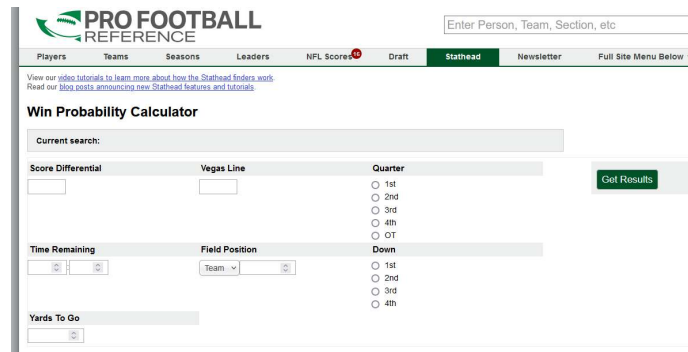**NFL Big Data Bowl 2025: Pre-Snap Play Prediction**

By Matthew Rabin

Introduction

The National Football League has been working in partnership with Amazon Web Services since 2017 to capture "real time location data, speed and acceleration for every player, every play on every inch of the field". Win Probability (WP) is an existing metric that is useful for in-game decision-making. WP can be calculated from details about the game state such as field position, time remaining, and score differential. This project uses machine learning based on player tracking data and the inputs for WP to create a new metric for pre-snap play prediction.
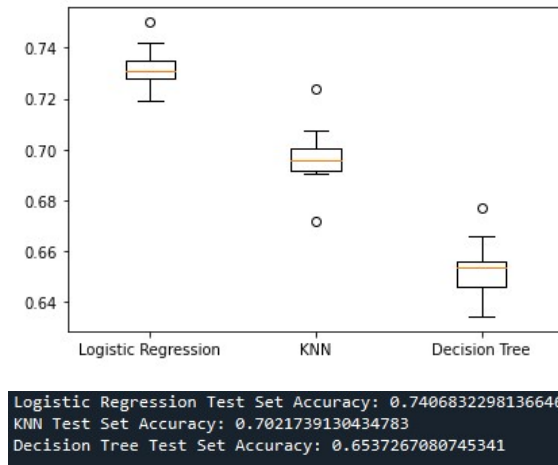

Data

The plays dataset was used to identify offensive formation, receiver alignment, field position, down and distance, game time remaining, and score differential. The game dataset was used to identify the home team. A binary classification of pass or run (identified from game data) was used as the outcome variable for the predictive model.

Methodology

Python was used to prepare the data and build the predictive model (see appendix). Data was limited to only plays that resulted in a pass or run. Game time remaining in seconds was calculated from the quarter and gameClock fields. The home team was identified by joining play and game data. Home_team_advantage was calculated as a function of home_team (1/0) and game time remaining. This was modeled after pro-football-reference's WP formula which accounts for the "diminishing amount of time remaining in the game". One-Hot Encoding was used to create dummies for categorical variables (offensive formation, receiver alignment, and team). A standard scaler was applied to convert the features to the same scale. The outcome was a binary classification for whether the play was a pass or a run.

The data was randomly split into two groups for training (80%) and testing (20%). The training data was further divided ten non-overlapping times into 90% training and 10% validation groups. This allowed for 10-fold cross validation to optimize model parameters and prevent overfitting. Three different machine learning models were built and tested to determine which had the best performance. Models were built

using Logistic Regression, K-Nearest Neighbors and Decision Tree classification. The Logistic Regression model was selected after returning the best results.


```
Logistic Regression Test Set Accuracy: 0.7406832298136646
KNN Test Set Accuracy: 0.7021739130434783
Decision Tree Test Set Accuracy: 0.6537267080745341
```

Discussion/Conclusions

This research provides a data-driven approach to predicting plays in football. It combines variables from an established metric for in-game decision-making (Win Probability) with player location data from AWS Next Gen Stats to predict plays. The Logistic Regression model performed best among the three classification models with 74% accuracy. The model could be used by defensive coaches to predict whether an offense will run or pass and react accordingly with their play calls.

Improvements/Future Work

Pro-football-reference's Win Probability formula uses Vegas Line rather than home field advantage. This was not available in the dataset. The model's accuracy may improve if these points spreads were found from another source and added as a feature. Only 136 games of data were used (the first half of the 2022 season). The model's accuracy could be improved with more training data.

Acknowledgements
Professors Brian Hall and Matt Manocherian from the Artificial Intelligence and Machine Learning and Gridiron Analytics courses at New York University.

References
Next Gen Stats quote: https://nextgenstats.nfl.com/glossary
Python resources: https://app.datacamp.com/learn/courses/supervised-learning-with-scikit-learn
Pro-football-reference Win Probability formula: https://www.pro-football-reference.com/about/win_prob.htm

Appendix
Python code: https://github.com/matthew-rabin/NFL-Big-Data-Bowl-2025