Predicting NFL Play Calling Using Historic Play-By-Play Data

Hans Missenheim

Table of Contents

Abstract

Professional football is a sport in which any single play can make a significant impact on the game's outcome. An accurate prediction of what their opponents next play would be is crucial for National Football League (NFL) teams in moving the play's impact to their advantage. If teams were able to make more accurate predictions, they would not only increase their chances of winning the game but increase the chances for a close and more exciting game for fans. In this project, I aim to use machine learning models to create an accurate predictor for NFL teams that can predict what the next play is prior to the snap.

The dataset used contains play-by-play information for every play in the past 24 NFL seasons. Each sample of the data represents an individual play, each with variables of the current game state and team choices. The game state includes the quarter, time remaining, down, distance, and other factors of the position within the game. Team choices within the data include whether the formation is in shotgun, and no huddle was used before the play. When only using the pass and run plays, the dataset contains over 800,000 unique plays to be used for training, validation, and testing.

Various models are constructed and compared using machine learning algorithms such as naïve bayes, random forests, and neural networks. The results found that encoding of data through normalization and standardization had a positive effect on naïve bayes model accuracy, while balancing had no effect. Conversely, the encoding of data had no effect for random forest, and balancing of data is negative to the model's performance. While the naïve bayes model was found to be acceptable for the implementation requirements, it was found that accuracy was improved with model complexity for both the random forest and neural network, reaching levels of 71.6% and 72.2% respectively. The findings of this project and the use of the model can

provide insights for coaches and teams to make more informed decisions and improve their game strategies.

Project Plan

April 2, 2023

Profile of the organization and background of the opportunity

Primary Company Details:

Founded – September 17, 1920
Headquarters – New York, NY
IPO – N/A (Private)
Categories – Sports Teams & Leagues, Media Production

Address:

345 Park Avenue
New York, NY 10154
United States

Company Communication:

Phone Number: (212) 450-2000
Website: www.nfl.com

Business Description:

The National Football League is a professional American football league that is considered the highest level of professional football in the world. It is comprised of 32 football teams based in the United States which are divided into 2 conferences, the American Football Conference and the National Football Conference.

The NFL operates as an overarching body to the teams, in charge of establishing rules for the sport, overseeing media distribution, and the overall management of the operations between teams. The individual teams themselves are owned and operated separately by private owners or investment groups. The league as a whole generates every year billions in revenue through national merchandise, sponsorship, and television deals, with the profits being shared across the members of the league.

This financial success is driven from the NFL's immense popularity in American culture and its scope outside of the United States. Multiple networks reserve multiple time slots every week to air NFL games live during the season, with many being nationally televised during "primetime" hours. The annual championship, the "Super Bowl", played between the AFC and NFC champions is consistently the most watched television broadcast of the year. Of the 30 most watched broadcasts in US history, only eight are events other than the Super Bowl, with nine of the top ten being the NFL's championship game.

Financials:

Latest Financial Data – 2022
Revenue – $18 Billion
No. of Employees – 3,595

Key Executives:

Roger Goodell, 64, Commissioner
Brian Rolapp, Chief Media and Business Officer
Troy Vincent Sr., 52, Executive Vice President, Football Operations
Dawn Aponte, 52, Chief Football Administrative Officer

Major Competitors:

National Basketball Association
Major League Baseball
National Hockey League
Major League Soccer
Canadian Football League

Business/Analysis Opportunity:

Using a dataset collected from publicly available NFL game data consisting of 1,148,717 plays from 6418 games within 24 seasons, I will analyze the provided research question and create predictions on it using statistical models of classification. These models will not directly be based on making business outcome predictions, but focused on team-level in-game decisions that will affect the outcome of the game. If teams use data to create better decisions, then games may become more exciting and gather higher viewership and popularity for the league. Thus, the use of this analysis will indirectly lead to higher financial success of the league as an entertainment product.

Research Questions

A major key to in-game decision making in the sport of American football is based on what you believe your opponent will choose to do. This is especially important on defense, where you must play differently against either of the offense's options, running or passing the football. Thus, it is crucial for teams to make accurate predictions on what type of play their opponent will call. The following question will be the focus of this project's research.

RQ: What is the next type of play (run/pass/etc.) given the current game state and previous plays?

It can be crucial for defenses in football to predict what type of play is coming next. However, over committing to a run play when it is a pass and vice versa can lead to a fatal error. If teams can predict what type of play their opponent is likely to run, the likelihood of a successful defensive play should increase dramatically. Having a high probability prediction of an offense's tendencies allows for the defensive team's play-caller to choose plays that lean in success against one type of play over another. Based on the findings, this research question can be expanded to investigating from the offense's point-of-view with which type of play they should run.

Hypothesis

H: The overall tendencies of all teams, along with processing team-specific history, will allow teams to accurately predict an opponent's next play.

Considering the entirety of the dataset's variables on every play from every game will create an accurate predictive model of a team's next play. This accuracy will be improved when the model uses historic data pertaining to each play. This includes a team's run/pass tendencies in previous seasons, the team's run/pass tendencies in previous games, the team's run/pass tendencies in the current game, and the previous play calls during the current possession. Each will likely be weighted differently in importance, but each will improve the existing generic play calling model.

Data

The data for this project has been collected by using **nflfastR**, a package in the **nflverse** collection of R packages centered around the collection and analysis of NFL (National Football League) data. The specific dataset that will be used contains detailed play-by-play data for every

season between 1999 and 2022, and can be downloaded at the nflverse-data GitHub repository (https://github.com/nflverse/nflverse-data/releases/tag/pbp). Combining the entries of every year's play-by-play data cumulates to 1,148,717 individual plays over 24 seasons. Each entry contains specific data of the play, such as the current time, field position, play type (run/pass), and the outcome of the play, as well as overall game data of the teams playing, the predicted scores, and the outcome. The play-by-play dataset contains 372 different variables; however, a large portion are only used for distinct plays, leaving most entries to contain between 100 and 250 variables.

Game State

Many variables for each play consist of the current state of the game during the play. This includes the current quarter, time remaining, current down, yards until first down, current yard line, and current score differential. Differences in each piece of the current game state will affect how aggressive or conservative an offense is willing to be in their play calling.

Offensive Decisions

Other factors captured by the play-by-play dataset are the decisions made pre-play by the offense. These include whether the current formation is in shotgun, where the quarterback stands several feet back from the offensive line, and whether the offense is running hurry-up, where the team decides not to huddle before the play. Both of which are highly predictive of the future play call, but the defense has much more limited time to gauge these variables compared to game state.

Historic Tendencies

By using the previous plays in the entire dataset, information can be added to each individual play, such as the previous play calls in the current possession. Along with that, the coach of the team is also within each database, which can be found to have their own tendencies throughout the data.

Measurements

For this research, what is being measured is relatively straight-forward. This project will be focusing on the future play calls of teams. This play call can be classified as a kickoff, pass, run, no play, punt, extra point, field goal, QB kneel, or QB spike. Some of these plays, such as kickoffs, no plays, extra points, and QB kneels only occur due to specific game instances that are forced, so no prediction is needed. Other plays, like the QB spike, happen in limited instances and again do not need a prediction since there is no preventative measure needed against such a play. The play types that will be focused on in this research are the pass and run plays that consist of 73.5% of all plays within this dataset.

Methodology

For this project's research question, we are trying to predict whether the next play will either be a run or a pass. Since this is a prediction of one of two categories, classification models should be used in order to create a predictive model. Predictive models will be made using a few popular classification methods and compared. Such models include naïve bayes, random forests, and feedforward neural networks. Once a predictive baseline is reached for a play and its own variables, variables from previous plays related will be introduced. This includes retraining the previous models after appending historic play calling data to each play, and dividing the plays

using distinct groupings (possession, game, season, team) and training a recurrent neural network using it.

Computational Methods and Outputs

For the initial classification, I believe that each model will be similarly accurate. This is likely due to the number of entries to work with in the dataset, and the likely high collinearity of the variables that impact the classification of the play. However, the similar accuracy won't be an incredibly high accuracy because of the missing historical context of each play.

For most of the classification models used, the historical context for a play will have to be appended to each data point. One variable, the team's coach, is already added to each play. But each play in the current series, the play-calling tendencies of the current game, and the team's tendencies before this game must be added as additional variables for each play.

The output for the research question is the binary prediction of whether the next play is a run or pass. Based on the accuracy of other, less-frequent play calls, this binary classification may be extended to a prediction of each of the play types found in the dataset. It may also be beneficial for both model analysis and team utilization to output the confidence percentage for each prediction, rather than just the classification result.

Output Summaries

The classification models will be able to accurately predict what the next play to be called by the offense, whether it be a run or a pass. Additionally, this predictive model will be run on the latest NFL season's data to determine the predictability of each team and compare them.

Campaign Implementation

As previously discussed, it is crucial for each team to predict the likely moves of their opponent during the game. This most importantly includes the offensive team's play call and whether it is a run play or pass play. Information allowing teams to predict what the offense will attempt to execute will lead to better countermeasures in terms of defensive play calls. Additionally, teams can use a similar analysis on their own play calling to diagnose their own tendencies and become less predictable. The predictive models of this research could be used for both implementations, and lead to increased team success.

Literature Review

The adoption of data analytics has been slower in the sports world compared to other industries. However, the effectiveness seen in the "sabermetrics" of baseball in the 21st century has led to further investigation in other sports. The sport of focus, American football, has recently seen an increase in statistical analyses due to more refined data collection. Since the publication of aggregated play-by-play data for analysis purposes, offensive play calling in the National Football League has become a main research topic.

One key vector of play calling analysis is the amount of value gained or lost by teams based on whether a run or pass play was chosen. Research investigating NFL play calling risk by Benjamin C. Alamar in the *Journal of Quantitative Analysis in Sports* found that teams in fact were sub-optimal in decided play calls. It is the case that "if team's passed more, they would increase their probability of winning", yet it is often seen that team's run the ball when they should instead pass. It is often presumed by coaches that this considered over-reliance on running the ball is taken as it is seen as minimizing the risk of negative plays. However, the

previous research by Alamar found that running the football results in a “lower expected outcome with increased risk”.

Another question researched in the realm of American football play calling is what play the offense *will* run, rather than what play *should* they run. One might assume that the answer to the latter question would lead to the answer to the former question, but as shown by previous research, NFL teams frequently do not choose between running and passing optimally. Thus, in order to predict offensive play calls, the optimized decision in terms of expected points added or win percentage is unlikely to be consistent in effectiveness.

Dr. Marius Ötting, PhD researcher at Bielefeld University, used hidden Markov models (HMMs) to predict NFL play calls with a higher success rate to the optimal-play-call method. The hidden Markov model, a statistical model that uses observable outputs to make inferences about underlying states, uses historical information in its process to better predict the future outcome. In this case, the observable outputs entered in the HMM are each play call prior to the one being predicted, and the underlying state given those outputs is the team’s tendency in play calling. By including this past historical context, each play’s prediction by the model is focused on that specific game state provided by the previous plays, rather than the general likelihood of every teams’ tendency in that same game state.

It would be crucial for any predictive model with the goal of correctly predicting American football play calls to consider the previous tendencies of an offense. A popular statistical method that is used in concurrent time series predictions is a Recurrent Neural Network (RNN).

RNNs are a specific type of neural network in machine learning where a node's input data is processed within the node several times before exiting the network layer. Early into the research of RNNs, researchers J. T. Connor, R. D. Martin, and L. E. Atlas found that using recurrent networks in creating predictions for time series data had advantages over the simpler feedforward neural networks. This use case of RNNs can be seen today, as they are the leading model in any data that is represented in order, such as stock prices or language modelling where the order of words is important.

For this project, the usefulness of RNNs can be investigated in prediction of NFL offensive play calling. With the significance of previous plays in the probability of the future play call having been supported, the inclusion of such data is crucial. It is possible to align the data of previous plays as a time series like the use cases of RNNs provided above. With the data prepared in such a way, the use of an RNN can increase the predictive abilities of a model for run/pass play calling tendencies.

Research Question

What is the next type of play (run or pass) given the current game state and previous plays?

It can be crucial for defenses in football to predict what type of play is coming next. However, over committing to a run play when it is a pass and vice versa can lead to a fatal error. If teams can predict what type of play their opponent is likely to run, the likelihood of a successful defensive play should increase dramatically. Having a high probability prediction of an offense's tendencies allows for the defensive team's play-caller to choose plays that lean in success against one type of play over another. Based on the findings, this research question can

be expanded to investigating from the offense's point-of-view with which type of play they should run.

Exploratory Data Analysis

The dataset for this project can be accessed and downloaded at the nflverse-data repository on github.com. The data that comprises the dataset is collected using the nflfastR package, which collects publicly available NFL play-by-play data for the seasons between 1999 and 2022. This includes 1,148,717 individual plays in total over the 24 seasons. However, since the research question of this project is interested in predicting a pass or run play from the defense's perspective, not every play is relevant. For instance, take a scenario where the offense reaches 4th down and sends out their punting unit. In this case, the information is already provided to the defense by the offensive personnel that the play is a punt, rather than a run or a pass, so no prediction is needed. It is possible for the offensive to fake the punt, and instead run or pass, but this is a low sample action making an accurate prediction more difficult. Similarly, other types of plays are scripted by the rules of the game, such as extra points and kickoffs, where again no prediction is necessary. Filtering the dataset for only the pass and run plays we have an interest in creates a dataset with 803,966 total plays. The dataset also features 372 different variables, but most of which are pertaining to the outcome of a given play, and a majority of those are unfilled as they are only filled for specific play types (like **kicker_player_name** that will only have data on kicking plays). Since we are trying to predict the type of play before it happens, any such outcome variables can be filtered out. The pre-play variables that will be beneficial to this analysis are the following:

play_type: Type of play ran by the offense (pass, run)

posteam: Offensive team in possession of the ball

drive: The current play's drive number in relation to the game (1 – 38)

qtr: Number quarter of the game (1, 2, 3, 4)

quarter_seconds_remaining: Number of seconds remaining in the current quarter (0 – 900)

down: Number of down in the current series (1, 2, 3, 4)

ydstogo: Number of yards until the first down marker or endzone (0 – 50)

yardline_100: Distance in yards from the endzone (1 – 99)

shotgun: Whether the offense lined up in shotgun formation (0, 1)

no_huddle: Whether the offense lined up in formation without huddling (0, 1)

score_differential: Difference between the current offensive possessing team's score to their opponents (-59 – 59)

season: Year of the current season of the play (1999 – 2022)

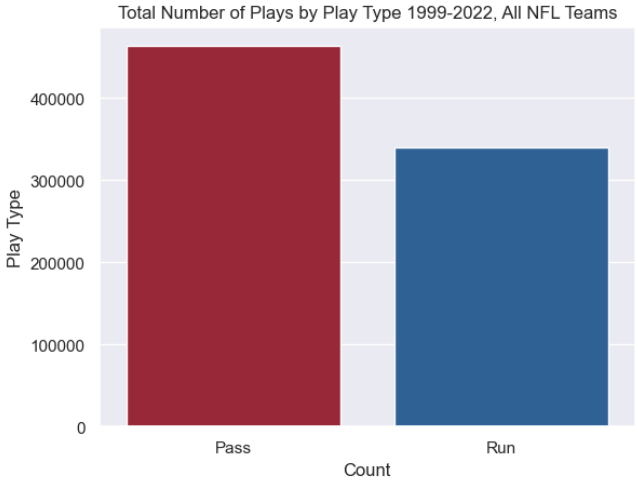
Two additional variables must be created, since the variables in the dataset are coded by home and away, without referencing the team with possession. These are created by comparing **posteam** with **home_team** and **away_team**, and then selected by the home and away variant of each. In addition, **play_in_drive** was created to count the index or position of each play within its individual drive. These created variables are listed below:

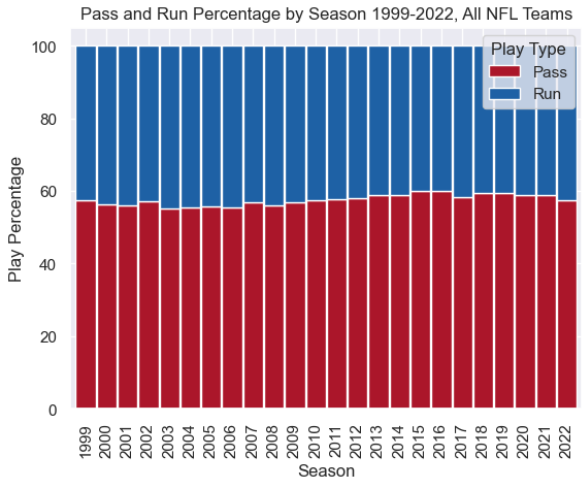
coach: Name of the possessing team's head coach

spread_line: The pre-game predicted spread of points between the teams by sportsbooks

play_in_drive: Index of current play in the current drive (1 – 23)

To gain a general understanding of how often teams choose to pass or run, the total number of each is plotted below, along with an additional graph of pass or run percentage by each season.





By looking at the at the total amount of pass and run plays over the dataset, the observation that teams will run pass plays more often than runs are shown, but there is not a large enough disparity between the proportions to consider this a highly unbalanced selection. This disproportion still must be considered when creating predictions, since a model of less than about 60% accuracy would be less accurate than predicting pass every time. When the pass-to-run percentage is plotted by season, we can see that there was a slight upwards trend in the percentage of pass plays to runs between most of the years of the dataset. However, the last few years have seen a downwards trend to where the last season, 2022, had a nearly identical proportion to the first year of the dataset, 1999 (57.56% to 57.50%). Because of the trend in the middle years, the **season** variable may help in validating the model in those years. But with the

overall trend over the entire dataset seemingly linear right below 60%, it is important not to overfit this trend for future predictions.

The previous plots set a base level understanding of the league-wide tendency of running and passing. It isn't unexpected that the total proportion of passes to run plays to level out at a standard limit for the entire league across thousands of games. We can do the same analysis, but at a per team basis, to see any trends that occur within their play calling. In order to investigate this, a heatmap divided by season for every team was created and is shown in appendix C. The darker red shows a higher propensity for passing for that team during that season, and darker blue shows a higher propensity for running. Note that the Houston Texans (HOU) are blank for the first three years since the team was formed for the 2002 season. The heatmap provides a better look at specific team-wide trends in play calling, as well as larger variety created by greater extremes. For example, the 2004 Pittsburgh Steelers (PIT) were a full percentage point below the second lowest season in passing percentage. This can be explained by having more run calls to protect their rookie quarterback, Ben Roethlisberger, a hall of fame running back, Jerome Bettis, and an overall season record that had them ahead of their opponent often, 15-1. Another trend shown is a higher run play percentage for teams with mobile quarterbacks, which include the 2002-2006 Atlanta Falcons (Michael Vick), 2019-2022 Baltimore Ravens (Lamar Jackson), and 2022 Chicago Bears (Justin Fields). The opposite effect can also be seen with a quarterback acquisition. From 2017-2019, the Tampa Bay Buccaneers were already above the league average pass rate, with an average of 63.2% between those three years. After acquiring six-time super bowl winning quarterback Tom Brady, the Buccaneers pass play percentage increased from 63.4%, to 67.4%, to 68.9%.

While the previous examples showcase how player personnel may affect the play call decision making, another factor of the team-grouped data is the head coach. While not every head coach is the one calling offensive plays, they are likely to hire a coordinator that follows a similar play-calling "philosophy" if they aren't. In this case, we can group the data by head coach, who may be in the dataset for multiple teams. After compiling the data and sorting the coaches by pass and run percentage, the top ten results are displayed in the tables found in appendix D and E. In these tables, we can see further extremes from the league-wide average, suggesting that play calling may have a statistical correlation with specific coaches, which can be used for predictive purposes. It is important to distinguish that some of the most extremes shown in this table are from coaches with low samples in terms of seasons coached, it is also noteworthy that just as many are coaches with many seasons of sample data, with some spanning multiple teams.

Next, we'll look at how various game state variables affect league wide play calling. The first of these is the quarter that the play is called under. Football is played in four periods or quarters, split into two halves. This distinction is shown in the plot in appendix F, where we can see the pay percentages mimicked between the two halves. In each, teams appear to start the half more conservatively by running the ball more or equal to the league average. But as the half continues into the second and fourth quarters, teams increase their passing percentage to above 60%. The conservative opening half play-calling is again apparent when a game goes to overtime, likely with coaches trying to avoid the risk of negative pass plays in a crucial next-score-wins scenario.

Football plays are separated into a series to four downs, in which the offense must reach either the first down marker to gain a new series of downs, or the endzone in which they score a

touchdown. By plotting the play percentage by each down, the graph in appendix G shows another trend that is important for predictions. Teams often start a series running on 1st down, increasing the rate they pass on 2nd down, and are critically high on 3rd down reaching nearly 80%. Most often, the 4th down is reserved for punts and field goals, but when teams do attempt to run or pass, they are more aggressive than their second down, passing on more than 60% of plays.

Another variable that is likely to change play calling is where the ball is located on the field. Ranging from 0 (the offense's own endzone) to 100 (the defense's endzone) should result in different play calling, expecting teams to be more conservative near their own side in order to minimize the risk of turnover. This trend is seen in appendix H, where teams are more likely than the average to run on their side of the field. This lower rate of passing is then increased the further the offense travels down the field. However, there is a distinct drop starting near the defense's 20-yard line (80 in the dataset). This is commonly referred to as the "red zone", and we can see a severe drop in offensive pass plays in this area. In fact, within five yards of the opponent's endzone is the lowest spot in the entire field for passing percentage.

As mentioned with the down variable, teams have a first down marker that they must reach to gain a new set of downs. At the start of a series, this is ten yards away, but can decrease with positive plays, or increase with negative plays or penalties. We can see how the distance needed for the first down marker can affect play calling in the plot I. Most obvious is the sudden valley at 10 yards to go. This is the most common start of a set of downs, and as seen in appendix G, teams most often start series with run plays. There is a similar smaller valley at 15, another common series starting point after a 5-yard penalty. Outside of the common starting spots, we see a trend from 0 to 15 yards where teams are more likely to run the closer, they are to

the marker. This is likely explained by the shorter average gain of running plays that are deemed safer. Since all that is needed is a shorter gain, the safer play is opted for. From 15 yards and further, the inverse is true. As teams move further back from the first down marker, their run percentage increases. The reason for this is probably due to coaches accepting the odds of their team reaching the first down marker is lower, thus they'll take the safer and shorter yards to pin the opponent further back after a punt. It is also the case that teams may think that running is unexpected and may catch the defense off guard.

The final variable plotted against the average play percentage of teams is the current score differential between the team in possession's score to their opponents score. In the plot of score differential in appendix J, the greatest trend so far is seen. In tied games, where the score differential is 0, teams run plays at a rate near the league average. But once a difference in scores is made, an opposite effect is made to both sides. The team that is now losing increases their pass rate, in order to make up the score, and the team now winning increases their run rate, opting for safer plays which also run down the clock. This is an expected outcome for a football viewer or just game theory strategist but is important to verify through this analysis. We can see this effect in a significant negative linear correlation between the score differential and the pass play percentage.

By conducting this exploratory analysis of the many variables in the dataset, we have found likely scenarios and reasons for play calling decisions. Looking at the different variables, we could easily predict a pass play on third down, down 14, and in the fourth quarter. But to make an accurate prediction on any play, we'll need to explore the deeper relationships between the different variables with more advanced forms of analysis.

Methodology

Binary Classification

The result of this type of research question should be a model that could predict whether the next play belongs to a certain type (run or pass). In this case, it would follow that a model would need to be created in the type of classification analysis, which is used to classify data (in this case run or pass). Since the output will be belonging to one of two types, binary classification is the specific task to complete this research question. In order to start the creation of the classification model, certain variables that are likely to influence the classification must be chosen. These variables were looked at closely during the previous exploratory data analysis and relate to the current game state before the play occurs. This includes the game quarter, the location of the ball, the distance to the first down and the score differential. Using these variables, different models will be created using a variety of methods listed below.

- Naïve Bayes
    - A probabilistic algorithm using Bayes' theorem with an assumption of variable independence to classify data into different categories.
- Random Forests
    - A machine learning algorithm in which decision trees are constructed under training and the output classification category is selected by the most common tree selection.

- Feedforward Neural Network
    - A type of artificial neural network where multiple variables are input through one or more hidden layers to predict the probability of a given output.

Analysis

The dataset for this project was created by creating a pandas DataFrame from each season's csv file, and concatenating each into a singular DataFrame. Two variables that were of interest, the team's coach and the game's spread, were coded for either home and away or just the away team. Since the prediction is based on the offense's play calling, these two variables had to be transformed for each play into both the possessing team's coach and pre-game spread. Additionally, two variables were created by counting a previous number of plays in the current game for a given team. These were drive, the collection of plays for a team in each possession, and play_in_drive, the number in the order of the drive for the current play.

To investigate how modifications in the dataset affected the performance of the machine learning algorithms, four models were constructed after dataset additions for the naïve bayes and random forest algorithms. These modifications included normalizing and standardizing the variables, adding the identifying variables (coach and team), and balancing the number of run plays and pass plays in the training data, along with using the unmodified dataset. While creating and training over a hundred different neural network models with different parameters and input data, it became clear the importance of properly encoded and balanced input data. Thus, models were long-term trained only using data that was previously balanced and normalized/standardized.

To validate the predictive capabilities of each model, the dataset was separated into three parts: one dataset for training the model, one dataset for validating the learning ability of the model and comparing against other models, and one final dataset to get the actual accuracy for the last NFL season for each model. If the dataset was only split into two parts, it is likely that the hyperparameter tuning and final choice of model would be biased to just predicting the plays from the latest NFL season (2022). However, the best model should be the one most accurate for

all future seasons, thus we reserve the final testing to gain an accuracy level of the final models, and not to compare the best approach in creating them.

It's also crucial to understand how the order of the dataset matters for predicting NFL plays. For instance, the application of using a model to predict what play was run in a game played in 2002 would be pointless as the game has already been played and thus a prediction is not needed. Instead, the application for this type of predictive model would be to predict what play in the future would occur from the plays known in the past. To properly validate this time-sensitive application, the first 23 seasons will be used for the training and validation datasets, with the final 2022 season to be used as the final test dataset.

Algorithm	Encoded	Coach/Team Data	Balanced	Accuracy	AUC
Naïve Bayes	No	No	No	64.90%	0.694
Naïve Bayes	Yes	No	No	66.77%	0.722
Naïve Bayes	Yes	Yes	No	66.71%	0.724
Naïve Bayes	Yes	Yes	Yes	*66.71%*	*0.725*
Random Forest	No	No	No	70.88%	0.781
Random Forest	Yes	No	No	70.70%	0.780
Random Forest	Yes	Yes	No	***71.22%***	***0.786***
Random Forest	Yes	Yes	Yes	70.98%	0.785

The above table lists the outcome validation accuracy and AUC for the four models created from each algorithm. Each model was created with a different combination of input data configurations to test the effect the data preparation would have on the outcome. The encoded column expresses whether the data was normalized (scaled between 0 and 1) and standardized (fit to a mean of 0 and a standard deviation of 1) or input to the machine learning algorithm raw. We can see from the table above that proper encoding had nearly a 2% increase in naïve bayes models' accuracy and was by far the most important data preparation step. However, the effect

on the random forest models was imperceivable. The coach/team data column expresses whether the possessing team and the coach were included in the input data for each play. In contrast to the last column, this was the most impactful step for random forests, while having little effect on the naïve bayes. The included identifying data did lower the naïve bayes model's accuracy, but improved the AUC so the marginal effect was kept for the final model. The final column, balanced, represents whether the number of each play_type (pass or run) was equal. It was found during the EDA that passes had 60% of all play shares, so this could affect some models' predictions. The results in the table find almost no effect to the naïve bayes model, but a significant decrease in the random forest model. This shows that the random forest model might be more practical in applications when handling unbalanced data, since it accurately creates predictions using the known unbalanced data.

Hidden Layers	Nodes Per Layer	Activation	Accuracy	Loss
3	256	relu	*0.7190*	*0.5471*
2	256	relu	0.7173	0.5498
1	32	relu	0.7139	0.5593
1	128	tanh	0.7070	0.5692
3	32	relu	0.7157	0.5594

This table shown above displays the outcome of long training five different neural networks for over 200 epochs. An epoch is one iteration for the network to ingest through the entirety of the training dataset, and the models showed limited to no learning improvement after the 200-epoch mark. The first column represents how many layers make up the network for the data to be processed through. The next is the number of nodes that make up each layer in the previous column. The third column is the activation function used by each node to use on the input data multiplied with its weight or bias. While the data here might suggest that a higher

number of hidden layers and nodes per layer is the key, it is important to note that many more models were trained, and no such correlation was significantly found. For this task, it appears that the relationship between the hyperparameters of neural network architecture is more correlated and complex. However, it was clear that the rectified linear unit (*relu*) function was superior in this task to the *tanh* function.

Algorithm	Accuracy
Naïve Bayes	69.24%
Random Forest	71.60%
Neural Network	*72.20%*

The final table shows the outcome accuracy of the best three models of each algorithm type. Here we can see that the 2022 season appears to have been more predictable than the previous seasons in the dataset, since each of the three models scored higher accuracy in their predictions than they did on the validation dataset. However, the neural network was the clear best model, with an accuracy of over 72% on the 2022 season, which is more accurate than not only the other two models selected in this research, but of the hidden Markov-model used in previous research.

Data Visualizations

Naïve Bayes Classifier

As explained in the methodology, the analysis was to begin with creating a naïve bayes classifier on the dataset. This is because of the simplicity of the model, which makes it an easy baseline to compare more advanced methods to. When creating a naïve bayes model for binary classification, the researcher has no hyperparameters to tune and can only change the model by manipulating the training data. This manipulation was investigated and resulted in the four

separate models seen in both Appendix K and L. The first naïve bayes model used the untouched data from the dataset, containing many different ranges and units. This was still able to create a model with a modest accuracy of 64.9%, which is acceptable for a model that took minimal compute power and no tuning required. The AUC score is more acceptable 0.694, but the appearance of the ROC looks mostly linear with few points and offset towards the bottom-left corner. When implementing feature engineering, in the terms of min-max normalization for the positive integer variables and standardization for both normally distributed variables, we can see these attributes change in the ROC. The curve's focal point appears to have shifted towards the (1, 0) point of the plot, and the curve appears much smoother than before. This encoding improved the accuracy by nearly 2%, as can be seen in its confusion matrix in Appendix L and increased the AUC score by over .02 as well. The third model presented has the coach and team identifier variables added to the input data encoded as one-hot variables. Neither the accuracy or AUC changed significantly, with the accuracy actually lowering, but again the ROC is smoothed further to appear more like a perfect curve. The last model was created by balancing the training data before fitting the model to it, which involved equaling the number of runs to the number of passes for training in order to avoid overfitting from guessing passes. However, this can be seen to not be necessary for the naïve bayes method, where it marginally affected the ROC and AUC and accuracy was constant. By looking at the results in the model's confusion matrix, the model simply shifted pass predictions into more run predictions, both correct and incorrect. This is expected since we had altered the training data from a pass-favored proportion to a more pass-run-neutral split. From these results, we can gather the importance of normalizing and standardizing the data for a naïve bayes model, lack of importance in balancing training data, and created a baseline to measure the next modelling techniques.

Random Forest Classifier

The next machine learning modelling technique that will be explorer is the random forest. A random forest is constructed of a variable amount of decision trees, the number being selected at the model's creation. While there is this configuration available, the random forest is also seen as a more complex black box, with increased accuracy from minimal configuration. Hence, the same four training data manipulations are executed to create four random forest models, just as was done for the naïve bayes models. The ROC plots and the confusion matrices for each model are presented in Appendix M and N respectively.

Here, we can see a significant increase in each model's effectiveness in accomplishing the binary classification task. The deep forest without balancing the dataset, any standardization or normalization, or any identifying variables was able to gain an accuracy of 70.88% when predicting plays from the 2022 season. It is important to note that no code was changed in either the preparation of the dataset, or the training and testing of the model. Only the constructor of the model was changed from the BernoulliNB object from scikit-learn to the RandomForestClassifier from the same package. Thus, this improvement is simply from how the different algorithm is equipped to handle this binary classification task.

Adding team and coach identifier information to the dataset improves the accuracy of the model by over 1%, with the AUC seeing a very slight improvement as well. The shape of the curve, however, does not see any noticeable change. Interestingly, the other two additional models show an interesting difference between the data preparation between random forests and naïve bayes. The opposite effect can be seen in random forests as they were in a naïve bayes model, regarding variable normalization/standardization and target variable balancing. With the random forest model, standardizing and normalizing the different variables shows no real affect

in accuracy or AUC, but balancing the data improves the model substantially. By balancing the number of runs and passes in the training data (which had no effect on the naïve bayes model), the random forest model decreases in accuracy to near the level of the model without team and coach data. The AUC, in contrast, barely changed as can be seen in Appendix M.

This difference in accuracy between the two non-tuned algorithms could potentially be explained by the independence of the variables in the dataset, or rather lack thereof. The naïve bayes model implies an independence in the input variables where there is not always. In contrast, the architecture of a random forest allows for the capturing of complex relationships between the many variables of the dataset. It could be possible that the variables given in the dataset create a web of dependence between all other variables that is uniquely capturable from the deep forest algorithm. In any case, the extreme accuracy difference against the other models in this project as well as prior research calls for further investigation.

An added benefit to random forests over other machine learning techniques is their architecture of decision trees for different variables allows for display of each variable's importance. The feature importance value was pulled from each non-hot-one encoded variable from the best performing random forest model and shown in the plot at Appendix O. Here we can see that whether the offense approaches the play without huddle has nearly no effect, which makes the application of the model easier since no huddle plays would give the defense less time to utilize it. It is also interesting to see the next two least important variables, while still carrying some level of importance, are the down and quarter variables since those are often used in prior research's models limited number of variables. It is also peculiar that the most important feature is the time remaining in the quarter, since one would expect that this variable would make little

impact outside of when it is nearing the end. However, it's possible that that margin of the variable's range is in fact responsible for most of the feature's importance.

Feedforward Neural Network

The final model tested in this project was created from a feedforward neural network. Because of the known importance of normalizing and standardizing data, as well as prior balanced data, no full models were trained without either used on the appropriate variables. On a cursory exploration, it was also found additionally useful to include the identifying variables of coach and team as hot-one encoded variables in the data provided to the models.

With the same dataset following the previous preprocessing, over 100 feedforward neural networks were created using keras-tuner in order to find the layer architecture best suited for this task. Since this random search was completed using a consumer desktop computer, only one epoch was completed for each model in the search of those tested, the best performing model after one epoch consisted of an input dense layer composed of 128 nodes, 4 additional dense layers with 64 nodes each, and a final dense layer with a sigmoid activation to predict the binary variable. Each dense layer used the common rectified linear unit activation function. However, expanding this search to longer epoch-lengths found the highest accuracy model to have 3 hidden layers, comprised of 256 nodes using the relu activation function.

After ideal architecture was found, five models were tested to 200 epochs. The top performing model after 200 epochs's accuracy and loss graphs are shown in Appendix P. Loss is the function that the model tries to minimize in training, in this case "binary_crossentropy" was used. After around 100 epochs, we can see that the model continues to increase in training accuracy but stays stable in validation accuracy. This is because the model has started to overfit

to the specific training data. This phenomenon can be seen in the collection of all long-trained models in Appendix Q. To try and prevent this, dropout was added to some models, which randomly disables a percentage of nodes during training, however the overfitting still occued. However, even with this overfitting, the model was able to achieve accuracy levels above the naïve bayes and random forest with 72.2%. This is impressive when compared to the 71.5% of prior research, given that these models lack the time series data utilized in that model. It is possible that there is a complex dependence between the variables that the feedforward neural network is unable to grasp causing the overfitting from memorization. Therefore, further investigation in different neural network models may yield higher results.

Ethical Recommendations

When creating any predictive model such as the one in this project, many considerations must be discussed before proper usage. The widespread adoption of similar models could change how the game is played, as well as the careers of those in it. It is in the interest of both those creating and maintaining the model, as well as the personnel actively using it in the game to understand these recommendations.

In the game of basketball, teams found through statistical analysis that three-pointers have a high but under-represented expected value when compared to other shot types. This observation led to teams shifting their offense more towards three-pointers to the point that they are shot at a higher frequency now than ever in the history of the sport. While not immediately apparent in this current play prediction model, it is possible for a similar finding to be found that could change how league-wide play calling is handled. For instance, from the data we can see that teams run at a disproportionately high frequency on 1st and 10 than other downs and distances. If teams use the model and find that they would be less predictable passing on these downs, it may cause a similar shift to the three-pointer.

From a defensive perspective using the model, it is important to understand the accuracy of the model and its predictive properties. A single mis-prediction can not only have a massive effect on the outcome of a game, but also on one's career. In 2020, the New York Jets defensive coordinator was fired after calling his defense to blitz on a last second play that ended up costing the Jets a victory. If this model were to be applied to actual game situations, it is fundamental that the team understands the accuracy level is not perfect, and instead uses it as a tool in assisting play calling. It is also crucial for teams to consider the difference between the model's

overall predictions and the predictions it has for some teams, as the difference can result in substantial changes in probability.

With the offensive use case, it is also important to distinguish that what the model predicts, whether the play will be a run or pass, is not necessarily what the play should be. It would be common thinking that if the defense is predicting that you'll call a run or pass, that it would be optimal to call the opposite. However, the impact of unpredictability on the success of a play hasn't been measured. It's possible for the instance of a 3rd and long play to have a high predicted pass rate, but it be unadvisable to call the opposite when a predicted run play has a lower success rate than a predicted pass.

The effect that a play-predicting tool can have on a game has far reaching consequences. Not only have the fans and their entertainment value changed, but possibly the careers of the coaches and players alike. As seen previously, play calling decisions can lead to the firings of coaches. Similarly, players' future contracts are often impacted by their previous team's success, even without considering their contribution, so game-altering decisions may negatively shape their career path as well. This influence is why these considerations must be acknowledged before using any model like the one created in this project.

Challenges

A few problems were faced during the completion of this project that had to be addressed. The first was understanding and discovering the different methods of encoding the variables in the dataset to work with the selected machine learning methods. As shown from the difference in results between the models trained using encoded and raw datasets, the algorithms greatly benefit from proper encoding. Because of how important this is from a performance

perspective; it is crucial to make sure each individual variable is properly accounted for. This is particularly true in this dataset, which includes categorical, nominal, and ordinal data, along with each nominal variable being measured on a separate scale. By using my understanding of proper encoding techniques with different datasets to verify, the best results found were with the categorical variables hot-one encoded, the numerical variables with only positive values to be normalize, and the remaining numerical variables with both positive and negative values standardized.

Another challenge faced with working with this dataset was ensuring that the data was properly handled before being used between models. It is crucial in any machine learning applications to separate the training data from the testing data to ensure the model is creating predictions and not simply recalling its past training data. This is often done by simply splitting the data at a preferred percentage point. In this case, I dictated that the last season was the testing data, with the prior 23 seasons to be used for training. However, a syntax issue caused plays from 2022 to be included in the training dataset, and with over 200 variables once the dataset had been encoded, the problem was difficult to discover.

Recommendations

While progress from the prior algorithmic solutions to the research questions has been made, a number of improvements could be made on the research done in this project. The first of which is investigating a further number of machine learning techniques and their performance in the task. As can be seen from the difference in results between the current project and existing research, machine learning methods establish a further improvement in the predictions of NFL plays over simpler statistical approaches. But not all machine learning was found to be equally performant, as shown when comparing the deep forest to the neural network. Thus, the models that should be looked at should be those that are able to accurately model the dependent relationships between variables like deep forests. One such model is a gradient-boosted tree, which iteratively improves an ensemble of weak learners into stronger predictors, often outperforming the deep forest technique in exchange for much higher computational requirements.

Another improvement would be further investigating the addition of historical information into each play's features. This project included variables often removed in other approaches (coach and team) that allow for the model to create relationships due to their historic play-calling tendencies. Along with that, this project also created other variables including information from the play's history in both the number drive the current drive is in the game, and the current play the play is in the drive. These could be expanded on by including not only values for the location of the play in the game from play and drive number, but also what plays were run for previous plays. This may feature an improvement on the current dataset, without requiring additional information to be collected.

References

Alamar, B. (2010). *Measuring Risk in NFL Playcalling*. Journal of Quantitative Analysis in Sports, 6(2). https://doi.org/10.2202/1559-0410.1235

Axson, S. (2023, February 14). *Super Bowl 57 averages 113 million viewers, third most-watched television show in history*. USA Today. Retrieved April 23, 2023, from https://www.usatoday.com/story/sports/nfl/super-bowl/2023/02/14/super-bowl-57-becomes-tvs-third-most-watched-show-history/11211106002/

Cimini, R. (2020, December 7). *New York Jets fire defensive coordinator Gregg Williams after controversial Blitz call.* ESPN. Retrieved April 16, 2023, from https://www.espn.com/nfl/story/_/id/30469684/source-new-york-jets-fire-defensive-coordinator-gregg-williams-controversial-blitz-call

Connor, J. T., Martin, R. D., & Atlas, L. E. (1994). *Recurrent neural networks and robust time series prediction*. IEEE Transactions on Neural Networks, 5(2), 240–254. doi:10.1109/72.279188

Gough, C. (2023, March 30). *NFL League and team sponsorship revenue 2022*. Statista. Retrieved April 23, 2023, from https://www.statista.com/statistics/456355/nfl-league-team-sponsorship-revenue-worldwide/

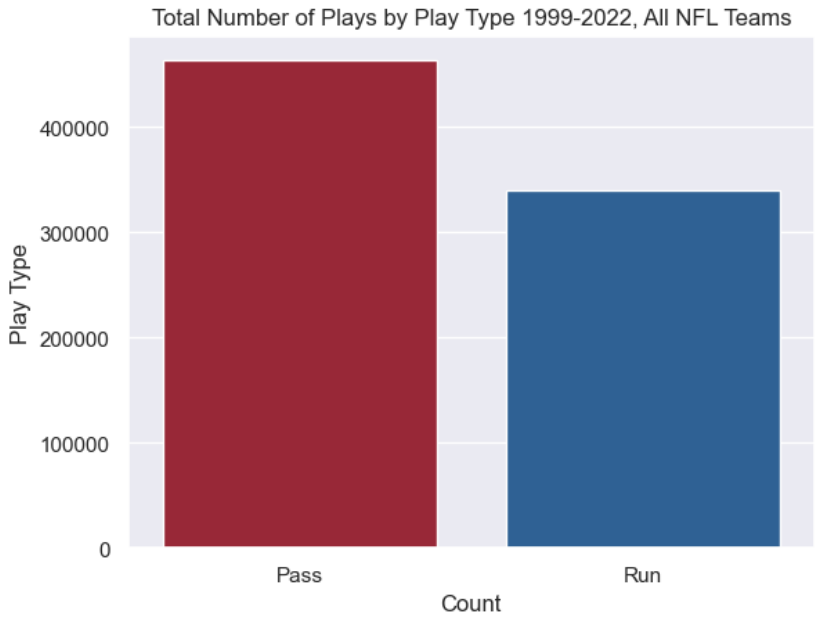
NFL Communications. (n.d.). Retrieved April 23, 2023, from https://www.nflcommunications.com/

The NFL Ops Team. NFL Football Operations. (n.d.). Retrieved April 23, 2023, from https://operations.nfl.com/inside-football-ops/nfl-operations/the-nfl-ops-team/

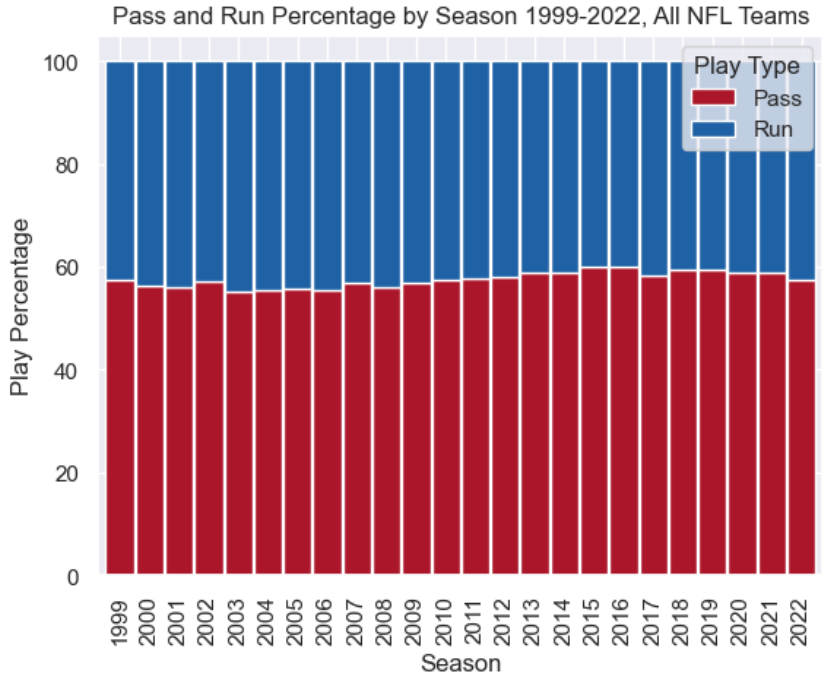
Ötting, M. (2020). *Predicting play calls in the National Football League using hidden Markov models*. ArXiv [Stat.AP]. Retrieved from http://arxiv.org/abs/2003.10791

Appendix

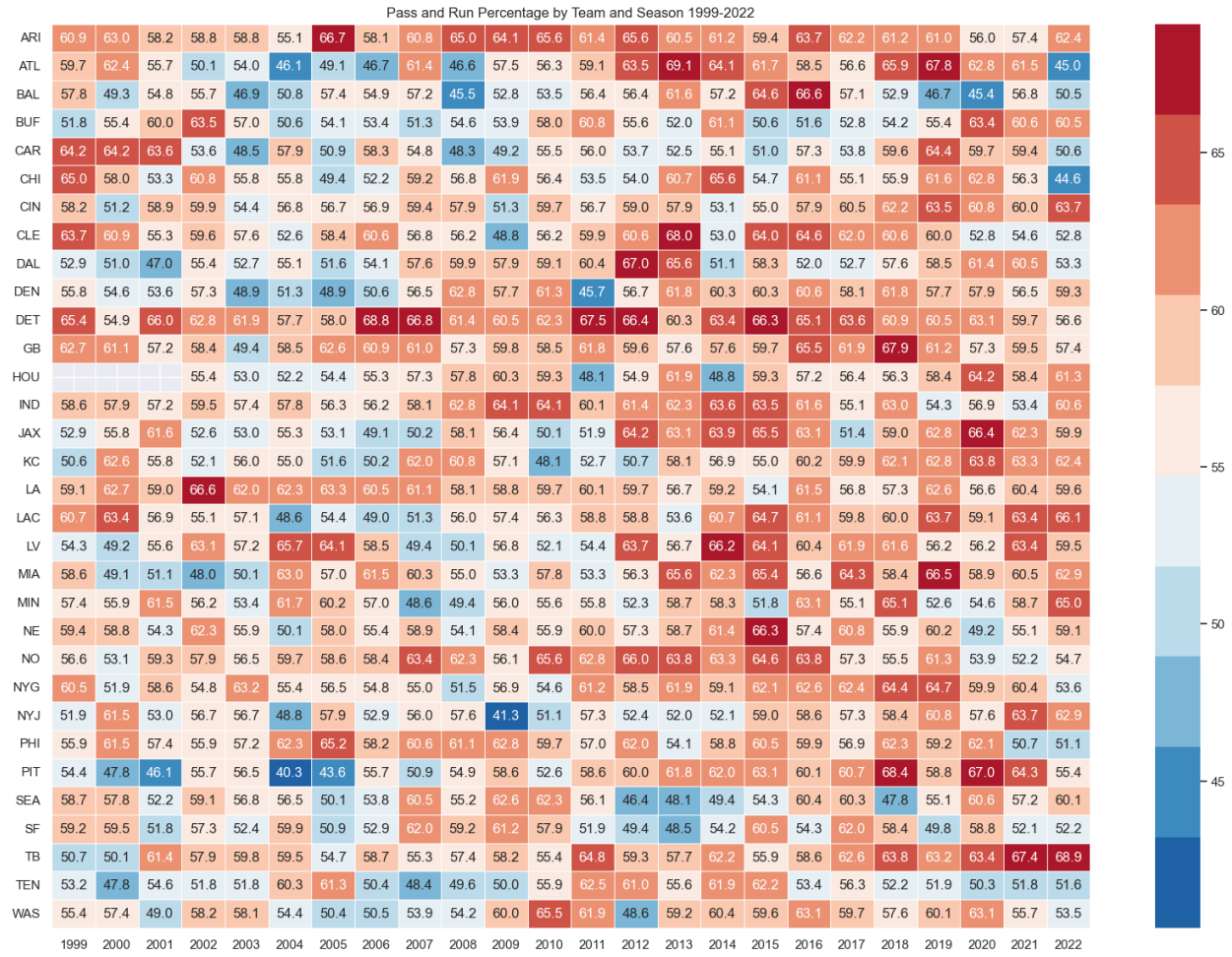
A. Total Number of Plays by Play Type 1999-2022, All NFL Teams



B. Pass and Run Percentage by Season 1999-2022, All NFL Teams



C. Pass and Run Percentage by Team and Season 1999-2022



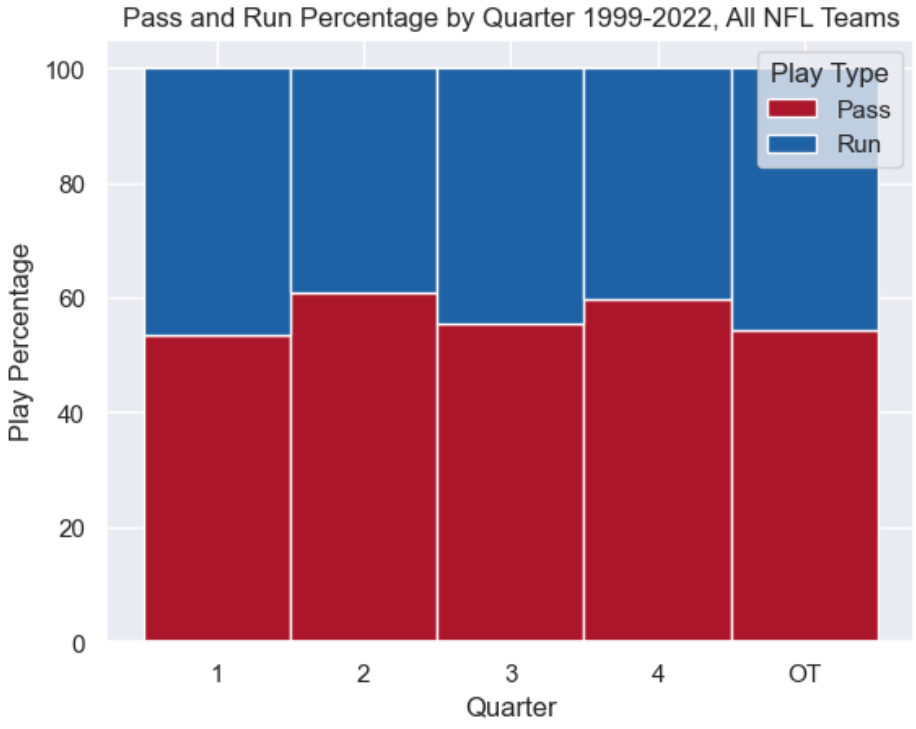
D. Ten Highest Pass Percentages by Coach 1999-2022

Coach	Pass Percentage	Run Percentage
Aaron Kromer	70.40%	29.60%
Rob Chudzinski	68.00%	32.00%
Rod Marinelli	65.72%	34.28%
Jim Bates	65.66%	34.34%
Kevin O'Connell	64.96%	35.04%
Brandon Staley	64.78%	35.22%
Marty Mornhinweg	64.45%	35.55%
Gus Bradley	64.04%	35.96%
George Seifert	63.99%	36.01%
Jim Caldwell	63.86%	36.14%

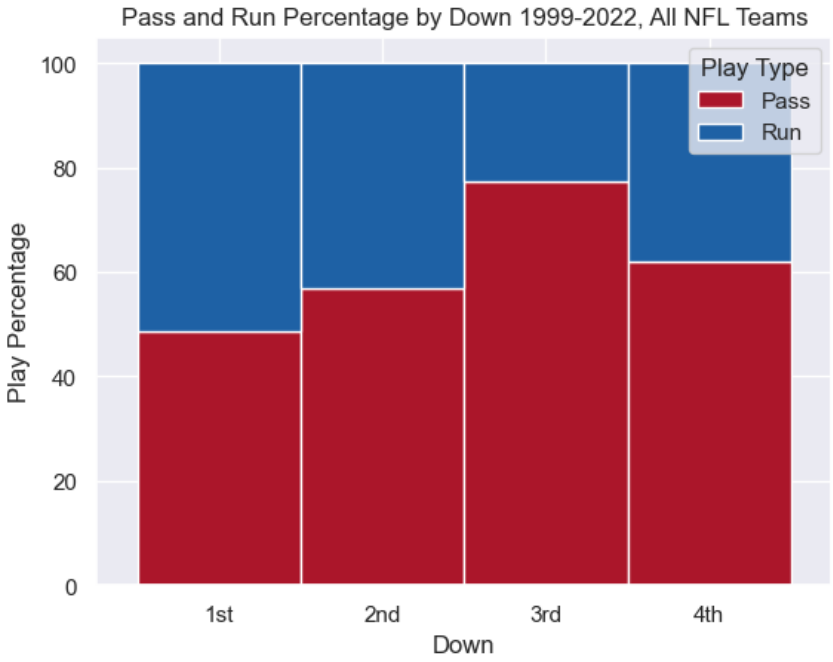
E. Ten Highest Run Percentages by Coach 1999-2022

Coach	Pass Percentage	Run Percentage
Matt Eberflus	44.57%	55.43%
Lane Kiffin	48.63%	51.37%
Bill Cowher	49.86%	50.14%
Rex Ryan	50.85%	49.15%
Jim Harbaugh	50.89%	49.11%
Nick Sirianni	50.90%	49.10%
Dave Wannstedt	50.95%	49.05%
Dave Campo	51.09%	48.91%
Eric Studesville	51.20%	48.80%
Mike Vrabel	51.54%	48.46%

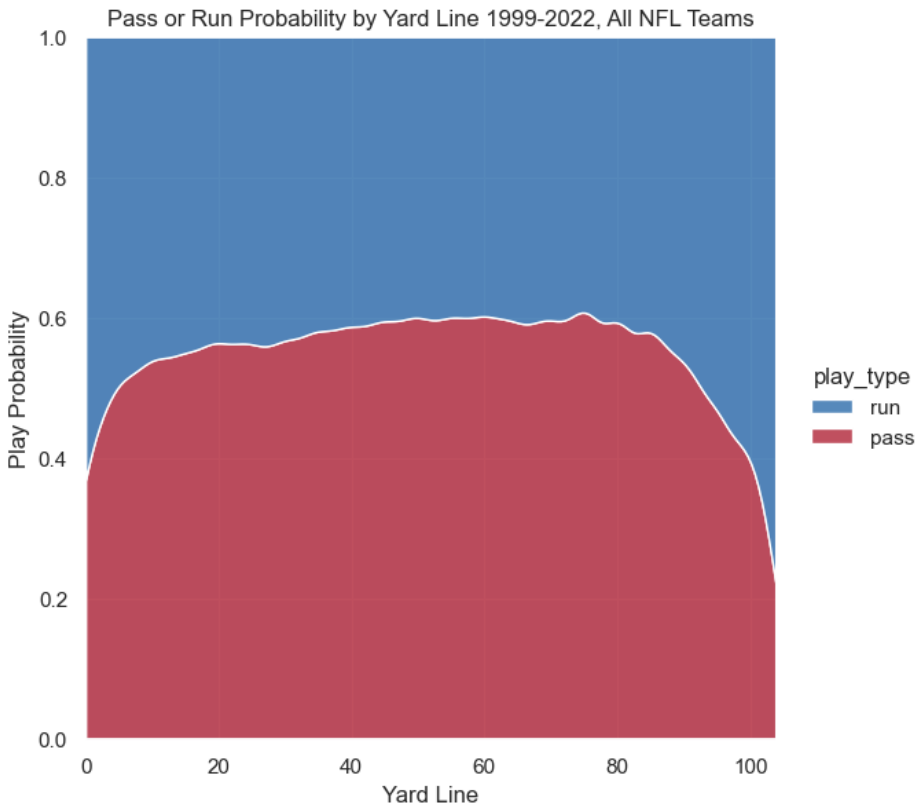
F. Pass and Run Percentage by Quarter 1999-2022, All NFL Teams



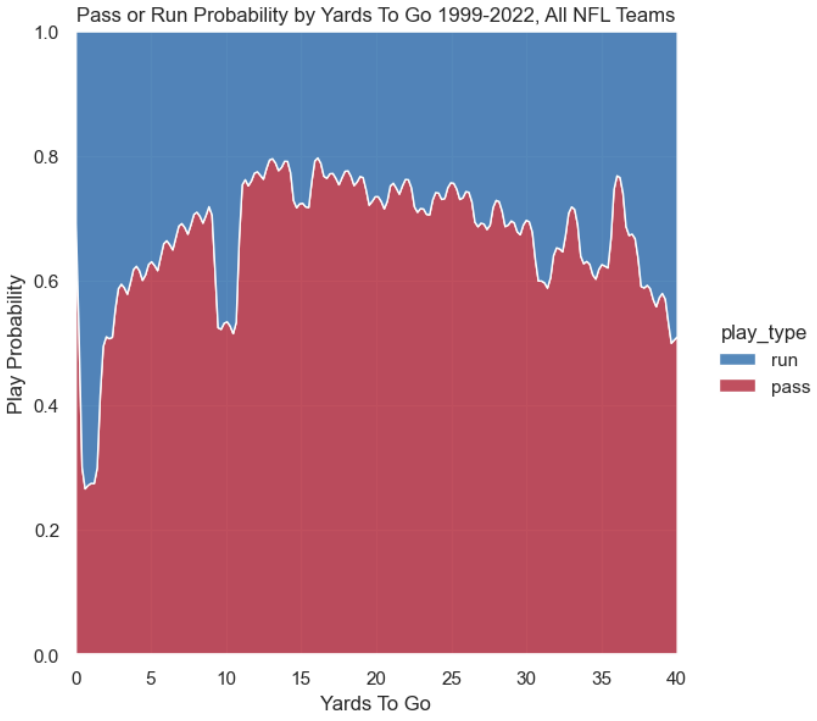
G. Pass and Run Percentage by Down 1999-2022, All NFL Teams



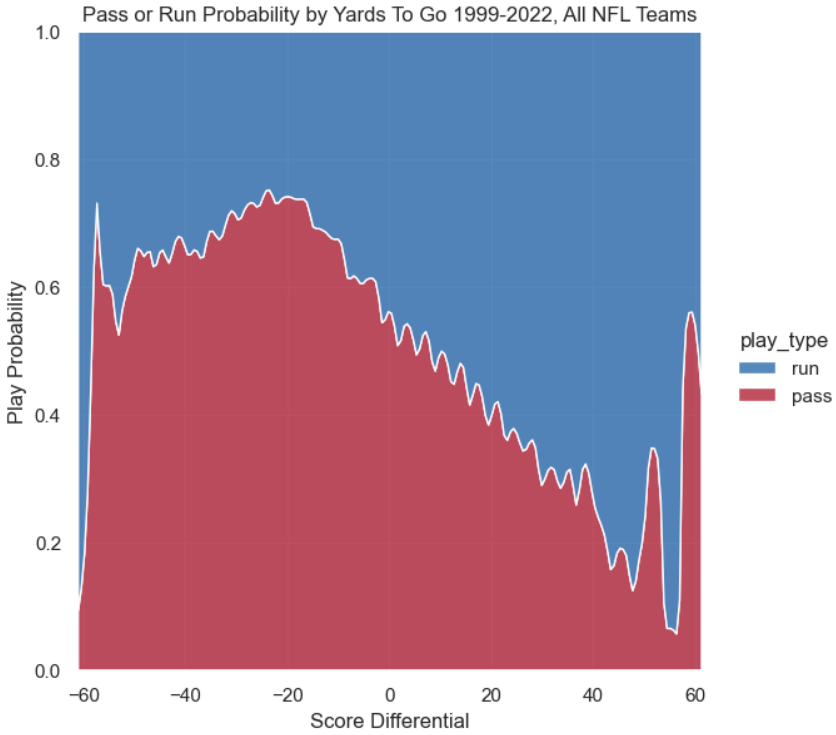
H. Pass or Run Probability by Yard Line 1999-2022, All NFL Teams



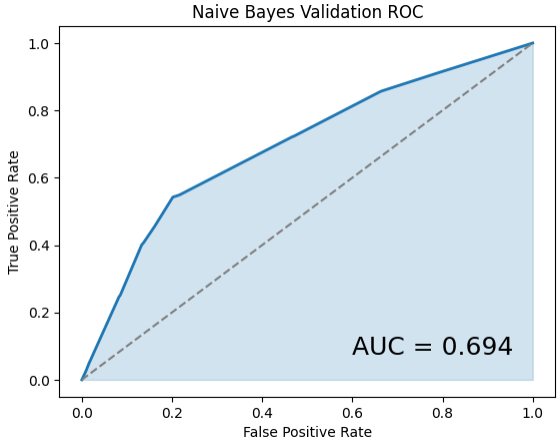
I. Pass or Run Probability by Yards To Go 1999-2022, All NFL Teams

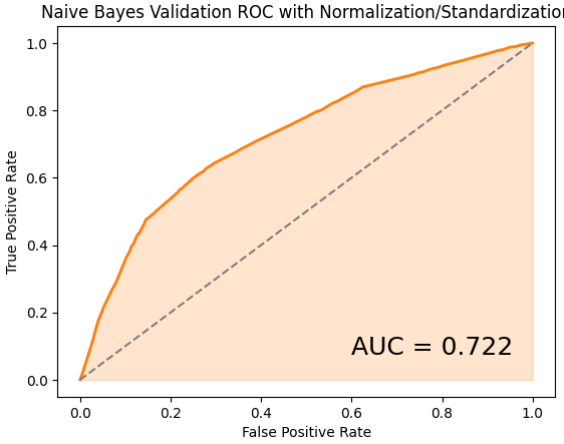


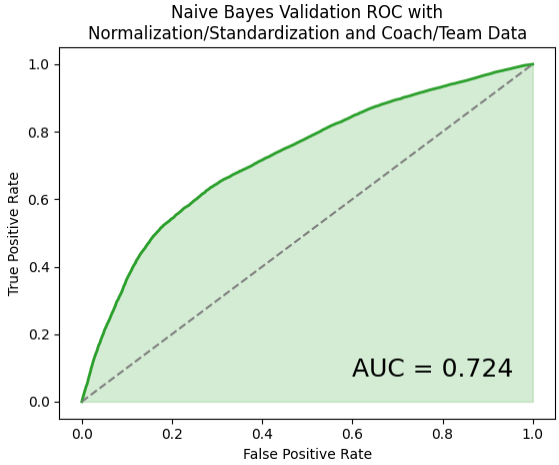
J. Pass or Run Probability by Yards To Go 1999-2022, All NFL Teams

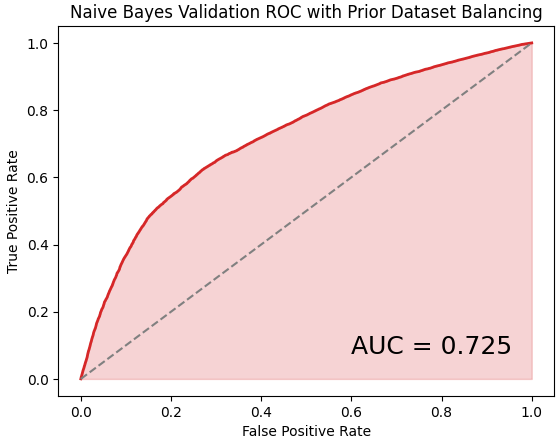


K.

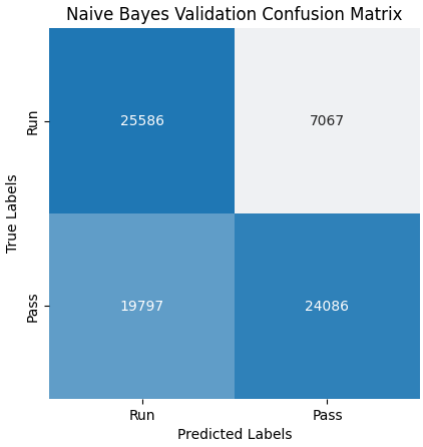

Naive Bayes Validation ROC
True Positive Rate
False Positive Rate
AUC = 0.694

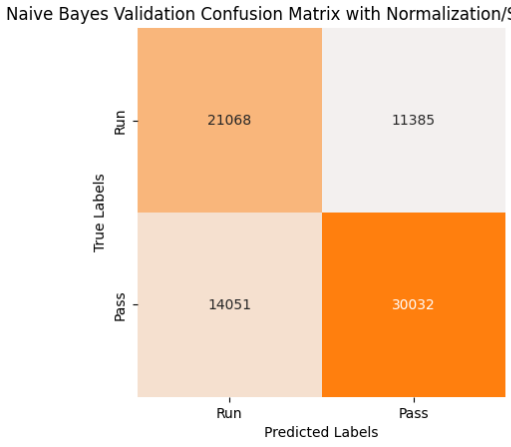

Naive Bayes Validation ROC with Normalization/Standardization
True Positive Rate
False Positive Rate
AUC = 0.722

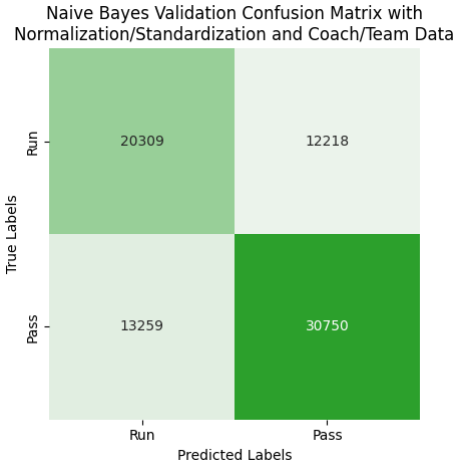

Naive Bayes Validation ROC with
Normalization/Standardization and Coach/Team Data
True Positive Rate
False Positive Rate
AUC = 0.724

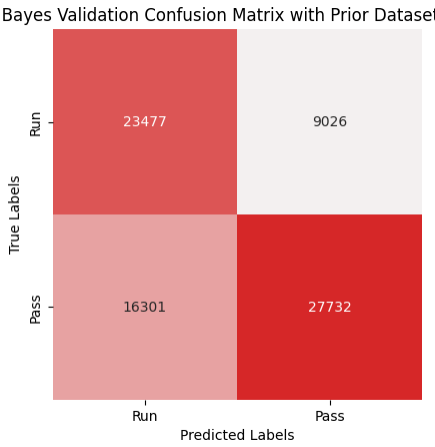

Naive Bayes Validation ROC with Prior Dataset Balancing
True Positive Rate
False Positive Rate
AUC = 0.725

L.

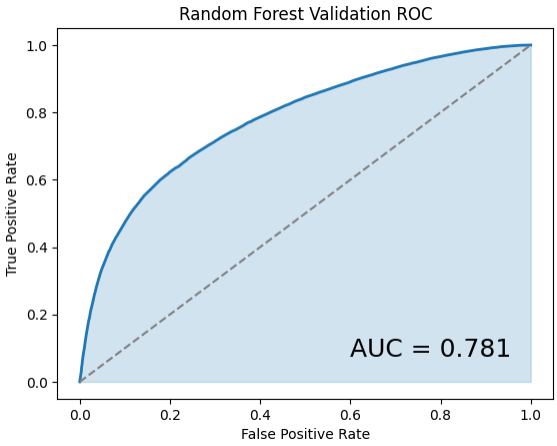

Naive Bayes Validation Confusion Matrix
True Labels
Run
Pass
25586
7067
19797
24086
Predicted Labels

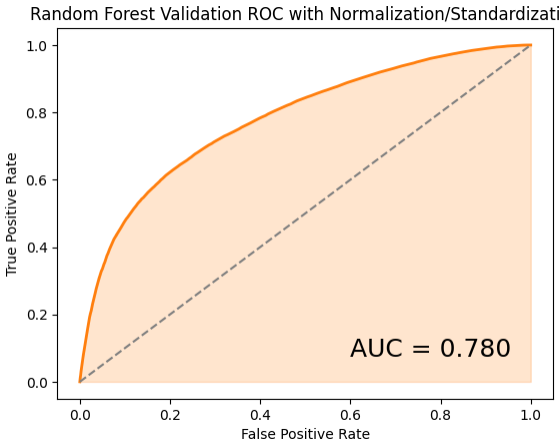

Naive Bayes Validation Confusion Matrix with Normalization/Standardization
True Labels
Run
Pass
21068
11385
14051
30032
Predicted Labels

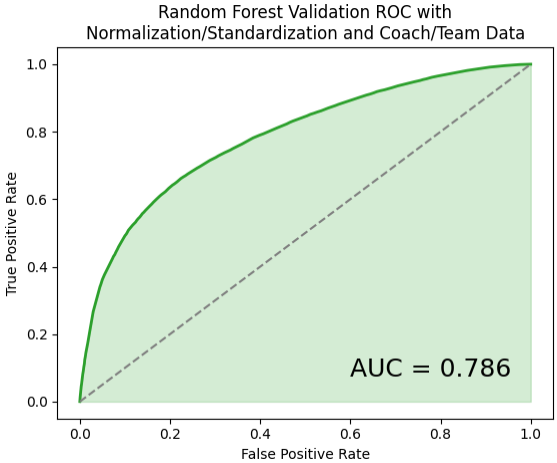

Naive Bayes Validation Confusion Matrix with Normalization/Standardization and Coach/Team Data
Run
Pass
True Labels
20309
12218
13259
30750
Run
Pass
Predicted Labels

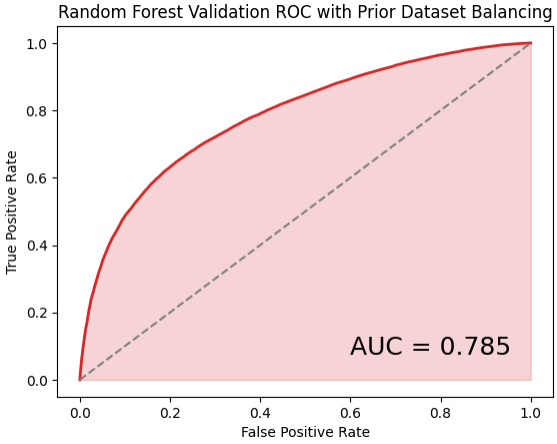

Naive Bayes Validation Confusion Matrix with Prior Dataset Balancing
Run
Pass
True Labels
23477
9026
16301
27732
Run
Pass
Predicted Labels

M.

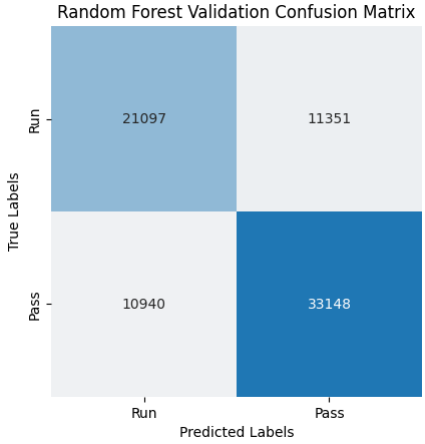

Random Forest Validation ROC
True Positive Rate
False Positive Rate
AUC = 0.781

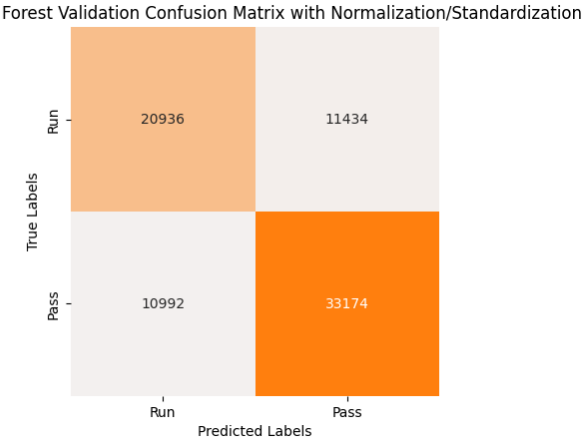

Random Forest Validation ROC with Normalization/Standardization
True Positive Rate
False Positive Rate
AUC = 0.780

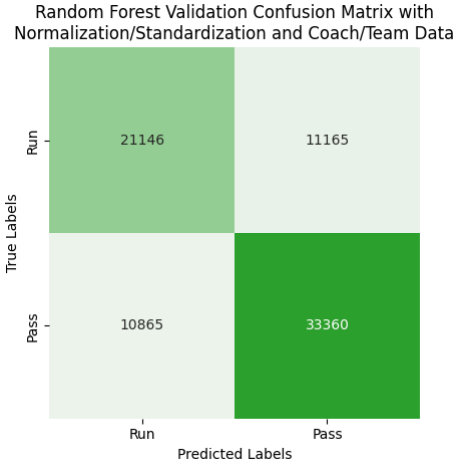

Random Forest Validation ROC with
Normalization/Standardization and Coach/Team Data
True Positive Rate
False Positive Rate
AUC = 0.786

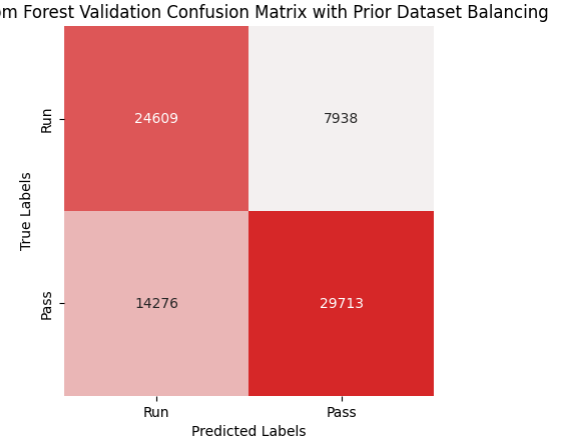

Random Forest Validation ROC with Prior Dataset Balancing
True Positive Rate
False Positive Rate
AUC = 0.785

N.

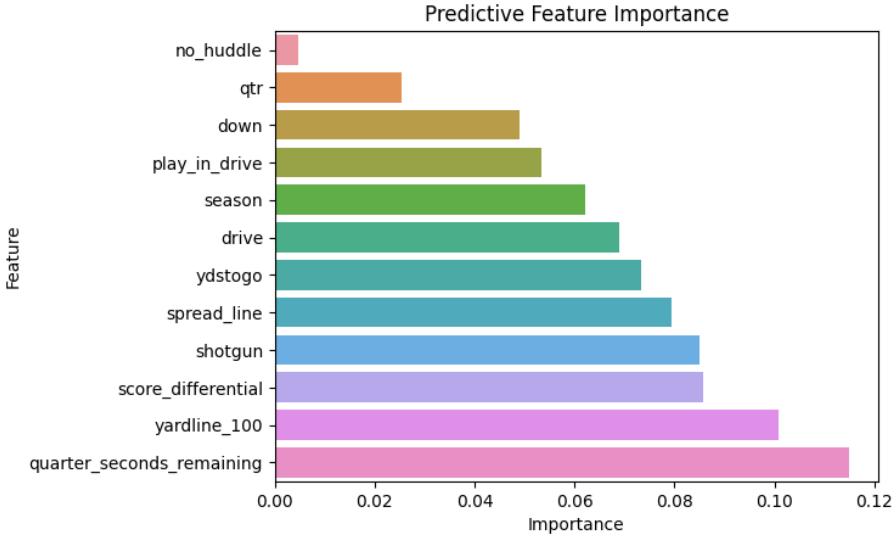

Random Forest Validation Confusion Matrix
21097
11351
10940
33148
Run
Pass
True Labels
Predicted Labels


Random Forest Validation Confusion Matrix with Normalization/Standardization
20936
11434
10992
33174
Run
Pass
True Labels
Predicted Labels


Random Forest Validation Confusion Matrix with Normalization/Standardization and Coach/Team Data
21146
11165
10865
33360
Run
Pass
True Labels
Predicted Labels

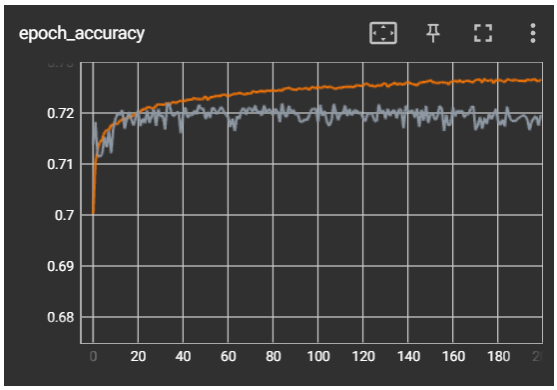

Random Forest Validation Confusion Matrix with Prior Dataset Balancing
24609
7938
14276
29713
Run
Pass
True Labels
Predicted Labels

O.


Predictive Feature Importance
no_huddle
qtr
down
play_in_drive
season
drive
ydstogo
spread_line
shotgun
score_differential
yardline_100
quarter_seconds_remaining
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance

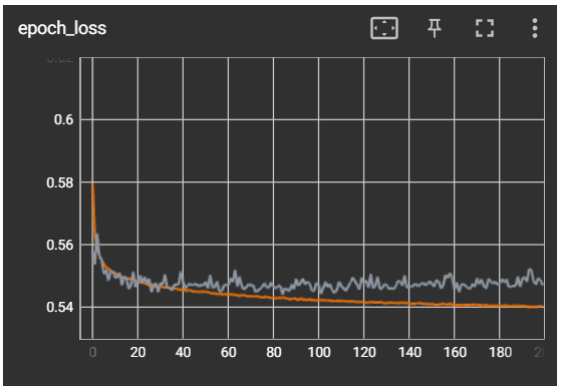
P.

Accuracy vs. Epoch



Orange = Training, White = Validation

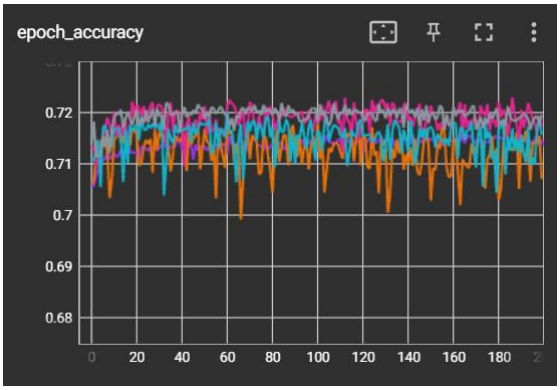
Loss vs. Epoch



Orange = Training, White = Validation

Q.

Validation Accuracy vs. Epoch



Validation Loss vs. Epoch