Colette Barca, Keith Osani, Nisha Srishan, and William Wulster

Professor Tweneboah

MATH 570-01

11 December 2020

College Football Recruiting

College football is big business. As the highest grossing college sport in the United States, "football brings in [approximately] $31.9 million in revenue" per school.[1] On average, colleges spend anywhere from $700,000 to $2 million on football recruiting costs.[2] To create a winning team that will continue to generate revenue, a Division I Football Bowl Subdivision (FBS) school needs to choose its recruits wisely. In this regard, we built a model to predict whether or not a high school recruit will commit to a school in one of the Power Five Conferences. Such a model could allow a school to optimize its recruiting process, maximizing its return on investment. To build the model, we chose a dataset from github.com that included over 66,000 high school football recruits of the 2010 – 2019 graduating classes. We narrowed this down to only include recruits who are ranked by 247sports.com, committed to an FBS school, and of the graduating classes 2015 – 2019.

Our modified dataset contains 114 distinct Division I FBS colleges. Each of these teams belong to one of ten FBS conferences, except for Army West Point, Notre Dame, New Mexico State, and Liberty, as they are FBS independents. Five of these ten FBS conferences are referred to by industry professionals as the elite of college football. These "Power Five" conferences are: the Southeastern Conference (SEC), the Atlantic Coast Conference (ACC), the Big Ten Conference (B10), the Pac-12 Conference (Pac-12), and the Big 12 Conference (B12); 61 of the teams in our dataset belong to this group. The five remaining conferences in our dataset are the Mid-American Conference (MAC), the Sun Belt Conference (SBC), the Conference USA (CUSA), the Mountain West Conference (MWC), and the American Athletic Conference (AAC) and are subsidiary to the aforementioned conferences (Appendix Table A).

[1] https://finance.zacks.com/much-money-college-sports-generate-10346.html

[2] https://247sports.com/Article/college-football-recruiting-rankings-spending-budget-134080367/

The github dataset provided specific information on each recruit and college. It described recruits using both physical (height, weight, high school, etc.) and performance (national rank, star count, position, etc.) attributes. College features detailed the schools' locations and conferences. In an attempt to formulate a more comprehensive analysis, we chose to avoid only including features that describe each recruit's football skills. We conjectured that certain information on each recruit's hometown might be significant when predicting if they will commit to a particular type of school.

The github dataset contained 21 unique football positions. We can segment these positions into nine distinct position groups, corresponding to the information contained in collegefootballdata.com's API.[3] These seven groups can be succinctly described by three categories, contingent on the side of the field the athlete plays on (offense or defense) and if the position is industrially labeled as "skilled". This concept is explained in further detail in Appendix Table B.

We used each recruit's homestate to create two dummy predictors: hotbed state and college in home state. In football recruiting, hotbed states refer to the four states that approximately 50% of all FBS recruits hail from: Texas, California, Georgia, and Florida, evident in Figure 1.[4] The bar chart in Figure 2

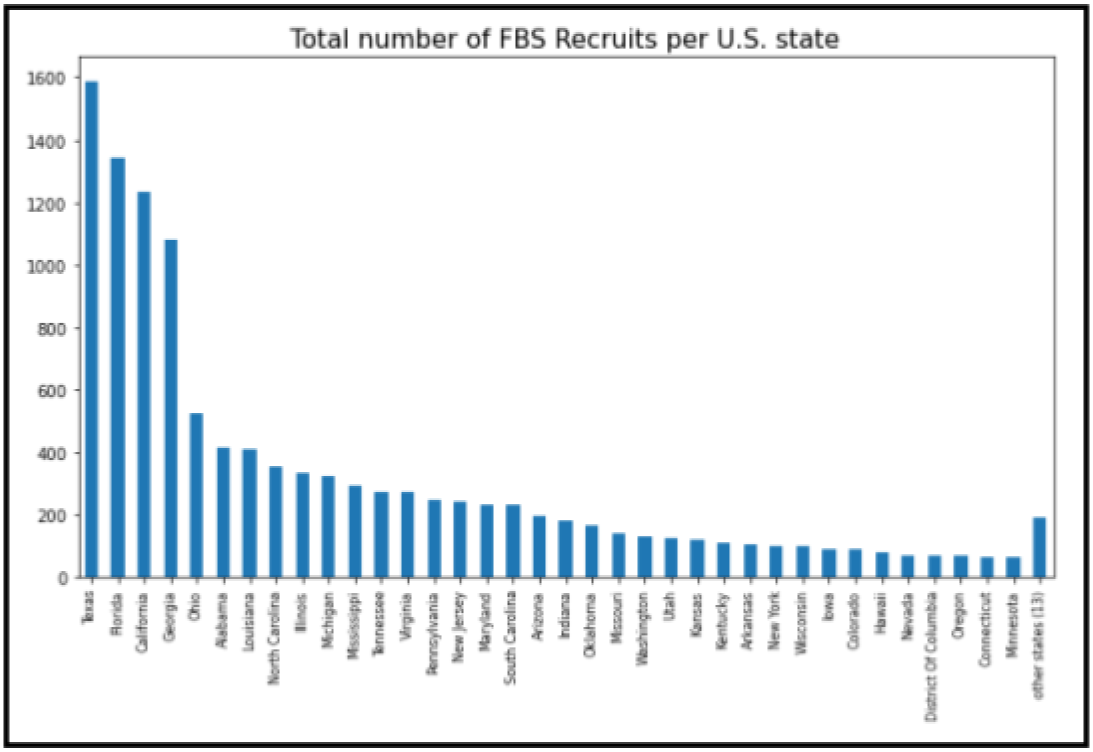


Figure 1: Distribution of Recruits by State

[3] https://api.collegefootballdata.com/api/docs/?url=/api-docs.json

[4] https://www.bannersociety.com/2020/2/4/21111828/college-football-recruits-by-state

displays the distribution of recruits from hotbed states (orange) versus other states (blue) for both Power Five and other FBS school commitments.The addition of this dummy variable made us speculate if a recruit's home state influenced his commitment decision. Hence, the college in home state predictor indicates if a recruit is an in-state student at his respective school. Table 1 shows the distribution of recruits who are committed to in-state and out-of-state schools.

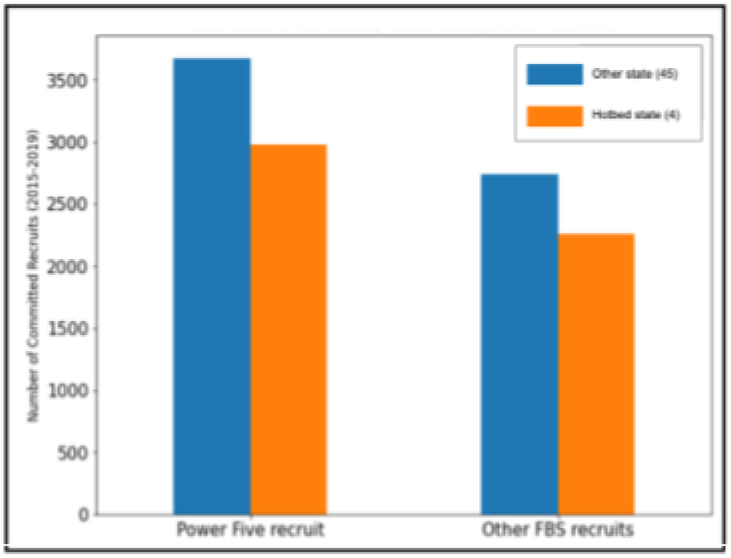
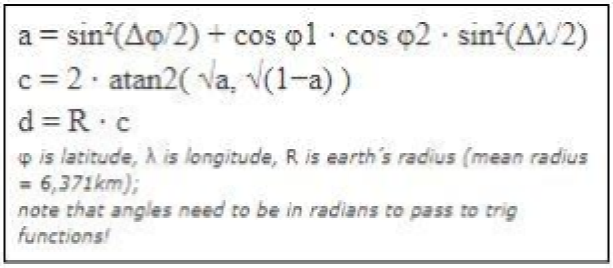


Figure 2: Distribution of FBS Recruits, Hotbed States, and the Power Five

Table 1: Distribution of In-State and Out-of-State Recruits

	Total Recruits	Power Five Commits	Other FBS Commits
In-State College Commitment	**3,718** recruits *32% of dataset*	**1,545** recruits *41% of in-state commits*	**2,173** recruits *58% of in-state commits*
Out-of-State College Commitment	**7,941** recruits *68% of dataset*	**3,700** recruits *47% of out-of-state commits*	**4,241** recruits *53% of out-of-state commits*

Using each recruit's hometown data, we acquired the pertinent 2010 census data, grouped by U.S. zip codes. We used this information to quantify the average income per household and average house value for each recruit's hometown, believing they could be potentially significant attributes. From a student's perspective, the traveling distance between his home and a college could alter which school he commits to. To calculate this distance, we needed the latitude and longitude of each recruit's high school and college. We utilized the geocode endpoint from the Google Maps API to get the zip codes, latitudes, and longitudes of the aforementioned locations. We wrote a Python script to generate the addresses from the data we already had and then passed those addresses through the API to obtain the desired data. We then calculated distance via the Haversine formula (Figure 3).



$$a = \sin^2(\Delta\varphi/2) + \cos\varphi 1 \cdot \cos\varphi 2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \mathrm{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km);
note that angles need to be in radians to pass to trig functions!

Figure 3: Haversine Formula

Typically, various seasonal team statistics from the previous year influence a recruit's commitment decision. Consequently, we felt including such information would result in a more authentic model. We used sports-reference.com to obtain each college's seasonal strength of schedule, total offense, total defense, and final Associated Press (AP) ranking, while their yearly number of commits by position group was collected from the collegefootballdata.com API. Such data were aggregated for the 2014 – 2019 seasons, as our recruits are from the 2015 – 2019 graduating classes.

When initially reading in the data, our dataset contained 1,948 observations with missing data; however, this assessment was conducted on all 42 variables. Several of these variables, such as a school's zip code or a college's division, were highly correlated; they were utilized to create other predictors. When performing an analysis, it is imperative that only independent variables are used for modeling; hence, we removed many features. Once we reduced the dataset to 21 variables, only 817 observations contained missing data. Instead of removing these instances, we chose to impute values for the missing height, weight, average house value, and income per household entries. We wanted to ascribe values as authentically as possible. Thus, we made logical groupings prior to calculating and inserting the replacement values. For height and weight, we imputed the mean, grouped by recruit position, class, and star count. For average house value and income per household, we imputed the median value grouped by state and class. Following these imputations, six observations with missing data remained. For simplicity, we dropped these instances, leaving our final dataset with 11,659 samples and 16 variables.

We used our dataset to analyze a supervised classification problem with fifteen predictors in which the target variable is commitment to a Power Five school. 6,656 of the recruits in our dataset committed to a Power Five school, while the remaining 5,003 recruits committed to some other FBS school. Figure 4 supports the notion that the dataset is an example of relatively balanced classification. Among the fifteen predictors, five are categorical and ten are continuous. A table detailing the variables is displayed in Appendix Table C.

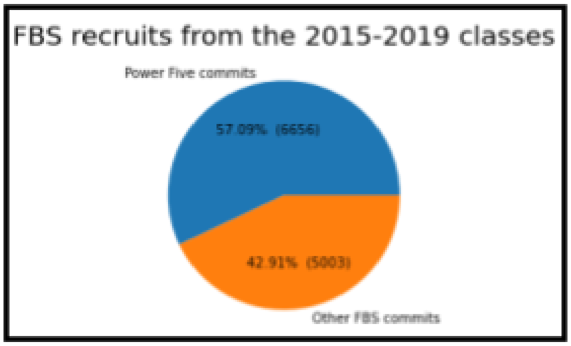


Figure 4: Distribution of Recruit Commitments

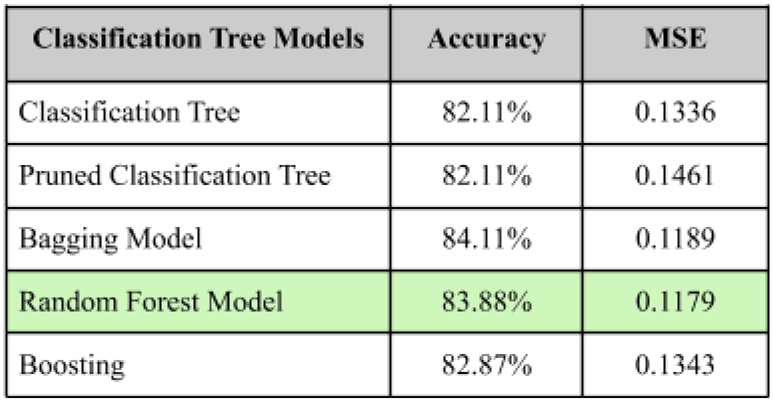
Before starting our analysis, all samples from the graduating class of 2019 were removed from the dataset and reserved for model deployment. This subset contained 2,391 recruits. We decided to set a seed prior to randomly splitting the remaining data into training and testing sets. This ensured the four of us used the same exact observations to fit our models and assess their accuracies. Ergo, we could objectively compare our resulting MSEs and accuracy percentages without having to account for sampling variation. We experimented with various training to testing ratios for dividing the remaining observations, ranging from using fifty to seventy percent of the 9,268 recruits for training. We assessed to see if a particular training to testing ratio yielded significantly more accurate results. Ultimately, we found minimal variability among the resulting MSEs, with the best and worst results differing by a few thousandths. Hence, we chose to proceed with the statistically common partitioning ratio of 2:1 observations for training data to testing data. Our training and testing sets contained 6,116 and 3,152 recruits, respectively.

The four classification techniques implemented in this analysis were: Classification Trees, Support Vector Machines (SVM), Logistic Regression, and Generalized Additive Models (GAM). Our initial models performed with 98 percent accuracy on the testing set. The nearly perfect accuracies spurred further investigation of our dataset. We found that the variables describing the commit school (the college's total offense and defense, its strength of schedule and final AP ranking, and the number of recruits committed within each position group) were either too highly correlated with the target variable or were not correlated with the recruits themselves. Even though these pieces of information are thoroughly reviewed by recruits, their inclusion in our models produced skewed and overly accurate results. Once we removed these variables, the resulting models were far more appropriate for predicting if a recruit would commit to a Power Five school.

Since tree models can be used for both regression and classification, we fit a Classification Tree to our dataset in order to predict our target variable. When using the `tree()` function, we set the target variable to be a factor and used all the predictors. In the produced Classification Tree, national rank was the only variable of importance, as it was the only variable used to split the tree. This tree consisted of four terminal nodes and had a training set misclassification error rate of 17.53 percent. We then performed cross-validation to determine the optimal level of tree complexity in order to assess if a pruned tree created a more accurate model. Plotting the cross-validation result revealed that two terminal nodes produced similar results to the tree with four terminal nodes. The pruned Classification Tree with two terminal nodes also split based on national rank. The percent accuracy of the pruned Classification Tree on the test data is identical to that of the unpruned tree; however, the MSE of the former is increased in comparison to the latter. Therefore, the unpruned Classification Tree proves to be a better fit for the dataset.

For Bagging, we used `mtry=15`, as this technique requires all predictors be used for modeling. This model's training misclassification error was about 17%, meaning it correctly classified the recruits approximately 83% of the time. Some important variables with Bagging included national rank, distance to school, and star count. Next, we moved to a Random Forest model. To choose `mtry`, we tried fitting Random Forest models with integer values ranging from 1 to 14, inclusively. `mtry=5` resulted in the best model, providing an accuracy of approximately 83.03% for the training data. These results are similar to those of our Bagging model. While Random Forest's most important variable was also national rank, its secondary and tertiary predictors were composite rating and position rank, respectively. For the Boosting model we used the `gbm()` function and set the distribution to `"bernoulli"` to signify a classification problem. For this model, national rank outperformed the other variables in terms of relative influence with a value of 37.0874; the second variable, distance to school, had a value of 11.0595. Composite rating's value of 10.0464 made it the third most influential Boosting variable. A comprehensive summary of the results of these five models applied to our test data are displayed in Table 2 below.

Table 2: Classification Tree Summary



Classification Tree Models	Accuracy	MSE
Classification Tree	82.11%	0.1336
Pruned Classification Tree	82.11%	0.1461
Bagging Model	84.11%	0.1189
Random Forest Model	83.88%	0.1179
Boosting	82.87%	0.1343

When comparing all five models, Bagging and Random Forest are highly comparable in their results when being applied to the test data, providing very similar numbers in accuracy percent and MSE. However, as Random Forest has a slightly lower MSE, this model proves to be the best of the five models.

We also trained our binary classification problem on various SVM models. The SVM was trained to classify our dataset using different algorithms to separate the set of observations with a hyperplane of varying shapes (linear, radial, polynomial). We then passed in a set of parameters to each of the models to optimize hyperplane development. The first parameter, `cost`, controls the margins; smaller `costs` create larger margins, allowing more misclassifications. Conversely, a larger `cost` creates a narrower margin, permitting fewer misclassifications. The `gamma` parameter defines how far the influence of a single training example reaches, where low values translate to "far" and high values indicate reaches should be "close".

For each of the three kernel types, we fit the training set with all the predictors and converted the target variable to a factor. We also set the probability parameter to `TRUE` in order to calculate the MSE later in the analysis. After training a model, we ran it against the test set, generated a confusion matrix, and calculated both the percent accuracy and MSE. We then refit the training data using 10-fold cross-validation to try different `cost` and `gamma` values with various training set combinations. From the output of each cross-validation, we identified the parameters that constituted the best-performing model and applied those to the test set. Lastly, we produced the confusion matrix, accuracy score, and MSE for this "best" model. These final model statistics are detailed in Table 3.

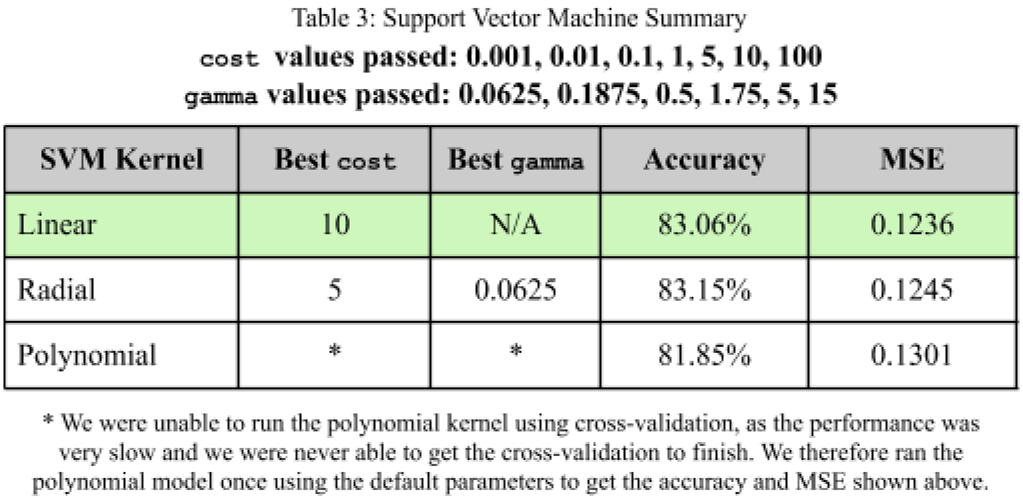


Table 3: Support Vector Machine Summary

`cost` values passed: 0.001, 0.01, 0.1, 1, 5, 10, 100

`gamma` values passed: 0.0625, 0.1875, 0.5, 1.75, 5, 15

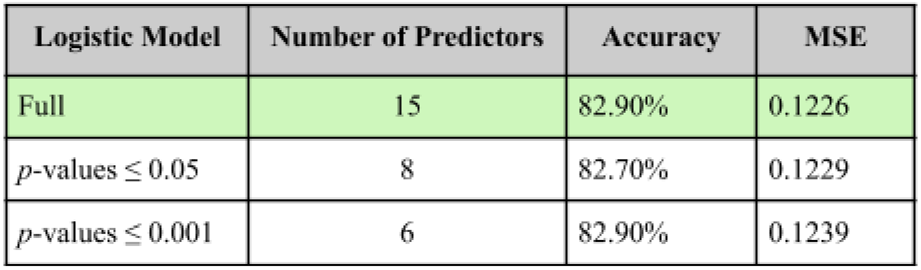
SVM Kernel	Best `cost`	Best `gamma`	Accuracy	MSE
Linear	10	N/A	83.06%	0.1236
Radial	5	0.0625	83.15%	0.1245
Polynomial	*	*	81.85%	0.1301

* We were unable to run the polynomial kernel using cross-validation, as the performance was very slow and we were never able to get the cross-validation to finish. We therefore ran the polynomial model once using the default parameters to get the accuracy and MSE shown above.

The SVM with the lowest MSE was the linear model, but the radial model was just slightly less accurate. Ultimately, there was not much of a difference between the linear and radial models. The overall accuracy of our SVM model was approximately 83%.

For the Logistic Regression technique, we first fit a generalized linear model (glm) on our training data using all 15 predictors, ensuring the `family` parameter was set to `binomial`. As with the previous models, we also converted the target variable to a categorical one using the `factor()` command. The most significant predictors, which all had *p*-values less than or equal to 0.001, were a recruit's national and state ranking, if he played in an offensive or defensive position, his estimated income per household, and his distance from the school. Slightly less significant predictors, with *p*-values less than or equal to 0.05 were his composite rating and if his position was a skilled one. The model revealed that each of these predictors influenced whether a recruit committed to a Power Five school. We initially conjectured that a player's height and weight held statistical significance in regards to the response, but the logistic models negated this notion. Instead, among other things, the results highlighted that increasing a recruit's distance to school by 100 units is associated with an increase in the logarithmic odds of committing to a Power Five school by approximately 0.03143, regardless of the other predictors. The numeric results of each model when applied to the testing data are shown below (Table 4). We determined that, of the three Logistic Regression models, the most accurate was the model that included all 15 predictors, because it had the lowest MSE value.

Table 4: Logistic Regression Summary

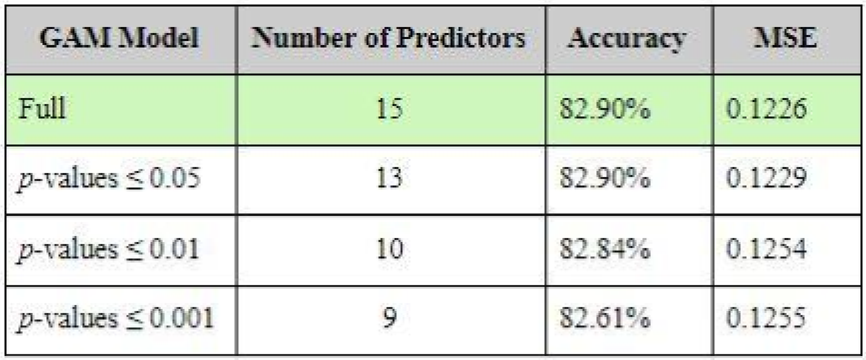


Logistic Model	Number of Predictors	Accuracy	MSE
Full	15	82.90%	0.1226
p-values ≤ 0.05	8	82.70%	0.1229
p-values ≤ 0.001	6	82.90%	0.1239

$$y = \frac{e^{\beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)}}{1 + e^{\beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)}}$$

Figure 5: Classification Equation of the Generalized Additive Model

Lastly, we applied the Generalized Additive Model (GAM) to our dataset. The equation of GAM in the classification setting is shown in Figure 5, where $_0$ is the intercept, f_i is a linear function, and x_i is a predictor. Using the `gam` library, we fit four distinct GAM models to our training data; one included all fifteen predictors, the second only included predictors with *p*-values less than or equal to 0.001, the third only included predictors with *p*-values less than or equal to 0.01, and the fourth only included predictors with *p*-values less than or equal to 0.05. We used the summary of the model fit with the full set of predictors to determine which features to include in the three subsequent models. All attempts required the implementation of the wrapper function `I()` to create a binary response variable as well as the application of the `gam()` function with the `family` parameter equal to `binomial` to specify the error distribution and link function that should be used in each model. We then applied each of these four models to the testing data in order to assess their accuracies. Surprisingly, model accuracy decreased as fewer predictors were included. A table detailing the specifics of each model can be seen in Table 5.



GAM Model	Number of Predictors	Accuracy	MSE
Full	15	82.90%	0.1226
p-values ≤ 0.05	13	82.90%	0.1229
p-values ≤ 0.01	10	82.84%	0.1254
p-values ≤ 0.001	9	82.61%	0.1255

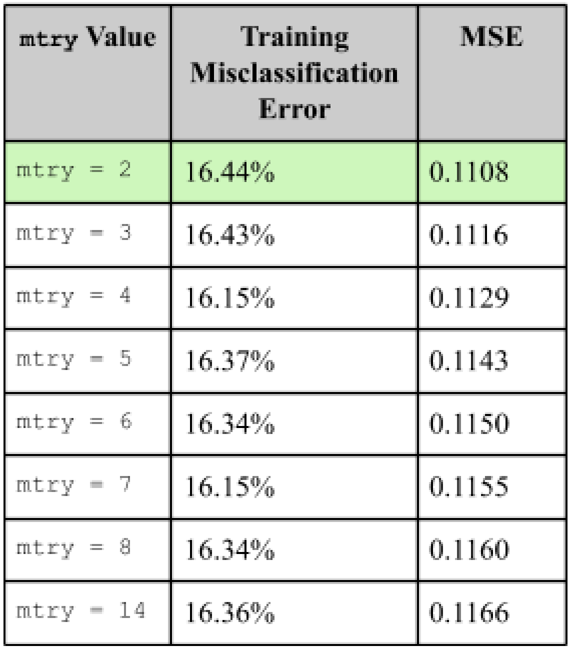
Table 5: GAM Model Numerical Results

Clearly, the full model yielded the best results. Unexpectedly, these numerical results were identical to those obtained with the best Logistic Regression model. Careful analysis of the underlying concepts of these two techniques reveals that Logistic Regression is actually a special case of GAM. Specifically, Logistic Regression occurs when $f_i = \beta_i$. In this instance, the GAM functions used to produce the best fit were identical to the Logistic Regression coefficients. Despite their identical numeric results, the two models differed in their identification of significant predictors. The GAM model expressed that a recruit's star count, composite ranking, and if he plays in a defensive position were the three most statistically significant predictors when determining if a player committed to a Power Five school. Meanwhile, a recruit's weight and whether his position is deemed to be skilled had minimal or no statistical significance on the predictions.

After fitting multiple models using the classification techniques detailed above, we compared the MSEs and model accuracies of the best models for each technique (Classification Tree with Random Forest, Linear SVM, Logistic Regression with all predictors, and GAM with all predictors). As a group, we determined the Random Forest model produced the best results when applied to the testing set. We chose to proceed with a Random Forest model rather than a Bagging model, despite the latter's slightly higher model accuracy percentage, because the former had the lowest MSE of all the models. A Random Forest model splits observations using only the most influential predictor from a subset of the predictors for each internal node and decorrelates its produced trees, enabling it to create the best models. Thus, we fit a Random Forest Classification Tree on the union of the training and testing datasets. It is important to note that, although the Random Forest Model outperformed the three remaining models, its prediction accuracy is higher by less than one percentage. Similarly, the range of MSE values for these four models is a mere 0.0057; hence, the four models' predictive abilities are extremely comparable for this particular dataset.

In an attempt to find the best possible model, we experimented with different values of `mtry` to discover the numerical subset of predictions that produced the lowest MSE when applied to the saved 2019 data. Ultimately, `mtry = 2` yielded the lowest MSE and highest model accuracy percentage of 0.1109 and 84.90%, respectively. Detailed results of all attempts are shown in Table 6. Producing the numerical measures of variable importance along with their respective plots revealed a recruit's national ranking is indisputably the most statistically significant predictor of whether a recruit commits to a school in the Power Five conference. These results support intuition, as the highest ranked athletes are most likely to be recruited by Power Five schools, as they want to create a winning team. Since these Power Five schools have the best teams, they are the most frequented schools by NFL scouts, making these schools highly desired by recruits. Consequently, individuals recruited by Power Five schools are highly likely to sign with them. Conversely, students who are nationally ranked lower are less likely to be recruited by Power Five schools, making it unlikely for them to commit to one of these schools.

Table 6: Summary of Final Model `mtry` Attempts



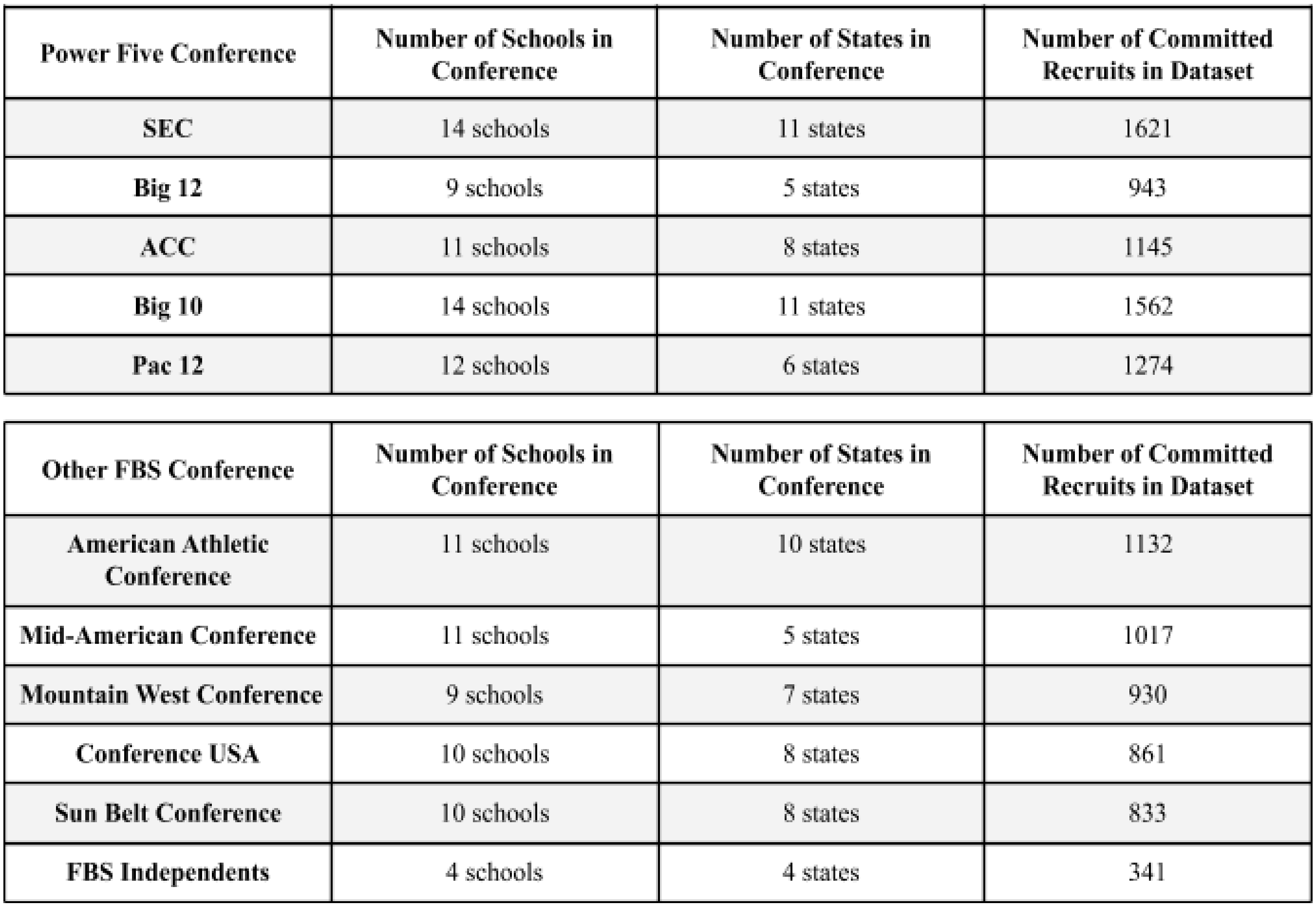
`mtry` Value	Training Misclassification Error	MSE
`mtry = 2`	16.44%	0.1108
`mtry = 3`	16.43%	0.1116
`mtry = 4`	16.15%	0.1129
`mtry = 5`	16.37%	0.1143
`mtry = 6`	16.34%	0.1150
`mtry = 7`	16.15%	0.1155
`mtry = 8`	16.34%	0.1160
`mtry = 14`	16.36%	0.1166

In the future, it would be interesting to take the current 2020 recruits and apply our model to them, predicting if each recruit will or will not attend a Power Five school. After these individuals commit to their respective schools later in the school year, we could review our predictions and assess their accuracies. It would also be interesting to modify our model to predict from the perspective of a particular college, rather than that of a group of 61 schools.

Appendix

NOTE: Any visuals or outputs not included in the above report nor in the appendix are available in the pdf version of the R Markdown file

Table A: FBS Conference Breakdown



Power Five Conference	Number of Schools in Conference	Number of States in Conference	Number of Committed Recruits in Dataset
SEC	14 schools	11 states	1621
Big 12	9 schools	5 states	943
ACC	11 schools	8 states	1145
Big 10	14 schools	11 states	1562
Pac 12	12 schools	6 states	1274

Other FBS Conference	Number of Schools in Conference	Number of States in Conference	Number of Committed Recruits in Dataset
American Athletic Conference	11 schools	10 states	1132
Mid-American Conference	11 schools	5 states	1017
Mountain West Conference	9 schools	7 states	930
Conference USA	10 schools	8 states	861
Sun Belt Conference	10 schools	8 states	833
FBS Independents	4 schools	4 states	341

Table B: Breakdown of Position Groupings

Position Group	Position	Number of Power Five Recruits	Number of Other Recruits	Offensive Position	Defensive Position	Skilled Position
Running Back	Running Back	381	317	Y	N	Y
	All-Purpose Back	40	30	Y	N	N
	Full-Back	28	14	Y	N	N
Quarterback	Dual-Threat Quarterback	146	98	Y	N	Y
	Pro-Style Quarterback	189	163	Y	N	Y
Offensive Line	Center	100	84	Y	N	N
	Guard	397	259	Y	N	N
	Tackle	675	435	Y	N	N
Receiver	Wide Receiver	873	697	Y	N	Y
	Tight End	345	266	Y	N	N
Defensive Back	Safety	557	371	N	Y	N
	Cornerback	704	583	N	Y	Y
Linebacker	Inside Linebacker	321	331	N	Y	N
	Outside Linebacker	497	330	N	Y	N
Defensive Line	Defensive Tackle	506	348	N	Y	N
	Strong-Side Defensive End	355	234	N	Y	N
	Weak-side Defensive End	321	237	N	Y	N
Athlete	Athlete	77	108	Y	Y	N
Special Teams	Place Kicker	63	54	N	N	N
	Punter	45	22	N	N	N
	Long Snapper	36	22	N	N	N
				11	8	5

Table C: Detailed Variable Explanations

cs_power_conf	*categorical (binary)*	1 → the recruit is committed to an FBS school within of the Power Five conferences 0 → the recruit is committed to an FBS school that is NOT in one of the above conferences
height_in	*continuous*	-- the recruit's height (in inches) ; anywhere from 65 to 83 inches, with an average of 74.05
weight	*continuous*	-- the recruit's weight (in pounds) ; anywhere from 65 to 83 inches. 137 to 400 pounds, with an average of 225.6
star_ct	*continuous*	Number of stars assigned to recruit by the 247sports composite rtg and other expert analysts. Ranked from lowest level to highest level: 1 → will likely not contribute to any FBS school and have the lowest NFL draft capability 2 → "limited potential as NFL prospects" and unlikely to contribute to a Power Five program" 3 → potential NFL prospects and have the ability to contribute to a Power Five program 4 → recruits more than likely "produce college careers that get them drafted into the NFL." 5 → "the top 32 players in the country for to mirror the 1st-round picks in the NFL Draft"
composite_rtg	*continuous*	-- The 247Sports Composite Rating is the industry's most comprehensive and unbiased prospect/recruiting ranking and is also used to generate Team Recruiting Rankings. Assigned to each recruit a decimal between 0.001 and 1.0
nat_rk	*continuous*	national ranking by class ; A recruit's national ranking is a number that indicates where he ranks in terms of talent level relative to the rest of the recruits in the country for his graduation class. Between the 5 recruiting classes, the national rankings range anywhere from #1 to #4545.
pos_rk	*continuous*	(national ranking by position per class) ; A national-level player ranking (1 to 577), derived from the national ranking, assigned to each recruit according to the recruit's class and position ;
st_rk	*continuous*	(national ranking by high school state per class) ; A state-level player ranking (1 to 577), derived from the national ranking, assigned to each recruit according to the US state of the recruit's high school (*Note: not his hometown)*.
incomeperhousehold	*continuous*	Average income of a family in the recruit's hometown.
averagehousevalue	*continuous*	Average value of a house in the recruit's hometown.
distance_to_school	*continuous*	Distance in miles from the recruit's hometown to the college to which they committed.
cs_in_homestate	*categorical (binary)*	1 → recruit is committed to an FBS school located within his home state 0 → recruit is committed to an FBS school located outside of his home state
state_hotbed	*categorical (binary)*	1 → recruit home state is either Texas, California, Georgia, or Florida 0 → recruit's home state is not one of the above
pos_off	*categorical (binary)*	1 → recruit's position is one of the 10 offensive positions or is listed as an "Athlete" 0 → recruit's position is not one of the above positions
pos_def	*categorical (binary)*	1 → recruit's position is one of the 7 defensive positions or is listed as an "Athlete" 0 → recruit's position is not one of the above
pos_skilled	*categorical (binary)*	1 → recruit's position is one of 5 possible "skilled" positions 0 → recruit's position is not one of the above