NFL Success

Colin Seiler, Aidan Collins, Nayan Paliwal, Abdul Salam
Team 8 / EAS 508 / University at Buffalo

***Abstract*— This study examines whether play success can be predicted using only pre-snap information. Using play-by-play data from the 2016–2023 NFL seasons, we frame success prediction as a binary classification problem and evaluate Logistic Regression, Random Forest, Gradient Boosting, and a stacked ensemble model using features derived from game context, formations, and personnel. Performance improves substantially on third and fourth downs. Feature importance analysis identifies yards-to-go as the dominant predictor of success, with rushing plays predicted more reliably than passing plays.**

1 Introduction

Over the past two decades, the National Football League (NFL) has undergone a drastic transformation in how teams leverage data to gain a competitive edge. What was once driven almost entirely by intuition, film study, and experience has expanded into a landscape shaped by large-scale data collection and analytical methods. Teams now rely on comprehensive datasets that capture the structure and outcome of plays, drives, and games in new and exciting ways.

Within this newly data-rich environment, one major frontier that has proved useful is the identification of tendencies and mismatches, and how they relate to success. Offenses seek to exploit defensive weaknesses through things like personnel mismatches or formation shifts while defenses aim to counter through coverage disguises and forcing offenses out of their comfort zones. Data and modeling give us the ability to quantify these different tactics and understand how often they lead to a positive outcome for both the offense and defense. In turn, this understanding allows teams to adapt their strategy and planning, evolving the "chess-like" game that happens between every play.

In recent years, 'success' in football analytics has gained prominence as an alternative to raw yardage and other standard box-score stats. Rather than measuring how many yards a play obtains, 'success' is about whether a play achieves meaningful progress within a set of downs. A play is a 'success' if the offense gains 40% of the yards to go on first down, 60% of the yards to go on second down, and 100% of the yards to go on third and fourth. This framework, compared to raw stats, is much more context-aware and stems from the importance and emphasis of controlling the flow of the game while on offense, better aligning with how both coaches and analysts measure performance on the field.

In this project, we take a data-driven approach to classifying offensive plays as a 'success' from pre-snap information, as well as examine whether machine learning tools can offer value in their prediction and find characteristics that indicate positive offensive tendencies. This topic has emerged repeatedly in analytics, with analysts working with success rate on fourth down to help coaches better understand aggressiveness risk vs reward in fourth down scenarios. While analysts do good work pushing these bounds, few use machine learning tools in the manner we hope to do. We hope that using their ideas alongside our domain knowledge will help in our understanding within our models and provide new actionable insights for teams and the league alike.

2 Literature Survey

While predicting specific outcomes based on pre-snap information is a staple of NFL modeling, there isn't much published data in terms of predicting a success based only on pre-snap indicators. Our literature review will draw on studies that build the understanding of how success is influenced as well as draw upon other recent reports that are some of the foundational building blocks of the NFL data scene. We offer a brief report of a few of those papers below.

The Hidden Game of Football (1988), is the first instance of success being mentioned as a metric.

They developed it with the idea that a play should not be judged by raw yardage alone but by whether the offense's chances of scoring increases. The team introduced 'success' how we outlined it above to contextually quantify this. This idea directly influenced later analytics systems such as FootballOutsiders' 'success rate' and 'DVOA', which compares success vs the league average.

Another metric that builds on the idea of contextualizing success is Expected Points (EP) and its derivative Expected Points Added (EPA). EPA is another stat and an often-used metric to describe the value of a play by measuring how much a play increases or decreases a team's chance of scoring. Yurko, Ventura, and Horowitz (2018) was the first peer-reviewed article that built on historically established EPA methods to estimate how each play changes a team's expected points. The paper's broader goal was to evaluate player EPA. While this will not be relevant to our project, their process underlines important features for us to consider as well as the impact of certain players and positions on success.

A further extension of this is in Brill, Yee, Deshpande and Horwitz (2024). They took the prior EPA framework and tried to account for situations where variables might appear related due to an underlying factor or third variable, rather than a true link. This is important to acknowledge and understand within our work.

One issue with both papers is their reliance on play-by-play data only. In the following three research papers, Caldeira et al. (2023), Burke et al. (2020), and Koshida et al. (2021), live player location data is used in a variety of different ways. While the idea of using player location data is intriguing, we currently do not have the ability to get enough of that data for free. Thankfully NFLVerse does provide extracted data such as position counts and offensive formations that is built using the systems proposed in these papers.

A paper like these that may be more useful is Villar et al. (2023). It builds out full models using player location stats for a play. While we will not be using those specifically, we can build out a lot of similar models but with time instead cut at pre-snap and focused on Success Rate and EPA. Horn et al. (2023) also explores the impact of live movement stats on the success factor of a play. They found there was no statistical impact through their models. This is important for us, telling us that pre-snap predictors might be the most important factor in determining success of any given play.

In Sandholtz et al. (2023) they use Markov Decision Processes to model 4th down decision making. They show teams often act differently than a pure expectation model would predict due to game state. While their focus was on 4th downs, we can extend their process of probabilistic thinking and modeling to help predict Success Rate. On first and second down, teams are more likely to act unpredictably due to the 'openness' of a first or second vs a third or fourth. They also focused on the dangers of passive strategies, showing more aggressiveness would generally lead to more wins.

In Davis et al. (2024), the authors specifically examined evaluation challenges in sports analytics, so this will provide us a perfect framework for how to evaluate our model's success. The article can get very broad at points and covers a lot of sports, therefore we will need to adapt a lot of the concepts instead of applying them directly. There are some risks to our project, including overfitting of models leading to improper predictions as well as the potential noise from the unpredictable human element of football. We can monitor this and look to other papers such as Beal et al. (2020). While this paper was about predicting a game winner based off traditional stats, the successes of its different models may correlate to the success of ours given our shapes might be similar for Success Rate.

Finally, in Putnam and Tolhurst (2025) they found that generally teams do not pass as much as they should, saying teams should pass about 11% more often. They found score and position impact this more than anything else. We need to be conscious of their findings, as the conservative approach could affect our outcomes, especially since in our preliminary findings, passing plays are 20% easier to classify as a success or fail.

In all, these papers give us a good foundation for our problem. Our strategy will stem from these and a further paper by Lee, Chen and Lakshman (2022). This paper focuses on applying machine learning to predict pass or run based on prior formation. We can extend their system directly to ours.

3 Data

As stated previously, play-by-play data will be our main data source, containing year, game, and play information as well as plenty of information for each of those plays including personnel, formations, down, and distance. This data will also be extended to contain specific player or skill group information if it is required. Thankfully, due to the growth of NFL statistics and analytics, this play-by-play data is readily available from nflverse (nflverse.com). From this dataset we chose and developed the following features:

1. Current Quarter
2. Time Remaining
3. Current Down
4. Distance to First
5. Over/Under line
6. Home/Away
7. Score Difference
8. Count of Off Pos
9. Count of Def Pos
10. Formation Marker
11. Defenders in box

All from the years 2016-2023. We also hope to extend our dataset with further domain knowledge by considering the strengths of given teams. We will be incorporating Madden Player ratings for offensive positions as well as rate stats.

While Madden Ratings are not extremely accurate, they are a consistent, stable proxy for player evaluation. We must be careful about how we insert them, and how we use potential masking strategies. In the case that some player may be missing from the ratings system, a generic 50 rating will be filled. If that position is not used, a 0 will be filled instead. We plan to incorporate the following ratings from the beginning of the year as well as team stats from the previous season:

11. QB
12. RBs
13. WRs
14. TEs
15. OL Sack Rate
16. Rushing Yards
17. Passing Yards
18. Pass/Run Ratio
19. Success Rate
20. 3rd/4th Down Rate

4 Methods

Utilizing Python's `scikit-learn` library, we trained 3 separate models of varying bias and variance in an attempt to properly predict success outcomes based on the previously outlined features. We also used `optuna`, a hyperparameter optimization framework, to help hyperparameterize each model to its feature space. All 3 models were chosen for a few unique reasons.

4.1 Logistic Regression

While we consider our problem to be non-linear, a natural and often-used algorithm for binary classification is logistic regression. Despite its non-linearity, logistic regression can work well with noisy data that needs strong regularization for its noise or collinearity.

As for tuning, `optuna` found using no collinearity regularization was the best and we confirmed this through our testing as well. It also found a low C was best which strongly regularizes our coefficients. This is surprising given the assumed non-linear and highly correlated nature of our problem.

4.2 Random Forest

While a decision tree may be easy to understand and simple to use, a single one is not very good at making any sort of prediction on its own. To combat this issue we can use many trees. Random Forest is an ensemble algorithm, combining n different decision trees each trained independently on its own bootstrap sample. The averaging of these bootstrapped samples helps deal with collinearity and non-linearity.

Again with tuning, `optuna` evaluated that deep trees with a max depth of 10 was best with around 300 estimators being the sweet spot. This depth of 10 trees tells us more of what we know, that our data is highly non-linear and requires more variance in outputs to get the most correct outputs.

4.3 Gradient Boosting

Finally, Gradient Boosting is another tree ensemble method that uses boosting instead of bagging. Instead of building a system of independent models, Gradient Boosting builds a sequence of models. Each model built in the sequence learns the loss from previous models and adjusts its nodes and branches based on which variables can best reduce the overall loss. The predictions are then given again to the majority vote in the classification.

With tuning we see `optuna` focused on a bal-

anced split, giving us a max tree depth of 6 while with a learning rate of 0.44, allowing our trees to learn some over time but not overfit too strongly with our decent depth size.

4.4 Ensemble Models

After training and tuning the other models, we will also examine whether some weighted combination of predictions could be used to develop a better classifier. In our case, the linearity of Logistic Regression combined with Gradient Boosting or Random Forest non-linearity may allow us to find the best possible prediction process. We used the `StackingClassifier` from the same `scikit-learn` library. This classifier takes outputs from our other three classifiers and fits those outputs to some model, in our case logistic regression with 'l2' penalty, to determine a final output. To avoid data leakage, we've trained and tuned our initial three classifiers on 2016-2020, trained and tuned our stacked classifier on 2021-2022, and evaluated on 2023.

5 Error Analysis

When working with classification, the most common and simple method to resort to is to simply calculate *Accuracy*, the number of predictions that match the true cases. In our case, accuracy tells us some of the story but not all of it due to the slight imbalance in True and False values. Due to this imbalance, as well as trying to focus specifically on Type I error reduction (guessing success when not successful), we will focus heavily on Precision, Recall, and F Score. Formally, Precision (P), Recall (R) are defined as follows:

$$P = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$R = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

It should be clear that Precision focuses on how many Positive Predictions are correct out of all positives while Recall focuses on how many Real True values are correctly predicted. F score is a combination of those:

$$F = (1 + \beta^2) * \frac{\text{Precision} * \text{Recall}}{(\beta^2 * \text{Precision}) + \text{Recall}}$$

F1 is used to understand the balance between Precision and Recall. Since we are focusing more on Precision, we should choose some β where $\beta < 1$ to weight it more.

Finally, we will also use Area Under the Curve (AUC) of our ROC Curve. ROC plots the true positive rate vs false positive rate as the classification threshold varies. From this, AUC can be interpreted as the probability that a random sample with a success label of 1 will be assigned to that class. This will be our most important measurement for verifying that our model is understanding our data.

6 Results

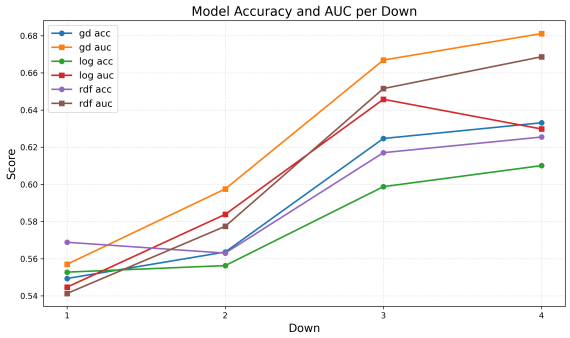
We will break our results section into the following two subsections.

6.1 All Down Results

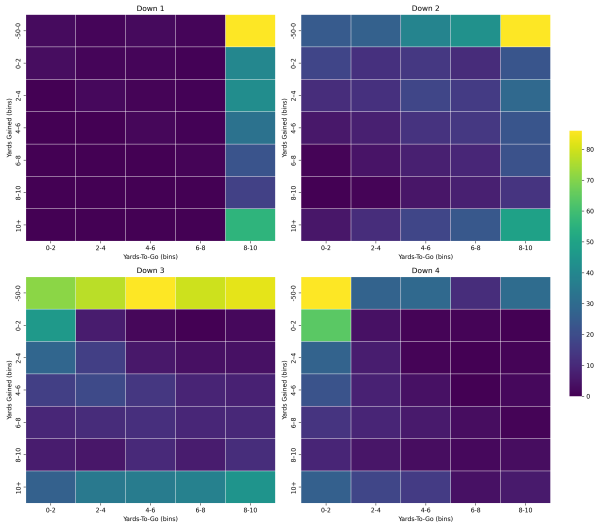
After training our algorithms, initial measurements and tests were done versus a simple baseline. On average, a given NFL play is a success about only 40% of the time, thus 60% accuracy would be our baseline if we guessed fail every time. We processed the models on all plays and produced the following test accuracy and AUC on the year 2023:

Model	**Accuracy**	**AUC**
Baseline	0.600	0.500
Logistic	0.560	0.586
RandForest	0.580	0.585
XGBoost	0.570	0.604

All of which do not meet the standard baseline. There is some information to be gleaned from these initial models, as their AUC is slightly above the baseline of 0.5, but overall they fail generally when measuring success over all downs. While these overall results are not what we desired, we do have some useful data that comes from our predictions.



Throughout our models, as we approach 3rd and 4th down we see a consistent theme, the outcomes become more and more predictable. When it comes to football this actually makes a lot of sense. While every play of every drive matters, third and fourth down are the last chances to get a first down in a set of downs. These downs have some meta strategy involved but generally have a set, distinct goal. First and second down in comparison are much more chaotic. First down is a chance to set up the remaining downs by playing for yards or by running it to set up play action. Second and short is often seen as an opportunity for aggressive offenses to call their shot or go for a more aggressive throw. In either scenario, with 2 more downs to work with, teams sometimes take chances or instead play to set up their final downs. Third and fourth down do not give teams the chance to be different. It is do or die.



The previous graph is the perfect example of this. First down usually starts from 1st and 10 barring a penalty, but despite that you see the density of plays flip from being above the diagonal to below it. This signifies that teams understand the situation they are in on 3rd and 4th downs and how it affects their play calls and decision making. Because of this more rigid definition of a goal, teams will then have more rigid approaches to reaching this goal.

Two more things to point out is located in second and third downs density locations. Second down has less gains of ≤ 0 as the yards to go decreases, signifying a higher chance to run and gain at least a yard or 2. Meanwhile third down has the highest densities to gain ≤ 0 yards by a long shot. This increase can be explained by teams leaning on passing on third down. These passing plays often result in incomplete passes or sacks meaning no or negative gains.

6.2 Final Down Results

Digging into our other results, let's instead focus specifically on third and fourth downs and call these downs 'final' downs. For our final down efforts we will go through our whole process again, training only on 3rd and 4th down and tuning our models using `optuna`.

Our results were much much better. Please note that because of our change in down, our baseline changes as well. We've added Precision for these models as our tuning focused on improving this and AUC specifically.

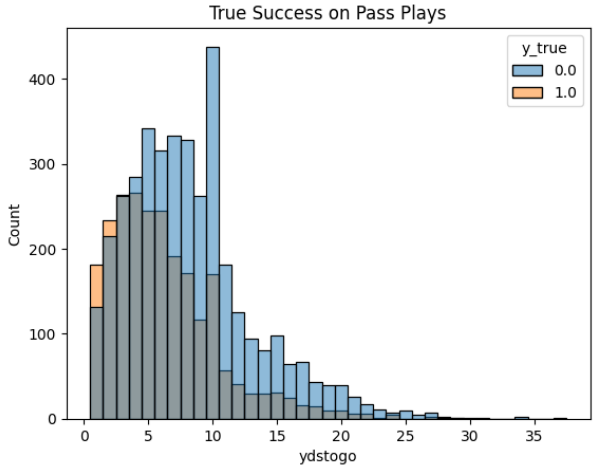
Model	Precision	Accuracy	AUC
Baseline	0.000	0.572	0.500
Logistic	0.600	0.636	0.668
RandForest	0.570	0.634	0.671
XGBoost	0.560	0.625	0.670
Ensemble	0.610	0.636	0.670

Our improved results indicate that our models and feature set possess some meaningful predictive power, giving us some insight into what exactly goes into successful third and fourth down plays. Across all three algorithms we see *ydstogo* emerges repeatedly as a dominant predictor of play success. This aligns with our domain knowledge and intuition, validating that the models are indeed capturing the core structural relationships inherent to our system. However, this single variable dominating our predictions suggests the need for more rich features or additional on-field and live context. Our current additional features contribute to some

degree, but are strongly overruled, indicating plenty of room for improvement.

Feature	XGBoost	RandForest
ydstogo	0.149	0.347
under_center	0.063	0.004
shotgun	0.040	0.002
ydstotd	0.026	0.067
bet_line	0.025	0.043
QB1	0.024	0.030
defenders_in_box	0.024	0.013
WR1	0.024	0.025
WR3	0.024	0.021
time_remaining	0.023	0.072

With *ydstogo* dominating our features, we broke up our data into run and pass plays to see how it handled each split. Rushing plays we have increased precision and recall while passing plays are not determined well at all by our system. This further aligns with our domain knowledge given our dominating feature. Passing plays have a greater chance of gaining large amounts of yardage but a chance of being incomplete making it a dud, these low yardage situations may not predict those scenarios very well then. Below is a graph of all passing plays true success vs failures.



Again this is all reflected in our precision and recall splits for each run and pass. Increasing precision and recall among passing plays would be the best opportunity to extend and perfect our work, where our precision and recall among running plays is actually quite impressive for NFL data.

Play Type	Precision	Recall
Run	0.650	0.805
Pass	0.510	0.480

7 Conclusion

In our project, we applied multiple machine learning techniques to predict play success in the NFL using a combination of rich datasets. Through data cleaning and manipulation, as well as training 4 different machine learning models, we were able to build a predictive model that provided insight into the different pre-snap features that matter most regarding 3rd and 4th down success. While our Ensemble method worked the best, our Logistic model working best out of all models was a surprise despite its linear decision boundary and was most likely caused as a result of the strength of *ydstogo*.

In the future there are a few distinct ways we would first like to extend our work. For starters, we would like to stack a model with our current system, predicting a play as a run or pass and adding that prediction as a new feature in our system. We believe the systems inability to predict success on pass and run plays could be greatly improved with this addition.

Instead of using predicted run/pass as another predictor, we may also choose to build separate models for both pass and run. Our hope was to build a system that relied only on pre-snap data that could be used in real time by offenses and defenses to provide insights into the best approaches in each situation. Splitting across run/pass would give some leakage, but would still allow our model to be useful, giving probabilities of success across different play types. We can also do further extensions of this split idea, classifying different types of passing or running plays such as screens, go plays, rub routes, draw plays, or counters.

In either effort, building out and documenting the system at each iteration mentioned would allow teams to understand broader ideas at each level as well as dig in in-depth. Using these systems, we would be interested in building out a full application that coaches and teams could use during live games, accounting for both teams, trying to understand their strengths and weaknesses and suggest a plan of attack.

References

Ryan Beal, Timothy J Norman, and Sarvapali D Ramchurn. A critical comparison of machine learning classifiers to predict match outcomes in the nfl. https://rb.gy/wxwgnf, 2021.

Matthew Brill, Stephanie Yee, Arjun Deshpande, and Benjamin Wyner. Moving from machine learning to statistics: The case of expected points in american football. *arXiv preprint arXiv:2409.04889*, 2024.

Brandon Burke et al. Going deep: Models for continuous-time within-play valuation of game outcomes in american football with tracking data. *Journal of Quantitative Analysis in Sports*, 16(2):163–182, 2020.

Nelson Caldeira, Rui J Lopes, Dinis Fernandes, and Duarte Araujo. From optical tracking to tactical performance via voronoi diagrams: Team formation and players' roles constrain interpersonal linkages in high-level football. *Sensors*, 23(1):273, 2023.

Jesse Davis et al. Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned. *Machine Learning*, 113(9):6977–7010, 2024.

Hayley Horn, Eric Laigaie, Alexander Lopez, and Shravan Reddy. Using geographic information to explore player-specific movement and its effects on play success in the nfl. *SMU Data Science Review*, 7(2), 2023.

Shinji Koshida and Kalle Kytölä. The quantum group dual of the first-row subcategory for the generic virasoro voa. *arXiv preprint arXiv:2105.13839*, 2021.

Peter Lee, Ryan Chen, and Vihan Lakshman. Predicting offensive play types in the nfl. Stanford CS229 Final Project, 2016.

Daniel S Putman and Tor N Tolhurst. Tackling endogeneity: Estimating optimal pass rate in the nfl using instrumental variables. Working Paper, 2025. SSRN: https://ssrn.com/abstract=5159158.

Nathan Sandholtz, Lucas Wu, Martin L Puterman, and Timothy CY Chan. Learning risk preferences in markov decision processes: An application to the fourth-down decision in the nfl. *arXiv preprint arXiv:2309.00756*, 2023.

Ximena Villar et al. Article title. *Journal of Sports Analytics*, 2023.

Derrick R Yama and Michael J Lopez. Modeling field goal success in the national football league. *Journal of Sports Analytics*, 2020.

Joseph P Yurko, Samuel L Ventura, and Maksim Horowitz. nflwar: A reproducible method for offensive player evaluation in football. *arXiv preprint arXiv:1802.00998*, 2018.