Technical Report

Lincoln Moehn, Zachary Rau, Curtis Leaver, and Xinsheng Tang

Abstract

The goal of this report is to analyze how a college football team's winning percentage is impacted by their turnover margin, penalties taken, and what conference they are in. To create our model, we eliminated variables that we felt were not useful in predicting a team's performance. We also decided to not use common predictors like points scored or points allowed for the purposes of researching predictors that are not frequently explored to discover new insights. The applications of this model are plentiful. One application could be to describe a team's performance relative to expectations and see if they are winning more or less than expected. This model can also help coaches form and develop game plans and strategies to benefit their teams and maximize their talent to win as many games as possible.

Introduction

College football busted on to the scene in 1869, when Rutgers played the College of New Jersey, which is now known as Princeton. Around 70 years later, in 1936, the first College Football rankings were established. Fast forward to today and there is a total of 11 different conferences in Division 1-A. In this Division there are a total of 133 schools. Every year these schools compete from late August to January, to see who will come out on top. Every game these teams play is extremely important because one loss, for most teams, means you are out of the chase for the playoffs. There is also a business side to college football, since good teams make their college millions of dollars each year in revenue. Consequently, we felt it would be important to know what factors contribute to a team's performance.

Data

```
library(tidyverse)
library(readxl)
library(skimr)
library(dplyr)
```

```
library(nnet)
library(car)
```

The data we obtained was from statista.com. It is a collection of the statistics and win/loss records of 130 division 1 college football teams from the 2022 college football season. For each team, there are about 50 statistics available to describe each team's performance for the 2022 season. These statistics range from a team's wins and losses during the season to how many points they scored during the season, to how many times they sacked the opposing team's quarterback throughout the year. The response variable for our model is the team's winning percentage (winPer). The explanatory variables we were most interested in were Points_per_game (numeric), Avg_points_per_game_allowed (numeric), Off_yards/play (numeric), Yards/play_allowed (numeric), Penalty_Yards_per_game (numeric), and AvgTurnoverMarg (numeric).

The name of each team had two problems: they included spaces and the conference they were in. For the first problem, we cleaned the data by replacing all spaces with underscores. To solve the latter complication, we created the new explanatory variable conference by splitting the team's name into team and conference. Another problem we had was that a lot of our variables were not needed, so we had to select the variables we wanted and get rid of the ones we didn't. One last idea we had was to add a variable confRank. We added this variable because historically, we know that some conferences have better teams on average than others. We looked at the conferences that historically performed the best and ranked them accordingly. The process and code we used to solve these problems can be seen below.

```
football <- readxl::read_excel("football.xlsx")
football <- football |>
  mutate(AvgTurnoverMarg = Avg_Turnover_Margin_per_Game) |>
  select(-Avg_Turnover_Margin_per_Game) |>
  separate_wider_delim(Team, names = c("TeamName", "conference"), delim = "(", too_many =
  mutate(TeamName = str_replace_all(TeamName, "_", "")) |>
  mutate(conference = str_replace_all(conference, "/", "_")) |>
  rename_all(~str_replace_all(., "/", "_per_")) |>
  mutate(conference = str_remove_all(conference, "\\)")) |>
  select(TeamName, conference, winPer, AvgTurnoverMarg, Sacks, Penalties)
#create numeric rank variable for conference, rankings by bleacherreport
football <- football |>
  mutate(confRank = if_else(conference == "SEC", 1,
                    if_else(conference == "Big_Ten", 2,
                    if_else(conference == "Big_12", 3,
                    if_else(conference == "Pac_12", 4,
                    if_else(conference == "ACC" | conference == "FL_(ACC"
                            | TeamName == "NotreDame", 5,
```

```
                    if_else(conference == "AAC", 6,
                    if_else(conference == "Sun_Belt", 7,
                    if_else(conference == "Mountain_West", 8,
                    if_else(conference == "C_USA", 9,
                    if_else(conference == "MAC", 10, 10))))))))))
```

Apart from these issues, our dataset seemed to be clean and manageable. We narrowed it down to 8 variables with no missing values.

```
skim(football)
```

Table 1: Data summary

Name	football
Number of rows	131
Number of columns	7
Column type frequency:	
character	2
numeric	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
TeamName	0	1	3	18	0	130	0
conference	0	1	3	15	0	13	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
winPer	0	1	0.52	0.20	0.08	0.42	0.54	0.69	1.00	
AvgTurnoverMarg	0	1	0.03	0.55	-1.58	-0.33	0.07	0.41	1.57	
Sacks	0	1	27.16	10.82	5.00	19.00	27.00	35.00	59.00	
Penalties	0	1	76.53	14.49	46.00	67.00	76.00	86.00	111.00	
confRank	0	1	5.60	2.98	1.00	3.00	6.00	8.00	10.00	

Furthermore, we wanted to explore variables that have not been explored before. It is commonly known that points scored, and points allowed were very significant predictors of a team's winning percentage, and there are numerous studies on the topic. This allowed us to eliminate the variables Points_Per_Game, Avg_Points_per_Game_Allowed, Touchdowns, Field_Goals, and Opp_Field_Goals_Made from our model.

```
fbmodel <- lm(winPer ~ AvgTurnoverMarg + confRank + Penalties, data = football)
summary(fbmodel)
```

```
Call:
lm(formula = winPer ~ AvgTurnoverMarg + confRank + Penalties,
    data = football)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3669 -0.1136 -0.0138  0.1208  0.4300

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.414513   0.080965   5.120 1.10e-06 ***
AvgTurnoverMarg  0.182738   0.026209   6.972 1.52e-10 ***
confRank        -0.014562   0.004881  -2.983  0.00342 **
Penalties        0.002392   0.001004   2.384  0.01862 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1653 on 127 degrees of freedom
Multiple R-squared:  0.3365,    Adjusted R-squared:  0.3209
F-statistic: 21.47 on 3 and 127 DF,  p-value: 2.578e-11
```

According to our model summary, about 32% of the variability in winning percentage can be explained by our variables. This is not a super high r-squared value, but given the data set that we selected and our choosing not to include points, yards, and touchdowns in our model, this is the best that we could come up with.

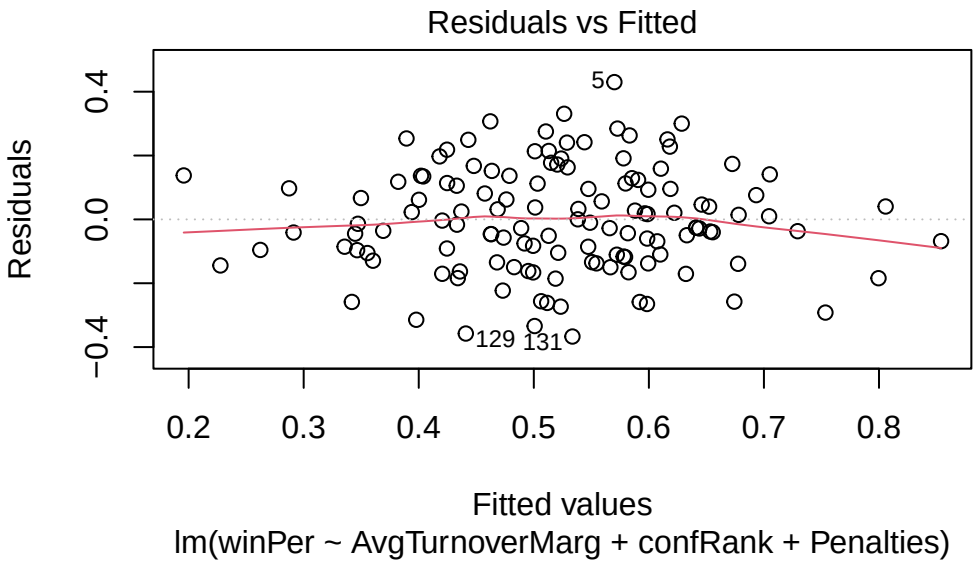
```
vif(fbmodel)
```

```
AvgTurnoverMarg        confRank       Penalties
       1.001899        1.006957        1.005315
```

Since we thought the coefficient for penalties was counter intuitive, we thought one possible explanation for this could be multicollinearity. So, we got the VIF values of the variables, but they look normal (all less than 5). Therefore, we can reject the idea that any multicollinearity was the cause of this effect on our model.

Once the final model was created, the conditions for inference had to be checked before any conclusion could be reached. The first condition for inference that was addressed for the model was normality, using the residuals vs. fitted plot.

```
plot(fbmodel, which = 1)
```



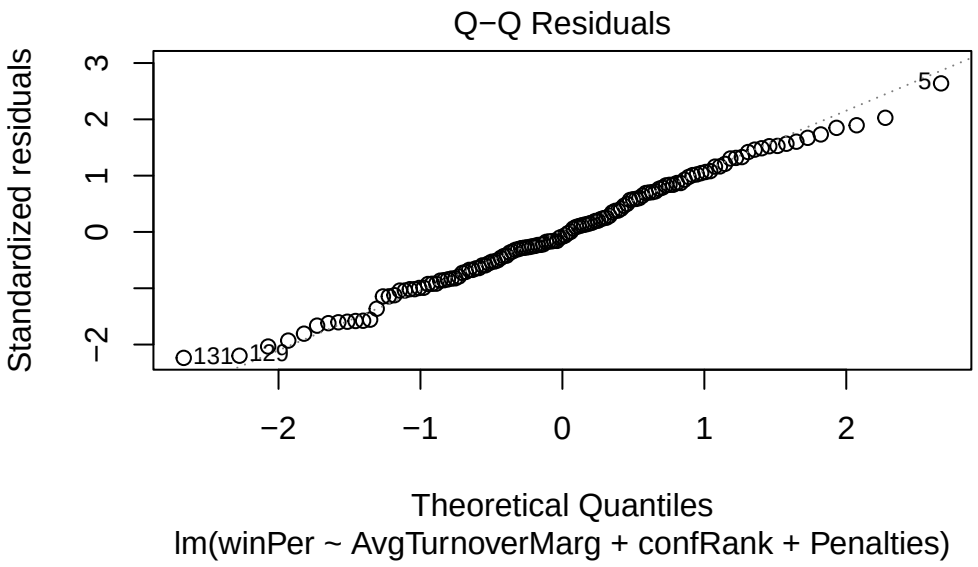
The linearity condition is upheld because the dots in the residual vs fitted plot seem have a relationship and the red line is very linear and follows the dotted line, which is enough evidence to say that our model passes this condition. No transformation is needed here.

Then, we needed to evaluate the independence condition by our own intuition, and thinking about whether the observations are independent of one another.

The independence condition is upheld because each school or team is independent of each other. There is a little bit of crossover since the games played between the teams are played against each other, but we still believe that the independence condition holds true.

Then, we had to evaluate normality using the Q-Q plot.

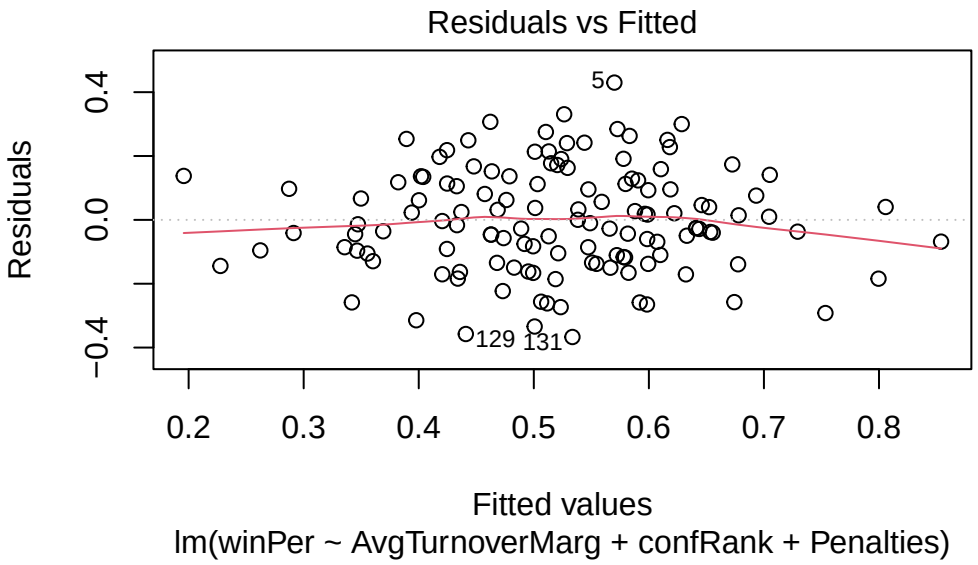
```
plot(fbmodel, which = 2)
```



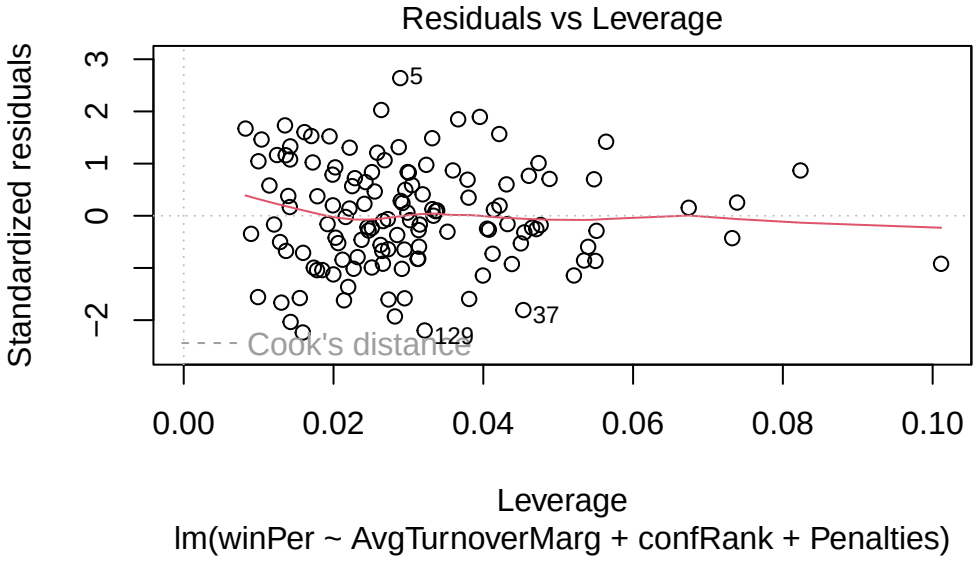
The Q-Q plot indicates that the normality condition is upheld. The standardized residuals follow the line given to us very closely. On the right side, there seems to be a little aberration from this trend, but it doesn't appear significant, and the last value corrects this trend. No transformation is needed here either.

Finally, we had to evaluate equality of variance using both the residuals vs. leverage plot and the residuals vs. fitted plot.

```
plot(fbmodel, which = 1)
```




```r
plot(fbmodel, which = 5)
```



Lastly, the equality of variance condition also passes since the residuals do not appear to form any fan-like shape in the residuals vs. fitted plot and no abnormalities are evident. We can also use the residuals vs. leverage plot to support this claim because the red line follows very closely along the dotted line, and also there are not any outliers according to Cook's distance. No transformation is needed here, making our work a lot easier.

Results

```
summary(fbmodel)
```

```
Call:
lm(formula = winPer ~ AvgTurnoverMarg + confRank + Penalties,
    data = football)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3669 -0.1136 -0.0138  0.1208  0.4300

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.414513   0.080965   5.120 1.10e-06 ***
AvgTurnoverMarg  0.182738   0.026209   6.972 1.52e-10 ***
confRank        -0.014562   0.004881  -2.983  0.00342 **
Penalties        0.002392   0.001004   2.384  0.01862 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1653 on 127 degrees of freedom
Multiple R-squared:  0.3365,    Adjusted R-squared:  0.3209
F-statistic: 21.47 on 3 and 127 DF,  p-value: 2.578e-11
```

A one-turnover increase per game may seem like a small shift, but it's akin to a ripple effect that echoes throughout the season. It's not just a statistic; it's a reflection of a team's ability to seize opportunities and beat opponents, altering the trajectory of success. We know based on the model that for a one-turnover increase in average turnover margin per game, the predicted winning percentage for a team is expected to increase 0.183, holding all other variables constant.

When it comes to conference rank, each shift represents a change in competition. A minor descent in rank isn't just a number; it's a subtle yet impactful change in predicted winning

percentage, illustrating the nature of sports hierarchies. We know that in our model, for a 1-rank increase in conference rank, the predicted winning percentage of a team is expected to decrease about 0.015, holding all other variables constant.

Penalties, too, are more than just numbers on a sheet. The slight increase in predicted winning percentage per penalty is counter intuitive, but we decided to include it anyways. It is counter intuitive to what you would think, because you would think that the more penalties a team has, the worse, because penalties are bad. However, our model shows that for a 1-penalty increase a team has in a season, the predicted winning percentage is expected to increase about 0.002, holding all other variables constant.

Conclusion

The goal of this project was to aim at any relationships between winning percentage and variables other than yards, points, and touchdowns scored. The relationship that we found is between winning percentage and average turnover margin, penalties, and rank of the team's conference. The multiple linear model that we formed does a good job of providing insight into a team's performance.

The model revealed that turnover margin significantly influences a team's winning percentage, with a 1-turnover increase leading to a predicted increase of approximately 0.183 in winning percentage. Conference ranking also played a notable role, where a 1-rank decrease was associated with a predicted decrease of about 0.015 in winning percentage. Surprisingly, an increase in penalties was linked to a slight increase in the predicted winning percentage, though the effect was deemed basically negligible.

The linearity, independence, normality, and equality of variance conditions were examined by us thoroughly and found to be satisfied. This ensures the strength of our model's predictions.

The applications of this model move beyond this analysis, as college coaches or programs can use a tool like this to maximize their team's potential. The model can also be utilized to measure a team's performance relative to expectations, providing a measure of success.

Overall, this report contributes to the understanding of the factors that influence the success of a college football team and provides an insightful model that could be implemented in the world of sports data analysis.