Using Recruiting Rankings and Returning Team Measurements to Predict College Football Team Success

by

Sydney Lynn Singleton

Honors Thesis Project

Appalachian State University

Submitted to the Department of Mathematical Sciences

in partial fulfillment of the requirements for the degree of

Mathematics, Bachelor of Science

May, 2019

Approved by:

Ross Gosky, Ph.D., Thesis Director, Mathematical Sciences

Rene Salinas, Ph.D., Second Reader, Mathematical Sciences

William J. Cook, Ph.D, Honors Director
Department of Mathematical Sciences

Eric Marland, Ph.D. Dean, Chair
Department of Mathematical Sciences

Abstract

This paper proposes and compares a set of models of college football team performance for teams in major conferences during the years of 2006 – 2018. The outcome measure of team performance is the team's standardized Sagarin Ranking at the end of the season after the postseason bowl games and, in recent years, playoff games are complete. Potential predictor variables include several variables taken from the team recruiting rankings at the website www.rivals.com, and other attributes of the team compiled from an annual college football prediction magazine. Models considered include models screened via traditional forward, backward, and stepwise model selection methods, as well as a regression tree model. These candidate models are first compared using a cross-validation technique where each individual season is used successively as a test data set, and the predictive accuracy of the candidate models are compared after these successive comparisons. We find that the model chosen via stepwise selection performs the best in this cross-validation comparison but that other models have comparable error rates. We further consider refinements of the forward selection model when quadratic terms and a piecewise approach is taken for two predictors, and compare the prediction error rates for these models using the same cross-validation technique. Our findings from these analyses suggest that teams with higher recruiting rankings are predicted to perform better in a given season, but that other factors about the team are also significant predictors of performance.

1. Introduction

College football recruiting is a source of significant interest, especially for fans of teams in the largest conferences, such as the Southeastern Conference (SEC), the Big Ten, the Big Twelve, the Pac-12, and the Atlantic Coast Conference (ACC). As a sign of the popularity of evaluating recruits who join teams every year, several college-sports related websites publish rankings of team recruits annually, and the largest ones have historically included www.rivals.com, www.scout.com, www.espn.com, and www.247sports.com.

Each recruit who signs a letter of intent to accept a scholarship to play football at a Football Bowl Subdivision (FBS) school is assigned a recruiting ranking, which is based on the opinion of analysts about the recruit's football potential at the college level. To an extent, the following distinctions are slightly different across the recruiting sites, but all of them use similar, and common language to rate each recruit. This rating system is known as the star-system, which functions similarly to movie ratings in that players with higher star ratings are considered better college prospects. Five-star recruits are generally regarded as among the best 25-50 players in the entire country, regardless of position. Four-star recruits are generally regarded as players who are not five-star recruits, but nonetheless possess significant potential, and are generally among the best 250-300 players in the country. Three-star recruits are defined similarly to four-star recruits, but are regarded as among the best 750 or so players in the country. Two-star recruits are regarded as outside the best 750 or so players in the country, although they are good enough players to have earned a college scholarship. There is not generally a designation below two-stars for a recruit, although some recruits who are not well known can potentially be unranked.

College football teams at the highest (FBS) subdivision can have a total of 85 players on scholarship at any given time. Typically, teams offer scholarships to approximately 20-25 players per season, based upon the number of scholarships available, and each team's recruiting class will be ranked by the websites mentioned above. In the Rivals.com case, these rankings are based upon a calculated number of total recruiting points summarizing the team's recruiting class in a given year. The team recruiting rankings on the Rivals.com website for the year 2017 are listed at https://n.rivals.com/team_rankings/2017.

The effect of recruiting rankings on predicting player and team performance has been the subject of curiosity and analysis in recent years. One study used recruiting data from the years 2002 – 2012 to show that teams that recruit higher-rated players do generally achieve higher performance on the field in terms of wins, and they found statistically significant effects of recruiting after accounting for school effects on performance (Bergman and Logan 2014). Additionally, Dronyk-Trosper and Stitzel (2017) also found some evidence of associations between recruiting rankings and win percentage, but also suggested that these effects may be program specific in that successful teams show a stronger association between recruiting rankings and team win percentage than do weaker teams. Other articles have examined recruiting effects on team performance as well, and these are intended as two relatively recent, peer-reviewed research examples on the subject.

Many popular press articles have been written on the subject, such as Hinton (2014), Pettigrew (2015), and Boyd (2015). Each of these writers expresses somewhat different views on the usefulness of the recruiting rankings in predicting team success. Specifically, Hinton argued that aggregated recruiting ratings alone can predict the winner of many head to head matchups between teams from the largest conferences, Pettigrew looked at how well teams have performed compared with regression model based predictions based upon their recruiting rankings, and Boyd uses the success of certain teams to argue that recruiting rankings are flawed due to the ability of these noted teams to find players who fit their system and who perform well despite not being elite recruits.

Preseason predictions also are an important aspect of any sport, and many magazines and websites make these predictions before the beginning of any given season. Some of the most popular prediction magazines are Athlon (https://athlonsports.com/college-football), Lindy's Sports (http://www.lindyssports.com/) , Sporting News (http://www.sportingnews.com/), ESPN (www.espn.com), Sports Illustrated (www.si.com), and others. These magazines predict the ranks of the teams each upcoming season, taking into account whatever available information they choose. Some of the more detailed preview magazines, such as Lindy's Sports, also list aspects of each team, such as the number of returning starters per team, the number of years the coach has been with the team, and other information. Most of these magazines and previews will predict the top 25 ranked teams in the upcoming season, and others will provide a predicted rank of all 128 teams currently in the Football Bowl Subdivision.

Measuring a team's success in a given season can be done in many ways, including binary metrics such as a team reaching a bowl game, winning its conference, or being ranked in the postseason top 25 teams via a common ranking poll such as the Associated Press (AP) or Coaches polls that are widely available. Most analyses, including the papers mentioned previously, focused on a team's winning percentage as the outcome measure of success. Winning percentage is certainly a useful measure of team performance, but by itself it fails to take strength of schedule into account. Many other ranking systems exist that attempt to numerically differentiate between the performance of teams regardless of their win-loss record, and two of the most popular are produced by Jeff Sagarin (www.sagarin.com), and Kenneth Massey (www.masseyratings.com). Both of these ratings attempt to quantify the strength of a team in a given season in a manner that takes both team performance and strength of schedule into account. In the case of the Sagarin ratings, each team receives a numeric score using a computational formula, which typically ranges between about 70 and 100 for most teams in the largest college football conferences, where a higher rating is better. This overall rating is driven by three different sub-ratings, but the difference in two teams Sagarin composite ratings in a given season is roughly comparable to the point differential between the two teams quality in a given season. In other words, a team that is 10 points higher in the Sagarin rankings than another would be rated as being roughly 10 points better than the other team on a neutral field.

Our study has a few different goals. First, we want to predict which teams from the major conferences will be successful in a given season. Additionally, we want to make inferential conclusions about the role of recruiting rankings and other predictors in these models. We will also consider some additional team-related factors, such as returning starters, coach experience, and the team's previous year's performance, among others, in our models to determine whether recruiting rankings are statistically significant in a model which already takes some team characteristics and recent performance into account. In other words, do recruiting rankings matter as a differentiating predictor between two teams who were equally good the previous season and with with similar team characteristics? Furthermore, given the many components of recruiting rankings, we also wanted to determine which components (if any) of the recruiting rankings were important in prediction of team performance. Given our balanced goals of inference and predictive accuracy, we focused on models for which both predictive accuracy could be assessed and for which clear inferential conclusions about the predictors in the model could be clearly assessed.

In Section 2, we describe the variables we collected to conduct the analysis. In Section 3, we describe how we chose a set of initial candidate models to predict team performance via some variable screening techniques and the creation of a regression tree model. In Section 4, we describe the results of some model comparisons to evaluate the performance of the different candidate models through a cross-validation process. In Section 5, we consider some refinements of the multiple regression model that performed the best in the cross-validation analysis in Section 4, with the refinements made to account for some nonlinear effects of some of the predictor variables. When considering these refinements, we conducted some additional cross-validation analyses to determine the final model. Finally, in Section 6, we provide some discussion and conclusions.

2. Data Collection

We collected recruiting data from the Rivals site and team information from the Lindy's Sports college football preview magazine for the years 2006 to present, and have specifically focused on teams in the largest conferences, specifically the SEC, Big Ten, Big Twelve, Pac-12, ACC, and the current AAC, which historically was called the Big East Conference. Notre Dame was also included although they are historically independent in football. We focused on teams from these conferences because recruiting rankings tend to vary the most among teams from the major conferences. Generally, within the smaller Football Bowl Subdivision (FBS) conferences, many recruits are ranked at the two-star level, leading to more homogeneity among recruiting rankings than we prefer for our analyses. Furthermore, predictions tend to focus attention toward the top teams, which belong to the conferences that we have included in our analysis.

Our data set had the following variables (and variable names) measured for each team for each season:

- Team Sagarin rating at the end of the season (both raw and standardized to account for different mean ratings per season)
  - Sagarin (raw), Zsagarin (standardized)
- Yearly rivals.com recruiting measurements for the most recent five years (Freshmen, Sophomore, Junior, Senior, Redshirt Senior), including the number of total recruits, the number of five, four, and three star recruits, and the average star rating for the class:
  - Frnbrrecruits, Sonbrrecruits, Jnrnbrrecruits, Snrnbrrecruits, Rssrnbrrecruits, Fr5star, Fr4star, Fr3star, So5star, So4star, So3star, Jr5star, Jr4star, Jr3star, Sr5star, Sr4star, Sr3star, Rssr5star, Rssr4star, Rssr3star, Fravg, Soavg, Jravg, Sravg, Rssravg
- Conference affiliation
  - Binary values for the variables : BigTen, SEC, ACC, BigTwelve, Pacten, Bigeast
- Team Sagarin rating at the end of the previous season (both raw and standardized)
  - Lysagarin (raw), z_lysagarin (standardized)

- Team Sagarin rating at the end of the season two years prior (both raw and standardized)
  - Tyasagarin (raw), z_tyasagarin (standardized)
- Returning offensive and defensive starters for the season as determined by Lindy's Sports Magazine
  - Retoff, retdef
- A binary variable to indicate whether the team returns its starting quarterback from the previous year
  - qbret
- A binary variable to indicate whether the team participated in a bowl game in the previous year
  - bowl
- The number of bowl games the team won the previous season (note: in almost all cases this is 0 or 1, but the national champion in the recent college football playoff system can technically win 2 bowl games)
  - bowlwin
- Number of years of head coaching experience for the team's head coach, both at the school, and overall as a college football Division 1 head coach
  - coachexp_school, coachexp

In order to differentiate the recruiting ratings for the five previous seasons, we refer to the familiar class-year designation in college football, where Freshmen, Sophomores, Juniors, Seniors, and Redshirt Seniors for recruiting rankings refer to the most recent five years of recruiting classes, respectively. It is worth noting that our recruiting rankings are taken from the

Rivals website and fixed once the freshman class is signed for each team. We did not account for transfers in and out of the program, for graduations among most seniors after four years in the program, or for injuries. Our recruiting rankings summary is primarily intended to measure a rolling five year performance in recruiting, with the recognition that the majority of the team roster is comprised of players recruiting within the previous five seasons.

We are not using the total Rivals.com recruiting points for each class as a predictor variable in this analysis for two reasons. Firstly, the formula from which these recruiting points are calculated was modified in 2013, and has not been perfectly consistent for the duration of this data set, although higher point values indicate a stronger recruiting class in all cases. Secondly, the updated formula is largely driven by factors already accounted for in the rest of the variables about recruit level quality that we used in our analyses.

3. Candidate Models

Due the large number of potential predictor variables, we performed some classical variable screening procedures on the data to determine an initial set of candidate regression models.

For each model, our response variable was the team's standardized Sagarin score when compared with all the teams in our study for that particular season. For example, a team with a standardized Sagarin score of +1.5 would have a Sagarin score 1.5 standard deviations above the mean Sagarin score for all the teams in our data for that particular season. Potential predictor variables included all of the other variables mentioned previously in Section 2.

We used the R version 3.5.1 software (https://cran.r-project.org/) to perform forward, backward, and stepwise variable screening and used Akaike's Information Criterion (AIC) (Akaike, 1973) as the method of model comparison. The final model chosen by each method was the one that obtained the minimum AIC value among the models screened at this stage.

Tables 3.1 and 3.2 illustrate the results of forward and backward variable screening methods:

Table 3.1: Results of Forward Variable Screening Method

Step	# of Variables	Formula	AIC
1	0	Zsagarin ~ 1	-26.3
2	1	Zsagarin ~ z-score for last years sagarin (z_lysagarin)	-520.45
3	2	Zsagarin ~ z_lysagarin + Fravg	-592.52
4	3	Zsagarin~ z_lysagarin + Fravg + retoff	-612.7
5	4	Zsagarin~ z_lysagarin + Fravg + retoff + retdef	-623.59
6	5	Zsagarin~ z_lysagarin + Fravg + retoff + retdef + Jnrnbrrecruits	-633.16
7	6	Zsagarin~ z_lysagarin + Fravg + retoff + retdef + Jnrnbrrecruits + Jravg	-640.83
8	7	Zsagarin~ z_lysagarin + Fravg + retoff + retdef + Jnrnbrrecruits + coachexp_school	-645.24
9	8	Zsagarin~ z_lysagarin + Fravg + retoff + retdef + Jnrnbrrecruits + qbret	-648.1
10	9	Zsagarin~ z_lysagarin + Fravg + retoff + retdef + Jnrnbrrecruits + qbret +Fr5star	-650.03

11	10	Zsagarin~ z_lysagarin + Fravg + retoff + retdef + Jnrnbrrecruits + qbret + Fr5star + Jr5star	-651.56
12	11	Zsagarin~ z_lysagarin + Fravg + retoff + retdef + Jnrnbrrecruits + qbret + Fr5star + Jr5star +So5star	-652.56

Table 3.2: Results of Backward Variable Screening Method

Step	**# of Variables**	**Formula**	**AIC**
1	40	Zsagarin ~ FrNbrRecruits + Fr5star + Fr4star + Fr3star + Fravg + Sonbrrecruits + So5star + So4star + So3star + Soavg + Jrnbrrecruits + Jr5star + Jr4star + Jr3star + Jravg + Srnbrrecruits + Sr5star + Sr4star + Sr3star + Sravg + Rssrnbrrecruits + Rssr5star + Rssr4star + Rssr3star + Rssravg + z_lysagarin + z_tyasagarin + retoff + retdef + qbret + bowl + bowlwin + coachexp_school + coachexp_total + BigTen + SEC + BigTwelve + ACC + PacTen + Bigeast	-615.03
2	39	Removed Sr3Star	-617.02
3	38	Removed Rssr3star	-619.02
4	37	Removed Rssr5star	-621
4	36	Removed Fravg	-622.97
5	35	Removed BigTwelve	-624.83
6	34	Removed Rssravg	-626.68
7	33	Removed SEC	-628.4
8	32	Removed PacTen	-630.16

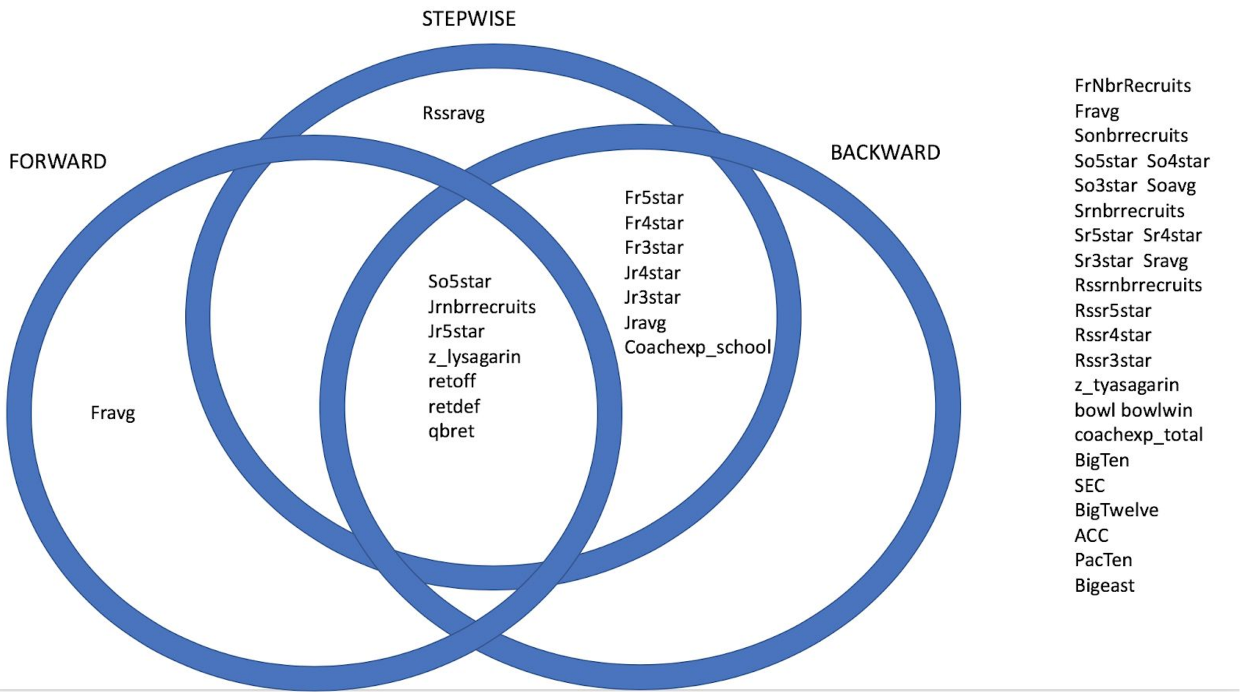
9	31	Removed BigTen	-631.98
10	30	Removed ACC	-633.74
11	29	Removed Sonbrrecruits	-635.33
12	28	Removed So3star	-636.84
13	27	Removed Soavg	-638.42
14	26	Removed So4star	-640.01
15	25	Removed Sr4star	-641.54
16	24	Removed Sravg	-643.34
17	23	Removed z-tyasagarin	-644.82
18	22	Removed Bigeast	-645.87
19	21	Removed Srnbrrecruits	-646.66
20	20	Removed Frnbrrecruits	-647.47
21	19	Removed bowl	-648.25
22	18	Removed Sr5star	-649.06
23	17	Removed coachexp_total	-649.54
24	16	Removed bowlwin	-649.85
25	15	Removed Rss4star	-650.12
26	14	Zsagarin ~ Fr5star + Fr4star + Fr3star + So5star + Jrnbrrecruits + Jr5star + Jr4star + Jr3star + Jravg + z_lysagarin + retoff + retdef + qbret + coachexp_school	-650.14

		(Removed Rssrnbrrecruits)	

Performing stepwise selection on this dataset yielded very similar results to the backward selection screening method. The only difference is that stepwise added one more step at the end where the redshirt senior average variable was added back into the model after being removed earlier in the process. The resulting AIC of the final stepwise model is -650.3.

To summarize these table illustrations, refer to the venn diagram below that visually compares and contrasts the significant variables in the three final models. The list on the right shows the variables that did not show up in any of the models.

Figure 3.1 Venn Diagram for Model Comparison



4. Model Comparisons

In order to evaluate the predictive accuracy of the different models in Section 3, we evaluated the performance of each of the models through a cross validation process. Our cross-validation procedure created a sequence of training and test data sets, by keeping each season successively as a test data set and the remaining seasons as the training data set at each step of the sequence. Table 4.1 illustrates how the training and test data sets were created for the first four of the 13 comparisons.

Table 4.1: Test and Training Datasets for First Four Cross-Validation Model Comparisons

Comparison Number	**Test Data Set**	**Training Data Set**
1	2006 season	2007 - 2018 seasons
2	2007 season	2006, 2008-2018 seasons
3	2008 season	2006-2007, 2009-2018 seasons
4	2009 season	2006-2008, 2010-2018 seasons

For each comparison, we fit each candidate model to the training data set, and used that model to predict the standardized Sagarin Score for each team in the test data set. We measured the predictive accuracy for each model, for each comparison, by using two metrics: Mean Absolute Prediction Error and a Mean Square Error. These metrics are given by:

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{n}$$

$$MAPE = \sum_{i=1}^{n} \frac{|(y_i - \hat{y_i})|}{n}$$

Equations 4.1 and **4.2**

where $\hat{y}$ is the predicted value, ***y*** is the actual value, and ***n*** is the number of observations in the test dataset. These two values, while they yielded similar results, differ in some ways. The MAPE simply measures the average distance between the predicted and actual value. The MSPE measures the squared difference of the predicted and actual values; therefore, the MSE penalizes more for differences between ***y*** and $\hat{y}$ that are larger than one in absolute value**,** while penalizing less for smaller differences.

The following tables show the MSPE (Table 4.2) and the MAPE (Table 4.3) that were calculated from the predictions of each season. The bolded numbers show which model had the lowest MSPE/MAPE of the three for that year. We also show the average error according to each measure at the bottom of the table. We first notice that the performance of each model in this cross-validation exercise is comparable. In Table 4.2, both the Forward and Stepwise models had the smallest MSE an equal number of times, and the Stepwise average MSE across all the seasons was slightly smaller than the other two models. In Table 4.3, the Forward selection model had the lowest MAPE value in six of the twelve seasons analyzed, and also the smallest average MAPE value overall. Given the comparability in performance of these models, one can reasonably choose either the Forward or Stepwise models as the best choice. We will focus on the Stepwise model as the best choice due to its performance in Table 4.2 for two reasons. The MSE metric penalized the models for larger inaccuracies which are more than one standard deviation in absolute value, and also, our regression models are chosen according to a least-squares philosophy, which is more consistent with the MSE criterion.

Table 4.2 MSE of Models in Cross Validation

Year	Forward	Backward	Stepwise
2007	.532	.505	**.501**
2008	**.480**	.515	.514
2009	.418	**.415**	.425
2010	.674	**.662**	.665
2011	.459	.452	**.447**
2012	.4587	.462	.**4586**
2013	**.473**	.482	.485
2014	.495	.492	**.489**
2015	**.416**	.422	.419
2016	.456	.444	**.440**
2017	**.413**	.422	.425
2018	**.358**	.363	.360
Avg	.4693	.4696	**.4691**

Table 4.3 MAPE of Models in Cross Validation

Year	Forward	Backward	Stepwise
2007	.581	.569	**.567**
2008	**.585**	.610	.608
2009	.5119	**.5115**	.514
2010	.659	**.6427**	.6431
2011	.556	.545	**.541**
2012	**.561**	.566	.565
2013	**.574**	.575	.578
2014	.562	.559	**.558**
2015	**.501**	.512	.510
2016	.564	**.5508**	.5514
2017	**.504**	.519	.519
2018	**.474**	.482	.482
Avg	**.5528**	.5536	.5530

5. Further Model Modification

Once the model selections were made, we analyzed the relationship of each predictor with the response variable. The visualizations led us to consider that several variables had the possibility of a nonlinear relationship with the response variable. We first explored this possibility by adding quadratic terms to the model chosen by the stepwise selection procedure. However, the only variables that were close enough to statistical significance to merit possible inclusion as predictors in a modified final model were the coach's experience at their current school and the team's number of 4-star freshman recruits, which were present in the backwards and stepwise models. The coach experience variable looked as if it could either be modeled with a quadratic term or a piecewise linear function because its relationship with the response variable increased linearly until around year ten, then it plateaued with modest evidence of a decline. We inserted a piecewise linear function that modeled the behavior of this variable with respect to time to allow the effect of coaching experience to change after ten years at the school.

In creating these two models, we discovered that the variable added to create the piecewise function was close to statistical significance with a p-value of 0.05881. However, the quadratic term for the coach experience was deemed significant in the quadratic version of the stepwise model with a p-value of 0.04828, while the quadratic term for the number of 4-star freshman had a p-value of 0.06121.

Tables 5.1 and 5.2 summarize the results of cross-validations including these new potential models, using the same methodology as described in Section 4.

Table 5.1 MSPE of Models in Predicting Season Outcomes

Year	Forward	Backward	Stepwise	Piecewise	Quadratic
2007	.532	.505	.501	**.493**	.513
2008	**.480**	.515	.514	.506	.530

2009	.418	**.415**	.425	.451	.447
2010	.674	.662	.665	**.660**	.650
2011	.459	.452	.447	**.439**	.447
2012	.4587	.462	.4586	.457	**.447**
2013	.473	.482	.485	.476	**.470**
2014	.495	.492	**.4889**	.496	.4894
2015	.416	.422	.419	**.412**	.405
2016	.456	.444	**.440**	.442	.446
2017	**.413**	.422	.425	.4135	.415
2018	**.358**	.363	.360	.387	.376
Avg	.4693	.46962	**.4691**	.4695	.46964

Table 5.2 MAPE of Models in Predicting Season Outcomes

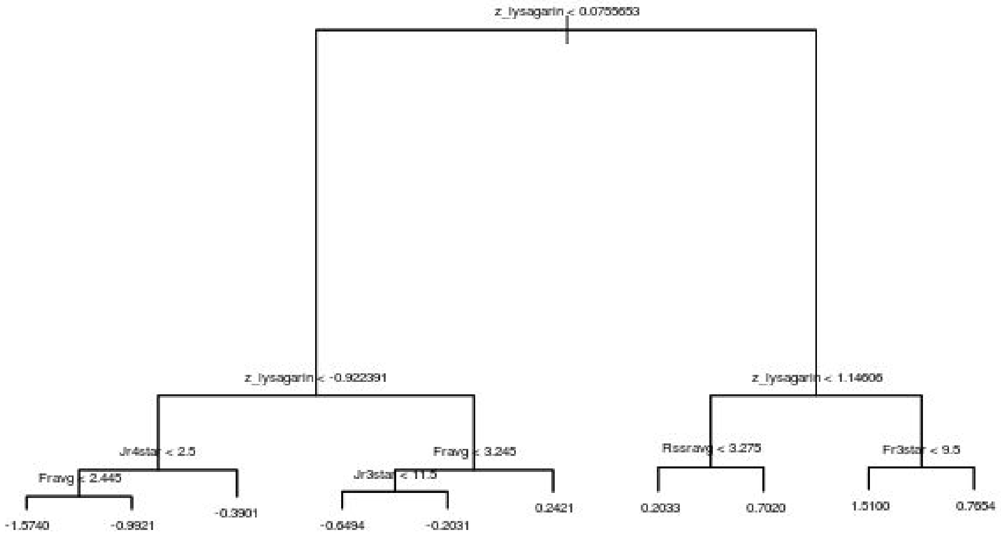
Year	Forward	Backward	Stepwise	Piecewise	Quadratic
2007	.581	.569	.567	**.559**	.569
2008	**.585**	.610	.608	.599	.613
2009	.5119	**.5115**	.514	.531	.525

2010	.659	.6427	.6431	.640	**.637**
2011	.556	.545	.541	**.531**	.540
2012	.561	.566	.565	.566	**.560**
2013	.574	.575	.578	.567	**.564**
2014	.562	.559	**.558**	.564	.562
2015	.501	.512	.510	.497	**.495**
2016	.564	**.5508**	.5514	.5509	.559
2017	**.5035**	.519	.519	.5039	.5041
2018	**.474**	.482	.482	.502	.494
Avg	.5528	.5536	.5530	.5510	**.5518**

As you can see, the model with the quadratic terms and the model with the piecewise linear portion did not drastically improve the error margins. The original stepwise model had still had the lowest MSE; however, the model with the piecewise function did have the lowest average MAPE of the five models.

To provide another prediction model for comparison, we also considered a decision tree model. Decision trees find different nodes in the data that are predictive of the chosen response variable. The top three nodes were conditions based on a teams performance in the previous year, quantified by the Z-score for last year's Sagarin score. Figure 5.1 shows the tree diagram created based on the entire dataset.

Figure 5.1 Tree Model Created from Entire Dataset

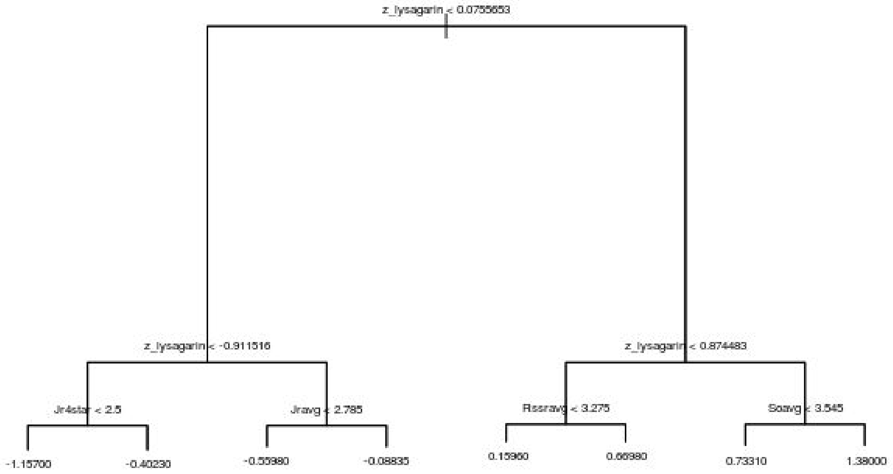


One should interpret this tree model by reading the statement at the top node, and moving down and to the right if the statement is true in the dataset, or moving down and to the left if the statement is false. Repeat this method until you reach the result found at the bottom of the model.

From Figure 5.1, we see similarity between the regression tree model and the multiple regression models we established in Section 3. Specifically, the first two nodes in the diagram are based upon the prior year's standardized Sagarin score, which corresponds with that variable's being selected first in our forward model selection in Section 3. After the first two nodes, the predictor variables at each node are recruiting related, with an emphasis on the Freshman and Junior classes, which is also generally similar to our multiple regression models from Section 3. In total, the first two nodes roughly split teams into four groups based upon last year's standardized performance: teams more than one standard deviation away from the mean on each side (positive and negative), and teams within one standard deviation from the mean on each side (positive and negative). Based upon this distinction, different recruiting predictor variables were selected in later nodes to further explain the difference in performance.

In addition to producing the tree diagram for the whole dataset, we also performed cross-validation for this model. This process was slightly different from the cross-validation on the multiple regression models. Rather than fitting a model to the remaining data after it has been subset, we produced a new model tree based upon the training data for each iteration. For example, Diagram 5.2 shows the predictive tree diagram for the data when the 2018 season was held out as a test dataset.

Diagram 5.2 Tree Model Created in Cross-Validation when 2018 Season Omitted



Note that the limitation of using tree diagrams to predict a continuous quantitative variable, such as the Z-score for a team's Sagarin score, is that the diagram only predicts a finite amount of outcomes, one for each terminal node in the diagram. Therefore, several teams will receive the same prediction, which structures the variable categorically, yielding high prediction error metrics as you can see when we incorporate the decision trees into the cross-validation results in Table 5.3 and Table 5.4.

Table 5.3 MSE of Models in Predicting Season Outcomes

Year	Forward	Backward	Stepwise	Piecewise	Quadratic	Tree
2007	.532	.505	.501	**.493**	.513	.636
2008	**.480**	.515	.514	.506	.530	.584
2009	.418	**.415**	.425	.451	.447	.450

2010	.674	.662	.665	**.660**	.650	.848
2011	.459	.452	.447	**.439**	.447	.589
2012	.4587	.462	.4586	.457	**.447**	.486
2013	.473	.482	.485	.476	**.470**	.564
2014	.495	.492	**.4889**	.496	.4894	.527
2015	.416	.422	.419	**.412**	.405	.430
2016	.456	.444	**.440**	.442	.446	.578
2017	**.413**	.422	.425	.4135	.415	.591
2018	**.358**	.363	.360	.387	.376	.441
Avg	.4693	.46962	**.4691**	.4695	.46964	.5603

Table 5.4 MAPE of Models in Predicting Season Outcomes

Year	Forward	Backward	Stepwise	Piecewise	Quadratic	Tree
2007	.581	.569	.567	**.559**	.569	.587
2008	**.585**	.610	.608	.599	.613	.603
2009	.5119	.**5115**	.514	.531	.525	.529

2010	.659	.6427	.6431	.640	**.637**	.734
2011	.556	.545	.541	**.531**	.540	.618
2012	.561	.566	.565	.566	**.560**	.590
2013	.574	.575	.578	.567	**.564**	.608
2014	.562	.559	**.558**	.564	.562	.557
2015	.501	.512	.510	.497	**.495**	.534
2016	.564	**.5508**	.5514	.5509	.559	.619
2017	**.5035**	.519	.519	.5039	.5041	.612
2018	**.474**	.482	.482	.502	.494	.514
Avg	.5528	.5536	.5530	**.5510**	.5518	.5921

While it is known that several enhancements to the regression tree modeling procedure, such as boosting or random forests, can improve the predictive performance of these models, we did not pursue those options for two reasons. First, because our cross validation approach was done on a season-by-season basis and not on a random selection of observations from the overall data set, using the regression tree was easier to implement in a comparable fashion for cross validation. Second, these enhancements can create a loss in the interpretability of the specific model predictors beyond their overall importance, and this interpretability was an important consideration in our modeling.

6. Conclusions

Based on the models that were selected (Figure 3.1) and the cross-validation results described in Section 4, we achieved the following conclusions:

- Recruiting does matter:
    - The number of Freshman 3, 4, and 5-star recruits were significant predictors in 2 out of the 3 multiple regression models.
    - The Junior class is very important in determining a team's performance: the size of the junior class and the number of 5 star Juniors were both present in all three multilinear models. The variables for the number of 3 and 4-star Juniors were present in the backward and stepwise models. The variable for 4-star Juniors also showed up in the tree diagram along with the average number of stars for the Junior class.
- Returning starters are a significant predictor of performance: the number of returning offensive and defensive starters, along with the binary representation of a returning quarterback, were present in all three multiple regression models.
- A team's performance in the previous year is an obvious indicator of the current year's performance: the standardized Sagarin score from the previous year showed up in every model that we produced.

To conclude, our findings suggest that how well a team recruits is a significant predictor of their on-field performance, not only when the recruits are freshmen, but also how they develop by their junior year. While recruiting is an important predictor of team success, we also note that the forward selection procedure selected the previous year's standardized Sagarin score as the single best predictor of team performance. This suggests that while recruiting is important, the team’s most recent performance is the best single predictor of their future success. But, the significance of several recruiting-related predictor variables in our final model also says that if two teams were equally good the previous year and have equal team characteristics in terms of returning starters and coaching experience, the team with higher recruiting ratings across their classes will be predicted to perform better. The fact that some recruiting variables for the freshman class are included in our final model (compared with, say, the senior class) may be initially surprising, because more team starters will be seniors rather than freshman on most teams. But, most of the senior class recent performance and ability is already quantified indirectly in the team’s standardized Sagarin score from the previous year. The recruiting rankings of the freshman class describes players who were not on the team the previous year, so bringing in a strong incoming freshman class would sensibly be associated with higher performance among two teams who are otherwise equal. Following this logic, the recruiting rankings of the Freshman through Junior classes are most likely describing players who had less to do with the team’s success the previous season than the rising Senior class, so the inclusion of more recruiting rankings for those classes makes sense as well.

We also used our models to assess conferences indirectly, to see if there were any conference affiliations that (all else being equal) led to higher or lower predicted performances. There are frequently discussions about conference superiority during the bowl and playoff season, but our model found that conference affiliation was not a significant predictor of success once other team factors and recruiting rankings were taken into account.

Another idea that can be evaluated with our models was possible carryover effects regarding bowl participation and bowl wins in the previous season. A team that participates in a bowl game has additional end-of-season practices to prepare for that game, and those practices may benefit some of the younger team players, which could presumably benefit the team further in the following season. Similarly, a bowl win the previous season may help the confidence of the team heading into the next season. However, our model did not find these effects to be significant once the other factors were taken into account. Regarding bowl participation, the lack of significance of this predictor is likely because teams that do not make a bowl game the previous season have many differences from teams that do. Thus, any benefits of extra bowl practices would most likely be seen in teams that were among the weaker bowl teams or the strongest non-eligible bowl teams from the previous season, and these differences did not seem to make a significant difference the following season. Likewise with bowl wins, which would occur for half of the bowl teams the previous season. We did not see evidence that winning a bowl game the previous season had carryover predictive effects the following season.

The model with the inserted piecewise linear function, while having the smallest MSE, only had an R^2 of .5327. This tells us that in our data, despite all of the available information about recent performance, returning team characteristics, and recruiting rankings, our model explains about 53.3% of the variation in team performance. This fact suggests that despite attempts to accurately predict season outcomes every year, there remains a significant amount of variation in performance, which likely is a component helping the popularity of the sport in the long-run.

References:

Dronyk-Trosper. T. and Stitzel, B. (2017). Lock-In and Team Effects: Recruiting and Success in College Football Athletics. Journal of Sports Economics, 18(4), 376-387.

Bergman S. A. and Logan T. D. (2014). The effect of recruit quality on college football team performance. Journal of Sports Economics, 17(6) 578-600.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory (Petrov BN, Csaki F, eds). Budapest:Akademiai Kiado, 267–281.

Hinton, M. (2014, February 5). Recruiting Matters: Be they ever so humble, the rankings (still) get it right [Web log post]. Retrieved February 5, 2014, from http://www.footballstudyhall.com

Boyd, I. (2015, February 10). Why college football recruiting rankings are flawed metrics [Web log post]. Retrieved February 10, 2015, from http://www.footballstudyhall.com

Pettigrew, S. (2015, February 4). After Signing Day, Wisconsin Makes The Best Of Its Recruits [Web log post]. Retrieved February 4, 2015, from https://fivethirtyeight.com