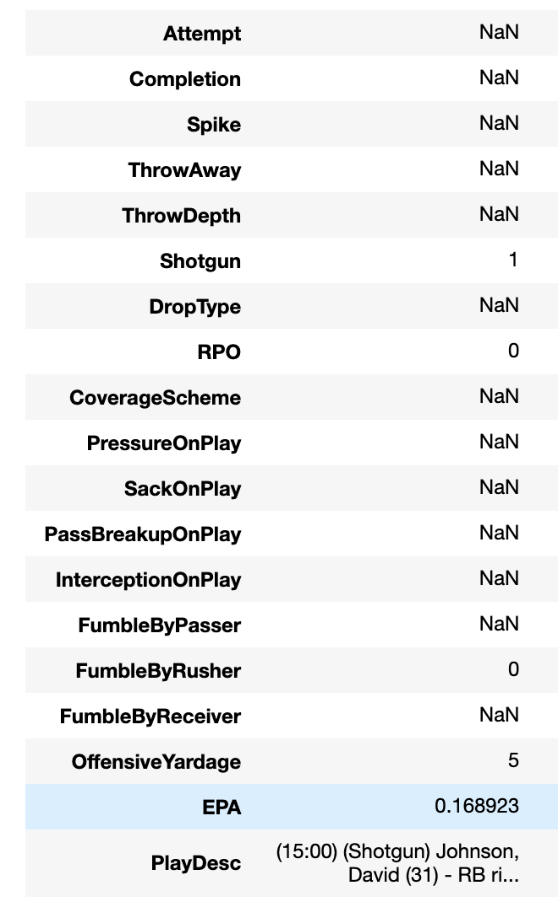
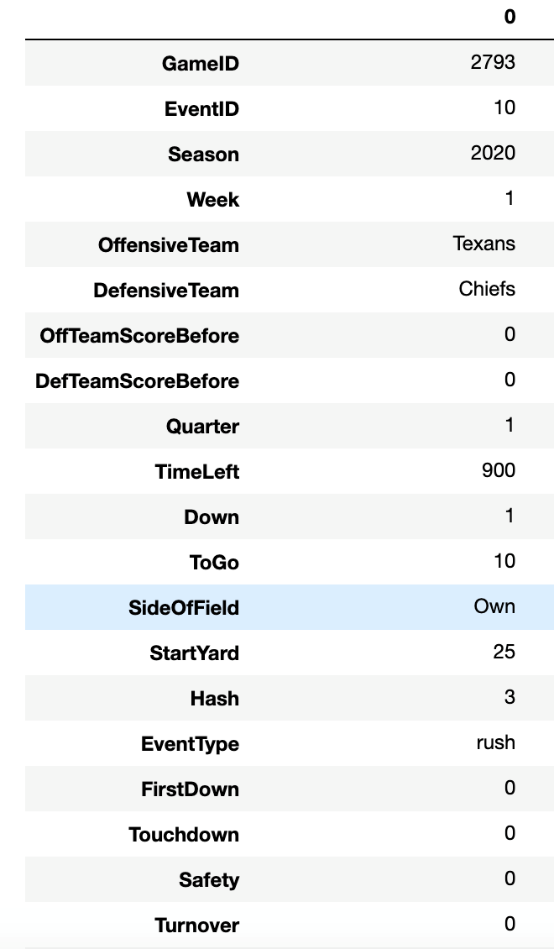
DSC 540: Group 3 Final Project Proposal

Sam Hoppen, Christopher Wieneke, Daniel O’Brien, West Xie

1. Our dataset comes from [Sports Info Solutions](https://github.com/SportsInfoSolutions/AnalyticsChallenge2021). The dataset includes all plays run by all NFL teams in the 2020 season. The data was collected by recording certain information from each play, and we intend to use this information to draw conclusions from specific receiver routes and defensive coverages in the NFL.
2. There are 192988 rows and 50 columns. We started with two separate datasets, one that documented each play that occurred throughout the season, and one dataset that documented what each skill position player was doing during each play.

Merged Dataset First Column:



1. Each observation is a play in the NFL in the 2020 season.
2. Below is a table explaining each of the variables in our dataset:

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Variable Type* | *Description* |
| *GameID* | *Nominal* | Unique ID for the game |
| EventID | Nominal | An ID for each play of each game, typically (not always) starting at 5 and incrementing by 5 |
| Season | Nominal | Year of Season |
| Week | Nominal | Week of Season |
| OffensiveTeam | Categoric | Name of Offensive Team |
| DefensiveTeam | Categoric | Name of Defensive Team |
| OffTeamScoreBefore | Numeric | The offensive team’s score prior to the given play |
| DefTeamScoreBefore | Numeric | The defensive team’s score prior to the given play |
| Quarter | Numeric | Quarter of Game |
| TimeLeft | Numeric | Time remaining (in seconds) in the current quarter |
| Down | Numeric | Down for current play |
| ToGo | Numeric | Yards to go at start of play |
| SideOfField | Categoric | The side of the field the play started on (“Own” or “Oppo”) |
| StartYard | Numeric | Yard line the play started on (paired with SideOfField to get field position) |
| Hash | Categoric | The hash from which the ball was snapped for the play |
| EventType | Categoric | The type of play, in terms of how it ended up. That means that QB scrambles are called runs. Plays that were modified as a result of a review have “challenge” preceding the event type. |
| FirstDown | Nominal | 1 if the play gained a first down, 0 otherwise |
| Touchdown | Nominal | 1 if there was a touchdown on the play (for either team), 0 otherwise |
| Safety | Nominal | 1 if there was a safety on the play (for either team), 0 otherwise |
| Turnover | Nominal | 1 if there was a turnover on the play, 0 otherwise |
| Attempt | Nominal | 1 if there was a pass attempt on the play, 0 otherwise |
| Completion | Nominal | 1 if pass was completed |
| Spike | Nominal | 1 if the QB spiked the ball to kill the play, 0 otherwise |
| ThrowAway | Nominal | 1 if the QB threw the ball away intentionally, 0 otherwise |
| ThrowDepth | Numeric | The number of yards downfield a pass was thrown |
| Shotgun | Nominal | 1 if the QB lined up in shotgun pre-snap, 0 otherwise |
| DropType | Categoric | Categorical variable describing the type of drop the QB took on a pass play. Most are in terms of the number of steps back he took, but there are other items for designed rollouts or screens. |
| RPO | Nominal | 1 if the play featured a (pre-or-post-snap) run-pass option. SIS defines this as the offensive line and backfield executing a run concept while at least one receiver is running a route. |
| CoverageScheme | Categoric | The coverage scheme being employed by the defense on a pass play |
| PressureOnPlay | Nominal | 1 if the QB was hurried, hit, knocked down, and/or sacked on a pass play |
| SackOnPlay | Nominal | 1 if sack on the play |
| PassBreakupOnPlay | Nominal | 1 if the pass was intercepted or in some other way broken up by a defender |
| InterceptionOnPlay | Nominal | 1 if the pass was intercepted |
| FumbleByPasser | Nominal | 1 if the passer fumbled the ball |
| FumbleByRusher | Nominal | 1 if the rusher fumbled the ball |
| FumbleByReceiver | Nominal | 1 if the receiver fumbled the ball |
| OffensiveYardage | Numeric | Yards gained or lost by the offense (passing or rushing) on the play |
| EPA | Numeric | Expected Points Added on the play (by SIS’s model). Importantly, on plays with a penalty following them (e.g. roughing the passer), the penalty’s value is not included. |
| PlayDesc | Categoric | Description of the play |

1. Overall, our goal of the project will be to evaluate and identify what the most effective passing strategies in the NFL are. This will include evaluating different route concepts and coverage schemes to see which offer the highest expected value for an offense. This could help NFL teams optimize their passing-game strategy while providing some insights into
2. We plan to use a plethora of machine learning techniques to help answer our problem. Specifically, using classification or clustering to help identify whether a play will result in a touchdown or not or if the play will result in a turnover or not. Additionally, regression will be used to build our own expected points model that will be compared against the expected points values that our data provides. As we continue to explore the data, we will introduce different machine learning techniques into our analysis to further improve our analysis.
3. The biggest challenge for this dataset is the limitation to one season. Each team will only have two games of sample sizes to predict or to classify the optimal route or scheme for each of their opponents. In addition, not all teams fall for the same route or scheme, so the possibility of low prediction or classification accuracy is relatively high.
4. Below is a summary of the research questions we plan to attack:
   1. Which routes (or route combinations) are most effective? Which route combinations most likely lead to touchdowns? Which route combinations lead to highest yardage gains?
      1. Possible analysis: build a model to evaluate route effectiveness and then explore which routes are the best.
   2. Which coverage schemes are most effective on defense?
      1. Possible analysis: similar to the prior problem, build a model to evaluate coverage scheme effectiveness and then explore which schemes are the best. In addition, insight from plots could help solidify conclusions from the model.
   3. Do different passing nuances (i.e., drop type, RPO) significantly impact a play’s level of success?
      1. Possible analysis: run a correlation or ANOVA analysis to see if there are major differences in play success for each passing nuance. There is also room to run either regression or classification for either yards or whether it was a touchdown or not depending on the definition of success
   4. Which route (or route combinations) are most likely to be intercepted? Additionally, what factors are most likely to lead to an interception (i.e., coverage patterns, route combinations or something else)?
      1. Possible analysis: create a random forest model to identify which plays will be an interception or not. Logistic regression may also provide some insights when looking at probability of an interception for added insight.
   5. Does route variety affect player receiving totals or points (Mix of routes v. all slants)?
      1. Possible analysis: run a cluster analysis on various routes and route combinations to see which routes have similar levels of success and traits (depth of target, yards gained, etc.).