Capstone Report Submission Final Submission | Gino Sacco | March 29th, 2020

In the NFL, defensive coordinators have the responsibility of defending their team's endzone and are always trying to predict the opposition's next moves. Meanwhile, offensive coordinators look to maximize their efficiency on any given play by earning the most amount of yards per play. Being unpredictable on the offensive side of the ball can lead to consistent yards gained which ultimately leads to moving the ball down the field into the end-zone for a touchdown.

My objective with this assignment is to see if I can predict on any given offensive play in the NFL whether a team will pass or run. This information can lead to providing the defensive team with a strategic advantage to better prepare themselves for the upcoming play. Football is unpredictable by nature. It's not possible to always foresee your opponent's plays, but you can try to maximize the accuracy of your predictions. In this project, I will train machine learning models to predict whether upcoming plays will be passes or runs.

In order to do this. I needed to source data of past NFL plays. Luckily, user Max Horowitz uploaded a dataset of every play occuring in the NFL from 2009-2018. This data was collected using an R package called nflcrapR which uses an API maintained by the NFL to scrape and output data with each row corresponding to a play, and 255 features of the plays that have occured. I also found additional data of the 2019 season on github which was also collected and compiled by the nflscrapR package.

Since the dataset was fairly clean, and both datasets I acquired from both sources had the data formatted identically, I jumped into EDA. Initially I was only working with data across the 2009-2018 seasons, however I managed to include the 2019 season later as a test data for the models I trained.

At first,I was curious to see what teams have had the most success over the given sample of data. I measured success as the amount of wins. It was no surprise that the New England Patriots had by far the most wins with 121 through 2009-2018. It's important to note that each team plays 16 regular season games per season. New England maintained a 76% win rate over 10 seasons which is quite remarkable. The New England Patriots have made the playoffs in every season from 2009-2018, winning the AFC East division all 10 years, and ultimately won the superbowl 3 times in the given time span (2014, 2016, 2018), and runners up on 2 other occasions (2011, 2017).

Next I wanted to see what the distribution of play calling in the NFL was over from 2009-2018. It turns out passing tends to be a more popular play call across the NFL at 58.4% vs 41.6% runs. This makes sense as passing on average tends to get more yards then running (6.3 yards vs 4.3 yards) but also has a much higher standard deviation (10.2 vs 6.5). Passing in general is more of a high risk high reward type play and is good when a lot of yards are needed on a given play, while runs tend to be more conservative and are effective in short yardage situations.

Lastly I needed to reduce the dimensions of my data as it consisted of 255 features, but many of the features in the data were descriptive aspects of the outcome of the play, which could not be used in a predictive model. After checking the features that existed in the data set, and engineering a couple of my own, I ended up reducing my dataframe to 16 features including my target.

Once the data was all numerical, I split the data set into train and test sets. I used the 2009-2017 data to train my models, and used the 2018 and 2019 seasons as my testing data. The idea behind this was to see if I could use historical data to predict future plays. After testing 10 different algorithms such as logistic regression, decision trees, KNN, etc. I was getting fairly consistent accuracy results across both seasons of test data between 68% - 73%. My best initial results came from an XGBoost classifier. Since my accuracy results were pretty consistent across all the algorithms I used, I decided to stick with XGBoost and hyper parameter tune it to see how much accuracy I could get from this algorithm.

It's important to note that while I was optimizing my model, I kept my 2018 and 2019 test data separate. My logic behind doing this was to try to find a good model that would get consistent results across multiple seasons. This would then theoretically lead to getting similar results in the upcoming 2020 season that is set to start in September.

After hyper-parameter tuning the XGBoost classifier, I got the following results;

2018 Season Test Set :

Accuracy - 72.75%
F1 Score - 78.01%
Recall - 81.46%
Precision - 74.86%

2019 Season Test Set :

Accuracy - 73.16%
F1 Score - 78.51%
Recall - 82.59%
Precision - 74.81%

With having very consistent results across this board, my next step was to interpret the results. Since the NFL consists of 32 teams, I wanted to see if there were teams that were more predictable than others. It turns out this happened to be the case.

In 2018, teams specific predictions ranged from lowest being Seattle Seahawks at 61.97% and the highest being the New England Patriots at 78.49%. I found this to be a very interesting result as the New England Patriots ended the season with the second best record in the league with 11 wins 5 losses, and ended up winning the superbowl. They were the best team in the league

this season yet were the most “predictable” team in the NFL according to my model. This initial result went against my assumption that predictable teams would be less successful on average.

Moving into the 2019 season predictions, the accuracy range was even larger as the lowest was the Balttimore Ravens at 58.34% and the highest being the Tampa Bay Buccaneers at 78.81%. The Ravens were the most dominant team during the 2019 season winning 14 games and only losing 2, while ultimately getting upset in the first round of the playoffs. It was interesting to see that in 2018 the best team is the most predictable, while the following season the best team was the least predictable according to the model..

Next step was to figure out how the model was making its predictions so I checked out feature importance. By a large margin the most important feature was “shotgun” which is a type of formation an offense lines up in prior to the play occuring. A more detailed description of this formation can be found at https://en.wikipedia.org/wiki/Shotgun_formation.

This is fairly intuitive as shotgun formation is mainly used when a team is planning on passing the ball, however in the case of the Baltimore Ravens during the 2019 season, they lined up in shotgun formation on 95% of all their offensive play calls, yet only passed the ball 54% of the time. It’s important to put these numbers into context. League wide through the 2019 season across all 32 teams, shotgun formation occurred in ~65% of all offensive plays and the distribution of passes was 74%. The Ravens started nearly all of their plays in this formation and passed on just over half of these plays. Also another interesting finding was that in both 2018, and 2019 respectively the Seattle Seahawks and the Baltimore Ravens were the only teams to run the ball more than 50% on all their play calls throughout the entire league. These factors ultimately led to the Seahawks and Ravens being the least predictable team in the NFL during the 2018 and 2019 season respectively.

This leads me to future work. I would like to continue working on a more robust model that can pick up differences between teams such as offensive tendencies, or even build new features including quarterback tendencies. This would be challenging for teams with either rookie quarterbacks with no previous data, or teams with undefined starting quarterbacks, however a player like Tom Brady who has spent the last 20 seasons with the New England Patriots and just signed a new contract this spring to join the Tampa Bay Buccaneers, would likely change the offensive dynamic of both his new and former teams.

And lastly, my final goal for this project is to have real time predictions as if we were in a real game situation. These predictions could be relayed to the defensive coordinator to provide insights into defensive play calling. Since I am not employed by an NFL team, the next best thing would be to create a twitter bot that sends out predictions during an NFL game so fans like myself and hopefully many others would follow along while watching a game.