DSC 530 Data Exploration

Wyatt Rasmussen

My topic for the final project was trying to determine whether there was statistical data that a college football team could utilize that could maximize the amount of wins they got in one season. I performed this research using college football data from the 2020 regular season and used variables such as first downs, turnovers, and penalties.

My EDA was rather interesting. I found a negative correlation between first downs and wins, which at first glance may seem weird. My estimation as to why I am seeing that is good teams have explosive plays that go for much more than 10 yards while lesser teams are fighting for every yard and first down. I also found, unsurprisingly, that creating turnovers for you team leads to winning games, in fact teams with less than the mean for turnovers created were all less than 9 wins in a 12 (give or take because of covid) game season. And inversely found that giving the ball away via turnovers leads to losing games with a correlation of 0.557. Overall, I found what college coaches have been saying for decades, don’t throw interceptions or fumble the ball and force the other team to do so.

One thing I feel that was missed during the analysis was probably a wider variety of datasets. I think if I would have used data over the last 5 seasons, for example, my sample size would have been higher and I likely would have been able to draw more unique conclusions than I was able to.

I had a very large dataset with 154 different variables, with a little more knowledge of how to approach this and experience with data science I could have used a number of those extra variables such as First Down Runs vs First Down Passes to see if one led to more success or even using Red Zone stats to figure out what areas correlate with the most success for a team.

I didn’t have a ton of assumptions, most of them I ended up correcting in my dataset. One that I found early on that seemed to be an outlier and invalid assumption was penalty yards. With Covid-19 changing the length of seasons for some teams a total number of penalty yards seemed in accurate, of course the teams who played more would have more penalty yards. I tried to make those assumptions more unbiased.

I think the biggest challenge that I faced was the shear amount of variables I had at my disposal. As I talked about earlier on, it was difficult for me to decide what variables I should use and what I should leave out for my sort of beginners sake. I would have loved to use all or almost all of the variables, but settling on less than 10 was the hardest part for me.