Tim Hulak

Prof. Jillian K. Lando

IST-718 Big Data Analytics

11/07/2021

(Data) Mining for Gold: Football Analytics for San Francisco 49ers

Introduction

Each football season, the goal of every NFL team is to win the Super Bowl. In order to even have a chance to appear in the Super Bowl, let alone the playoffs, a team must first play well enough to win their division. The San Francisco 49ers are an American NFL football team based out of the San Francisco Bay Area in California. Their offensive performance has been declining in recent years and to complicate matters further, they are in the fiercely competitive NFC West division. The goal of this analysis is to identify strengths and opportunities that may lead to a divisional title, a deep playoff run, and hopefully a Super Bowl victory.

Business Problems

The ubiquitous business problem that football teams attempt to answer is *"How do we win football games and ultimately win the Super Bowl."* For the 49ers in particular, the area of the game that appears to be lacking is passing the ball. The 49ers seem to perform well on defense and when running the ball, but the passing game has been a struggle for the team in recent years. This analysis will aim to answer the following questions (as well as any questions that arise from exploratory data analysis):

- What passing plays/formations achieve the greatest success (measured in yards gained) against each of the defensive formations?
- Which Wide receiver routes achieve the greatest success (measured in yards gained) against each of the defensive formations?
- How do down, distance, and game situation (score, time left on the clock, etc.) effect passing and how can the 49ers improve in certain situations?
- How can the 49ers surpass divisional opponents in order to win the NFC West Division and appear in the playoffs?

The Data

The primary data has been provided by the NFL via Kaggle (https://www.kaggle.com/c/nfl-big-data-bowl-2021/data). This data contains *games.csv*, *players.csv*, *plays.csv*, and 17 CSV datasets for performance in each of the 17 regular season weeks of games. In addition, records for historical wins and losses have been scraped from the NFL website (https://www.nfl.com/standings/division/2021/REG). Lastly, supplemental performance data has been obtained from the NFL Savant website (http://www.nflsavant.com/about.php) in order to fill in some gaps that the NFL datasets may have. Data from the 2018 season will be analyzed since each of the data sources converge on that year. This will likely present the best picture of the conditions for a typical NFL season.

Dataset Descriptions

games.csv – High level data that reflects the game schedule for the 2018 season

253 Rows by 6 columns

- 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)

players.csv – Data on the players in the NFL

1,303 Rows by 7 columns

- 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.csv – Detailed play-by-play information on games

19,239 Rows by 27 columns

- 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. For purposes of this contest, the most important penalties are Defensive Pass Interference (DPI), Offensive Pass Interference (OPI), Illegal Contact (ICT), and Defensive Holding (DH). 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)

week[n].csv - Player tracking data from all passing plays

Between 932,240 – 1,231,793 Rows by 19 columns

- 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. See Figure 1 below. (numeric)
- y: Player position along the short axis of the field, 0 - 53.3 yards. See Figure 1 below. (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)

NFL.db

A SQLite3 database that contains data from the NFL Savant website and data scraped from the NFL website using pandas pd.read_html() function. The tables are as follows:

- fulldata: All data combined
- year2013: Data from the year 2013
- year2014: Data from the year 2014
- year2015: Data from the year 2015
- year2016: Data from the year 2016
- year2017: Data from the year 2017
- year2018: Data from the year 2018
- year2019: Data from the year 2019
- year2020: Data from the year 2020
- year2021: Data from the year 2021
- records: Data for team records (wins, losses, etc.)

The tables that will be used from this database are:

- year2018: Data from the year 2018 (*45,016 Rows by 42 columns*)
  - GameId *(int64)* : Identifier of a particular game
  - GameDate *(object)* : Date of the game
  - Quarter *(int64)* : Quarter where the play took place
  - Minute *(int64)* : Minute of the Quarter where the play took place
  - Second *(int64)* : Second of the Quarter where the play took place
  - OffenseTeam *(object)* : Offensive team on the play
  - DefenseTeam *(object)* : Defensive team on the play
  - Down *(int64)* : Down of the play (1-4)
  - ToGo *(int64)* : Yards to go for a First Down
  - YardLine *(int64)* : Yardline the ball is placed on
  - SeriesFirstDown *(int64)* : Is this play the first down on the series
  - NextScore *(int64)* : ?
  - Description *(object)* : Text description of the play
  - TeamWin *(int64)* : ?
  - SeasonYear *(int64)* : Year of the season
  - Yards *(int64)* : Yards gained on the play
  - Formation *(object)* : Formation of the offense
  - PlayType *(object)* : Play type (pass, run, etc.)
  - IsRush *(int64)* : Is the play a rushing play?
  - IsPass *(int64)* : Is the play a passing play?
  - IsIncomplete *(int64)* : Was the throw completed?
  - IsTouchdown *(int64)* : Was the result of the play a touchdown?
  - PassType *(object)* : Length of pass (long, short, mid)
  - IsSack *(int64)* : Was the QB sacked?
  - IsChallenge *(int64)* : Was the paly challenged?
  - IsChallengeReversed *(int64)* : Did the challenge result in reversal of the ruling on the field?
  - Challenger *(float64)* : Which team challenged the play?
  - IsMeasurement *(int64)* : ?
  - IsInterception *(int64)* : Was the offense intercepted on a pass?
  - IsFumble *(int64)* : Was the ball fumbled?
  - IsPenalty *(int64)* : Was there a penalty on the play?
  - IsTwoPointConversion *(int64)* : Was the play a 2-pt conversion?
  - IsTwoPointConversionSuccessful *(int64)* : Was the 2-pt conversion successful?

    - RushDirection *(object)* : Which direction dod the running back rush (left, middle, right)
    - YardLineFixed *(int64)* :
    - YardLineDirection *(object)* :
    - IsPenaltyAccepted *(int64)* : Was the penalty accepted and applied?
    - PenaltyTeam *(object)* : Which team commited the penalty?
    - IsNoPlay *(int64)* : Was the play a No-Play?
    - PenaltyType *(object)* : Waht was the penalty type?
    - PenaltyYards *(float64)* : How many yards were lost on the penalty?
    - Week *(object)* : Week of the season
    - GameDay *(object)* : Day of the week the game occurred
- records: Data for team records (wins, losses, etc.) (*288 Rows by 30 columns*)
    - Team *(object)* : Full team name
    - Abbrv *(object)* : Team abbreviation
    - Year *(int64)* : Year of the record data
    - Win *(int64)* : Number of wins
    - Loss *(int64)* : Number of losses
    - Tie *(int64)* : Number of ties
    - Win_Pct *(float64)* : Winning Percentage
    - Pts_For *(int64)* : Points For (total points the team has scored in the season)
    - Pts_Against *(int64)* : Points Against (total number of points scored against the team in the season)
    - Net Pts *(int64)* : Pts_For minus Pts_Against
    - Home_Win *(int64)* : Number of wins at home
    - Home_Loss *(int64)* : Number of losses at home
    - Home_Tie *(int64)* : Number of ties at home
    - Road_Win *(int64)* : Number of wins on the road
    - Road_Loss *(int64)* : Number of losses on the road
    - Road_Tie *(int64)* : Number of ties on the road
    - Division *(object)* : Which Division a team belongs to
    - Division_Win *(int64)* : Number of wins against divisional opponents
    - Division_Loss *(int64)* : Number of losses against divisional opponents
    - Division_Tie *(int64)* : Number of ties against divisional opponents
    - Div_Pct *(float64)* : Winning Percentage against divisional opponents
    - Confrence *(object)* : Which conference a team belongs to
    - Confrence_Win *(int64)* : Number of wins against conference opponents
    - Confrence_Loss *(int64)* : Number of losses against conference opponents
    - Confrence_Tie *(int64)* : Number of ties against conference opponents
    - Conf_Pct *(float64)* : Winning Percentage against conference opponents
    - Non-Conf *(object)* : Win-Loss-Tie record against opponents outside of teams conference
    - Streak *(object)* : Most recent win/loss streak
    - Last_5_Win *(int64)* : Number of wins in last 5 games of the season
    - Last_5_Loss *(int64)* : Number of losses in last 5 games of the season
    - Last_5_Tie *(int64)* : Number of ties in last 5 games of the season