Fourth Down Decision Making: Challenging the Conservative Nature of NFL Coaches

Will Palmquist[1], Ryan Elmore[2], Benjamin Williams[2]
[1]Student Contributor, University of Denver
[2]Advisor, Department of Business Information and Analytics, University of Denver

Abstract

This thesis analyzes the hypothesis that coaches in the National Football League are often too conservative in their decision making on fourth downs. I used R Studio and NFL play-by-play data to simulate actual football plays and drives according to different fourth down strategies. By measuring expected points per drive over thousands of simulated drives, we are able to evaluate the effectiveness of different fourth down strategies. This research points to a number of conclusions regarding the nature of NFL coaches on fourth downs as well as the complexity of modeling and simulating decision making in a complex sport such as professional football. While we are able to demonstrate areas where a more aggressive fourth down strategy could be utilized to a team's advantage, this research demonstrates that fourth down decision is not a simple binary choice and that making this critical decision must be taken in context. In other words, further research should be done that takes into account additional variables and their impact on a team's decision to "go for it" or not on fourth down.

Keywords: sports analytics, data analytics, statistics, simulations, football, fourth downs

1 INTRODUCTION

The 2009 season marked a major turning point in the use of analytics in the National Football League[1]. For the first time, a large pool of data was available to teams and statistical modeling gained popularity in front offices and on the field. One particular game during that season is credited with bringing analytics into the spotlight. Coach Bill Belichick's New England Patriot's led the Indianapolis Colts by six points and faced a fourth-and-2 on their own 28-yard line. In these situations, teams have three choices: punting the ball, giving possession to the other team but leaving them further from a scoring opportunity; attempt a field goal for 3 points (if within range); or go for it with the intention of earning a first down but run the risk of a turnover. In this case, the Patriots were out of field goal range and going for it could result in giving the Colts a very strong field position. To the typical coach, a punt would seem like the obvious choice. Instead, Coach Belichick chose to go for it: they turned the ball over, the Colts gained possession deep into the Patriots' territory, and ultimately scored the game-winning touchdown on the ensuing drive. Though Belichick's decision proved costly, analytics proved that the decision gave the Patriots the greatest probability of victory[2]. This particular play, dubbed the "Belichick fourth-and-2", ultimately marked one of the first times modern analytics gained widespread coverage in the NFL.

A few years prior, the decision-making tendencies of NFL coaches began to attract analytical scrutiny. For example, using play-by-play data and dynamic programming, Romer[3] asserted that NFL teams' behaviors on fourth downs often fail to maximize their overall chances of winning the football game. It was suggested that coaches are too passive on fourth downs and elect to punt or kick a field goal too often. One drawback of this research was that it only covered fourth down scenarios with one yard to go for the first down. Though limited in scope, Romer's research began to show that coaches and teams were lacking in their ability to select plays, especially on fourth downs.

Building on Romer's research, Causey, Katz, and Quealy[4] developed what has since been termed the "New York Times 4th Down Bot". This expands on Romer's research to include optimal decision making at every yard line and every yards-to-go distance on the field (see Figure 1). Their research sought to provide an impartial strategy for approaching fourth downs and yielded results that would be considered very aggressive when compared to the typical coach's decision making. For example, their model suggests a team should

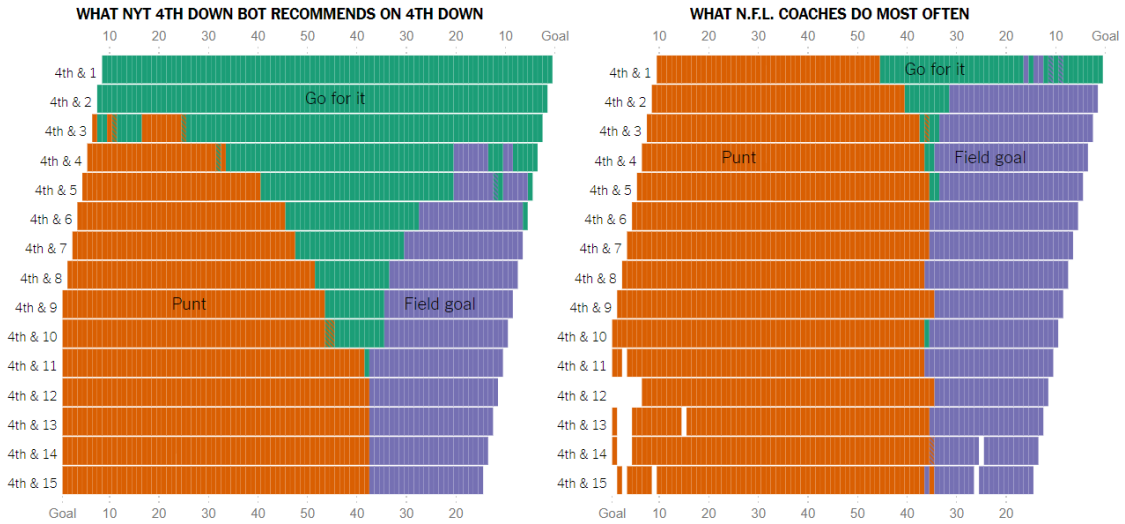

Figure 1. NYT 4th Down Bot

go for it on fourth-and-2 anywhere beyond their own 28-yard line, an idea that would seem extreme to the vast majority of coaches. This is exactly the strategy that was employed by Coach Belichick in the game versus the Colts described earlier. Their research supports, in summary, the idea that coaches are far too conservative on fourth downs.

More recently, Yam and Lopez[5] found that teams miss out on an extra 0.4 wins per year by not implementing a more effective fourth-down strategy. In their research, Yam and Lopez account for additional factors like time remaining, point differential, and the relative offensive and defensive strengths of each team. They draw a similar conclusion to past research: NFL coaches are too conservative.

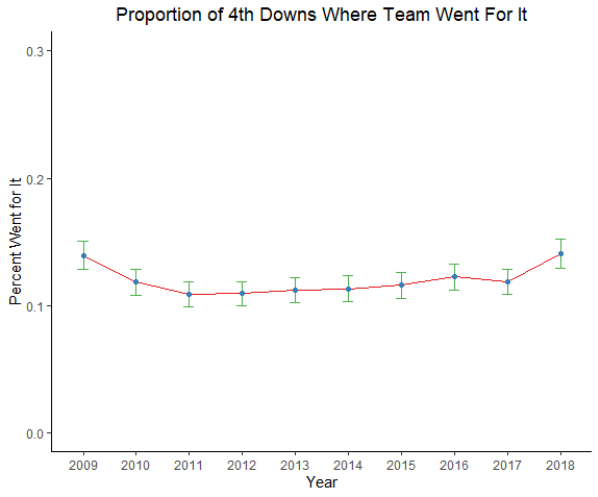

Figure 2. Proportion of fourth downs where the team went for it from 2009 to 2018. In other words, the proportion of run or pass plays on fourth down.

Despite convincing research, the conservative nature of the NFL on fourth downs has not changed in any meaningful way in recent years. Figure 2 demonstrates the relatively steady trend of fourth down decision making, with coaches electing to go for it between 11% and 14% of the time over the past 10 seasons. Despite an uptick in 2018, the data shows little to no changes in decision making on fourth downs. This begs the question: why is it that teams and coaches make decisions, particularly on fourth downs, that decrease their chances of winning? The answer may lie more in behavioral psychology than statistical analysis. A paper by Urschel and Zhuang[6] points to risk and loss aversion as the culprits of poor decision making, specifically on kick-off decisions. Their prospect theory-based model suggested that coaches tend to be overly cautious due to a greater sensitivity to losses relative to wins. This potentially comes from external factors like the ridicule faced when an aggressive play fails. In other words, the conservative choice, though suboptimal, will yield far less criticism. This idea of risk aversion could be one of the many factors influencing the strategic decision making of NFL coaches on fourth downs.

The purpose of this thesis is to test differences in a team's expected points by utilizing a very aggressive fourth down strategy. Past research has shown a tendency of NFL coaches to act conservatively on fourth downs. Despite a plethora of statistically significant findings, coaches have continued to execute suboptimal decision making in these situations. The 4th Down Bot points out that more than half of all plays result in a gain of 4 or more yards[7] so one would think that going for it on fourth down would be more common, especially when a team is within a couple of yards of a first down. However, coaches have largely stuck to their traditional

strategies and the use of an optimal decision-making strategy, supported by analytical theory, to choose plays is not widely adopted. One of the primary goals of this paper is to further demonstrate the need for coaches in the National Football League to approach fourth downs in a less conservative manner. With the nflsimulator package[8], we are able to simulate drives according to different strategies with the purpose of evaluating the effectiveness of a rather extreme fourth down strategy. The strategy is simple – regardless of field position, a team will choose to go for it on fourth down. Along different starting field positions, we aim to evaluate the difference in expected points per drive in comparison to a traditional fourth down strategy.

2 RESEARCH BODY

2.1 Data

In order to address this problem, we use NFL play-by-play data from the National Football League's (NFL) publicly available Application Programming Interface (API). In particular, this API is accessed via the well-known R statistical software[9] package called nflscrapR[10]. This package allows users to analyze an extensive library of NFL data on the single play, game, and season level. The functions within this package not only parse and clean the data from NFL.com but provide detailed metrics to enhance data analysis. In addition, Elmore and Williams[8] developed an R package called nflsimulator that is used to simulate plays and drives using the data in nflscrapR. Note that a drive is defined as a series of plays when the offensive team has possession of the ball. A drive ends when the team's possession of the ball ends, either through a score, punt, turnover, or the clock expiring.

The dataset used for simulating drives in this project includes almost fifty thousand plays from the 256 games in the 2018 NFL season. For each recorded play, nflscrapR provides over 250 individual variables ranging from simple items like the current yard line to more complex data like the expected points added from air yards on a pass play ("air_epa"). The level of granularity of the data allows for the nflsimulator functions to capture necessary data to simulate drives.

2.2 Methodology

2.2.1 Stimulating Drives

When we began the project, we wanted to simulate different fourth down strategies on a single drive basis and estimate the average expected points for each. Using the nflsimulator R package and the sample drive function in particular we are able to simulate individual drives according to different strategies. The function itself takes the starting yard line (measured as yards from the team's own goal), the data set used (in this case 2018 play-by-play data), and the scenario or strategy being tested. Figure 3 shows an example play used in the simulations.

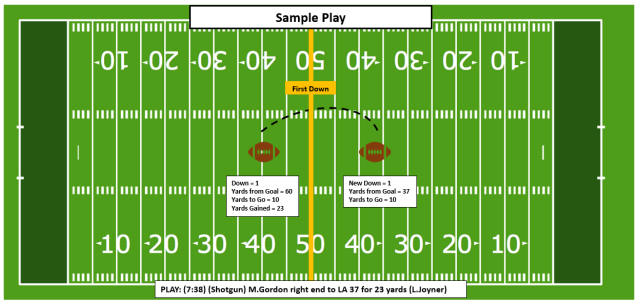

Figure 3. A sample play from the NFL's API, accessed by the nflscrapR package[10]. Shows a random run play starting 60 yards from goal on first down. Melvin Gordon ran for a gain of 23 yards to the opponent's 37-yard line.

Acting on the notion that NFL coaches are historically too conservative, we wanted to test a very aggressive fourth down strategy in comparison to what NFL coaches typically decide on fourth downs. The aggressive strategy in this case involves a team electing to go for it on fourth down regardless of the usual decision-making variables such as down, field position, and yards to go. It is important to note that the two strategies differ in the way they sample from the play-by-play data being used. The normal strategy will sample from the collection of actual plays as one would expect, matching variables like down, yards to go, yard line, etc for each new play in the drive. However, the "going for it" strategy samples from first and second down plays on second down, second and third down plays on third, and third and fourth down plays on fourth. In doing so, we eliminate some of the psychological effects of choosing to go for it on fourth down (as there is no alternative in this strategy) while allowing for a larger sample of plays.

A summary of the results comparing the two strategies is given in Figure 4. The results are based on running 1000 drive simulations for every five-yard increment between 5 and 95 under each respective strategy. We are able to compare the two strategies directly with regard to expected points per drive. A drive that ended without a score (turnover or punt) results in zero points, a touchdown is 7, a field goal is 3, and safety is 2. The final expected points value for each yard line is the average points per drive over all the simulations.

The results show that for the aggressive strategy (G), the expected points per drive is higher from the 5-yard line all the way to about the 40-yard line. At that point, the normal NFL strategy (N) yields a higher expected points value for all yard lines until the 95 (95 yards from the team's own goal). These results make sense in context as the data represents singular drives, so a

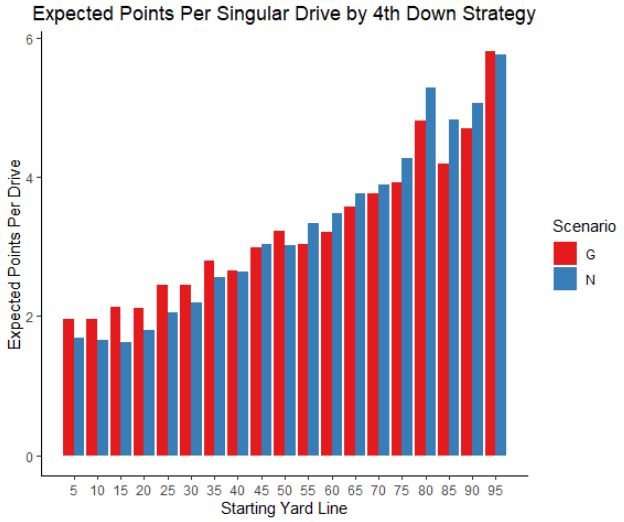

Figure 4. Graph is based on 1000 simulations of singular drives at every five-yard increment on the field for two different fourth down strategies. G represents the home team's strategy of going for it on fourth down no matter what, while N represents a team acting in accordance with normal fourth down decision making. Expected points per drive is the average points per drive of the simulation data.

simulation ends with either a turnover, a punt, or a score. From the 5- to 40-yard line, going for it would yield a higher amount of points, yet this does not factor in the opposing team's field position in the event of a turnover. This is important to consider since turning the ball over on a team's own 10-yard line would almost certainly lead to a score for the other team. However, comparing the two at the 95-yard line is worthy of discussion. After trailing the N strategy for a majority of the starting yard lines, the G strategy takes over again at the 95. In other words, when a team starts a drive 5 yards from goal, going for it on fourth down would yield greater expected points than acting in accordance with the typical NFL coach.

2.2.2 *Stimulating Until Score*

After analyzing the two strategies on a singular drive basis, we thought it was important to account for the opposing team's chances of scoring on the ensuing drive. For example, the opposing team's expected points after turning the ball over on your own 10 should be factored into the drive simulation. In order to account for this more realistic scenario, we utilized the sample drive until score function from the nflsimulator package. This function simulates drives until a team scores. The function takes into account factors such as field position in the event of a turnover, punt distance when a team elects to punt, and allows for the testing of the two strategies against each other. By taking a home team strategy and an away team strategy, we can assess how a strategy stacks up against the other. With regard to calculating and storing expected points, the only change comes from the fact that a score for the opposing team (strategy 2) results in a negative points value for the home team (strategy 1). The results of the simulations are shown in Figure 5. Note that these simulations are based on 1000 drives at each yard line from 5 to 95 in increments of five yards.

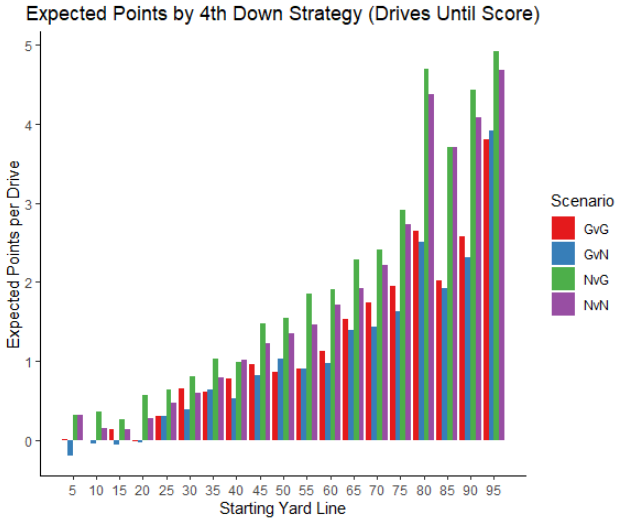

Figure 5. Using the drive until score function, we were able to run 1000 simulations at each five-yard increment on the field and for all four possible combinations of two fourth down strategies. The G in GvN represents the home team's strategy of going for it on fourth down no matter what, while N represents a team acting in accordance with normal fourth down decision making. On the y-axis is expected points which is the average expected points per simulation.

Figure 5 shows it is important to account for the opposing team when testing a fourth down strategy. In particular, the figure shows that a normal strategy for fourth downs is optimal at the vast majority of yard lines on the field when tested against the aggressive strategy. The aggressive strategy is indeed overly aggressive and lacks nuance in its decision making. In other words, home teams have higher expected points at almost all yard lines when using a normal strategy against a team that is using this very aggressive strategy. This makes sense when considering the fact that going for it on fourth down every time, regardless of field position or time of game, is not a realistic strategy. Not only can it give good field position to the opposing team, but it also could result in going for it in situations that are difficult to convert such as fourth down and 25 yards to go. Overall, it became clear that simply going for it on fourth down no matter what would likely lose out to the average NFL coach's fourth down play calling.

2.2.3 *Drives Per Score*

Next, we looked at an aspect of the simulation data other than expected points, namely, the number of possessions before a score takes place (drives per score).

Figure 6 shows the distribution of drives per score for each strategy at 10-yard increments on the field. These yard increments are the starting yard line of the simulation. The overall distributions are as one would expect. We see that the closer the starting yard line is to the goal line, the fewer number of drives typically take place before a score. Looking closer at the comparison between strategies shows the two scenarios that start with the normal strategy (NvG and NvN) have a higher count of drives with only one possession before a score. This is likely due to the ability to kick a field goal, resulting in teams having to travel a shorter distance before recording a score. Another aspect of the charts worth noting is the spike in the dark blue line (GvN) at the 10-, 20-, and 30-yard lines. For all 3, the GvN scenario has its highest frequency at a value of 2 drives per score. This suggests again that going for it close to your own goal line will often lead to the opposing team scoring on the following drive.

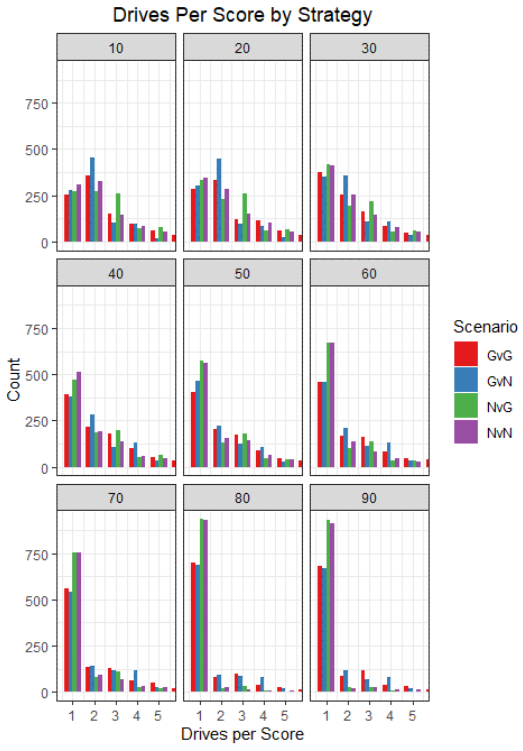

Figure 6. Shows the frequency of drives per score at 10-yard increments for all four possible combinations of two fourth down strategies. The G in GvN represents the home team's strategy of going for it on fourth down no matter what, while N represents a team acting in accordance with normal fourth down decision making.

2.2.4 *Testing a New Strategy*

After testing the rather extreme strategy of going for it on fourth down no matter the circumstance, we wanted to test another slightly different approach. The simulation data clearly pointed to the fact that a more effective fourth down strategy would need to take into account additional factors. This would include things like time of game, opposing team's skillsets and tendencies, or the yards to go to the first down. Because a fourth and 15 and a fourth and 1 are very different scenarios in the NFL, we chose to modify the existing fourth down strategy to a cutoff point of 5 yards to go. For example, a fourth down and under 5 yards to go results in going for it, while a fourth down of 5 yards or over results in the typical coach's decision (usually a punt or field goal). Figure 7 reflects this minor change in strategy on the singular drive level.

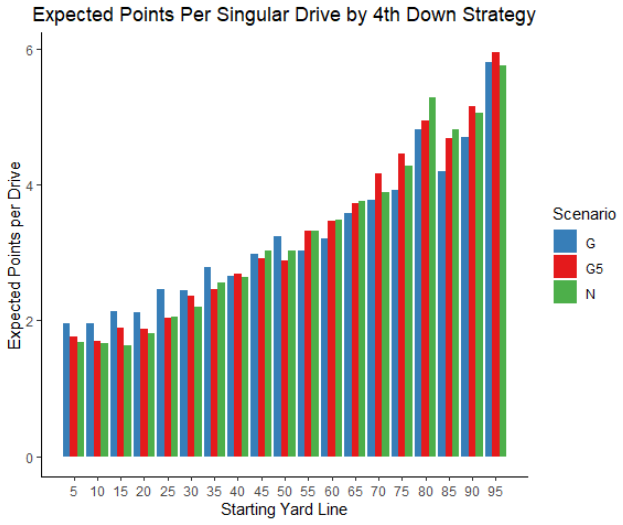

Figure 7. Graph is based on 1000 simulations of singular drives at every five-yard increment on the field for three different fourth down strategies. G represents the home team's strategy of going for it on fourth down no matter what, G5 represents a strategy of going for it no matter what when there is less than five yards to go, and N represents a team acting in accordance with normal fourth down decision making. Expected points per drive is the average points per drive for the simulations.

The figure shows that the new strategy (G5) yields a higher expected points per drive than the original aggressive strategy (G) at every yard line after 50. Not only does it outperform our more aggressive fourth down strategy, it has higher average expected points than the normal NFL coach's decision (N) at the 90- and 95-yard lines. As we mentioned earlier, the earlier yard lines (5-50) are harder to analyze without taking into account the opposing team's ensuing drive, so the underperformance of the strategy at those yard lines is not important to our results.

Figure 8 also shows an improvement over the original going for it strategy when reevaluating the drive until score simulations.

By comparing all three strategies when they are facing the typical decision making of an NFL coach, N, we can compare the three on a level scale. Besides a couple of interesting outliers at the 30- and 70-yard lines, the normal strategy still performs the best when put against

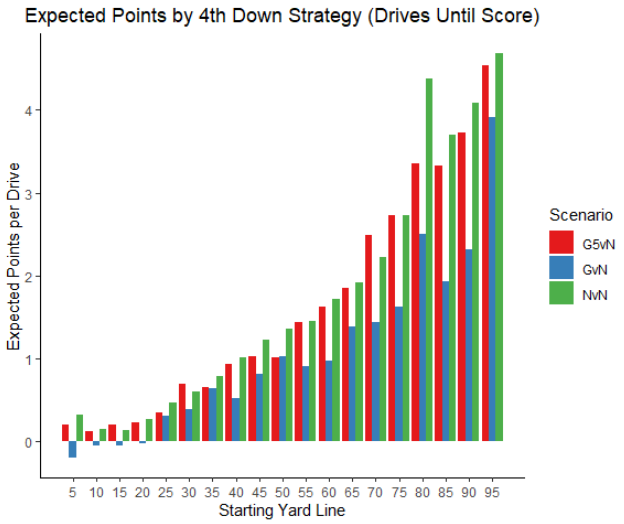

Figure 8. Uses the drive until score function to run 1000 simulations at each five-yard increment on the field and for combinations of opposing fourth down strategies. The G in GvN represents the home team's strategy of going for it on fourth down no matter what, the G5 in G5vN represents going for it only when there is less than five yards to go to a first down, and N represents a team acting in accordance with normal fourth down decision making. On the y-axis is expected points which is the average expected points per simulation.

an opponent with a normal strategy. However, the G5 strategy performs significantly better than the G strategy. This would suggest that the addition of this "under 5 yards to go" criteria to our going for it strategy leads to a noticeable improvement in fourth down decision making.

3 CONCLUSIONS AND FUTURE DIRECTIONS

Overall, this project revealed interesting conclusions as well as possible avenues for future research. By simulating NFL drives according to different fourth down strategies, we were able to show that coaches are likely too conservative on fourth downs in certain situations. Starting with a simple approach of simulating single drives, this research suggests that it is more beneficial to always go for it when starting at the 95-yard line. Furthermore, when simulating until a team scores, we can see that going for it every time on fourth down does not account for certain important aspects of the game. We found that a fourth down decision must be taken in the context of additional factors like the distance to the next first down or score. By adding a single criterion to the strategy where a team will go for it when it is fourth down and less than 5 yards to go, we saw a large improvement in expected points per drive. This suggests a possible idea for future research in which other factors are tested to create a better fourth down strategy. We could test other "yards to go" cutoffs or variables, like time remaining and opposing team defensive ability, to create a more realistic fourth down strategy.

Additionally, the project demonstrated a common trend regardless of the strategy or functions being used for simulation. In all data sets and for all strategies, there is a noticeable drop off in expected points from the 80- to the 85-yard line. This drop off is then followed by a return to the normal trend at the 90-yard line (see Figures 3, 4, 6, and 7). This could suggest a tendency towards coaches acting more conservatively when 15 yards from goal then when they are 20 or 10 yards from goal respectively. This odd finding is certainly something to look into in future research.

In conclusion, simulating fourth down strategies in the NFL has been the subject of extensive research often suggesting that coaches are too conservative in their decision making. The research in this paper supports some aspects of this argument and reveals very promising paths for future research. The strategies employed in our simulations demonstrate that a more aggressive strategy is warranted at certain yard lines and that fourth down decisions should not be made with one simple strategy, but with one that accounts for the unique context of each and every play.

4 LESSONS LEARNED

While conducting research for this thesis, I learned a lot of valuable lessons outside of fourth down decision making in the NFL. First and foremost, I was able to greatly improve my knowledge and skill in the R programming language. My knowledge of gathering data, manipulating data, constructing complex loops, and creating informative data visualizations were just some of the many aspects of the language that I was able to add to my repertoire. I learned about the time-consuming nature of running thousands of simulations on a computer for hours on end, as well as ways to utilize additional computing power through the Parallel package[11]. My advisors taught me the convenience of using Git in order to directly link to an online repository, share code with colleagues, and, in general, control a project under a revision control system. Last and certainly not least, I learned the importance of debugging as I assisted my advisors (Dr. Elmore and Dr. Williams) with the early stages of the nflsimulator package. In doing so, I acted as a test user to find and help correct errors in a package made to handle complex problems and large amounts of data. Overall, the experience provided me an invaluable learning experience and exposure to the application of analytics in the sports industry.

5 ACKNOWLEDGEMENTS

I would like to thank my advisors Dr. Ryan Elmore and Dr. Ben Williams for their time and assistance in conducting this research. The two of them helped me at every turn from de-

bugging code to brainstorming fourth down strategies and everything in between. I am especially grateful that I took Dr. Elmore's sports analytics class which opened up my eyes to the industry. In working with these two, I found an area that combines two of my biggest interests.

6 EDITOR'S NOTES

This article was peer reviewed.

REFERENCES

[1] Burke, B. Fourth-down decisions changed for good 10 years ago: How the Patriots innovated (2019). URL https://www.espn.com/nfl/story/_/id/28073660/fourth-decisions-changed-good-10-years\protect\discretionary{\char\hyphenchar\font}{}{}ago-how-patriots-innovated.

[2] Burke, B. Defending Belichick's Fourth-Down Decision (2009). URL https://fifthdown.blogs.nytimes.com/2009/11/16/defending-belichicks-fourth-down-decision/.

[3] Romer, D. Do Firms Maximize? Evidence from Professional Football. *Journal of Political Economy* **114**, 340–365 (2006).

[4] Causey, T., Katz, J. & Quealy, K. A Better 4th Down Bot: Giving Analysis Before the Play (2015). URL https://www.nytimes.com/2015/10/02/upshot/a-better-4th-down-bot-giving-analysis\protect\discretionary{\char\hyphenchar\font}{}{}before-the-play.html.

[5] Yam, D. & Lopez, M. Quantifying the Causal Effects of Conservative Fourth Down Decision Making in the National Football League. *SSRN Electronic Journal* (2019).

[6] Urschel, J. & Zhuang, J. Are NFL Coaches Risk and Loss Averse? Evidence from Their Use of Kickoff Strategies. *Journal of Quantitative Analysis in Sports* **7** (2011).

[7] NYT 4th Down Bot. 4th Down: When to Go for It and Why (2014). URL https://www.nytimes.com/2014/09/05/upshot/4th-down-when-to-go-for-it-and-why.html.

[8] Elmore, R. & Williams, B. nflsimulator: Simulating NFL Drives using NFLScrapR Data (2020).

[9] R Core Team. R: A language and environment for statistical computing (2019).

[10] Horowitz, M., Ventura, S. & Yurko, R. nflscrapR: Compiling the NFL Play-by-Play API for easy use in R (2019).

[11] Microsoft Corporation & Weston, S. doParallel: Foreach Parallel Adaptor for the 'parallel' Package (2019). URL https://cran.r-project.org/package=doParallel.