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. 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. The difference between the prediction values and 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 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 1.978.

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_Pass_Attempt” 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, 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.

**RESULTS & DISCUSSION**

**INTRODUCTION**

**METHODS**

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable</th>
<th>Removed</th>
<th>R-Square</th>
<th>R-Square</th>
<th>C(p)</th>
<th>AIC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Points_Per_Play</td>
<td>addition</td>
<td>0.591</td>
<td>0.590</td>
<td>1287.6600</td>
<td>3238.7194</td>
<td>1.9516</td>
</tr>
<tr>
<td>2</td>
<td>Opponent_Points_Per_Game</td>
<td>addition</td>
<td>0.818</td>
<td>0.817</td>
<td>149.2000</td>
<td>2651.3778</td>
<td>1.3065</td>
</tr>
<tr>
<td>3</td>
<td>Sacks_Per_Game</td>
<td>addition</td>
<td>0.830</td>
<td>0.829</td>
<td>91.2760</td>
<td>2545.0325</td>
<td>1.2639</td>
</tr>
<tr>
<td>4</td>
<td>Third_Down_Conversion_Percentage</td>
<td>addition</td>
<td>0.833</td>
<td>0.832</td>
<td>48.6260</td>
<td>2521.7484</td>
<td>1.2795</td>
</tr>
<tr>
<td>5</td>
<td>Completion_Percentage</td>
<td>addition</td>
<td>0.841</td>
<td>0.840</td>
<td>57.0810</td>
<td>2514.6220</td>
<td>1.2238</td>
</tr>
<tr>
<td>6</td>
<td>Punt_Attempts_Per_Game</td>
<td>addition</td>
<td>0.844</td>
<td>0.843</td>
<td>21.0390</td>
<td>2501.9975</td>
<td>1.2138</td>
</tr>
<tr>
<td>7</td>
<td>Time_of_Possession_Percentage</td>
<td>addition</td>
<td>0.845</td>
<td>0.844</td>
<td>20.1210</td>
<td>2498.0490</td>
<td>1.2073</td>
</tr>
<tr>
<td>8</td>
<td>Opponent_Yards_Per_Pass_Attempt</td>
<td>addition</td>
<td>0.846</td>
<td>0.845</td>
<td>18.9420</td>
<td>2494.3799</td>
<td>1.1940</td>
</tr>
<tr>
<td>9</td>
<td>Opponent_Yards_Per_Game</td>
<td>addition</td>
<td>0.848</td>
<td>0.847</td>
<td>12.3600</td>
<td>2490.2662</td>
<td>1.1977</td>
</tr>
<tr>
<td>10</td>
<td>Yards_Per_Pass_Attempt</td>
<td>addition</td>
<td>0.848</td>
<td>0.847</td>
<td>11.0000</td>
<td>2488.8597</td>
<td>1.1978</td>
</tr>
</tbody>
</table>

**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 1.978.

- 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_Pass_Attempt” 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, 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**

```r
k <- 5
pl1 <- na.omit(CFB_Stats_3)
pl1$kfold <- sample(1:5,nrow(pl1),replace=T)
mse <- vector("double",length=k)
mae <- vector("double",length=k)
betaz <- list(length=k)
for(i in 1:k){
df_train <- pl1[kfold != i]
df_test <- pl1[kfold == i]
mae[i] <- mean(abs(df_test$Wins - pl1$Wins))
mse[i] <- mean((df_test$Wins - pl1$Wins)^2)
betaz[[i]] <- coef(k_mod)
}
```