SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

College Football

Can the Public Predict Games Correctly?

#SASGF

College Football: Can the Public Predict Games Correctly?

Matthew B. Collins and Taylor K. Larkin

The University of Alabama

Introduction

Thanks to advances in technologies that make data more readily available, sports analytics is an increasingly popular topic. A majority of sports analysts use advanced statistics and metrics to achieve their goal, whether it be prediction or explanation. Last year's highly anticipated NCAA College Football Championship game between Ohio State and Oregon broke ESPN and cable television records with an astounding 33.4 million viewers (College Football, March 2016). Given the popularity of college football, especially now with the inclusion of the new playoff system, people seem to be paying more attention to the game than ever. The ESPN online game, College Pick'em, gives fans an opportunity to compete with their friends and colleagues on a weekly basis to see who can correctly pick the winners of college football games. Each week, ten close matchups are selected, and users must choose which teams they believe will win their games and rank those picks on a scale from 1 (lowest) to 10 (highest), according to the user's confidence level. For each team, the percentage of participants who picked that team and the national average confidence level are shown. Ideally, one could use these variables in conjunction with other information to enhance one's own predictions. The analysis described in this work explores the relationship between public opinion data from College Pick'em and the corresponding game outcomes via visualizations and logistic regression using SAS® Enterprise Guide and Base SAS®.

ESPN College Pick'em

Based upon the information presented in College Pick'em (shown in figure 1), we extracted four pieces of key information to use in trying to predict the outcome of games (variable name in parentheses) based on the last two years of data (2014-2015):

1. Home or away *(Home)*
2. Percentage of participants who picked each team *(Picked)*
3. Average confidence level for a given team *(Confidence)*
4. Whether or not a team is ranked *(Ranked)*

We also collected the betting lines for each of the games at the time of kickoff and added that information as a variable called *Line.*

TIME	STATUS	TEAMS		% PICKED	PICK	CONFIDENCE
Sat, Nov 28 7:30PM FOX	Final	#6 Notre Dame Fighting Irish (10-1) at	36	6.1	✗	1
		#9 Stanford Cardinal (9-2)	38	5.9		
Sat, Nov 28 3:30PM CBSSN	Final	Brigham Young Cougars (9-3) at	51	4.5	✓	2
		Utah State Aggies (6-6)	28	2.6		
Sat, Nov 28 7:30PM ESPN	Final	#13 Florida State Seminoles (10-2) at	27	5.6	✓	3
		#12 Florida Gators (10-2)	2	5.4		
Sat, Nov 28 7:15PM ESPN 2	Final	#18 Ole Miss Rebels (9-3) at	38	4.9	✓	4
		#21 Mississippi State Bulldogs (8-4)	27	4.6		
Sat, Nov 28 3:30PM ESPN 2	Final	#22 UCLA Bruins (8-4) at	21	4.3		5
		USC Trojans (8-4)	40	4.1	✓	
Sat, Nov 28 12:00PM HD	Final	#8 Ohio State Buckeyes (11-1) at	42	2.9		6
		#10 Michigan Wolverines (9-3)	13	2.3	✗	
Sat, Nov 28 8:00PM HD	Final	#3 Oklahoma Sooners (11-1) at	58	7.1	✓	7
		#11 Oklahoma State Cowboys (10-2)	23	6.0		
Sat, Nov 28 3:30PM ESPN	Final	Penn State Nittany Lions (7-5) at	16	3.1		8
		#5 Michigan State Spartans (11-1)	55	6.8	✓	
Sat, Nov 28 7:30PM SECN	Final	Texas A&M Aggies (8-4) at	7	6.2		9
		LSU Tigers (8-3)	19	7.1	✓	
Sat, Nov 28 3:30PM CBS	Final	#2 Alabama Crimson Tide (11-1) at	29	7.7	✓	10
		Auburn Tigers (6-6)	13	3.6		

Figure 1: Screenshot of College Pick'em (ESPN College Pick'em, March 2016).

Visualizations

Matthew B. Collins and Taylor K. Larkin

The University of Alabama

Below are some descriptive visualizations created in SAS® Enterprise Guide 6.1.

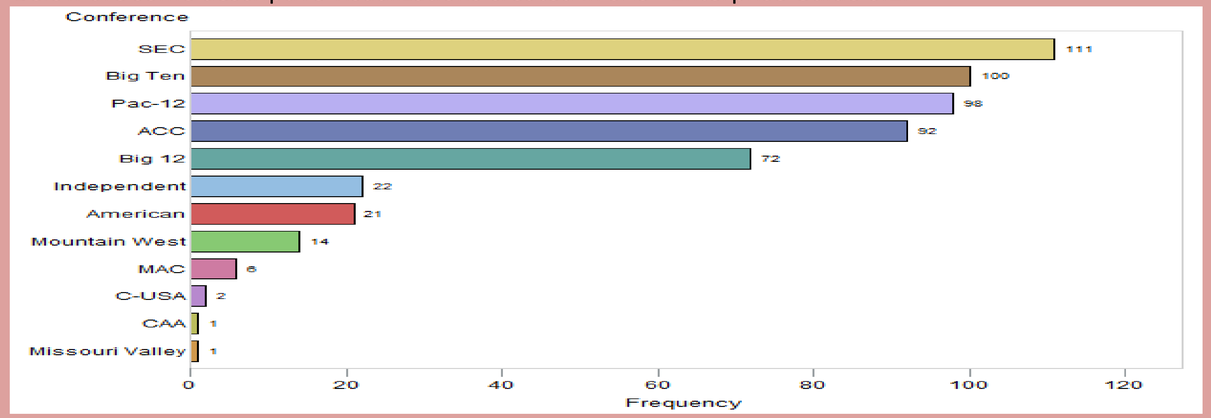

Figure 2: Number of appearances by conference

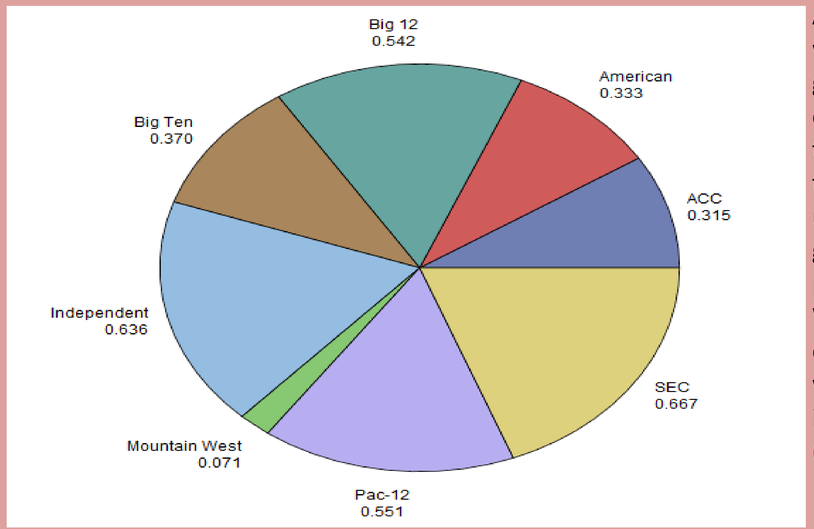

Figure 3: Percent of time teams from each conference were ranked

As shown in figure 2, SEC teams were chosen for College Pick'em games more often than any other conference. This may be due to the fact that more SEC teams tend to be ranked, or that SEC teams usually have a high following, given the low number of professional teams in this region. We also looked at the percentage of time teams for each conference were ranked when they were included in College Pick'em (shown in figure 3).

Figure 3 indicates that SEC teams also led the way in percent of times that their teams were included in College Pick'em while being ranked, at 66.7%. This coincides with our theory from the last graph about a potential reason SEC teams were selected more often than any other conference. Another interesting note from this graph is that the Independents were the second highest percent in teams being ranked. However, this is certainly explained by the fact that Notre Dame is usually ranked and featured in College Pick'em, as well as the small number of teams included from the Independent conference in this analysis (three teams in this dataset).

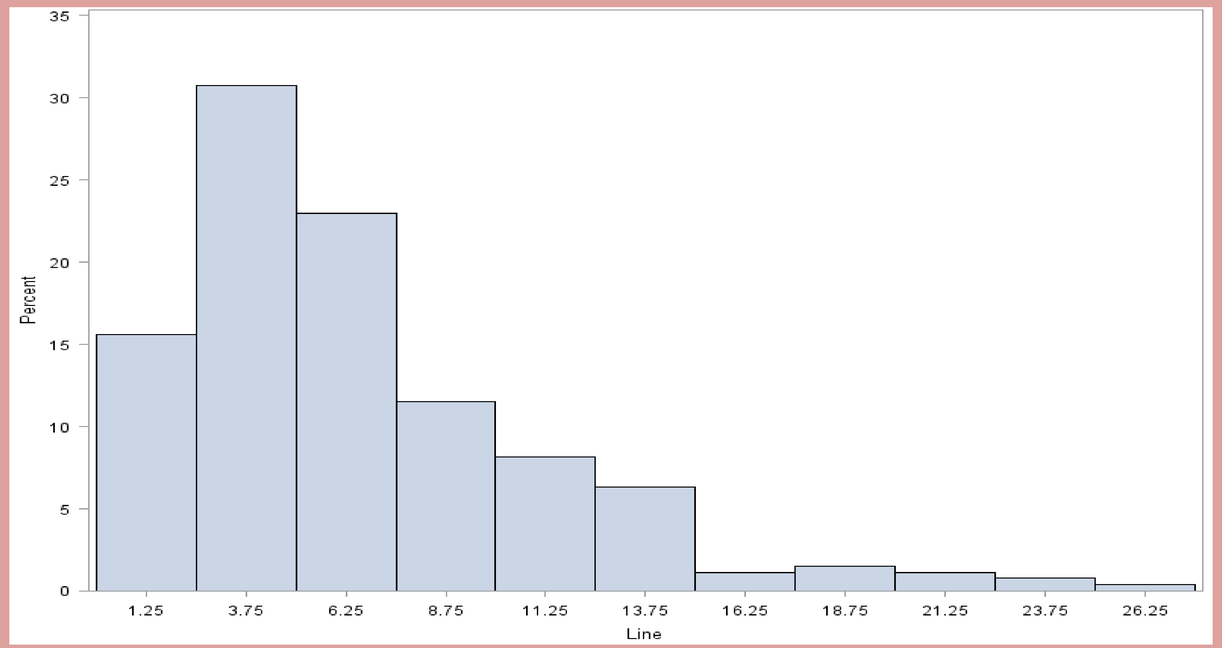

Figure 4: Percent of time teams from each conference were ranked

Figure 4 above shows the distribution of *Lines*. 93.4% of the games selected ended with a line of 11 points or less, and 68% of the games had a line of less than a touchdown, so these are typically close games that are difficult to consistently choose the winner.

Are the College Pick'em Variables Useful?

Matthew B. Collins and Taylor K. Larkin

The University of Alabama

Probability modeled is Win = 1

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	−0.8556	0.3160	7.3311	0.0068
Picked		1	1.9013	0.3488	29.7188	<.0001
Confidence		1	−0.0194	0.0681	0.0808	0.7762
Ranked	0	1	0.00533	0.0999	0.0029	0.9574
Home	0	1	0.1516	0.0928	2.6660	0.1025

adding *Line* into the model

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	−0.2168	0.3910	0.3073	0.5793
Picked		1	0.7340	0.5428	1.8280	0.1764
Confidence		1	−0.0304	0.0687	0.1960	0.6580
Ranked	0	1	−0.00743	0.1005	0.0055	0.9411
Home	0	1	0.2573	0.1016	6.4175	0.0113
Line		1	−0.0653	0.0239	7.4741	0.0063

Figure 5: Results from each logistic regression model. Null hypothesis for the Hosmer-Lemeshow Test is not rejected for either model which indicates a good fit to the data. Neutral site games are omitted.

Statistical Analysis

To study the effects of the College Pick'em variables, logistic regression models are built in Base SAS® 9.4 using the PROC Logistic command. Figure 5 displays the initial model with the four main variables. According to the p-values, *Picked* is the only significant variable at the conventional 0.05 significance level. Using only this first model, one may conclude that *Picked* is the only useful piece of information from College Pick'em. However, once the variable denoting the betting line of each team at game-time is included, *Picked* becomes insignificant, while *Home* and *Line* are driven to be statistically significant in the presence of the other predictors. Two main finding can be drawn:

1. Including information about the betting line neutralizes the effect of public perception on who they believe will win the game.
2. The coefficient for *Home* is positive, meaning that teams are less likely to win when they play at home. Specifically, according to the odds ratio in figure 6, the odds of winning increase by 67.3% when teams are not at home.

Upon further inspection, the former point makes sense since the correlation between *Picked* and *Line* is approximately -0.82. The public is sure to use betting lines to make their own prediction as to who will win games. The latter point, which may be an even greater finding, is that teams playing at home are less likely to win. According to this dataset, teams who were at home only won about 47% of the time. Though this seems counter-intuitive, the fact that ESPN chooses difficult match-ups each week makes this conclusion naturally feasible.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Home 0 vs 1	1.673	1.124	2.491
Line	0.937	0.894	0.982

Figure 6: Odds ratio estimates for the statistically significant variables in the latter regression model from figure 5.

Concluding Remarks

Matthew B. Collins and Taylor K. Larkin

The University of Alabama

Discussion

The SEC was the most featured conference in the College Pick'em games, according to this data. This is likely due to the fact that the SEC had more ranked teams represented in College Pick'em than any other conference. Additionally, an overwhelming percent of these games selected had lines under ten points, emphasizing the small disparity between the teams. On the statistical modeling side, we only studied four variables which could easily be derived from the main College Pick'em page as well as the betting lines. Such Information as which conference each team plays and the order of their rankings are not considered. These could be potentially useful variables for predicting outcomes. Future work consists of estimating predictive performance with a variety of different statistical models and machine learning algorithms. In addition, the findings about the insignificance of *Picked* and *Confidence* as well as the interpretation of *Home* reinforce the difficulty in predicting outcomes in sports, especially in the college realm. Many other factors, including intangible traits such as emotion, play a role in deciding the outcomes of games. Ultimately, for this dataset, little evidence supports that the public's opinion is a useful predictor for modeling outcomes in the presence of the other variables.

Conclusion

Given the popularity of college football and the play-off system (see figure 7), people have become more invested in college football. As a consequence, online games such College Pick'em, are becoming more popular. This game in particular provides an interesting avenue into analyzing data about the public's perception on who they think will win close match-ups. This work focuses on visualizing the data gleaned from College Pick'em using SAS® Enterprise Guide and the statistical relevance of each variable in Base SAS®. Moving forward, one could expand the analysis from this type of data to make their own predictions about which teams they believe will win using the power of SAS®.

References

1. College Football Championship Sets A New Cable Ratings Record. (n.d.). Retrieved March 14, 2016, from http://www.npr.org/sections/thetwo-way/2015/01/13/377060626/college-football-championship-sets-a-new-cable-ratings-record.
2. ESPN College Pick'em. (n.d.). Retrieved March 14, 2016, from http://games.espn.go.com/college-football-pickem/2015/en/?addata=promo_cpickem_2015.
3. [Rique]. (2016, March 12). *Kenyan Drake Returns Kickoff for 95 Yard Touchdown*. [Video File]. Retrieved from https://youtu.be/luDv4EE-Mtg.

Figure 7: Video of Kenyan Drake returning a kick-off late in the 4th quarter of the National Championship game.

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF