Big data bowl 2022

Massimo Hong
Student ID: 2020280082
Lohan Meunier
Student ID: 2021280356

Abstract

NFL is one of the biggest and most prestigious sport league in the United States, with the highest average attendance of any sport in the world.
The kaggle competitions has a list of compulsory, but not limited, tasks to complete. In our project we will also gather and filter data from the seasons 2018-2020, analise the various statistics and feed the obtained parameters to the neural network in order to get the win probability of a team against a rival. In the end, we will give the same inputs to a logistic regression model and compare the respective outputs. Other than win probability, it could also be possible to predict whether a certain play or strategy will be successfull or not.
A neural network *should* be more accurate than a logistic regression, that is essentially just a "subset" of the former.

1 Introduction to NFL

The National Football League (NFL) is a professional American football league consisting of 32 teams, divided equally between the National Football Conference (NFC) and the American Football Conference (AFC). The NFL is one of the four major North American professional sports leagues, the highest professional level of American football in the world.
The NFL's eighteen-week regular season runs from early September to early January, with each team playing seventeen games and having one bye week. Following the conclusion of the regular season, seven teams from each conference (four division winners and three wild card teams) advance to the playoffs, a single-elimination tournament culminating in the Super Bowl, which is usually held on the first Sunday in February and is played between the champions of the NFC and AFC[1].

2 Dataset

The dataset is divided into multiple csv files:

- game.csv, contains teams playing involved in each game.
- plays.csv, contains information regarding the plays of a game.
- players.csv, contains information regarding the players.
- tracking.csv, contains player tracking for each season.
- PFFScoutingData.csv, contains play level scouting for each game.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

3 Data analysis

We can use the gathered information to determine factors like:

- what play causes a team to win or lose the most points.
- the most important attacking player on the team.
- if the team has an advantage against the rival team based on the players, or the overall playstyle (e.g. difference in height and weight can mean a lot in this sport).
- what could be the contribution of a certain player to the team.

All these factors can be used to better predict the outcome of a certain play, strategy, or the match itself.

4 Required tasks of the competition

- Create a new special teams metric.
- Quantify special teams strategy. Special teams' coaches are among the most creative and innovative in the league. Compare and contrast how each team game plans. Which strategies yield the best results? What are other strategies that could be adopted[2]?
- Rank special teams players. Each team employs a variety of players (including longsnappers, kickers, punters, and other utility special teams players). How do they stack up with respect to one another[2]?

5 Which neural network to use

We will divide the data into training set and test set. The training set will be approximately 2/3 of the total data set.
A possible choice for the neural network is the Probabilistic neural network (**PNN**), as these kind of networks usually generate an accurate target probability score.
Anothere option is to use a classification network to predict whehter a play was successfull or not.

6 Regression Model

It is more natural to use a logistic regression approach rather than a linear model, even though the output is not exactly binary. A few tweaks might be necessary in order to make it work correctly, such as using odds instead of probabilities, for an easier interpretation. The advantages of a regression model is its easy implementation and ready to use methods offered by the python *sklearn* library.

7 Conclusion

After getting the results from the neural network and the regression model, we can proceed to compare them. If everything is implemented correctly, the neural network should have a higher accuracy score than the regression model.

References

[1] Wikipedia. *National Football League*.

[2] Kaggle. *NFL Big Data Bowl 2022*.