KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Predicting College Football Team Wins Using Stepwise Regression

Brett Martin

Faculty Advisor: Dr. Austin Brown

INTRODUCTION

The goal of this project is to build a model which can be used to predict the number of wins a college football team achieved during a certain season. There have been several "eras" of college football including the Bowl Championship Series (BCS), and more recently, the College Football Playoff (CFP). The 2014 season marked the inaugural season of the CFP. This season presented a good beginning point for the study.

The seasons considered in this study were the inaugural season of the CFP (2014) until the 2019 season. The 2020 season was omitted since it was heavily influenced by the COVID-19 pandemic. For instance, all teams played substantially fewer games compared to the standard ~12 games per season. Furthermore, the number of games played by each team varied, therefore contributing further to the potential for complications regarding the collection of meaningful data.

METHODS

- Initially, 26 statistics were gathered based on their suspected correlation to the success of a team. There were 774 total observations, each observation listing a team's statistics for a given season. Despite already being recorded, significant consolidation was required given the data had to be collected from numerous different tables.
- The goal of this analysis was accomplished through utilizing stepwise regression to select the ideal variables for predicting the number of wins a team accomplished for a given season. Variable selection via stepwise regression suggested that 11 of 26 variables should be included in the model; however, due to concerns regarding multicollinearity, "Points_Per_Game" was removed.
- The prediction values were generated once the final model was determined. These values were then placed in a table along with their respective team and season. This table was then left joined with the original table by "Team" and "Season" such that the prediction values were added to the table with the other variables. The difference between the prediction values and the actual number of wins were calculated. A bar chart was used as a visual representation of the count of differences within certain ranges.
- K-fold cross validation was used to assess the predictive capability of the model. A correlation heat matrix was also used to study relationships between predictors.

Table 1: Stepwise Selection Summary

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Points_Per_Play	addition	0.591	0.590	1287.6600	3238.7194	1.9556
2	Opponent_Points_Per_Game	addition	0.818	0.817	149.2000	2615.3278	1.3065
3	Sacks_Per_Game	addition	0.830	0.829	91.2760	2565.0325	1.2639
4	Third_Down_Conversion_Percentage	addition	0.839	0.838	46.6260	2523.7244	1.2298
5	Completion_Percentage	addition	0.841	0.840	37.0810	2514.6220	1.2218
6	Punt_Attempts_Per_Game	addition	0.844	0.843	25.0390	2502.8997	1.2118
7	Time_of_Possession_Percentage	addition	0.845	0.844	20.1210	2498.0490	1.2073
8	Opponent_Yards_Per_Pass_Attempt	addition	0.846	0.845	16.9420	2494.8799	1.2040
9	Opponent_Yards_Per_Game	addition	0.848	0.846	12.3650	2490.2662	1.1997
10	Yards_Per_Pass_Attempt	addition	0.848	0.846	11.0000	2488.8597	1.1978

Table 2: Beta Coefficients for Linear Model

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-5.098235	1.358663	-3.752	0.000188	***
Sacks_Per_Game	0.619401	0.083879	7.384	4.02e-13	***
Completion_Percentage	3.039113	0.929738	3.269	0.001129	**
Opponent_Yards_Per_Pass_Attempt	-0.263891	0.081344	-3.244	0.001229	**
Opponent_Points_Per_Game	-0.195435	0.015382	-12.706	< 2e-16	***
Third_Down_Conversion_Percentage	7.013855	1.221106	5.744	1.34e-08	***
Time_of_Possession_Percentage	6.168447	1.597918	3.860	0.000123	***
Opponent_Yards_Per_Game	0.005118	0.001982	2.582	0.009999	**
Punt_Attempts_Per_Game	0.291555	0.079270	3.678	0.000252	***
Yards_Per_Pass_Attempt	0.114620	0.062479	1.835	0.066965	.
Points_Per_Play	15.722503	0.906791	17.339	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3: First Five Rows of Final CFB Stats Table

Team	Season	Prediction	Wins	Difference
<chr>	<db1>	<db1>	<db1>	<db1>
Ohio State	2019	15.1	13	2.05
Alabama	2018	14.5	14	0.453
Alabama	2019	14.3	11	3.29
Clemson	2018	14.3	15	-0.745
Clemson	2019	13.8	14	-0.200
Penn State	2017	13.3	11	2.30

Figure 1: Correlation Heatmap Matrix for Selected Variables

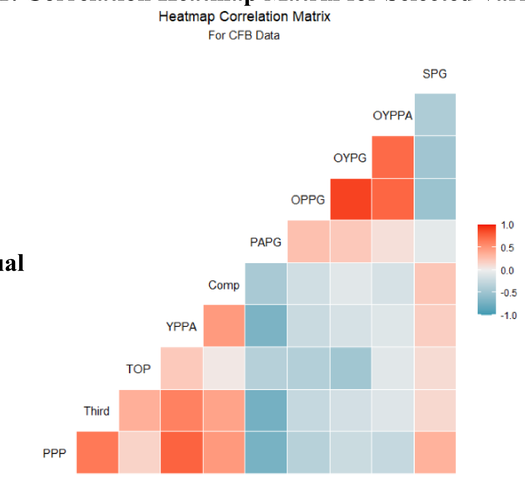
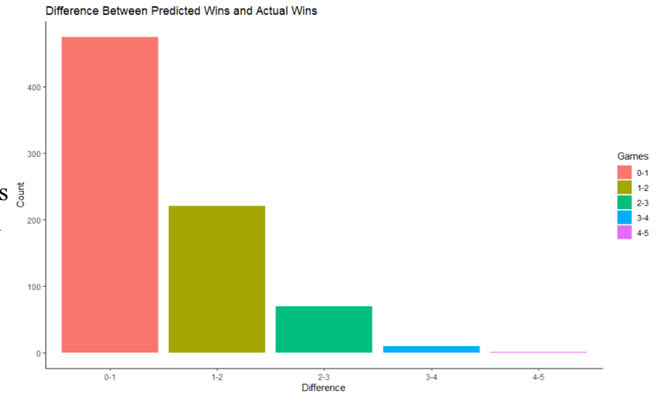


Figure 2: Bar Chart for Difference in Prediction and Actual



RESULTS & DISCUSSION

- The steps of the stepwise selection process are indicated in Table 1. Adjusted R-Square for the final model equals 0.846, C(p) equals 11.0000, AIC equals 2488.8597 and RMSE equals 1.1978.
- Table 2 shows the coefficients for the final model. The beta values for "Third_Down_Conversion_Percentage", "Time_of_Possession_Percentage", and "Points_Per_Play" indicate these variables have the greatest influence on the response variable ("Wins"). On the other hand, "Opponent_Yards_Per_Game" has a very small beta value which indicates very little influence on the response.
- Table 3 shows the first five rows of the data set which includes "Season", "Team" and "Prediction". The difference between "Prediction" and "Wins" is given by "Difference".
- Figure 1 gives a visualization for how the predictor variables in this study are correlated. Opponent points per game ("OPPG") and opponent yards per game ("OYPG") appear to share the highest correlation. Both these predictors also appear to be strongly correlated with opponent yards per pass attempt ("OYPPA").
- Figure 2 shows the counts of the difference between the number of predicted wins and the number of wins a team achieved. The number of wins were predicted within 1 for ~61% of the observations, within 2 for ~90% of the observations, within 3 for ~99% of the observations, and within 4 for ~99.9% of the observations.
- K-fold cross validation indicates the mean MSE is roughly 1.45 and mean MAE is roughly 0.96.

R CODE

```
k <- 5
p1 <- na.omit(CFB_Stats_3)
p1$kfold <- sample(1:5,nrow(p1),replace=T)
mse <- vector("double",length=k)
mae <- vector("double",length=k)
betaz <- list(length=k)
for(i in 1:k){
  df_test <- p1 %>% dplyr::filter(kfold == i)
  df_train <- p1 %>% dplyr::filter(kfold != i)
  k_mod <-
lm(Wins~PPP+Third+TOP+YPPA+Comp+PAPG+OPPG+O
YPG+OYPPA+SPG,data=df_train)
  pvs <- predict(k_mod,newdata=df_test)
  mse[i] <- mean((pvs - df_test$Wins)^2)
  mae[i] <- mean(abs(pvs - df_test$Wins))
  betaz[[i]] <- coef(k_mod)
}
```