Football Play Type Prediction and Tendency Analysis

by

Karson L. Ota

B.S. Computer Science and Engineering
Massachusetts Institute of Technology, 2016

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTER SCIENCE AND ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2017

Signature of Author: _____
Department of Electrical Engineering and Computer Science
May 25, 2017

Certified by: _____
Christina Chase
Lecturer
Thesis Supervisor

Accepted by: _____
Christopher Terman
Chairman, Master of Engineering Thesis Committee

Football Play Type Prediction and Tendency Analysis

by

Karson L. Ota

Submitted to the Department of Electrical Engineering and Computer Science
on May 25, 2017 in Partial Fulfillment of the
requirements for the Degree of Master of Engineering in
Computer Science and Engineering

ABSTRACT

In any competition, it is an advantage to know the actions of the opponent in advance. Knowing the move of the opponent allows for optimization of strategy in response to their move. Likewise, in football, defenses must react to the actions of the offense. Being able to predict what the offense is going to do before the play represents a tremendous advantage to the defense.

This project applies machine learning algorithms to situational NFL data in order to more accurately predict play type as opposed to the widely used and overly general method of general statistics. Additionally, this project creates a way to discern tendencies in specific situations to help coaches create game plans and make in game decisions.

Thesis Supervisor: Christina Chase

Title: Lecturer

I. INTRODUCTION

Today, analytics are changing the way that sports are played. In the past, much of the preparation for a game came in the form of scouting and film analysis. However, with the growth of technology, more and more data is becoming available. One example of this is football. Each play is now documented, meaning there are more statistics available than ever. Now, information is available about the game situation, players, and even player location for every single play in every single game, though some types of data are proprietary or have restricted public access.

Football is played in the form of discrete plays. The offense attempts to advance the ball down the field. It has four attempts to advance the ball ten yards, with the eventual goal of advancing the ball into the end zone. Each time the offense advances the ball the required ten yards, it gets a new set of downs to again advance the ball.

Each offensive play takes the form of one of two types: a run play or a pass play. On a run play, the offense simply tries to advance the ball down the field by running forward with it. On a pass play, the ball is thrown down the field to another member of the team. Not only are these plays very different, they typically have very different outcomes. Run plays produce a lower gain in yardage than pass plays on average. However, they also have a much lower variance in yardage gained than pass plays as well. Variance can be considered a proxy for risk, as pass plays provide a more volatile sequence of gains. Additionally, pass plays carry a higher risk of turning the ball over to the other team. Because of these aspects, run plays are generally considered safer, while pass plays have significantly more risk, but higher gain on average. Table 1 describes the average and standard deviations of gains on run and pass plays.

	Run	Pass
Average Gain (Yards)	4.185	6.359
STD Gain (Yards)	6.330	10.036

Table 1

As the types of possible plays are so different, they are defended very differently. On a run play, the defense converges toward the line of scrimmage to tackle the ball carrier. On a pass play, the defense typically spreads out away from the line of scrimmage in order to cover the receivers running down the field. Due to different defensive strategies in response to the different play types, the defense would benefit significantly if it were able to accurately predict what the offense is going to do.

A. Problem Statement

Play type prediction is hugely beneficial to defensive coaches in football. One way to produce play type predictions is by using machine learning to model play calling. The discrete nature of football plays makes situational data perfect to apply machine learning to. Each play can act as a single data point. What's more, as they are separate, they do not depend on each other, meaning that it is not necessary to analyze a sequence of plays as a whole. Therefore, each play can be viewed as its own instance, distinct from the plays that come before and after it. This is an ideal setup for most machine learning algorithms. Each play functions as a sample, with a binary prediction of run or pass.

Play type prediction has many applications. One such application is to better understand the tendencies of coaches in aggregate. For example, by modeling all play calls through machine learning, it is possible to predict actions that an offense will take in any given situation. This allows us to understand the general thinking and actions taken by coaches, and allows us to better understand the way that coaching decisions are made. One approach to gain such an understanding would be to simply create a database of all plays, query the specific situation, and utilize a basic statistical breakdown of play types in that scenario. However, this approach is flawed in a few critical capacities. First, while there is a wealth of data from a full season of NFL plays, each specific scenario has relatively few samples. For example, there are relatively few third down plays with seven yards to go from midfield with the offense ahead by ten and twelve minutes to go in the second quarter. Such a scenario would yield very few samples with which to do a statistical analysis. Another flaw in using raw statistics is that they often yield discontinuous results. For example, all else being equal, a play from the twenty yard line should have nearly identical results as a play from the twenty-one yard line. However, using raw statistics could very easily yield very different results. Therefore, viewing coaching tendencies from a purely statistical viewpoint is a less than ideal way to understand the true nature of coaching decisions. Instead, using machine learning to model the problem in aggregate would solve these problems, offers the opportunity to significantly improve prediction accuracy over baseline probability distributions, and provide valuable insights into coaching decisions.

Another application of play type predictions is to produce a viable tool for coaches to use to adjust their strategy, both before and during a game. Understanding the

tendencies of an opposing team during the week of preparation leading up to the game would allow a defensive coach to create a game plan that better incorporates the behavior of the opponent. Furthermore, play type prediction could be used within the game itself. An understanding of how an offense is expected to behave in a specific situation could allow a coach to appropriately incorporate this information into each and every play call. However, while plays are discrete and there is a pause between each play, the pause is typically forty seconds at most. Therefore, being able to make a prediction quickly is of the utmost importance. Today, most predictions are made off of probability distributions that simply describe how often an offense chooses to run or pass. These predictions have a high degree of error, and thus are used mostly as a guideline for defensive play calls. It would be reasonable to meet the time constraints needed to make a prediction between plays. A model could easily be trained before a game, and then a prediction could be made in a matter of seconds between plays. Such a tool would be a huge advantage to a defensive coach, and could significantly improve defensive strategy. A higher degree of accuracy in play type prediction would significantly lift the confidence that defensive coaches have using the tool to influence strategy.

II. DATA

Data for the 2016 NFL season was scraped from pro-football-reference.com [FOR]. Though there are other sources of play-by-play data, they typically have numerous errors and irregularities. In comparison, the pro-football-reference data is very reputable and the site is well known throughout the sports realm. Additionally, large

sports networks, such as ESPN, often reference their analysis and data. The data is stored in a large index separated by run plays and pass plays, searchable by each team in the NFL. I was able to collect data by each team and play type, then compile it into a single large data set. I was able to get every recorded play from scrimmage for the entire 2016 season, a total of 32,994 plays. Unfortunately, pro-football-reference did not have detailed formation, personnel, or player location data. Had these been available, they could have been incorporated. A play from scrimmage is defined as a play run by the offense that results in either a pass play or a run play. This excludes all special teams plays and all plays that are nullified by a penalty. I chose to incorporate different statistics describing the game situation based on their relevance to football play calling. By manipulating the structure of each statistic, I was able to compile a data set that would function well as input to an algorithm.

A. Down

The down describes which of the allotted four attempts is being performed to gain the required ten yards. Typically, offensive coaches change their strategy to account for the amount of attempts they have remaining to gain the necessary yardage. Fourth down is usually not a play from scrimmage, as strategy usually dictates that the offense punts the ball or kicks a field goal. Both are types of special teams plays, and thus are not counted as plays from scrimmage.

B. Distance

The distance describes how far the offense needs to advance the ball to gain another first down. This has a significant impact on offensive play choice. As described

earlier, a pass play is much riskier, but also averages a higher amount of yards gained. Therefore, with a long distance remaining to achieve a first down, an offensive play caller might be more inclined to call a pass play, especially on later downs. Conversely, for very short distances remaining, a run play might be called more often as it is safer and more consistent.

C. Score Differential

Score differential represents the difference in score between the two teams. This differential could either be positive (the offensive team is winning) or negative (the offensive team is losing). This has a huge impact on choice of offensive play. Winning teams have an incentive to allow the clock to run and maintain longer drives with less risk by running the ball. In a similar vein, losing teams tend to take larger risks to try to get back into the game.

D. Time Remaining in the Game

Play calling is often affected by the amount of time remaining in the game. Depending on the score differential, coaches typically react to the time remaining in the game in two ways. Teams that are winning the game tend to run the ball more, while teams that are losing tend to throw the ball more. Part of this choice corresponds to the riskiness of each play type teams that are losing are more willing to take more risk to get back into the game. Additionally, each play type has a different rule regarding the way the game clock operates. An incomplete pass stops the clock from running until the start of the next play. Running the ball keeps the clock ticking. This incentivizes losing teams to pass the ball more, as it preserves the time remaining the game. Conversely, teams

that are winning are incentivized to run the ball in order to allow the clock to keep running.

E. Field Position

Field position refers to the location of the line of scrimmage on the field. The field is 100 yards long. This is impactful on play calling, especially at either end of the field near either end zone. Offenses that are very close to their own end zone must be careful of the risks of being tackled in their own end zone, which gives the other team two points and the ball. Often times, coaches opt to be extra conservative in theses situations, choosing to try to gain a few yards to improve field position before punting the ball away.

If the offense is close to the other teams end zone, the dimensions of the field tend to have an impact on the choice of play call. For example, an offense that is very close to the goal line has a limited distance down the field that they can throw the ball. The end zone is ten yards long. Therefore, if an offense is ten yards away from the goal line, it is limited by the fact that it can only throw the ball 20 yards down the field into the end zone. This limitation reduced the amount of area that a defense needs to cover. This makes passing near the goal line significantly more difficult compared to the middle of the field.

III. AGGREGATED VIEW PROCEDURE

An aggregated view of football play types allows analysis of the NFL as a whole.

Likewise, predictions in this context are more about determining how to model play calling tendencies. It can help identify when certain actions are common, and how coaches generally react to given situations.

In order to make predictions, I first encoded the situational football data as feature vectors. Each of the individual components in the feature vector was normalized to a [0,1] range in order to avoid scaling bias. I developed several different models based on down. There was a separate model for first, second, and third down, as well as a model with all downs included. This was done because each down has significant and distinct play calling tendencies. Each is its own category, with very different distributions of play calling, as shown below.

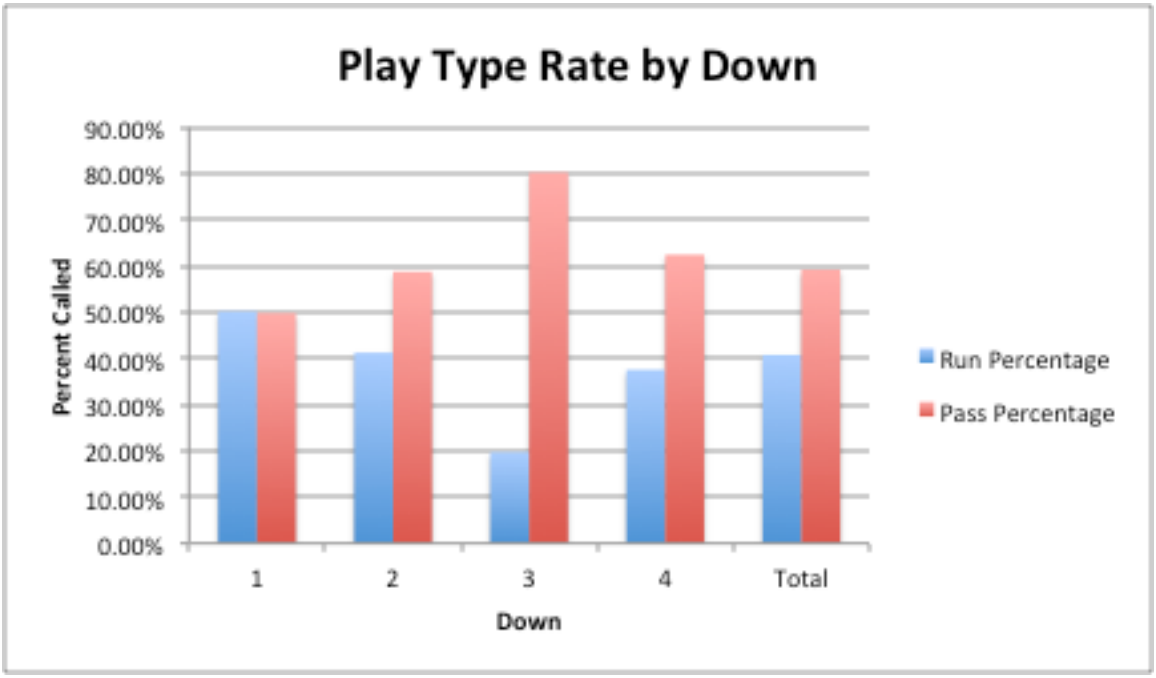

Figure 1

Fourth down was not included as its own model, as there are very few plays from

scrimmage on fourth down. Typically, an offensive team either punts or kicks a field goal on fourth down, so there were less than 500 total fourth down plays.

A. Creating Feature Vectors

For use in a neural network, the data needed to be combined into feature vectors with corresponding labels. Each individual component of each vector was normalized to a [0-1] range. Distance began as a positive integer, and was transformed to the proper scale by normalizing on the total range of distance values in the data set. Score differential was attained by taking the difference of the offensive team's score and the opponent's score, and then normalized. Time remaining in the game was transformed from the time remaining in each quarter to the number of minutes remaining in the total game. In the case of overtime, the time remaining is measured as the total time remaining in overtime, regardless of whether or not the game ended early due to a score. For games that went to overtime, the regulation time is still measured as the amount of time until regulation time ends. Field position was measured as the total distance remaining until the goal line. The labels were created by using a binary variable, where a run play is represented as 1 and a pass is represented as -1.

Once all of the data had been vectorized, it was separated into training, validation, and test data sets. In order to ensure that the data in each set would be representative of the total play distribution as a whole, the test and validation sets were composed of 1000 plays each, where each play was randomly selected from the total data set. The training set was then the remaining 30,994 plays. This allowed for a significant amount of plays to use to train the network, but also left reasonable amounts of data to be used to validate

and test the accuracy of the model. For each of the models by down, the validation and test sets were made to be 500 plays each, as the total number of plays was reduced. Each down still had at least 6000 examples in it.

B. Neural Net Architecture

I used the scikit-learn [Ped+11] implementation of a neural network for this project. The implementation is that of a multi-layer perceptron, with fully connected layers. This implementation allows the user to specify several parameters, including the activation function of the hidden units, the solver to iteratively minimize log-loss, the number of hidden layers and, number of units in each layer, L2 regularization parameter, and whether to use early stopping.

C. Validation

The various parameters to the neural network allowed for significant performance improvement through validation. Each of the following parameters was tested on the validation set to achieve the best performance possible. Early stopping, a technique that stops iteration once log-loss no long improves, was tested but ultimately not used in the final implementation. Use of early stopping resulted in a decrease in accuracy on the validation of .03 for each model. Furthermore, it did not offer any improvement as a regularization technique, as it did not significantly improve the ratio of train accuracy to validation accuracy. The implementation offers three activation functions: tanh, logistic, and relu. Relu achieved the best performance, improving prediction accuracy on the validation set by .05 over both of the other activation functions. Of the possible solver functions to reduce loss - stochastic gradient descent and limited memory BFGS

(LBFGS) - LBFGS produced a score .04 better than stochastic gradient descent. The L2 regularization parameter that optimized performance is .0001, which was found using a search over different orders of magnitude from 0.1 to .000001. Finally, numerous architectures were tested. In general, larger layers with 100 units outperformed smaller layers with ten units. Furthermore, a flat architecture outperformed both architectures with increasing and decreasing number of units. Additionally, increasing the number of hidden layers produced increased accuracy through three layers, and then decreased. Therefore, a final architecture of three hidden layers with 100 units each was chosen.

IV. AGGREGATED VIEW RESULTS

Each model was tested against a baseline of always predicting the most common play type in each scenario. Thus, for the first down model, run was predicted every time, whereas for the second down, third down, and total models, pass was predicted every time. In each category, the model showed improvement over the baseline, with improvements of more than ten percent in the first down, second down, and total models. The table below shows the scores in for each model compared to the baseline.

	Total	1st	2nd	3rd
Model	68.90%	61.80%	64.80%	86.80%
Baseline	61.30%	51.20%	59.80%	83.60%

Table 2

Since the test set for the total model was 1000 plays, the increase represents 99 more plays that were correctly predicted. Likewise, since the test sets for first down were 500 plays each, the increases in accuracy represent 79, 46, and 30 more plays that were

correctly predicted respectively. Relative to the overall test set size, these are large increases that represent a significant opportunity to affect strategy for a defense. The chart below shows the models performance relative to the baseline.

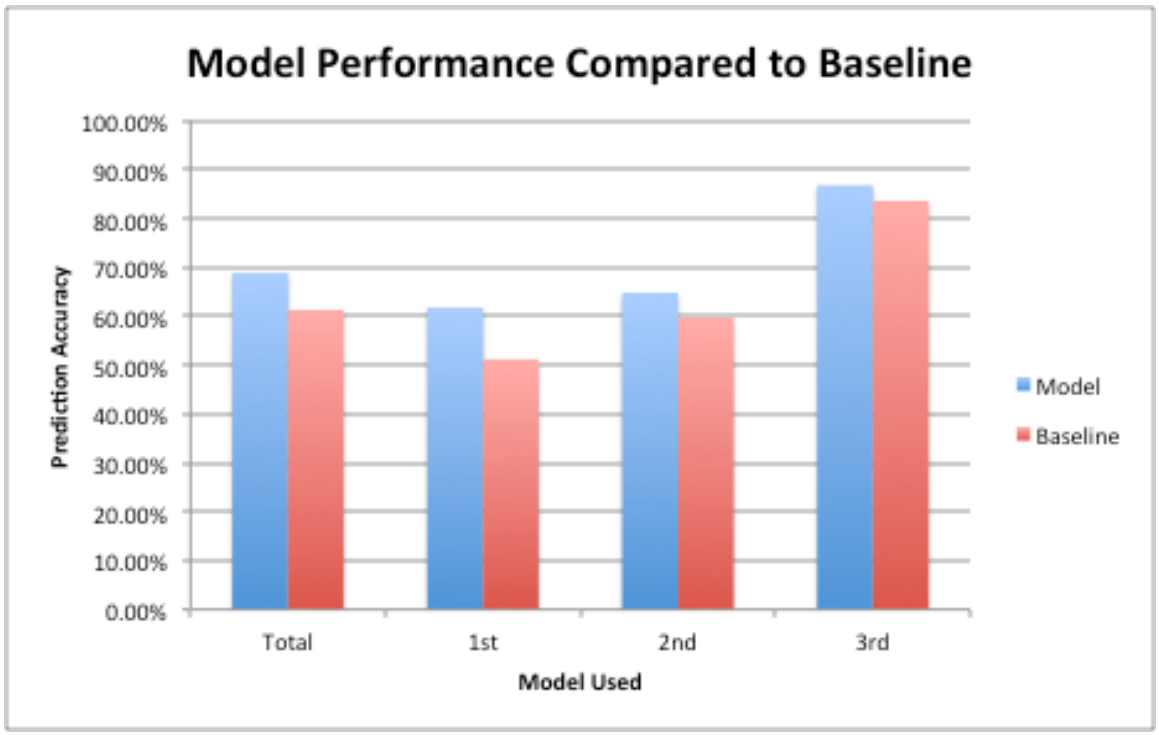

Figure 2

V. CONSULTATION WITH COACH CHAD MARTINOVICH

I met with MIT Head Football Coach Chad Martinovich in order to gain insight into how he creates a game plan for the week, what information sources he uses to create his game plan, and what would be most useful to him.

While Coach Martinovich coaches at the NCAA DIII level, there are several parallels that transcend all levels of coaching. First and foremost, coaches rely heavily on game film and film analysis to formulate a game plan. Data regarding each individual play is derived from film. Secondly, coaches design a game plan around the tendencies

of their opponent. Martinovich broke down the term tendencies into two categories, one with respect to players, and one with respect to schematics. He mentioned how he might take special notice of special skills players may possess, or areas of a game a player may excel in. For example, a quarterback may be exceptionally accurate with his passes, but might not possess the arm strength to throw the ball deep down the field. Similarly, Martinovich took special interest in schematic tendencies. This ranged from a high level such as how much a team runs or passes the ball, all the way down to the blocking schemes and route combinations. He said that any improvement in these tendencies, or ability to be more specific and situational in any sort of tendency analysis would be helpful, especially in a way that would not be especially burdensome to his small staff.

Meeting with Coach Martinovich was a critical step forward. He confirmed that the problem of tendency analysis has room for improvement and has a practical application. Furthermore, any way to do such analysis in a way that would minimize a burden on the staff would be helpful. This idea also applies to big NCAA D1 and NFL teams. While they certainly have larger budgets and more manpower than MIT, they do not have unlimited resources, and any way to do something better or more efficiently could free up those resources to be used elsewhere.

One takeaway from meeting with Coach Martinovich was that in order to build a practical tool, it is important to view analysis from the viewpoint of a coach. While an aggregated viewpoint is useful for deriving general views and tendencies, it is more useful to a coach to have a more fine tuned, deeper analysis for very specific situations. For starters, it is important to have specific analysis for in individual team instead of on a league-wide basis. Teams vary widely in terms of general strategy, roster construction,

and talent level. For example, the Cleveland Browns had the worst record in 2016 while the New England Patriots won the super bowl. As a result of the obvious differences in the teams, they have very different play calling tendencies. Furthermore, within an analysis of each team, it is important to have analysis for various scenarios.

VI. VIEWING DATA FROM A COACH'S PERSPECTIVE

Another major takeaway is that many coaches view situational data differently than academics might. Variables that are seemingly discrete across a contiguous range such as score differential, time remaining within a game, and yardage remaining are not necessary continuous to a coach, but rather are often viewed in categories. When viewed in this context, it alters the way in which plays should be modeled. Rather than using larger datasets of all plays, it makes more sense to view subsets of plays within each category. This will allow the model to train on data in the same way that a coach would normally go about breaking down film. In order to provide meaningful insights to coaches, it is important to provide information and tools that can help coaches within the structure of how they already operate.

One example of how coaches categorize certain aspects of situational data is for yardage remaining. For yardage remaining to gain a first down, coaches tend to view three categories: short, medium, and long. These are defined as 1-3 yards remaining, 4-6 yards remaining, and 7+ yards remaining respectively. While this is by no means a scientific viewpoint, it is a widely held view, and therefore is worth considering in this research.

Another example is field position. Rather than viewing the field as continuous, coaches typically break the field up into zones, such as the goal line, red zone, and middle of the field. As coaches have various different definitions of these zones, I have created five categories that generally capture several zones that most coaches agree upon. The first is the goal line, ten yards or less to the end zone. The second is the scoring zone, from thirty yards to ten yards to go. The middle of the field is between each thirty yard line. The fourth zone is from the offense's ten to thirty yard lines. Finally, the fifth zone is within ten yards of the offense's own goal line.

For time remaining in a game, coaches tend to care the most about the last few minutes in each half. There is little distinction between twenty-eight minutes remaining versus twenty-five minutes remaining. Either amount represents a large amount of time, and there is little way to predict the game flow beyond that point. Coaches begin to take time remaining into account near the four minutes remaining mark. However, while there is certainly some correlation between play calling and time remaining, there is no way to create a rule for how coaches view time remaining.

Similarly, many coaches view score differential in ranges, especially late in a game. The categories respond to common scoring increments. Field goals (worth 3 points) and touchdowns (worth 7-8 points including the extra point) are far and away the most common ways to score. Therefore, when his team is behind, there is no difference to a coach whether his team is losing by one versus two points – any score puts his team in the lead. Likewise, a team being down by four, five, or six points makes little difference to a coach, since a touchdown would take the lead. Furthermore, beyond an eight point score differential begins to be measured in the number of scores it would take

to tie the game. For example, since the maximum points that can be scored in a single possession in eight, a nine-point deficit is a "two score game", whereas a seventeen-point deficit is a "three score game". Typically, this distinction is with respect to the points needed to tie a game.

Looking at specific situations may be a more practical approach that may even yield more accurate results. However, looking at subsets of data presents its own challenges. Using subsets of data requires using smaller data sets. This means that applying a neural network is no longer practical. While it may be possible to use multiple years of data in an aggregate view where league-wide changes in strategy tend not to drastically change, individual teams change fairly drastically each year. Rosters have a significant turnover rate. Additionally, many teams will have a change in coaching staff that will impact strategy. Therefore, it is unreasonable to use data from a previous year to augment datasets. As a result, it is necessary to use a new approach for play type and tendency analysis.

VII. SITUATIONAL VIEW PROCEDURE

In order to handle the smaller datasets that come with looking at specific situations, I tested several algorithms: logistic regression, linear regression, support vector machines, and the forest of random trees. Each of these models accommodates small sample sizes, but can also handle larger datasets as well. I decided to separate the data by team, down, distance, and field position. Each dataset was then partitioned into a training set containing 80% of the data, while the test set contained 20%. The partition was done randomly to make sure that the plays represented the same distributions. Given

the size and number of the situational datasets, I did not validate the models, but rather used the default parameters of the scikit-learn [PED+11] implementations. Optimization of the model through validation can over fit the model to the training set, and thus is inadvisable in this situation. Of the models tested, logistic regression returned the highest average accuracy across all of the datasets. Accuracy is reported as the percentage of plays that were correctly predicted by the model on the test set. The accuracy of the model is compared to the naïve statistical model, where the most common play in that situation is chosen as a prediction.

VIII. SITUATIONAL VIEW RESULTS

This section discusses and analyzes the accuracy and prediction results of the model. We look at the results through the lens of several different ways to compare accuracy. Additionally, we will discuss how the results impact how the model can be used to influence coaching decisions.

A. Tendency vs. Predictability

In order to properly analyze the results from the model, there needs to be a distinction between tendency and the predictability of a certain type of play. For example, the chart below shows the tendencies in down and distance scenarios. For example, on average, an NFL team passes about 51% of the time on 1st and long and runs about 49% of the time. This is a tendency, as it represents what a team tends to do. However, since the two play types are somewhat balanced, it is not very predictable. In contrast, 3rd and medium is very predictable, as NFL teams pass the ball over 90% on

average. Strong tendencies are predictable, whereas situations that have weaker tendencies are not as predictable. Because of this, coaches often trust stronger tendencies that they see from statistics, yet will hold weaker tendencies in less regard.

Down & Distance	Pass Percentage	Run Percentage
1st & Short	29.61%	70.39%
1st & Medium	39.45%	60.55%
1st & Long	50.69%	49.31%
2nd & Short	37.71%	62.29%
2nd & Medium	51.88%	48.12%
2nd & Long	66.23%	33.77%
3rd & Short	57.41%	42.59%
3rd & Medium	90.45%	9.55%
3rd & Long	87.86%	12.14%

Table 3

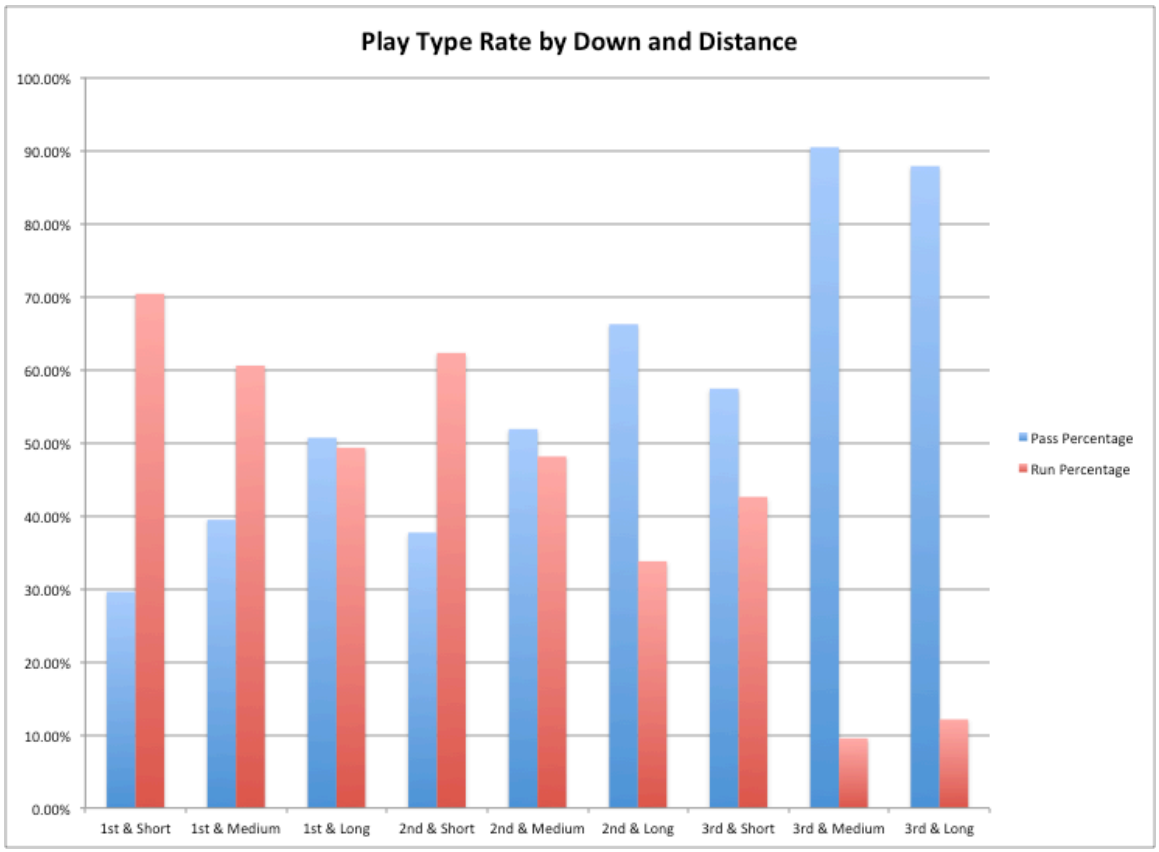

Figure 3

B. Results

There are several ways to view prediction accuracy. I decided to look through each of data aspects that coaches utilize to look at specific situations. Therefore, I broke the data down into groups by down, distance, field position, and down combined with distance. Each accuracy measurement is the average accuracy across all teams.

The first view is how to model compares on a down by down basis. On each of the three downs, the model showed improvement over the naïve statistical method. The model showed significant improvement on first and second down. One takeaway from this is that first and second down are much less predictable than third down, as can be seen in the overall tendencies in Figure 3. The improvement over the basic statistical

model in these situations means that the model is more useful in the situations with less predictability.

Down	Average of Model Accuracy	Average of Naive Accuracy
1	56.85%	51.84%
2	61.17%	57.81%
3	78.75%	78.44%
Grand Total	65.67%	62.86%

Table 4

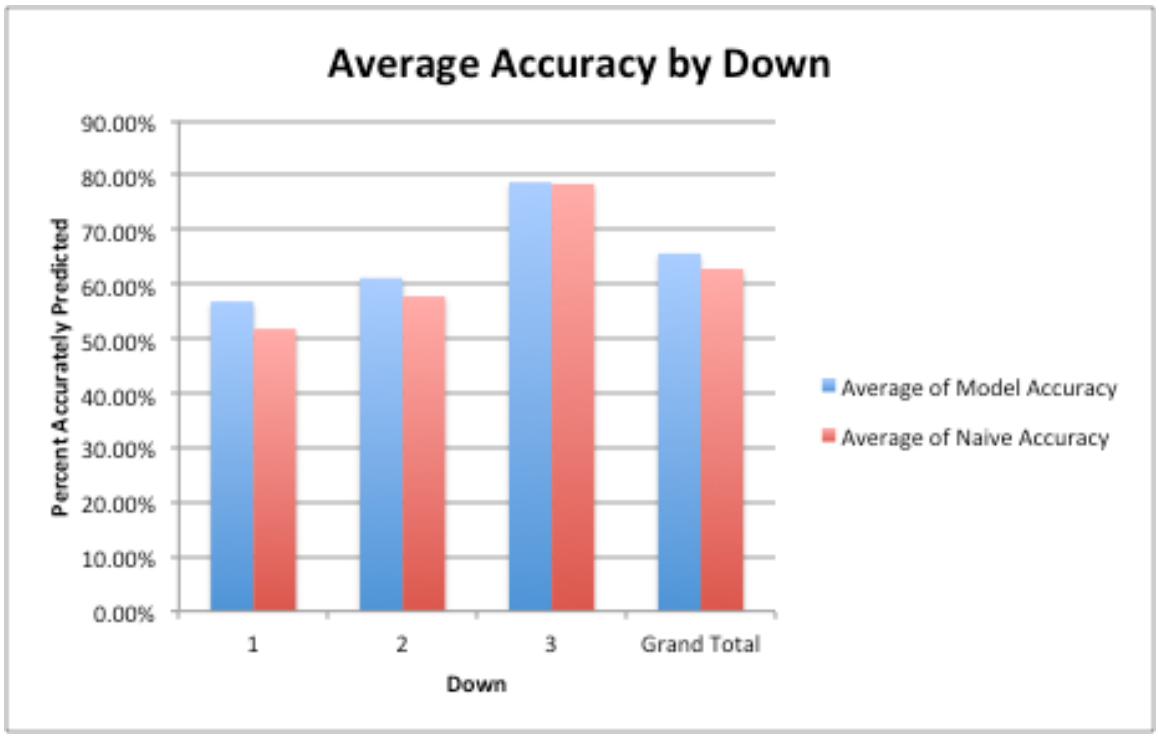

Figure 4

Another way to view model accuracy is by comparing accuracy on the possible distances: short, medium, and long. Short is defined as three or less yards to go, medium is four to six yards to go, and long is seven or more yards. Once again, the model

outperformed the naïve view in all instances. Similar to the down view, the model outperformed the naïve view the most on short distance. Once again, short distances are less predictable than long or medium distances.

Distance	Average of Model Accuracy	Average of Naive Accuracy
Short (1-3 yards)	62.95%	57.96%
Medium (4-6 yards)	63.88%	60.19%
Long (7+ yards)	67.60%	66.07%
Grand Total	65.67%	62.86%

Table 5

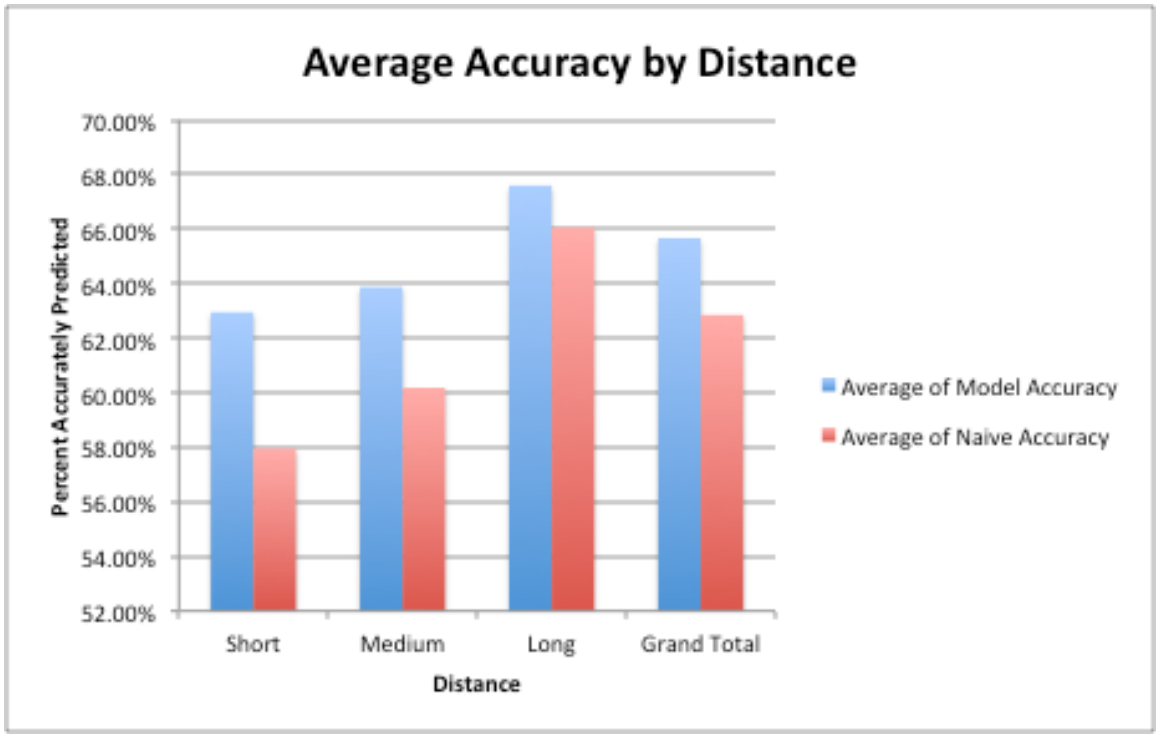

Figure 5

Yet another way to view model prediction accuracy is through field position. Below, the field is segmented into five different areas. [1, 10] represents the goal line

area, where the team is within 10 yards of the goal line. [11, 30] represents the scoring zone area. Traditionally, some coaches think of the 'Red Zone' as within the twenty yard line. However, this has become somewhat outdated in the NFL. Today kicker accuracies are at all time highs. It is reasonable in most weather conditions and game situations that kickers can reliably be called upon to kick a field goal with the ball inside the thirty yard line, which results in about a forty-seven yard field goal. [31, 70] represents the middle of the field. This is the area between the thirties, which is considered fairly neutral. The team is not necessarily close to scoring, yet is not backed up to the point of having poor field position. The [71, 90] range is close to the team's own end zone. This is generally considered weak field position. Finally, [91, 99] represents the team being inside its own 10 yard line. This is very poor field position. The likelihood of scoring from this field position is low, and many coaches choose to try maneuver into better field position or simply give the punter more room to do his job instead of aggressively trying to move down the field.

As shown below, the model outperforms the naïve method in all areas except for one, [11, 30]. One interesting takeaway from the overall accuracy measurements is that both the model and naïve methods are more accurate in the middle of the field as opposed to close to either end zone. This is fairly intuitive, as most coaches conform to general play calling norms in the middle of the field, whereas they may begin to exploit specific game plan notes in more specialized situations. Furthermore, the scoring zone from [11, 30] has some interesting properties. Within this region on the field, a field goal has a reasonable chance of succeeding. Barring some sort of catastrophic outcome, such as a turnover, there is increased incentive to aggressively pursue a first down in order to try to

score a touchdown. Likewise, since a field goal is likely to end the possession as opposed to a punt, there is less incentive for an offensive coach to consider improving field position.

Field Position	Average of Model Accuracy	Average of Naive Accuracy
[1, 10]	58.11%	54.18%
[11, 30]	65.89%	66.79%
[31, 70]	69.76%	66.00%
[71, 90]	65.07%	61.04%
[91, 99]	56.13%	51.43%
Grand Total	65.67%	62.86%

Table 6

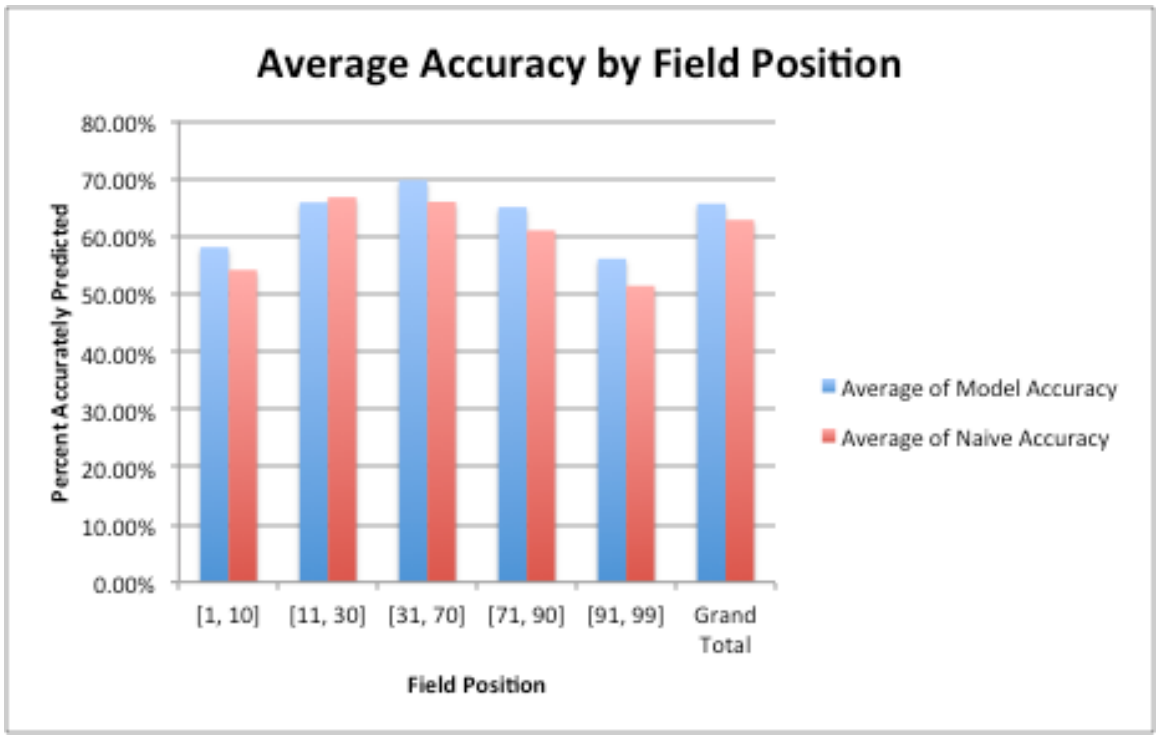

Figure 6

Another important view to take is that of down and distance paired together, which is the most common way for coaches to look at situational data. Perhaps the most interesting aspect illuminated by accuracies grouped by down and distance is the combinations in which the model outperforms the naive statistical approach. The four cases in which the model fails to outperform the naïve approach are also the four cases in which the naïve approach performs with the highest accuracy, which means that there is less variation on these types of plays. In other words, these plays are the most predictable. The four instances in particular are 1st and short, 2nd and long, 3rd and medium, and 3rd and long. From a football standpoint, this makes total sense. 1st and short plays happen almost exclusively near the goal line. This means that the ball is three yards or less away from scoring. Most teams are fairly certain that with three tries, they can run the ball and score without assuming the risk of throwing. The other three cases, 2nd and long, 3rd and medium, and 3rd and long, all fall into the same category – without a significant gain, it is unlikely to pick up a first down. Therefore, teams would be more willing to use a pass play, which is slightly riskier but more likely to pick up big chunks of yardage.

Conversely, the model outperforms the naïve approach when the play type is less predictable. For instance, on 1st and medium and 2nd and short, an offense has lots of flexibility. Therefore, the offense tends to be less predictable. These two instances are the cases when the model outperforms the naïve method the most. This indicates that the model has the most value in areas of the game with the least certainty. As such, the model is most effective in the moments when a coach might be most willing to use it.

Down & Distance	Average of Model Accuracy	Average of Naive Accuracy
1st & Short	56.55%	64.88%
1st & Medium	47.92%	31.25%
1st & Long	57.41%	51.71%
2nd & Short	66.45%	57.96%
2nd & Medium	53.57%	48.80%
2nd & Long	65.07%	65.91%
3rd & Short	61.13%	56.35%
3rd & Medium	86.56%	87.16%
3rd & Long	86.88%	89.21%

Table 7

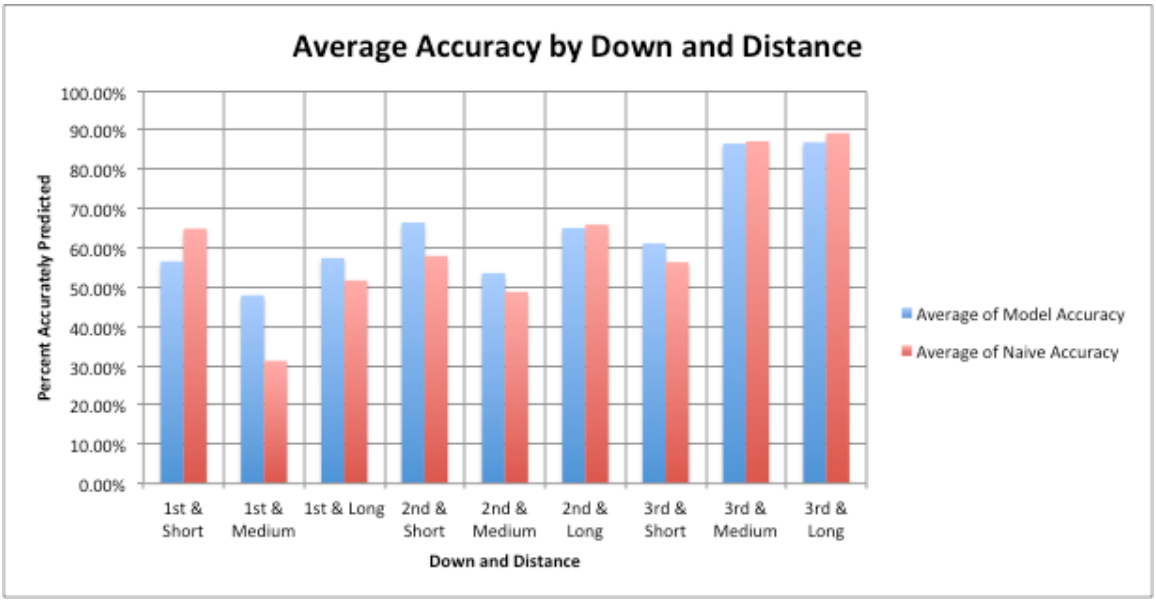

Figure 7

IX. CONCLUSIONS

Prediction of football offensive play calling represents an opportunity to significantly improve defensive strategy. The improvements shown in this paper demonstrate that employing machine learning models can drastically improve prediction

accuracy over the baseline, even when controlling for each down, distance, and field position. Perhaps the most interesting case is the first down case. It is not unusual for coaches to attempt to be especially balanced in offensive play calling on first down, demonstrated by the near 50/50 split in run versus pass on first down. However, this was the case that the model most improved upon. This clearly signifies the importance of other situational game factors upon play calling strategy beyond the most basic down and distance metrics. Additionally, by looking at the predictability across all situations, the results show that machine learning models have a significant advantage in prediction plays with weaker tendencies. Intuitively, this means that machine learning models are better at making predictions in difficult, less obvious situations.

Tendencies are very important to coaches. They paint a picture of how the opposing team behaves in certain situations. They are so vital to coaching decisions that coaches often carry charts with them on the sidelines. These charts may be very basic, covering only down and distance, or they may be more in depth. Given that coaches want to understand tendencies of a certain team in numerous situations, there is a problem that there simply is not a lot of data. By breaking situations down into so many subsets of data, the size of data sets for each situation can become extremely small, even going to zero in some situations. For example, below is a tendency sheet for third down by field position and yardage to first down for the New England Patriots, similar to something a coach might use to game plan and may reference during a game.

Field Position	3rd & Short	3rd & Medium	3rd & Long
10 – Goal	67% Run 33% Pass 12 Plays	43% Run 57% Pass 21 Plays	0% Run 100% Pass 4 Plays
30 – 11	58% Run 42% Pass 26 Plays	32% Run 68% Pass 28 Plays	14% Run 86% Pass 14 Plays
(-30) – 31	41% Run 59% Pass 29 Plays	30% Run 70% Pass 53 Plays	14% Run 86% Pass 58 Plays
(-10) – (-29)	43% Run 57% Pass 7 Plays	30 % Run 70% Pass 20 Plays	15% Run 85% Pass 26 Plays
Goal – 9	N/A - 0 Plays	N/A - 0 Plays	33% Run 67% Pass 3 Plays

Patriots Tendency Sheet - Normal

Notice how there are some situations, such as when the offense is inside its own ten yard line, where there isn't any data at all. In these situations, the model can be used to predict the action of the opposing team by simulating hundreds of plays randomly created within the situational category. Thus, for a sample of 100 predicted plays, it would be possible to create statistics for how often an offense is predicted to run or pass based on the how the model responds. This would allow the coach to have an idea of how the opposing team may perform, even without concrete data. Furthermore, in situations with very little data, simulations could augment the existing data to inform coaches on how the model believes a team will behave. For example, below is a revised tendency sheet where the situations with no data have been modeled and simulated. In

this instance, the model was trained on all patriots third down data, which allows the model to make a decision for the categories of plays within a certain subsection.

Field Position	3rd & Short	3rd & Medium	3rd & Long
10 – Goal	67% Run 33% Pass 12 Plays	43% Run 57% Pass 21 Plays	0% Run 100% Pass 4 Plays
30 – 11	58% Run 42% Pass 26 Plays	32% Run 68% Pass 28 Plays	14% Run 86% Pass 14 Plays
(-30) – 31	41% Run 59% Pass 29 Plays	30% Run 70% Pass 53 Plays	14% Run 86% Pass 58 Plays
(-10) – (-29)	43% Run 57% Pass 7 Plays	30 % Run 70% Pass 20 Plays	15% Run 85% Pass 26 Plays
Goal – 9	65% Run 35% Pass Simulated	45% Run 55% Pass Simulated	33% Run 67% Pass 3 Plays

Patriots Tendency Sheet - Simulated

Not only is the improvement in prediction accuracy an interesting achievement, but a machine learning model is also extremely practical for use in a football game. While models may take a long time to train, predictions can be made very quickly. This means that a model could be trained before a game, and then used to inform coaches of potential opportunities to improve strategy. In an age when every team now has an analytics department to find tendencies and patterns in nearly every aspect of the game, machine learning clearly has a place in the game of football.

X. NEXT STEPS

While the models developed in this paper provide a significant improvement over the baseline in prediction accuracy, there are still opportunities to further increase prediction accuracy. There is a large amount of data that could be added to feature vectors to provide a broader picture of a football situation.

Many coaches use personnel groupings (number of players at each position) of players to predict offensive play calling. For example, if an offense substitutes more wide receivers (generally pass catchers) onto the field, a pass is typically more likely than if the offense substitutes extra linemen (generally used to block for a run play). This is so important that defenses almost always substitute their defenders when they see an offense substitute. While this information is not easily available to the public, it would be readily available to a football organization.

Furthermore, the NFL has begun to track positional data of players on the football field using RFID chips embedded in the players' shoulder pads. This means that there is now an X-Y coordinate for where each player is located on the field at all times. This could be used to encode offensive formations before the snap into features. Again, while this information is not publicly available, NFL teams would have access to it. This could provide yet another route to improve the model for increased accuracy.

Another way to improve prediction accuracy would be to explore the use of alternate encodings of data. For example, this paper has discussed using the normalized form of data across the entire spectrum of data, as well as some basic categories used by

coaches. However, there are near infinite ways to encode the data, some of which may be helpful. For example, while it was mentioned that time remaining was important to coaches in some capacity and score differential was important with varying degrees based on the time remaining in the game, it proved difficult to model in a way that actually represents how coaches approach these variables. If encoded properly, these have the opportunity to lift accuracy even further.

Additionally, coaches mentioned that the specificity of game planning is not limited to situations. They also focus heavily on players. While it would be extremely difficult to incorporate which specific players were on the field at all times, it might be reasonable to model based on quarterback. Quarterback is the one position that almost never changes within a game, barring injury. However, there are a few instances that come to mind that may be enlightening for coaches. For example, during the 2016 season, Tom Brady did not play in the first four games due to suspension. During the four games that he did not play, the Patriots certainly called plays differently than the remaining twelve games that he played in. Similarly, some teams, such as the Browns, started several different quarterbacks with vastly different skill sets. Modeling by quarterback represents an opportunity to improve prediction accuracy while also providing a way to provide situation specific insights to coaches.

REFERENCES

[PED+11] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[FOR] S Forman. *Pro Football Reference*. http://www.pro-football-reference.com/. Accessed: 2016-12-13.