Predicting College Football Gambling Lines Using 247 Composite Recruiting Data, Conference Data, and Rolling Averages of Basic Football Stats

Ted Henson

Abby Williams

Brett Shanklin

Alex Matthew

Samantha Ferraro

Data information

The first dataset gathered was 247 composite recruiting data [1]. We decided to pursue this data first as the teams with the top recruiting classes are usually the best teams. Data was copy and pasted from 247 composite rankings for all relevant seasons. For every game, we considered data on the prior 5 recruiting classes for each team. Any team names that were spelled differently (UCF versus Central Florida) were substituted through brute force. All teams without data were double checked after this process to ensure the missing data was not due to non-missing team names. We included the number of recruits for each class, as we would expect that teams with more upperclassmen to perform better. Total recruiting points and average points per recruit was also included for each class, as the former probably favors teams that have more depth while the latter favors teams that are top heavy. If a team did not have any recruiting data, then they were given 0s in those columns, indicating they did not have any ranked commits. We did not throw those rows out as that would penalize teams with one ranked commit per season.

Conference data was gathered through copying and pasting from the NCAA standings website [2]. Brute force substitution was also applied to the team names, as well as double checking. If a team moved conferences, we used their current conference as that is a more accurate depiction of the team's quality and strength of schedule going forward. The complexity added by engineering prior conference memberships would not be worth the improvement in prediction. We included both the conference as a variable, and the interaction between the conference and their division as categorical variables. We included this as many divisions are lopsided (the SEC West has been much stronger than the East in recent years) and teams within the same division play each other every year. These variables were included to try to regress all stats towards their conference mean. We also created two binary variables: one indicating whether it was a conference game, and/or an inner division game to account for the role familiarity plays in who wins, and by how much. Understanding the opposing team's scheme and players can expose team's weaknesses and generate different results than expected.

Rolling averages from the GameStats dataset were engineered based on all prior games for that team within a season. Many opposing teams did not have more than 1 game of both offensive and defensive game stats to create a prior belief about their value. Therefore, if an opposing team did not have 4 games of prior defensive stats (as this was the limiting factor), those rows were thrown out. As a result, all first weeks of the seasons were thrown out. We justified this as we will be predicting on games very late in the year. Games early in the year, especially the first game, are much more influenced by the prior years performance. Most of these teams were FCS level teams, which we will not be predicting on anyway, so we felt comfortable throwing them out based on the prediction task. For the two teams that do have FCS games, we did not have adequate stats for 2019 on them. Therefore, offensive and defensive rolling averages were used from prior seasons as estimates. The following stats were included for both teams, from the offensive and defensive point of view: yards per attempt, rushing yards per attempt, passing yards per attempt, completion percentage, total turnovers

and turnovers per play, interceptions per passing attempt, total points, points per play, touchdowns per play, rushing TDS, passing TDs, and winning percentage. We included both total yards per attempt and broken down into passing and rushing as the former represents overall success while the latter gives a more comprehensive look at the team's strengths. Completion percentage was included as the interaction between that and passing yards per attempt could be significant. Total turnovers and turnovers per play were both included. Some teams may turn the ball over a lot, but play at a faster pace, and we tried to estimate that based on number of plays. We included both as our estimation of the number of possessions is tenuous. Total points was included to capture their overall strength, and points per play was created to measure explosiveness. Rushing TDs and Passing TDs were included as passing ability may be more predictive of future success than rushing, as it has been shown to be in the NFL. Touchdowns per play was also included to measure a different type of explosiveness (how often you reach the end zone per play). The prior winning percentage in that season was also included as another measure of overall strength.

Other non-team specific information was included such as the Game Type, as some post season games such as bowl games tend to be blowouts. The game number for each team was also included as teams with greater depth, measured by recruiting points, may be better late in the year; however, this caused worse out of sample prediction, and caused issues with joining our predictions dataset, so we elected to leave it out. The game location was also included as there is no home field advantage effect at neutral locations.

For our game stats and the recruiting data, we took the difference in stats between the home and away team. This was done primarily to reduce computation and ease interpretability. There is no significant reason to believe that a given statistic (such as points) is more important for the home team than the away team. Any difference found between home and away stats would either be due to random chance or be marginally significant. This transformation reduced our errors slightly, eased interpretability, and reduced computational cost. We also felt more confident in the predictions as they are less complex and the variables chosen make more sense.

After seeing some of the significant variables for the spread, we engineered a few lag variables from the prior 4 games to give an index of a team's performance recently for our final predictions on ACC games. We considered the prior 4 games for offensive and defensive points per play, and winning percentage. In order to control for some teams having poor schedules early in the year, we further subset our data to only train on games October and later. After retraining our data, the metrics for our games were more or less the same. We realized the rolling averages for prior games will be identical to the moving averages for half of our training games. Therefore, we left it out: the added complexity was not worth it. We instead focused on more advanced modeling techniques to reduce our errors, rather than introducing more covariates. We only trained on games after September, as that will be more representative of our test set, and it did reduce our test errors.

Methodology for Spread

We performed similar processes for predicting all of our response variables; however, knowing our relationships appear to be primarily linearly related based on our outcomes on predicting the result, we trained mostly linear models for the spread. We added an Earth model to see if there were any important interaction or higher order terms. We trained our models both on normalized and unnormalized data. The normalization did not improve predictions, even on methods such as neural nets which typically benefit from it, so we elected to not normalize our data. Other models trained include a bagged Earth model, simple linear regression, stepwise regression, extreme gradient boosting, elastic net, Bayesian additive regression trees, and others. The ones shown below were the best, and are relatively simple.

At first, we trained our models using 5 fold cross validation; however, bootstrap random sampling reduced our minimum RMSEs and reduced our errors across our test samples. Below we provide a summary of the RMSE across our 1000 bootstrap samples for our best models: partial least squares, earth (multivariate adaptive regression splines), lasso regression, and ridge regression. Although the model with the corresponding minimum RMSE is used to select the optimal hyperparameters, we will display summary statistics of the out of sample RMSEs across all our bootstrapped samples for good measure.

RMSE	Min	1st Q	Median	Mean	3rd Q	Max	NAs
pls	16.43975	17.29781	17.61071	17.60081	17.91624	18.88724	0
earth	16.45214	17.49713	17.83777	17.87083	18.16739	49.88780	0
lasso	16.20758	17.12313	17.42886	17.44672	17.74571	19.31604	0
ridge	16.33358	17.39467	17.69157	17.68750	17.98259	19.06807	0

All of our models had performed similarly, with a minimum out of sample RMSE of between 16.2 to about 16.45. Although the Earth model did not perform the best, looking at the inputs included in the final pruned model could yield some insight, as it tends to produce sparse models. Below are the variables used in the final model ranked by the reduction in out of sample cross validation error (scaled 0 to 100):

Variable	Reduction in Out of Sample Error
Dif.pts.def.roll.avg	100.00
Dif.Recruit.Mean.Rating.Freshman	74.274
Dif.pts.off.roll.avg	55.301
Dif.rush.ypa.def.roll.avg	17.124
Dif.pass.ypa.off.roll.avg	12.809
Dif.Recruit.Num.Junior	7.333

Surprisingly, the earth model found that the difference in the average rating of freshman players was highly predictive. This seems counterintuitive since seniors and juniors usually make up the bulk of the players, but perhaps after accounting for the overall quality of the teams in terms of points and yards, freshman talent can start to come on late in the year and play a role in blowout game spreads. Additionally, freshmen may play more later in the season as injuries begin to reshape rosters. Transfers have also started to run rampant in college sports, so perhaps after a few years, most of the players in a given class have moved onto other programs. Notably, the model does include the difference in the number of junior commits as a variable so our conventional wisdom that upperclassmen are more important is captured to an extent.

Our lasso model choose similar, but many more variables to include. Our conference variables were not included in the Earth model and were not highly significant for the lasso. Conventional wisdom would think that a bad team in a power five conference would be better than a similar team statistically than a group of 5 team. We even engineered a binary variable indicating if the team was in the power 5 or not, and this was not highly significant either. Although this analysis has been quite basic, preliminary results show that a team's conference is not as important for predicting their performance compared to other variables. Whether or not a team is FCS or not; however, does appear to be significant, as a wildly inaccurate prediction on our test set was generated because it thought Miami was an FCS team. Therefore, we kept the conference variables in the models, knowing that it may be significant, just less predictive than other information.

In addition to the models presented in the table, we trained a stacked model, which was a linear regression on the predictions of the other four on 100 bootstrapped random samples. We discussed using different weighting schemes to average our predictions, but this was the more conservative approach as methods such as greedy optimization could lead to overfitting. When we previously instituted a simple train and test split, the stacked model performed the best, so we will use that for our final predictions, and it will also be the most conservative approach.

Although we will mostly relied on our stacked model for predicting the spread, we made some manual changes to our predictions as we did not collect statistics or results from games since the final week in the GameResults dataset. We decided not to collect recent stats as it is very late in the season, so our predictor variables would not be highly influenced by these

outcomes anyway. It also might be computationally expensive. We would account for significant recent outcomes through manual changing of spreads. The main teams that have had been particularly horrible and since our GameResults dataset have been NC State and Syracuse (they've both lost 3 in a row, sometimes badly). Miami and Virginia Tech have also been pretty good recently, the former winning 3 straight and the latter hanging with Notre Dame and beating Wake by 19 points. In addition to not incorporating recent weeks, our model treats the spread as a continuous response variable when it is actually discrete. We used manual changing to account for this as well. ESPN FPI [3] was used heavily in addition to recent results for both changing and rounding the spread to a common football spreads. When in doubt, we rounded to the nearest common football spread (such as 7 or 10). The only large changes made were in extreme cases that we discussed above, such as where the team has lost by 3 touchdowns or more in 3 straight weeks (NC State) or when a team was given too much credit for being in a good conference (Liberty was liked by our model because of being an independent). Additionally, for the first weekend in our predictions, we regressed our predictions to the Bovada Vegas line [4].

Methodology for Total Points

For predicting the total points, histograms showed that the response variable might be skewed. Therefore, we performed a log transformation on our response in order to achieve a lower out sample error rate. We ran regressions with and without this transformation to see which achieved a lower out of sample error. This transformation did not help the predictions so we elected to leave it out. Additionally, we reversed the transformation of the variables that we made for the spread and result: instead of taking the difference in stats, we took the sum of the game stats and the recruiting data. We trained models doing both the difference and the sum of our variables, and the sum proved to be more predictive for both types of variables (game stats and recruiting). Below are the summary statistics of the out of sample MAEs across all our 100 bootstrapped samples.

MAE	**Min**	**1st Q**	**Median**	**Mean**	**3rd Q**	**Max**	**NAs**
lasso	13.05144	13.84467	14.04593	14.04704	14.25769	15.06964	0
ridge	13.09662	13.99516	14.20427	14.19776	14.43697	15.23037	0
earth	13.31093	14.01549	14.20062	14.22600	14.47839	15.01988	0
pls	13.21159	13.85927	14.15223	14.10453	14.34602	15.12871	0

We created a stacked model here as well, and chose it for our final total points predictions as it will be the most conservative. When we previously instituted a simple train and test split, it performed the best. As with our spread predictions, we will list the variables used by the Earth model scaled from 0 to 100 by the reduction in cross validation error for predicting total points:

Variable	Reduction in Out of Sample Error
total.pts.def.roll.avg	100.00
total.pts.off.roll.avg	83.11
total.ypa.off.roll.avg	49.36
total.rush.ypa.off.roll.avg	26.45
total.rush.tds.def.roll.avg	21.18
total.rush.tds.off.roll.avg	14.93

Points were more predictive overall than yards per attempt in the Earth model. Although one might think yards would play a greater role in the total points, pace of play varies more so in college than in the NFL. Points can account for both the overall offensive and defensive strength of the team and the pace of play to some extent. Additionally, as in the case for the spread, we rounded the total points to common college football totals [5], and regressed our totals towards the vegas line for the first week. Common point totals were also determined by looking at the distribution of total scores in our dataset. Based on comparing our totals to the vegas totals for week one, we tended to predict higher than the Bovada total, so we kept that in mind for future rounding.

Methodology for Result

For classifying the result, we trained a variety of models through 5-fold cross validation on 1615 games (all after September), and generated our predictions based on the final model outputted by cross validation. Shrinking our training set to only games post September did decrease our accuracy in our folds (as it is inherently easier to predict on FBS versus FCS matchups); however, we anticipate greater performance on our future games as many games early in the year are not conference games, which is what we will primarily be predicting on.

We trained a variety of models, but have only included the four best. Some of the models we trained, but are not shown were single and multiple layer neural nets, simple logistic regression with and without interaction terms, qda, lda, and polynomial and linear support vector machines. The support vector machines predicted well, but they do not support predicting class probabilities, which we will need in order to generate stacked predictions. The best models were an averaged neural network (many networks trained over many random seeds; final prediction is average over all seeds), gradient boosted classification trees (gradient descent is used to prune the decision trees), L1 penalized logistic regression, and a Bayesian additive regression

tree model (similar to the gradient boosting, but substituting a prior for the learning rate). All of them were trained through 5-fold cross validation. We did train these models using bootstrapping as that was superior for our other outcomes; however, in this scenario, cross validation performed better. Below is the table of out of sample accuracy rates for our best models over 5 folds:

Accuracy	Min	1st Q	Median	Mean	3rd Q	Max	NAs
avNNet	0.6780186	0.6996904	0.7120743	0.7083591	0.7151703	0.7368421	0
xgbDART	0.7120743	0.7120743	0.7213622	0.7244582	0.7337461	0.7430341	0
glmnet	0.7027864	0.7058824	0.7213622	0.7207430	0.7337461	0.7399381	0
bartMachine	0.7058824	0.7089783	0.7120743	0.7201238	0.7244582	0.7492260	0

Although the models we trained were complex, they did predict much better than some of the models mentioned prior. 2% is a big difference in classification rate in this scenario. We feel confident that our combined model was a linear combination through 50 bootstrapped samples of 4 other well performing models. The most important variables in these models were similar to our Earth and lasso variables for spread. Most of the variables included in all of these models do seem to be significant to some extent. We will list the most important variables in our regularized logistic regression, although none of them are particularly interesting:

Variable	Reduction in Out of Sample Error
Dif.tds.rate.off.roll.avg	100.00
Dif.pts.rate.off.roll.avg	54.65531
Dif.win.pct.roll	37.59611
Dif.rush.tds.def.roll.avg	12.05079
Dif.ypa.off.roll.avg	9.38194
Dif.rush.ypa.def.roll.avg	3.48807
Dif.Recruit.Mean.Rating.Sophomore	2.62982
Dif.Recruit.Mean.Rating.Freshman	2.49456
Dif.pts.def.roll.avg	1.83742
Dif.pts.off.roll.avg	1.05459
Dif.Recruit.Points.Junior	0.16889
Dif.comp.pct.off.roll.avg	0.11466
Dif.Recruit.Mean.Rating.Junior	0.11431
Dif.Recruit.Num.Junior	0.10330
Dif.Recruit.Points.Freshman	0.01054

References

[1] "2016 Football Team Rankings", *247Sports*, 2019. [Online]. Available: https://247sports.com/Season/2016-Football/CompositeTeamRankings/. [Accessed: 22- Oct- 2019].

[2] "NCAA College Football FBS Standings | NCAA.com", *Ncaa.com*, 2019. [Online]. Available: https://www.ncaa.com/standings/football/fbs. [Accessed: 31- Oct- 2019].

[3] "ESPN Football Power Index - 2019 - ESPN", *ESPN.com*, 2019. [Online]. Available: http://www.espn.com/college-football/statistics/teamratings. [Accessed: 14- Nov- 2019].

[4] *Bovada.lv*, 2019. [Online]. Available: https://www.bovada.lv/sports/football. [Accessed: 14- Nov- 2019].

[5] J. Boyd, "Handicapping College Football Over/Under Lines: Key Numbers for Totals", *Boyd's Bets*, 2019. [Online]. Available: https://www.boydsbets.com/key-numbers-for-college-football-totals/. [Accessed: 12- Nov- 2019].