Visualizing NFL Team Performance*

Mei Ming[1,2], Meixian Wu[3,4], Yongkang Liu[3], and Yixuan Chen[3]

University of Southern California, Los Angeles CA 90007, USA
mming@usc.edu
meixianw@usc.edu
yongkang@usc.edu
ychen513@usc.edu

Abstract. This project is aiming at providing a thorough guide for football lovers to learn about history NFL game data by visualizing points gained by each team and predicting actions taken by the offense team. Our website contains a map, bar chart, pie chart and line chart that presents the historic play records for some popular NFL teams. It can be used for everyone who wants to learn more about different teams and their performances, and help them to decide which team they should support next time.

Keywords: NFL Team · Super Bowl · Quarterbacks · Performances.

1 Introduction

1.1 Motivation

American football is by far the most popular sport in America. The National Football League, known as NFL, is the most professional football league that consists of 32 teams. [1] Every year, the NFL championship is decided by the Super Bowl game which has become a tradition in America to celebrate this most watched sporting event. Thus, as a football lover, it is necessary to learn about the game's rules and regulations, as well as the wise strategies to support favorite teams.

We chose the NFL game as our final project topic because we want to analyze historic play dataset and visualize the performance of different teams. By analyzing the historic NFL dataset, we can have a better understanding of each team's win/lose times and the trends of the quarterback position. Our goal for this project is to make predictions about the championship in the next Super Bowl. We also hope to provide more straightforward information for people who know little about football through graphs and maps.

1.2 Scope

Our dashboard is designed to provide an overview of the game records of NFL teams. For example, we have shown the team win-loss records by year, and

* Students final project for DSCI 554 Fall 2022 in USC.

the average points a team gained from 2019 to 2022. We provide geographic information on the map to show the introduction and location for each team.

1.3 Introduction of Sections

We will thoroughly go through the paper in different sections. In Section 2, we will discuss how we prepared for the dashboard and analyzed the offense play type by predictive models. In Section 3, we will introduce the data we collected to build the dashboard. In Section 4, we present the architecture and approaches we used to build the website. In Section 5, we show the main aspects of our websites. Finally, we make conclusions and bring forward future improvements about the project in Section 6.

2 Preparation and Predictive Model Analysis

2.1 Preparation

In this section, we are going to analyze the play type ("passing" or "rushing"). We know there are two types of actions taken by the offense team to either directly pass the ball or run with it. Our goal is to predict the offense play type with a historic plays dataset from the NFL. [2]

The model we choose is XGBoost and random forest. We will compare the performance of these two models and make further improvement on the predicted results.

2.2 Predictive Model Analysis

In the data preprocessing part, we removed the quarter that is 5 since in normal games, the quarter number is less than 4 and removed downs that equal 0. We converted the gameClock from minutes to seconds, then transformed the quarter to half. Instead of yardlineNumber, we were more curious about the yards to the end zone so that we can know how far the offense team would go. The string value currently shown in personnel offense is not conducive to input, so we converted each personnel position to its own column to indicate the number present on the field during the play. We also encoded the categorical variable offense formation into different numbers. We converted the play outcome into a single column called play_type represented by either a 0 for running or a 1 for passing.

In the data visualization part, we did an exploratory data analysis. First, we draw a bar plot to see the count of the play type. From the bar plot below, we know that most play actions are passing instead of rushing. However, the number of passes is much more than that of rushing, so our label play_type is imbalanced. A down is a period where a team can attempt a play. We draw a bar plot to see whether the count of play actions will change with different down. From the bar plot below, most actions are taken during the 1st, 2nd and 3rd downs. We drew a basic regression line to see if there is a correlation between

yardsToGo and numericPlayType. From the regression line below, the larger the value the yardsToGo is, the action is more prone to be rushing.

In the predictive model part, the first model we used is called random forests. The core of this model is begging which is a very powerful ensemble method. Random Forest classifiers can handle both categorical and numerical variables. The input is all variables excluding the play type. We used all the features to predict the play actions that are taken by the players. The second model we use is called XGBoost. XGBoosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. It takes numerical variables and we also split the dataset into train, validation and test datasets to avoid overfitting.

The precision is from all the classes we have predicted as positive, how many are actually positive. It should be as high as possible. And the recall is from all the positive classes, how many we predicted correctly. It also should be as high as possible. The f1-score compares precision and recall at the same time. It uses harmonic means to integrate precision and recall. The accuracy is from all classes, how many of them we have predicted correctly. Higher accuracy sometimes means better model performance (we should avoid overfitting). Overall, random forests classifier performs better than XGBoost, but the results are very close. Thus, in our predictive model analysis, either random forests and XGBoost are good to predict the play type (passing or rushing).

Table 1 gives a summary of the model performances.

Table 1. Model Performances

Model	Precision	Recall	F1-Score	Accuracy
Random Forests	0.463095	0.498398	0.480099	0.923442
XGBoost	0.457888	0.500000	0.478019	0.915777

3 Data

Data was mainly collected from two sources: 1)2020-2022 Big Data Bowl Dataset published by The National Football League for Kaggle competition; 2) the statistics of players and teams published on the Official Website of The National Football League [3]. The Big Data Bowl Dataset contains a detailed summary of games, plays, players, and the player position tracking of games. We focused on the game's result, and extracted a subset of the "Games" dataset which includes only the teams that we were interested in, and aggregated the game results with respect to different factors. More datasets are collected from the official website of the NFL which provided detailed and authentics statistics about the players, teams and games. For the lollipop visualization in "Quarterbacks", the dataset consists of the total number of wins and lost games that the selected quarterbacks played upon 2021. The responded Multi-Line chart presented two statistics

of the player, the number of touchdowns and the percentage of the total successfully completed passing, over the decade from 2012 to 2022. The detailed Players table consists of 15 core statistics that measure the performance of a Quarterback player, all provided by the NFL official stats collection.

4 Approach

American Football has been such a popular sport nationwide and even worldwide for a long history. There could be numerous aspects to share the information and analyze the fun observations from the games. But for the web pages we created this time, we mainly focused on introducing some of the Teams and Players of the NFL in recent years. In this case, anyone new to the field of football can quickly get into the games and enjoy the teams and players they liked.

4.1 NFL Teams Truth

The page "Team info" showed the geographical location of the teams, distinguished by their belonged conference. At the top of the page, the basic rules, timelines, and timeframe of the NFL season and the postseason Super Bowl Games are explained beforehand. This gave the audience a starting point of understanding the structure of the League, how teams in different conferences can combat and what they are winning for. Then the geographical distribution of the teams provided information about the sport headquarters and the supportness of the states. The complementary color gave a good sense of the conference. We can notice that teams do not have geographical preference in terms of the conference, but rather a truly national sport. A table of detailed information about the team will show up when clicked on the team in the map, which includes a link to the team official page, description, location, active time, conference, Star Players, and historical rewards.

4.2 Teams performance Analysis

Two pages for team performance analysis are presented. The first one uses a collection of data of the total number of wins and losses from 2000 to 2021. An interactive bar chart is made using green and brown bars to represent wins and losses. Using the select bar could filter data for different years. Also, teams are sorted from ones with most wins to most losses. It is easy to detect relatively strong and weak teams for a specific year. The second one is a two-line time series plot for total points earned and lost with green and brown color again for each year. For this plot, the select bar is created to filter different teams. Change for each team could be found clearly by noticing the trend for their points statistic. For example, Team Los Angeles Rams keeps a stable trend for points lost and an increasing trend for earning points since 2016. Overall, these two pages present a general idea for each team's performance and provide a stage for digging deeper.

4.3 Special Position Analysis - Quarterbacks

There are three visualization techniques used to analyze the performance of Quarterbacks players. For time consideration, only 7 of the famous QB players are selected to perform analysis. First, lollipop was used to represent the cumulative number of wins and loses in 2021. The two ends of the lollipop, marked in complementary color blue and orange, gave clear indication of the features that encoded. The further the orange end is reached, and the longer the stick is, the better the performance. We can observe that the performance of Tom Brady is remarkable and far from the average performance. His 'Win' end is about 2 standard deviations more than the average win count of 89, while the number of lost games was still kept within average range. Secondly, corresponding to the selected QB player, a multi-line chart is used to display the percentage of completed passing and the total touchdown count over the years. Such a line chart is straightforward in analyzing the stability of the player's performance under different game situations, and also is referable to see whether the player maintained the performance and managed this physical condition professionally. However, to evaluate a QB player there are much more to consider. In the table below the plots, it provided more core statistics that measures the performance of the selected players. The description of the variables are also provided to help the audience understand the statistics. Our further analysis would be to construct 3-dimensional bubble plots with number of attempts(ATT) as x-axis, total receiving yards(YDS) as y-axis, and the overall throwing effectiveness rating as the bubble radius to investigate possible relationship between the number of attempts and effective rate. (See Fig 1)

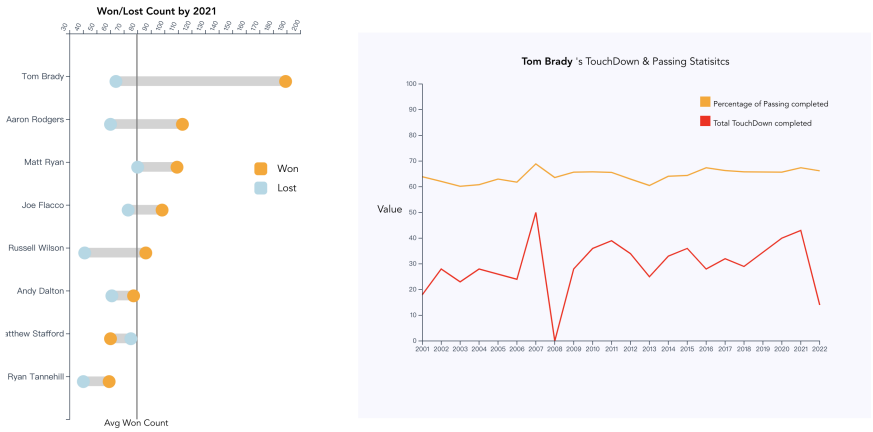

Fig. 1. A figure that shows the win/lose records of the quarterback players and corresponding player's performance over time.

5 System

This visualization application used Node.js and Vue.js as front-end framework, and Bootstrap as CSS styling framework for responsive, and organized front-end web development. All visualizations and tables are built with D3.js, which makes use of Scalable Vector Graphics and HTML5 to produce dynamic and interactive data visualizations in web browsers. D3.pie was used to form control and built an arc path for the donut chart.

During the implementation of the teams geographical location in map, there are difficulties in which the teams address locations could not directly map to the topojson projection file. After failing in matching the data with county name, we switched gear to match the FIPS id, which is the county-level identifier, with the address. By looking up the intermediate FIPS id for all team addresses, we are able to match up the team location in the projection map. There were also some hardships in getting the tooltips moved with mouse position. After debugging and reading documentation referring to the correct usage of functions, the tooltips were able to move along accurately with the mouse pointer.

6 Conclusion

6.1 Contribution

We divided the project work by sections. Meixian, Yongkang and Yixuan are responsible for building the websites and making different plots. Mei is mainly responsible for predictive model analysis. We worked together to solve the technical problems and had meetings each week to talk about the next step.

As for the individual tasks, Meixian created the d3 map and the lollipop chart. Yongkang plotted the pie chart and bar chart, and Yixuan plotted the stacked bar chart and line chart. We focused on our own parts while writing the final paper.

6.2 Improvements

We still need to elaborate on our project by adding more useful UI design of the web pages. It is very important to reach out to the users and make them more involved. For example, we could add more buttons or mouse over to make users learn about what they want and keep the page as clean as possible. The other suggestion is that we can add some links in the web pages for guiding users to different pages.

As for the functionality, it is better that we have more dimensions of datasets. The current dataset we have collected only contains information which is not comprehensive for a beginner to know much about football. We may consider adding more features that help users with different knowledge to learn about football from our website.

References

1. Super Bowl, https://en.wikipedia.org/wiki/Super_Bowl.
2. NFL Big Data Bowl 2023, https://www.kaggle.com/competitions/nfl-big-data-bowl-2023/data?select=plays.csv.
3. Official site of the National Football League, https://www.nfl.com/teams/.