# College Football

## How much talent do you need to recruit to win a national championship?

**Andreea Serban & Chris McAllister**

Masters of Applied Data Science,
Milestone I
Winter 2024

### Introduction:

Over the past few years, college football has quietly become a multi-billion dollar industry. Conferences are signing massive TV deals, coaches are acquiring record-breaking contracts, and student-athletes are now earning millions of dollars [3].

However, as the money increases so do the expectations for players, coaches, and athletic directors. Success and championships have never been more important for student-athletes and the universities they represent.

There are hundreds of D1 college football teams, but only a handful of universities have won a championship in the past few years. One key aspect of college football is recruiting; convincing the most talented high school players to attend your university is arguably the most important part about being a head coach. It's also expensive. **Choosing the right players to sign (and avoiding the wrong ones) can save universities millions of dollars.**

### Objective:

In this project, we're going to analyze the last 9 years of college football games and recruiting data to find out what type of players schools need in order to compete at the top of the sport.

### Questions:

To answer the question above, we will look into the following:

1. How should "success" be measured in college football?
2. Does building a talented roster matter?
   a. What's the minimum talent threshold required to compete at the top?
   b. Is team success correlated with the proportion of players who get drafted to the NFL?
   c. Are high school rankings a good indicator of an athlete's performance?
3. Where do the best recruits come from and where do they go to school?

[3] New York Times, College Athletes May Earn Millions from Their Fame

# A new success metric: ELO Rating

## Context

There's no shortage of ways to measure a college football team's success. Because there are so many teams, all of whom play vastly different calibers of schedules, there is a large need to evaluate teams through some type of data driven approach. For example, if Michigan finishes its regular season 12-0 after playing a very difficult Big Ten schedule, that would be much more impressive than someone like Eastern Michigan having the same record over a much easier schedule.

Our solution to that is something called an ELO rating. The ELO rating system was designed by a physicist named Arpad Elo[2]. The original intent was to create a way to rank chess players, but the same algorithm can be applied to anything where two parties compete against each other.

The idea is simple: if you beat a player you should beat, then you add very few points to your rating. If you lose to a player you should beat, then you will lose a lot of points. The majority of 'players' will have a score between 1,000 - 2,000 (see figures 1a and 1b).

We will be exploring the effectiveness of this metrics within our analysis to understand if we can effectively use this as a measurement of a team's success. **Simpler metrics like total victories and point differential don't take into consideration an opponent's skill.**

The ELO rating system is maybe the most trusted open source algorithm for rating competitors while considering the skill of their opponents. But we will **trust but verify** this within our dataset as well.

**Figure 1a:** Traditional Distribution of Elo Rating in Chess
Reference 1 - bkgm.com

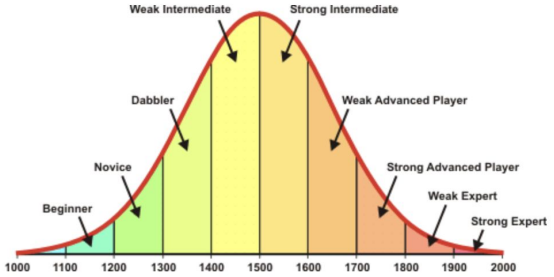
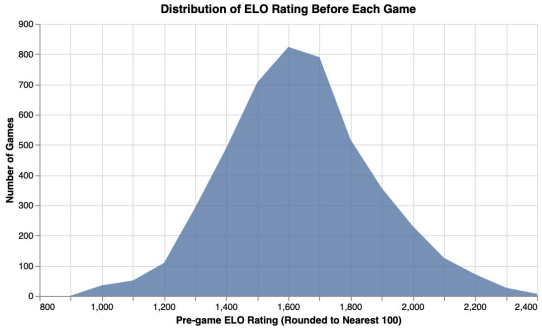



**Figure 1b:** Distribution of ELO Rating before Each Game




[2] The Proposed UCSF Rating System, Arpad Elo

# The Data

## Source/Extract/Combining

There are two data sources that we leveraged to help us understand what it takes for a college team to be successful:

### College Football Data (CFBD)

Source: https://api.collegefootballdata.com/api/docs/?url=/api-docs.json

This source contains the following data:

1. **College Game Outcomes and Statistics:** Every college football game played from the 2015 - 2023 seasons.
    a. Each game will produce two records: one from each team's perspective.
2. **Team Talent:** Overall Talent rating of a college football team.
3. **Team Roster:** Every college football player and the school they played for.
4. **High School Recruiting:** every high school player's recruiting rating and the college they committed to.

We used the College Football Data (cfbd) library and an API key from the site to get this data, and then saved the data locally into multiple CSVs so we can then access the files quickly and easily.

### National Football League (NFL) Draft Data

Source: https://www.pro-football-reference.com/years/2022/draft.htm

Every draft pick from 2015 - 2023. We extract the key features: (1) the player's name, (2) the College Team that player came from and (3) the year they got drafted.

**Figure 2:** Preview of Draft Dataset

| | Player | team | year |
|---|---|---|---|
| 0 | Justin Johnson | Fresno State | 2015 |
| 1 | Jamaal Jackson | Delaware State | 2015 |
| 2 | Johnny Jackson | Arizona | 2015 |
| 3 | Jason Johnson | Western Kentucky | 2015 |
| 4 | Nicholas Peoples | Grambling | 2015 |

We downloaded the dataset into a local CSV location.

### Combining the datasets

We joined the CFBD data sources tables within each of the other CFBD by the college name. We also joined the CFBD with the Draft Data by leveraging the player's name, along with some deduplication logic detailed on slide 5.

# The Data

## Data Manipulation - Data Cleaning - Which ELO rating to use?

Before diving into the analysis, we performed some Exploratory Data Analysis (EDA). We were particularly interested in exploring the fields that we hope to be using in the analysis; specifically the ELO rating.

There is a different **ELO rating for each of the 12 games a team played in a season.** This is expected since a team's rating will change as we learn more about a team throughout the season. Ideally, to measure the success of a team as a whole, we would like to have just one ELO rating for each team in each season. So we had to make a decision on which game's ELO rating within a team's season we should use.

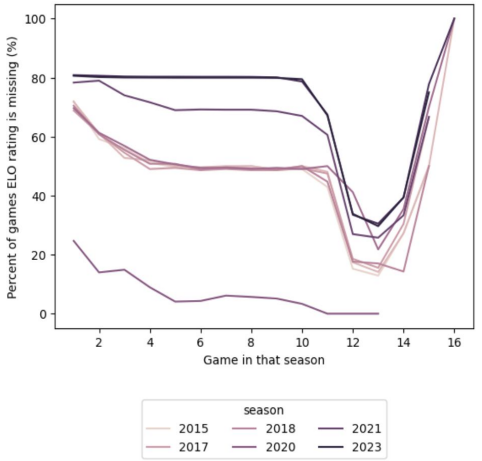
As we stated before, we care about two things for this metric:

1. Which game gives the best overall representation of a team's success in that season:
    a. After some initial analysis, we learned that the later games give a better representation of a team's overall performance.
2. We also care about the data completeness (ie exploring missing values in the ELO column) and making sure that we can compare ELO rating across all different teams.

In our initial EDA, we realized that our data is more likely to be complete for games later in the season. As we can see in Figure 3, throughout the seasons, we can see a high percentage of teams missing the ELO rating up until the 12th game.

This makes sense since this is typically the last game of the regular season and the team's ratings are then used to evaluate their position in the whole league. Due to this finding, we decided to leverage every team's 12th game's ELO rating as our primary success metric for a given team.

**Figure 3:** ELO missing by game in that season



# Data Manipulation:

## Joining datasets on non-unique identifiers

To get the correlation between a player's recruiting rating and their draft likelihood, we had to join the recruiting data to the draft data. These came from two completely different sources and therefore had no joining key. Instead, we decided to use the player's name as the key.

Unfortunately, this created a many-to-many relationship (since it was common for multiple players to share the same name in both datasets - see **Figure 4** below). We tried creating a unique identifier of sorts by joining on the player's name *and* the university they went to since both of these columns were available in the two datasets.

Unfortunately, the join was complicated because the two sources did not use the same naming convention for the university that the player committed / was drafted from. For example, the naming convention for Mississippi State University was denoted as "Mississippi St." in the recruiting dataset, and as "Mississippi State" in the draft dataset (again, see **Figure 4**).

To filter down the output of the many-to-many join, we created a column called "CommonSequence" that returns a 1 if the university column in the recruiting dataset and draft dataset share at least 4 sequential characters in common (see **Figure 5, index #19 for this flag in action**). Lastly, to de-duplicate, we created a column called "Duplicate_Count" that counts the number of times that name and rating have appeared (basically creating a unique ID from these two fields - see **Figure 5**). If the draft school matches the commitment school we prioritize that record as 1 in Duplicate_Count (see **Figure 6** - the correct Chris Jones, Miss St. record was chosen.)

**Figure 4:**
Start with recruit.csv and draft.csv datasets

Combine datasets by joining on name in many-to-many relationship

**Figure 5:**
Exploded dataset where every combination of recruit and draftee name is in the same dataframe.

De-duplicate using the "CommonSequence" and "Duplicate_Count" columns.

**Figure 6:**
Final output is a table containing their rating and an "IsDrafted" flag, which is the CommonSequence column renamed

Two distinct Chris Jones' in our draft dataset, only of whom appears in the recruiting dataset:

| | Player | draft_year | College/Univ | Rnd | Pick |
|---|---|---|---|---|---|
| 202 | Chris Jones | 2013 | Bowling Green | 6 | 198 |
| 821 | Chris Jones | 2016 | Mississippi St. | 2 | 37 |

Six distinct recruits named Chris Jones in our recruits dataset:

| | id | name | year | committed_to | latitude | longitude | rating |
|---|---|---|---|---|---|---|---|
| 18 | 24974 | Chris Jones | 2013 | Mississippi State | 33.898446 | -88.999227 | 0.9912 |
| 3139 | 28166 | Chris Jones | 2013 | Coastal Carolina | 39.983162 | -75.823836 | 0.7398 |
| 4100 | 28805 | Chris Jones | 2014 | Wisconsin | 38.952944 | -76.940865 | 0.8603 |
| 4847 | 29515 | Chris Jones | 2014 | Nebraska | 30.332184 | -81.655651 | 0.8233 |
| 6800 | 31219 | Chris Jones | 2014 | UTEP | 32.091299 | -96.464682 | 0.7441 |
| 28463 | 62478 | Chris Jones | 2020 | Florida Atlantic | 26.237860 | -80.124767 | 0.8572 |

| | name | rating | stars | committed_to | Pick | draft_year | College/Univ | CommonSequence | Duplicate_Count |
|---|---|---|---|---|---|---|---|---|---|
| 3141 | Chris Jones | 0.7398 | 2 | Coastal Carolina | 198 | 2013.0 | Bowling Green | 0 | 1 |
| 3142 | Chris Jones | 0.7398 | 2 | Coastal Carolina | 37 | 2016.0 | Mississippi St. | 0 | 2 |
| 6809 | Chris Jones | 0.7441 | 2 | UTEP | 198 | 2013.0 | Bowling Green | 0 | 1 |
| 6810 | Chris Jones | 0.7441 | 2 | UTEP | 37 | 2016.0 | Mississippi St. | 0 | 2 |
| 4853 | Chris Jones | 0.8233 | 3 | Nebraska | 198 | 2013.0 | Bowling Green | 0 | 1 |
| 4854 | Chris Jones | 0.8233 | 3 | Nebraska | 37 | 2016.0 | Mississippi St. | 0 | 2 |
| 28482 | Chris Jones | 0.8572 | 3 | Florida Atlantic | 198 | 2013.0 | Bowling Green | 0 | 1 |
| 28483 | Chris Jones | 0.8572 | 3 | Florida Atlantic | 37 | 2016.0 | Mississippi St. | 0 | 2 |
| 4105 | Chris Jones | 0.8603 | 3 | Wisconsin | 198 | 2013.0 | Bowling Green | 0 | 1 |
| 4106 | Chris Jones | 0.8603 | 3 | Wisconsin | 37 | 2016.0 | Mississippi St. | 0 | 2 |
| 19 | Chris Jones | 0.9912 | 5 | Mississippi State | 37 | 2016.0 | Mississippi St. | 1 | 1 |
| 18 | Chris Jones | 0.9912 | 5 | Mississippi State | 198 | 2013.0 | Bowling Green | 0 | 2 |

| | name | rating | committed_to | IsDrafted |
|---|---|---|---|---|
| 3141 | Chris Jones | 0.7398 | Coastal Carolina | 0 |
| 6809 | Chris Jones | 0.7441 | UTEP | 0 |
| 4853 | Chris Jones | 0.8233 | Nebraska | 0 |
| 28482 | Chris Jones | 0.8572 | Florida Atlantic | 0 |
| 4105 | Chris Jones | 0.8603 | Wisconsin | 0 |
| 19 | Chris Jones | 0.9912 | Mississippi State | 1 |

# How effective of a metric is the ELO Rating in our dataset?

## Analysis

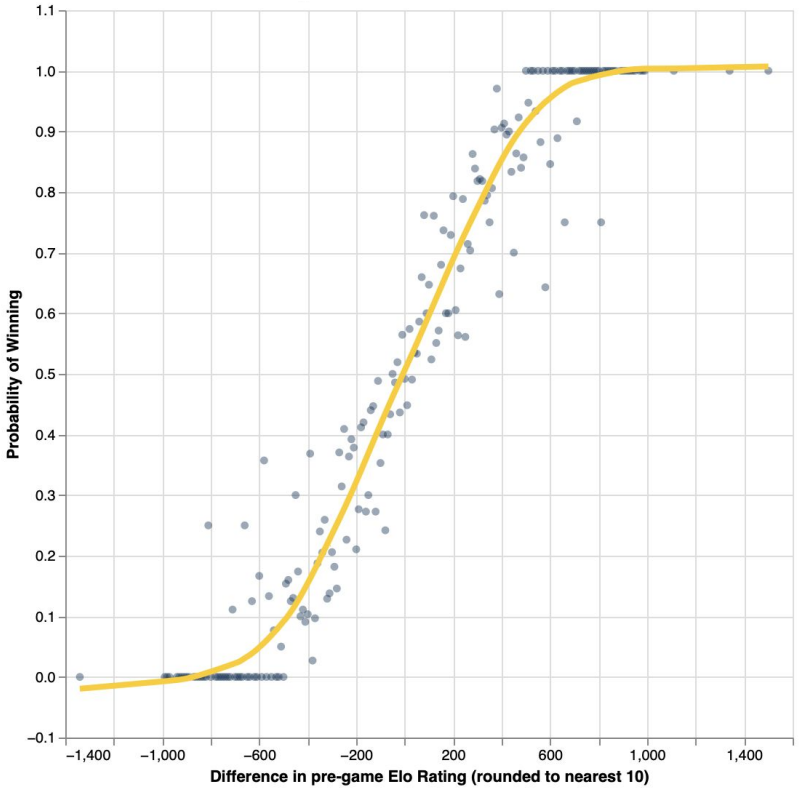
To understand whether the ELO rating system is a reliable metric to quantify a college football team's skill, we analyzed the past 9 years of college football games to see if a team's ELO rating was a good predictor of likelihood to win.

Specifically, we calculated the difference in pre-game ELO rating (on the x-axis), and then plotted the win percentage for each rating difference (on the y-axis).

As you can see from the chart on the right, there is a clear correlation between difference in pre-game ELO rating and probability of winning.

We leveraged all the game's with a pre-game ELO rating for this analysis; however, going forward, we'll restrict usage of the ELO rating to the 12th game since we do not need all the ELO ratings for each game, but rather just one to evaluate a team's effectiveness. As a refresher on how we got to this 12th game decision, we've gone more in depth in slide 5.

**Figure 7:** Pre Game ELO Rating vs. % Chance to Win
(Includes all College Football Games after Week 7 from 2013-2023)



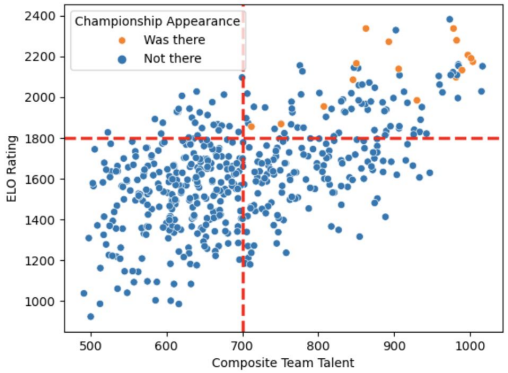
## What's the minimum talent threshold required to compete at the top?

**Analysis**

After confirming the ELO rating validity, we charted Team Talent rating against ELO rating to uncover their relationship and pinpoint a potential minimum threshold for national championship contention. Figure 8 illustrates a moderately positive correlation between team talent and ELO rating. Moreover, teams aiming to compete at the championship level seem to require a team **talent rating exceeding 700** and an **ELO rating surpassing 1,800.**

Typically, only 17% of teams meet these criteria.

**Figure 8:** ELO Rating vs. Team Talent Rating by Championship Appearance




## Is a team's success correlated with the proportion of players who get drafted to the NFL?

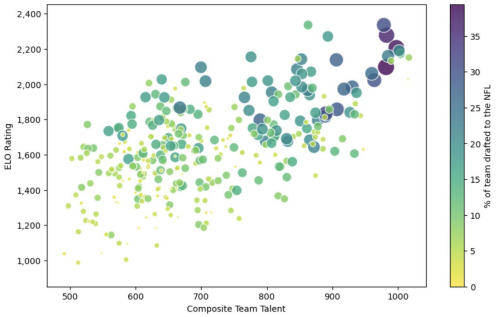
**Analysis**

Now that we've observed a positive relationship between ELO rating and Team Talent, let's explore how these factors correlate with the percentage of players drafted from a team.

Figure 9 reveals a relationship between ELO Rating, Composite Team Talent, and the percentage of players drafted to the NFL. **As Team Talent and ELO rating increase, a higher percentage of the team is likely to be drafted to the NFL.**

Additionally, examining Pearson's correlations between numerical variables, we find a moderately positive correlation (Pearson's correlation: 0.34) between the percentage of team drafted and the team's talent, and a slightly higher positive relationship (Pearson's correlation: 0.39) between the percentage of team drafted and the ELO rating.

**Figure 9:** ELO vs. Composite Team Talent by Draft Percentage color scale



# Are high school ranking a good indicator of an athlete's performance?

## Analysis

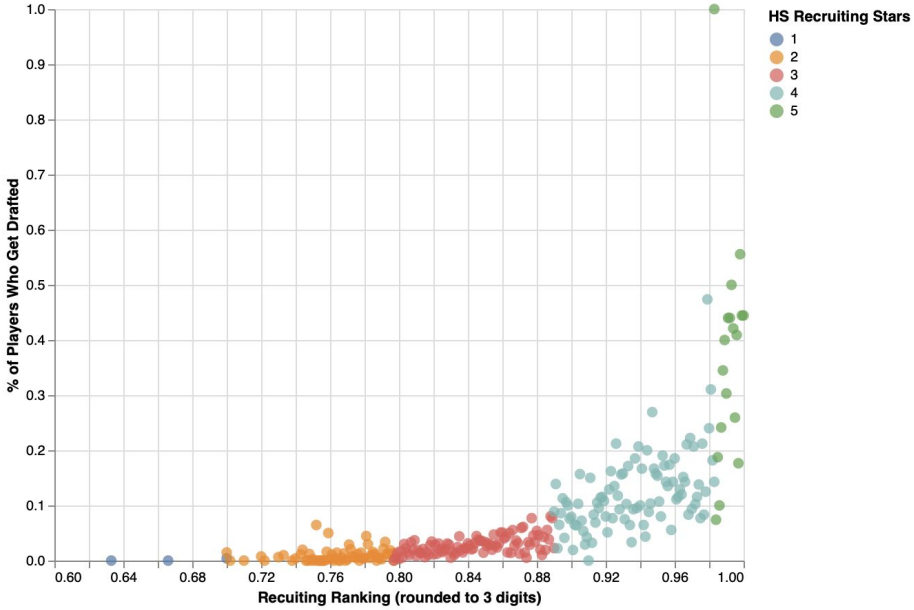
High school football players are scored on a rating scale from 0.50 to 1.0. The ratings are then segmented into 5 categorical groups on the star-rating scale (1-star to 5-star players).

From the previous slide, we know that building a talented roster in aggregate is important to finding team success.

However, we wanted to understand whether the recruiting rankings of individual players accurately predict that athlete's future success. As we can see from the chart on the right, there's a clear correlation between a player's high school recruiting rating and their likelihood to get drafted.

In fact, a **5-star recruit** is **10 times more likely to be drafted than** their **3-start counterpart.**

**Figure 10:** High School Recruiting Ranking vs. Draft Likelihood



# High school talent source and destination.

## Analysis

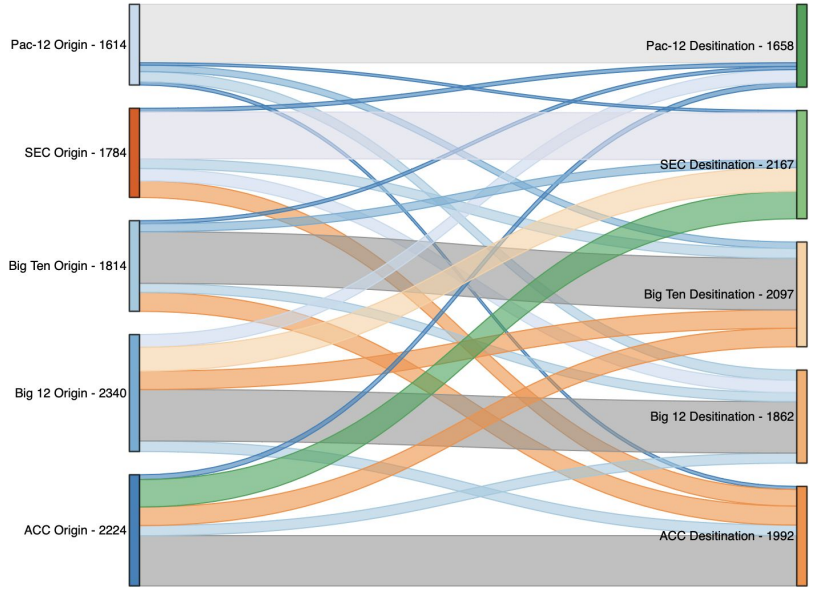
While the last slide unveiled the importance of high school recruiting rankings, it didn't show where the most talented players are from and where they end up going to school. This diagram shows the geographical start and end points for high school recruits.

The left side of the diagram represents the college football conference closest to the athlete's hometown, and the right side represents the conference of the university they ultimately committed to.

We identified their conference of origin by finding the closet Power 5 school to them geographically and then joining on the teams dataset to see what conference that team is in.

Note that about **60% of recruits** will typically **remain in the same area where they're originally from.**The most common conference-to-conference path is for a recruit to go from the ACC region to the SEC (see green ribbon in Figure 12). The least common is for a Pac-12 player to go to the SEC. In fact, in general the Pac-12 had fewer recruits from other regions.

**Figure 12:** A Breakout of Where Recruits Come From and Where They Commit



# Conclusion

## Conclusions, Next Steps, and Ethical Considerations

### Conclusions and Insights:

1. ELO is a reliable metric to use when predicting outcomes in college football games (figure 7, slide 6).
2. A team's composite talent rating is a solid (but not perfect) predictor of their season-ending ELO rating (figures 8-9, slide 4).
3. There is a minimum talent threshold to compete for National Championships - no team in the past 9 years has had a below average composite talent rating and appeared in the National Championship game (slide 7, figure 8).
4. Individual player recruiting rankings are a strong predictor of their draft likelihood (slide 8, figure 10).
5. Players are most likely to stay in their geographical region when choosing a school, but there are some pipelines that exist region to region - such as ACC to SEC (slide 9, figure 12).

### Next Steps and Additional Questions:

1. If we have more time we would have liked to understand what teams had a lower draft % but high team performance.

2. We'd like to better understand the recruiting pipelines between recruits to schools. We learned that most stay in the same region, but we believe there are more detailed insights to unpack, such as:

   a. Are the pipelines of highly rated players different than others?
   b. Do individual schools have regions they're particularly successful recruiting in?
   c. What do teams with high ELO ratings have in common in terms of recruiting tendencies?
   d. How far do recruits typically travel from their hometown and does it vary by their rating?

### Ethical Considerations:

- ***Ethical Misconceptions #10 (from SIADS 503)*: *Ethics is someone else's problem.*** Since a recruit's rating is a strong predictor of NFL success, it could create an incentive for players to pay scouts for a higher rating. This could lead to an unfair playing field and cause talented players to be overlooked.
- ***Ethical Misconceptions #15 (from SIADS 503): Numbers are objective***. Just because we boiled down a recruit's rating to a single number doesn't mean it's objective. Potential biases could easily exist among the recruiting scouts themselves (such as race or geographical location).

# References / Contributions

**Contributions:**

1. Project proposal - Chris & Andreea.
2. Slide template creation - Chris
3. Github maintenance - Andreea
4. Initial visual EDA - Andreea
5. Data manipulation / joining - Chris & Andreea
6. ELO rating overview - Chris & Andreea
7. ELO as a Predictor of game outcomes - Chris
8. Minimum Talent Threshold Analysis - Andreea
9. Team Talent Rating and NFL Draft - Andreea
10. High school rating and draft likelihood - Chris
11. Recruit source and destination - Chris

**Articles and References:**


1. https://www.bkgm.com/faq/Ratings.html
2. https://uscf1-nyc1.aodhosting.com/CL-AND-CR-ALL/CL-ALL/1967/1967_08.pdf#page=26
3. https://www.nytimes.com/2021/06/30/sports/ncaabasketball/ncaa-nil-rules.html


**Data Sources:**

- https://api.collegefootballdata.com/api/docs/?url=/api-docs.json
- https://www.pro-football-reference.com/years/2022/draft.htm

**Github:**

- https://github.com/aserban13/SIADS593-MilestoneI