Finding Fantastic Feats of Football in the NFL

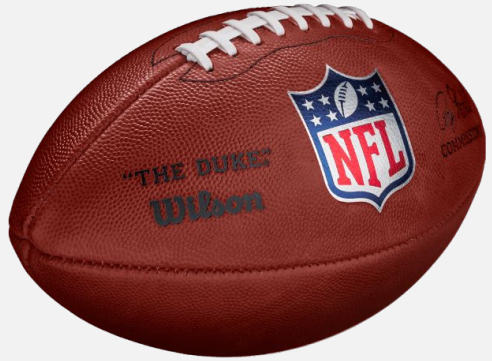
Alex Plazas, Benjamin Price, Raul Ramos

CSPB 4502

Prof. Kristy Peterson

DESCRIPTION

Football is a dynamic, complex sport rich in statistical information. Detailed play-by-play data enables a granular analysis of player performance and how that performance impacts team success. How does field position impact win probability? Does having top players always guarantee success? In what situations should you "go for it" on fourth down? Our project aims to answer these questions in addition to uncovering unique relationships and examining trends in football data gathered from the National Football League (NFL).



PRIOR WORK

"...members of the NFL's football data and analytics team will share updates on league-wide trends in football data…and provide an inside look at how the NFL uses data-driven insight to improve and monitor player and team performance."

- THE EXTRA POINT, NFL Football Operations
https://operations.nfl.com/gameday/analytics/stats-articles/

"It's about translating that data ASAP and being very, very in tune with the numbers. You can't be a year behind, you can't be a month behind…"

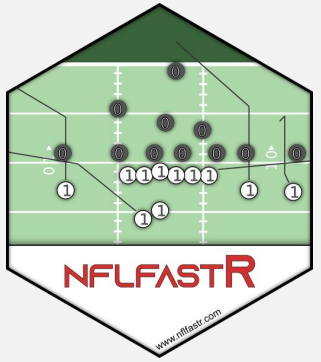
- The NFL's Analytics Revolution Has Arrived, The Ringer
https://www.theringer.com/nfl/2018/12/19/18148153/nfl-analytics-revolution

DATASETS

Our primary dataset is comprised of NFL play-by-play data for the 2009 to 2018 seasons. The set contains 700,000+ data points with nearly 250 unique attributes. While the data can be easily downloaded from Kaggle, the team has decided to save the data in a group Google Drive, as well as the team Github:

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

Additionally, the team has identified alternative data sources in the form of an R package to scrape and output play-by-play data: https://www.nflfastr.com/



PROPOSED WORK

- Data preprocessing:
  - Part of the preprocessing work will involve loading the play-by-play .csv file as a DataFrame, enabling further data automations.
- Data cleaning:
  - Many of the attributes for each data point are N/A or missing altogether, requiring selective cleaning depending on the attribute.
- Data aggregation:
  - If a question we seek to answer requires data outside of our current season range, the team will have to combine sets and ensure set attributes are aggregated correctly.
- Attribute combining:
  - Some of the attributes given for each play-by-play point are similar in nature (i.e. forced fumbles, fumble forced, fumble not forced), and can be combined to simplify the analysis process.

LIST OF TOOLS

- Our primary language will be Python (>3.0).
- Python packages Numpy, Pandas, and Matplolib for data processing, analysis, and visualization.
- Secondary/alternative language will be R for fetching data from other sources.
- Tableau, PowerBI, and Excel for visual exploratory analysis
- Google Slides for presentations and project updates

EVALUATION

- Create a regression model using the data from the 2009-2016 seasons as "training" data, and test the model using the 2017-2018 seasons.
- Create 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.
- Assign each team a probability of winning the Super Bowl for that season, and test if the model is accurate in predicting playoff success.