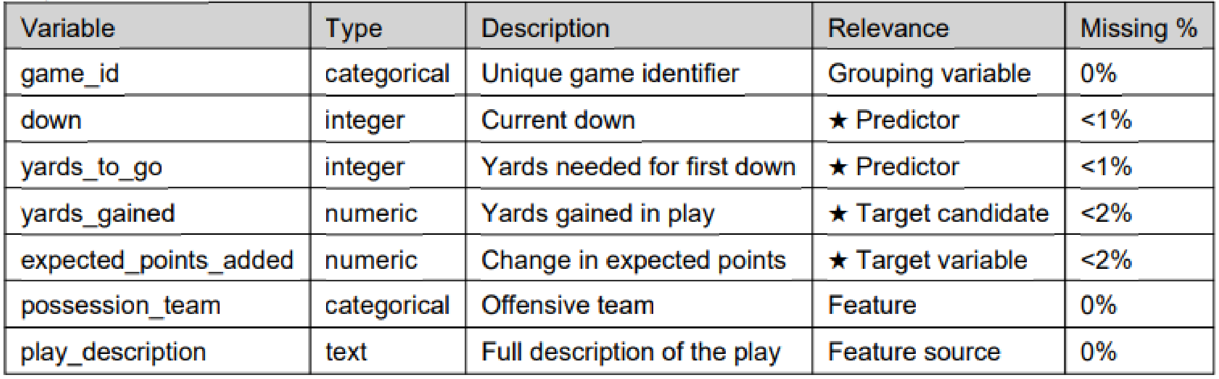
Sprint 2 Report: Predicting Positive Expected Points Added in Play-by-Play Data - Jeremiah Isichei

1. Data Acquisition & Description:
    a. I was able to obtain the primary dataset titled supplementary_data.csv, which is the primary data source for this project. The data includes 18,009 observations and 41 distinct variables, each providing detailed information on individual play-level events that were recorded during football games. The observations include a wide range of contextual and performance-dependent variables, including team identifiers, in-game scores, play outcomes, field position, and situational factors such as down and yards to go. In addition, the data include some higher-level probability-based measures, such as expected points and win probability measures, that provide insight into the statistical and strategic elements of play. Together, these variables form a comprehensive foundation for examining in-game decision-making and result outcomes, and therefore this dataset is complete and highly suitable for predictive modeling and exploratory data analysis.
2. Source:
    a. User-provided CSV file, consistent with publicly available play-by-play data from a Kaggle 2026 NFL Big Data Bowl
3. Collection Method:
    a. Derived from administrative/sensor-based game logs or play-tracking APIs.
4. Temporal Coverage:
    a. The NFL Seasons of 2023–2024.
5. Spatial Coverage:
    a. U.S.-based professional football teams.
6. Unit of Analysis:
    a. Each row represents a single play within a game. Sample Size: 18,009 observations after cleaning.
7. Sample Size
    a. The sample size consists of 18,009 information.

Key Variables:

8. 

Variable	Type	Description	Relevance	Missing %
game_id	categorical	Unique game identifier	Grouping variable	0%
down	integer	Current down	★ Predictor	<1%
yards_to_go	integer	Yards needed for first down	★ Predictor	<1%
yards_gained	numeric	Yards gained in play	★ Target candidate	<2%
expected_points_added	numeric	Change in expected points	★ Target variable	<2%
possession_team	categorical	Offensive team	Feature	0%
play_description	text	Full description of the play	Feature source	0%

9. Data Quality Assessment & Cleaning
    a. The dataset contained relatively fewer missing values in its numeric columns, with the overall percentage well below two percent. To preserve the completeness and integrity of the dataset for subsequent analysis, I performed median imputation for the numeric columns, replacing missing values with the median of their respective columns. This method was selected as it minimizes the effects of outliers without changing the general distribution of data. For categorical variables, missing values were replaced by the dummy label "MISSING" that did not remove these observations from the dataset and enabled models to treat them as a distinct, comprehensible category. Aside from imputing missing data, duplicate rows were also identified and removed to eliminate redundancy and potential bias when conducting the analysis. Column names were also standardized for consistency in formatting and for ease of referring to them in the course of the data cleaning and modeling exercises. Finally, a new binary feature named “target_positive_epa” was added to capture whether each play was a positive Expected Points Added (EPA). This metric is the primary measure of outcome for the predictive modeling phase, distinguishing plays which improved a team's scoring ability from those that did not.
        i. Missing Data: Numeric imputed by median; categorical imputed with 'MISSING'. Missingness is apparently systematic in advanced-tracking categories.
        ii. Outliers: Present but mostly real
        iii. Duplicates: Removed exact duplicates.

        iv. Feature Engineering: Added binary target, normalized dates, and imputed missing values.

    b. Cleaning Pipeline: Data cleaning process followed a sequential and systematic workflow for consistency and reliability of all observations. Loading of the raw CSV file with the parameter *low_memory=False* was the starting point to facilitate optimal type inference and prevent data type fragmentation during importation. Next, duplicate records were identified and removed to eliminate redundancy and ensure dataset integrity. Next, all the column names were normalized for clarity and ease of reference throughout the analysis. The "game_date" field was converted into a datetime field to enable accurate temporal analysis. A new binary feature, "target_positive_epa", was created that classifies plays according to whether their Expected Points Added (EPA) was positive. Missing numeric features were replaced with the median of their columns, and missing categorical features were replaced with the placeholder "MISSING" in order to preserve record completeness. Finally, the fully cleaned dataset of 18,009 rows and 42 columns was saved for additional exploratory and modeling.

        i. Potential Bias: Median imputation can compress variance and bias results for subgroups or older seasons.

10. Exploratory Data Analysis (EDA) The majority of Sprint 2 was spent exploring the cleaned dataset. I made histograms, bar charts, and correlation heatmaps.

    a. Univariate Analysis: Yards gained and EPA were right-skewed in distribution, illustrating the occasional long plays. "Down" and "yards_to_go" exhibited expected discrete patterns.

    b. Bivariate/Multivariate Analysis: Correlation analysis indicated that field position, down, and pre-snap win probability are highly correlated with EPA. Scatterplots indicated that as field position improves, EPA also shifts in a positive direction.

    c. Visualization Best Practices: All figures were labeled and generated for readability. Only 5–6 key visuals were used for clarity.

    d. Surprising Findings: Pre-snap win probability was more strongly correlated with play success than expected. Text features in play descriptions will require parsing to obtain predictive signals.

    e. Data Limitations: No player or betting data explicitly stated; lacked EPA data for certain games. Advanced tracking metrics were sometimes unavailable.

11. Refined Problem Statement & Analytical Plan
    a. Revised Problem Statement: Analysis now forecasts whether a play possesses positive EPA, rather than betting outcomes, due to data availability.
    b. Updated Analytical Approach: Baseline models for classification problems “target_positive_epa” are Logistic Regression and Random Forests, and metrics are ROC-AUC and PR-AUC. For regression on continuous EPA, RMSE and MAE will be used. Data will be split by season to prevent leakage. Feature selection will focus on field position, down, and play context features.
    c. Challenges & Mitigation: Imputation and regex extraction will address missing EPA fields and text parsing. In the event of poor classification accuracy, regression models on EPA magnitude will be employed as a fallback.
12. Progress Tracking & Next Steps Accomplishments: Dataset cleaned, target defined, EDA completed, and modeling plan determined.
13. Sprint 3 Plan: The primary work of Sprint 3 will be to enhance the dataset through advanced feature engineering and set up the first stage of predictive modeling. This process will involve building additional variables that capture significant gameplay mechanisms, such as binary indicators for “is_pass”, “is_run”, and “is_touchdown”, which are derived from the play descriptions. These extra features will allow the models to make a better distinction between offense plans and situations in every play. In addition, interaction terms between fields like “down”, “yards_to_go”, and “field_position” will be designed to model the joint impacts of game strategy and field position. At the same time as feature generation, additional detailed missing data analysis will be conducted to identify patterns across teams and seasons in order to ensure that systematic missingness is properly accounted for in order to prevent bias in subsequent modeling stages. Following cleaning of the dataset, baseline machine learning models such as logistic regression and random forest classifier will be employed to predict if a play results in a positive Expected Points Added (EPA). Model performance will be gauged using standard parameters such as accuracy, ROC-AUC, and precision-recall scores to establish stable baselines for future comparison. Visualization of model outputs,

including feature importance plots and confusion matrices, will also be performed to provide insight into salient predictors driving model performance. These qualitative evaluations will serve as the foundation for additional interpretability and stability testing in Sprint 4.

14. Sprint 4 Plan: Sprint 4's primary area of interest will be to look beyond predictive performance and explore in greater depth the interpretability and reliability of the models developed in Sprint 3. In this stage, a lot of attention will be given to performing a comprehensive explainability analysis using techniques such as feature importance evaluation, SHAP (SHapley Additive Explanations) values, and partial dependence plots to gain insight into the individual variables' effect on the model's predictions. In addition to this, robustness testing shall be performed to determine the generalizability of the models to other subsets of data, e.g., other teams, other seasons, or other types of games. This is to ensure that the inferences gained are not a result of overfitting or sample bias. Finally, the results will be synthesized to determine the application of the predictors to real-world problems, transmuting statistical relationships into conclusions that express gameplay strategy and performance drivers. These conclusions not only will validate the analytical findings but also will provide a bridge between data-driven modeling and real-world decision-making in sports analytics.
15. Self-Assessment: On track. Biggest risk is lack of betting data; mitigation is pivoting to predictive modeling of in-game outcomes.