

Mind the Gap: Using Pre-Snap Data for Insights Into the Run Game

Josh Moore, Michael Rom, and
Liz Giel

December 2024




Welcome to our presentation for the Big Data Bowl 2025. We have been tasked with using data from Next Gen Stats powered by Amazon. Since we elected to pursue the coaching track, we aim to create a scouting report that will help teams gain an edge when game-planning. We take a dive into the run game and what tendencies and predictions can be obtained from pre-snap player tracking data. Through the data science process, we will show which features are most important; particularly, whether running back depth is significant in predicting a play type.

Background



The Big Data Bowl (BDB) is in its seventh year of competition. With the advancement of data analytics and its availability, more and more focus is being placed on developing insights from that data. This year the BDB has challenged competitors to develop actionable insight from pre-snap player tracking data. When one watches a game, there will be times where it is fairly obvious what the offense is about to do. Think about the Philadelphia “Tush-Push” or the ball is at the 50 and time is about to expire, and the offense needs a Hail Mary Miracle. However, in the normal progress of a game, it may not be as obvious, and this is where the real insight begins.



Purpose of Analysis

- Run or Pass
- Pre-Snap Tracking Data
- Important Variables
- Predictability of running back depth

Our goal is to predict whether a play will be a run or pass by analyzing pre-snap player tracking data that includes formations, motion, shifts, personnel type, down and distance, and specific player positioning. Furthermore, we seek to test our theory that running back depth is significant in determining play type. For this analysis we will be using the data provided by the competition.



Problem Statement

- Complex Schemes
- Fast-Paced Offenses
- Real-Time & Halftime Adjustments
- Tendencies

This image by Unknown Author is licensed under CC BY-NC

When Phil Mickelson sat down with David Feherty for his weekly TV show, he went through the numerous parameters he goes through for every shot from tee to green. Now picture having to assess a relatively similar number of factors to call a play before the offense snaps the ball. There in lies the challenge. Coordinators must make these decisions in real-time, and sometimes even quicker when facing a face-paced offense. Our challenge is to sift through the different variables to include formations, down and distance, play shifts, player motion, motion at the snap, receive alignment, time left in the game, score, and so on. Anyone has a 50/50 chance of guessing whether a play is a run or pass, but coaches have to be better than that. Likewise, offensive coordinators develop schemes to appear balanced and unpredictable but can they out-run the data and hidden tendencies that might exist. Again, this report can serve useful to both defensive and offensive coordinators.

Importance of Predictive Analytics



Predictive analytics can reveal hidden patterns with formations, alignments, and motion, offering insights that coaches can use to outsmart opponents. These insights can help teams optimize play calls, adapt strategies, and make data-informed adjustments on the field. Additionally, coaches can develop insights from predictive modeling by testing different strategies and engineering new features to gain an even bigger edge that opponents may not realize they're doing. Sometimes the tendency can be so easy to spot such as a quarter back who has a playbook armband on each arm – one for running plays and one for passing plays. However, with the sophistication of the NFL, it is unlikely to find tendencies easily.



Dataset Overview

- 13 CSV Files
- 9 Weeks of 2022 Player Tracking
 - ~58.7 million records
- ~150 Columns of Data
- Games, Plays and Players
- Tracking Data Must be Used

The BDB provided 13 different tables in the form of comma separated value files with approximately 150 combined columns and about 58.7 million records. Four of the files consisted of player data, game data, play data, and player play data. The bulk of the data was found in the 9 weeks of player tracking data for every play in every game during that time. When combined there were over 50 million records with a size of about 8.2 GB.

Data Cleaning, Preparation, Preprocessing, & Transformation



We began our analysis by cleaning, preparing, preprocessing, and transforming the datasets. Next, we removed features that were associated with post-snap data, removed data where the play would result in a penalty like having two men in motion at the snap, and iteratively cycled through each week of tracking data by aggregating, encoding, engineering features, and merging the tables. Lastly, we concatenated the 9 different aggregated tables to develop a final aggregated table that consisted of just over 16000 rows and 35 columns – a much more manageable file at 4.2 MB. At this point, we developed descriptive visuals that could possibly reveal trends and break-downs the are more readily accessible in the data.

Down and Distance & Rushing Probability

Trend:

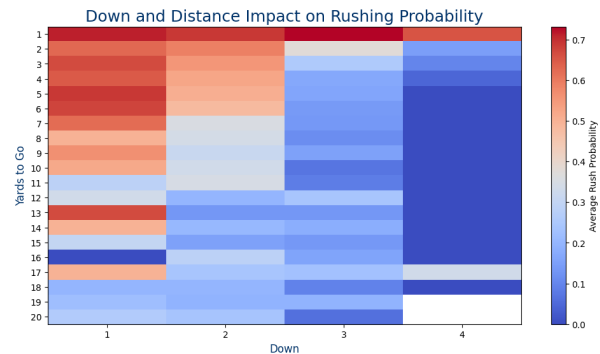
- On early downs (1st and 2nd), rushing probability is higher.
- In short-yardage situations (e.g., fewer than 2 yards to go, rushing becomes the preferred play.
- Teams prioritize rushing plays to maintain manageable scenarios or exploit defensive weaknesses.

Strategic Tendencies:

- Coaches favor rushing to control the clock and set up better passing opportunities on subsequent downs.
- Longer distances (10+ yards) reduce the likelihood of rushing.

Conclusion:

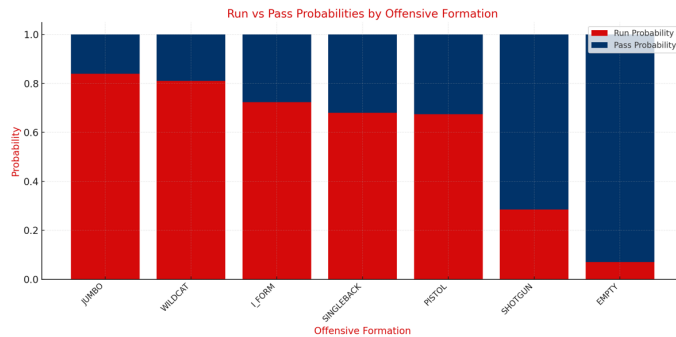
Down and distance are crucial predictors of play-calling strategy, offering actionable insights for game planning and predictive modeling.



It will come as no surprise that as the down and distance increases, the probability of a pass play coming increases. The same can be said for when the distance to get a first down decreases, the probability of seeing a run play increases. Although that trend is noticeable in the visual, it is revealing that during that time frame in 2022, teams were more likely to run the ball on first down with 10 or less yards to go, and second down with 6 or less yards to go. A few anomalies were found like the tendency to run the ball on first down and 13, 14, or 17 yards to go when passing may produce a better outcome in down and distance for second down.

Run vs. Pass Probabilities by Formation

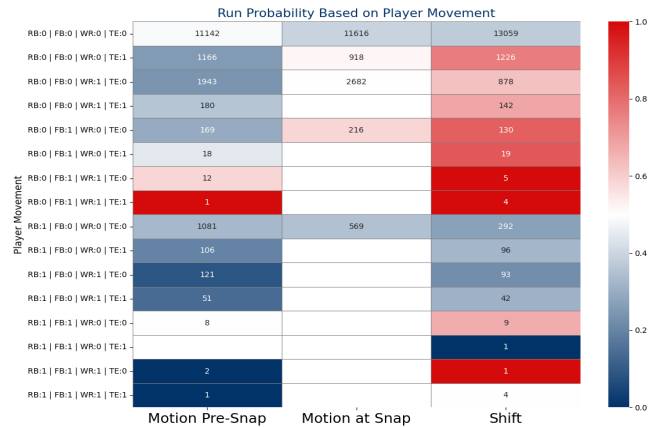
- Run Heavy:
 - Jumbo
 - Wildcat
- Pass Heavy:
 - Empty
 - Shotgun
- Run Likely:
 - I-Formation
 - Singleback
 - Pistol



Again, the descriptive analytics confirms that when teams are in certain formations like Jumbo or Wildcat, they are more likely to run the ball, and when no running backs are in the backfield (Empty) they are more likely to pass. However, the empty set can be tricky when teams have explosive quarterbacks like Lamar Jackson, Justin Fields, or Patrick Mahomes. These teams will purposely spread out the defense with numerous receivers to give running quarterback a better chance of running the ball successfully. It would be wise to have at least one spy-back or stress to the defensive lineman to maintain there rushing lanes and don't pass up the quarterback in their rush.

Player Movement & Play Type

- Motion Pre-Snap
 - More likely to pass
- Motion at Snap
 - Balanced, but pass likely
- Shift
 - Run likely if FB shifts
 - Pass likely if no shift



Here we begin to see some patterns emerging that may not have been so evident like the more obvious indicators. First, when a fullback is in the game, and especially if he shifts, there is a high probability it is a run play. Likewise, if the team doesn't shift at all there tends to be a slightly higher probability of a pass play. Wide receiver motion is very popular in the dataset and was very balanced and showed little variance between pass and run; most likely this is used to induce defenses into making shifts of their own that may actually help the offense.

Modeling Approach

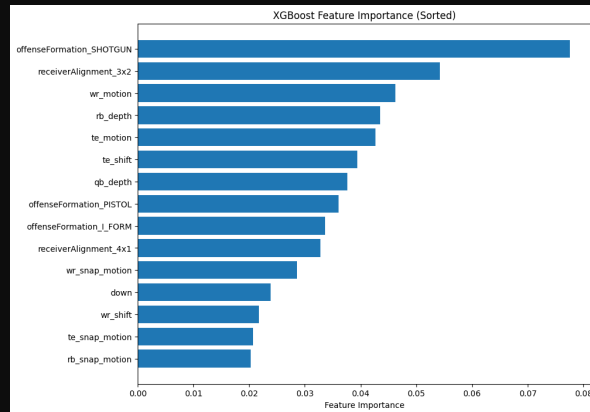
- Target Variable
- Predictive Models Used
- Optimization
- Fine-Tuning
- Performance Metrics



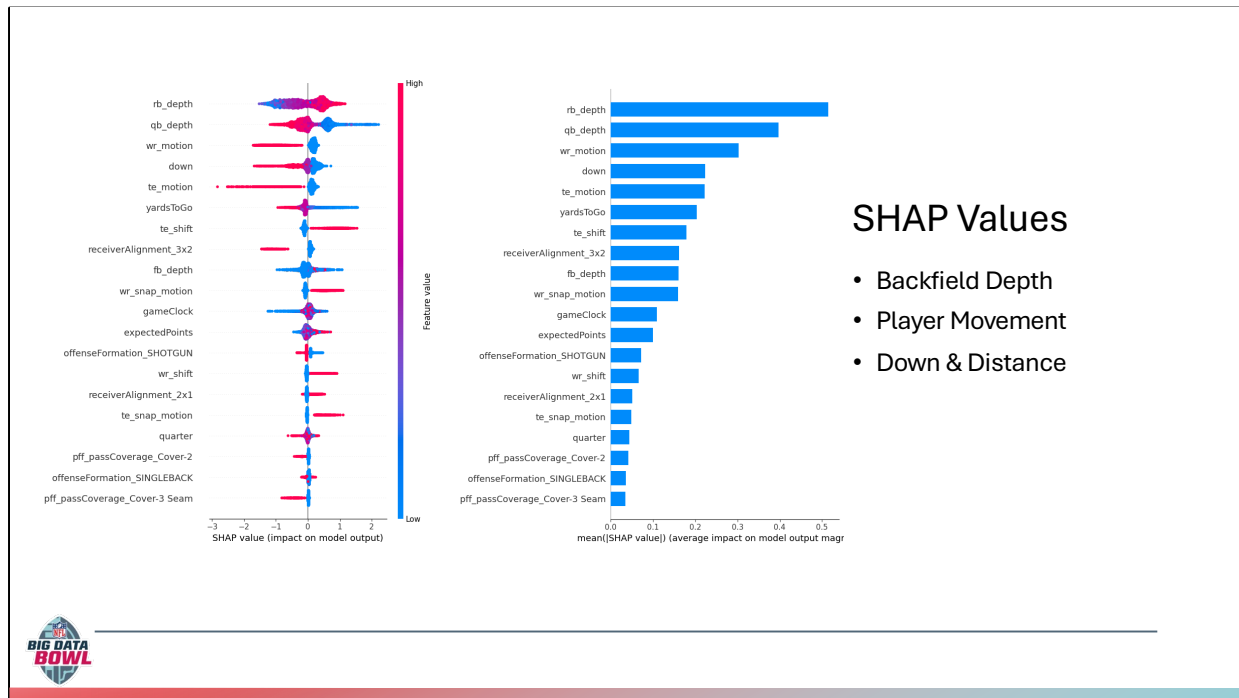
We chose the `rushLocationType` variable as our target because if the play was a run, it had a value and if it was a pass play it was null. We engineered this variable to be a binary classifier with 1 being a run play and 0 being a pass play. One-Hot encoding was used for shift, motions, and motion at snap, and additional features were engineered measuring backfield players' depth at the time of the snap. Continuous variables were standardized, and Boolean variables were created as an integer type. We utilized XGBoost and LightGBM predictive models. Feature importance was conducted with the XGBoost model. Bayesian optimization was used to fine-tune hyperparameters, but GridSearchCV and RandomizedSearch were also performed to ensure we obtained the most accurate parameters. LightGBM proved to be the best training model, correctly predicting whether a play was a run or pass 78.6% of the time. Additionally, it achieved a high ROC-AUC score of 85.6%, which describes how well the model distinguishes between run and pass plays. This means the model is reliable not just in accuracy but also in understanding subtle patterns.

Feature Importance

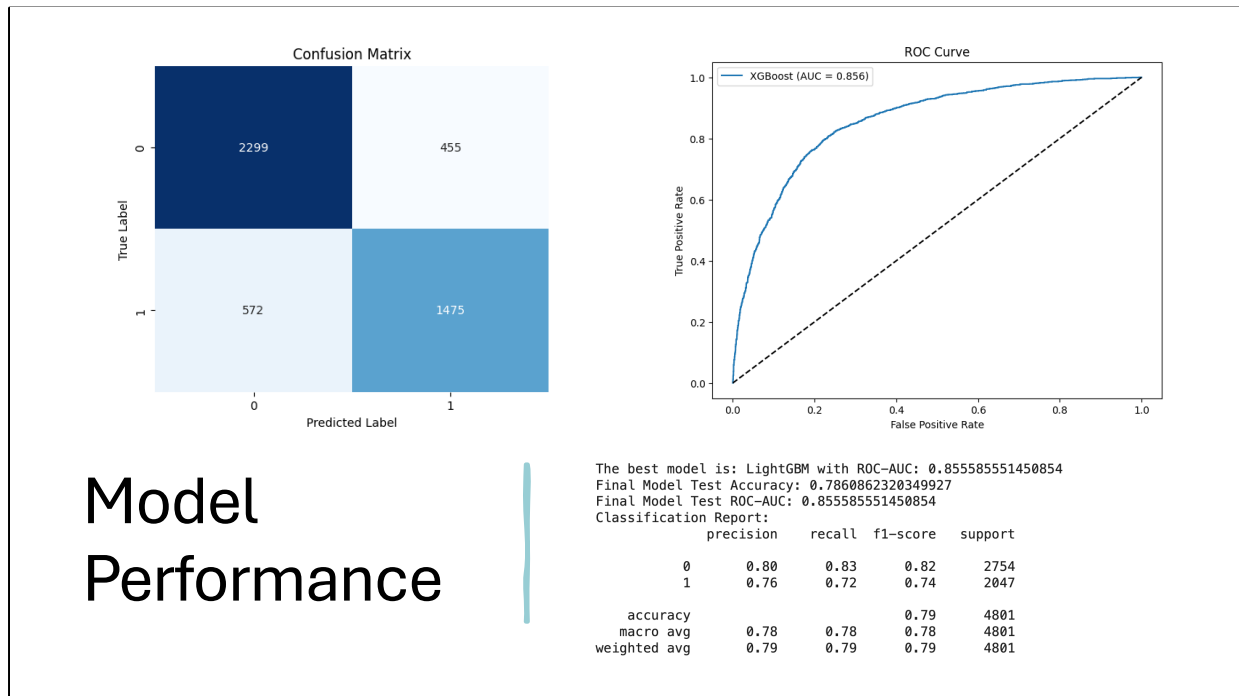
- Which variables provide the most gain
- Running Back (rb_depth) significant
- Shifting creates predictive gain



Using XGBoost, we also identified the most important factors contributing to predictions, such as offensive formations, wide receiver alignment, and running back depth. Don't be fooled when looking at this chart and thinking "Shotgun formation means run?" That would be a mistake. This is simply saying these features, or variables, in the data had the most gain for the predictive model. As we can see, rb_depth is near the top, which is a good indicator for our hypothesis that running back depth is statistically significant in predicting whether a play will be a run or pass.



As predicted at the beginning of the presentation, running back depth had the most impact on the model output. Furthermore, we confirmed our descriptive findings of player movement and down & distance. Here we use SHAP Values to visualize the impact each variable is having on the model.



After testing both XGBoost and LightGBM models, we found that the LightGBM model tested best. We found the accuracy to be 78.6% and ROC-AUC to be 85.6%. This is basically saying that with our model, one can have a much better prediction than a random guess at 50% and the model will be correct more than 3/4ths three quarters of the time. Those metrics are promising, considering in 2021, Otting's Markov model had an accuracy of 71.6%, while Lee, Chen, and Lakshman had 75.9% accuracy with a ROC-AUC of 84.9% in their publication.

Coaching Recommendations



The future is here and incorporating machine learning into your game planning should be happening. All 32 NFL teams have data analysts and technology is only becoming more powerful. It is worthwhile to evaluate available metrics and try to uncover a tendency that could be a significant tell. This report has shown that running back depth is very significant and can be an indicator for a defensive audibles. One thing to remember about running back depths, defensive players may have a harder time seeing the depth difference on the field and may need a signal or call-out from the sideline to notify them. Additionally, player shifting and of course down and distance can also reveal meaningful information in an offense's scheme.

Future Consideration



As technology and generative artificial intelligence continues to be popular, it is not beyond the realm of possibilities to use computer imaging to train predictive models that will be able to provide real-time analytics and insights. Imagine the advantage in scouting your opponent if all you had to do was run their game film through a model and let the computer use its advance capabilities of picking up on even more hidden nuances then once believed possible. We believe we have only hit the tip of the iceberg when it comes to the intersection of professional football and data science, and the future is bright. Thank you!

References

Saquon Barkley (image slide 2): https://commons.wikimedia.org/wiki/File:Saquon_Barkley_Giants_2018.jpg

Eagles & Redskins (image slide 3): <https://www.flickr.com/photos/imatty35/6262282662>

Lee P., Chen R., & Lakshman V. (unknown) Predicting Offensive Play types in the National Football League.

Marius Ötting, Predicting play calls in the National Football League using hidden Markov models, IMA Journal of Management Mathematics, Volume 32, Issue 4, October 2021, Pages 535–545, <https://doi.org/10.1093/imaman/dpab005>

Lee, P., Chen, R., & Lakshman, V. (n.d.). Predicting offensive play types in the NFL. Retrieved from <https://cs229.stanford.edu/proj2016/report/LeeChenLakshman-PredictingOffensivePlayTypesIntheNFL-report.pdf>

[GitHub Repository](#)

