Final Project Proposal

Matt Anikiej, Russell Spence, Andrew Shen

For our final project we wanted to be able to predict which type of play an NFL team will run, given the current game state. Part of the skill of a good offensive/defensive play-caller is being able to understand and read their opponents gameplan. This prediction informs their choices then on how to stop that play, and prevent them from getting a first down or score. Some plays are obvious and easy to predict, but others are much more nuanced. If a model can predict these types of plays, it will be invaluable to a play-caller since they will always have the best information available.

In order to predict future data, we need past data. Fortunately, the website NFLsavant has data sets for every single play that has been run all the way back to 2013. Due to team turnover over time, both within players and their skills and coaches and their aggressiveness, we will only use the data from 2022, because we want to predict plays in 2023. Despite only using one season's worth of data, there were 38599 plays and rows of data to train/ test on. This should be enough to train a decent model, and more data can be added if needed. All of their data is gathered from publicly available NFL play-by-play data.

The data set does have many features that are irrelevant to our prediction, or a direct result from our prediction. Likewise we don't care about the outcome of a play because we're trying to predict what it is before it happens. The most important features will be the Quarter, Minute, Second, Down, ToGo, and YardLine as that gives a clear picture of the game state. Different team tendencies and skills are also important, so OffensiveTeam and DefensiveTeam will also be necessary.

The data is on a pretty similar scale, and also includes some categorical data. We can use one-hot encoding for the categorical data, and combine the Minute and Second features into one feature as they are directly related to each other. We will also be adding features for the current score for each team. The model will be measured on its accuracy to correctly predict the play that is going to be run. It would be interesting to see how these predictions can be made from a supervised and unsupervised learning perspective. If there is a pattern to the type of play and the game state, then K-Means should be able to cluster together similar game states with similar plays. However, if the game states aren't so easily separable, a neural network might be able to find correlations that K-Means was not able to.