ISYE 6740 - Summer 2020 - Final Report
Team 206 - The Kick and the Pick

Alex Song (asong49), Mo Berro* (mberro3), Nikki Cross (bcross3)

July 2020

Problem Statement

Football is well-known to be a team sport, requiring the skills of a wide variety of players and careful coordination of all teammates. The ultimate goal of the team is to win as many games as possible, hopefully becoming the top team in all of college football. But how do the stats associated with each player contribute to the odds of reaching the top 25 by the end of the year?

In this project, we dive into how different players' stats predict the ranking a team will have at the end of the season. Can we forecast a top 25 finish with just the quarterback's rating or perhaps a few offensive statistics? Or, as is the case on the field, will we need the full team's effort to accurately predict the outcome for the season?

This problem is one of interest and significance for a number of reasons. For one, it was an intellectual curiosity for the group, simply to determine if we could. We all like the sport and picking who'll do the best is always the subject of much discussion. There is additionally a market for this among gamblers - being able to model the likelihood of a team performing well could be a lucrative opportunity in Vegas. Finally, with all the crazy things going on in 2020 and the very real possibility there won't be a 2020 football season or it will be a very unusual one, it was nice to have a reason to think about more normal times and relive some fun football memories of past years.

Data Source

The metrics used in this analysis are the result of significant screen scraping from espn.com. We pulled statistics for 5 key players from all NCAA FCS division football teams across the 2018 and 2019 seasons, along with the team's conference and whether they finished in the AP Top 25 that year. The player stats used were for the players with the most:

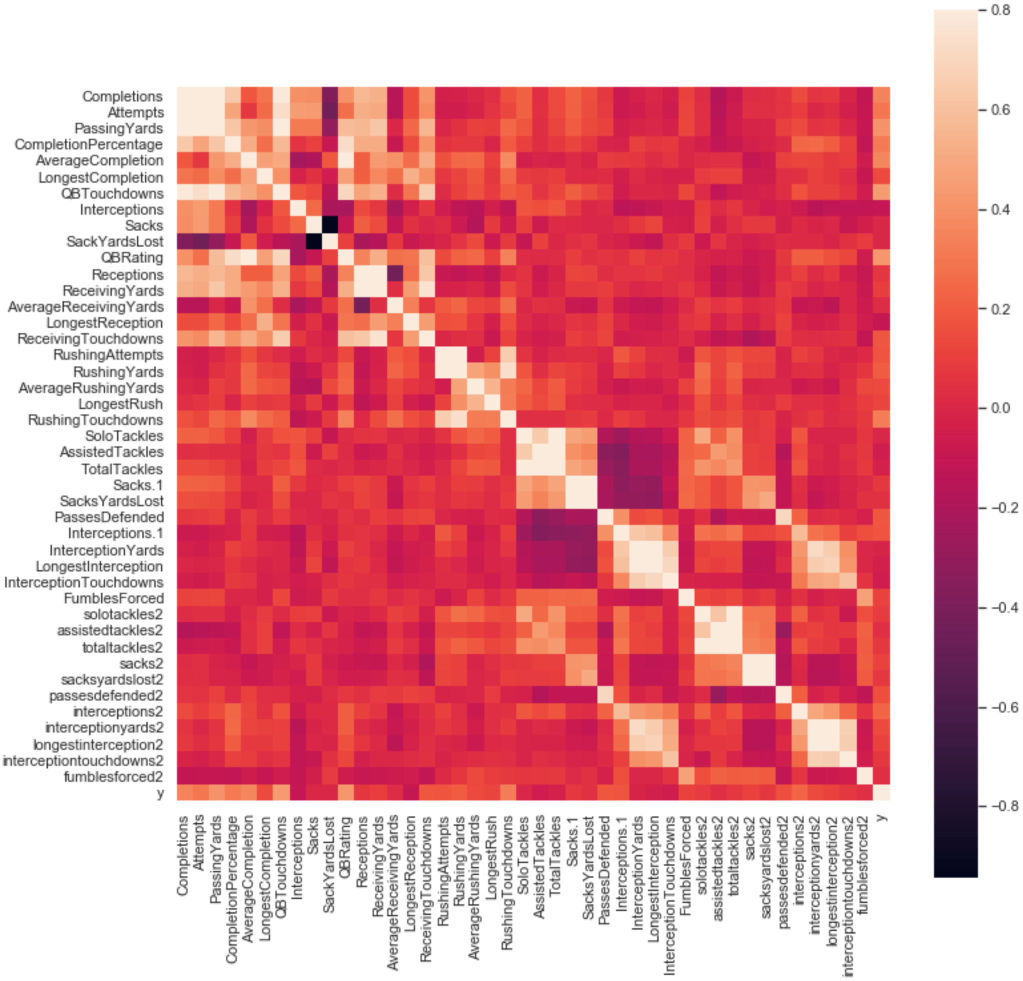
- passing yards
- receiving yards
- rushing yards
- tackles

- interceptions

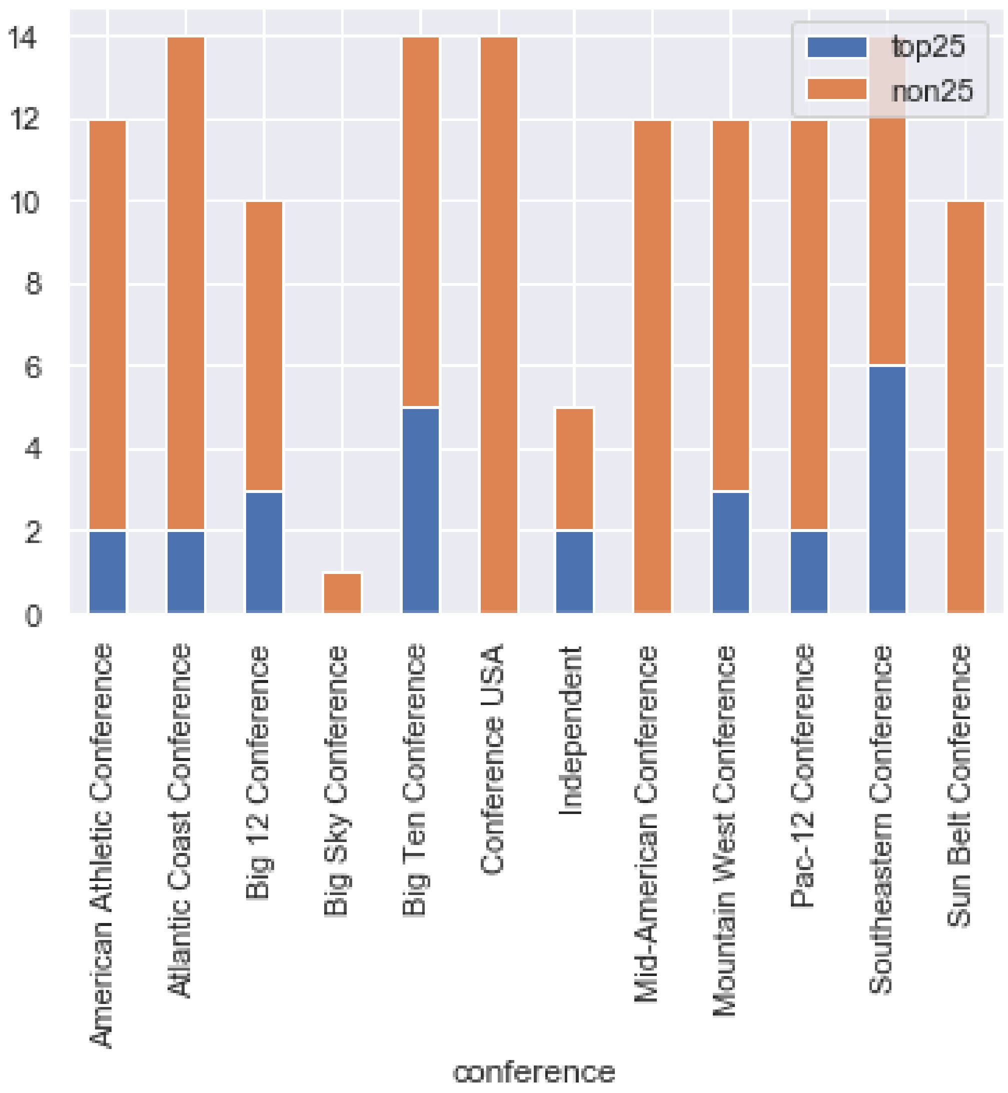
This yielded a data set of approximately 130 teams (teams do occasionally come into or move out of the division) and 44 predictors with which to predict the season's final outcome.

Exploratory Data Analysis

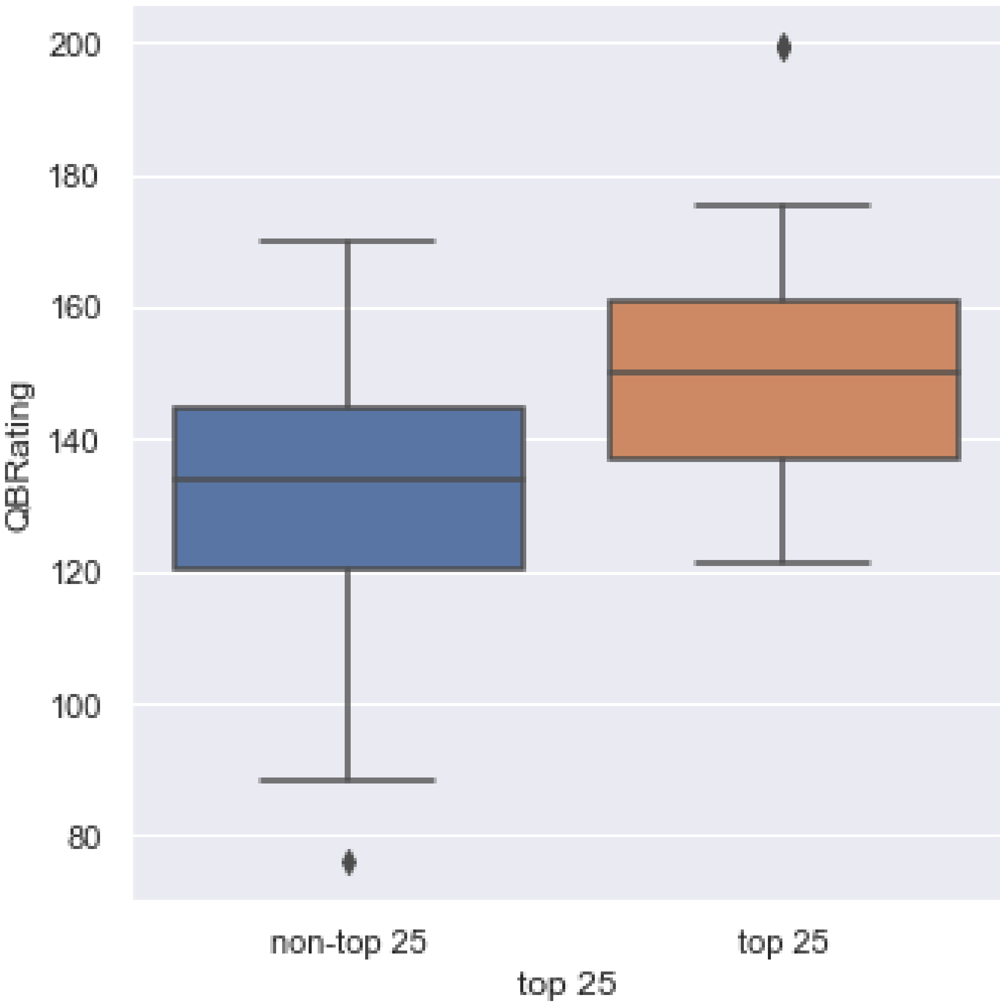
Note: While we reviewed a wide variety of feature traits, we have limited their presentation here for brevity and clarity. We present a sample of our research and items of note, rather than an exhaustive list.
We started data exploration with a heat map to see the correlation of the predictors with each other and the target field. While much of the broad heat map is true red (indicating nearly 0 correlation), there is a distinct pattern of squares along the diagonal, indicating significant correlation between fields. This tends to indicate a single player's statistics, which are highly correlated with each other. Additionally, the top left contains a larger square, where the quarterback and primary receiver's statistics show high levels of correlation. Finally there are two diagonal patterns above and below the center diagonal in the bottom right of the chart. These are high correlations in the same stats between the two defensive stats leaders evaluated; in some cases, these could be the same player, leading to high correlation.



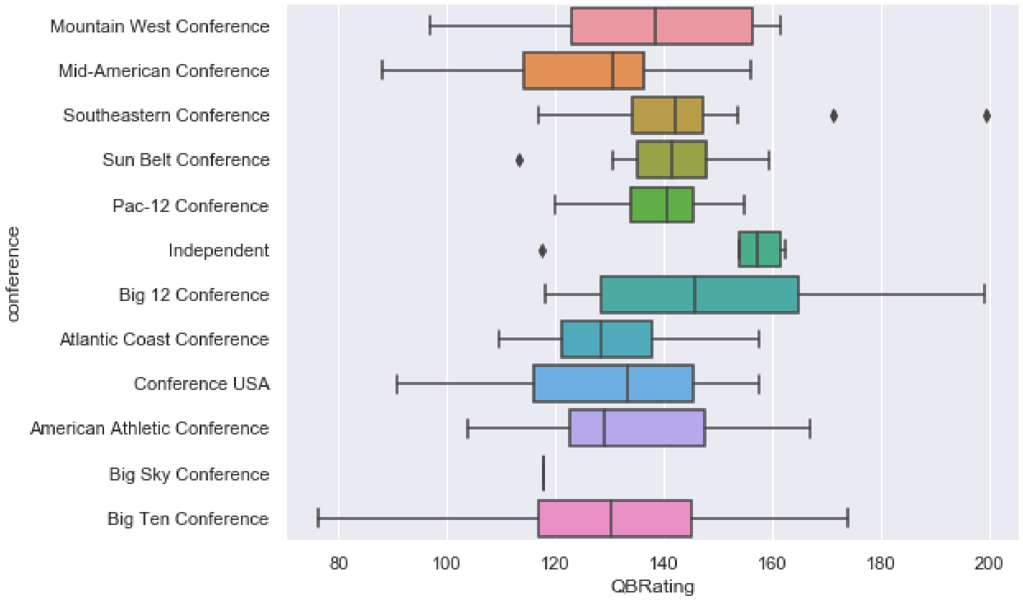
Conference is an important consideration in the team's strength of schedule and can be a considerable factor in post-season bowl assignments; it is the sole categorical field in our data set. You can see that the Big Ten and the SEC were particularly strong in this year, accounting for 11 of the top 25 schools.



It quickly became evident that QBrating, itself an aggregated score based on other factors, was a nice summary field and quite predictive. In the box plots below, you'll see the 25th percentile QB rating for a top 25 team was still higher than the 50th percentile for a non-top 25 team, showing the differentiation strength in this field.



To understand the combined strength of the QB rating and conference, we have done a combined box plot. When compared with the team's conference, we see the wide variation in skill level that some conferences have. The Big 12 and Big Ten, for example, have nearly 100 point ranges for the rating, while Pac-12 and Sun Belt Conferences are quite compact.



Methodology

We applied a variety of techniques to the data set to determine which was the best differentiator of our 0/1 target, inclusion in the AP Top 25. Modeling techniques attempted include:

- adaboost
- decision tree
- k nearest neighbors clustering
- logistic regression
- naive Bayes algorithm
- neural network
- random forest
- support vector machine

We set the data up such that the 2018 season was the training data set and the 2019 was our out of time validation. Predictive accuracy shown below is on the 2019 predictions, based on models created with the prior season.
One point that we would like to make clear - we are examining correlation and not implying causation in this analysis. The statistics credited to any given player are not achieved by that player alone; a quarterback cannot put up large passing yardage numbers without an effective offensive line blocking or a wide receiver who can get open and make the play. As such, we seek to understand the ability of these statistics to predict the team's ranking, but are not suggesting that, absent the remaining team members, these metrics and therefore the team outcomes are possible.

Evaluation and Final Results

When the model of the 2018 season is used to predict the top 25 finishers of 2019, we find that the various techniques tested had 79-87% accuracy on the test sample. These results are reasonable, but not particularly strong. This shows that the individual stats are insufficient to generate the predictive power we would like.

Technique	Test Data Accuracy Rate
adaboost	0.817
decision tree	0.832
k nearest neighbors clustering	0.855
logistic regression	0.817
naive Bayes algorithm	0.786
neural network	0.809
random forest	0.870
support vector machine	0.817

While the existing data is relatively unbiased (it contains all teams from the 2018 season, but we have not tested if the season itself could be biased in a meaningful way), it is unclear if the data itself is sufficient to generate the results we desire. If we were to attempt to further improve the model, there are a number of things we could try to make a more robust training sample:

- Increase sample size: It's likely that increasing the number of seasons used in the training would result in a more robust model being created.

- Increase the breadth of data: By including other team-level statistics, we could build a stronger model. This might include strength of schedule statistics, win-loss records, and perhaps some information on coaching staff, for example.

- Increase the depth of data: While we currently have statistics on 5 players within each team, we could certainly pull some stats on the rest of the team and likely create additional leverage for the model.

While this project was largely one of research and team interest, there are few real-world applications. This could certainly be used to inform discussions with friends or perhaps gambling wagers, but the statistics used are for the full season. Predicting how the final results will appear after the entire season has been played is far less interesting than attempting to do so mid- or even pre-season.

In conclusion, we were able to construct several reasonable but not incredibly powerful predictors of the top 25 finishers for a football season. Most techniques performed similarly, with random forest having the strongest out-of-time validation with 87.0% accuracy. Several data expansion recommendations would likely improve the strength of the model and could be considered for future testing.

Team Composition

Team Member	Primary Project Contributions
Alex Song	Project ideation, EDA, modeling, evaluation
Mo Berro*	Project ideation, data collection
Nikki Cross	Project ideation, documentation, modeling, evaluation

*Mo dropped this course, but we want to ensure he is acknowledged for his contributions to the project.