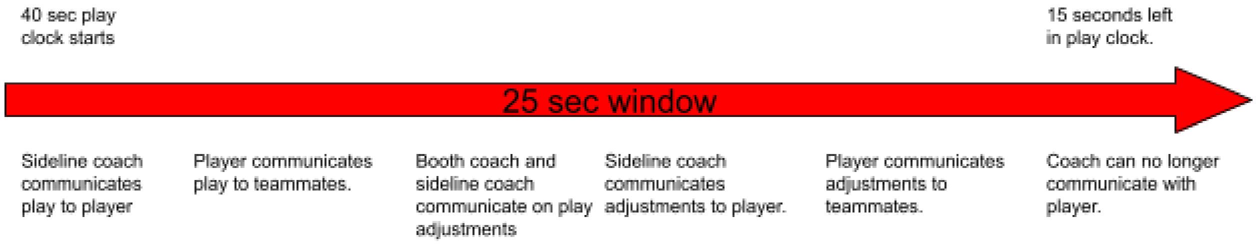
A Data Science Approach to NFL Pre-Snap Evaluation

Project Proposal
Ranier Gran
DS 504 - Fall 2021

Abstract:
Sports analytics is a relatively new and constantly growing field. With the help of advanced camera technology and a new emphasis being put on data gathering and management, more insight can be gained than ever before. This project aims to analyze captured game and player data in an attempt to predict play outcome for coaching decision making.

Background:
Data and analytics have been a major talking point in the world of sports for the past few years. From illegal actions like videotaping sidelines and banging on trash cans to less frowned upon, analytical decision making approaches, like going for it more frequently on 4th down, teams across all professional leagues will do nearly anything to gain the slightest competitive edge. Many league rules have been specifically designed to allow coaches to have more control over their team. Do you ever wonder why there is a green dot on the back of some player's helmets when you're watching an NFL game? This dot is to mark which player has a speaker in their helmet for their coach to communicate with them. In order to make this more manageable and fair, the NFL only permits one player on defense and one player on offense (almost always the quarterback) this one-way listening device, allowing the player to hear their coach without the ability to communicate directly back themselves. It is also important to note that communication is automatically shut off with 15 seconds left in the play clock to prevent the coach from micromanaging their players on the field. Lastly, the coach speaking with the player through their headset must be a sideline coach, not a coach in the booth ensuring they do not have a perfect birds eye view of the game. With this information in hand we can create a clear picture of the tight decision window for play calling and adjustment communication.


40 sec play clock starts
15 seconds left in play clock.
25 sec window
Sideline coach communicates play to player
Player communicates play to teammates.
Booth coach and sideline coach communicate on play adjustments
Sideline coach communicates adjustments to player.
Player communicates adjustments to teammates.
Coach can no longer communicate with player.

Working with only 25 seconds, play recognition must be fast and efficient for coaches to communicate changes to their team.

This is where big data analytics comes in. On any given play there are dozens of variables to take into account. Advanced analytics allows teams to better prepare before games in scouting, player selection, and gameplanning, but can also be utilized to make better decisions on the fly. The main focus of this project will be to improve in game decision making with play success prediction with minimal computation time. The primary objective of our decision algorithm will be focused on pre-snap, defensive adjustments but analysis will also be done for offense and pre-game planning.

Dataset:
This project will utilize a kaggle dataset from the NFL Big Data Bowl 2021 (https://www.kaggle.com/c/nfl-big-data-bowl-2021/overview). This includes four distinct data structures.

Dataset	Description	Key Value(s) Contained
games.csv	Contains data on which teams are playing	gameID
players.csv	Contains player physical attributes and other qualitative variables.	nflID (playerID)
plays.csv	Contains information on plays in the games	playID, gameID
week[#].csv	Contains player tracking data for plays broken into multiple files for each week	gameID, playID, nflID (playerID), frameID

Games Data

Columns	Description
gameId	Game identifier, unique (numeric)
gameDate	Game Date (time, mm/dd/yyyy)
gameTimeEastern	Start time of game (time, HH:MM:SS, EST)
homeTeamAbbr	Home team three-letter code (text)
visitorTeamAbbr	Visiting team three-letter code (text)
week	Week of game (numeric)

Player Data

Columns	Description
nflId	Player identification number, unique across players (numeric)
height	Player height (text)
weight	Player weight (numeric)
birthDate	Date of birth (YYYY-MM-DD)
collegeName	Player college (text)
position	Player position (text)

displayName	Player name (text)

Plays Data

Columns	Description
gameId	Game identifier, unique (numeric)
playId	Play identifier, not unique across games (numeric)
playDescription	Description of play (text)
quarter	Game quarter (numeric)
down	Down (numeric)
yardsToGo	Distance needed for a first down (numeric)
possessionTeam	Team on offense (text)
playType	Outcome of dropback: sack or pass (text)
yardlineSide	3-letter team code corresponding to line-of-scrimmage (text)
yardlineNumber	Yard line at line-of-scrimmage (numeric)
offenseFormation	Formation used by possession team (text)
personnelO	Personnel used by offensive team (text)
defendersInTheBox	Number of defenders in close proximity to line-of-scrimmage (numeric)
numberOfPassRushers	Number of pass rushers (numeric)
personnelD	Personnel used by defensive team (text)
typeDropback	Dropback categorization of quarterback (text)
preSnapHomeScore	Home score prior to the play (numeric)
preSnapVisitorScore	Visiting team score prior to the play (numeric)
gameClock	Time on clock of play (MM:SS)
absoluteYardlineNumber	Distance from end zone for possession team (numeric)
penaltyCodes	NFL categorization of the penalties that ocurred on the play. Multiple penalties on a play are separated by a ; (text)
penaltyJerseyNumber	Jersey number and team code of the player commiting each penalty. Multiple penalties on a play are separated by a ; (text)
passResult	Outcome of the passing play (C: Complete pass, I: Incomplete pass, S: Quarterback sack, IN: Intercepted pass, text)
offensePlayResult	Yards gained by the offense, excluding penalty yardage (numeric)
playResult	Net yards gained by the offense, including penalty yardage

	(numeric)
epa	Expected points added on the play, relative to the offensive team. Expected points is a metric that estimates the average of every next scoring outcome given the play's down, distance, yardline, and time remaining (numeric)
isDefensivePI	An indicator variable for whether or not a DPI penalty ocurred on a given play (TRUE/FALSE)

Tracking Data	
Columns	**Description**
time	Time stamp of play (time, yyyy-mm-dd, hh:mm:ss)
x	Player position along the long axis of the field, 0 - 120 yards. (numeric)
y	Player position along the short axis of the field, 0 - 53.3 yards. (numeric)
s	Speed in yards/second (numeric)
a	Acceleration in yards/second^2 (numeric)
dis	Distance traveled from prior time point, in yards (numeric)
o	Player orientation (deg), 0 - 360 degrees (numeric)
dir	Angle of player motion (deg), 0 - 360 degrees (numeric)
event	Tagged play details, including moment of ball snap, pass release, pass catch, tackle, etc (text)
nflId	Player identification number, unique across players (numeric)
displayName	Player name (text)
jerseyNumber	Jersey number of player (numeric)
position	Player position group (text)
team	Team (away or home) of corresponding player (text)
frameId	Frame identifier for each play, starting at 1 (numeric)
gameId	Game identifier, unique (numeric)
playId	Play identifier, not unique across games (numeric)
playDirection	Direction that the offense is moving (text, left or right)
route	Route ran by offensive player (text)

The dataset has been curated already focusing on passing plays. As the primary focus of the data is on the passing game, offensive and defensive line players are not included in the tracking set. Significant work will be required to join the above datasets in an intelligent manner.

The initial plan of action will be to consolidate the tracking data to gather information on player speed and acceleration. Additional data may be gathered for further analysis such as team standings, weather, and playoff implications.

Methodology:

As mentioned, work on this project will begin with a data cleaning and data engineering phase. Necessary text fields will be cleaned and one-hot encoded. Tracking data will be manipulated to gain relevant player level data such as average/max speed and acceleration. Finally, data tables will be joined together for analysis. From here, exploratory data analysis will be used to identify interesting data features and play trends. After diving deep into the data, several prediction exercises will be discussed. Play result prediction from a defensive perspective, and offensive perspective will most likely behave differently. Feature importance will be analyzed to determine coaching strategies in pre-game meetings.

Project work will follow the below general schedule.

Week	**Goal**	**Deadline**
1	Data downloaded or accessed for size and streaming	
2	Data cleaned and one hot encoded	Final Project Status Report #1
3	Tracking data analyzed and consolidated	
4	Data joined, EDA begins	
5	EDA completed	
6	Prediction exercises begin	Final Project Status Report #2
7	Finalize predictions and cross validation	
8	Finalize presentation deck and submission	Final Project Deliverable (Strategy Consulting Presentation Deck)

Data is currently expected to be gathered from kaggle and stored locally. Depending on size, data streaming may be required to facilitate efficient analysis. Generally, python will be utilized for data analysis and prediction exercises however, if runtime and data size becomes an issue, distributed, parallel processing such as spark or hadoop will be used to accelerate computing.

Resources:

Background information has been gathered from previous research as a part of humanities and arts capstone as well as other sources. More background research will be considered as work continues.

Capstone - Sports Analytics The Story Behind the Numbers:
(https://drive.google.com/file/d/13F2bQRG8Zz80xTZIVGL0ldU-rpSGrcgT/view?usp=sharing)
Headset Rules:
(https://gethypedsports.com/do-football-players-have-speakers-microphones-in-their-helmet/#:~:text=There%20are%20a%20set%20of,what%20the%20coach%20is%20saying.)