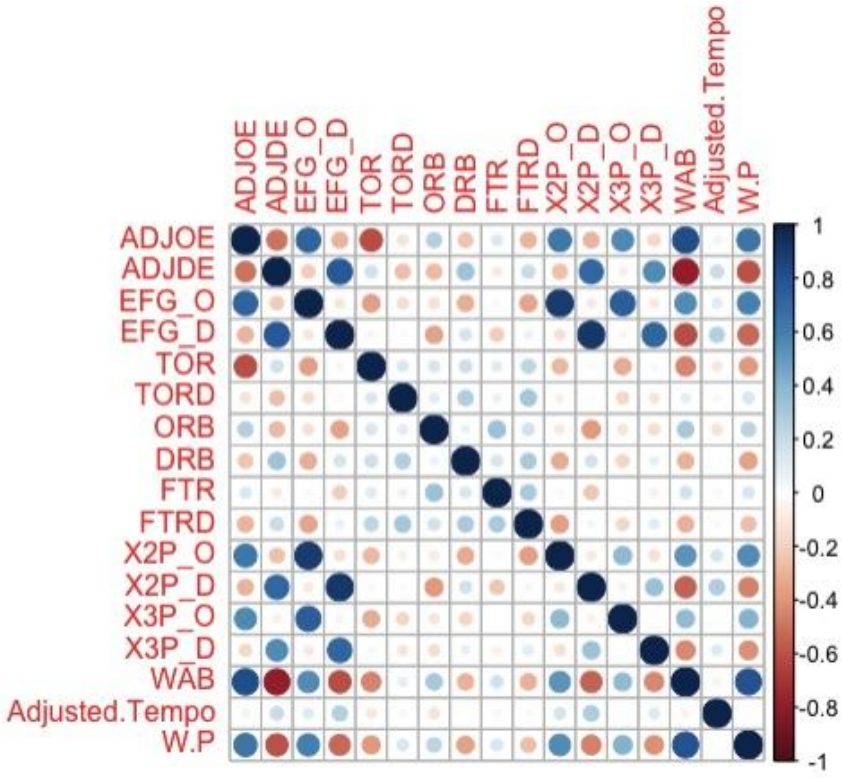
Table 2: Correlation with W.P

Numerical Predictor	Correlation with W.P
ADJOE	0.644
ADJDE	-0.591
EFG_O	0.590
EFG_D	-0.526
TOR	0.387
TORD	0.146
ORB	0.245
DRB	-0.355
FTR	0.148
FTRD	-0.266
X2P_O	0.554
X2P_D	-0.457
X3P_O	0.415
X3P_D	-0.412
WAB	0.806
Adjusted.Tempo	-0.007

Below is a scatterplot matrix between the top 10 numerical variables with the highest correlation with W.P



Below is a visualization of the correlations between all numerical predictors



Based on the correlation matrix and the visualization above, we notice that WAB has a high correlation with W.P. The next highest correlated predictors with W.P are ADJOE, ADJDE, EFG_O, X2P_O and EFG_D. ADJOE and ADJDE have a high correlation with WAB, thus creating a concern for multicollinearity. These numerical predictors all show normal patterns in the histogram with no improvement through the use of Box-Cox transformation, such as using logarithm on the response variable.

Below is a first attempt approach at using all numerical variables:

Table 3: Summary of using all numerical variables

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.2866399	0.0983722	2.914	0.00361	**
ADJOE	-0.0112760	0.0009898	-11.393	< 2e-16	***
ADJDE	0.0119161	0.0009747	12.225	< 2e-16	***
EFG_O	0.0412886	0.0076122	5.424	6.54e-08	***

EFG_D	-0.0155923	0.0101045	-1.543	0.12296
TOR	-0.0183821	0.0015997	-11.491	< 2e-16 ***
TORD	0.0222409	0.0014450	15.392	< 2e-16 ***
ORB	0.0085420	0.0007630	11.196	< 2e-16 ***
DRB	-0.0118508	0.0008366	-14.166	< 2e-16 ***
FTR	0.0018835	0.0004222	4.461	8.63e-06 ***
FTRD	-0.0025830	0.0003955	-6.531	8.27e-11 ***
X2P_O	-0.0122354	0.0048306	-2.533	0.01139 *
X2P_D	-0.0025025	0.0064750	-0.386	0.69918
X3P_O	-0.0072074	0.0040364	-1.786	0.07432 .
X3P_D	-0.0041452	0.0053626	-0.773	0.43963
WAB	0.0227623	0.0008943	25.452	< 2e-16 ***
Adjusted.Tempo	0.0019723	0.0006612	2.983	0.00289 **

Table 4: Summary statistics for all numerical variables

$Observations$	$Residual\ Std.\ Error$	R^2	$R^2\ adjusted$
2000	0.08593	0.7999	0.7983

From the full model, the p-values of predictors indicate that X2P_D, X3P_O, and X3P_O should be removed as they have a p-value larger than 0.05 indicating overfitting.

Removing these predictors, we form a reduced model as shown below.

Table 5: Summary of reduced numerical model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2561454	0.0974217	2.629	0.00862 **
ADJOE	-0.0111200	0.0009834	-11.307	< 2e-16 ***
ADJDE	0.0121535	0.0009701	12.528	< 2e-16 ***
EFG_O	0.0279423	0.0019861	14.069	< 2e-16 ***
EFG_D	-0.0210526	0.0014894	-14.135	< 2e-16 ***
TOR	-0.0180895	0.0015936	-11.351	< 2e-16 ***
TORD	0.0227485	0.0014206	16.013	< 2e-16 ***
ORB	0.0082309	0.0007542	10.914	< 2e-16 ***
DRB	-0.0119254	0.0008277	-14.407	< 2e-16 ***

FTR	0.0017412	0.0004187	4.159	3.34e-05 ***
FTRD	-0.0026497	0.0003931	-6.741	2.06e-11 ***
X2P_O	-0.0038184	0.0013239	-2.884	0.00397 **
WAB	0.0228221	0.0008907	25.623	< 2e-16 ***
Adjusted.Tempo	0.0020840	0.0006592	3.161	0.00159 **

Table 6: Model with Numerical Predictors 2

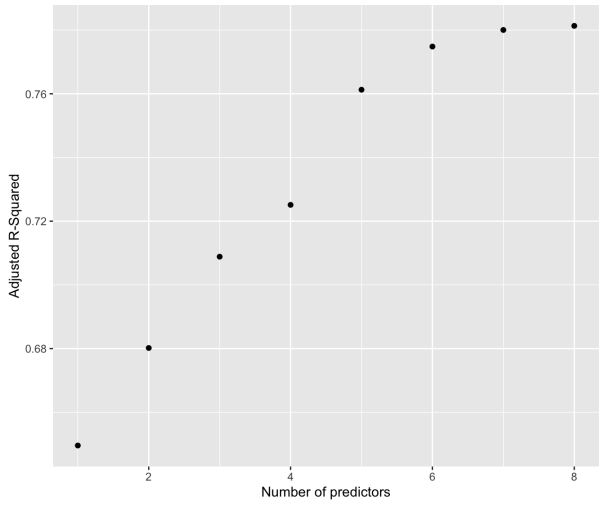
$Observations$	$Residual\ Std.\ Erro$	R^2	$R^2\ adjusted$
2000	0.08601	0.7992	0.7979

A variance inflation factor was then run on the reduced linear regression, which still indicates multicollinearity. The VIF is shown below:

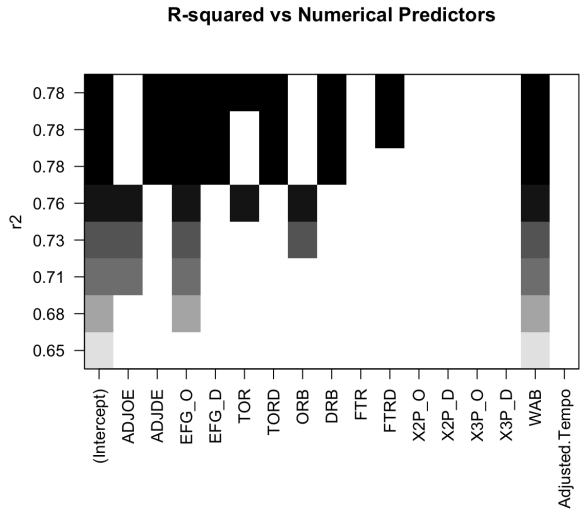
Table 7: Model with Numerical Predictors 2

Numerical Predictor	VIF
ADJOE	14.177638
ADJDE	10.740417
EFG_O	10.316132
EFG_D	5.033485
TOR	2.966624
TORD	2.645083
ORB	2.709254
DRB	1.996184
FTR	1.392320
FTRD	1.661289
X2P_O	5.485971
WAB	10.076451
Adjusted.Tempo	1.155035

In order to see which of these numerical predictors are most useful in building our model, we run regsubsets and plot the number of numerical predictors that will maximize our adjusted R-squared. From the plot, we can see that the adjusted R-squared for the model is maximized at 6 numerical predictors, adding more predictors would not improve the model significantly.



We used regsubsets again to find the 6 most useful predictors in our set of all numerical predictors. From the plot below, we can see that the predictors ADJDE, EFG_O, EFG_D, TORD, DRB, and WAB contribute significantly to the value 0.78 in R-squared.



Using these predictors, we form a model as shown below.

Table 8: Summary of the model using regsubsets

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.4153616	0.0659503	-6.298	3.7e-10 ***
ADJDE	0.0178323	0.0008424	21.169	< 2e-16 ***
EFG_O	0.0111078	0.0009169	12.115	< 2e-16 ***

EFG_D	-0.0284790	0.0012300	-23.155	< 2e-16 ***
TORD	0.0273352	0.0011590	23.585	< 2e-16 ***
DRB	-0.0157243	0.0007866	-19.990	< 2e-16 ***
WAB	0.0227946	0.0006574	34.673	< 2e-16 ***

Table 9: Model with Numerical Predictors 3

$Observations$	$Residual\ Std.\ Erro$	R^2	$R^2\ adjusted$
1985	0.08581	0.7996	0.7989

A variance inflation factor was then run on the reduced linear regression which still indicated multicollinearity. The VIF is shown below:

Numerical Predictor	VIF
ADJDE	7.268094
EFG_O	1.973169
EFG_D	3.080797
TORD	1.580209
DRB	1.618042
WAB	4.926852

Since ADJDE has VIF greater than 5, we remove the variable and create a new model with the 5 remaining predictors, as shown below:

Table 10: Summary of the model using regsubsets

Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1865791	0.0658398	2.834	0.00465 **
EFG_O	0.0175722	0.0009565	18.370	< 2e-16 ***
EFG_D	-0.0116711	0.0010393	-11.229	< 2e-16 ***
TORD	0.0150021	0.0011086	13.532	< 2e-16 ***
DRB	-0.0077930	0.0007653	-10.183	< 2e-16 ***
WAB	0.0135343	0.0005430	24.926	< 2e-16 ***

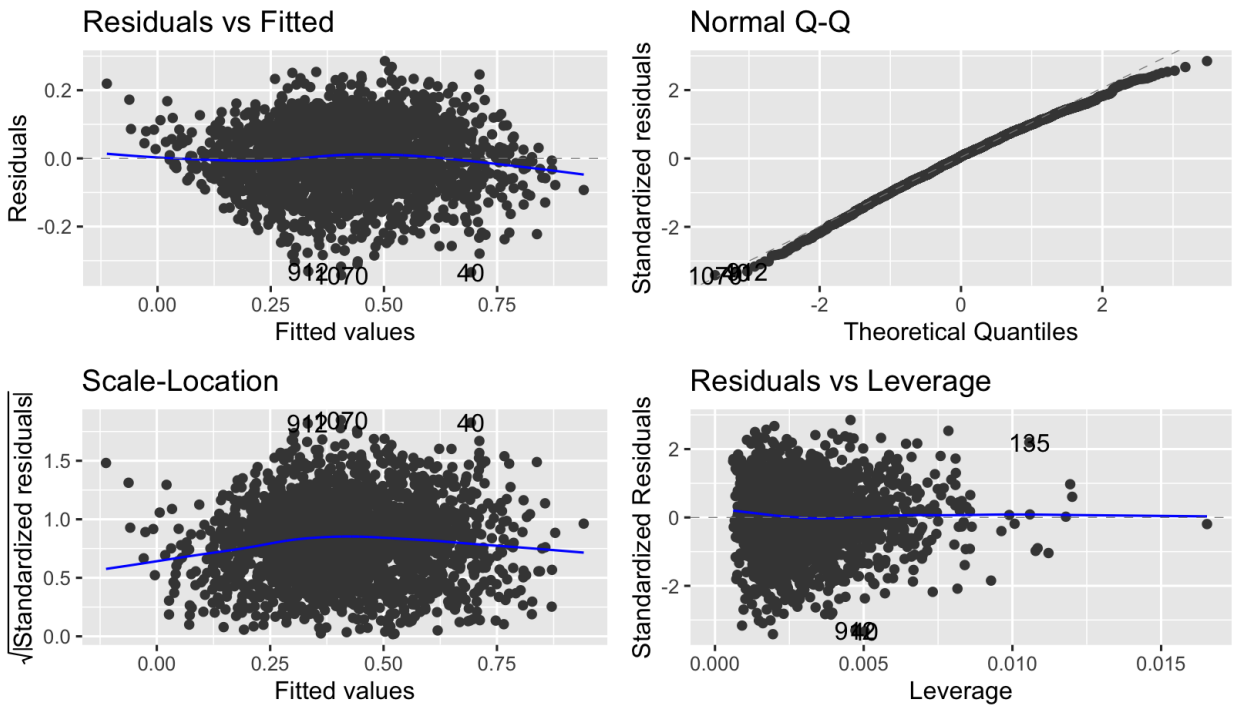
Tabel 11: Model with Numerical Predictors 4

$Observations$	$Residual\ Std.\ Erro$	R^2	$R^2\ adjusted$
2000	0.1005	0.725	0.7243

The VIF for this model is as shown below:

Numerical Predictor	VIF
EFG_O	1.754290
EFG_D	1.796956
TORD	1.180937
DRB	1.250978
WAB	2.745400

None of the VIF for the numerical predictors exceeds the value of 5. Therefore, we conclude that the numerical predictors do not indicate the violation of multicollinearity. Next, we check the diagnostic plots:



From the Residuals vs. Fitted values, we see that the line is roughly horizontal and that there is no pattern in the fitted residuals. Therefore, the assumptions of independent residuals and equal variance are satisfied. The Normal Q-Q plot has some small deviations at the top left corner but still satisfies the assumption of normal residuals. Since all the assumptions for the diagnostic plots are not violated, thus we state that the numerical predictors we would use are EFG_O, EFG_D, TORD, DRB, and WAB.

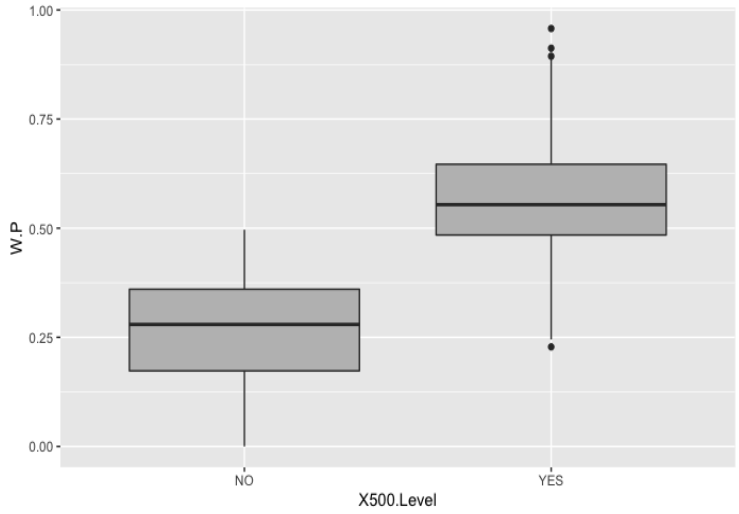
The Categorical Predictors

Below is a table of all categorical predictors in the NCAA data and the number of categories they have.

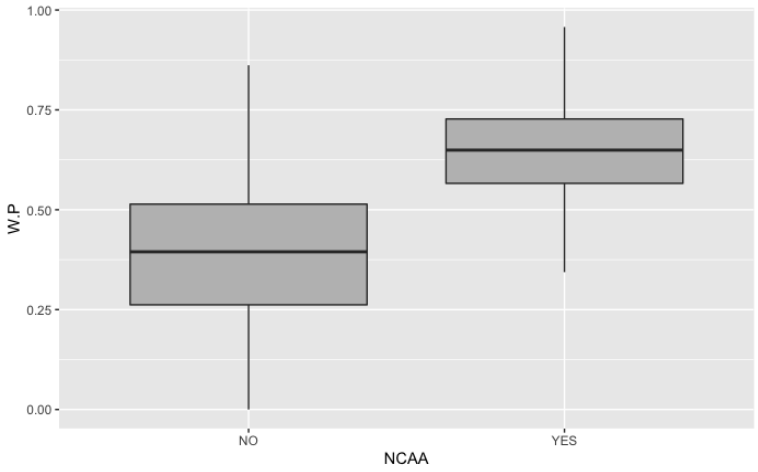
Table 12: Categories and Levels

Categorical Predictor	Number of Categories
X500.Level	2
NCAA	2
Power.Rating	3
YEAR	9

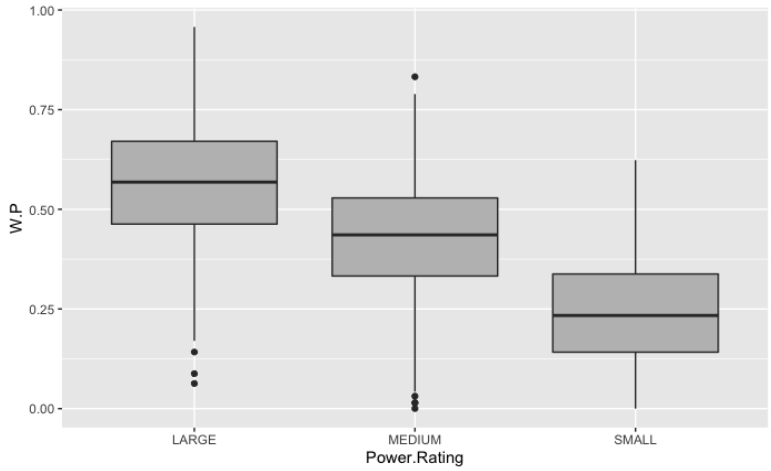
X500.Level: Whether or not a team has more wins than losses. (Yes means more wins than losses or No means more losses than wins)



NCAA: A categorical predictor that indicates whether the team made it to the NCAA or not (Yes or No)



Power.Rating: A categorical predictor that indicates the level of the chance of beating an average Division I team (Small, or Medium, or Large)



YEAR: College Basketball Seasons 2013 - 2021

This is followed by a scatterplot matrix of W.P along with the categorical predictors. The scatterplot matrix shows that X500.Level would be good predictors of W.P.

Scatterplot Matrix

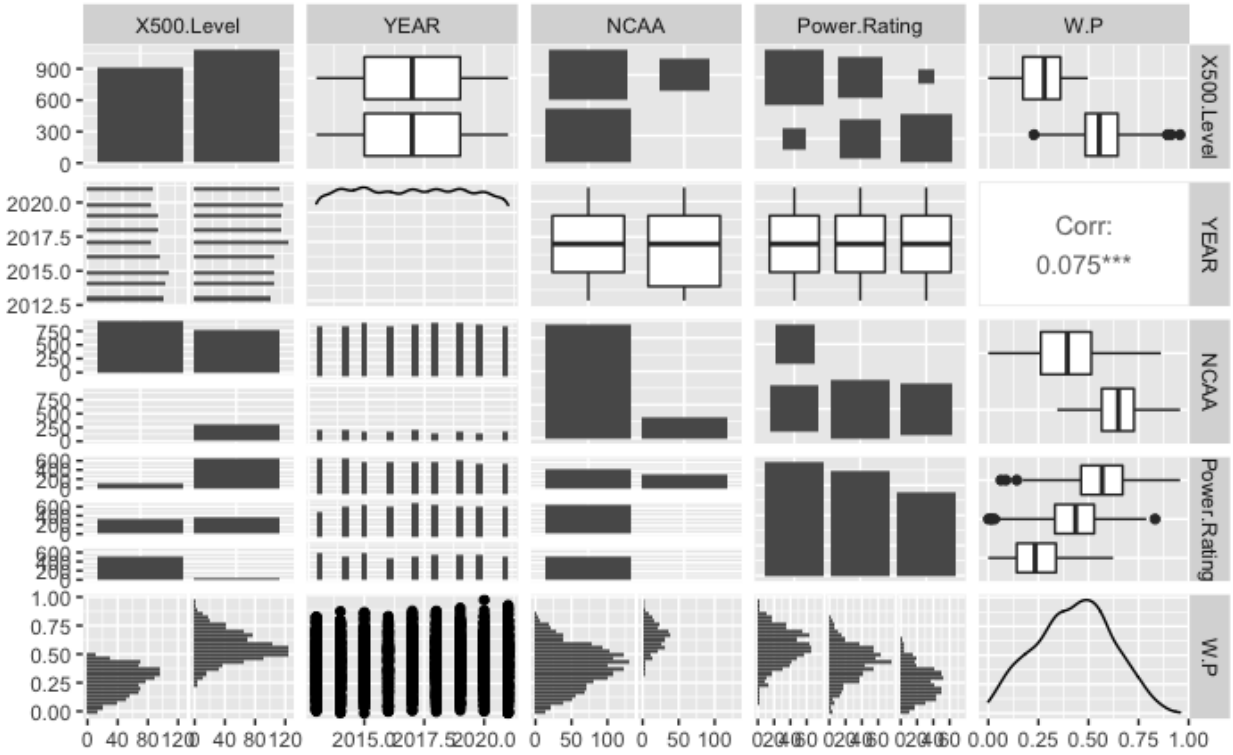


Table 13: Categories and Chi-square multicollinearity

Chi-square test	P-value
X500.Level, NCAA	2.2e-16
X500.Level, YEAR	0.2181
X500.Level, Power.Rating	2.2e-16
YEAR, NCAA	0.6839
YEAR, Power.Rating	0.9624
NCAA, Power.Rating	2.2e-16

Based on the Chi-square tests above, we can remove the categorical variables NCAA and Power.Rating due to their multicollinearity with the variable X500.Level. Additionally, the variable YEAR may be kept because it does not violate the multicollinearity in the model. However, since the YEAR in the dataset does not explain anything regarding winning proportion (the winning proportion of a team is not correlated with what year they were playing in based on the scatter plot above), we will use X500.Level as the only categorical predictors in our model.

Building the Model

Initial Attempt

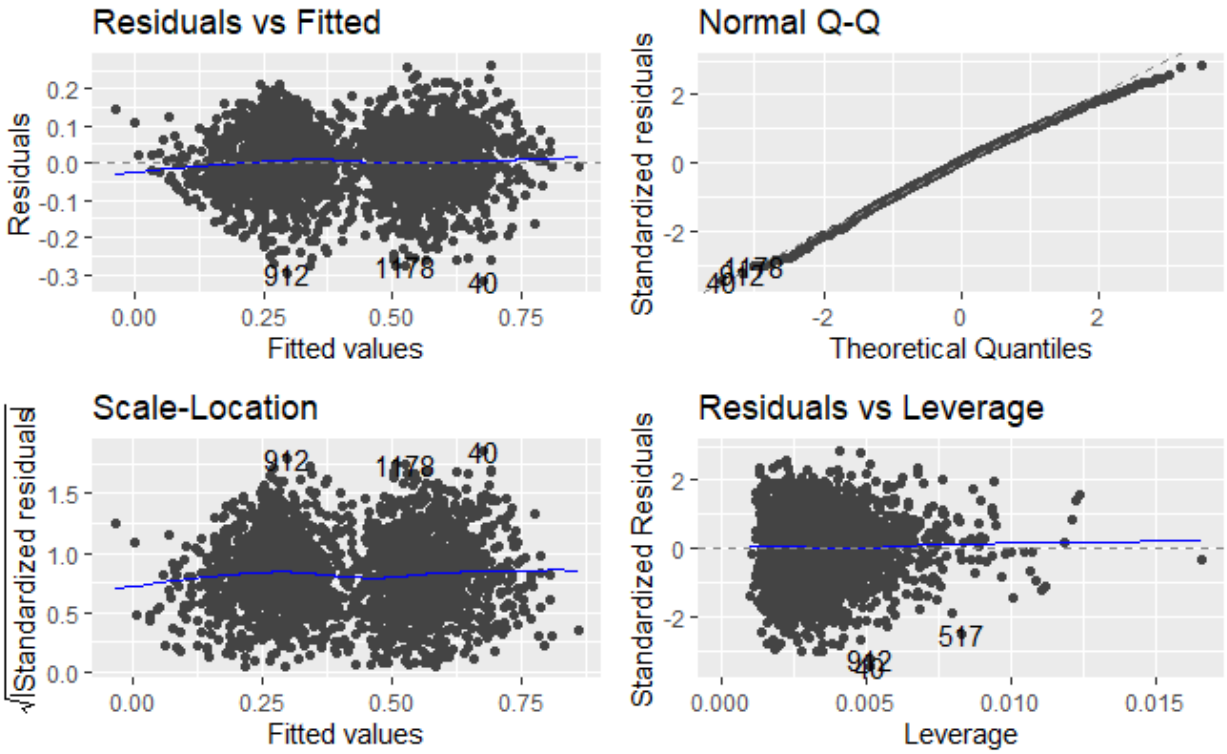
We first build a model by combining the most important numerical variables from regsubsets methods (EFG_O, EFG_D, TORD, DRB, and WAB) with a notable categorical variable (X500.Level) found above. Since our variables are selected carefully using statistical methods, the result below shows that all of our variables are statistically significant. The diagnostic plots do not show major violations of assumptions. Namely, the residual plots do not show a pattern and it indicates that we have constant variances. We need to note that there is a separation in fitted values due to the categorical variable in the model and some potential outliers and leverage points.

Table 14: Initial model with numerical and categorical variables

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.1293409	0.0603728	2.142	0.0323	*
EFG_O	0.0131275	0.0009050	14.506	< 2e-16	***
EFG_D	-0.0076568	0.0009737	-7.864	6.07e-15	***
TORD	0.0115655	0.0010304	11.224	< 2e-16	***
DRB	-0.0062305	0.0007055	-8.832	< 2e-16	***
WAB	0.0095638	0.0005370	17.810	< 2e-16	***
X500.LevelYES	0.1228588	0.0062693	19.597	< 2e-16	***

Table 15: Model of Initial Attempt

Observations	*Residual Std. Erro*	R^2	R^2 *adjusted*
2000	0.09201	0.7695	0.7688



Improving the model

To improve our prediction model even further, we added interaction terms and explored transformations. By considering different combinations of numerical variables, we have determined that the interaction between ADJOE and TOR, and ORB and FTR provide statistically significant contributions to the prediction model. Furthermore, after performing a box cox transformation, we found that the interaction between (ADJOE)^(-0.2) and ADJDE provides a prominent contribution in increasing R-squared. This leads to our final model in the next section.

5 and 6. Results / Discussion

Final Model:

Our final model consists of EFG_O, EFG_D, TORD, DRB, WAB, X500.Level, (ADJOE^-0.2):ADJDE, ADJOE:TOR, and ORB:FTR. The summary is provided below:

Table 16: Summary of Final Model

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3.464e-01	7.013e-02	-4.940	8.47e-07	***
EFG_O	1.437e-02	9.345e-04	15.372	< 2e-16	***
EFG_D	-1.937e-02	1.160e-03	-16.699	< 2e-16	***
TORD	2.011e-02	1.067e-03	18.850	< 2e-16	***
DRB	-1.189e-02	7.314e-04	-16.261	< 2e-16	***
WAB	1.754e-02	8.415e-04	20.850	< 2e-16	***
X500.LevelYES	8.944e-02	6.027e-03	14.840	< 2e-16	***
newADJOE:ADJDE	2.979e-02	2.046e-03	14.558	< 2e-16	***
ADJOE:TOR	-9.887e-05	1.333e-05	-7.415	1.79e-13	***
ORB:FTR	5.747e-05	9.747e-06	5.897	4.35e-09	***

Table 17: Final Model

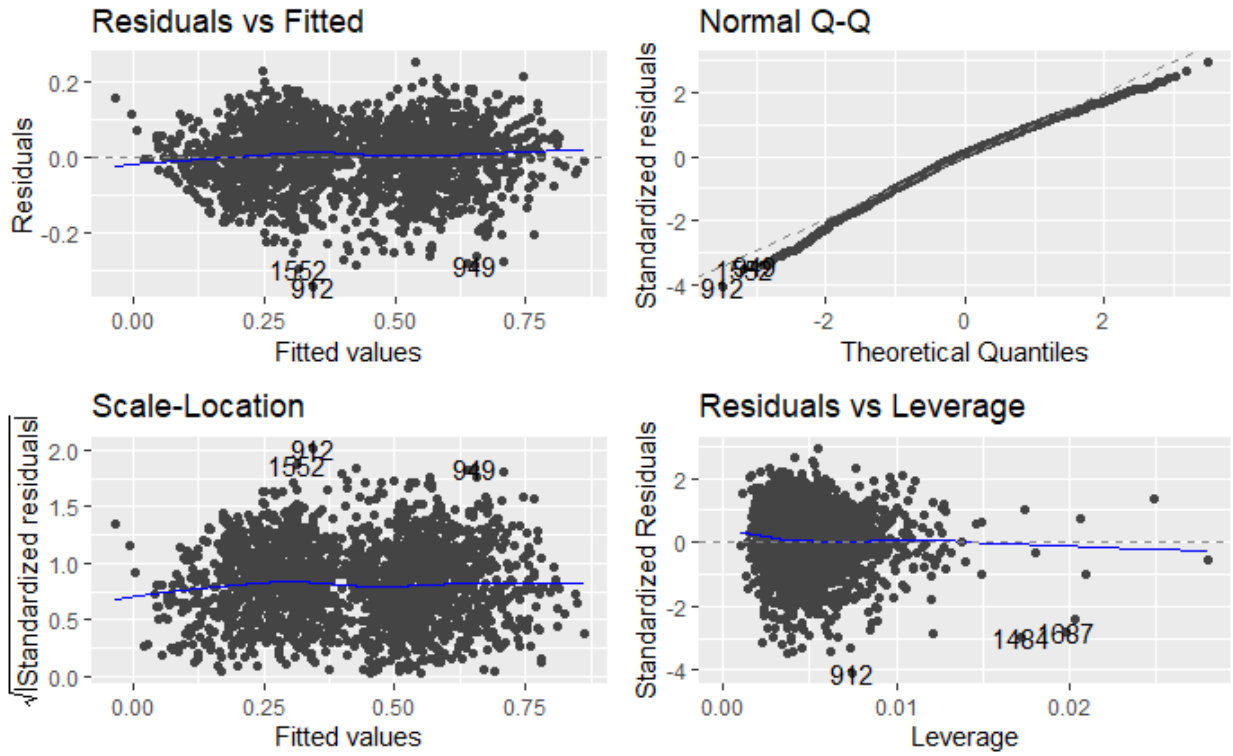
$Observations$	$Residual\ Std.\ Erro$	R^2	$R^2\ adjusted$
2000	0.0846	0.8054	0.8045

The equation of final model is given by:

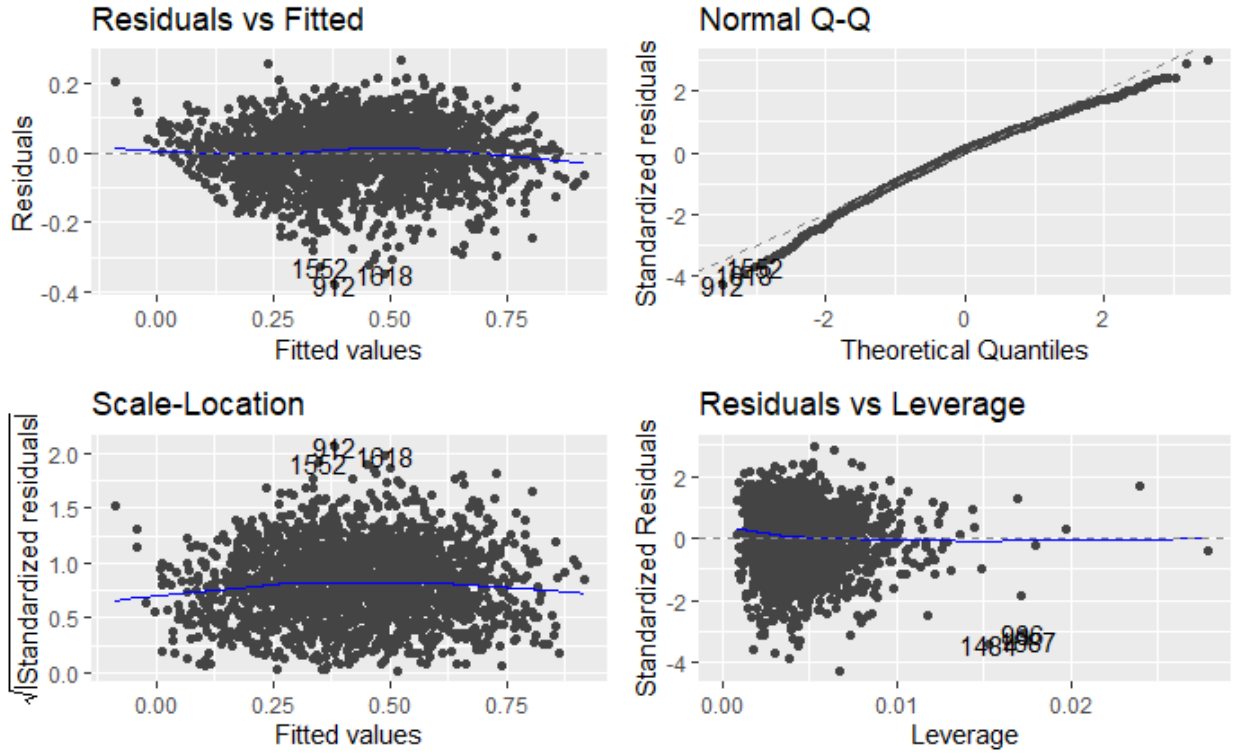
$\widehat{W.P}$ = -3.464e-01 + 1.437e-02 (EFG_O) + -1.937e-02 (EFG_D) + 2.011e-02 (TORD) + -1.189e-02 (DRB) + 1.754e-02 (WAB) + 8.944e-02 (X500.Level) + 2.979e-02 (ADJOE^(-0.2) * ADJDE) + -9.887e-05 (ADJOE * TOR) + 5.747e-05 (ORB * FTR)

We see that all our coefficients are statistically significant and R-squared have improved from previous attempts.

Diagnostic Plots



Diagnostics plots: The diagnostic plots with all predictors



Diagnostics plots: The diagnostic plots with only numerical predictors

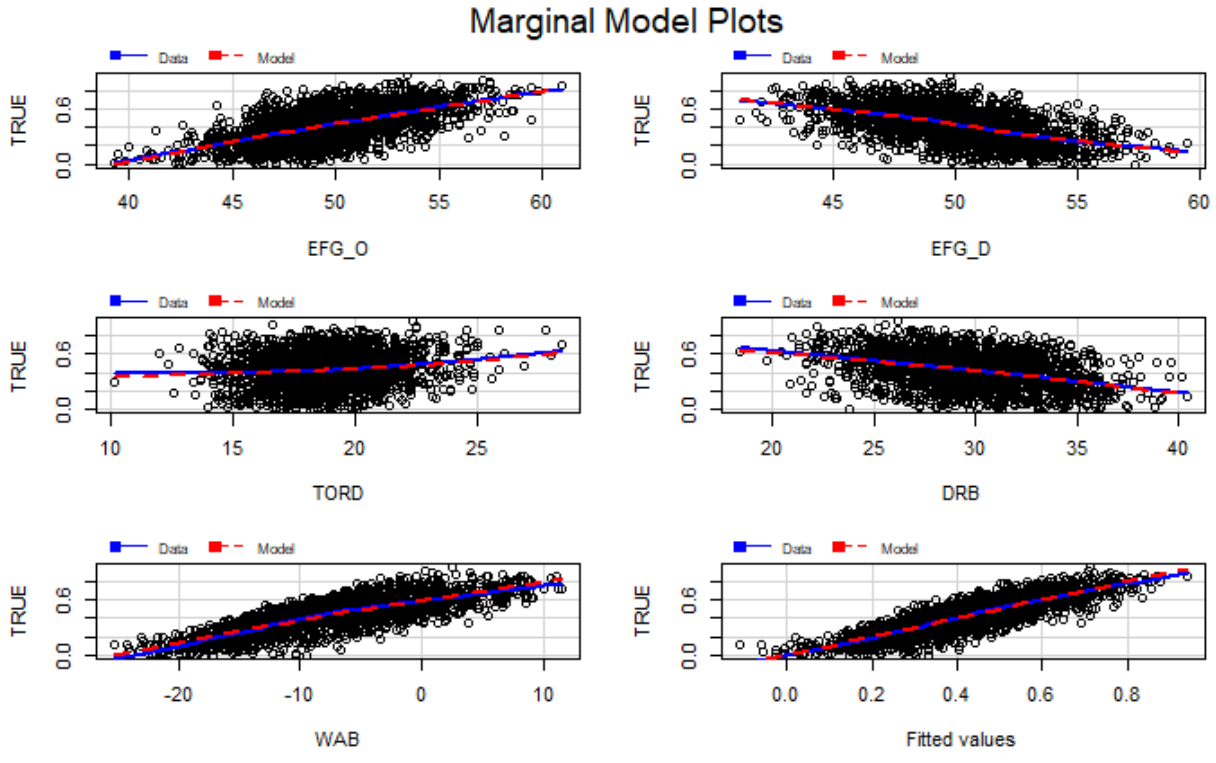
From the diagnostic plots above, we see that there are no major violations of assumptions. Residuals appear to be independent, relatively constant and normally distributed. By comparing two plots, we observed that the separation of residuals is due to the addition of a categorical variable (X500.Level), which is an important indication that the predictor is making a statistically significant impact on our model.

Table 18: VIF (Variance Inflation Factor) Table

Numerical Predictor	VIF
EFG_O	1.754290
EFG_D	1.796956
TORD	1.180937
DRB	1.250978
WAB	2.745400

The VIF table above depicts that we have no serious multicollinearity issues because the variables without interaction terms do not have a variable that is greater than 5 (corresponding to 80% R-squared for each variable).

Marginal Model Plots



The marginal plot above portrays that our final prediction follows the pattern of data; therefore, appropriate variables and transformations are used for constructing the model. This provides further evidence for the validity and meeting the diagnostic requirements.

Leverage and Outliers

Table 19: Leverage and Outliers from the model

Leverage/Outlier	Yes	No
Yes	5	60
No	90	1845

The table above shows that our model has 5 bad leverage points and 60 good leverage points. There are 90 outliers, but most of our dataset are ordinarily points. This does not raise a significant concern because the bad leverage points are only 0.25% of overall training data. We decided to keep them to our data because it reflects the unusual data points, and it did not improve our model by removing it.

7. Limitations and Conclusion

Some of the limitations in creating our model is the inability to include some predictors highly correlated with W.P in the multiple linear regression due to high VIF and multicollinearity. Many of the predictors in the training dataset are highly correlated. This causes the model to overfit and explain deviation that is already explained by another variable (Almohalwas 2020). By removing those variables we fixed the multicollinearity issues.

The diagnostic plots in our result section show that we addressed major assumptions for multiple linear regression, and none of them has a serious problem. While there are several minor issues, such as a small number of bad leverage points and a slight deviation from normal qq-plot line, those are not major concerns. Other plots such as residual plots and marginal model plots verify the validity of our model.

By looking at the coefficients of our model, the number is expected based on the context of basketball. For instance, higher EFG_O, effective field goal percentage, or higher shooting percentage lead to better players and teams (Sampaio et al.). Despite some of the interaction terms being statistically significant, the interpretation of those terms is hard to define.

Our final kaggle R-squared was 0.8054, which compared to our training model R-squared was .80935. The consistency in R-squared shows that the final model was able to generalize to newer testing data, and we did not overfit our data by including more variables.

8. References

Almohalwas, Akram. 2020. *Chapter 6 Winter 2022*.

Almohalwas, Akram. 2020. *Chapter 7 Updated Winter 2022*.

2022. *STAT 101A Winter 2022 Kaggle Competition*: Predicting Winning Proportion Using College Basketball Data.

Kaggle. 2021. "cbdTrain.csv." *https://www.kaggle.com/c/predicting-winning-proportions/data*.

Sampaio et al. 2015. Exploring Game Performance in the National Basketball Association Using Player Tracking Data. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4501835/.*