NFL Big Data Bowl 2021

Group 3 || Matthew Fritze, Krupal Patel, Mital Patel, Junyi Peng, Mohit Ramnani, Yanyan Zhang

Source: https://www.kaggle.com/c/nfl-big-data-bowl-2021/data

The data source for this project is provided through a current Kaggle competition called NFL Big Data Bowl 2021 sponsored by the NFL. The dataset contains games, players, plays and tracking data for all passing plays during the 2018 regular season.

Data Preparation : The dataset was 2.17GB across 20 csv files which were manually uploaded to Databricks platform and then converted and saved to a table for faster reads.

Although the data is of a high quality, there is missing data. Lineman data is not provided. The dataset is focused on passing plays; running plays data are not included.

Objectives: What is your thesis that you set out to prove or disprove?

The objective of this analysis is to determine whether or not strong offensive plays made an impact on the subsequent defensive adaptations. We applied ML models to make predictions on play results using key information such as player position, size, and setup.

Analysis or Model: Explain the model/analysis and how you confirmed the validity of the approach. What were the challenges? How did you overcome them?

Following four classification and regression methods were used to analyze the dataset.

1. MLLib - Random Forest Classifier

    A multi-class Random Forest Classifier was built with Spark MLLib in an attempt to predict the pass result of a given play.

    The model was unsuccessful in accurately classifying the labels based on the features provided. The classifier was unable to find any meaningful patterns in the data therefore making any useful predictions was impossible.

Another Random Forest Classifier was built to try to predict a defensive formation based on personnelO data and absoluteYardlineNumber to create a suggestive model.

The model with the best hyperparameters was used for test but the results were not satisfactory, giving only a prediction accuracy of just greater than 1%. It was concluded that a set defensive formation cannot be predicted based on just those features since every team uses a different tactic and what they choose to go with also may depend on the situation of the game, and other external factors.

2. MLLib - Decision Tree Regression

   Decision Tree Regression is used to explore and explain whether there's any relation between the playtype variables and the outcome (playResult), and the feature importance of the playtype variables for determining and predicting the outcomes.

   A decision tree regressor was the algorithm chosen for modeling.

3. MLLib - Linear Regression

   Linear regression was used to predict the number of yards gained on the play which, in this dataset was encoded as a float.

   The linear regression model was a poor predictor for net yards gained. Further analysis could be done by breaking down the yards gained by each offensive formation, since yards-gained is more likely to be linearly related to individual play-types.

4. MLLib - Logistic Regression

   Logistic Regression analysis was used to predict how factors such as players speed, acceleration, weight, height and age affected the performance of their offense plays. Our hypothesis is that players' speed, acceleration, weight and height will positively impact the likelihood of successful offensive plays, whereas age will have a negative impact on the result of the offensive plays.

Conclusions: What did you learn?

Finding consistent advantages turned out to be more difficult than we hypothesized. This aligns well with our original hypothesis that any successful strategies would be correspondingly counter-strategized.