

Here are things you need to have:

- Do you have at least 3 data visualizations? Include them in your write up.
  - Tables of summary statistics, correlation coefficients, linear regression coefficients all count as a visualization as long as they are well-formatted and have a title. Screenshots of raw data don't count unless you collected your own data.
  - Examples: scatter plots, histograms, bar charts, line plots, visualization of your decision tree classifier
- Do you have at least 2 custom functions in your code? Don't include your code in the write up, just submit the notebook file.
- Did you build and evaluate at least one model?
- Did you investigate a hypothesis or a question?
- Do you work with at least 3 variables?
  - Fitting a model with 3 features works!
  - Even if you only end up investigating 1~2 features for your model, if you did data exploration with 3+ features and mention why you narrowed the list down for further analysis, that's ok.
- Did you discuss (but not necessarily implement) the 7 stages of the data science lifecycle? See below:

## **1. Define the problem**

For our research project, we chose to investigate the factors that contribute to passing plays in the National Football League (NFL) resulting in positive expected points added (EPA), to identify which factors ultimately impact EPA the most. Expected points (EP) is a metric used in the NFL that represents the number of points a team is expected to score on the current drive. The expected points *added* metric measures how a team's expected points value changes on an individual play. Each play in an NFL game has many different moving parts, such as offensive and defensive formation, personnel, down, and distance to the first down marker. We were interested in developing a method to determine which factors on the football field are most important to determining a successful passing play, using EPA as our measure of success.

## **2. Data collection**

We worked with a dataset from the popular data analytics website Kaggle. The author of the dataset is Arya Shah. In particular, we used the file named 'plays.csv', which contained information on individual plays during the 2018-19 NFL regular season. This play-by-play data was likely collected by statisticians and analysts from various NFL organizations during actual games, and then made publicly available through league databases or popular sports analytics

websites. This dataset initially contained 19,239 samples and 27 different features. Again, in this case, each sample represents an individual play run by an NFL team during the 2018-19 regular season. Some of the seemingly most relevant features in the original dataset were the current down on the play ran, the number of yards needed for a first down, the absolute yard line number, the result of the pass thrown (completion, incompleteness, interception), and both the offensive and defensive personnel. Some features such as the description of the play were difficult to quantify, and others, like ones regarding identification of the specific game and play, were irrelevant to predicting EPA on any given play. One of the major limitations of the dataset is that it solely contains information on passing plays! We cannot draw any conclusions regarding rushing or special teams plays and how they relate to EPA.

### **3. Data preparation**

#### **a. If you dropped any data, describe why and how you did it.**

We removed data regarding penalties because that isn't something an offense can always control. While technically an offense can cause a defense to commit a penalty, that isn't always the case, and much of the time it's purely because of a mistake of a defender, so we decided that it was better to drop all plays that resulted in penalties. We did this by only including the rows where there was a null value in penaltyCodes.

#### **b. If you normalized any data, describe why and how you did it.**

We normalized all of our data because, with over 30 variables used, we have many different scales on our data. We subtracted each data point by the column's mean and divided it by the column's standard deviation to get each variable on the same scale.

#### **c. If you converted any categorical data to numeric, describe why and how you did it.**

We converted a few categorical variables to numerical variables. We expanded offenseFormation and typeDropback to binary columns of all possible values in the dataset using the pandas function `get_dummies()`. We also changed personnelO and personnelD into variables with the number of RBs, WRs, and TEs for offense and DBs, LBs, and DLs for defense. We did this with a custom function that we wrote called `read_personnel()`. This function splits the string of personnel into the given positions, takes the number listed before the position, and uses that as the value. Another categorical variable we changed to numeric was passResult, where we changed the types of results of a pass (incompletion, completion, sack, and interception) into 0, 1, 2, and 3 to make it a usable variable. We made these changes because we thought these would be valuable variables to take into account in our model, as different formations, personnel, and types of pass dropbacks have different sets of plays called with them so we wanted to see their effect on our model.

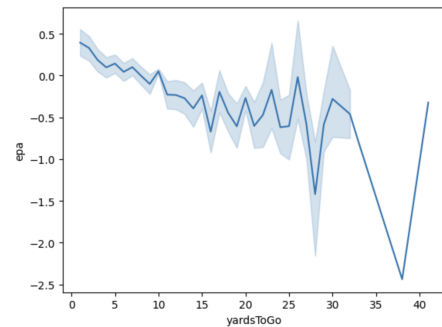
**d. Describe any other data preparation steps you may have taken.**

We did other data preparation by dropping columns that had no impact on predicting EPA, such as gameId and playId. We also created multiple functions to help build our model. One function, called to\_seconds(), took the gameClock variable and changed the value into strictly seconds remaining in a quarter to make it easier to work with. Another function, called read\_personnel(), extracted the key information in the variables 'personnelO' and 'personnelD' to use for the regression model.

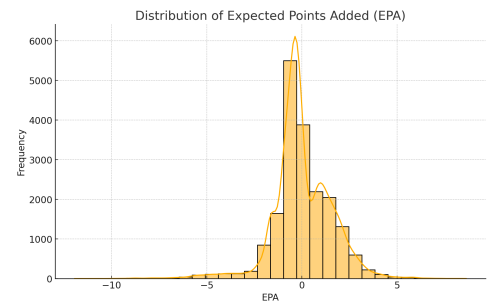
**4. Data exploration**

**a. Include your data exploration from your project proposal.**

This line plot models the relationship between 'yardsToGo' and 'epa.' We were interested in how 'epa' relates to other variables, so we created a simple line plot with only one of our potential predictor variables to help understand our response variable. In this plot, there appears to be a negative correlation between 'yardsToGo' and 'epa.'



This histogram shows the distribution of EPA values across all plays in the dataset. The right side of the histogram has a shooter tail meaning that plays with high EPA are less frequent. There is a negative skew showing that plays with negative EPA occur frequently. Most of the plays hover around 0 EPA meaning that the plays have a neutral effect on scoring potential.



**b. Include any new data exploration you may have done to help inform your model choice, but not required to receive full credit.**

**5. Model building**

- a. This step and the Model Evaluation step should be the heart of the final write-up.**
- b. Describe the model you chose to use and why.**
- c. Describe the features or variables you used and why. Remember that you need to investigate at least 3 variables and their relationships. Clearly state which variables (out of all of them in your dataset) you chose to investigate.**

**d. You are also welcome to build multiple models and compare how they perform, but not required to receive full credit.**

Since we wanted to predict EPA values, we chose to use a regression model rather than a classification model. We chose to use a multiple linear regression because football is a complicated sport and many variables have an impact on each play, as our model ended up having 33 variables:

- **yardsToGo, down\_2, down\_3, down\_4, absoluteYardlineNumber:** We used these variable because, together, they extremely important in determining the likelihood of a plays success. For example, converting a 2nd & 1 on a team's own 41 yard-line is going to result in a much lower EPA than converting a 3rd & 13 on an opponents 49 yard-line because the former is far more probable and a successful conversion has far less of an effect on a team's win probability.
- **defendersInTheBox, numberOfPassRushers:** These two variables are important for setting the context of a play. If there are many defenders in the box, a play-action pass is likely to be more successful because defenses are expecting run. The number of pass rushers is critical to know because whether or not a defense blitzes can change the likiehood of an explosive play and a sack. Coaches want to know how to beat blitzes and how to get defenses out of heavy boxes in rushing downs, so these variables are extremely useful.
- **preSnapVisitorScore, preSnapHomeScore:** The score in the game greatly affects playcalling decisions and how teams operate.
- **passResult, offensePlayResult:** These two variables relate to the result of the play, which is also what EPA measures. Since EPA considers more context than yards or a binary result of a pass, like a completion or incompletion, it's worth looking at how correlated these variables are with EPA and if these basic stats have a significant difference in a certain play result compared to more advanced stat like EPA.
- **quarter\_2, quarter\_3, quarter\_4, quarter\_5, secondsRemaining:** Time remaining and quarter greatly influence an offense's playcalling. For example, a team trailing by a touchdown in the fourth quarter with under two minutes remaining is likely to pass most, if not every, play. A defense therefore expects this and will play adequate personnel and coverage to match this expectation.
- **O\_RB, O\_WR, O\_TE, D\_DB, D\_LB, D\_DL:** These variables track positions that vary in numbers based on the personnel and playcall, so these are relevant to track. For example, defenses are likely to play more DBs when they expect offenses to pass and offenses are likely to play more WRs in obvious passing downs. Offenses also want to catch defenses by surprise and pass in personnel groupings that would lead defenses to expect run, so we figured these variables would be important to our model.

- **typeDropback\_DESIGNED\_ROLLOUT\_RIGHT, typeDropback\_SCRAMBLE, typeDropback\_SCRAMBLE\_ROLLOUT\_LEFT, typeDropback\_SCRAMBLE\_ROLLOUT\_RIGHT, typeDropback\_TRADITIONAL:** These variables track the type of dropback for the offenses' QB. Generally, it is assumed that rolling a quarterback out of the pocket makes it easier for him than in a traditional dropback because it shortens the field and makes reads easier, so we wanted to see if this assumption holds in our model.
- **offenseFormation\_I\_FORM, offenseFormation\_JUMBO, offenseFormation\_PISTOL, offenseFormation\_SHOTGUN, offenseFormation\_SINGLEBACK, offenseFormation\_WILDCAT:** These variables classify an offenses' formation. Offenses are gravitating more towards the shotgun in the modern NFL, as many veteran QBs like to dissect a defense from the shotgun. It's generally thought of as a passing formation, so defenses are likely to combat this by expecting pass and playing coverage. Offenses are constantly looking for innovative ideas, so seeing what formations have the largest effect on EPA could be extremely useful for coaches.

## 6. Model evaluation

- This step and the Model Building step should be the heart of the final write-up.**
- Describe the hypothesis you tested or question you investigated. What does your model say about your hypothesis or question?**
- For evaluation, you can use a technique appropriate for the model you have chosen, such as simulation, sampling, comparing distributions, or calculating accuracy.**
- If you're looking for more analysis to do, you might also evaluate the effects of outliers, model complexity (e.g. tree depth, degree for polynomial regression), or effect of random train/test split. Not required for full credit.**
- You are also welcome to build multiple models and compare how they perform, but not required to receive full credit.**

For our research question, we looked at many variables and how they affect Expected Points Added on passing plays. Looking at the results of our model, we can see a few observations. First of all, our model performed decently but was overall unspectacular. The training dataset had an  $R^2$  value of .584 and the test data had an  $R^2$  value of .575. This means that our model has some predictive value but has its share of flaws. Looking at the coefficient values, we can see that playResult is by far the most predictive of EPA, which matches what we'd expect, as the more yards a play results in, the more expected points a play should be worth. Looking at other variables that displayed larger magnitudes, we can see defensive personnel seem to have a major impact on EPA. All of D\_DL, D\_LB, and D\_DB have coefficients of a high magnitude relative

to the rest of the variables. This suggests that if defenses put unbalanced personnel groupings on the field, with a lot of DL for example, offenses tend to take advantage of it and generally have more successful EPA. 3rd and 4th downs tend to have a relatively high effect relative to other variables in this model, which makes sense considering how much a successful conversion could impact a team's win probability. One trend we found surprising was that the different offensive formations tended to be near the bottom in terms of coefficient magnitude, suggesting a relatively small effect on EPA for passes.


 Train R-squared score: 0.584 Test R-squared score: 0.575	
	Coefficient vals
offensePlayResult	1.169472
D_DL	0.631025
D_LB	0.627979
D_DB	0.446275
passResult	-0.229141
yardsToGo	-0.198160
typeDropback_TRADITIONAL	-0.085853
down_3	-0.076700
down_4	-0.070661
typeDropback_SCRAMBLE_ROLLOUT_RIGHT	-0.066420
typeDropback_SCRAMBLE_ROLLOUT_LEFT	-0.048861

Table 1. Train and Test R-squared, most significant coefficient values in terms of magnitude

An overall trend for our model suggests, due to the nature of the sport of football and how interconnected and complicated it is, that none of the variables are obviously much more significant in terms of predicting EPA. Except for playResult, which measures the result of the play in terms of yards, all of the other variables have coefficients that are very close to one another. playResult also differs from these other variables because it measures the result of the play, so it isn't a variable that can be used to predict EPA presnap or during a play, like formation or dropback type can be used to.

Our dataset is so large that outliers do not have a large effect on our model, and our model isn't likely to improve with more data. To improve our model, we could look for even more specific

data, such as data regarding pre-snap motion or defensive coverage type. We could also try a different, more complex regression model.

## **7. Model deployment**

- a. Briefly describe how you could envision this model being used or deployed.**
- b. Briefly describe any issues, risks, and or challenges you see with deploying your model.**

This model could be used to help teams game plan and improve their passing schemes by examining what specific factors have the most weight on an individual play's success. Some of the risks regarding the deployment of the model relate to the context needed to understand the impact of the model's results. For example, offenses pass out of formations like the shotgun far more often than out of jumbo, and when teams do pass out of run-heavy formations like jumbo, they can catch the defense off-guard with play-action because the defense is expecting a run. Even though our model suggests pass plays out of jumbo lead to more success than out of the shotgun, an offense can't just implement a lot more of these passing concepts out of run-heavy formations because defenses will be less fooled by play-action if an offense continues to pass out of those formations. This model also classifies variables generally and doesn't factor in player alignment or other important factors resulting in play success such as pre-snap motion. There are other formations and dropback types that aren't included in our model, as we have the most common ones listed. If an offense gets extremely creative, it might not be properly accounted for in our model.

## **8. Code**

- a. As part of your write-up, please submit your Python notebook.**
- b. This should include at least 2 custom functions written by your team**
- c. Code will not be graded on efficiency, complexity or style, just that it was used to generate the figures and/or tables in the write-up.**