Chris Murdter

Springboard Data Science Bootcamp

1/28/20

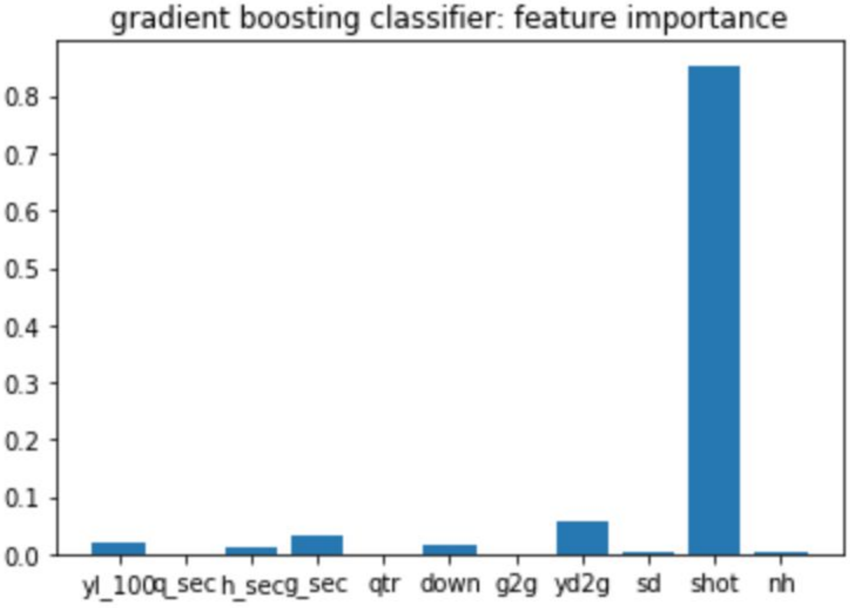
Capstone Project 2 Final Report

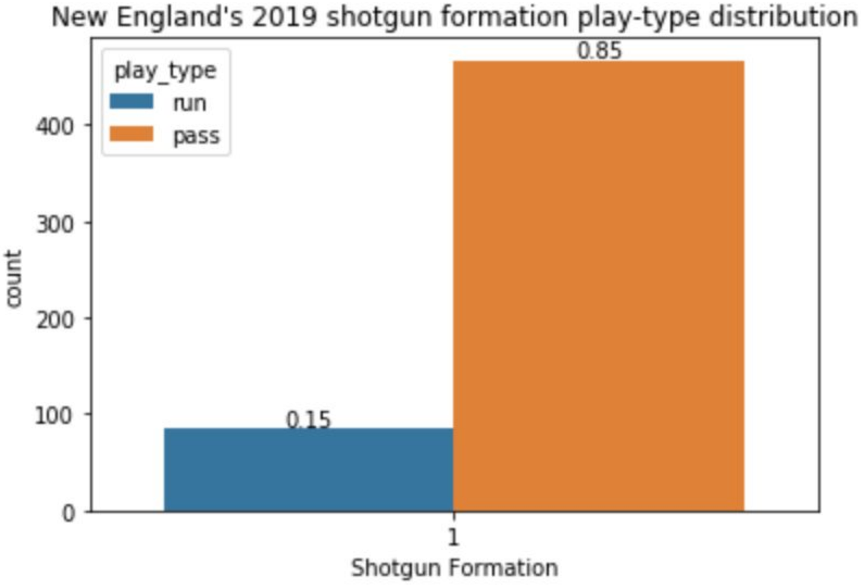
This second capstone project is taking a dive into NFL data. Using yearly play-by-play data from 2009-2019, my goal is to make a machine learning prediction model to predict offensive play calling as well as look at offensive patterns to find trends in these plays. Data analysis is becoming more and more popular with NFL teams and fans. Data analytics was first used in the NFL to help teams draft players and is now becoming more popular being used for on the field analytics. The NFL has recently hired Amazon Web Services to gather data and implement machine learning and AI to gather insights into the players ability or game predictions. The data for my project was taken from Github user ryurko nflscrapeR-data repo and will be linked in the description.

My first task was to take the yearly data and concatenate it into one big data set. This would make cleaning it much easier. Using panda’s concat method I then had a data set of 256 columns and almost half a million rows. Printing out all the columns was next as I needed an idea of what columns would be good features to use for my model. Next I made a dataset that included the columns that I needed for the project. These columns included the team, time of the game, field position, description, play type and a few others.

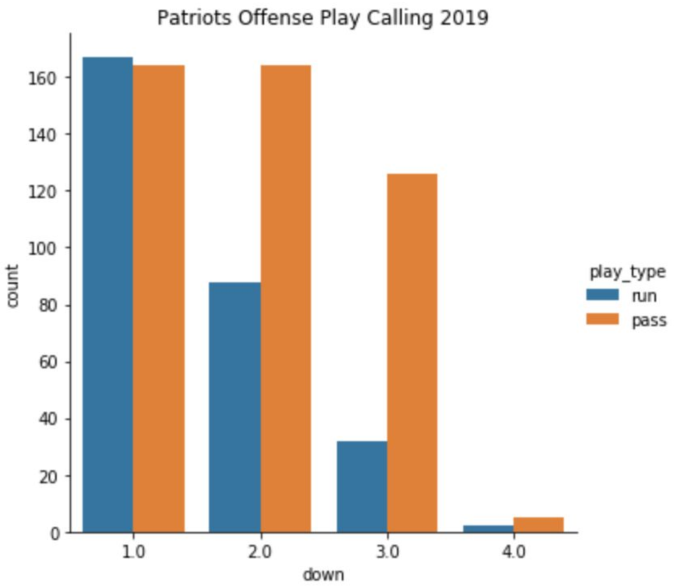
Next I made visualizations to see which play has been used the most from 2009-2019. To no surprise, the pass play is the most used. Then I plotted on which play was most used on a certain down and found that teams run the ball on first down and tend to lean more on throwing on second down and especially third down. On a regression plot I found that as the yards to the first down increase the more teams rely on passing plays. This makes sense as passing plays can result in bigger yards gain more successfully than running. This was useful information as I wanted to use these same methods to focus on a couple teams later.

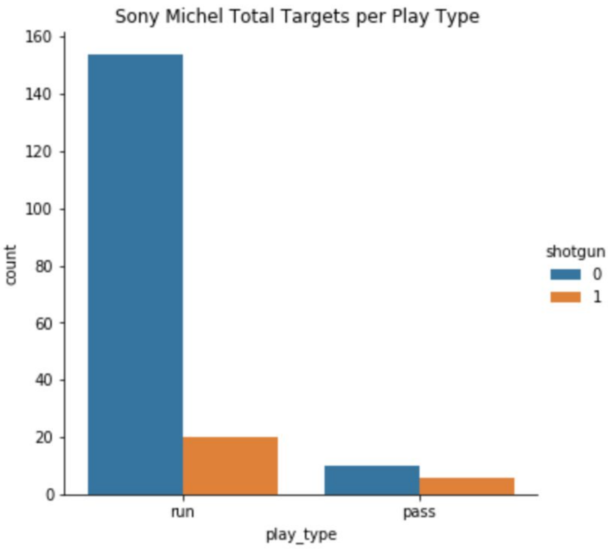
The first team I looked at was the New England Patriots. They are a pass heavy team and have had the same coach and Quarterback for almost two decades. I made a training data set of 2009-2018 plays and a testing data set consisting of 2019 plays. Creating label and feature data sets, a created a Gradient Boosting Classifier using SKlearn. This model would predict if an offensive play was a run or pass. Very easy to implement. It resulted in an accuracy of 76.12% which is an improvement over the base prediction which is around 50%. The base prediction is found from looking at the data and how many run/pass plays there are over the total amount. My next task was to figure out which feature from my feature list was the classifier using the most. Using feature_importances and plotting features vs feature_importances It was shown that the shotgun column is by far the most important data column. The next important feature was yards to go until a first down. But this feature was not even close to the importance of shotgun. I found that the shotgun formation is the most important feature in my model because when the New England Patriots are in the shotgun formation they pass 85% of the time.

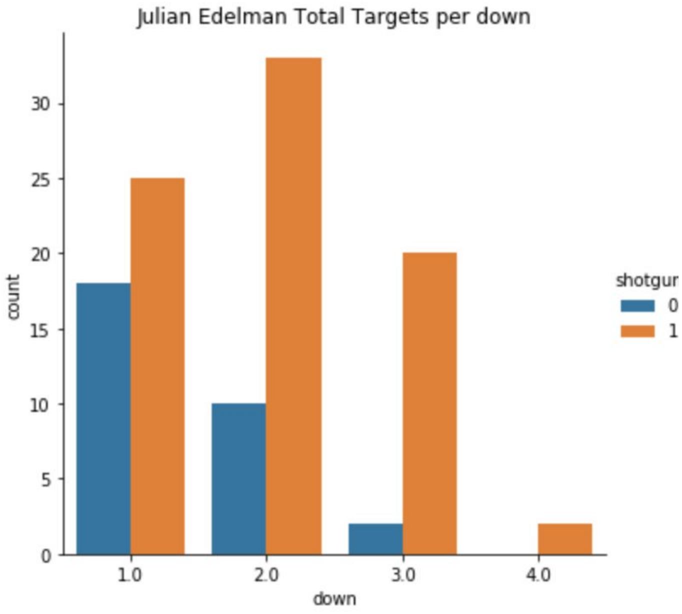


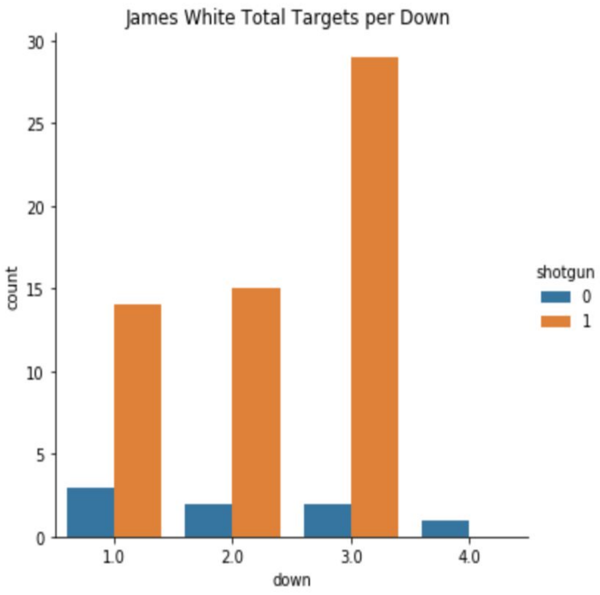


Next step was to look at Patriots offense personnel. Who are the patriots running the ball with? Or throwing the ball to? Finding trends here could improve our model or find more answers to figuring out the offense. I found that when in the shotgun formation the Patriots target Julian Edelman (WR) a lot on first and second down and then target James White (RB) a lot on third down. This is because Edelman will most likely be double covered on third down as he is Tom Brady's (QB) go to reciever. I also found that when Sony Michel (RB) is in the Patriots rarely run shotgun formation and its most likely going to be a run play.

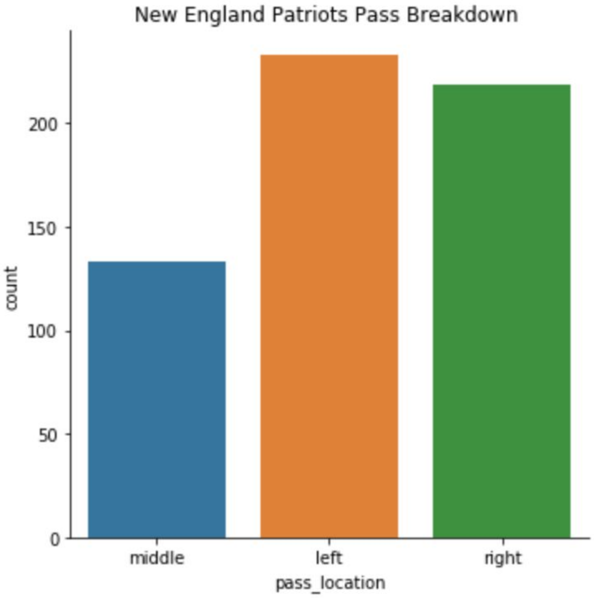

Patriots Offense Play Calling 2019

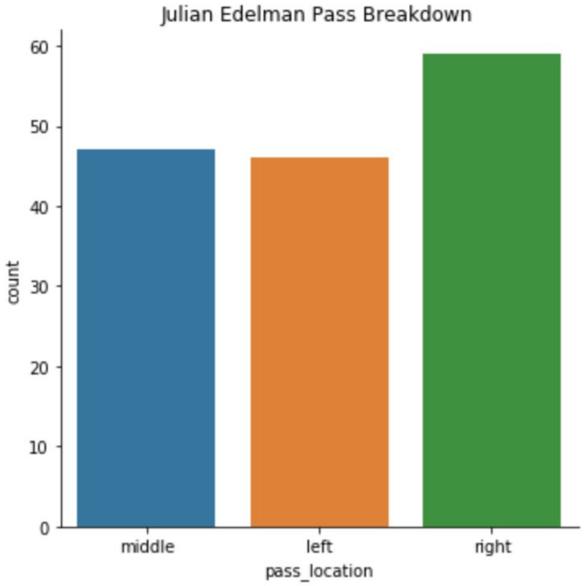

Sony Michel Total Targets per Play Type

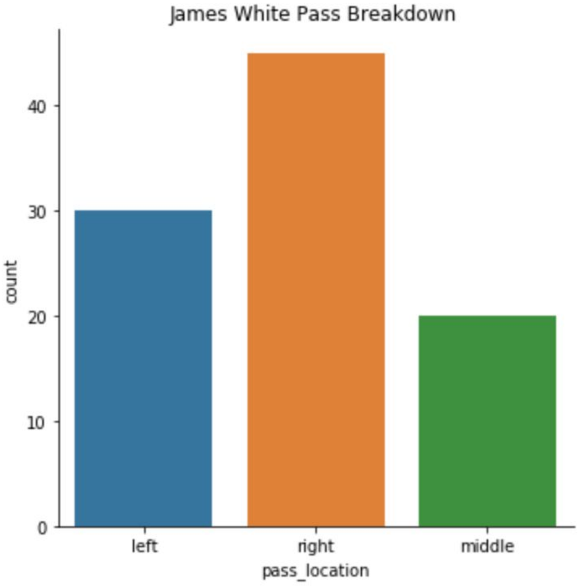

Julian Edelman Total Targets per down


James White Total Targets per Down

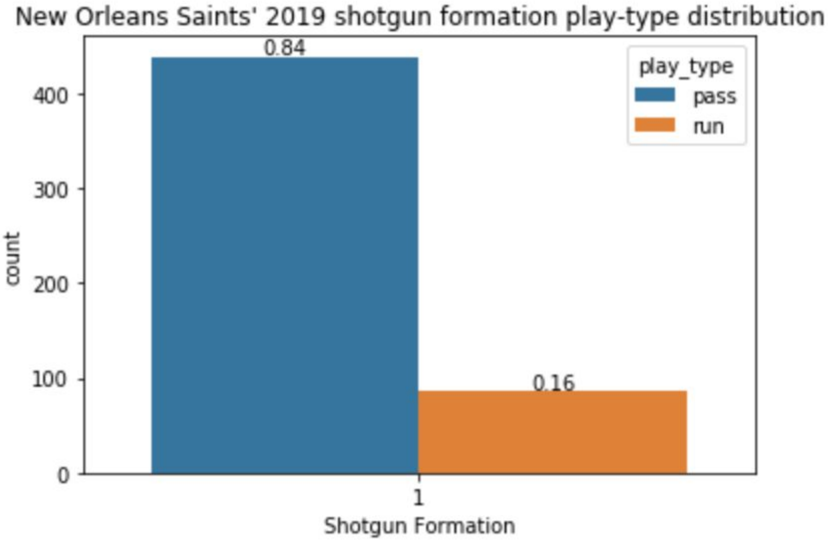
Next I wanted to know where the Patriots were throwing the ball on the field and who they were targeting. Pandas makes selecting this data very easy. I found that the Patriots throw more to the left and right sides of the field than in the middle. Julian Edelman is targeted the most on the right side, most likely because he is lined up on the right side of the field. But his targets in the left and middle are very similar. Julian Edelman likes to target weak spots in the defense in the middle of the field. James White gets most of his targets on the right side. This is because most of James Whites' targets are screen plays or very short passes behind the line of scrimmage.





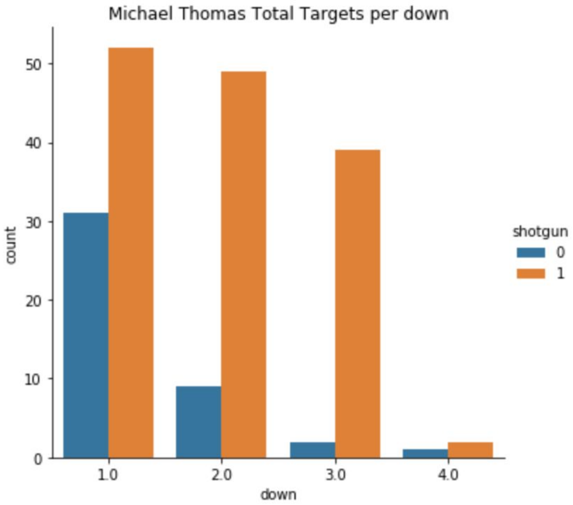


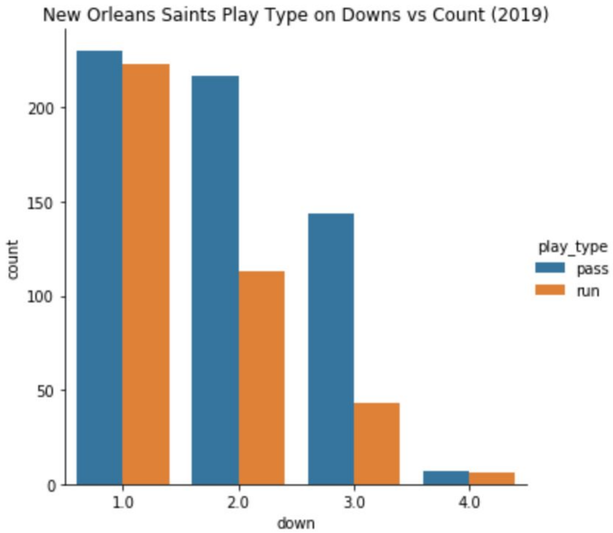
I did this same research on the New Orleans Saints and found that Drew Brees (QB) had similar targets with Michael Thomas and Alvin Kamara (RB) but ran the ball when Latavius Murray (RB) was in. I noticed that Michael Thomas is having such a great year and is a very talented wide receiver that Drew Brees likes to target him pretty much on every down. This is slightly different from the Patriots where Julian Edelman is usually double teamed on third down and so Tom Brady has to look elsewhere. The Patriots offense was not as effective this year as in previous years in that no wide receiver besides Julian Edelman really broke out. My model for the New Orleans Saints was 75.89% accurate as the shotgun formation was once again the most important feature due to Saints passing 84% of the time in the shotgun formation.

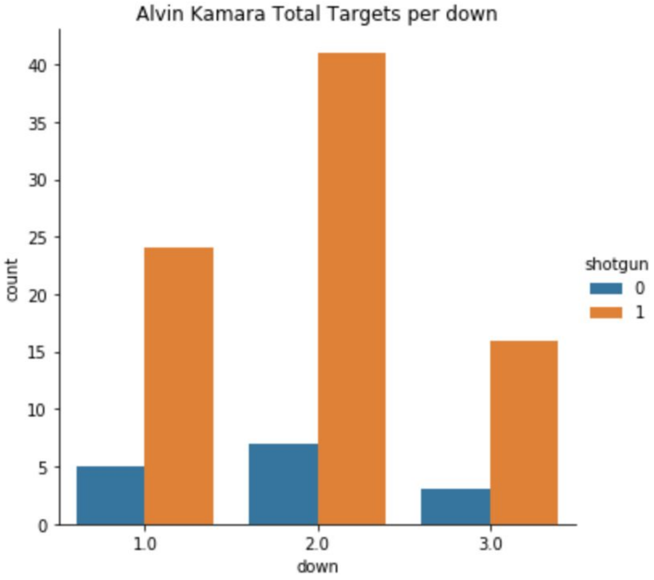


Now let’s look at the visualizations for how the Saints offense structure and how the key players are targeted. The Saints throw the ball slightly more on first down than run the ball. This is a difference from the Patriots offense were they tend to run the ball more on first down. Then on second and third down the Saints pass more just like the Patriots. As mentioned above, Michael

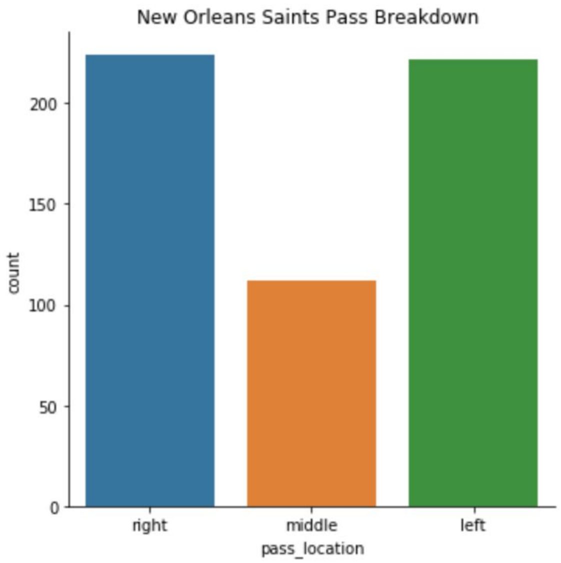
Thomas gets targeted as much as possible and Alvin Kamara a big 2nd down target. Its interesting because teams try and stop Michael Thomas and it looks like it doesn't work as well as they want.

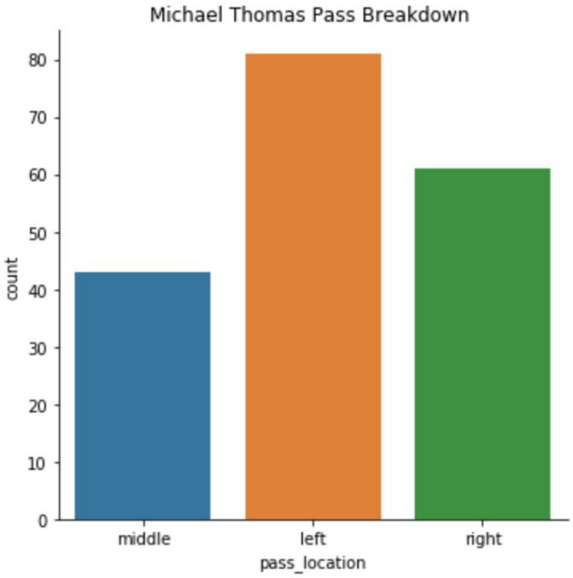


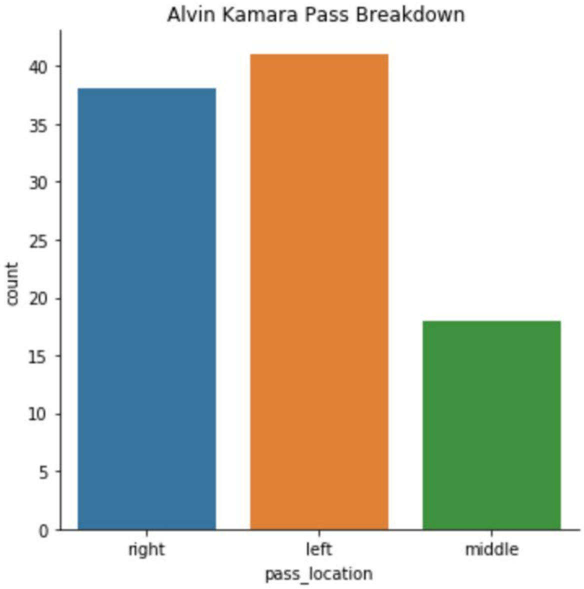




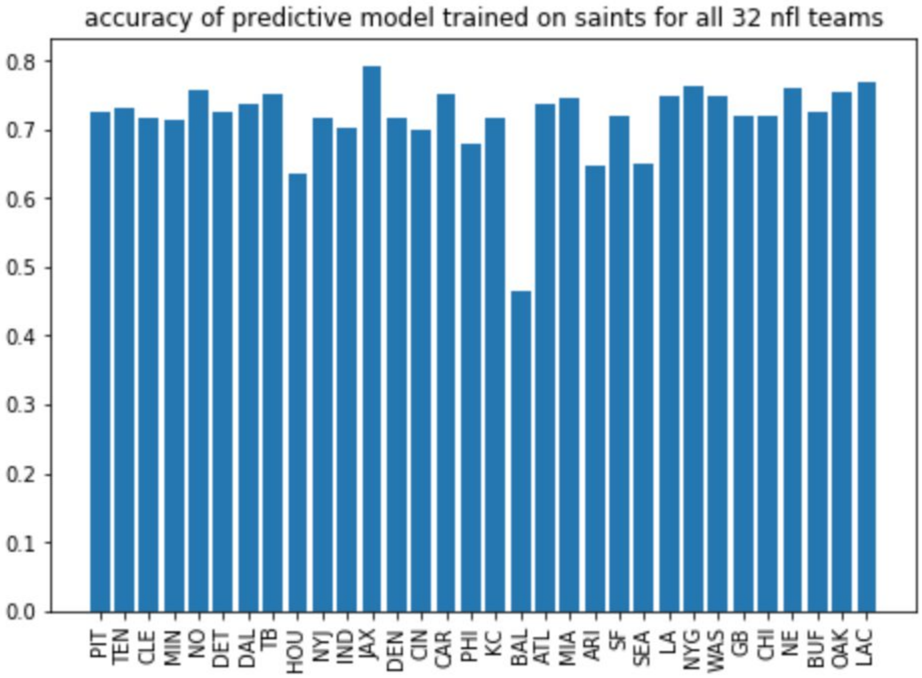
Breaking down the Saints offense even more we see that even more similarities come up. Saints pass the most to the left and right. Michael Thomas gets targeted the most on the left side of the field where he lines up and Alvin Kamara gets targeted the most on the left and right due to screen plays or other short passes.



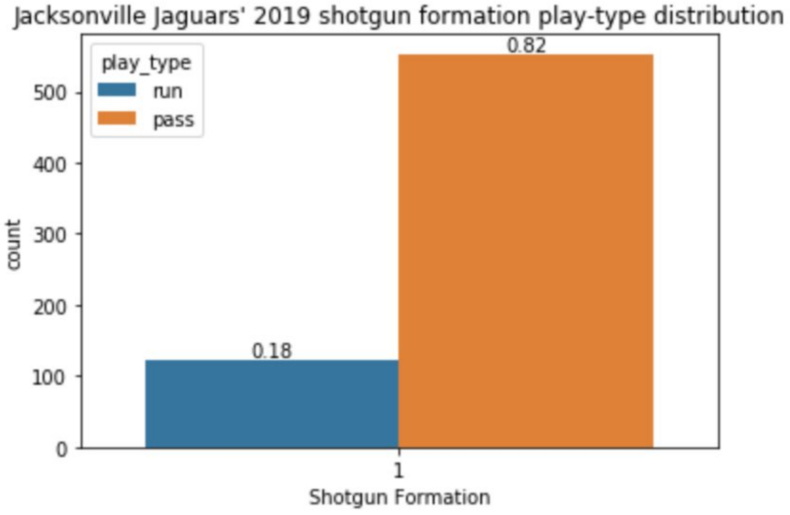




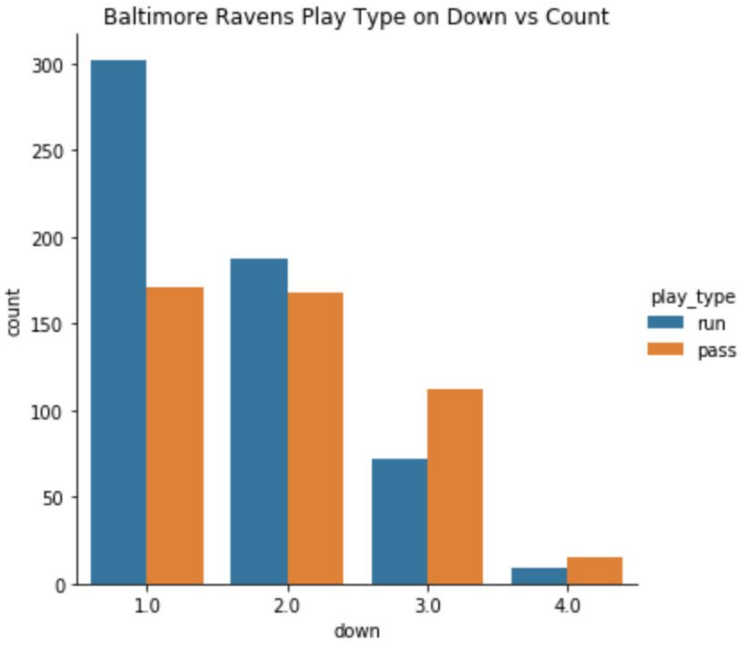
Because the Saints and Patriots were so similar in shotgun passing I wanted to use New England data (2009-2018) to predict Saints (2019) play calling to see if there was any improvement. I managed an accuracy of 75.79% which was a very small difference. The next step was to use Patriots data to predict all teams. I did this by iterating through each team's data and extracting the features. Plotting the accuracy of the model for all teams on a plot lead to some very interesting results.



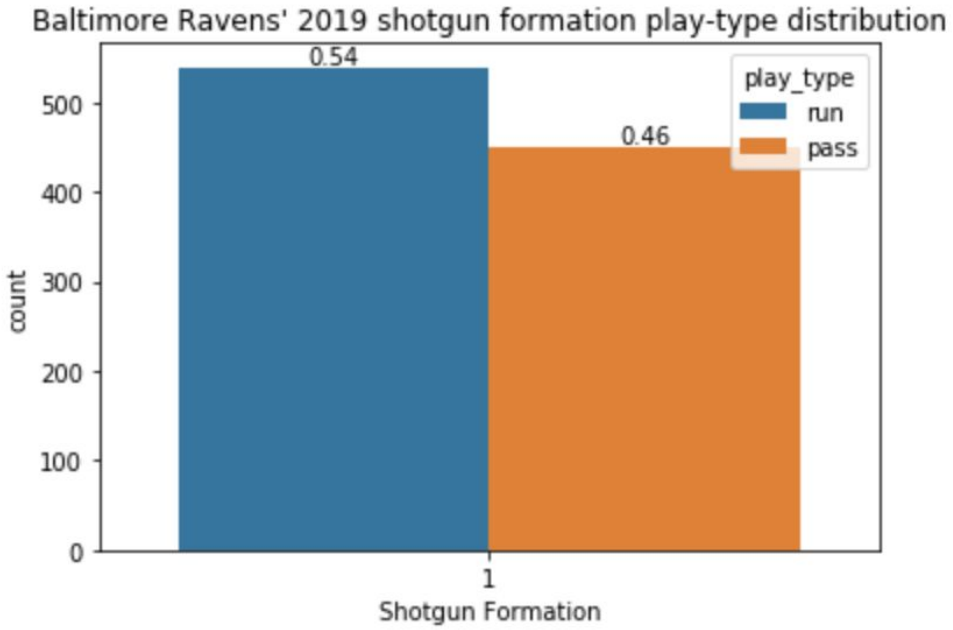
As seen above, most teams are around the 65-75 percent in accuracy score. The highest accuracy score from a team was achieved from the Jacksonville Jaguars. I was surprised by this result as well but mainly because I do not pay much attention to this team. They weren't very competitive or entertaining in my opinion. But nonetheless I took a deep dive into their offense. The Jags achieved a score of 79.3%. I plotted shotgun formation play type for the Jags and found that the reason the accuracy score is so high is due to the high percentage of pass plays in shotgun formation. It is much higher than any other team in the NFL.



But one team really sticks out. The team with the lowest score was the Baltimore Ravens. I looked deeper into the offense of the Ravens just as I did with the New England Patriots and New Orleans Saints. Most teams on first down, have a very close distribution of pass vs run then as second and third down come around, they rely more on the pass play. But not the Ravens. The Ravens offense was unlike any other team in the NFL this year. As seen in the graph below, the Ravens were very run heavy first down, and then became even on second and finally relied on the pass in third down situations. Compare this with Patriot/Saints Play Type on Down and you will see the difference.



But it doesn't just stop there. We know the model heavily depends on the shotgun formation. Well unlike every other team, when the Ravens are in shotgun they are more likely to run the ball then pass it!



This is the main reason why my model failed so miserably while trying to predict the Ravens offense. Ravens use a type of shotgun called the "pistol formation". This formation is called that because the quarterback is closer to the line of scrimmage than the shotgun but still a couple yards behind the center and the running back is behind or next to the QB. So the pistol formation is a type of shotgun formation but it has the advantage that the QB can make downfield reads and the running back is further back and build momentum. Another advantage is that it keeps the defense guessing if the play is going to be a run or a pass play. That's exactly what the Ravens did this year. They kept the defense guessing and when they ran often they could switch to a passing play and catch the defense off guard. They were very effective this year as an offense and their QB Lamar Jackson would go on to win the MVP.

In conclusion, I am very happy with the results of this project. I proved that machine learning can be used to predict offensive play calling in the NFL with success. I learned that the shotgun formation is a very important feature and one that most teams use when passing the ball. However, I also used that this very formation can be modified to confuse defenses and become very effective in different ways. Most coaches in the NFL, if not all of them, already know these kinds of trends from watching hours and hours of film. Some teams have data analyst departments where this kind of project would fit. I think this kind of work can save coaches hours and hours of time preparing for a team by finding trends in data and then reinforcing them with film study. Coaches only have less than a week to prepare for the next opponent so this kind of analysis can be crucial. To expand upon this project, an interesting idea would be to combine this kind of analysis with the AWS player tracker system which uses RFID chips in the players shoulder pads to track them. That way when the offensive players step onto the field, coaches can get a real time estimation of the types of plays that are most likely in certain scenarios. Or coaches can use this data for the next time they prepare for the same team to make another model that predicts WR routes or RB plays in certain scenarios using the player tracking data. This project would be a major expansion on the work done in this project but something I could see be the next step in NFL analytics.