COLLEGE FOOTBALL PREDICTIVE ANALYTICS

Using data and analytics to predict outcomes of college football games and to create an unbiased ranking system to determine the best teams.

Wyatt Rasmussen

Bellevue University, Master's of Data Science | DSC 630 Predictive Analytics

Executive Summary:

In this project I aimed to create an unbiased ranking system and begin to predict the outcome of the games that were being played. I created a ranking system that allowed us to get the top 4 teams in each season that rewards teams for playing tough competition and for winning by larger margins of victory. This will help increase the competitiveness of the regular season of football by encouraging tough competition.

I also created a system that would predict the end margin of a football game by taking into account the rank I gave each team and the conference that each team belonged to. By doing this we can get a simple, yet effective predictive system that I believe was effective in predicting the margin. This process of predicting the games proved to be really difficult and would likely require a lot more research and a lot more factors to allow us to predict more accurately.

Lastly I created a system that would predict the outright winner of each game. This system would take into account the pregame ranking of each team and return a guess of who won the game. After working on this, I was able to create a system that would accurately predict the outright winners of games in about 70% of games that are played. This would be considered a rather sizeable success due to the use of just the pregame ranking. This number would likely improve with more factors such as weather, injuries, etc.

Overall I think this process was very beneficial and could provide excellent results if implemented into the college football playoff.

Abstract:

College football is a hot topic every Fall for a number of fan bases around the country. There are currently 130 teams in FBS (Football Bowl Subdivision) that are competing on a yearly basis. Teams play their games and only the top four teams are selected for the playoff at the end of the season. Currently those four teams are selected by a committee of 13 athletic directors, former coaches, and former players. This committee is on rotation, so those 13 members won't be the same people year over year. Since this committee inherently has a rooting bias, having been involved with the teams that they are ranking, this often comes with a lot of controversy as to who the top teams are. Top 25 rankings are released throughout the season in order to give the teams and fans an understanding of where your team is ranked.

In this paper I will run through the process of creating the unbiased Elo ranking system as well as creating multiple models that will help us predict the margins of the games and ultimately predict the winner of the games as well. This can be used to predict the outcome of two teams in the future playoff discussion.

Introduction:

College football, inevitably, has many teams that are up in arms on a yearly basis due to the bias of the committee and the results of the playoff selection. By taking the human decision element out of the process, we can hopefully remove some of the rooting bias that the committee has in order to create a more balanced and results based ranking system.

In this project I will create a unbiased Elo based ranking system that will determine the top 25 and teams in an order based on the results of the games that they

will play. With those Elo ranking that is applied to each team, I will then create a few different machine learning models that will allow me to determine the predicted margin outcome of a game and also to predict who the winner and loser of a given game will be. This will allow college football to use an unbiased system to determine the winner of a game.

Methods:

The data that I am using for this project is supplied at collegefootballdata.com. There is a wide variety of different statistics to choose from including, but not limited to game by game data, play by play data, game results, etc. From collegefootballdata.com I am able to use an API that they have set up with my unique API key in order to begin pulling the data in that manner. Using this API I set up a call that would get the results of football games that were played in the time frame that I specified. For the purpose of this project, I pulled in every football game that was played between 2017 and 2021. This gives 5 seasons of data to work with and build off of. This API returns a lot of data, but I have chosen to limit it to a few features that seem of the most importance for my purposes. Those features are as follows: game_id, start_date, season, home_team, home_conference, home_points, away_team, away_conference, and away_points.

This data will be used to create the ranking system that I am creating. Using the data, I am able to create an Elo based ranking system. An Elo based ranking system is a method for calculating the relative skill levels of players in zero-sum games. This system was created by a Hungarian-American physics professor, Arpad Elo. (Mittal, 2020). The formula for Elo rankings is:

$$R'_A = R_A + K(S_A - E_A).$$

- R'_A is the new rating for team A.
- R_A is the current, pregame rating for team A.
- K is the K value, which we'll be setting to the absolute value of margin of victory.
- S_A is the actual score for team A.
- E_A is the expected score for team A. (Bill, 2021)

Using the Elo ranking system will allow us to look at the results of the games and unbiasedly select the top teams. Teams will be rewarded for winning by a larger margin and docked for losing by a large margin. Teams will also be rewarded for beating better teams. This causes games between good teams to really push the rankings towards the end of the season and thus encourage teams to play tougher strength of schedule as beating all bad teams wouldn't result in a top 4 finish.

Once we have created a ranking system and applied a beginning Elo ranking to each team and adjust the Elo after each game they play we can begin to do a little machine learning for predicting the outcomes of the actual games. We will do this by using the 2017-2020 seasons to train our model and use the 2021 season to test. That gives over 3000 data points for the training set. There are two key ways to accomplish this, predicting the margin in a game and simply determining the winner of each game.

Predicting the margin of each game will be difficult as getting the exact correct margin will likely require many more features than we are currently working with. This could be factors such as weather, matchups, injuries, etc. To do this we are training our model to predict the target variable of the margin using a wide variety of different features to find a model that works for us.

Predicting the outright winner of each game will prove to be a little easier than the margin. This is because we are predicting one team to win and another to lose. The primary machine learning algorithm to use for this will be Logistic Regression since that is generally used to predict things that have 2 outcomes.

Results:

Creating an Elo System

To begin creating an Elo system I first had to create a few functions that would determine the Elo rating of the two teams playing a game. The first function created is called get_expected_score, this function would bring in the rating of the two teams prior to the game and give an expected win percentage using a logistic curve function. Two teams with the same rating would return .50000.

The next function that I created was called get_new_elos. This function would bring in the home_rating, away_rating, and margin of the game. This would give us a new elo rating for each team after the game was completed. The calculation for this was the equation mentioned in the methods. Losing teams rankings went down and winning teams rankings went up. The reason for including the margin in this calculation was to reward teams for having a large win or keeping it close against a team that was much better than them in the Elo ranking prior to the game.

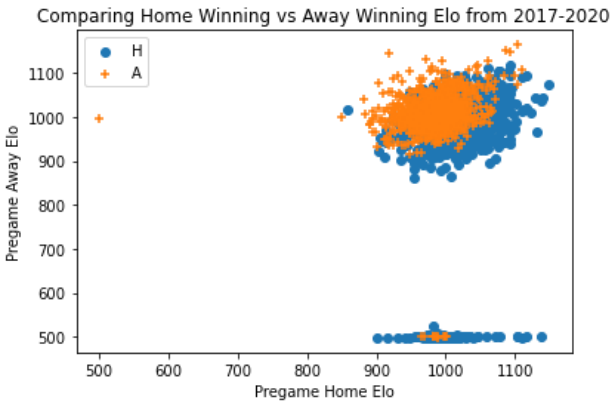
Since I wanted to create an unbiased system, I decided that each year should calculate separate of every other year. By doing each season individually every team started on a level playing field to eliminate bias. Each FBS team started with an Elo rating of 1000 and FCS teams (the lower division) would start at 500. Then I loop

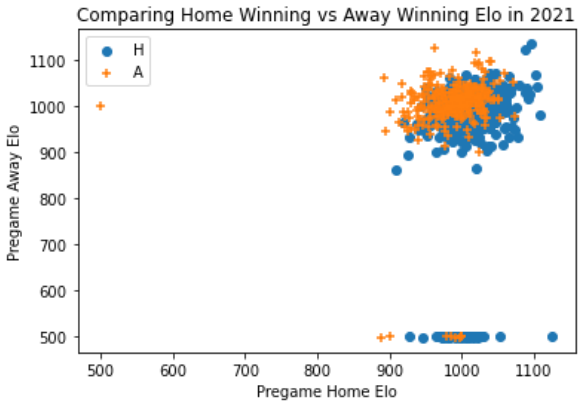
through every game and attach a new Elo based on the result of the game. The 2017-2021 season top 25 rankings ended up looking like so:

2017		2018		2019		2020		2021	
1118	Penn State	**1150**	Alabama	**1169**	Ohio State	**1129**	Alabama	**1125**	Georgia
1110	Alabama	**1136**	Clemson	**1154**	Clemson	**1107**	Clemson	**1114**	Ohio State
1109	Ohio State	**1097**	UCF	**1123**	LSU	**1106**	BYU	**1112**	Michigan
1107	Oklahoma	**1097**	Ohio State	**1109**	Alabama	**1080**	Coastal Carolina	**1109**	Cincinnati
1106	Wisconsin	**1094**	Oklahoma	**1102**	Oregon	**1077**	Cincinnati	**1106**	Alabama
1105	Georgia	**1091**	Georgia	**1098**	Wisconsin	**1073**	Notre Dame	**1083**	Notre Dame
1104	Clemson	**1091**	Michigan	**1092**	Utah	**1063**	Oklahoma	**1081**	Pittsburgh
1096	UCF	**1086**	Notre Dame	**1086**	Notre Dame	**1060**	Florida	**1079**	Coastal Carolina
1088	Auburn	**1084**	Utah State	**1084**	Appalachian State	**1060**	Buffalo	**1078**	Utah
1087	Washington	**1081**	Fresno State	**1080**	Oklahoma	**1059**	North Carolina	**1069**	Louisiana
1073	Oklahoma State	**1075**	Penn State	**1079**	Georgia	**1058**	Ohio State	**1069**	Oklahoma State
1068	Florida Atlantic	**1072**	Appalachian State	**1078**	Memphis	**1057**	Iowa	**1066**	Western Kentucky
1064	Stanford	**1071**	Boise State	**1076**	Penn State	**1056**	Iowa State	**1062**	Appalachian State
1064	Notre Dame	**1067**	Ohio	**1076**	Michigan	**1052**	Georgia	**1062**	Houston
1063	South Florida	**1064**	Washington State	**1072**	UCF	**1051**	Liberty	**1057**	NC State
1063	Memphis	**1064**	Mississippi State	**1072**	Louisiana	**1050**	Texas	**1057**	Iowa State
1060	Virginia Tech	**1063**	Cincinnati	**1071**	Boise State	**1050**	UCF	**1056**	Baylor
1056	TCU	**1062**	Iowa	**1071**	Baylor	**1048**	Appalachian State	**1054**	Wisconsin
1055	Toledo	**1061**	West Virginia	**1062**	Auburn	**1047**	Louisiana	**1053**	Kentucky
1055	Northwestern	**1058**	North Texas	**1061**	Florida	**1044**	San José State	**1051**	Boise State
1051	Boise State	**1054**	NC State	**1060**	Florida Atlantic	**1043**	Texas A&M	**1051**	Air Force
1051	USC	**1053**	Missouri	**1060**	Navy	**1042**	Tulane	**1051**	UT San Antonio
1047	Iowa	**1051**	Temple	**1059**	Air Force	**1039**	Marshall	**1051**	Texas A&M
1046	Miami	**1050**	Washington	**1056**	SMU	**1035**	Indiana	**1050**	UCLA
1046	Troy	**1048**	Army	**1055**	Minnesota	**1032**	Arizona State	**1047**	Wake Forest

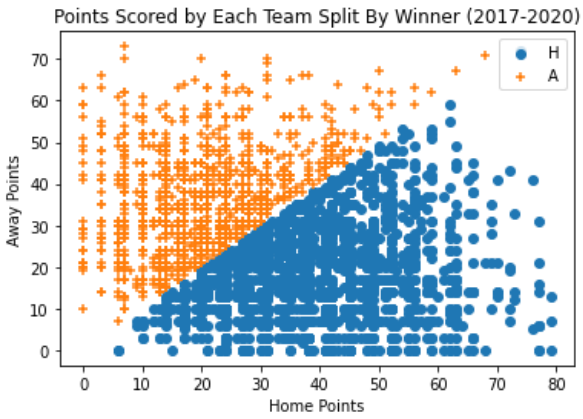
Overall, these rankings look pretty good.

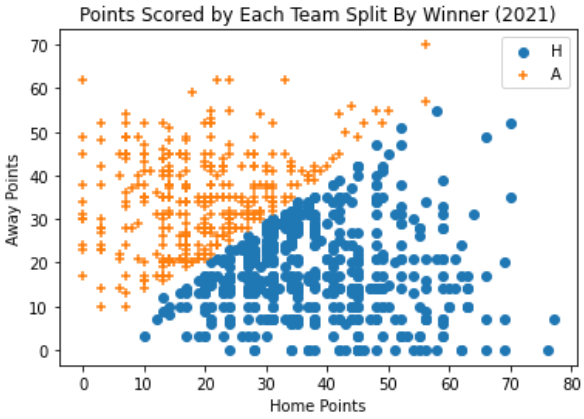
Once I completed the process of making rankings for each of the seasons, the next step was to begin to visualize this. I created a few data visualizations that would show the matchups that we saw in games. I was also able to split a few of them into the winning teams, orange as the Away team winning, and blue as the Home team winning. Those charts are below.



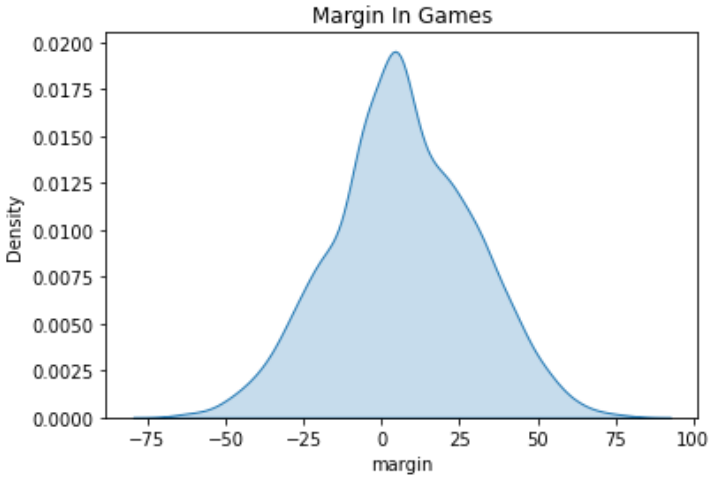


These two charts show the pregame Elo of the teams and the colors correlate with the team that won. It seems in a lot of the games between higher ranked teams, the home team came away with the victory.





These two charts show the points scored in the games. I found it interesting that it seems to have a pretty heavy concentration of data towards the middle of the chart. I also found it interesting that home teams tend to score more with the x-axis going beyond 80 and y-axis only going up to 70.



The last visualization to see is the margin in the games. Negative numbers would mean the away team winning, and positive numbers would mean the home team winning. This ended up forming a pretty normal distribution.

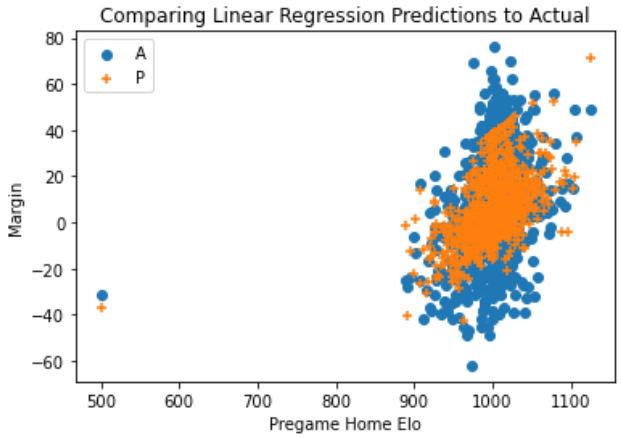
Building a Prediction Model

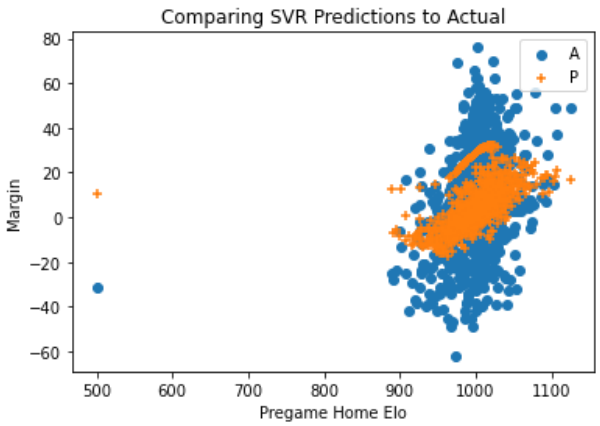
The next step is to create a prediction model based on the results of the games. To do this I start with training my model from the 2017 – 2020 seasons. As I described in the methods section, there are two different ways to do this. First, I will cover predicting the margin of the games.

Since this problem primary deals with numbers the most logical approach to this would be to use a regression-based model to predict the final margin. I tried a few different regression-based models such as Linear Regression, Random Forest Regressor, and Support Vector Regressor. Along with using these different regressions I also tested using different features to predict with. In all of the models I used pregame Elo of each team. To test features I also began using the conference that each team was affiliated too. Below is a table of the R^2 Scores that I got for each model.

MODEL	LESS VARIABLES	MORE VARIABLES
LINEAR REGRESSION	.2686	.3826
RANDOM FOREST REGRESSOR	.1401	.2653
SUPPORT VECTOR REGRESSOR	.3319	.2796

As you can see the Linear Regression that used more variables was the most successful of these algorithms, although the support vector regressor with less variables gave us a pretty good score as well.





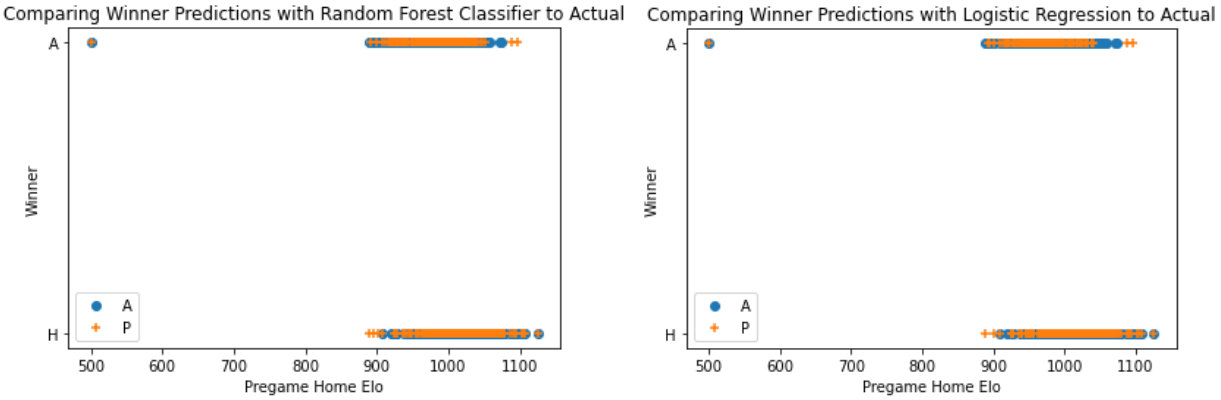
They each accomplished this in different manners. The linear regression was fairly accurate on a lot of predictions, but has a few that are way out there. The SVR predictions tended to be a lot more conservative in comparison. This likely lead to a lot more incorrect guesses, but generally pretty accurate.

Next, I tried to predict the winner of the games. This process was a bit easier since I am not trying to accurately guess the margin. For this kind of model it makes sense to use Logistic Regression, Random Forest Classifier, and Decision Tree

Classifier. Below is a chart of the accuracy scores of these classifiers with the same process of less and more variables included.

MODEL	LESS VARIABLES	MORE VARIABLES
LOGISTIC REGRESSION	.6961	.7079
RANDOM FOREST CLASSIFIER	.6867	.6867
DECISION TREE CLASSIFIER	.6408	.6396

As you can see the model tended to perform a lot better which was to be expected. There really wasn't much different in using more variables here, so I would likely recommend using less variables for this kind of prediction. Below are the charts of the logistic regression and random forest classifier with less variables.



Conclusion:

The goal of this project was to create an unbiased ranking system and begin to determine what teams would win when playing in a game. I believe the usage of the Elo ranking system allowed us to accurately get the top 4 teams in the sport. By using only the results of the game, including team Elo ranking margin we were able to get the top teams to rise to the top. I believe I was also really successful in determining the

outcomes of games. Using a predictive model to accurately get the margin in the game proved to be rather difficult, but we saw that using more features was able to help us create a better model. We tended to see the inverse happen when we were trying to predict the outright winner of the game. The model with less features was just as accurate and lead to a better model since it was using less features.

Acknowledgments:

I would like to acknowledge Deependra Dhakal for taking the time to peer review my code and help me along the way. I would also like to acknowledge Professor Fadi Alsaleem for overseeing the project and sharing knowledge.

References:

Bill. (2021, October 7). Talking tech: Calculating Elo ratings for college football. CFBD Blog. Retrieved February 6, 2022, from https://blog.collegefootballdata.com/talking-tech-elo-ratings/

Mittal, R. (2020, November 6). What is an ELO rating? Medium. Retrieved February 6, 2022, from https://medium.com/purple-theory/what-is-elo-rating-c4eb7a9061e0

Overview. College Football Playoff. (n.d.). Retrieved March 3, 2022, from https://collegefootballplayoff.com/sports/2016/9/30/overview.aspx

NCAA.com. (2022, January 11). *College football championship history*. NCAA.com. Retrieved March 3, 2022, from https://www.ncaa.com/news/football/article/college-football-national-championship-history