Evaluating Reinforcement Learning as a Technique for Optimizing and Automating NFL Play Calling

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Citable link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364636

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Evaluating Reinforcement Learning as a Technique for Optimizing and Automating NFL Play Calling

A THESIS PRESENTED
BY
MANAV J. KHANDELWAL
TO
THE DEPARTMENTS OF STATISTICS AND COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECTS OF
STATISTICS AND COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2019
Evaluating Reinforcement Learning as a Technique for Optimizing and Automating NFL Play Calling

Abstract

Offensive coaches in the NFL are tasked with one of the sport’s most difficult jobs: making successful play calling decisions in real time. While play calling requires acute tactical nous and a thorough understanding of the game, some of the league’s brightest minds have turned to analytics in order to improve. We propose the use of reinforcement learning methods, namely off-policy approaches such as Q-learning, to augment and ultimately automate certain aspects of play calling. We design a Markov Decision Process to simulate offensive drives by sampling existing plays from the 2017 season and train optimal policies for deciding between running and passing. Among our findings is that, across the board, teams should be more aggressive than they currently are by adopting pass-heavy strategies. In addition, we find that considering the time remaining and score is important in order to achieve better offensive outcomes. Ultimately, we are confident that improved reinforcement learning methods can serve as a platform for generating better play calls on average and the approach’s flexibility can be tailored to fit the needs of specific personnel and coaching philosophies.
Contents

1 INTRODUCTION 1
   1.1 Motivation ................................................. 1
   1.2 Literature Review ............................................. 8
       1.2.1 Observational Reinforcement Learning ................. 8
       1.2.2 Play Calling Prediction ................................. 9
       1.2.3 play calling Optimization .............................. 10
       1.2.4 Game Theoretic Models ................................. 11
       1.2.5 Generalized Matching Law ............................. 16
       Reinforcement Learning ................................... 18

2 METHODS 21
   2.1 Markov Decision Processes .................................. 21
   2.2 Q-Learning .................................................. 24
   2.3 Implementation ............................................. 26
       2.3.1 Baseline Model ........................................... 26
       2.3.2 Models of Higher Complexity ......................... 30

3 DATA ANALYSIS 32
   3.1 Data Sets ..................................................... 32
   3.2 Data Cleaning Process ....................................... 34
   3.3 Data Summary ............................................... 35
   3.4 Exploratory Modeling ........................................ 38
       3.4.1 Considering Time and Score Effects on Play Prediction .. 38
       3.4.2 Serial Correlation ..................................... 40

4 RESULTS 42
   4.1 Baseline Model ............................................. 42
   4.2 Time and Score Model ...................................... 46
   4.3 Serial Correlation Effects Model .......................... 49
   4.4 Model Comparison ......................................... 52
Acknowledgments

I would like to extend my most sincere thanks to my advisor, Kevin Rader, who has been an invaluable resource and mentor for me throughout not only the thesis-writing process but my academic career at Harvard. His passion for sports analytics has been an important motivating factor as I have tackled a difficult applied problem and it has been wonderful to be able to connect on so many levels with my advisor.

The Philadelphia Eagles, and in particular Alec Halaby, took a chance on me by giving me the opportunity to intern in the Football Analytics department this past summer. Without that experience and Alec’s guiding hand, it is unlikely that I would have developed the requisite skills or knowledge of the domain area to write this thesis.

Thank you to Pro Football Focus (PFF) and Carnegie Mellon for the play-by-play data that I used. In particular, Ryan Smith and Bryan Hall at PFF helped me get set up.

Most importantly, I would like to thank my family and friends for their unwavering support. Without my parents, Manoj and Meena Khandelwal, and my sisters, Medha Agarwal and Mallika Khandelwal, I would not have had the opportunity to attend Harvard and pursue my interests.
Make no mistake, [advanced analytics] is going to be a separator in terms of your competitiveness, both in personnel and on Sundays.

NFL Executive (anonymous)

Introduction

1.1 Motivation

Football games may be won and lost on the field, but it is often split-second decisions made on the sideline or in a coach’s box that influence the outcome of a particular game. Offensive play calling, the choice to run a specific play with a specific set of personnel on the field, is a particularly important subset of those decisions. Many teams in the National Football
League (NFL) try to differentiate themselves from the competition and stay ahead of their opponents by hiring the most competent and innovative play callers as head coaches or offensive coordinators.

![Image of Doug Pederson and Andy Reid](image.png)

**Figure 1.1:** Doug Pederson (right) and Andy Reid (left), two of the most innovative offensive strategists in the NFL, have embraced analytics.

The level of complexity involved in the creation and execution of a call sheet, the offensive bible on game days, is fascinating; a great play caller is meticulous in his or her preparation. Plays are based off of schemes, which are general philosophies about how to design a successful offense. Schemes range from the “spread” offense—a hybrid system reliant on a mobile quarterback—to something like the “West Coast” offense—a pass-first system characterized by short, horizontal routes—to anything in-between. Play callers draw up plays
that fit their scheme but also take advantage of their personnel and specific match ups they can exploit against opposing defenses.

All in all, there are a number of components of a game plan that go into how a team calls plays. The first is the team’s own strengths and weaknesses; this involves the overall quality and depth at different positions, the specific positional attributes that the team’s personnel possess (e.g. mobility for a QB), and the players’ familiarity with certain schemes. The second is an in-depth understanding of the opposing team’s strengths, weaknesses, and tendencies, which may include how good they are at defending certain types of runs and passes, how often they line up in certain formations—and then what defensive schemes they execute out of those formations—and the craftiness of the defensive coaching staff. The third is the situation in which the play is being called; the quarter, down, distance, field position, and score, among other factors, can significantly influence an offense’s decision.

The NFL is inhabited by a number of coaches that are considered “geniuses” because of their seemingly unique and effective play calling abilities. Bill Belichick was long considered an elite thinker among NFL coaches, scheming up defensive game plans on his way to a number of Super Bowl rings. However, a new generation of offensive tacticians has taken the league by storm, from Doug Pederson, Super Bowl-winning head coach of the 2018 Philadelphia Eagles, to Sean McVay, the mastermind behind the Los Angeles Rams’ recent resurgence. But what makes these people truly special? And how can an in-depth evalua-
tion of play calling analytics, and in particular reinforcement learning, help propel these kind of strategists even further forward?

At its core, play calling is a hybrid of planning and adaptation. The best coaches use scouting and data to identify league-wide and opponent-specific defensive trends, and then craft plays to take advantage of certain weaknesses. This results in the development of a set of rules that influence how certain offensive minds approach play calling in specific situations and against particular defensive alignments. However, this can change rapidly in the middle of the game, and thus pre-planning is rarely sufficient in winning games consistently. The best play callers can take advantage of what they see early in a game to craft specific strategies to beat a defense later on.

The most forward-thinking offensive minds and organizations seem to understand that analytics can help them immensely in this regard. Teams in the NFL have recognized the importance of embracing data in building rosters, evaluating talent, signing contracts, and planning in-game strategies. Ultimately, these teams are looking to become more data-driven in the way they approach game planning and play calling especially. By harnessing the power of advanced analytics, made possible by seismic forward leaps in data availability and computing power, these coaches are showing that play calling is equal parts art and science. As information becomes available to team analysts and third-party data scientists in greater quantity and quality, the possibilities for turning aspects of the highly complex,
inter-connected game of American football into a solvable statistical problem continue to grow.

Offenses have already begun to evolve rapidly thanks to a combination of data-driven aggressiveness and rule changes, with teams scoring at historic clips during the 2018 NFL season.\textsuperscript{1} Initial work in football analytics has shown that teams have historically passed less than they should, and over the past two decades the trend towards becoming a passing league has continued and even accelerated, as seen in Figure 1.2. Overall, running has become more of a way to set up a more successful passing game rather than a team’s go-to system.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{rushing_passing_attempts.png}
\caption{A league-wide comparison of rushing and passing as a percentage of team’s yards over time.\textsuperscript{2}}
\end{figure}
I was able to witness the transformative impact of data science on football thinking first-hand. While working for the Analytics department of the Philadelphia Eagles, I found that analytics can directly impact high-level decision-making in a number of ways; coaches and player personnel experts are beginning to trust the insights that models give them and are now using data in every part of their job. These range from identifying simple defensive alignment trends to replacing or augmenting different ‘creative’ aspects of coaching. It is the latter scope that I hope to tackle in this paper: can we use reinforcement learning methods to learn more about situational play calling, infer strategies, and ultimately automate some of the preparation and decision-making?

There is historical precedence for analytics influencing offensive decision-making and game-planning. Teams have responded to early work in the field around risk-reward analysis by being more aggressive on first down, passing the ball more often and continuing to do this even as they approach the red zone. Teams have also used studies of the efficacy of third-down play calling to reduce their reliance on running in 3rd-and-medium situations. The analytics department of the aforementioned Eagles became famous during their Super Bowl run for fueling head coach Doug Pederson’s aggressive nature; he merely used simple analytics to pass more often on third down and go for it on fourth down and two-point conversions, but was hailed as a pioneer. As Eagles owner Jeffrey Lurie put it, “But what we found is, there’s been so many decisions over time that are too conservative for the odds of maxi-
mizing your chance to win at the opportunity. I mean, you’ve seen certain coaches that are deemed more aggressive because the math leads them there. That’s all it is.”

Additionally, there is certainly a desire within the league for advanced analytics to take an even greater role in play calling. As in 1996, when IBM Watson beat its first grand chessmaster, the time will likely come when a play calling algorithm, trained using a combination of simulation and empirical results, can out-coach the NFL’s best. Now, football may often likened to a chess game, but it is significantly more complex, and more importantly, stochastic—at least from the perspective of the play caller. The complexity means that decision-making algorithms cannot operate in real-time yet, at least given current computation and data limitations. It is expected, however, that those concerns will be mitigated as more money is dumped into the space.

That said, reinforcement learning and other similar methods can only go so far in helping play callers with their job. play callers rely on their personnel for execution. Unlike chess, which is highly deterministic, the outcome of a football play depends on many post-snap factors. Additionally, there is still the “art” of play calling, the creation and design of new plays, that is incredibly difficult to automate. So what can play callers hope to gain from using decision-making algorithms? Coaches can help establish parameters around when it is most profitable to make certain decisions, understand the factors that statistically influence these decisions, and identify the trends that should guide their thinking. As such, this paper
will examine to what extent reinforcement learning can improve on current advanced play calling analytics and then identify its uses and shortcomings.

1.2 Literature Review

1.2.1 Observational Reinforcement Learning

Generally, the setting of reinforcement learning assumes a world with infinite iterations and the ability to simulate any state-action pair \((s, a)\). This provides the opportunity to test counter-factuals—what would have been the result if I had taken a different action—and ultimately, in the long-run, estimate the true optimal policy.

In this analysis, of course, we are relying on existing football data. These are plays and outcomes that have been observed, not simulated for our benefit. This creates a number of challenges, including a lack of counter-factuals, a selection bias towards decisions that were made using policies that we do not know, and confounding factors in the success of a play.

The most important issue we run into using observational data is that we cannot obtain information about the efficacy of alternative play calls; if the team passed in a given situation, we cannot determine exactly what the effect of a run would have been. Thus, we are forced to rely on using the data in aggregate to fill in holes and infer success. This makes it quite difficult to estimate the true optimal policy.

There has been some theoretical work around the statistical properties of observational
reinforcement learning algorithms. We in particular are focused on policy optimization rather than outcome estimation, and there is some literature that explores estimation while acknowledging that actions $A$ are not independent from the feature set $X$ (which can refer to determinants of the state space). In this setting, the algorithm calculates $m$, the expected outcome of action $a$ given feature $x$, and hopes to find a policy $\pi$ that maximizes the policy outcomes (where $D$ is the marginal distribution of features): $V(\pi) = E_{x \sim D}[m_h(x)]$.

This study gives us theoretical estimation bounds between the estimated and actual policy outcomes.

1.2.2 Play Calling Prediction

Research focused on the area of predicting play calls is not widespread. Michael Dickey and William Burton of North Carolina State University built the first machine learning model to predict run or pass plays based on NFL data. Using data from 2000 to 2014, Dickey and Burton used logistic regression to determine significant predictors; the pair ended up including down, time remaining, score difference, points by both teams, cumulative fumbles and interceptions, an interaction between down and distance, timeouts remaining, and yards gained on previous play. They determined that variables like previous play call and weather were not significant. The two, however, tested their model on training data which led some to question the model’s efficacy.
Building off of the above study, a Stanford research project applied more advanced machine learning tools to classify runs and passes. The response variable studied is particularly relevant to the analysis done here. The authors used four different methods: logistic regression, linear discriminant analysis (LDA), random forest, and a gradient boosting machine (GBM). They considered a number of features, ranging from situation (down, distance, time, field position, score) to personnel (offensive, defensive, formation) to track record (past tendencies, past success, Madden ratings). The GBM achieved .757 test accuracy with an AUC of .84. The models tended to predict much better on third down and in the fourth quarter, suggesting more offensive randomness in earlier downs and quarters. This paper, however, did not go as far as identifying key predictors in play calling and how to build intuition around prediction or evaluation. Unfortunately, the use of machine learning methods eliminated opportunities for inference.

1.2.3 play calling Optimization

As mentioned in the introduction to this thesis, play calling is a subtle art combined with a dose of science. The literature on optimal play calling is primarily focused on identifying specific aspects of decision-making (“go for it” or not, personnel, run or pass, etc) and describing intuitive models for training an optimal strategy. This thesis, similarly, uses reinforcement learning (and broadly, a Markov Decision Process) to describe the environment in which the game is being played and is primarily focused on the decision to run or pass.
In anticipation of that, this section of the literature review will focus on the different methods used in the past to evaluate and optimize play calling strategies.

**Game Theoretic Models**

Early study in the area of play calling involved a significant amount of work using game theoretic models, often augmented by additional parameterization. A 1999 paper by Boronico and Newbert established the foundation for modeling football decision-making by using a two-player game in which the defense acts as an adversary to the offense. Using a dynamic programming model with a stochastic element (assuming certain probabilities of success), the authors maximized the probability of scoring a touchdown over a sequence of downs starting from 1st and Goal. They assumed both players followed the minimax criterion and thus achieve a Nash Equilibrium based on a mixed strategy of running and passing. In addition to showing obvious results that being closer to the end zone (and being earlier in the down sequence) increased the probability of scoring a touchdown, the paper led to the first creation of an “optimal” run-pass strategy by assigning a probability weight for each situation. The writers were able to show, albeit using touchdown probabilities developed by a college football coach, that an optimal strategy would lead to scoring a touchdown twice as often as a randomized strategy. This paper, despite its significant limitations and simplicity, set the stage for future analysis of optimal strategies by modeling a play caller’s decision set and describing how coaches can use the results in preparation of strategies and evaluation.
of opposing tendencies.

Nearly ten years later, a team made up of Air Force and University of Alabama operations researchers and statisticians hoped to replace subjective decision-making with optimized, quantitative strategies that used updated information as it became available. Using a two-player, zero-sum game theoretic model, the paper considers possible actions, opposing defensive alignments, and past results in such situations. Then, the model estimates a risk level for each coach given field position, down, and distance. Using past data about success of certain plays against certain defenses (and the likelihood of the defense to adopt a certain strategy) along with a play caller’s risk tolerance, a game is simulated to optimize the offense’s strategy. As seen in Figure 1.3 and Figure 1.4, the paper shows that teams should trend towards higher risk (a.k.a trading variance for expected value) strategies earlier in the game, with less information about the defense, than later in the game once they have more information about the defense. The authors believed that this model for analyzing play calling strategy should be the basis for future research, but besides the use of a two-player game, this has not been the case.7

At the same time, researchers from the National Bureau of Economic Research built off of this initial game theory paper with further exploration of the run-pass strategy employed by NFL play callers. In particular, they were focused on whether they truly followed Nash Equilibrium strategies, or the minimax criterion. Using data from the 2001-2005 NFL sea-
Figure 1.3: Initial $\rho$ interaction plot between players 1 and 2.

Figure 1.4: $\rho$ interaction plot between players 1 and 2 after one iteration.
sons, Kovash and Levitt used a two-player, zero-sum game with an early version of expected points as the response variable. Expected Points refers to the “net value a team can expect given a particular combination of down, distance, and field position among other factors,” which often includes information about the current drive and the next scoring event. This paper uses a crude version of expected points, which is modeled using just the three features above, along with yards gained as the response variables/rewards. Based on regressing play call along with other factors against the response variables, the writers showed that passing consistently outperformed running (implying that teams were under-utilizing the pass), which can be seen in Figure 1.5 below. The analysis went on to showed negative serial correlation in play calling; teams are more likely to run after passing and vice-versa, so it may be difficult to assume independence of decision in a sequence of play calls. Thus, the authors of the paper were able to show that NFL play callers, at least in the years studied, did not seem to follow optimal game theoretic strategies defined by minimax behavior.\(^6\)

Within a year, a researcher from the University of Toronto responded directly to the claims made in the above study with his own unique model of the two-player game for NFL play calling. He builds an adversarial game in which the offense gains zero reward if both offense and defense pick run or pass (“match pennies”) but gains a unique positive reward otherwise. The author of this paper includes two modifications: he includes 1) team-specific effects and then 2) the value of “running to set up the pass”/“running to wear down the
Figure 1.5: From the Kovash and Levitt study: a sensitivity analysis (in different situations) of the gap between running and passes with different response variables.

defense” (running increases future rewards). This paper begins to explain some of the “sub-optimal” decision-making that the above paper decried by including conventional football wisdom into the theoretical; the author shows that unsuccessful plays will lead to more randomized strategies but the model predicts positive serial correlation overall, which is empirically confirmed.9

A similar study in 2010 by McGarrity and Linnen used the “matching pennies” framework for a two-player, zero-sum game to determine the effect of a backup quarterback on the play calling decisions of an offense. Figure 1.6 shows the game theoretic representation of this situation. This paper showed a theoretical result that supported the analogous empirical result (which existed in the data): even though a backup quarterback results in lower
passing productivity, the defense responds by defending the run and therefore the offense does not change its overall strategy. This is based on solving for a mixed strategy equilibrium as before. However, the paper challenged an earlier result by positing that serial correlation (between passing and running in the past and present) does not hold empirically. Using a Runs Test, which looks for positive correlation (too many Runs) or negative correlation (not enough Runs) where a Run refers to a string of consecutive identical play calls, the authors determined that they could not reject the null hypothesis that play calls are random; their explanation is that play callers are removed from the heat and emotion of the game and thus are better able to game plan (unlike tennis servers, where serial correlation was shown).  

<table>
<thead>
<tr>
<th>Offensive Play</th>
<th>Defend Pass (y)</th>
<th>Defend Run (l − y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass (a)</td>
<td>(A − X_1), (−A + X_1)</td>
<td>(B − X_2), (−B + X_2)</td>
</tr>
<tr>
<td>Run (1 − a)</td>
<td>C, −C</td>
<td>D, −D</td>
</tr>
</tbody>
</table>

Payoffs: offense, defense.

**Figure 1.6:** The game theoretical model from McGarrity and Linnen that includes X_1 and X_2, the negative effects of a backup QB.

**Generalized Matching Law**

The Generalized Matching Law (GML) states that decision-makers tend to allocate between possible options proportionally to the reinforcement (reward) schedule. Where B_1 and B_2 are possible actions and R_1 and R_2 refer to respective reinforcements, the original
law states:

\[ \frac{B_1}{B_1 + B_2} = \frac{R_1}{R_1 + R_2} \]

The Generalized Matching Equation (GME) was introduced in the 1970s to account for slight deviations from exact matching, which accounted for nuance in decision-making processes. The equation is as follows (with \(a\) accounting for sensitivity in reinforcement differentials and \(\log b\) corresponding to bias or preference for one action over the other):

\[ \log \left( \frac{B_1}{B_1 + B_2} \right) = a \left( \frac{R_1}{R_1 + R_2} \right) + \log b \]

A 2006 study by Reed, Critchfield, and Martens wanted to examine if run and pass likelihoods were related to the results from those plays. This paper used yards gained as the reinforcement and run versus pass as the behavior set. Their study found that yardage gained influenced play calls, accounting for approximately three-quarters of the variance in play calls. This confirmed that as long as both types of plays produce positive yardage, it is theoretically optimal to “mix” strategy. There was also evidence for under-matching \((a < 1)\) as well as a bias for rushing (meaning teams ran more than expected based on reinforcement), but there was little evidence that teams operated on a “constant-yardage” assumption of running versus passing (constant reinforcement values). The study also showed that under-
matching was increasing year-over-year (smaller “coefficient” of the reinforcement differentials) while the bias for rushing was diminishing. However, it was difficult to nail down how teams approached this decision-making process with new information gathered over the course of the season.  

Western Michigan University researcher Jacob Bradley returned to this idea of the GME in 2018 in order to explore the efficacy of this model in the football context with analysis of new reinforcement measures. He considered Success Rate, which measures offensive efficiency by determining whether a certain play was “successful” or not depending on the down and distance. When modeling play calling using the GME with Success Rate as the reinforcement measure, a higher amount of variance in play calling was explained by the model, there was a bias for passing which coincides with league-wide beliefs, and there were significant differences in matching slopes (with Success Rate implying a closer to 1). This ultimately suggests that there is merit in pairing advanced analytical tools with strategic decision-making methods to optimize analysis.  

Reinforcement Learning

There is a dearth of reinforcement learning literature applied to a football context given the difficulty of using empirical data (rather than simulation) to train an agent. In addition, to fully capture the game’s nuances a large state space is required (which leads to significant
“over-fitting”) along with an adversarial setup that cannot be as easily modeled as in two-player games.

The main academic paper related to the use of reinforcement learning in football hoped to challenge the main problem identified above: the high dimensionality of the state space required to properly optimize a policy with limited bias. Molineaux, et. al from the Naval Research Labratory explored the use of plan recognition (identifying opposing strategy) to bolster case-based reinforcement learning techniques in football decision-making. Case-based reinforcement learning is a heuristic that relies heavily on domain-specific knowledge to derive a policy for a specific state $s$ based on knowledge gained about similar states $s', s'', \cdots \in S$. This paper used a physics-driven simulation of a toy football game as its environment, and hoped to optimize the decision of a quarterback during the play. It used clustering techniques to estimate defensive strategies based on post-snap player movement and this had a significant positive effect on offensive outcomes. However, there is little academic work in this area specific to pre-snap play calling.$^{13}$

Michael Alcorn, a professional machine learning engineer for Red Hat, published a piece online that helped provide some of the context for this thesis’ approach to setting up a play calling environment. In his post “Coaching Football with AI,” Alcorn creates a large state space (which includes players themselves) and uses empirical plays from the previous season to train the Q-table. This is a similar approach to what I describe below, but with a
larger state space. He then proposed taking the defensive approach into account by estimating yardage gained for certain types of plays against certain defenses with a formation-and team-specific regression model. He does not, however, provide statistical or intuitive justification for this extension of his reinforcement learning model.
If we all listened to the professor, we may be all looking for professor jobs.

Coach Bill Cowher (2008)

2

Methods

2.1 Markov Decision Processes

Reinforcement learning encompasses goal-oriented algorithms that learn optimal policies over many iterations in order to maximize rewards attained. These algorithms rely on the ability to make decisions about actions to take in certain situations. The basis for these types of decision-making problems is a Markov Decision Process (MDP). A Markov pro-
cess is a randomized process in which the future is independent of the past, conditional on the present. Mathematically, this implies that a state \( S_t \) is Markov if and only if:

\[
P(s_{t+1}|s_t) = P(s_{t+1}|s_1, \ldots, s_t)
\]

This is an important simplification that is often approximately true (or at least experimentally close over many iterations) and allows for important mathematical properties to emerge.

A MDP is the formal description of a fully observable environment for reinforcement learning. In an MDP, we have the following components: states \( s \in S \), actions \( a \in A \), a policy \( \pi(s) \), a reward function \( r(s, a) \), and a value function \( V(s) \).\(^1\!\) Here’s what they mean:

- A state \( s \) is information about the current environment (e.g. the game situation in football, like down, distance, etc). The set \( S \) is the set of possible states in the “world” that the model is imitating.

- An action \( a \) is a decision you can make given a certain state (e.g. run or pass). \( A \) is the set of all possible decisions that can be taken in a given state.

- A reward function \( r : s \times a \rightarrow \mathbb{R} \) is the immediate reward from taking a certain action in a certain state (e.g. gaining six yards or scoring a touchdown). \( r \)'s representation is quite subjective and requires domain expertise, and could be deterministic or random.

- A policy \( \pi : s \rightarrow a \) is a description of the action you would take for all states \( s \in S \) (e.g. run on 1st down, pass on 2nd/3rd down). The optimal policy \( \pi^* \) is the solution to a reinforcement learning problem.
• A value function \( V : s \rightarrow \mathbb{R} \) is the ‘long-term’ (in some cases, over an infinite time horizon) value of a particular state. This is what most reinforcing algorithms look to estimate.

![Diagram of Environment, State, Action, Reward, Agent](image)

**Figure 2.1:** A visualization of the iterative nature of MDPs.

To learn the optimal policy, and thus reach the goal of reinforcement learning, algorithms look to estimate the value function. This can be seen as the expected value of following policy \( \pi \) from the position of being in state \( s \). Here, we have \( R_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \) as the discounted sum of present and future rewards (given discount factor \( \gamma \), which is close to 0 when we are oriented towards the short term and close to 1 when we value future rewards significantly).

\[
V_\pi(s) = \mathbb{E}_\pi[R_t | s]
\]

The Bellman equations famously derived a more intuitive expression for the value function which lets us express the value of a particular state in terms of other states explicitly (where
\( \gamma \) is the discount factor and \( p \) is the transition probability):

\[
V^\pi(s) = \sum_a \pi(s, a) \left( \sum_{s'} p(s'|s, a)[r(s'|s, a) + \gamma V^\pi(s')] \right)
\]

2.2 \textit{Q-Learning}

Because we are working with observational football data, and do not have the ability to
test counterfactuals or simulate plays, we use a method of reinforcement learning called \textit{Q}-
learning. It is considered \textit{off-policy}, because we estimate future rewards by maximizing over
possible options rather than using the given policy that is being followed. We use off-policy
learning in this context because, in fact, we do not know the policy that was being used to
create the data (only the coaches themselves do). Therefore, we cannot assume a particular
policy when estimating future rewards or the value of particular states.

The actual implementation of a \textit{Q}-learning algorithm requires a \textit{Q}-table, a table of values
for each state-action pair \((s, a)\) which approximates the value function \(V\). As Figure 2.2
shows, \textit{Q}-learning algorithm then requires you to initialize values in your \textit{Q}-table (tradi-
tionally, to 0) and then to repeatedly choose an action, evaluate the action, measure the
reward, and update your \textit{Q}-table. We use the Bellman equations for the update step:

\[
Q(s, a) = Q(s, a) + \alpha \left( r + \gamma \cdot \max_{a'} Q(s', a') - Q(s, a) \right)
\]
where $\alpha$ is the learning rate (helps determine rate of converge) and $\gamma$ is the discount rate. To actually fill in our $Q$-table, since everything is originally 0, we use an $\epsilon$-greedy strategy, where we choose the optimal action $a = \pi^*(s)$ with probability $1 - \epsilon$ and a random action $a$ otherwise. We will have $\epsilon$ decrease over time to encourage more exploitation and less exploration.

There is a well-known theoretical result that if each action $a \in A$ is executed in each state $s \in S$ infinite times with an appropriately shrinking learning rate $\alpha$, $Q$ will converge to $Q^*$ such that the true value function $V^*(s) = \max_a Q^*(s, a)$. More specifically, you must have $\sum \alpha_t = \infty$ but $\sum \alpha_t^2$ is finite. Generally, the experimental convergence depends on a combination of the learning rate and the discount rate $\gamma$ (for example, a polynomial...
learning rate means the convergence rate will be polynomial is \( \frac{1}{1 - \gamma} \).\(^{15}\)

2.3 Implementation

2.3.1 Baseline Model

For the remainder of this paper, Model 1 will refer to the baseline model described below. The environment can be intuitively described as the simulation of many offensive drives. This approach does not acknowledge specific teams or specific games in the sense of quarters and time, and it therefore applies to a generic NFL offense on a play-by-play, drive-by-drive basis. Each “iteration” of the algorithm maps to a new drive, and the agent (play-caller) will receive a reward based on the result of the drive rather than each play individually. The agent will optimize decisions based on the approximation of a Q-table that is updated with each play but only receives a present reward at the end of the drive. Therefore, the training of the state space will occur from a sort of “back-propagation” from the agent desiring to get to certain valuable members of the state space.

In Model 1, the MDP/Q-learning problem can be described with the following parameterization (in relation to football):

- \( S \): down \times \text{distance to go} \times \text{field position}. We have that \(|S| = 90\) because we only consider three downs, three levels of distance to go \((\leq 3 \text{ yards}, 4-6 \text{ yards}, \geq 7 \text{ yards})\), and ten levels of field position \((\text{every} 10 \text{ yards from} 0 \text{ yards to go, the defense’s goal line, to} 100 \text{ yards to go, the offense’s goal line})\). Therefore a state \( s \) is made up of a down, distance level, and field position level.
\( A = \{ \text{Run, Pass} \} \). Very simply, we limit \( a \) to be either run or pass.

\( r : s \times a \rightarrow \mathbb{R} \) is stochastic and depends on the empirical results from the 2017 season. In this model, rewards are as follows for certain results (all others have \( r = 0 \)):

- Touchdown: 7
- Interception: -3
- Fumble: -3
- Made Field Goal: 3
- Missed Field Goal: -1
- First Down Achieved: 0.5

\( \pi : s \rightarrow a \) is based on the Q-table described above. The exact policy followed in this algorithm will be described below, but involves a combination of exploration and exploitation.

\( \epsilon \), the exploration hyperparameter, begins with a value of 0.5 and decays exponentially over time (more below).

\( \gamma \), the discounting hyperparameter, is 0.95.

\( \alpha \), the learning rate, is \( \frac{1}{\sqrt{\text{counts}}} \) where \( \text{counts} \) refers to the number of observations that have been seen in that state-action pair \( (s, a) \), inclusive.

The state space is chosen based on previous literature (not just reinforcement learning) and based on knowledge of the state space. In my conversations with members of the offensive staff and analytics team with the Eagles, along with other people who research football, teams generally see field position in ten-yard increments (occasionally 20) and
down/distance based on down number $\times$ long/medium/short. The exact yards were chosen based on my knowledge of football and what constitutes those factor levels. Additional factors can and will be considered, but the state space is limited to size 90 to ensure sufficient density in each element for the baseline model. The action space is similarly limited to run/pass in order to build off of previous study of this area and to reduce dimensionality of the state space. Finally, the reward function was created to model the discrete nature of football scoring (for touchdowns and field goals). For other plays, it was based on expected points/expected points added analysis of turnovers and first downs.

The environment (or setup) of the problem is as follows. For 5000 iterations (each representing a unique drive), we perform the following steps at the start of each drive:

- Choose the starting field position (yards from the end zone) by randomly sampling from a scaled $\beta$ distribution, specifically $\text{field}_\text{position} \sim 100 \cdot \text{Beta}(7, 3)$. The starting field position will average around 70 yards away with a left skew (to account for short punts, turnovers, and longer kick-off returns). Figure 2.3 shows a comparison of this sampling distribution with empirical data.

- Set the starting down and distance to 1st and 10.

Once the drive has been initialized as such, we performing the following steps until the drive ends (through a touchdown, field goal, turnover, or punt):

1. Pick run or pass (an action $a \in A$) using the $\epsilon$-greedy method, which prescribe a random action with probability $p = \frac{\epsilon}{\text{epoch}}$ and with probability $1 - p$ you choose the action that maximizes the value of the $Q$-table for the given state $s$. 

28
Figure 2.3: Visual comparison of empirical starting field positions from 2017 NFL season with the scaled $100 \cdot \beta(7, 3)$ distribution I sampled from in the reinforcement learning environment.

2. Once an action is chosen, we have a state-action pair $(s, a)$. We sample a play from our data that follows the prescribed action $a$ and exists in the state $s$. If the action has never been taken in that state, we choose the other action; if the state has never been explored in real life, we end the drive with no effect on the trained model.

3. The result of the play is then assessed and our $Q$-table and environment are updated accordingly.

- If a drive-ending play occurs, then we set a variable `new_drive = True` and assign the correct value to the variable `reward`, which represents the output of the reward function $r$. This information is then used to update the value $Q(s, a)$ and then we move to a new epoch/drive. If it was a failed third down, then we check to see if the sampled play was followed by a field goal attempt or not and assign a reward based on the result.

- If it is not a drive-ending play, then the yardage gained or lost is taken into account and the down is incremented by one. Thus the new state $s'$ has $dun^t =$
dwn + 1 and \( ydstogo' = ydstogo - yds_{gained} \). This is, of course, unless a first down is achieved in which case the field position is re-set to the previous yards to go minus the yards gained and the down/distance are re-set to 1st and 10.

2.3.2 Models of Higher Complexity

First we consider a model that takes into account time and score (Model 2). As will be shown in section 3.4, a categorical variable considering different times in the game, sometimes conditioning on how the offensive team is faring in that game, can be an important factor in determining play-calling. For the purpose of this model, we will consider an additional dimension of our state space, timescore, which has five levels and thus increases the size of our state space to \( 90 \cdot 5 = 450 \). The levels are:

- “Steady-state”: first quarter, second quarter (first 11 minutes), third quarter.
- “First half four-minute drill”: Last four minutes of the second quarter.
- Offense losing in the fourth quarter.
- Game tied in the fourth quarter.
- Offense winning in the fourth quarter.

The model environment itself is very similar to that of Model 1. The key difference is that two new variables are sampled at the beginning of each drive to account for the new state
dimension: first, \texttt{timesecs} is sampled from a discrete uniform (0, 3600). This is the number of seconds left in the game. Then, \texttt{scorediff} is sampled from a discrete uniform (0, 2) where 0 is offense losing, 1 is tied, 2 is offense winning. As the drive progresses, everything proceeds the same as \textbf{Model 1} except for two additional considerations

- If the play elapses longer than was remaining in the “half” (\texttt{timesecs} modulo 1800), then drive ends.

- After each play, \texttt{timesecs} is reduced by the time elapsed by the play.

Now we consider a model that takes into account whether the last play was a run or a pass (\textbf{Model 3}). As will be shown in section 3.4, whether the last play was a run or not can be an important factor in predicting play-calling. For the purpose of this model, we will consider an additional dimension of our state space (compared to the baseline model), \texttt{lastplay}, which has three levels and thus increases the size of our state space to $90 \cdot 3 = 270$. The levels are:

- 0: last play was a run
- 1: last play was a pass
- 2: last play was NONE (first play of drive)

The model environment is almost entirely identical to that of \textbf{Model 1}. The only difference is that we keep track of the last play and use that to determine the new play-call.
It’s one thing to have the da-ta or day-ta. It is another thing to know how to read the damn thing… I still think doing things the old fashioned way is a good way.

Jon Gruden, Raiders head coach

3.1 Data Sets

Two data sets were used in this thesis. The default data set that was used in all of the analysis, unless explicitly stated otherwise, is play-by-play data collected and published on Kaggle by researchers at Carnegie Mellon. This data set includes important play information, including situation and outcome columns, along with proprietary advanced analytics for each
play (including expected points, win probability, etc). The full data set includes 452,411 rows and 103 columns.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive</td>
<td># drive of the game</td>
<td>integer</td>
</tr>
<tr>
<td>qtr</td>
<td>quarter in which play took place</td>
<td>{1,2,3,4,5}</td>
</tr>
<tr>
<td>down</td>
<td>time left in quarter</td>
<td>{1,2,3,4}</td>
</tr>
<tr>
<td>time</td>
<td>seconds left in game</td>
<td>string (ex: “12:32”)</td>
</tr>
<tr>
<td>TimeSecs</td>
<td>seconds that play took to occur</td>
<td>integer (0-3600)</td>
</tr>
<tr>
<td>PlayTimeDiff</td>
<td>field position (yards to end zone)</td>
<td>integer</td>
</tr>
<tr>
<td>ydline100</td>
<td>yards needed for a first down</td>
<td>integer (0-100)</td>
</tr>
<tr>
<td>ydstogo</td>
<td>whether the team is in goal-to-go territory</td>
<td>{0,1}</td>
</tr>
<tr>
<td>GoalToGo</td>
<td>offensive team code</td>
<td>string (ex: “NE”)</td>
</tr>
<tr>
<td>posteam</td>
<td>defensive team code</td>
<td>string (ex: “PHI”)</td>
</tr>
<tr>
<td>DefensiveTeam</td>
<td>home team code</td>
<td>integer</td>
</tr>
<tr>
<td>ScoreDiff</td>
<td>net yards gained by offense on play</td>
<td>integer</td>
</tr>
<tr>
<td>HomeTeam</td>
<td>whether touchdown was scored on play</td>
<td>{0,1}</td>
</tr>
<tr>
<td>PassType</td>
<td>whether offense converted first down</td>
<td>{0,1}</td>
</tr>
<tr>
<td>PassOutcome</td>
<td>type of play by team with ball</td>
<td>run/pass</td>
</tr>
<tr>
<td>PassLength</td>
<td>outcome of a pass if attempted</td>
<td>incomplete/complete</td>
</tr>
<tr>
<td>AirYards</td>
<td>description of pass distance</td>
<td>deep/short</td>
</tr>
<tr>
<td>PassLocation</td>
<td>distance a pass traveled in the air</td>
<td>integer</td>
</tr>
<tr>
<td>InterceptionThrown</td>
<td>direction of pass</td>
<td>left/right/middle</td>
</tr>
<tr>
<td>RunLocation</td>
<td>whether interception thrown on play</td>
<td>{0,1}</td>
</tr>
<tr>
<td>FumbleLost</td>
<td>direction of run</td>
<td>{0,1}</td>
</tr>
</tbody>
</table>

Table 3.1: Column descriptions related to the game situation.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Touchdown</td>
<td>whether touchdown was scored on play</td>
<td>{0,1}</td>
</tr>
<tr>
<td>FirstDown</td>
<td>whether offense converted first down</td>
<td>{0,1}</td>
</tr>
<tr>
<td>PlayType</td>
<td>type of play by team with ball</td>
<td>run/pass</td>
</tr>
<tr>
<td>PassOutcome</td>
<td>outcome of a pass if attempted</td>
<td>incomplete/complete</td>
</tr>
<tr>
<td>PassLength</td>
<td>description of pass distance</td>
<td>deep/short</td>
</tr>
<tr>
<td>AirYards</td>
<td>distance a pass traveled in the air</td>
<td>integer</td>
</tr>
<tr>
<td>PassLocation</td>
<td>direction of pass</td>
<td>left/right/middle</td>
</tr>
<tr>
<td>InterceptionThrown</td>
<td>whether interception thrown on play</td>
<td>{0,1}</td>
</tr>
<tr>
<td>RunLocation</td>
<td>direction of run</td>
<td>{0,1}</td>
</tr>
<tr>
<td>FumbleLost</td>
<td>whether fumble lost by offense on play</td>
<td>{0,1}</td>
</tr>
</tbody>
</table>

Table 3.2: Column descriptions related to the play outcome.
<table>
<thead>
<tr>
<th>Column Name</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field_Goal_Prob</td>
<td>probability that offense scores field goal on drive</td>
<td>float: [0,1]</td>
</tr>
<tr>
<td>Touchdown_Prob</td>
<td>probability that offense scores touchdown on drive</td>
<td>float: [0,1]</td>
</tr>
<tr>
<td>ExpPts</td>
<td>offense’s expected points on drive (pre-snap)</td>
<td>float</td>
</tr>
<tr>
<td>EPA</td>
<td>change in expected points due to play</td>
<td>float</td>
</tr>
<tr>
<td>Win_Prob</td>
<td>offense’s win probability (pre-snap)</td>
<td>float: [0,1]</td>
</tr>
<tr>
<td>WPA</td>
<td>change in offense’s win probability due to play</td>
<td>float: [-1,1]</td>
</tr>
</tbody>
</table>

Table 3.3: Column descriptions related to advanced analytics.

Tables 3.1, 3.2, 3.3 show relevant columns, placed into three different tables separated by type of feature. These tables include the column name in the data, a description of the feature, and a description of the values in the column.

3.2 Data Cleaning Process

The data cleaning process was extensive and focused on preparing the data for the specific reinforcement learning methodology that comprises the primary focus of this paper.

The following steps were taken to clean the data up front:

- For PlayType, values of “Sack” were changed to “Pass” to account for the fact that sacks occur on pass plays.

- Column NextPlayFG created, which equals 1 if the next play was a successful field goal, 0 if it was an unsuccessful field goal, and is null if the next play was not a field goal attempt. This is to be used in the football simulation aspect of reinforcement learning so that failed third down attempts can be potentially converted into field goal attempts.

- Any row where PlayType does not equal “Run” or “Pass” is removed, then data limited to 1st-3rd downs. This matches the simulation.
• Nearly 70 columns were removed that had no relevance to the analysis.

• *Fumble*, which indicated a fumble on the play, was replaced by *FumbleLost*, which is 1 the offensive team lost possession via a fumble (by comparing the recovery team to the offensive team) and 0 otherwise.

• For the purpose of time analysis, only plays in regulation were kept (not overtime), so \( qtr \leq 4 \).

• *Touchdown* was adjusted so that only offensive touchdowns were considered (based on positive versus negative EPA).

Then the state space for reinforcement learning was created in each model as prescribed in the methods section. Thus, categorical variables like \texttt{OFFFIELDPOSITIONrl} and \texttt{YDSTOGorl} were created for the purpose of building the discrete state space on which the \( Q \)-table was based.

### 3.3 Data Summary

It is important to understand, in broad terms, the shape and makeup of the data that is being used. Below, Table 3.4 gives a simple overview of the cleaned data set using summary statistics.

First, it is worth noting the frequency with which certain types of plays occur at the bottom of the table; fumbles and interceptions are equally prevalent, while the offense is just as likely to score a touchdown on a given play as they are to give up a sack. One also sees
<table>
<thead>
<tr>
<th>Column Name</th>
<th>Mean ($\mu$)</th>
<th>Median</th>
<th>StdDev ($\sigma$)</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>down</td>
<td>1.77</td>
<td>2</td>
<td>.78</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>TimeSecs</td>
<td>1765 1815</td>
<td>1039</td>
<td>0</td>
<td>3600</td>
<td></td>
</tr>
<tr>
<td>PlayTimeDiff</td>
<td>23.14 27</td>
<td>16.53</td>
<td>0</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>yrdline100</td>
<td>52.86 57</td>
<td>24.15</td>
<td>1</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>ydstogo</td>
<td>8.76 10</td>
<td>3.95</td>
<td>1</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>ScoreDiff</td>
<td>-1.3 0</td>
<td>10.2</td>
<td>-44</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>EPA</td>
<td>-0.028 -0.237</td>
<td>1.39</td>
<td>-12.84</td>
<td>9.51</td>
<td></td>
</tr>
<tr>
<td>WPA</td>
<td>.001 -0.003</td>
<td>.042</td>
<td>-69</td>
<td>.79</td>
<td></td>
</tr>
<tr>
<td>YardsGained</td>
<td>5.42 3</td>
<td>8.82</td>
<td>-22</td>
<td>97</td>
<td></td>
</tr>
</tbody>
</table>

Proportion of Is ($p$)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Touchdown</td>
<td>3.9%</td>
</tr>
<tr>
<td>GoalToGo</td>
<td>5.4%</td>
</tr>
<tr>
<td>InterceptionThrown</td>
<td>1.3%</td>
</tr>
<tr>
<td>FumbleLost</td>
<td>1.3%</td>
</tr>
<tr>
<td>Sack</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

Table 3.4: Summary statistics of key quantitative and binary variables.

that, on average, Expected Points Added (EPA) for the offense is negative—but slightly right-skewed. Win probability added, however, is very slightly positive on average but left-skewed. The yards gained per play by an offense is 5.42 on average and right-skewed, which is unsurprising given the opportunity for big positive plays.

Given that this analysis is focused on play calling in certain situations, particularly run versus pass, we also want to understand the empirical data in an attempt to gain intuition about what real NFL coaches do. Below, Table 3.5 gives a simple summary of the likelihood and success of running and passing in different down-distance scenarios. “S” is short (1-3 yards), “M” is medium (4-6 yards), “L” is long (7+ yards). “FD” indicates that a first down
was gained.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>270</td>
<td>386</td>
<td>13317</td>
</tr>
<tr>
<td>Run %</td>
<td>67%</td>
<td>59%</td>
<td>52%</td>
</tr>
<tr>
<td>R Avg (yds)</td>
<td>0.58</td>
<td>2.56</td>
<td>4.32</td>
</tr>
<tr>
<td>P Avg (yds)</td>
<td>0.99</td>
<td>3.96</td>
<td>6.93</td>
</tr>
<tr>
<td>R SD (yds)</td>
<td>1.75</td>
<td>2.81</td>
<td>6.36</td>
</tr>
<tr>
<td>P SD (yds)</td>
<td>1.12</td>
<td>6.65</td>
<td>10.64</td>
</tr>
<tr>
<td>R FD %</td>
<td>42%</td>
<td>21%</td>
<td>12%</td>
</tr>
<tr>
<td>P FD %</td>
<td>60%</td>
<td>40%</td>
<td>29%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1538</td>
<td>3226</td>
<td>5837</td>
</tr>
<tr>
<td>Run %</td>
<td>63%</td>
<td>46%</td>
<td>35%</td>
</tr>
<tr>
<td>R Avg (yds)</td>
<td>3.46</td>
<td>4.23</td>
<td>4.62</td>
</tr>
<tr>
<td>P Avg (yds)</td>
<td>5.71</td>
<td>5.99</td>
<td>6.33</td>
</tr>
<tr>
<td>R SD (yds)</td>
<td>5.28</td>
<td>6.79</td>
<td>6.65</td>
</tr>
<tr>
<td>P SD (yds)</td>
<td>9.17</td>
<td>9.44</td>
<td>9.61</td>
</tr>
<tr>
<td>R FD %</td>
<td>58%</td>
<td>27%</td>
<td>13%</td>
</tr>
<tr>
<td>P FD %</td>
<td>50%</td>
<td>41%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Table 3.5: Summary statistics of down-specific play calls and outcomes.

The results of Table 3.5 have the major caveat that no situational information is accounted for—including field position, score difference, time, offensive capability, defensive capability, etc. It appears that, overall, passing yields significantly higher expected yards with higher “risk” (in terms of variation in yards gained). While some of this variance can be explained by the right-skewedness—due to long pass plays, which are good outcomes—some of it also comes from sacks and incompletions which are certainly negative.

Meanwhile, passing appears to significantly more profitable from a yards and conversion perspective on first and second down. On third down, however, running tends to yield higher conversion in short yardage situations, the two are similar in middle distance situations, and passing is weakly preferred in third and long (which is surprising). It is important to note that in many third and long situations, teams choose to run in situations where the
likelihood of success is low and therefore yards gained can be inflated.

3.4 Exploratory Modeling

3.4.1 Considering Time and Score Effects on Play Prediction

We want to consider the effect of the time (both in terms of quarter and time remaining) along with the score at that time. Are these effects relevant in considering a modeling of play calling? Do teams behave differently and should our reinforcement learning model thus consider interactive effects as part of the state space?

Specifically, we explored the hypothesis that a five-level categorical variable that captured this interactivity would better predict play calling than a simpler model. First, we fit a logistic regression model on the response variable \textit{PlayType} with \( Y = 1 \) correspond to “Pass” and \( Y = 0 \) corresponding to “Run”. This first model included just three predictors: down (as a 3-level factor), distance, and field position. The results are shown below in Table 3.6.

\[
\begin{array}{|c|c|c|}
\hline
\text{Predictor} & \beta_j & p\text{-value} \\
\hline
\text{(Intercept)} & -1.174693 & < 2e^{-16} \\
\text{as.numeric(yrdline100)} & .000346 & .0488 \\
\text{ydstogo} & .10706 & < 2e^{-16} \\
\text{as.factor(down)1.0} & 0.631 & < 2e^{-16} \\
\text{as.factor(down)2.0} & 1.8596 & < 2e^{-16} \\
\hline
\end{array}
\]

\textbf{Table 3.6:} The coefficient summary of the three-predictor logit model.

We then created a 5-level factor variable \textit{TIMESCOREr1} which takes on the following values
for certain conditions:

- \( \text{TIMESCOREl} = 0 \) is the "steady-state" (first quarter, third quarter, first 15 minutes of the second quarter).

- \( \text{TIMESCOREl} = 1 \) is the 4-minute drill at the end of the second quarter.

- \( \text{TIMESCOREl} = 2 \) is the 4th quarter when the offensive team is losing.

- \( \text{TIMESCOREl} = 3 \) is the 4th quarter when the game is tied.

- \( \text{TIMESCOREl} = 4 \) is the 4th quarter when the offensive team is winning.

We then ran a similar logistic regression as above and examined the results of this new, bigger model; a summary of the model can be found in Table 3.7. We then examined the efficacy and necessity of this new categorical variable by running a \( \chi^2 \) analysis of deviance test. The results of this test can be found in Figure 3.1.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>( \beta_j )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.174693</td>
<td>&lt; 2e(^{-16})</td>
</tr>
<tr>
<td>as.numeric(yrdline100)</td>
<td>.000744</td>
<td>0.149</td>
</tr>
<tr>
<td>ydstogo</td>
<td>.1146</td>
<td>&lt; 2e(^{-16})</td>
</tr>
<tr>
<td>as.factor(down) 1.0</td>
<td>0.692</td>
<td>&lt; 2e(^{-16})</td>
</tr>
<tr>
<td>as.factor(down) 2.0</td>
<td>2.013</td>
<td>&lt; 2e(^{-16})</td>
</tr>
<tr>
<td>as.factor(TIMESCOREl) 1</td>
<td>.59493</td>
<td>&lt; 2e(^{-16})</td>
</tr>
<tr>
<td>as.factor(TIMESCOREl) 2</td>
<td>1.1201</td>
<td>&lt; 2e(^{-16})</td>
</tr>
<tr>
<td>as.factor(TIMESCOREl) 3</td>
<td>.12794</td>
<td>0.191</td>
</tr>
<tr>
<td>as.factor(TIMESCOREl) 4</td>
<td>-.90147</td>
<td>&lt; 2e(^{-16})</td>
</tr>
</tbody>
</table>

Table 3.7: The coefficient summary of the logit model with time-score added.
Analysis of Deviance Table

Model 1: PlayType ~ as.numeric(yrdline100) + ydstogo + as.factor(down)
Model 2: PlayType ~ as.numeric(yrdline100) + ydstogo + as.factor(down) +
         as.factor(TIMESCORErl)

<table>
<thead>
<tr>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Dev Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31321</td>
<td>39607</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>31317</td>
<td>37798</td>
<td>4</td>
<td>1808.9 &lt; 2.2e-16 ***</td>
</tr>
</tbody>
</table>

Signif. codes:  0 ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure 3.1: R output of ANOVA comparing basic vs time-score GLMs.

The results are unsurprising but nonetheless interesting. First, it appears that while a tie game in the fourth quarter is not statistically different in predicting run versus pass when compared to the steady-state, the other situations are. In particular, we find (unsurprisingly) that running is more common when a team is winning in the fourth quarter while passing is more common when a team is losing in the fourth quarter, controlling for other factors. The ANOVA results, with a $p$-value of less than $2.2 \cdot 10^{-16}$, suggests that the model including the time-score effects has a statistically significantly better fit and thus these new predictors should be considered.

3.4.2 Serial Correlation

We want to consider the effect of the previous play call on the next one. We fit a logistic regression model with the same predictors as above but add a categorical variable with three levels:

- **LastPlayType** = 0 is first play of drive.
- **LastPlayType = 1** is last play was run.
- **LastPlayType = 2** is last play was pass.

Below in Figure 3.2 we see the summary of the model with last play added.

![Table of coefficients](image)

The results indicate that the last play being a pass is not significantly different from "random" choice (for plays that are unknown) but the last play being a run leads to a statistically significant result: a reduced chance of the next play being a pass, conditioning on other factors. The analysis of deviance on the time-score versus serial correlation model, with a *p*-value of *< 6.08 · 10⁻⁶*, finds that adding the previous play as a categorical factor improves the model significantly, so it should be considered.
We confirmed that there’s a competitive advantage in analytics in a league that is structured to prevent you from having a competitive advantage.

Joe Banner, former Eagles and Browns executive

4

Results

4.1 Baseline Model

In examining the results of the reinforcement learning models, we want to infer intuitive results about football while also evaluating their efficacy in achieving the goal. The former results from a comparison of models, the latter from examining the results of following Q-learning in the simulated environment. The baseline model, based on 100 independent
iterations (newly trained $Q$-table each time) of 5000 independent epochs, resulted in an
average of 5.36 plays per drive—taken for each of the 100 iterations, it resembles $\bar{x}$ — with a
standard deviation of 0.048. Similarly, we observed an average of 30.39 yards per drive with
a standard deviation of 0.41. Empirically, for the 2017 season, the average took 5.81 plays to
gain 30.04 yards (per Football Outsiders).

We also examine the number of touchdowns, interceptions, fumbles, and field goals to
understand the scoring and risk results of the $Q$-learning method using empirical data.
The Appendix includes Figures A.1, A.2, A.3, and A.4 which are histograms of those drive-
ending results, with the $X$-axis representing the occurrence per 5000 drives.

The mean and standard deviations over the 100 iterations can be found in the title of each
plot. A drive with a touchdown or field goal results in more points (and thus higher $Q$-
values), and vice-versa for a drive with a turnover. The results appear to somewhat normally
distributed, although there is no theoretical result to suggest that they should be. The rate
per drive, and its comparison to the empirical data from 2017, is noted in table 4.1 below.

For context, NFL offenses average approximately 11 drives per game.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>RL Model</th>
<th>2017 NFL Season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Touchdowns</td>
<td>.164</td>
<td>.197</td>
</tr>
<tr>
<td>Field Goals</td>
<td>.108</td>
<td>.152</td>
</tr>
<tr>
<td>Interceptions</td>
<td>.063</td>
<td>.075</td>
</tr>
<tr>
<td>Fumbles</td>
<td>.081</td>
<td>.043</td>
</tr>
</tbody>
</table>

Table 4.1: Comparing per-drive rates of statistics (simulation vs empirical).
In addition to actually simulating drives, the $Q$-learning model also trained an optimal policy, $\pi^*$, through iterative updates of the $Q$-table (which estimates the value function). A visualization of this optimal policy can be found below. First and ten will be considered in a later section, but Figure 4.1 represents 2nd down and Figure 4.2 represents 3rd down. Each field is split into field position and yards to go levels (the same way the algorithm does), so each $s \in S$ is displayed. First $\min(Q)$ was subtracted from each value in the $Q$-table to re-scale the values, and then each value seen below is calculated by

$$\kappa = \frac{Q(s, \text{pass})}{Q(s, \text{pass}) + Q(s, \text{run})}$$

for each state $s$. We have, thus, $0 \leq \kappa \leq 1$ and $\kappa = 0.5$ represents the indifference point between running and passing. Blue represents pass, with “stronger” shades of blue representing a greater preference for pass, and vice-versa for running with red instead.

![Baseline Q-Learning Model: 2nd Down Policy Heatmap](image)

**Figure 4.1:** The preference of pass versus run on second down for different field positions and distances to go (based on the baseline $Q$-learning model).
Figure 4.2: The preference of pass versus run on third down for different field positions and distances to go (based on the baseline $Q$-learning model).

These “football field” visualizations offer quick-and-easy insight into the policy learned by this reinforcement learning algorithm. First down usually involves 10 yards to go, which is why little data exists for “medium” and “short”. There appears to be a preference for passing on first down that declines over time as the offense approaches the goal line. Second down has few states where a strong preference exists, and is primarily populated by weakly-preferred passing states. Third down has significantly more variance in its scaled preference values; for example, with 7+ yards to go on 3rd and goal, it is strongly preferred to pass versus run. Meanwhile, in the middle of the field (but before obvious field goal range), when there are 4-6 yards to go, it is preferred to run.
4.2 Time and Score Model

In the Appendix you can find Figures A.5, A.6, A.7, and A.8 that show histograms of key statistics for Model 2. These “key events” are less normally distributed and have higher variance than the baseline model. We see more touchdowns and field goals, on average, with fewer fumbles and nearly a similar number of interceptions.

We also continue with the useful $Q$-table visualization on the football field for this model; below are representations of third down preferences for different situations. Figure 4.6 is for the steady-state, Figure 4.5 for the second quarter four-minute drill, Figure 4.3 for an offense winning in the fourth quarter, and Figure 4.4 for an offense losing in the fourth quarter.

Table 4.2 shows the per-drive statistics for Model 2 relative to the empirical NFL data, similar to what was shown for Model 1.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Time/Score RL Model</th>
<th>2017 NFL Season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Touchdowns</td>
<td>.178</td>
<td>.197</td>
</tr>
<tr>
<td>Field Goals</td>
<td>.123</td>
<td>.152</td>
</tr>
<tr>
<td>Interceptions</td>
<td>.064</td>
<td>.075</td>
</tr>
<tr>
<td>Fumbles</td>
<td>.069</td>
<td>.043</td>
</tr>
</tbody>
</table>

Table 4.2: Comparing per-drive rates of statistics for Model 2.

To compare Model 1 and Model 2, a two-sample $t$-test was performed on the number of touchdowns in each iteration of 5000 drives. Thus, the two samples were 100 values for
Figure 4.3: Third down run/pass heatmap in 4th quarter for winning offense.

Figure 4.4: Third down run/pass heatmap in 4th quarter for losing offense.
Figure 4.5: Third down run/pass heatmap during last 4 minutes of 2nd quarter.

Figure 4.6: Third down run/pass heatmap in "steady state".
each model. Because each iteration is independent and the number of touchdowns is approximately normal as seen in the histograms, the $t$-test was seen as an effective method for model comparison. The mean of the baseline model was 821.68 and the mean of the time-score model was 890.43; this resulted in a $t$-statistic of 10.446 and a $p$-value of less than $2.2 \cdot 10^{-16}$, well below the $\alpha = .05$ level. This indicates that the two models draw touchdown values from distributions with different means. This means that a model considering time and score results in significantly more offensive touchdowns than one without.

4.3 Serial Correlation Effects Model

In the Appendix you can find Figures A.9, A.10, A.11, and A.12 that show histograms of key statistics for Model 3. This model has slightly more touchdowns and field goals on average than the baseline but has fewer than the time-score model. This model results in a similar number of interceptions but more fumbles than the time-score model as well.

To compare the impact of having the last play be a run versus a pass, below are representations of second and third down preferences for different situations. Figure 4.7 and Figure 4.8 are for second down plays after a run and pass respectively, while Figure 4.9 and Figure 4.10 are for third down plays after a run and pass respectively.
Figure 4.7: Second down policy heatmap after a run.

Figure 4.8: Second down policy heatmap after a pass.
Figure 4.9: Third down policy heatmap after a run.

Figure 4.10: Third down policy heatmap after a pass.
4.4 Model Comparison

One way to compare the different models visually is to describe their optimal policies on first and ten. First and ten is a situation that all of the models will see consistently, and thus gives us a useful sample size on which to compare the general “philosophy” of each. The first row is the baseline model, the next five rows are from the time-score model, and the last two rows are from the serial correlation effects model.

![Model Comparison: 1st and 10 Policy Heatmaps](image)

**Figure 4.11:** Comparing each of the models (in each situation) on pass versus run preference for first and ten.

There is not a strong preference for either running or passing in most situations across the three models. Losing in the fourth quarter prescribes running in the red zone, but this is because “winning” is not the objective; most defenses are lined up to defend the pass when protecting a late lead and thus runs can end up being successful despite draining clock. Oth-
erwise, most models and situations exhibit a slight preference for passing which is consistent with recent literature that advocates for more aggressive first down play calling. Overall, the use of scoring and not winning as a reward pushes the different models to produce more similar results than they otherwise would, which is certainly a weakness of the approach that must be considered.
There are a lot of skeptics [of analytics]. And that’s honestly probably on the analysts and the statisticians. You have to be able to explain it to football people in their terms.

Tony Khan, Jaguars SVP of Analytics

Discussion

As Tony Khan, son of Jaguars owner Shahid Khan, so succinctly put it, the main barrier to widespread adoption of analytics in football is the way it is presented and communicated to the people making the decisions on the sideline and the field. In order to fully benefit from the incredible progress being made by economists, data analysts, and computer scientists in football research, it is important to bridge the gap between the language of statistics and the
language of football. The goal of incorporating data into football strategy creation and evaluation is not that its answers should be taken directly at face value; as George Box once said, all models are wrong but some models are useful. When probabilities suggest courses of action that deviate from conventional football wisdom, it is important to provide context for these recommendations or else one risks the credibility of analytics in football.

It is with this mindset that we hope to evaluate policies trained by reinforcement learning algorithms for the purpose of play calling. Up front, it is important to acknowledge that a Q-learning approach cannot replace play calling and that the goal of analyzing the results is to identify interesting trends and anomalies that teams and coaches may incorporate into future game-planning or analysis. There are inherent weaknesses to the reinforcement learning approach, including its Markovian assumption that the sequence of plays leading up to a decision should be independent of the decision itself, and these must be considered.

5.1 Model-Specific Analysis

First, let us consider the policy suggestions of the baseline model under certain circumstances. On first down, of course there are limited plays where a team might find itself in first and short or medium (devoid of a penalty, which is not considered in this model). On 1st and 10, there is a slight preference for passing that weakens as the offense gets closer to the goal line (even switching in the red zone). On second down, there are few significant
results. There is a strong preference for passing on second and short backed up inside your own goal line; this is likely due to the defense preparing for run plays on these situations. Surprisingly, second and long is no more favorable towards passes than medium or short. This likely results from the nature of second down, where offenses tend to set up a manageable third down situation empirically.

Third down offers a few interesting tidbits. Firstly, it recommends rushing in a number of third and short situations, particularly when the offense is backed up on its own goal line. Since first down conversion is the biggest driver of expected points (a.k.a the ability to score in the future), and 63% of runs are successful versus 50% of passes according to Table 3.5, these results make sense. Overall, passing becomes more favorable as the offense moves down the field; this again makes sense because passing’s risks become less pronounced in opposing territory (interception, sack) and due to the defense tightening up against the run with limited space behind them. This preference for passing is particularly strong on third down and long inside the 10 because of the difference between a touchdown and field goal along with the defense’s ability to stack the box effectively.

We then consider the time-score model, and focus on third down particularly due to the high-leverage nature of these situations. The steady-state policy resembles the baseline model’s preferences on first and second down: fairly conservative with a preference for running across the board (even in medium and long situations) until the offense gets close to
the end zone. The third down policy in the first half “four-minute drill” is a bit more aggressive, with more of a preference for passing especially in medium and long situations. This may be because teams tend to be forced into managing time more carefully in these situations and therefore are looking to convert without burning clock. Runs may gain yards but can reduce the probability of converting because of a running clock (which is considered in the time used in the model). It is interesting that offenses are encouraged to run more often in the red zone; while this goes against the aggressive nature of four-minute offensive principles, it does reduce the likelihood of a disastrous play that pushes the offense out of field goal range with a sack or interception. The Q-learning model is likely training a policy that is influenced heavily by this large, discrete negative reward change.

We then compare the optimal policies for winning and losing offenses in the fourth quarter. The most important fact to consider when assessing the results is the primary goal of the agent in the model as currently constructed: scoring provides rewards, not winning. Therefore, traditional strategies like running in the fourth quarter to burn clock are not adopted by the model. This is certainly a weakness to consider because any coach should be optimizing for victory. However, given the parameters of the reinforcement learning environment, it is possible to observe defensive tendencies (prevent defense versus aggressive blitzing) that can determine the success of certain offensive strategies.

Broadly, the winning offense tends to prefer running across the board to the losing offense.
In particular, it is worth noting that in the middle of the field, as the offense approaches the middle of the field or field goal position, the difference becomes more drastic. The winning offense strictly prefers running compared to the losing offense, which fits our existing intuition because the defense will be less aggressive and the winning offense is less concerned with scoring a touchdown. This same pattern continues into the red zone; in third and short/medium situations, the winning offense is told to run more than the losing offense. Again, the football intuition is that a winning offense wants to burn clock rather than score points, but this model’s preference likely results from defensive tendencies and offensive biases towards these particular strategies.

Next, let us examine the policy suggestions of the serial correlation model in different situations. On second down, we find that passing on first down makes running more advantageous compared to when a team ran on first down. This is particularly pronounced as the offense approaches its opponent’s goal line. This could result from the condensed nature of the field and the usefulness of “softening up the front seven” with a first-down pass to set up the run. Third down does not produce any noticeable differences or trends. Overall, this serial correlation effects model only barely outperforms the baseline and performs much worse than the time-score model; this suggests that the previous play may be helpful in prediction but is not as helpful for optimization.
5.2 General Takeaways

So, what are the broader takeaways from these model-specific recommendations? What shifts in thinking or attitude has this initial research suggested? The only way to answer such questions is to establish a baseline for comparison, for which we will use an Expected Points Added (EPA) approach which has become adopted in a widespread manner by the analytics community. In the Appendix we find $Q$-table visualizations that correspond to the average of the expected points added (could be positive or negative) for all run and pass plays for those specific circumstances, which is the current advanced analytic standard-bearer for measuring offensive success. However, this paper has aimed to develop a data-driven play calling methodology that extends beyond just studying plays in isolation; decision processes of the kind explored in this paper allow for experiments that resemble the environment that offensive decision-makers will find themselves in. Therefore, it is useful to compare reinforcement learning-based and expected points-based outcomes.

As a brief aside, EPA is preferred to simply empirical data, of the kind exhibited in Figure A.14, for a couple reasons. Firstly, empirical data merely describes what coaches currently do generally without additional nuance; meanwhile, our reinforcement learning approach attempts to contextualize each decision and thus is difficult to compare directly. While examining trends as this paper has done is interesting, reinforcement learning is de-
signed to be compared on a play-by-play basis with past coaching intuition, rather than in broad strokes. Secondly, because reinforcement learning is prescriptive and not predictive, it is more interesting to compare it with another, albeit simpler, prescriptive methodology.

On second down, per Figure A.15, expected points added suggests running in nearly all short situations, passing in almost all medium situations, and it prefers running in the middle of the field/in field goal range while preferring passing near the respective end zones. Most of this agrees with conventional wisdom; running on second and short is low-risk and provides a high likelihood of success, while teams should be more aggressive through the air on second and long when three points are not risk. Our reinforcement learning model, however, leans more heavily towards passing across the board. A notable difference occurs on second and long as the offense approaches the 50-yard line and beyond into the defense’s territory; the difference is that an EPA model will reward the offense for gaining three or four yard on a running play because it increases the probability of a field goal (or even a touchdown) in expected points terms. Meanwhile, since a reinforcement learning works in discrete terms, for both the reward and state space, a running play on second long will not lead to a higher-reward state in all likelihood and thus passing, while riskier, is seen as the better option. The latter agrees with the thought process of some top offensive minds, especially the more aggressive play callers the league has seen in recent years; second and nine is not very different from third and seven, and therefore the offense should be aggressive in
trying to significantly improve its standing in those situations rather than chipping away for a few yards.

Similarly, on third down, expected points added has a significant preference for running over passing. Interestingly, this is particularly true in the red zone and in third and medium (four to six yard) situations. The \( Q \)-learning approach agrees with the latter but definitely prefers passing on third down as the offense reaches its opponent’s 25-yard line and closer. Again, expected points prefers caution in the red zone because of the dual effect of a sack or interception causing the offense to miss out on points and giving the opposing team a better opportunity to score points. Meanwhile, the MDP has a discrete (and equal) negative reward for both fumbles and interceptions along with a stronger preference for converting higher-leverage third down situations. This accounts for why reinforcement learning will advocate for a more aggressive strategy in the red zone, which is consistent with recent secular shifts in the NFL.

In conclusion, we find that reinforcement learning proposes, on average, more pass-heavy play calling than EPA would suggest. This stems from the way reinforcement learning is trained, particularly how the value function is generated over the course of many iterations. Rewards are discrete and only realized at the end of drives, through scoring plays or turnovers; therefore, with a finite number of simulations, the values of the \( Q \)-table that represent states close to the goal line will be higher and propagate down to the other states.
through the $Q$-value update formula. The $Q$-learning process for the simulation environment in this study is like if a coach received a huge psychological shock when big plays happened and then slowly remembered over time which situations—based on field position, down, distance, etc.—resulted in those outcomes, positive or negative. For example, a coach would weigh the positive effects of scoring a touchdown far more than the negative effects of failing to convert a fourth down even if both occurred just as often based on a particular decision.

Unlike EPA, which will assign non-trivial values to incremental changes in the offense’s situation (based on the likelihood of scoring a certain amount of points), the agent in this reinforcement learning environment will seek out high-reward outcomes and states directly. This provides an alternative, unique way of evaluating and ultimately formulating play calling strategies; these types of aggressive policies have been adopted by some of the more successful offensive tacticians in recent years and thus gives additional credence to using reinforcement learning methods.

5.3 Model Comparison

In addition to providing recommendations on a model-by-model basis, it is important to establish a system for comparing the efficacy of different $Q$-learning methods (and broadly, other reinforcement learning techniques). There is not a significant amount of consensus
in the literature because of the problem-specific nature of reinforcement learning; therefore, the number of touchdowns scored in each iteration was used as an approximation of the model’s success. As the histograms show, touchdowns for each iteration of 5000 drives were sampled from a symmetric distribution which allowed us to use a t-test for statistical significance. This showed that the time-score model produced significantly better results; this cannot merely be explained away by over-fitting because unlike linear regression, additional predictors or variables do not automatically increase the training success of the model. Nonetheless, the time-score model’s superiority is unsurprising given that offenses can adapt better to defensive tendencies if they take into account that extra information.

Going forward, the ultimate aim of developing and perfecting reinforcement learning—both the simulation environment and the training method—techniques for football is to optimize offensive success. Thus, as additional dimensions are added to the state space and the methodology is adopted to include new technical add-ons (case-based reasoning, model-based learning, etc), there must remain an objective measure for model comparison. This measure will be environment and MDP-specific; in this case, we measured touchdowns but in the future it could be score differential or points per drive.
We have all this information but so does everyone else.

What advantage does it give us to get it? None. It’s what we do with it, the way we use it.

Kevin Colbert, Steelers GM

Conclusion

The big question one should be asking is: how can an NFL team, and more specifically its primary play caller (whether that be a head coach or offensive coordinator), use reinforcement learning methods like the ones described above in order to improve on-the-field outcomes? Right now, the models are not trained on a sufficient amount of data to justify using the results directly out of the box in calling plays. Rather, it should be used to gauge
heuristics and trends about how offenses should be approaching certain situations. These situations include all the different “dimensions” on which the above models were built, and can be tailored to fit the needs of a particular coach.

Reinforcement learning approaches, like Q-learning, will become particularly impactful in real-time when the state space begins to include information about offensive and defensive personnel; this can substitute for team effects. Coaches want to make play calling decisions not only based on the situation but while considering the players on the field on either side. By developing league-wide models that incorporate the talent and style of players at certain positions—likely starting quarterback—football analysts can help coaches feel more comfortable with algorithmic play calling.
In order to improve the application of reinforcement learning to play calling and achieve the lofty goals presented above, the methods and data used in this paper need to be augmented and improved upon. Firstly, it may be unavoidable to include data from multiple seasons to increase the sample size and avoid over-fitting. The state space will also need to be expanded to include information about personnel in terms of available positions, styles, and talent levels. This data can be drawn from a number of sources, whether it’s Madden ratings or past statistics. In order to accommodate this significantly larger state space, future play calling algorithms will likely have to adopt case-based reasoning which will train an optimal policy while acknowledging that current situations resemble past examples and should be treated as such. This allows the learner to more efficiently sample from an already limited data set and craft a “smoother” action set. Model-based approaches, which can include underlying designs from neural networks to Gaussian mixed models to anything in-between, can also allow the reinforcement learning algorithm to train its optimal policy more efficiently.

If these improvements can be made properly and the weaknesses of the approaches presented in this paper can be mitigated, there is no doubt that data-driven, algorithmic play calling techniques will continue to pervade the NFL. Reinforcement learning offers decision-making support trained in a compelling prototype of the typical environment that play callers find themselves in. MDPs offer the flexibility that coaches need to adapt the problem
to the specific pieces of information they care about, allowing them to only consider the factors that they find valuable. For example, they could update the reward function or simulation environment to match their philosophical approach to play calling and thus achieve a better match between model and practice. Most importantly, reinforcement learning can improve on current expected points-based models by better modeling the discrete nature of concrete rewards in football and thus capturing more of the psychology of offensive strategy. This combination of flexibility and improved efficacy indicates that football analysts should consider reinforcement learning as the standard bearer for play-calling optimization going forward.
Appendix
A.1 Baseline Model Histograms

Figure A.1: Histogram of touchdowns per 5000 drives in the baseline model.

Figure A.2: Histogram of field goals per 5000 drives in the baseline model.
Figure A.3: Histogram of interceptions per 5000 drives in the baseline model.

Figure A.4: Histogram of fumbles per 5000 drives in the baseline model.
A.2 Time-Score Model Histograms

**Figure A.5:** Histogram of touchdowns per 5000 drives in the time-score model.

**Figure A.6:** Histogram of field goals per 5000 drives in the time-score model.
Figure A.7: Histogram of interceptions per 5000 drives in the time-score model.

Figure A.8: Histogram of fumbles per 5000 drives in the time-score model.
A.3 Serial Correlation Model Histograms

Figure A.9: Histogram of touchdowns per 5000 drives in the serial correlation model.

Figure A.10: Histogram of field goals per 5000 drives in the serial correlation model.
Figure A.11: Histogram of interceptions per 5000 drives in the serial correlation model.

Figure A.12: Histogram of fumbles per 5000 drives in the serial correlation model.
A.4 Empirical $Q$-table Visualizations

Figure A.13: Empirical second down pass versus run preferences from NFL coaches.

Figure A.14: Empirical third down pass versus run preferences from NFL coaches.
A.5 EPA $Q$-table Visualizations

**Figure A.15**: Pass versus run on second down (based on Expected Points Added).

**Figure A.16**: Pass versus run on third down (based on Expected Points Added).
References

[1] RIP, NFL defenses: A new era of offense has changed the game we knew.  


