Finding Fantastic Feats of Football in the NFL

Using Python and statistical libraries to mine unique relationships from NFL play-by-play data

Alex Plazas
Computer Science
University of Colorado-Boulder
Boulder, CO, USA
chpl8693@colorado.edu

Benjamin Price
Computer Science
University of Colorado-Boulder
Boulder, CO, USA
benjamin.price@colorado.edu

Raul Ramos
Computer Science
University of Colorado-Boulder
Boulder, CO, USA
Raul.Ramos@colorado.com

PROBLEM STATEMENT/MOTIVATION

Sports serve society by providing vivid examples of excellence. From the basketball court to center ice, athletes and teams strive for success in their respective disciplines. While there are many definitions of success, statistics generated from competition tend to come first when measuring performance.

In recent years, innovations in computing and information management have ushered in a new era of statistics in sports. Sports leagues and the teams therein commit vast sums of money to collect and process data that could give their side an edge in competition. Insights gained from larger bodies of statistical information from a sport enable teams to learn from the past, apply to the present, and potentially predict the future. And the sea-change in sentiment toward statistics is clear in the National Football League (NFL).

The NFL, including American football in a broader sense, has embraced statistics for the better of the sport. Each of the 32 teams of the league, including the league itself, has invested millions in dedicated offices for gathering and mining statistical data. And for good reason. At stake are billions of dollars invested by ownership, sponsors, and fans, with all of the above wanting their money's worth. Thus, each office works tirelessly to find the next big prospect or craft a guaranteed-winning strategy backed by data-driven methodologies that go beyond “gut instinct”.

Beyond the monetary incentives for collecting and processing vast sums of data, statistics offices for NFL teams are typically concerned with one problem: finding outliers. More specifically, positive outliers depend on the statistical category or measurement implied. Granular play-by-play data enables team front offices to detect outliers and assess performance in deeper detail. Teams place emphasis on finding players or situations that increase their probability of winning, and the players and situations involved are typically out-of-the-ordinary. Certain environments might also lend themselves to outlier performance such as specific training routines or game-time decisions, both of which can be modeled and tested using game statistics.

Our stated goal is to find such fantastic feats and study their impacts. We wish to understand the confluence of statistical factors that lead to outlier performances. We also wish to build a model that incorporates data from a wide range of seasons to predict the outcomes of plays and games given the presence of outliers. In the end, the team wishes to emulate the work of an NFL front office in finding fantastic feats of football.

LITERATURE SURVEY

Many researchers have attempted to tackle the NFL outcome prediction problem by using a mix of historical and play-by-play data, with unique constraints and intentions. Some studies focus on player-specific metrics at certain positions such as quarterback. In a paper titled “Measuring Productivity of NFL Players”, Berri et al. sought to explore factors associated with the quarterback position and how quarterback productivity correlated with a team’s offensive ability[3]. The study outlined several statistical attributes including the derived attribute of Quarterback Rating (QBR), and standard statistical categories such as yards, plays, and interceptions. Using regression analysis, the findings of the study indicated that there are measurable positive increases in the predicted point differential for each yard thrown for, and a negative impact on point differential for every play attempted. It was also shown that from 2000 to 2010, Peyton Manning had the highest QB rating and nearly a quarter of the top 40 seasons during the time period[3]. At the conclusion of the paper, Berri and Burke suggested shortcomings and alternatives to their regression model, such as calculating expected points values and “success rate.”

Other studies take a broader approach by analyzing team performance over time given certain conditions and historical outcomes. In “A Hybrid Prediction System for American NFL Results”, Uzoma et al. propose a method for predicting NFL games using particular models and features[5]. A hybrid linear regression and k-Nearest Neighbors model were designed to increase the prediction accuracy of already weighted features. Several data attributes of interest were also identified in the study including points scored by both teams, the number of turnovers, and offensive/defensive rating[5]. Using the hybrid model, the authors were able to predict outcomes of games with nearly 80% accuracy during the 2013 regular season[5].

In a similar study titled “Predicting Margin of Victory in NFL Games...”, Warner proposes using a machine learning model to accurately predict games better than Las Vegas bookmakers. Aside from using regular counting stats and attributes, the study also included novel features such as the location of the game, stadium conditions, and even climate/weather data for the particular game[4]. To generate predictions, a feature set was defined which minimized cross-validation error, and features were passed to a Gaussian process predictive model to yield a final prediction. Results of the study indicated that the model did not out-perform Las Vegas prediction models at a 95% confidence interval, despite the model including novel attributes[4].

On the topic of important attributes, a study published in 2010 titled “What Makes a Winner?” sought to identify attributes that strongly correlated to the expected outcome of a game. Gifford et al. defined a model which employed a decision tree coupled with binary logistic regression to find key attributes in NFL play-by-play data[6]. Several parameters were tested including passing yards, rushing yards, and turnovers to find which impacted the expected outcome the most. The results of the study indicated that offensive turnovers were the most important team statistic in determining the winner of an NFL game, having a strongly negative effect on the outcome for the offensive team, with the inverse being true for the defensive team[6].

PROPOSED WORK

The project begins with the found data as described in the subsequent section. The data is abundant and will need to be cleaned and aggregated in some meaningful ways to facilitate the types of exploration this project seeks to perform without losing any information.

Many data mining efforts struggle with how to deal with N/A or missing data. While our strategy will vary depending on the data type of a given attribute, many of our missing values are related to binary attributes and can safely be set to 0. Similarly, N/As in nominal attributes, such as Passer Player Name can be included because the field is an implied binary; N/A means there was no passer.

Other forms of cleaning will need to be performed to reduce the number of attributes to sift through in the data. Some of the attributes given for each play-by-play instance are similar in nature (i.e. forced fumbles, fumble forced, fumble not forced), and can be combined to simplify the analysis process. Similarly, some attributes can be implied from others and can be safely removed, such as Game Half being implied from Qtr. Carefully reducing the number of attributes in this way will speed up the data mining process without the loss of information gain.

A number of pre-processing steps will also have to be performed once the data is cleaned. Our data is currently broken up into many dimensions, with each object representing an individual play. While some exploration will look at data per play, full game data will also be required. A roll-up of the data per game will be implemented which can take place by totaling up binary data as well as the other numeric data and removing any attributes that would not apply at that level, such as seconds remaining or the Qtr attribute. With this higher-level view, comparisons can be made at a per-game level as well as the per play level provided by the original data.

Once the data has been sufficiently cleaned and transformed, the project will aim to answer meaningful low-level questions such as ‘How does field position impact win probability?’, ‘Does having top players always guarantee success?’, and ‘In what situations should you "go for it" on the fourth down?’

In previous literature, examples of using football data have been shown to help create unique opportunities to improve performance, such as running plays farther from an opponent’s bench to wear-out defensive players quicker[1]. This project will aim to aid in the discourse by using classification techniques and regression modeling to open up additional unique opportunities that could be implemented on a per-play basis in future NFL games to improve success.

DATASET

The data set used in this exploration is from the Kaggle website from user Max Horowitz[2]:

https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016?select=NFL+Play+by+Play+2009-2018+%28v5%29.csv

Detailed play-by-play statistics per game have been collected in this data set for National Football League games from 2009 through 2018. There are 253 attributes in the data set, with 449,371 data objects each identified by a combined key made up of both a Game Id and a Play Id.

Most data types are represented within this data set given that there are so many attributes provided. A number of nominal attributes are included, such as Play Type, Home Team, Away Team, and Pass Location. Binary attributes are numerous within the data to detail the inclusion of certain types of plays, such as if there was a kickoff attempt or a punt attempt, and ordinal data is provided such as the quarter of the game in the Qtr attribute. There are plenty of numeric attributes as well, such as the ratio-scaled data of Yards Gained and Game Seconds Remaining.

The sheer robust nature of this data allows versatility in the knowledge discovery process and provides multiple windows of perspective through which to explore patterns in recent NFL seasons.

EVALUATION METHODS

1. Create a regression model using the data from the 2009-2016 seasons as "training" data, and test the model using the 2017-2018 seasons.
2. Create a regression model using all available season data as "training" data (2009-2018), and source up-to-date season data (2019-2022) as "testing" data for the model.
3. Assign each team a probability of winning the Super Bowl for that season, and test if the model is accurate in predicting playoff success.
4. Consider classification (i.e., confusion matrix, classifier model) and clustering (i.e., Euclidean distance, visual analysis, vector partitioning) for predicting where categories belong.

TOOLS

1. The primary language is Python 3 or above
2. Python packages NumPy, Pandas, and Matplotlib for data processing, analysis, and visualization.
3. The secondary/alternative language is R for fetching data from other sources.
4. Tableau, PowerBI, and Excel for visual exploratory analysis.
5. Google Slides for presentations and project updates.

MILESTONES

1. **Part 1: Project Proposals: February 28, 2022.** Create presentation slides that contain the following slides: title, team members, project description (2-3 sentence), prior work, datasets (include URL), proposed work (data cleaning, preprocessing, integration), list of tools, and evaluation. Create a group GitHub account with all members set as contributors. Submit the file(s) as requested by instructions.
2. **Part 2: Proposal Paper: March 14, 2022.** Write approximately 3 pages using ACM SIG paper format, and include the following sections: problem statement/motivation (what you hope to find), literature survey (previous work), proposed work (data collection and processing and how it differs from proposed work), dataset (URL), evaluation methods (i.e., metrics, existing solutions), tools, and milestones (phases and deadlines). Submit the file(s) as requested by instructions.
3. **Part 3: Progress Report: April 18, 2022.** Meet with the group to delegate weekly tasks among group members. Start cleaning the dataset, and discussing, developing, and testing the code needed for the project. If needed, find resources to learn how to write the code. Consider referring to Homework 1 thru 4. Make sure the code works before you update it on GitHub. Do weekly meetings (or as needed) and consider the following: what you did last week, what you are doing this week, what obstacles are stopping you, what went well, what didn’t work, and how to improve. Write approximately 6 pages using ACM SIG paper format, and include the following sections: an updated proposal from Part 2, milestones completed, milestones to-do, and results so far. Submit the file(s) as requested by instructions.

4. Part 4: Final Report: Thursday, 28 April 2022, 11:59 PM.
Continue meeting on a weekly basis or as needed, cleaning data, and developing code. Write approximately 6 pages using ACM SIG paper format, and include the following sections: abstract (interesting questions and results summary), introduction (question descriptions and importance), related work, dataset (where from, attribute features), main techniques applied (i.e. data cleaning and preprocessing), key results (discoveries), applications, meaningful interactive visualization (extra credit). Submit the file(s) as requested by instructions.
5. Part 5: Project Code and Descriptions: April 28, 2022. Submit to GitHub all source code used for the project, and create a README file on the main page with the following: project title, team members, description of the project, a summary of questions sought and answer, application of this knowledge, link to the video demonstration, link to the final project paper. Submit the file(s) as requested by instructions.
6. Part 6: Project Presentation: April 28, 2022. Create a project presentation that includes the following: project title, team members, questions sought to answer, data preparation work, tools used, main techniques applied (i.e., data cleaning and preprocessing), knowledge gained, how knowledge can be applied. The presentation must follow the correct, last only 6 minutes, and all members must participate.
7. Part 7: Peer Evaluation & Interview Question: April 28, 2022. Each team member must fill out an individual peer evaluation form and include comments. Submit the file(s) as requested by instructions.

REFERENCES

[1] Kevin Clark. 2018. The NFL's Analytics Revolution Has Arrived. (Dec. 2018). Retrieved March 14, 2022 from https://www.theringer.com/nfl/2018/12/19/18148153/nfl-analytics-revolution

[2] Mark Horowitz. 2018. Detailed NFL Play-by-Play Data 2009-2018. (Dec. 2018). Retrieved March 14, 2022 from https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016?select=NFL+Play+by+Play+2009-2018+%28v5%29.csv

[3] Berri D.J., Burke B. 2012. Measuring Productivity of NFL Players. In: Quinn K. (eds) The Economics of the National Football League. Sports Economics, Management and Policy, vol 2. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-6290-4_8

[4] Warner, J. 2010. Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.8415&rep=rep1&type=pdf

[5] Uzoma et al., 2015 A Hybrid Prediction System for American NFL Results. International Journal of Computer Applications Technology and Research Volume 4– Issue 1, 42 - 47, 2015, ISSN:- 2319–8656 https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1038.1982&rep=rep1&type=pdf

[6] Gifford, M. and Bayrak, T., 2020. What Makes a Winner? Analyzing Team Statistics to Predict Wins in the NFL. https://core.ac.uk/download/pdf/326836349.pdf