Amani Akkoub (ctf3un), Charli Ashby (arv2vp), Jonah Cicatko (zvh3zz), Anne Kumashiro (yxt7ue), Jack Nickerson (rze7ud)

DS 3001 - Foundations of Machine Learning

10/04/2024

Data Wrangling/EDA

The source we've elected to use for our project is an online API with data on collegiate level football (https://collegefootballdata.com/). The API contains data under a wide variety of different categories, including but not limited to: game and team results/records, historical rankings, various prediction metrics, even betting lines. This particular API is updated on a weekly basis with the most recent data.

The goal of this project is to predict the results of the college football 2024-25 season, focusing on the ACC conference. This data will be used to find common characteristics among successful teams of past years, then creating a prediction for the 2024 season as it gets closer to the championship game.

There are a few possible challenges in this analysis, especially those stemming from the change in the format in which instead of the top 4 teams going to the NCAA playoffs, they have expanded to 12 teams, made up of a combination of the 5 conference champions and the remaining 7 highest ranked teams. Attempting to predict the entirety results of the NCAA Division I (1,099 schools) would be an incredibly wide scope so we decided to combat that by focusing on the ACC, shrinking our analysis to 17 schools.

Looking at the ACC over time, Stanford, California, and SMU are all new additions to the conference in 2024. This limits our relevant data for these teams, as their results from previous years do not reflect their performance in this league, against these new teams. For example, Stanford was previously a member of the Pac-12 Conference, where they were a fairly

competitive player in that league, but have never played against the Virginia Cavaliers, giving us very limited information in terms of their predicted performance against new opponents.

Something that comes up with narrowing our focus to the ACC is the new issue of analyzing out-of-conference performance. While the majority of varsity games occur within conference, there are a few out-of-conference games per season, varying drastically for each team. Some teams in the conference may have top-level out of conference competition, whereas others' may consist solely of what are called "tune-up games." This disparity in competition and lack of comparability is another thing we have to keep in mind when designing the model, since it is focused solely on a single conference.

The final and perhaps most pressing issue with the data we are using is that we will need to scrape the data on a weekly basis throughout the college football season, so a significant portion of the information metrics that we want to use are not available yet. With a subject like college football, the scene can change drastically over a short period of time with unexpected events like injuries, transfers, or similar situations, so it is important that we retrieve the most recent data for our prediction model to maximize its accuracy.