Austin Cicale
January 15, 2022

SAIL Application Paper

INTRODUCTION

Whether you're a fan or player, sports are a huge part of most cultures. The number of sports that have been played is endless, and they all differ in their own unique ways. But the one theme that remains constant is the importance of winning. Basketball is one of America's most popular sports, and there are numerous parts of the game that contribute to a team's winning capability. To determine which elements of basketball are most influential on winning, this paper created a model to predict winning percentage using the variables provided in Andrew Sundberg's College Basketball Dataset. Understanding which aspects of basketball are the best predictors of winning can help teams determine strategy adjustment in an attempt of improving success.

DATA

The College Basketball Dataset created by Andrew Sundberg contains data from the 2013 to 2019 Division 1 college basketball seasons. The data includes 24 variables, ranging from the conference the school participates in to more advanced stats such as adjusted offensive efficiency. To account for dissimilarities in the number of games played by each team, a winning percentage variable (win_pct) was added to the data by dividing number of games won by number of games played. Before creating the model, variables in the data set that were redundant or unessential and therefore did not have potential to be predictors in the model were removed. These variables include team name, games played, games won, power rating, postseason elimination round, NCAA tournament seed, and year. The data was then randomly split into training and testing sets: 2000 observations were assigned to the training set and 455 observations were assigned to the testing set. This data separation process was implemented so cross-validation could later be used to evaluate the model's performance.

RESULTS

Using all variables from the filtered data set, three different methods were used to create a model predicting win percentage. The backward, forward, and stepwise selection methods that were used all resulted in the same model with thirteen predictor variables and a Mallow's Cp of 46.702. A Residuals vs Fitted plot was produced to check linearity and constant variance conditions for the model. The residuals appeared to fall along the horizontal axis fairly well, and there wasn't too much variability in the spread of residuals for different fitted values, signifying that these two conditions were met. Additionally, a Normal Q-Q plot was used to uphold the constant variance condition as there was minimal deviation from the qqline. The five largest absolute residuals of the model were analyzed to determine if the degree of their influence warranted removal of data points. The standardized and studentized residual values for these five points were all greater than the absolute value of three, which might create concern for potential influence; however, the similarity in standardized and studentized values for each residual doubts extreme influence. To further assess the five largest residuals, their leverages were calculated and compared to double and triple the average leverage. All of the leverage values for the five residuals were greater than double and triple the average leverage, indicating a greater potential for these points to be influential on the model. Influence estimations for the five largest residuals were computed using Cook's distance to collectively determine if any points were significantly impacting the model. With no Cook's distance values greater than 0.01, it was decided that there was no large concern for influence among the five largest absolute residuals and no data points would be removed.

There was experimentation with the previously created model to check if changes or additions could be made to improve predictions of win percentage. Several transformations were made to the response and predictor variables in an attempt to increase the percent of variance explained by the model's predictors (Adjusted R-squared). There did not appear to be any variable transformations that resulted in an increase in the Adjusted R-Squared. However, there were some variable interactions that showed potential for model improvement: Adjusted Offensive Efficiency (ADJOE) with Turnover Percentage Committed (TORD), ADJOE with Effective Field Goal Percentage Shot (EFG_O), TORD

with Adjusted Tempo (ADJ_T), Adjusted Defensive Efficiency (ADJDE) with Effective Field Goal Percentage Allowed (EFG_D), and ADJDE with Free Throw Rate (FTR). Within each of these variable combinations, there is logical potential for the relationship between the predictor and response to vary with values of the predictor it's paired with. For example, a higher turnover percentage committed (steal rate) could increase the likelihood to score points on fast break opportunities. This might decrease the importance that offensive efficiency has on the prediction of win percentage.

The five variable interactions listed above were added to the original model and the Adjusted R-squared increased from 0.8733 to 0.8759. The addition of these interactions appeared to improve the model according to the percentage of variance explained by the predictors, but there was no certainty whether all five of the interactions should be included. Backward selection was used to create the most efficient model incorporating the five variable interactions in addition to all variables in the original model as potential predictors of winning percentage. This selection method did not remove any of the potential predictor variables; all of the predictors from the original model and all five variable interactions were included. The coefficient and p-value of each predictor variable in the model, except for Conference (CONF), is represented in Figure 1. The listed variables and interactions, all statistically significant, are the best predictors of winning according to the model. Like before, a Residuals vs Fitted and Normal Q-Q plot were produced to check linearity, constant variance, and normality conditions. These plots were extremely similar to those made for the original model. All conditions appeared to be fairly met. To reassure the new model including the interaction terms is an improvement of the original model, ANOVA testing was performed to compare the two models. The test resulted in an extremely small p-value of 1.247e-08, concluding that the slope relating win percentage to the interaction terms is non-zero for at least one of the variable interactions. This is the result that was expected, and it supports continuation of the analysis using this new, finalized model.

Figure 1

Final Model Predictor Variable Summary

Predictor Variable	Coefficient	P-value
ADJOE	-2.462e+00	3.09e-05
EFG_D	2.346e-02	1.55e-10
ADJDE	1.167e-02	0.002695
EFG_O	3.763e-02	7.33e-12
Turnover Rate (TOR)	-1.412e-02	< 2e-16
Offensive Rebound Rate (ORB)	4.899e-03	1.62e-13
Adjusted Tempo (ADJ_T)	-4.867e-03	0.202948
FTR	1.609e-02	0.000222
Free Throw Rate Allowed (FTRD)	-4.135e-03	< 2e-16
TORD	1.549e-02	0.326147
Offensive Rebound Rate Allowed (DRB)	-1.042e-02	< 2e-16
Three-Point Shooting Percentage Allowed (3P_D)	-3.484e-03	9.63e-05
ADJOE:TORD	-2.282e-04	0.011910
ADJOE:EFG_O	-2.193e-04	3.57e-05
EFG_D:ADJDE	-2.343e-04	0.000187
ADJ_T:TORD	4.279e-04	0.032541
ADJDE:FTR	-1.329e-04	0.001519

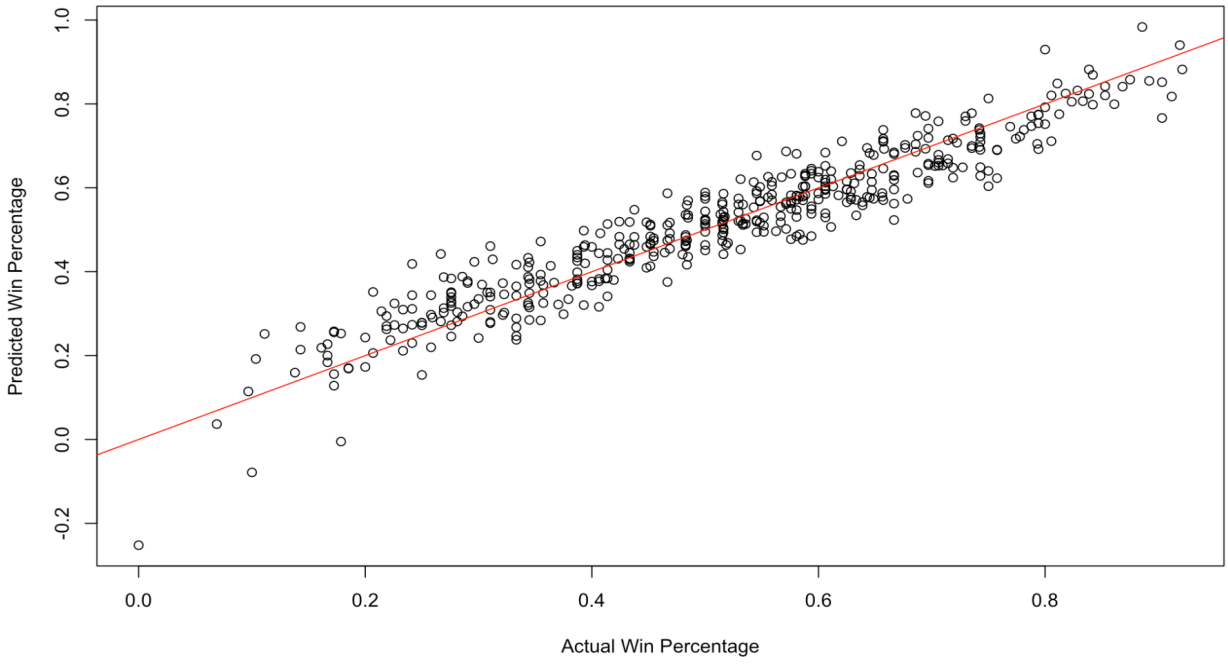
Note. Categorical variable Conference (CONF) not included for spacing reasons.

The final step is to see how the finalized model performs when it comes to predicting winning percentage for a new set of data. The model was used to compute residuals for the testing data. The mean and standard deviation of these residuals were calculated to address zero mean and variability conditions. With a mean value of about -0.002, the model is overpredicting on average, being off by about 0.2% of a team's true winning percentage. This value is extremely close to zero, so the zero mean condition is being upheld. The standard deviation of the holdout cases is very similar to the residual standard error of the

original model at just around 6%. The similarity in these values concludes that the variability condition appears to be met. Plots of the testing and trained data residuals were produced, and the distribution of residuals for both sets of data seem to have similar and minimal issues with normality. With these three conditions being upheld, we can proceed cross validating the model's capability of predicting values for complementary data. The shrinkage was calculated by squaring the cross-validation correlation and subtracting it from the Multiple R-squared of the training sample, resulting in an approximate value of -0.024. The shrinkage is fairly small and close to zero, which is a good indication that the model seems to similarly predict win percentage for the training and testing data. A graphical representation of the model's ability to predict win percentage for the testing data is shown in Figure 2. The red curve is a line of slope = 1 and intercept = 0. At every point on this line, the actual winning percentage is equal to the predicted winning percentage. If the model predicts efficiently, the raw data should relatively follow the red curve. This appears to be the case, further supporting the model's ability to accurately predict winning percentage for out-of-sample data.

Figure 2

Actual Win Percentage vs Predicted Win Percentage



CONCLUSION

While basketball may appear to be a relatively simple sport, there are various elements of the game that are commonly overlooked regarding their significance to winning. To determine which parts of the game have the largest impact on success, a model was created from a group of potential variables with the goal of predicting win percentage. After analyzing various selection methods and experimenting with variable transformations, the model was finalized. Thirteen predictors and five variable interactions were incorporated into the model, and their levels of significance can be seen in Figure 1. Not only are these variables capable of accurately predicting a team's win percentage, but they should be considered some of the most impactful parts of the game. Coaches should use these results to develop offensive and defensive strategies that focus on optimizing these variables to their benefit. Future direction to improve the model might include using data with more potential predictors.