CSCI 4502 - Project Proposal

Team Members:

Luke Campbell, Chaz Morton

Introduction/Motivation:

Professional football is a highly regarded sport that has captured the interest of millions of Americans. While many factors of any given football game are determined by skill and coordination, lots of the decisions made on the field are determined by statistics and follow a mathematical model. From the fans of both teams to the defensive coordinators and football players themselves, many people are invested in the outcome as well as the individual decisions made during the game. We chose to look at datasets from the NFL Big Data Bowl to create our own models of prediction in order to further our understanding of why certain decisions are made in the game of football.

Our team was browsing Kaggle for datasets when we noticed a competition from 4 years ago using the NFL Big Data Bowl dataset provided by the National Football League. The $75,000 challenge for the competition was to predict how many yards an NFL player will gain after receiving a handoff. Similar competitions were posted in the following two years to predict data involving passing plays and kicking plays respectively. We saw a lot of potential in combining these datasets to predict outcomes specifically relating to the play type.

Related Work/Literature Survey:

Several models have been proposed by various authors, published on the Kaggle website to predict how many yards a player will gain after receiving a handoff *(Featured code competition, Kaggle.com, NFL Big Data Bowl 2020)*, defensive performance on passing plays *(Featured code competition, Kaggle.com, NFL Big Data Bowl 2021)*, and special teams performance *(Featured code competition, Kaggle.com, NFL Big Data*

Bowl 2022), but we have not found any data models relating to prediction of play type (whether any given play is going to be a pass, run, or kick).

As previously mentioned, several datasets have been used as the basis of Kaggle competitions to predict outcomes of designated **play types** based on various metrics (**runs** in 2020, **passes** in 2021, **kicks** in 2022). Several others have used these datasets to create various other predictions related to the sport of football including offensive formations *(Predicting NFL Offensive Formations by Cole Peterman, Kaggle.com)*, final leaderboard predictions *(Final Leaderboard Predictions by Gert, Kaggle.com)*, etc.

Proposed Work:

As a preprocessing step, we will remove unnecessary attributes from the datasets to make sure that every attribute value is relevant to our project. Then, we intend to combine all three datasets from the 2020, 2021, and 2022 *(NFL Big Data Bowl, Kaggle.com)* competitions (after adding a PlayType attribute to each dataset) to create a "master dataset" We will use this to find similarities and differences between various metrics for each play type. This may include averaging various attributes and creating charts to find averages and compare them between the three play types. Our ultimate goal is to create a prediction model using this data to determine whether any given play will be a "pass, run, or kick" when given various play attributes. We will create an accuracy model of >= 67% to determine the validity of the prediction.

Evaluation

Our model will be trained on the following attributes:

- GameID
- PlayID
- PossessionTeam
- YardLine
- Quarter

- GameClock
- Down
- Distance
- FieldPosition
- HomeScoreBeforePlay
- VisitorScoreBeforePlay
- HomeTeamAbbr
- VisitorTeamAbbr
- Team

Our model will determine whether the play type is a run, pass, or kick using these listed attributes. The evaluation metric we will use to assess our model will be the percentage of correct prediction of play type. **This will be calculated by taking the number of play types correctly predicted, and dividing it by the total amount of predictions, giving us our percentage.** Our hope is that our final model will predict at least 67% of play types correctly. By using an accuracy model of >=67%, our results are validated based on majority, which leaves us room for potential outliers. We aim to train our model on the majority of the dataset, and test our model using the remaining data. With the possibility of predicting play types for live games if we are able to retrieve the attributes needed.

Milestones

Our project will have 6 milestones to achieve.

1) **(Oct. 9th, 2023) Create combined dataset**: We will first clean the 2020, 2021, 2022 NFL Big Data Bowl datasets.
    a) Removing unnecessary attributes.
    b) Make sure that every attribute value is relevant or not of null type. For example, the 2022 dataset contains kickoffs, punts, field goals, and extra points. Although we only want to see if the playtype will be either a punt or a field goal kick.

   c) Adding a playtype attribute to our combined dataset. The 2020 data will have playtype of run, 2021 will have a playtype of pass, and 2022 will have a playtype of kick.

2) **(Oct. 23rd, 2023) Create helper functions:** An example of this would be finding the score difference for the current team with the ball. We will use these helper functions to develop our model.
3) **(Nov. 6th, 2023) Create model:** Our model will find clusters of related data points that correlate to a certain play type based on our training data.
4) **(Nov. 20th, 2023) Analyze model:** Determine the model accuracy based on our test data and identify the relationships between the play attributes that determine the likelihood of any given play type.
5) **(Dec. 4th, 2023) Format findings and create presentation**: Create a powerpoint presentation outlining the steps we took to reach our conclusion, explanation of our model, the accuracy we achieved, and charts that visualize our findings.
6) **(Dec. 7th, 2023) Project Final Report & Presentation**