American Scholarly Journal for Scientific Research (ASJSR)
https://asjsr.org/ Volume 2026, Issue 1004 ISSN: 3143-2999

How AI Predicts NFL Plays

Author: Scott Reynolds
Institution: Naperville Central High School, Illinois
Mentor: Mike Stineburg

ABSTRACT

Before an NFL play begins, a defense has to prepare for several possibilities. Two of the most important are a run and a pass. This paper asks how well artificial intelligence can predict which of these two choices an NFL offense will make, and how a student could test that prediction fairly. It reviews published football research, public play records, and an NFL example that uses player tracking. It then proposes an experiment that would compare a historical pass rate, a simple prediction rule, and two machine learning models. The experiment would train on older NFL seasons and reserve a later season for the final test. Published research shows that a model can learn useful patterns from game situations. The evidence also shows why a single accuracy number leaves important questions unanswered. A model can miss many runs, make overconfident predictions, or lose accuracy when a team changes its offense. A fictional example explains how to measure some of these problems. The paper concludes that a useful model should give a clear probability, beat reasonable baselines on future games, and show where it makes its biggest mistakes.

Keywords: NFL, artificial intelligence, football, play prediction, machine learning, sports analytics

1. INTRODUCTION

Imagine watching the Chicago Bears from Naperville on a Sunday afternoon. The offense faces third down and needs eight yards. Two receivers line up on each side of the field, and the quarterback stands several yards behind the center. A viewer might expect a pass because the offense needs to cover a lot of ground in one play. But the offense could hand the ball to a running back, especially if the defense has spread out to cover the receivers. The same clues that make a pass seem likely can help create room for a run. This imagined Bears play is the kind of problem a prediction model would have to deal with over and over.

An NFL offense usually has four downs, or chances, to gain the yards needed for a new set of downs. On early downs a team can spread its plan across several plays. On third down the need to keep the drive going can narrow its choices. The score and the clock matter too. When a team trails late, the offense may need to gain yards quickly. When it leads, using time may be almost as valuable as moving the ball. The value of each clue depends on the others. Third and two with a ten-point lead feels different from third and two with a ten-point deficit and forty seconds to play.

In this paper, artificial intelligence refers to machine learning. A machine learning model uses examples to learn how inputs such as down, distance, and score relate to an outcome. A student supplies the examples and defines what counts as a correct answer. The model then uses patterns it learned to estimate what might happen in a different game. A baseline is a simple method used for comparison. If the model and the baseline see the same plays under the same test conditions, a student can measure how much the more complicated approach adds.

The main question is whether a model using information available before the snap could predict run or pass better than simple baselines on a later NFL season. Two smaller questions follow. Do the predictions improve when the model can learn that the effect of distance changes with the down or the score? And do predicted probabilities match what actually happens over many plays? Before answering either question, it helps to define what a prediction covers. The target in this paper is a broad type of offensive play. An exact route, the side of the field the ball will go to, and the chance that the play succeeds are separate questions. Each would require its own definitions, data, and tests.

2. REVIEW METHOD

This is a narrative review with a proposed experiment. A narrative review brings together selected sources to explain a question and compare evidence. The sources were chosen for three jobs. Research on play prediction provides examples of what models have achieved. Public documentation explains which game details a student can download and what those details mean. NFL tracking examples and machine learning documentation help explain how the experiment could be extended and how probabilities could be judged. The review emphasizes sources that describe their data or methods clearly enough for a reader to see where a claim comes from.

The most direct evidence comes from Fernandes and colleagues (2020), whose paper compares methods for predicting NFL runs and passes. A Stanford course report by Lee and colleagues (2016) offers a second example of how students approached offensive play types. Yurko and colleagues (2019) studied player evaluation rather than run and pass prediction, but their work shows how a public football record can be turned into a statistical analysis that other people can repeat. The play-by-play documentation from nflverse is used to design a possible new dataset. Published accuracy values belong to their original studies. The numbers in the worked example later in the paper come from an invented set of plays.

The important comparison across sources is the information each method can see, the time at which it sees that information, and the group of plays on which the prediction is scored. For example, a model built for all eligible offensive snaps could face many ordinary first downs. A model tested only on late third downs would face a different mix of runs and passes. A tracking model that sees a receiver move before the snap also has a different job from one that has to make a prediction before the offense lines up. These differences shape what each study can establish.

3. WHAT THE MODEL WOULD PREDICT

The word “call” sounds simple when a coach uses it, but it can be hard to recover from a public game record. The coach might send in a play that offers both a handoff and a short throw. The quarterback may make the final choice after watching a defender. The offense can also change the play at the line. A public record tells the reader what the offense did and sometimes helps explain how the play began. It usually cannot tell the reader what was said in the huddle. The closest target for a student project is therefore whether the recorded action was a dropback or a designed run.

A dropback covers plays that begin as attempts to pass, including throws, sacks, and quarterback scrambles that begin as passing plays. A sack ends with the quarterback being tackled behind the line before throwing. A scramble happens when the quarterback moves and ends up running, often after a pass opportunity closes. Treating a scramble as a designed run would mix two different starting actions. A planned quarterback keeper, on the other hand, belongs with designed runs when the record supports that label. The project would use the nflverse dropback

indicator and check disagreements against its play-type and scramble fields. Unclear examples deserve a written rule and a count so another student can repeat the decision (nflverse, n.d.-a). The study would use regular-season offensive snaps with a valid down from one through four. It would leave out kicks, punts, extra points, two-point tries, kneels, spikes, and plays marked as no play. Kneeling often uses time while protecting a lead, and spiking is usually done to stop the clock. These actions could be interesting targets in another project. Excluding them makes this study easier to understand and keeps easy clock-management cases from dominating parts of the test. Penalty records need extra care. A flag may erase a play or leave an official action on the record. In the proposed study, a penalty play can stay only if the official record retains a valid action and the run or dropback classification is clear.

The prediction would be made after the game situation is known and before the snap. For the basic model, the student would use only the scoreboard and play record fields that describe that moment. Player movement and formation would be saved for a possible follow-up. It also matters that the model is being tested on run or dropback snaps already identified after the game. In live use, a fourth-down model would have to consider a punt or field goal as well. The paper's prediction is best read as a conditional question. If the offense runs a regular play here, what is the probability that it begins as a pass?

There is a useful difference between a play type and a play result. A screen pass can gain almost no yards. A quarterback scramble can gain twenty. Predicting the play type correctly says nothing on its own about the yardage or whether the defense stopped it. The same is true for a suggested defensive response. Suppose a model forecasts a dropback and a coach puts extra defenders in coverage. The resulting play would unfold against that new defense. Past game records show the outcome of the defense the team actually used. Estimating what would happen under a different defensive call would take a separate study.

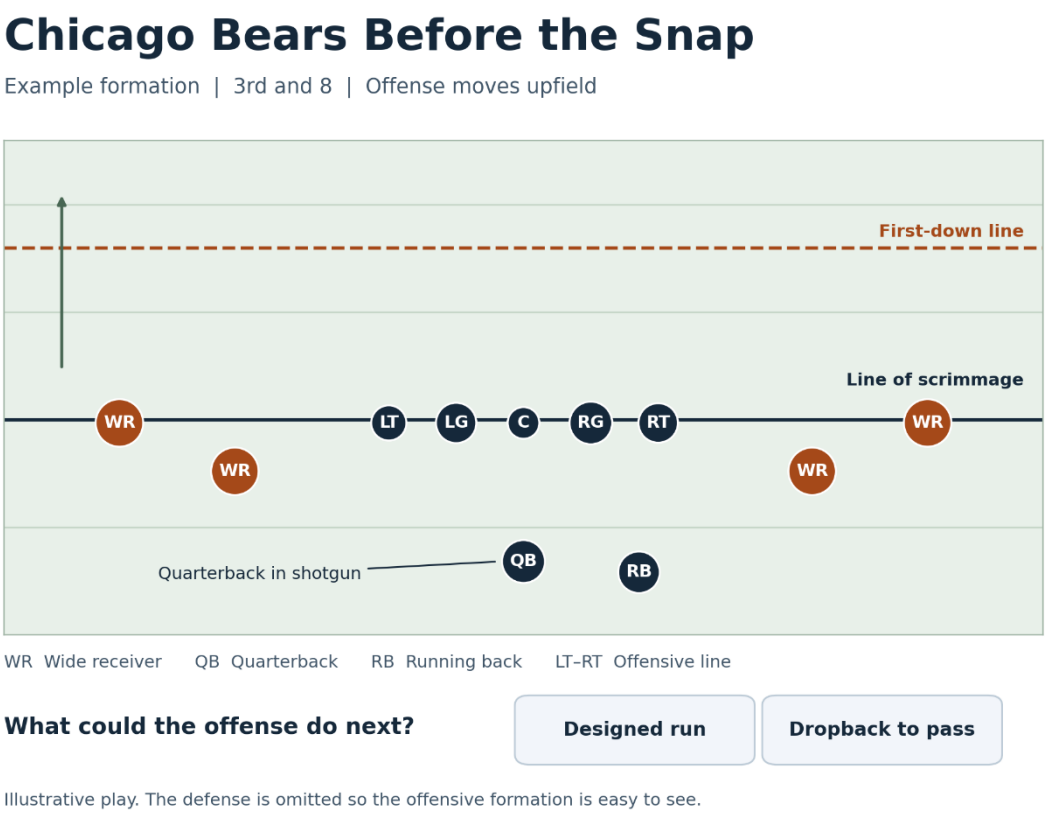

Figure 1. Chicago Bears before the snap. *In this fictional third-and-eight situation, the quarterback lines up in shotgun with two receivers on each side. From this formation, the offense could run or drop back to pass.*

4. WHAT EARLIER RESEARCH SHOWS

Predicting With Game Information

Fernandes and colleagues (2020) studied more than 130,000 regular-season pass and rush plays from four NFL seasons beginning in 2013. They compared several methods, including decision trees, random forests, and a neural network. A neural network is a model made of connected layers that learn relationships between inputs and an output. Their best reported neural network accuracy was 75.3 percent. A decision tree designed to be easier to follow used down, yards to go, and score difference and reached 65.3 percent overall in the authors' evaluation. Those results give a starting point for thinking about the value of complex models and the appeal of simpler ones.

The smaller tree is interesting because a coach or student can inspect how it reaches a prediction. A decision tree divides plays into groups by asking questions in order. One branch might separate short distances from long ones, then divide each group by down. At the end of a branch, the model can estimate a pass probability from the plays that landed there during training. A viewer can trace the questions on a diagram. A neural network usually draws on many internal calculations at once, so the reader generally needs separate checks to learn which clues affect its predictions.

The same paper reported different accuracy levels across game situations. Its simple model did better in some situations that offered clearer clues, such as later downs or long distances. This matters because an overall result gives frequent situations more influence. Consider a defense whose biggest problem is giving up passes on third and short. The defense may care more about that small group than about another five hundred correctly predicted first downs. A researcher can use the overall result as a summary and then open it up by down and distance to see how it applies to particular football questions (Fernandes et al., 2020).

Lee and colleagues (2016) also explored how machine learning could forecast the type of offensive play. Their Stanford course project is useful as an example of students framing a football problem around recorded inputs and comparing model choices. The course report combined two models and reported 75.9 percent test accuracy using its own setup. It used information about offensive formations, which is outside the first version proposed here. The two studies therefore give examples of what prediction can achieve under different conditions. The strongest reason to include it is to show that the project can begin with an ordinary question a fan understands and then turn that question into a problem a computer can score.

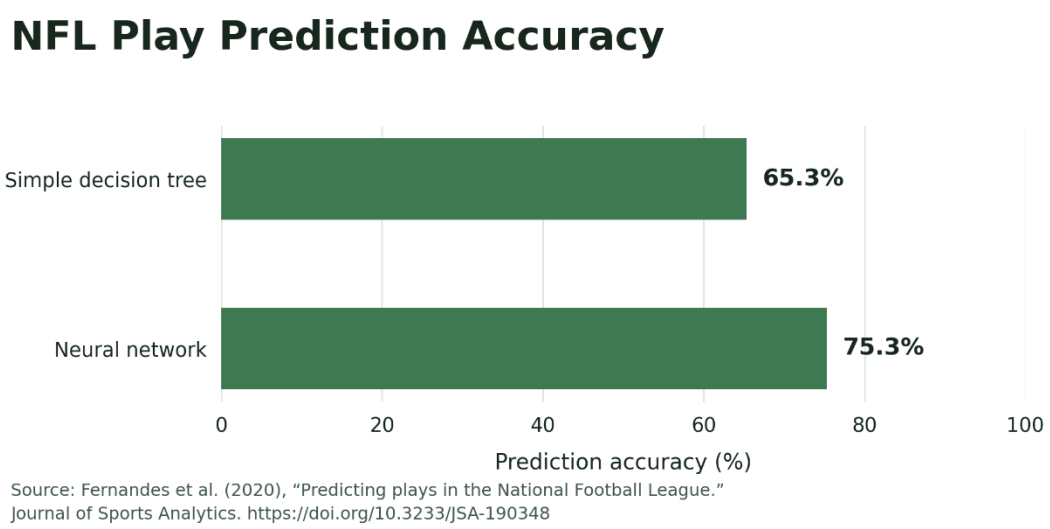

Figure 2. Published NFL play prediction accuracy. *Fernandes et al. (2020) reported 65.3% accuracy for the simple decision tree and 75.3% for the neural network.*
Source: https://journals.sagepub.com/doi/10.3233/JSA-190348

From Play Records To Player Movement

Public play records are useful because they organize a game into events and give each event a situation. The nflverse play-by-play data include fields that describe possession, down, yards to go, score, and location on the field. That makes it possible to compare situations across seasons even when the viewer does not have video for every play. Reproducible sports research such as the work by Yurko and colleagues (2019) shows the advantage of explaining how the raw record became the data used in an analysis. A reader can then tell which conclusions depend on an author's choices about the data.

Player tracking can add clues a scoreboard cannot capture. It describes where players are on the field at different moments. The NFL's 2025 Big Data Bowl asked participants to examine patterns before the snap using tracking data (NFL Football Operations, n.d.). The NFL later described a model that estimates defensive coverage before the snap and compares it with what happens after the snap to identify possible disguises (NFL Next Gen Stats, 2025). That example concerns defensive coverage. For this paper, its value is that it shows how a series of positions can describe what information people on the field have available at a given point in time.

Imagine that a receiver moves across the formation as the quarterback waits for the snap. A tracking model could use the receiver's starting position, the direction of movement, and how

much time passed before the ball was snapped. It might find that the same type of motion is followed by different play choices for different teams. The motion can also be used as a decoy, so it should not be interpreted as proof of a call. A fair test of the value of movement would compare the same plays with and without the tracking inputs. If the models see different games, any difference in scores could come from the games themselves.

5. HOW A STUDENT COULD TEST THE IDEA

Build A Dataset That Follows Game Time

A student could start with six completed NFL seasons available through nflverse. The proposed split is 2019 through 2022 for training, 2023 for validation, and 2024 for the final test. Training is where the model learns its patterns. Validation is where the student compares settings and choices. Testing is the final evaluation on data kept aside while those choices are made. Using time in this order is close to what a coach or analyst would face. Work on old games first, choose the model with a more recent season, and see how well it holds up when the next season arrives (nflverse, n.d.-b).

Putting all plays in a pile and randomly dividing individual rows would answer a less realistic question. Plays from the same game share teams, conditions, and plans. With a random split, the training set could contain many plays from a game whose other plays appear in the test set. It could also contain games from after the date of a test prediction. Either choice makes it harder to tell how the model would perform when given an upcoming game. The proposed test holds out an entire season. This lets the researcher report which weeks were easier or harder while keeping all plays from any one game together.

For the first version, inputs would include the down, yards to go, distance from the opponent's goal line, quarter, seconds remaining, and score difference from the offense's point of view. Using one point of view matters. If the Bears lead by seven with the ball, the score difference is positive seven. If the opponent has the ball in that same game state, its score difference is negative seven. A model given a home-team score difference would have to figure out which side has the ball before using the number. Standardizing the field lets one value mean the same thing in every game.

Monday's full game summary might be convenient, but many of its fields would give away information that was unavailable on Sunday. For example, current play yardage can provide a clue about whether a pass was completed. The result of a drive can tell the model about what happened several plays after the one being predicted. Even a team's season pass rate would leak future information if it used games from later in that same season. This problem is called data leakage. To prevent it, every input needs an answer to a short question. Could someone have known this value when the offense was getting ready to snap the ball?

A second model could use recent team tendencies. A practical version would add the offense's dropback rate over its last four completed regular-season games and the number of eligible plays behind the rate. The count tells the model how much evidence supports the percentage. For early weeks with fewer than four prior games in the same season, the rate could use the games available so far. When no earlier game is available, a league rate estimated from training data could take its place. The student should keep that rule consistent from one season to the next and decide it before checking the final test.

To make the project repeatable, the student would record the version or download date of the dataset, the list of fields, and the number of rows removed by each filter. Each game and play ID should belong to one row in the study, and every row should have an eligible class label. Missing values should be summarized by season before choosing how to handle them. A student could also inspect a small group of boundary cases, including sacks, scrambles, and plays with penalties. That check could catch an error before thousands of rows are labeled the same way.

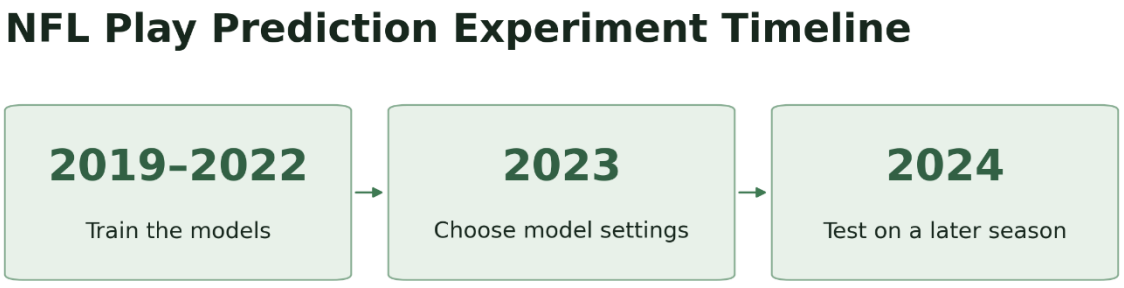

Figure 3. Proposed experiment timeline. *The models would train on games from 2019–2022, use the 2023 season to choose model settings, and test those settings on the 2024 season.*

Compare Simple And More Flexible Approaches

The first baseline would apply the overall training-season dropback rate to every play. If the rate were 58 percent, the baseline would give a 58 percent dropback probability throughout the test. At a threshold of 50 percent, where predictions at or above the threshold become pass guesses, it would choose pass every time. That makes the baseline simple to beat on some measures, but still useful. It tells the reader how high raw accuracy can be when a prediction never uses down, distance, or score. The rate in this example is hypothetical. The experiment would calculate its own from the selected seasons.

A second baseline would track dropback rates by down and yards to go. The student could sort distance into one to two yards, three to six yards, and seven or more yards. For each combination, the method would calculate the share of dropbacks in the training data. If a group were too small, the method could use the training rate for that down across all distances. Before validation, the student would set a minimum group size, such as 50 plays. This baseline reflects the kind of situation-based judgment a fan might already use, while putting an exact value on each prediction.

The first machine learning model would be logistic regression. It combines features and turns the result into a probability between zero and one. It can account for more information than the baseline table, including the score and the clock. A student might also give the model a term that combines fourth-quarter time and the offense's score difference. That term allows a seven-point lead to affect the prediction differently as time runs low. The proposed model would represent each down separately so that it does not assume third down is simply one unit higher than second in its effect on play choice.

A random forest would be the more flexible model. It builds many decision trees with some variation in the training examples and fields each tree sees. The forest averages their predictions. This can help capture patterns such as how late-game time, down, and score work together. It can also learn patterns that happen to exist in the older games and then fade. To limit this problem, the student would compare only a small number of settings on the validation season. For instance, the student could try several maximum tree depths and choose the setting with the lowest validation Brier score, a measure explained in the next section.

Every model would predict the same validation and test plays. The student would compare a version that uses only the game situation with a version that also uses recent team rates. If the

second version improved on the first, the test would show whether the recent rate helped for that particular period and setup. To see whether a gain is stable, the student could report results by week and use a confidence interval based on resampling whole games. Resampling means drawing games from the test set with replacement and recomputing the score difference. Keeping the plays from each game together respects their shared context.

6. HOW TO JUDGE THE PREDICTIONS

Accuracy is the number of correct guesses divided by the number of plays. It makes sense to start there because it is easy to calculate and explain. A second step is to count the kinds of mistakes. A confusion matrix is a small table with actual play types on one side and predicted play types on the other. It shows correct pass guesses, correct run guesses, runs mistaken for passes, and passes mistaken for runs. A model can look good overall and still fail at calling most of the runs. Comparing all four groups makes that weakness visible.

Consider an imaginary test with 100 plays. There are 60 dropbacks and 40 designed runs. An always-pass baseline correctly predicts 60 plays and misses all 40 runs. A sample model predicts pass on 48 of the 60 dropbacks and on 12 of the 40 runs. It predicts run on the other 40 plays, getting 28 of them right. Its accuracy is 76 percent, compared with 60 percent for the baseline. Its dropback recall, or the share of all actual dropbacks it identified, is 48 out of 60, or 80 percent. Its run recall is 28 out of 40, or 70 percent. In a real study the matrix would show counts as well as percentages so the reader could see the size of each group.

A probability gives the reader information beyond the final guess. A model that says 51 percent pass and another that says 95 percent pass would both pick pass at a 50 percent cutoff. They express very different levels of confidence. Calibration measures whether confidence matches observed frequency. If a model predicts close to 70 percent on many plays, a well-calibrated model should see a pass on roughly 70 percent of those plays. A calibration chart can group predictions into ranges and plot each group's average prediction against its actual pass rate. Reporting the number of plays in each range helps readers judge how much to trust each point (scikit-learn developers, n.d.).

The Brier score takes the difference between a probability and the actual outcome, squares that difference, and averages the squared values. The actual value is one for a dropback and zero for a

run. A prediction of 0.80 for an actual dropback receives an error of 0.04 because the difference is 0.20. The same prediction for an actual run receives an error of 0.64. Lower average error is better. If the imaginary baseline predicted 0.60 for every play and the actual share were 60 out of 100, its Brier score would be 0.24. The 76 percent model's Brier score depends on the probabilities behind its guesses. The four counts alone cannot determine it.

The choice of cutoff can depend on the goal. A shift from a 50 percent to a 40 percent cutoff would make the same model guess pass more often. That could catch more actual pass plays at the cost of calling more runs as passes. The student could compare several cutoffs in validation and report the tradeoff. The final test would then use the chosen cutoff without tuning it again on the test season. For a school project, a 50 percent cutoff is a clear starting point because it chooses whichever of the two plays the model thinks is more likely.

A result should also show where the model is useful. The student could report accuracy and Brier score for early downs, third and short, third and long, and late fourth-quarter plays. The Chicago Bears would make a good local example to illustrate the results after the league-wide test. The Bears example would need a play count and would describe how the model performed for one team in one season. A better score for one team might reflect its style, its opponents, its game situations, or some combination of these. A small set of unusual fourth downs could produce large changes in percentage from just a few corrected guesses.

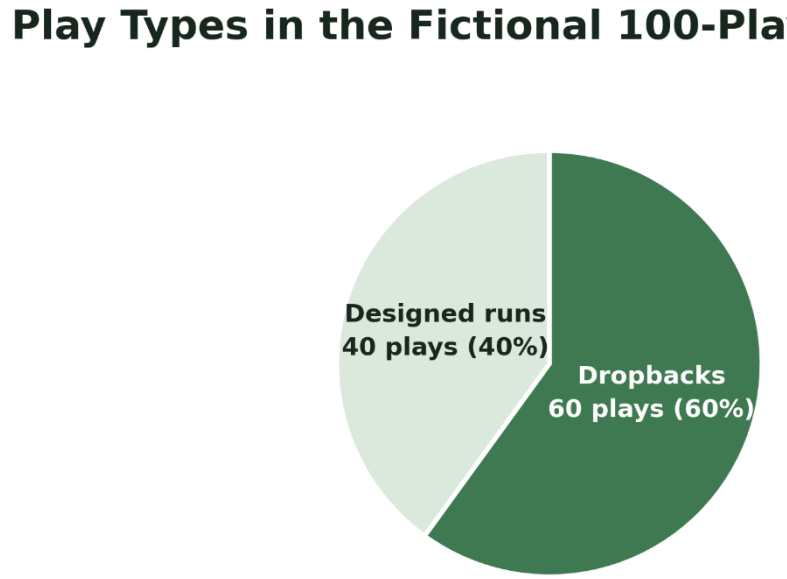

Figure 4. Play types in the fictional 100-play example. *The sample contains 60 dropbacks (60%) and 40 designed runs (40%)*

A Closer Look At Two Strong Guesses

The fictional example can be extended to show why two models with the same accuracy might be ranked differently. Suppose the sample model assigns 0.75 to every play it calls a dropback and 0.25 to every play it calls a run. For the 76 plays it gets right, the squared error is 0.0625 per play. For the 24 plays it gets wrong, the error is 0.5625 per play. The average is 0.1825. That is lower than the baseline's 0.24, so this particular model improves on both the accuracy of the labels and the quality of the probabilities.

Now imagine that a second model makes exactly the same guesses but uses 0.99 and 0.01 as its probabilities. Its accuracy is still 76 percent. Its Brier score rises to 0.2353 because each confident mistake carries a much larger error. The fictional test also makes it easy to see the issue without a formula. Of the 60 plays called passes, 48 actually began as passes, which is 80 percent. Of the 40 plays called runs, 12 began as passes, which is 30 percent. The first set of probabilities is closer to those observed rates than the second.

This example matters in games where the offense has several good options. An analyst might report, for instance, that a situation favors a pass but that the defense still needs a plan for a run. It would make less sense to express the same judgment with near certainty. In a larger study, a group of predictions near 0.99 could be tiny, so its actual pass rate would need to be read beside the group size. A good report makes it possible to judge both whether the model picks the more common choice and whether its confidence deserves to be trusted.

7. DISCUSSION AND LIMITATIONS

Football gives a prediction model patterns to learn because an offense has to respond to the same basic problems many times. Down, distance, time, and score describe part of those problems. Still, each offense chooses how to solve them. A coach can build a running game out of formations that other teams use to throw. A new quarterback might lead to more designed runs. A team may open a game with a plan that differs from its later approach. A model trained on older games can partly account for familiar patterns, but it has to be tested each time the people or strategies behind those patterns change.

An analyst could use a good model as part of film study. Suppose an opponent passes on several short-yardage downs in a game, even though league-wide predictions say those situations are close to even. A coach reviewing film could look at formations, defensive alignments, and the

order of the plays to see if the passes had anything in common. A season-wide average might have hidden that sequence. On the other hand, a few plays from one game could be coincidence. The model can help someone choose what to look at and the film can provide details that the data table leaves out.

A later version could test three prediction times on an identical set of tracking plays. The first would use the game situation when the offense begins to line up. The second would add the formation visible once the players are set. The third would add motion observed just before the snap. For each version, a student would need to record the last frame allowed into the model. Any movement after the snap would introduce information about the play's execution. By comparing the three versions one step at a time, the student could see how much earlier a useful guess was possible and whether waiting for motion improved the forecast enough to matter.

There are also limits to what this review can conclude. The published studies use specific samples and definitions, and their accuracy figures cannot be carried directly into the proposed test. Public play-by-play records describe actions that can be observed or inferred after the game, so some labels will remain uncertain. Changing the label rules could change the results. An additional test could remove all scrambles and sacks, then repeat the evaluation to see how sensitive the findings are to those choices. The narrower test would have its own question because it would focus more closely on plays that ended with a throw or a designed run.

One other limit comes from treating the opposing defense as part of the background. Even with the score and distance held steady, the offense may respond to which defenders are on the field and where they line up. On second and five, a team might run when the defense leaves room near the middle and throw when it packs defenders close to the ball. If a model lacks that information, some errors are reasonable given what it can see. Reviewing a few missed plays could help separate patterns the student could measure in a follow-up from details the public record simply cannot provide.

The best result would combine useful probabilities with an explanation of where they work. If the random forest beats logistic regression by a little on average but consistently underestimates runs on short yardage, the difference would matter to a coach who has to prepare for those runs. If a simple model does almost as well and its errors are easy to study, it could be the better tool for a first project. Another season would make it possible to see which result holds up. That kind

of comparison is more informative than choosing a model from one overall number and leaving the story there.

NFL Play Prediction Results

Actual play	Predicted dropback	Predicted run	Total
Dropback	**48**	12	**60**
Designed run	12	**28**	**40**
Total	**60**	**40**	**100**

Accuracy 76% (48 correct dropbacks + 28 correct runs out of 100 plays)

Fictional 100-play example. Shaded green cells show correct predictions.

Table 1. Prediction results for the fictional 100-play example. *The model correctly predicted 48 dropbacks and 28 designed runs. It misclassified 12 plays of each type, giving it an overall accuracy of 76%.*

8. CONCLUSION

AI can help predict whether an NFL offense is likely to run the ball or drop back to pass. Published research shows that ordinary game details contain useful clues and that different models trade simplicity for flexibility. A student can study the same question with public NFL records by defining the target carefully, checking what information was available before the snap, and comparing several methods on the same future season. Looking at both accuracy and probability quality makes it easier to see what the model actually learned. The most useful outcome would be a forecast that improves on what a fan could guess from down and distance, stays dependable on later games, and gives a clear account of the kinds of plays it still finds hard to read.

REFERENCES

Fernandes, C. J., Yakubov, R., Li, Y., Prasad, A. K., & Chan, T. C. Y. (2020). Predicting plays in the National Football League. *Journal of Sports Analytics, 6*(1), 37–48. https://doi.org/10.3233/JSA-190348

Lee, P., Chen, R., & Lakshman, V. (2016). *Predicting offensive play types in the National Football League*. Stanford University CS 229 final project report. View source

NFL Football Operations. (n.d.). *Big Data Bowl*. https://operations.nfl.com/platforms-initiatives/innovation/big-data-bowl

NFL Next Gen Stats. (2025). *Next Gen Stats: New advanced metrics you need to know for the 2025 NFL season*. NFL.com. View source

nflverse. (n.d.-a). *Data dictionary: Play by play*. nflreadr. https://nflreadr.nflverse.com/articles/dictionary_pbp.html

nflverse. (n.d.-b). *Load play by play*. nflreadr. https://nflreadr.nflverse.com/reference/load_pbp.html

scikit-learn developers. (n.d.). *Probability calibration of classifiers*. scikit-learn documentation. View source

Yurko, R., Ventura, S., & Horowitz, M. (2019). nflWAR: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports, 15*(3), 163–183. https://doi.org/10.1515/jqas-2018-0010