Academic Test Scores in NCAA Sports

Ethan Brown, Lucas Czarnecki, Grant Ripperda, Liam Waterbury, Harrison Wight
January 27, 2022

Abstract

We present an analysis of the factors that affect the Academic Progress Rate (APR) in NCAA programs. In particular, we investigate the correlation between academic and athletic performances. A predictive model was generated that has an accuracy ~**30%** more than the baseline. We find that female athletes have been consistently outperforming their male counterparts by an average of **16.13** APR points, but that gap has begun to lessen in recent years to **12** APR points.

1 Introduction

Our goals are to compare different sports and genders to see whether there are any trends in the athletic and scholastic performance in Division-I universities. We also want to explore the relationship between academic performance and athletic performance to see if we can reasonably predict a program's athletic ranking based on their academic performance. Our report focuses on the National Collegiate Athletic Association (NCAA) and the academic performances of each of its member colleges, using the Academic Progress Rate (APR), which is a score that measures a team's overall eligibility, graduation, and retention rate. The maximum APR is 100 points. We obtained our data through two different means; we used Kaggle (NCAA) to find a dataset containing various statistics such as the university, conference, and APR, and coupled this information with the rankings of all power 5 conference teams in football and basketball.

2 Data Preparation

Dataset

We obtained our dataset from Kaggle, which is linked below. The data set tracks all Division-I athletic teams and their respective APR performance. The dataset also tracks the team's four-year retention rate, four-year eligibility rate, and number of athletes. The data was obtained originally from NCAA reports.

General Data Cleaning and Preparation

Most of our dataset had good formatting and did not need a lot of cleaning. A few schools did not have APR data for specific years, which were shown with "-99" values. We replaced these with "NaN" in order to not include them in the reporting.

Web Scraping

An important aspect of our project is comparing academic performances among athletes. In order to compare athletic performance to academic performance, we decided to use a schools' conference rankings along with its APR. Our team scraped data from sports-reference.com for conference rankings in football and basketball.

Merging

We decided to merge this scraped rankings data set with our original dataset on the school names, which required a bit of formatting, as the rankings source had school names listed differently than the testing scores dataset. Doing this allowed us to compare the school's athletic performance with its academic performance.

Predict Football Ranking Based on Test Scores Data Preparation

Data was filtered down to just the football entries. Then, irrelevant columns such as SCHOOL_ID, SCHOOL_TYPE, etc. were all dropped. The original dataset had separate columns for each year and metric. This made it much harder to make a predictive model since teams perform differently each year. To fix this, the data frame was transformed so that the instead of there being a column for each year, there is now a row for each year entry.

Finally, the rankings were binned into "Good" and "Bad" buckets. This was done by grouping the data frame based on the year and conference so we know if it was a good ranking for that specific conference. It is important to group by conference and year because if for some reason there were 3 teams in a conference, the third place team would be considered "Bad" because they were in last place.

Analysis of Men's vs Women's Performance Data Preparation

The dataset did not already track whether a sport was Male or Female. Therefore, an attribute was added based on whether the sport name included "Men" or "Women". In addition, mixed sports were filtered out so as to not skew the data. Furthermore, any schools that only had sports for one gender were removed and schools with less than 6 sports were removed so that an outlying high performing sport wouldn't skew the data. Then, all of the female and male scores were averaged together so that each university contained an average score for their Female athletes and an average score for their Male athletes. Finally, the difference in the gender scores was assigned to each university using an aggregation function.

3 Methods

Predict Football Ranking Based on Test Scores Methodology

To predict the football rankings based on test scores, a data frame, as described in the data preparation phase with each year being a separate entry, was used. We did this in order to get the most data for the decision tree and naive Bayes models. It was also important that the RANKING column was cut so we could group the results into a "good season" and a "bad season" (See reference A1).

Both a decision tree and naive Bayes model were generated for the football ranking predictions. For each type of model, cross validation was used to determine the best hyper parameters for the respective models. The hyperparameter with the highest validation score was chosen.

For the decision tree, the hyperparameter range for max depth was between 1 and 10 inclusive. The best hyper parameter for decision tree was a max depth of 3.

For the naive Bayes model, the hyperparameter range for pseudo-counts was between 0.1 and 1000 inclusive. The best hyperparameter for the naive Bayes model was alpha=0.1. The value of alpha did not affect the validation accuracy when cross validating, meaning the validation curve was flat.

The best hyperparameters were used to generate the final models. In addition to cross-validation, a train/test split of **80%** training was used as an additional check against overfitting the data.

Analysis of Men's vs Women's Performance Methodology

Instead of comparing Men's and Women's performances directly, their performances were only compared within the respective university. This was in an effort to remove factors such as a university's relative academic prowess.

The key method was finding this difference in average performance and then aggregating these values based on the group that was desired (either Conference or University).

Initially, just the difference in performance was desired; however, this method was evolved to also determine the difference in performances over the tracked time.

4 Results

Predict Football Ranking Based on Test Scores Results

The two models performed similarly, with the decision tree having a score of **69.8%** when tested on the test data. The naive Bayes model had a slightly better score of **70.8%**. Both are similar enough to explore the results of each.

The decision tree was used using a depth 3 tree. The cross validation results shows that the decision tree performs much better than the baseline (**39.5%**) as shown in this visualization.

The decision tree's results show that the SCORE (APR score) is the most important attribute when determining whether a team will have a "good" or "bad" season.

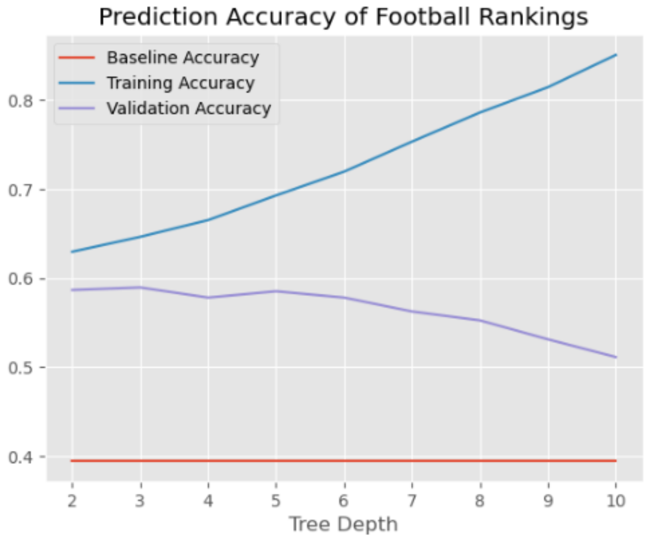


Fig. 2 - Football Rankings Decision Tree Cross Validation Results

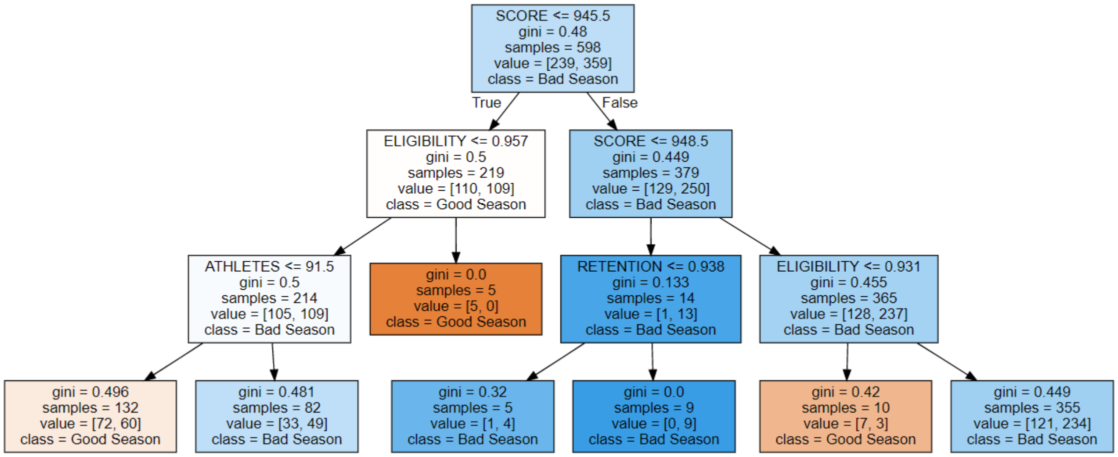


Fig. 3 - Resulting decision tree for predicting football ranking from a team's academic test score, number of athletes, eligibility percentage and athlete retention percentage

The naive Bayes model shows that the alpha hyper parameter was not an important hyper parameter setting. The validation score after cross validating was the same, no matter what the value of alpha was.

Overall, the predictive models performed well in predicting the football rankings based on the data from our initial data set. However, it seems that there are likely more factors in predicting whether a team has a good season or not. Academic performance is able to make a good amount of accurate predictions, but it seems that academics alone are not enough to predict the type of season a college football team will have.

Analysis of Men's vs Women's Performance Results

Using the 2014 scores, the women outperformed the men at their university by an average of 12 APR points. In addition, **81.5%** of the universities saw their women's athletes outperform the men in terms of their APR test score.

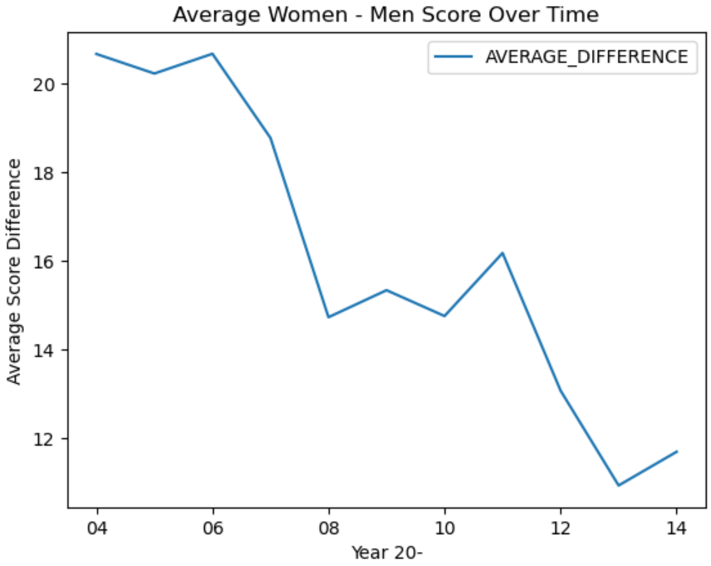


Fig. 4 - Gender Score Difference Over Time

This only represents a snapshot of the most recent year. Therefore, the same analysis was performed on the average values for the schools over the 2004-2014 period. After cleaning out universities with missing scores in the time period, the average difference between male and female scores was 16.13 and **98%** of universities had their women's sports have higher scores than the men.

This stark difference between the average and the most recent scores proposed the question, is the difference between the male and female athletes changing over time?

As the above graphic shows, the difference between the male and female scores is decreasing over time. Therefore, one could expect NCAA athlete academic performance to continue to become more similar between sexes in the future.

However, this will likely take a considerable amount of time as only 18 of 360 scores saw the men outperform the women in the majority of the years between 2004-2014.

Throughout the investigation, careful analysis was performed so as not to skew the data with higher performing teams. Therefore, the question of whether a university's overall APR score is correlated with the difference between men and women performance was proposed.

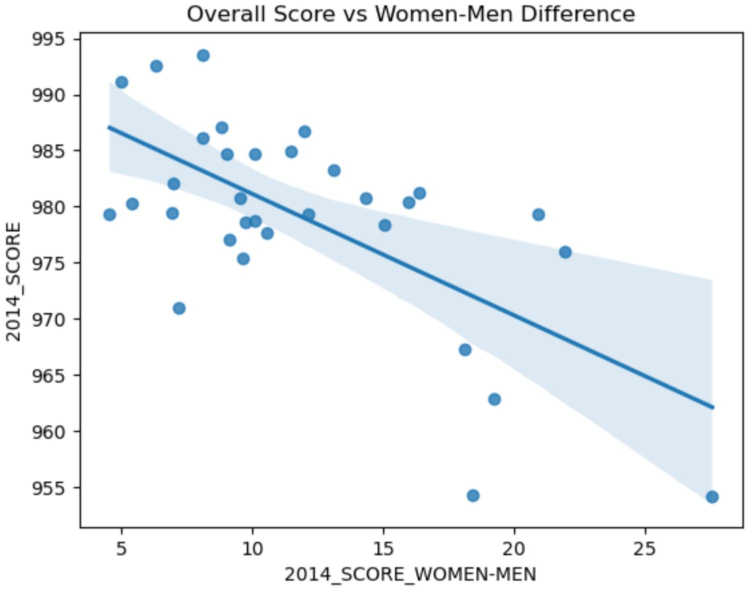


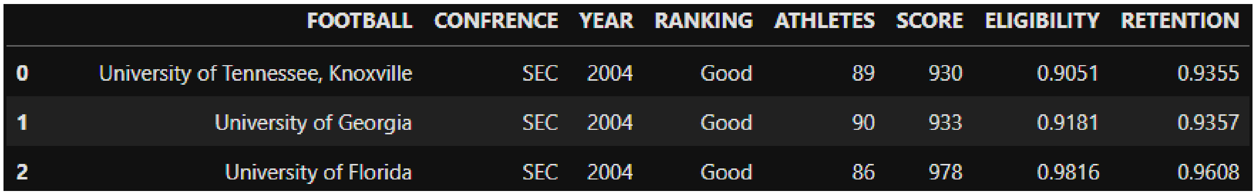
Fig. 5 - Correlation of Overall Score and Gender Difference

Figure 5 shows that there is a correlation between a university's overall score and the difference between men and women performance. The higher the overall score, the less the difference between the performance in the genders.

5 Conclusion

The increase to an accuracy of **70%** from the baseline of **39.5%** shows that the dataset may be appropriate for making predictions for the success of football teams using their APR scores. However, there are external factors that are not accounted for in the dataset that most likely will have an effect on ranking, such as a school's athletic prowess, coaching and recruiting. In addition, it is clear that the female athletes outperform the men consistently across all schools. There is also a strong trend showing that the difference in academic performance is decreasing over time. This may suggest that male and female athletes are beginning to perform more similarly in the academic scene.

References



	FOOTBALL	CONFRENCE	YEAR	RANKING	ATHLETES	SCORE	ELIGIBILITY	RETENTION
0	University of Tennessee, Knoxville	SEC	2004	Good	89	930	0.9051	0.9355
1	University of Georgia	SEC	2004	Good	90	933	0.9181	0.9357
2	University of Florida	SEC	2004	Good	86	978	0.9816	0.9608

A1 - Data Frame used for the predicting football RANKING based on ATHLETES, SCORE, ELIGIBILITY and RETENTION

Gallini, Jeff. "College Football Team Stats Seasons 2013 to 2022." Kaggle, 3 Feb. 2023, https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019?select=cfb13.csv.

NCAA. "Academic Scores for NCAA Athletic Programs." Kaggle, 16 Feb. 2017, https://www.kaggle.com/datasets/ncaa/academic-scores?resource=download.

Sundberg, Andrew. "College Basketball Dataset." Kaggle, 16 Mar. 2021, https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset.

"College Football Stats, History, Scores, Standings, Schedule & Records: College Football at Sports." *Reference.com*, https://www.sports-reference.com/cfb/.