**NFL Play Type Prediction Using K-Nearest Neighbours (KNN)**

**Predictive Analytics,**

**Introduction**

This analysis aims to predict NFL play types—either "Run" or "Pass"—by developing a K-Nearest Neighbors (KNN) classification model. By utilising features that describe play context, field position, and game situation, this model aims to enhance strategic decision-making by identifying patterns associated with different play types.

The following report outlines the steps involved in preparing and analysing the data, including handling outliers, encoding categorical features, balancing classes, and selecting relevant features.

**Methodology/Data Cleaning:**

**1. Data Loading and Initial Exploration**

Data Import: The dataset was loaded into a Pandas Data Frame, followed by an inspection of column names, data types, and missing values.

**2. Handling Missing Values and Dropping Irrelevant Columns**

The data cleaning process began with an initial inspection of the dataset to identify any missing values, irrelevant columns, and potential outliers. This step was essential to ensure that the dataset was well-prepared for analysis and model training. Columns irrelevant to predicting play types, such as game\_id, play\_id, and venue, were dropped to reduce noise in the data. These columns contained identifiers or location information that had no direct impact on whether a play would be a "Run" or "Pass."

**3. Encoding Categorical Features**

Categorical features, such as down and situational flags (e.g., in\_red\_zone, short\_yardage), required transformation to a numerical format suitable for machine learning models. Binary categorical features were label-encoded to convert them to a 0-1 scale, while multi-class features like down were one-hot encoded, creating a separate binary column for each possible value. This approach allowed the model to interpret these categorical distinctions without implying any ordinal relationship between the classes.

**4.Standardisation**

Standardisation of numerical columns was the final data preparation step. Standardisation scales features to have a mean of 0 and a standard deviation of 1, ensuring that each feature contributes equally to distance calculations in KNN. Without standardisation, features with larger ranges, such as yard\_line and score\_diff, could disproportionately impact the distance metrics in the KNN algorithm.

**Exploratory Data Analysis (EDA) and Statistics:**

The Exploratory Data Analysis (EDA) phase began with data cleaning and an initial exploration of key columns to assess their statistical properties. This process was essential for understanding the distributions, ranges, and potential issues in the data, such as outliers and class imbalance.

**Dropping Unnecessary Columns**

We first removed irrelevant columns that did not contribute directly to predicting play types. These included identifiers like game\_id and play\_id, as well as contextual fields such as venue, which were deemed unnecessary for our modelling goals. By retaining only relevant columns, we reduced potential noise in the data and focused on attributes that provide information about game situations, team performance, and play context.

**Summary Statistics**

For key columns, including yards\_to\_go, yard\_line, down, score\_diff, and play\_type, summary statistics were generated to examine the central tendencies, spread, and possible outliers. The statistics highlighted the following:

yards\_to\_go had a mean of approximately 8.45 yards, indicating that most plays started around a medium distance from the first down marker.

yard\_line had an average value of 51.17, meaning the typical play was close to midfield.

score\_diff ranged widely, with a mean close to zero, indicating that the dataset included plays from various game situations—some close and others with larger score differences.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| FEATURE | MEAN | STD | MIN | 25% | 50% | 75% | MAX |
| yards\_to\_go | 8.45 | 3.94 | 1 | 6 | 10 | 10 | 40 |
| yard\_line | 51.17 | 24.16 | 1 | 33 | 55 | 72 | 99 |
| down | 1.82 | 0.84 | 1 | 1 | 2 | 2 | 4 |
| score\_diff | 1.36 | 10.51 | -40 | -4 | 0 | 7 | 56 |

**Outlier Capping**

Outliers were identified in the score\_diff feature, which could distort model performance. We capped extreme values using the IQR method with custom bounds of -20 and 20 to contain excessively high or low values. This process helped stabilise the model by reducing the impact of extreme cases, especially in KNN models, where distance metrics can be heavily influenced by large values.

**Standardisation of Numerical Features**

The numerical features (yards\_to\_go, yard\_line, and score\_diff) were standardised to ensure each feature contributed equally to the KNN distance calculations. Standardisation transformed these features to a mean of zero and a standard deviation of one, which is crucial in KNN models to avoid features with larger ranges dominating the model.

**Summary Statistics for Standardised Features:**

FEATURE MEAN STD MIN 25% 50% 75% MAX

yards\_to\_go 0.00 1.00 -1.89 -0.62 0.39 0.39 8.00

yard\_line 0.00 1.00 -2.08 -0.75 0.16 0.86 1.98

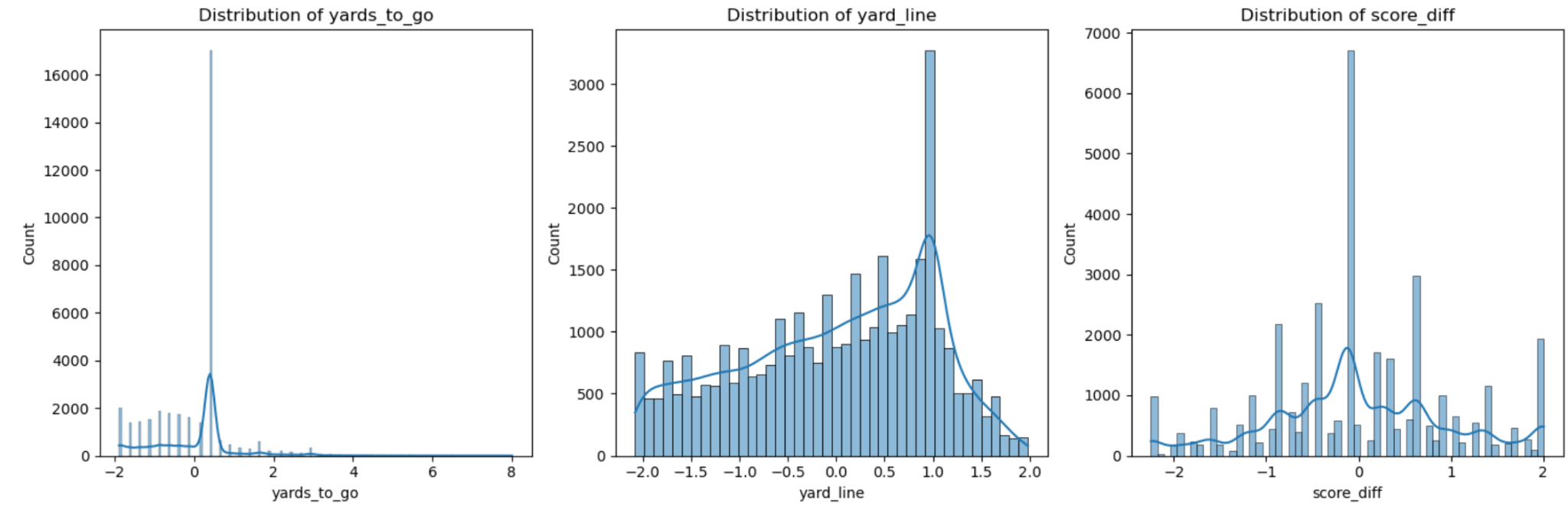
score\_diff 0.00 1.00 -2.25 -0.55 -0.12 0.62 2.00

The distribution of numeric features in this NFL play type prediction dataset provides valuable insights into the underlying characteristics and ranges of key variables, such as yards\_to\_go, yard\_line, and score\_diff. Understanding these distributions is essential, as they affect the model’s ability to learn meaningful patterns and make accurate predictions.

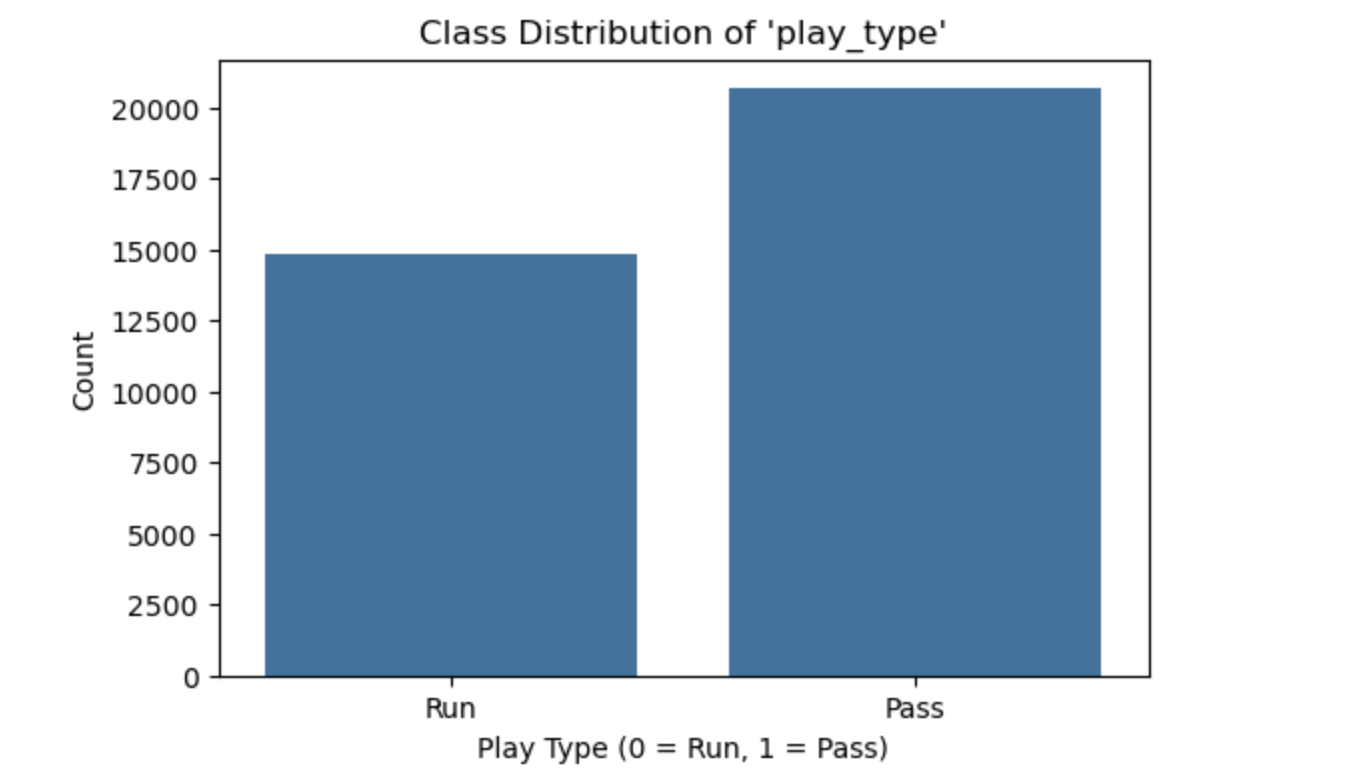
**yards\_to\_go:** This feature represents the yards remaining for a first down, with values ranging from 1 to 40. The distribution is relatively symmetrical, with a mean of approximately 8.45 yards. This indicates that, on average, plays start with a moderate distance needed for a first down. However, a concentration of values around 6 and 10 yards suggests a higher frequency of shorter and mid-range plays. This feature’s spread is somewhat constrained, with most values within a reasonable range, making it less prone to extreme outliers.

**yard\_line:** This feature reflects the field position where each play begins, with values ranging from 1 to 99. The mean value is 51.17, which suggests that plays are, on average, centred around midfield. The distribution is also fairly symmetrical, indicating that plays are distributed across the field rather than clustering in specific zones. However, values close to 1 or 99 may indicate plays starting close to the end zones, likely influenced by situational factors like punts for touchdowns.

**score\_diff:** Representing the difference in score between the offensive and defensive teams, score\_diff has a much wider range, from -40 to 56. This range reflects the diversity of game situations included in the dataset, with close games and blowouts represented. The distribution of score\_diff is less symmetrical than other features, with many values clustering around zero, indicating a tendency for closer games. However, extreme values at either end reveal cases where one team had a significant lead, which may impact play-calling decisions.

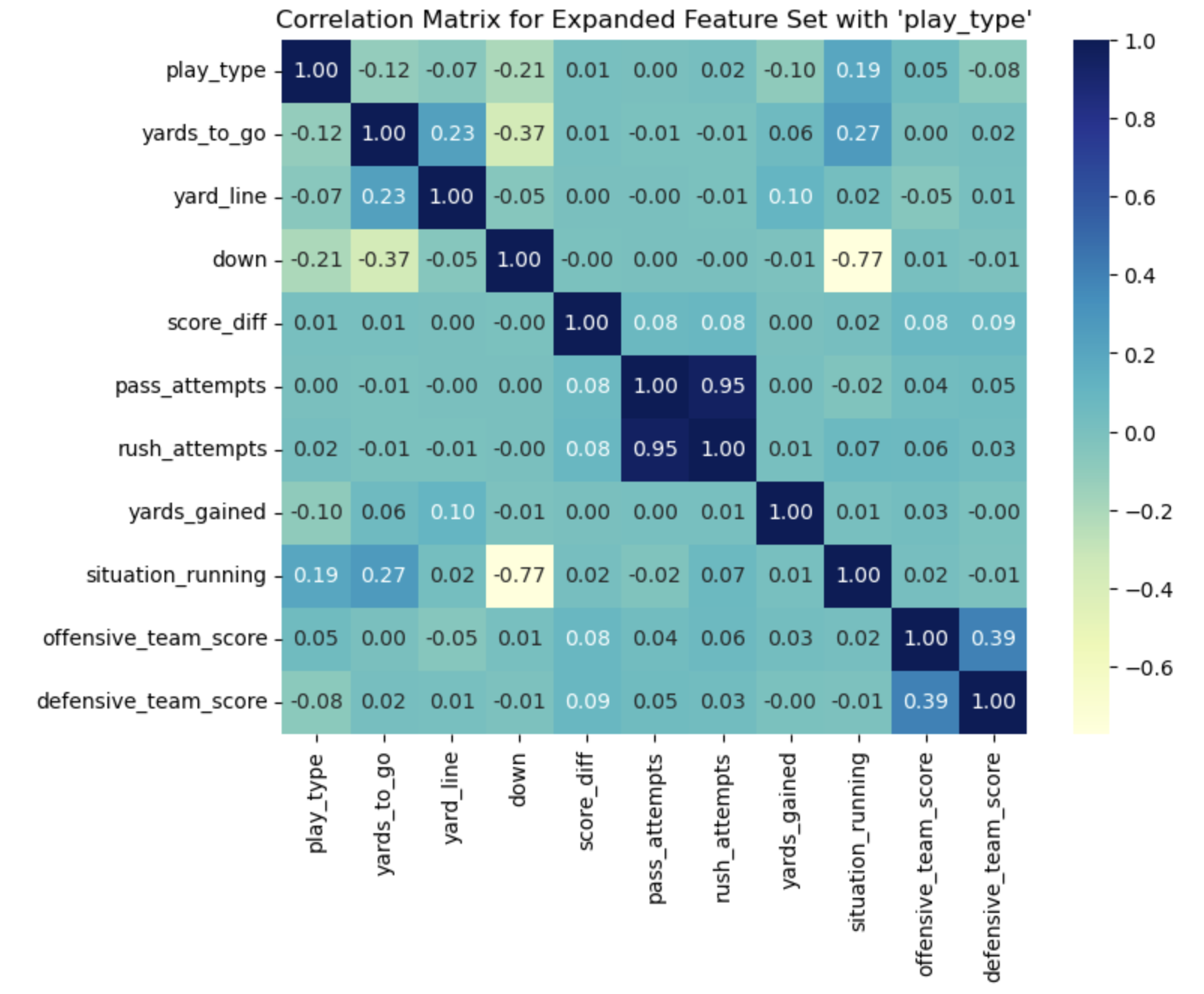


**Class Distribution of 'Play Type':** A count plot of play\_type shows a notable imbalance between run and pass plays. Specifically, there are 20,644 pass plays and 14,830 run plays in the dataset, indicating that passing plays are more common. This trend may reflect a strategic preference for passing, possibly due to situational demands or a general shift in offensive tactics within the games captured here. This imbalance suggests that teams are more likely to choose passing plays in various situations, providing valuable context for model training. Balancing these classes is essential to ensure that the model performs well across both play types without biassing toward the more frequent pass plays.



**Correlation Analysis:**

A correlation matrix was used to identify relationships among variables, revealing how each feature related to play\_type. This analysis confirmed the selection of yards\_to\_go, yard\_line, and score\_diff as primary predictive features due to their distinct influence on play type. Variables with high intercorrelation were examined to avoid multicollinearity, ensuring that each selected feature contributed uniquely to model predictions.



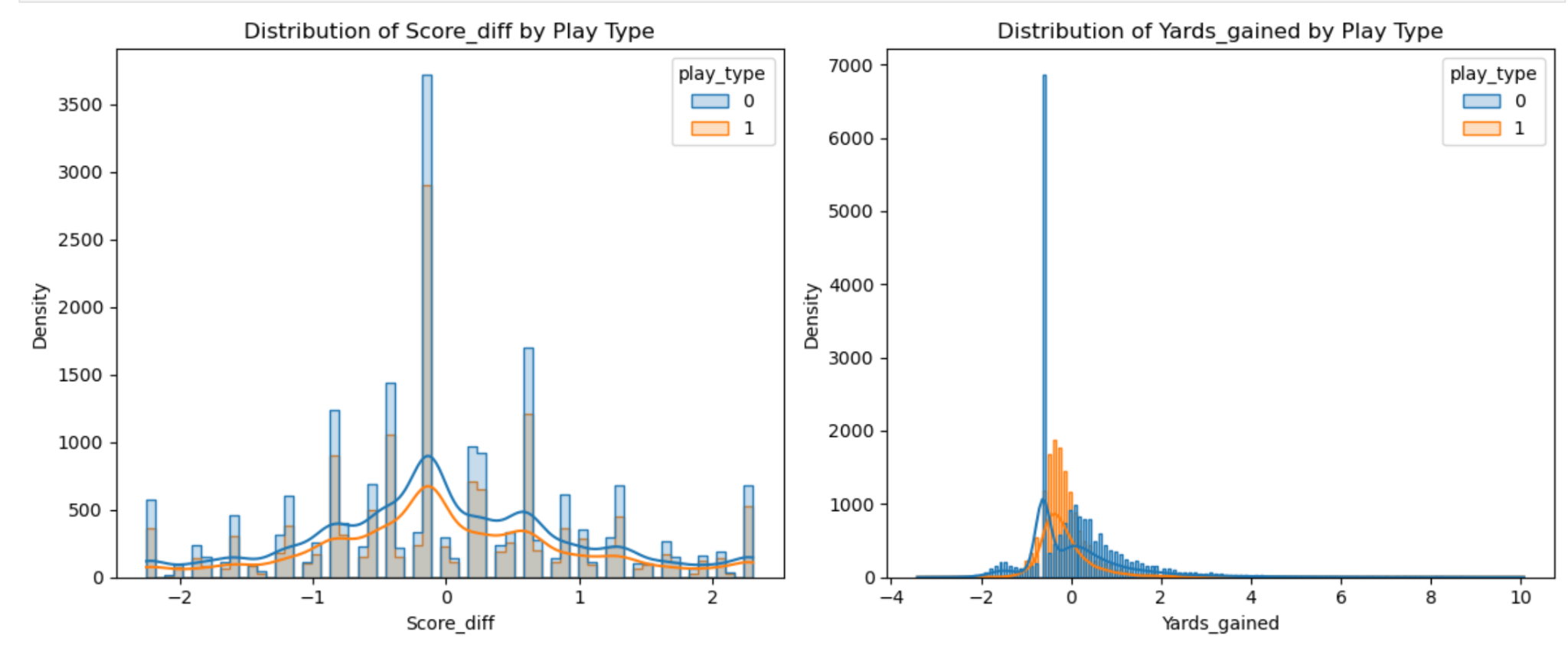
**Balancing Predictive Power and Simplicity:**

By focusing on a few high-impact features, the model remained efficient without overfitting to irrelevant data. This approach allowed for a streamlined model that leveraged key game insights, ensuring that predictions were both interpretable and strategically relevant.

To further explore the relationship between key features and play type, we analysed the distributions of score\_diff and yards\_gained using histograms with KDE (Kernel Density Estimation) curves. These visualisations were split by play type to observe any trends specific to "Run" and "Pass" plays.

Score Difference (score\_diff): The distribution of score\_diff centres around zero, with many games closely contested. This indicates that most plays occur in balanced game situations, where neither team has a substantial lead. However, for plays with larger positive or negative score differences, there are subtle shifts in play type tendencies, potentially reflecting strategic adjustments based on the game state.

Yards Gained (yards\_gained): The yards\_gained distribution shows distinct patterns for each play type. Pass plays tend to have a wider spread, with some plays resulting in higher yardage gains compared to runs. In contrast, run plays are more concentrated around lower yardage gains, reflecting the typical outcome of running plays in most game scenarios.



**K-Nearest Neighbors (KNN) with Varying K Values :**

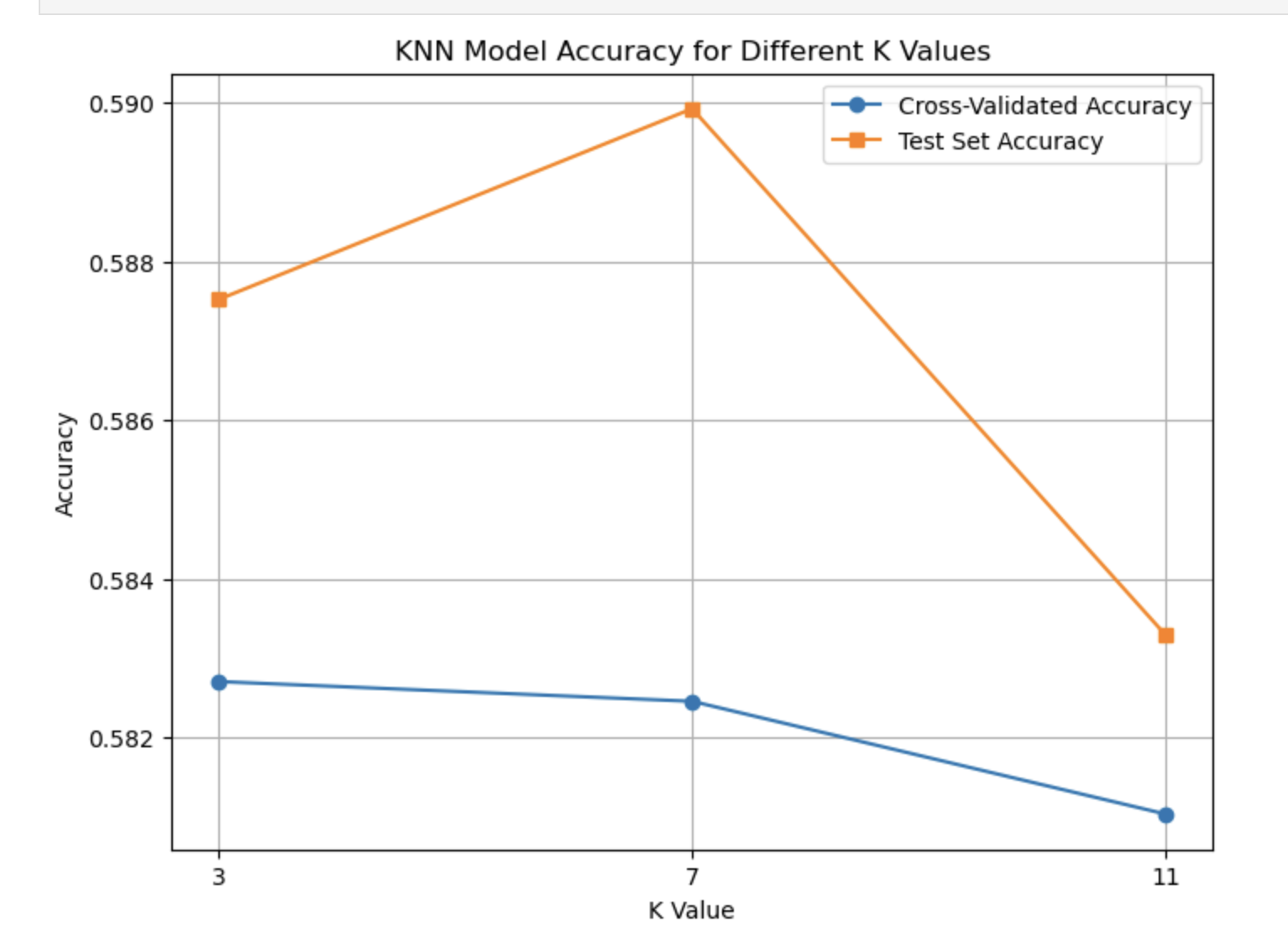
To evaluate the impact of different K values on the performance of the K-Nearest Neighbors (KNN) model, we tested K values of 3, 7, and 11. Both cross-validated accuracy (using 5-fold cross-validation) and test set accuracy were calculated for each K value to ensure robust model evaluation.

Cross-Validated Accuracy:

Cross-validated accuracy provides a stable performance estimate by averaging accuracy scores across five folds. The plot demonstrates that as K increases, cross-validated accuracy improves slightly, indicating that larger K values produce a more stable model that generalises well across different subsets of the training data.

Test Set Accuracy:

Test set accuracy also shows improvement with larger K values, with K=11 achieving the highest test accuracy among the values tested. This trend suggests that a higher K value may reduce the model's sensitivity to individual data points, providing a balanced and less noisy prediction. This finding aligns with the goal of using KNN to capture broader play-type trends in the NFL dataset, where a higher K smooths out minor variances.



K=3 provides a more localised decision boundary, capturing specific details but potentially increasing noise.

K=7 offers a balance, with improved cross-validated and test accuracy.

K=11 yields the highest accuracy, providing a generalised, stable model.

The analysis indicates that K=11 is the most effective choice for this dataset, achieving high accuracy on both cross-validation and the test set. This suggests that the model benefits from this higher K value's broader contextualization of play-type patterns. This plot highlights how tuning K can optimise the model’s predictive power and stability.

The model was evaluated using K values of 3, 7, and 11. For each K, cross-validated accuracy remained around 58%, with minor fluctuations, indicating consistent model performance. Test set accuracy was highest for K=7 at 58.99%, suggesting that this value may provide an optimal balance between capturing detailed patterns and reducing noise.

The classification report shows similar precision, recall, and F1-scores across both classes (0 for run and 1 for pass), with scores slightly favouring class 1 at higher K values. Overall, the model maintained balanced performance across both classes, with K=7 yielding the best accuracy.

**Conclusion:**

In this analysis, we developed a K-Nearest Neighbors (KNN) model to predict NFL play types (run or pass) based on features such as yards\_to\_go, yard\_line, and score\_diff. The model achieved its highest test accuracy of 58.99% with K=7, indicating that the selected features and K value provided a balanced approach to capturing game patterns. Data preprocessing steps, including outlier capping, standardisation, and class balancing with SMOTE, significantly contributed to model stability and performance.

The classification reports revealed that the model performed similarly across both play types, with precision, recall, and F1-scores around 58-60%. While the accuracy is modest, the model successfully identifies trends that could inform real-time decision-making. Future enhancements could involve experimenting with advanced models, such as ensemble methods, and adding more game-contextual features to improve predictive accuracy and better capture complex play-calling strategies.

**References:**

1.Deryckel, F. (n.d.). K-Nearest Neighbors (KNN) - Chapter on Machine Learning with R. GitHub Pages. <https://fderyckel.github.io/machinelearningwithr/knnchapter.html>

2.OpenAI. (2023, March 1). ChatGPT: Optimising language models for dialogue. OpenAI. <https://openai.com/index/chatgpt/>