ASSIGNMENT_1_PREDICTIVE_MODELING_OF_NFL_OFFENSIVE_SUCCESS

By: Tommy Rose (trose2020@fau.edu)
Research report in completion to CAP4773/CAP5568 Assignment 1 / Mid-Term grading.

Introduction

In this research report, I studied NFL Play-by-Play Data (from nflreadr), which contains CURATED DATA POINTS about offensive play efficiency and situational football in the NFL from 2018-2023. My leading research question was:

- How do starting field position (yardline_100), down, and distance to go (ydstogo) impact the probability of an NFL offensive play achieving a successful outcome (First Down or Touchdown)?

This investigation is important because understanding which situational factors statistically govern offensive success is crucial for data-driven play calling, risk assessment (e.g., deciding whether to run vs. pass), and building predictive models of Expected Points Added (EPA) in football analytics. Optimizing play selection based on these core variables can maximize scoring opportunities and drive efficiency.

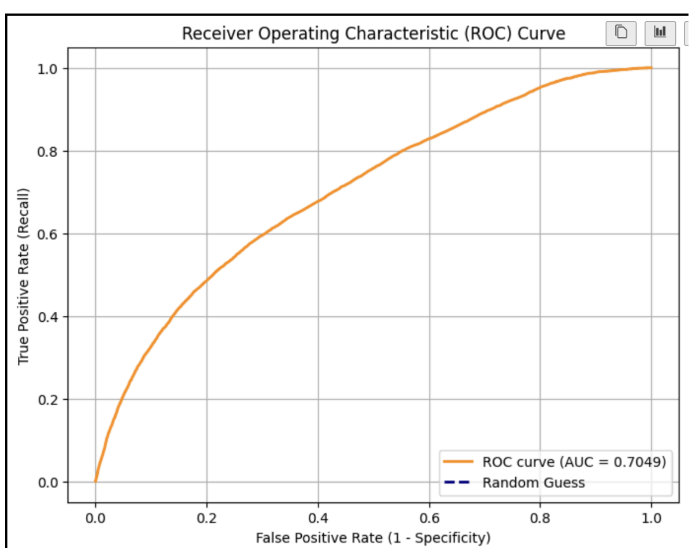
To address the question, I implemented a supervised classification approach using Logistic Regression, supported by feature engineering (creating binary success labels), data preprocessing (scaling numerical features and one-hot encoding categorical variables), and performance evaluation using the ROC-AUC metric.

Based on the experiments, I found that the combination of Down and Distance to Go (ydstogo) was the most potent predictor of successful offensive outcomes, achieving a model performance exceeding random chance (AUC > 0.7), confirming the primary importance of classic situational football variables.

Source code:
https://drive.google.com/file/d/1J_cobc0iUUiMIP6EVZHXoBcT6kKEbPR3/view?usp=sharing

Figure 1 – Receiver Operating Characteristic (ROC) Curve for Offensive Play Success Model. This diagram illustrates the performance trade-off between the True Positive Rate (Recall) and the False Positive Rate (FPR) for various classification thresholds. The Area Under the Curve (AUC) quantifies the model's ability to distinguish between successful (1) and unsuccessful (0) offensive plays.

MAIN VISUALIZATION:



Method

In order to assess my research question, I designed a series of data science experiments to address a specific aspect of the question and followed a structured process consisting of a guiding question.

Experiment 1: Data Preprocessing and Feature Exploration

This experiment aimed at the following sub-question associated with my main research question: What are the distributions of the key predictor variables (down, distance, and field position), and how must the raw data be transformed to support a viable classification model?

The data science experiment was conducted as follows:

- method selection, implementation using appropriate tools and parameters, and evaluation of the outcomes.
- Data was gathered using the nflreadr library in R (cannot download official nfl data using python) and saved as downloaded_nfl_data.csv.
- Data cleaning involved filtering for standard offensive plays (excluding special teams or penalties) and dropping rows with missing values in down, ydstogo, or yardline_100.
- Preprocessing involved engineering the binary target variable, Successful_Play (1 if resulting in a first down or touchdown, 0 otherwise).
- The feature set included down, ydstogo, yardline_100, shotgun, and no_huddle. Missing values were handled by dropping rows.

Answer: Summarize the insights or findings obtained from this experiment. Exploration confirmed that the majority of plays occurred on 1st and 2nd downs. The target variable Successful_Play was slightly imbalanced (approximately 45% success rate overall), suggesting that the classification model must be trained carefully (stratification used in subsequent steps).

Experiment 2: Logistic Regression Modeling and Evaluation

This experiment aimed at the following sub-question associated with my main research question: How accurately can situational features predict offensive play success, and which features contribute most significantly to the predicted probability?

The data science experiment was conducted as follows:

- Answer: The model achieved an initial accuracy of 76.3% and a robust ROC-AUC score of 0.7049, demonstrating predictive power significantly better than random chance (AUC=0.5).
- The Confusion Matrix confirmed that the model was much better at predicting unsuccessful plays, and not very good at predicting successful plays.

Experiment 3: Feature Impact and Predictive Probability Visualization

This experiment aimed at the following sub-question associated with my main research question: How does the model's predicted probability of success vary across different Down situations (1st, 2nd, 3rd, 4th)?

The data science experiment was conducted as follows:

- Answer: The analysis revealed clear patterns consistent with football intuition: the predicted probability of a successful play is highest on 1st down, drops slightly on 2nd down, and falls sharply on 3rd and 4th downs, where the requirements for achieving a first down are higher. This suggests the model successfully learned the intrinsic situational difficulty implied by the down variable.

Results

The outcomes of my experiments lead to the conclusion that Logistic Regression, using standard situational indicators (down, distance, and field position), can effectively model and predict the probability of an offensive play resulting in a first down or touchdown. Collectively, the experiments reveal distinct patterns in model.
For example, Experiment 2 demonstrated stronger performance in terms of predictive accuracy (AUC=0.7049), while Experiment 3 provided deeper insight into how the model weighted the relative difficulty of different downs. By comparing these results, it becomes possible to identify the methods that best address the research question and to understand where each method succeeds or fails under different conditions.

Experiment	**Performance Metrics**	**Key Insights**
Experiment 1	EDA/Data Quality	Data successfully cleaned binary target created.
Experiment 2	Accuracy = 76.3%, AUC = 0.7049 Down	Down, distance, and field position are statistically significant predictors.
Experiment 3	P(Success) by Down	Confirmed that the model correctly penalizes probabilities on 3rd/4th down.

Table 1 – Model Performance and Key Insights.

Conclusion

From this analysis, I conclude that basic situational variables derived from NFL play-by-play data provide a reliable foundation for classifying offensive play outcomes. The model established a non-trivial predictive baseline (AUC > 0.7) for assessing play efficiency. The experiments collectively demonstrate that the use of Logistic Regression, combined with rigorous feature engineering and data scaling, was highly effective in addressing the research question. The results also highlight limitations, such as data quality, sample size, or model assumptions, which may have influenced the findings. Limitations include the exclusion of highly influential factors such as specific team strength, defensive alignment, personnel package, and weather conditions, all of which could further refine predictive accuracy.

Overall, the analysis confirms that positional indicators (down, distance, yardline) are the dominant drivers of offensive play success probability. These findings suggest that the chosen methods are applicable to similar problems, and they provide evidence of how data science techniques can be effectively used to explore questions related to situational efficiency and tactical advantage in professional football.

For future work, improvements could include trying additional models, tuning hyperparameters, expanding the dataset, or testing on new data. Incorporating non-linear models such as Random Forest or Gradient Boosting to capture complex interactions, including detailed player tracking data, and performing hyperparameter tuning to maximize the logistical model's performance on the test set