Predicting Play Outcomes in the NFL

1st Tucker, Taylor
University of Virginia School of Data Science
Charlottesville, United States
rtt4fb@virginia.edu

2nd Shoriz, George
University of Virginia School of Data Science
Charlottesville, United States
pvq8hv@virginia.edu

3rd Lisman, Zack
University of Virginia School of Data Science
Charlottesville, United States
zjl2ue@virginia.edu

4th Casillas-Colon, Abner
University of Virginia School of Data Science
Charlottesville, United States
aec4hr@virginia.edu

5th Lotane, Charles
University of Virginia School of Data Science
Charlottesville, United States
aec4hr@virginia.edu

***Abstract*—This study focuses at the convergence of machine learning and American Football, leveraging the statistical revolution initiated by AWS and Amazon's growing involvement in the NFL. We aim to predict play outcomes, including expected yardage gains, play type classifications, and their impacts on scoring metrics like predicted points and victory probability. By utilizing a comprehensive dataset from Carnegie Mellon University, which includes over 449,371 observations and 255 attributes spanning from the 2009 to 2016 NFL seasons to perform rigorous data cleaning in preparation for advanced predictive modeling. Our methodology consists of three primary experiments: a sequence model utilizing recurrent neural networks (RNNs) to predict subsequent plays, a fully connected neural network to estimate play distance, and a binary classification model to distinguish between pass and run plays. The findings of this study have important implications for NFL teams looking to optimize offensive and defensive strategies, while also providing improved analytical tools for fans and stakeholders interested in the game's changing landscape. Ultimately, this work advances our understanding of play efficiency and player performance by proving the revolutionary power of predictive analytics in professional football.**

***Index Terms*—Data Science, Machine Learning, Artificial Intelligence, NFL, Predictive AI**

I. Introduction

With AWS and Amazon becoming more involved in the NFL than ever before, there is a statistical revolution occurring in football. Stats such as catch probability, separation distance, win probability, and many others have been introduced to football and are completely changing the way teams play the game.

In this study, we will be using our knowledge of machine learning and neural networks to attempt to predict outcomes of NFL plays. These outcomes include how many yards are expected to be gained or lost on a play, whether the play was a run, a pass, or a trick play, and how expected points and win probability are affected on a play. This is important for NFL teams because they would like to be able to run plays on offense that will gain lots of yards and points, and run plays on defense that will prevent yards and points. If a team knows that a particular play or formation works well on either the offensive or defensive side of the ball, it gives them a distinct advantage. Coaches and teams watch a lot of film to help them prepare for the games, but if we can take still images and videos and help provide insight on each play in addition to what is shown, it can add more helpful insight to teams.

Additionally, this analysis can be very useful to help analyze individual players. If a player performs a negative action such as dropping a pass or missing a tackle, or a positive action such as making a tough catch or covering a lot of ground to break up a pass, the impact of that player on winning can be more accurately quantified.

II. Motivation

As mentioned during the introduction, there is a large amount of interest for the ability to predict plays in the NFL. The ability to effectively evaluate the problem of play efficiency is an important problem at both the micro and macro level. Predicting plays alone is a problem that provides value in the NFL for teams and spectators who would want to judge how likely a pay is to succeed. This gives teams more tools to adjust accordingly against their opponents and fans more knowledge about the intricacies of the game even with less understanding on the technical elements of unique player positions. There are additional benefits of modeling this problem that can be applied to larger meta analyses. Some examples of how these are modeling win likelihood, evaluating the efficiency of plays in general to improve scheme, modeling player value by controlling for play effectiveness ,or in predicting outcomes for sports betting.

III. Dataset

Our dataset was compiled by a statistical research team from Carnegie Mellon University comprised of Ron Yurko, Sam Ventura, and Maksim Horowitz, who also developed `nflscrapR`, an R package which leverages the NFL's API to scrape data at the game, player, and play level. We downloaded the dataset as `.csv` file from the project post on Kaggle. [1]

The data is comprised of 255 features along the column space and more than 449,371 observations along the rows. The data consists of both numeric and categorical features and describe nearly all factors that likely affect football plays. The observations span NFL games from the 2009 season through the 2016 season. A subset of the features are:

- GameID
- Drive
- Down
- Pass Yards

A. Processing Requirements

Because of the size and variety of these data, it will be important to process the data for use with our models. We expect to perform various steps to prepare the dataset for analysis and prediction. We will need to perform general cleaning of the dataset, including dealing with missing values and outliers. We will need to perform feature regularization in order to avoid numeric computation issues during the modeling process. We expect to perform encoding for the categorical variables. We also expect to to perform feature selection – while having this much data for deep learning applications is great, it is likely that some of the data is not necessary or even harmful towards the goal of prediction.

IV. Related Work

A. Paper 1: *"A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression"*

While this paper by Gifford and Bayrak did not explicitly utilize a deep learning framework to predict outcomes in the NFL, the paper served as a useful nexus of related research. The primary uses of neural networks in the realm of sports predictions seem to have focused mainly on football (soccer), Australian Football, and horse racing. The paper described how, among machine learning models, ensemble methods, such as random forests, tend to outperform probabilistic models, such as naïve bayes models. Further, the paper introduced an important metric: `DVOA`, which stands for defense-adjusted value over average. This metric may be useful in our future experiments. [2]

B. Paper 2: *"Predicting Plays in the National Football League"*

Fernades et al. provide a machine learning strategy for predicting whether an NFL play will be a throw or a run. The primary goal is to create a comprehensible model that coaches and player can utilize in real-time during games. The scientists examined a variety of models. including classification trees, k-nearest neighbors, random forest, and neural networks, with the neural network obtaining the highest accuracy of 75.3 %. One of the study's important achievements is the creation of a straightforward decision tree model that maintains 86 % of the neural network's accuracy while being understood enough to be used in a game's fast-paced decision-making environment. The study also examined team-specific models, which achieved prediction accuracies ranging from 64.7% to 82.5% depending on the team. For coaches who want to modify their tactics in response to opponent treads, this degree of personalization may be beneficial.

While our group's effort will focus on forecasting specific game outcomes and efficiency measures, the insights in this article on predicting play type and using real-game data will help us develop a prediction model for NFL games. [3]

C. Paper 3: *"A deep learning framework for football match prediction"*

While the title of this paper has football in it, the deep learning framework applied in this article applies to what Americans refer to as soccer. However, most of this study is based on image analysis and feature extraction, which also applies to our analysis as well. Although the domain expertise is slightly different due to the differing nature of the games, a lot of the concepts will be similar. Both sports have the same number of players on the field at all times, both sports have advantages for the home teams, both sports have recent results that shape how the team is playing, both sports make changed to their lineups during the games based on injuries or tactical shifts, along with many other similarities. This study uses stochastic gradient descent with learning rate decay to help create the neural network along with using a softmax classifier, which is something that we anticipate we will implement in the network as well.

Lastly, the end goal of this study is to predict who will win the game, which I anticipate being something we will analyze as well, as that is the ultimate goal of any sporting match. Even if our focus shifts in a different direction, the classification and prediction used in this study can be applied to classifying plays as passes or runs, and predicting how many yards will be gained on a certain play. [5]

D. Paper 4: *"Neural Network Models for Predicting NFL Play Outcomes"*

Guo's research investigates the use of neural networks to predict the result of NFL plays based on game state information like field position, score, and time remaining. The results didn't generate great predicted accuracy (22%), the method is relevant to what our group intends to undertake in terms of developing neural network models for NFL games predictions. One important component of this article is its multi-class classification technique, which categorizes outcomes such as yards gained, turnovers, and touchdowns. Although, our study focuses on play type categorization and game result prediction, this methodologies and neural network typologies like the usage of ReLU activations and softmax outputs, are applicable to our model construction. The use of mini-batch gradient descent and hyperparamter adjustment provides useful insights into optimizing our own models for performance. The difficulties encountered in this paper, especially the problem of over-generalization with the model forecasting the most prevalent category (zero grain), suggest areas where our group would want to focus on improvements, like, including additional data elements or changing the architecture. Also, the authors propose that including player-specific data may increase model performance, which we intend to investigate in our own by leveraging relevant player information. Thought this study didn't achieve great accuracy, its strategy to forecasting outcomes using neural networks and pre-processed NFL data provided a core understanding on which to build as we design our own NFL game predictive models. [4]

V. Intended Experiments

We currently plan to perform three primary experiments, although throughout the course of the project we will likely find ourselves performing more.

A. Experiment 1: Sequence Model

We intend to develop a sequence model (`RNN`, `LSTM`, `Encoder-Decoder`, etc.) to predict the next-play vector based on previous plays in the drive. The next-play vector will consist of all of the features from the previous plays in the drive. The goal and motivation of this will be to predict the subsequent play in a drive in order for a hypothetical coaching staff to be able to scheme and strategize against the predicted play; American football can be a complicated and multidimensional sport, and strategy has played an ever-increasing role in the modern game.

B. Experiment 2: Predicting Play Distance

We intend to develop a neural network of a fully-connected architecture to perform play distance. The play distance will be a continuous variable. The goal and motivation of this experiment will be to provide defenses with tools to allow them to predict the success of a particular defensive formation and scheme against a given offense. Once again, this will allow defenses to improve their play by using offensive patterns to predict what the next play might look like.

C. Experiment 3: Pass/Run Classification

Perhaps the simplest experiment we plan to perform is classification of play type (Pass/Run) based on other features at the play level. The play type variable will take the form of a binomial variable. In the game of American football, teams can advance the ball down the field using a pass or a run. Being able to predict which play type might be seen next can allow defenses to properly assign personnel and scheme against the subsequent play.

References

[1] Horowitz, Maksim, et al. "Detailed NFL Play-by-Play Data 2009-2018." Kaggle, 22 Dec. 2018.

[2] Matt Gifford, Tuncay Bayrak, "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression", Decision Analytics Journal, Volume 8, 2023, 100296, ISSN 2772-6622, https://doi.org/10.1016/j.dajour.2023.100296. (https://www.sciencedirect.com/science/article/pii/S2772662223001364)

[3] Joash Fernandes, Craig et al. "Predicting Plays in the National Football League". 1 Jan. 2020 : 35 – 43.

[4] Rahman, M.A. "A deep learning framework for football match prediction." SN Appl. Sci. 2, 165 (2020). https://doi.org/10.1007/s42452-019-1821-5.

[5] Guo, Xuyi. "Neural Network Models for Predicting NFL Play Outcomes." Stanford C230, Stanford University, https://cs230.stanford.edu/projects_spring_2020/reports/38964602.pdf.