**Football Data Analysis**

**Predictive Analytics,**

**Introduction:**

Predicting play types in NFL games is crucial for optimising team strategies and improving performance. Coaches rely on understanding play tendencies to refine offensive and defensive tactics, while analysts and fans seek deeper insights into game dynamics. Accurate predictions can also enhance live commentary and sports betting experiences. This study focuses on developing a predictive model to classify play types, such as passes and runs, using game features like down, yard line position, score differential, and remaining time. By leveraging data analytics and machine learning, the research aims to provide actionable insights, demonstrating the value of data-driven strategies in professional football.

**Data Cleaning:**

For cleaning the training dataset, I prioritised data accuracy and interpretability by removing irrelevant columns such as 'Points Scored by Either Team' and 'Yards Gained'. These columns were deemed non-essential for predicting play types. For the 'Down' column, which had less than 1% missing values, I filled the gaps using the mode to maintain the dataset’s integrity without significant data loss.

Outliers in numerical columns like 'Down', 'To Go', and 'Score Differential' were addressed using the interquartile range (IQR) method. Extreme values were capped at the nearest valid boundary, ensuring data consistency while avoiding distortions. Categorical variables, including 'Offensive Team' and 'Defensive Team', were transformed into numerical codes using label encoding to ensure compatibility with machine learning models. Additionally, numerical columns such as 'Yard Line 0-100' were standardised to achieve uniformity in scale and prevent skewed model outcomes.

Feature engineering significantly enhanced the predictive power of the dataset. The 'IS\_RED\_ZONE' feature, a binary indicator, was introduced to identify plays occurring within 20 yards of the end zone, emphasising high-pressure scenarios. The 'ABS\_SCORE\_DIFFERENTIAL' feature was created to represent the absolute score difference, highlighting game situations where urgency might influence play type decisions. These features added contextual depth, improving the model’s ability to discern patterns.

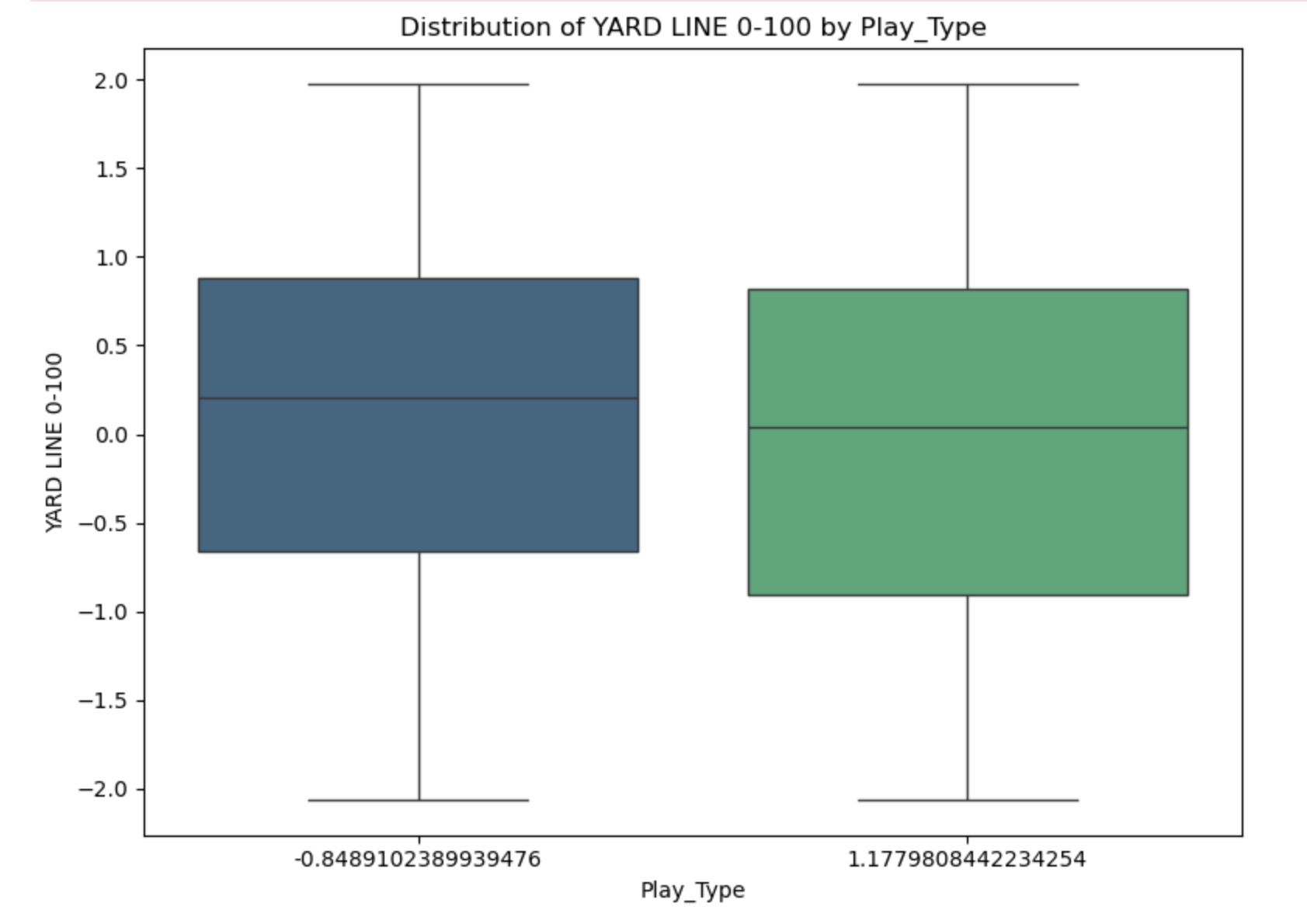
For the testing dataset, I applied the same cleaning process to ensure alignment with the training dataset. Columns like 'Road Team', 'Home Team', and 'Touchdown Details' were removed to match the training dataset's structure. Missing values in the 'Down' column were filled with the mode, and outliers in numerical features were capped using the IQR method. Additionally, the 'Remaining Time in the Quarter' column was converted into seconds, and categorical variables were label-encoded for consistency.

By harmonising the training and testing datasets, standardising numerical features, and incorporating relevant engineered variables, I ensured the data was prepared for robust predictive modelling. These steps minimised biases, improved model generalizability, and preserved the integrity of the datasets.

**EDA:**

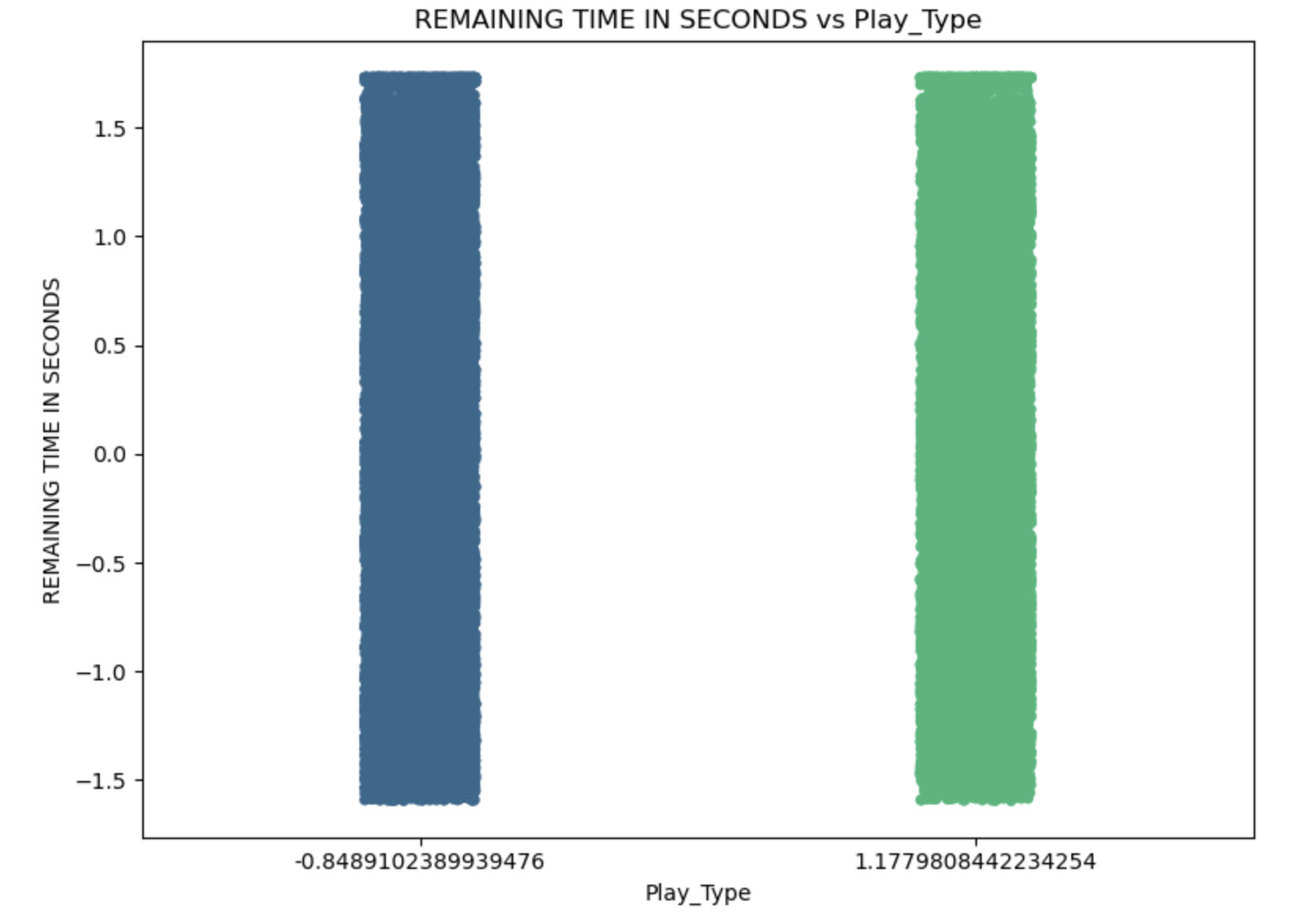
For this exploratory data analysis, I explored the relationship between various game features and play types (pass or run). The goal was to uncover patterns and trends in how factors like downs, yard lines, and score differentials influence play-calling decisions. Visualisations and statistical summaries were used to provide insights into these relationships.

**PLAY TYPE DISTRIBUTION**

The overall distribution of play types reveals a higher frequency of passing plays compared to running plays. Passing plays are labelled as 0, and running plays as 1. This distribution aligns with the general trend of modern football strategies favouring passing due to its versatility and higher potential for significant yard gains. Running plays, on the other hand, are situational and often used to manage time or maintain field control.

**HYPOTHESIS 1: Teams closer to the end zone are more likely to run than pass.**

I hypothesised that plays in the red zone (defined as within 20 yards of the end zone) would favour running over passing due to the reduced field length and increased focus on ball security. The binary IS\_RED\_ZONE variable was created to test this hypothesis.

The results supported this hypothesis: running plays were more frequent within the red zone, emphasising the tactical shift to minimise turnovers and capitalise on scoring opportunities. Outside the red zone, passing plays dominated, reflecting a strategy to cover more ground.

**HYPOTHESIS 2: Later downs favour passing plays due to higher pressure.**

I hypothesised that as downs progress, the likelihood of passing increases due to the urgency to gain yardage. The analysis confirmed this trend, with first downs showing a higher proportion of running plays, aligning with the strategy to establish manageable yardage early in the sequence. On third and fourth downs, passing plays were significantly more frequent, highlighting the pressure to secure a first down or score.

**FEATURE SELECTION**

In the correlation matrix, the most notable relationship with play type was observed with down (-0.18), indicating a slight preference for passing plays on later downs. Yard Line 0-100 showed a weak correlation (0.10), suggesting a minor tendency for passing plays to occur farther from the end zone. Variables like ABS\_SCORE\_DIFFERENTIAL and REMAINING TIME IN SECONDS had negligible correlations, indicating that scoring pressure and time constraints do not heavily influence play type decisions.

During the VIF analysis, several features demonstrated high multicollinearity, particularly the home team's accumulated score, the road team's accumulated score, and the score differential. These features had extremely high VIF values (above 50), indicating redundancy due to their strong correlation, as confirmed in the correlation matrix. Since the score differential was directly derived from the accumulated scores of the two teams, it was removed to address multicollinearity and simplify the feature set.

After eliminating the redundant features, the VIF values for the remaining features significantly improved. Features like 'Down' (VIF = 1.21), 'To Go' (VIF = 1.25), and 'Offensive Team Venue' (VIF = 1.02) exhibited minimal collinearity and were retained for modelling. These features contribute uniquely to the predictive model by capturing different aspects of game dynamics.

The finalised feature set includes 'Quarter', 'Down', 'To Go', 'Yard Line 0-100', 'Road Team's Accumulated Score', 'Home Team's Accumulated Score', 'Offensive Team', 'Defensive Team', 'IS\_RED\_ZONE', and 'Offensive Team Venue'. Collectively, these features provide a comprehensive view of the game context while maintaining low multicollinearity, ensuring the robustness and interpretability of the predictive model.

**Model Evaluation:**

**Model 1 – Logistic Regression**

Logistic regression was chosen for its robustness in modelling categorical outcomes, such as predicting play type (pass or run). Using the selected features, the logistic regression model achieved an accuracy of 60.1%, demonstrating moderate performance. This provides a solid baseline for predicting play types, though there is room for improvement.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Recall** | **F1-Score** | **Support** |
| 0 (Pass) | 0.55 | 0.62 | 3955 |
| 1 (Run) | 0.70 | 0.60 | 2837 |
| **Overall Accuracy**: |  |  |  |

The confusion matrix revealed that the model correctly classified 2167 passing plays (0) and 1976 running plays (1), but misclassified 1788 passing plays as running plays and 861 running plays as passing plays. This imbalance is evident in the precision, recall, and F1-scores for both classes. Passing plays (0) had a higher precision of 0.72, recall of 0.55, and F1-score of 0.62, outperforming running plays (1), which had a precision of 0.52, recall of 0.70, and F1-score of 0.60. The model's superior performance in predicting passing plays is likely due to their higher frequency in the dataset.

importance.

|  |  |  |  |
| --- | --- | --- | --- |
| Actual/Predicted | Predicted: 0 (Pass) | Predicted: 1 (Run) | Total |
| Actual: 0 (Pass) | 2167 | 1788 | 3955 |
| Actual: 1 (Run) | 861 | 1976 | 2837 |
| Total | 3028 | 3764 | 6792 |

The logistic regression model identified key features such as 'Down', 'To Go', and 'Yard Line Position' as significant predictors of play type. These features were crucial in the model's ability to differentiate between passing and running plays. However, the performance metrics indicated that the model is more effective at predicting passing plays, likely due to their higher frequency in the dataset. This highlights the need for additional techniques, such as further feature engineering or alternative modelling approaches, to improve the model's accuracy in predicting running plays.

**Model2:Stepwise Logistic Regression**

Using forward selection, I iteratively refined the logistic regression model by adding features with significant p-values below 0.05. The final model included key features such as TO GO, DOWN, REMAINING\_TIME\_SECONDS, YARD LINE 0-100, OFFENSIVE TEAM, and DEFENSIVE TEAM, along with the intercept (const). The model achieved an accuracy of 60.2%.

Classification Report Summary:

Class Precision Recall F1-Score Support

0 (Pass) 0.61 0.91 0.73 292

1 (Run) 0.55 0.15 0.23 201

Overall Accuracy: 60.2%

Macro Avg Precision: 0.58

Weighted Avg F1-Score: 0.53

The confusion matrix highlights that the model performs better on passing plays (0):

Precision: 0.61

Recall: 0.91

F1-Score: 0.73

In contrast, its performance on running plays (1) is weaker:

Precision: 0.55

Recall: 0.15

F1-Score: 0.23

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| Actual/Predicted | Predicted: 0 (Pass) | Predicted: 1 (Run) | Total |
| Actual: 0 (Pass) | 266 | 26 | 292 |
| Actual: 1 (Run) | 171 | 30 | 201 |
| Total | 437 | 56 | 493 |

Model Comparison Table:

|  |  |  |
| --- | --- | --- |
| Metric | Logistic Regression | Stepwise Logistic Regression |
| Accuracy | 60.99% | 60.24% |
| Precision (Class 0) | 0.72 | 0.61 |
| Recall (Class 0) | 0.55 | 0.91 |
| F1-Score (Class 0) | 0.62 | 0.73 |
| Precision (Class 1) | 0.52 | 0.55 |
| Recall (Class 1) | 0.70 | 0.15 |
| F1-Score (Class 1) | 0.60 | 0.23 |

When comparing the hyperparameter-tuned logistic regression model to the stepwise logistic regression model, notable differences emerged in class-specific performance. While overall accuracy was similar (60.99% for the tuned model and 60.24% for the stepwise model), the models exhibited different strengths depending on the play type.

For passing plays (Class 0), the stepwise model excelled with a high recall of 91%, correctly identifying most actual passing plays. However, its precision was lower at 61%, indicating frequent misclassification of running plays as passing plays. This combination resulted in a strong F1-score of 73%. In contrast, the tuned model demonstrated more balance, achieving a higher precision of 72% but a lower recall of 55%, leading to a moderately lower F1-score of 62%.

For running plays (Class 1), the tuned model outperformed the stepwise model significantly, achieving a recall of 70% compared to only 15% for the stepwise approach. This indicates the tuned model was better at identifying running plays, with an F1-score of 60% versus 23%.

The stepwise model simplifies analysis by focusing on key features like TO GO, DOWN, and field position, making it easier to interpret. However, the hyperparameter-tuned model's balanced performance and ability to handle class imbalance make it the preferred choice for predicting both play types effectively.

**Testing Results on Test Dataset**

I tested the trained logistic regression model on the test dataset to evaluate its performance in predicting passing and running plays. The model achieved an accuracy of 60.1%, meaning it correctly predicted the play type in 60.1% of the cases.

#### **Metrics Summary:**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Class 0 (Pass)** | **Class 1 (Run)** |
| **Precision** | 74% | 50% |
| **Recall** | 52% | 72% |
| **F1-Score** | 61% | 59% |
| **Accuracy** | **60.1%** | - |

The confusion matrix shows the model correctly predicted 514 passing plays and 470 running plays but misclassified 472 passes as runs and 182 runs as passes. These results indicate the model struggles to achieve a balance in predicting both play types, performing better on running plays in terms of recall but excelling in precision for passing plays.

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix | Predicted: Pass | Predicted: Run | Total |
| Actual: Pass | 514 | 472 | 986 |
| Actual: Run | 182 | 470 | 652 |
| Total | 696 | 942 | 1638 |

**Final Insights of the Model:**

**Key Variables and Their Impact**

The analysis highlights Play Type as the primary outcome, influenced by variables such as TO GO, DOWN, and YARD LINE 0-100. These predictors provide critical insights into whether a play is more likely to be a pass or a run. TO GO represents the yards required for a first down, with passing plays more common in long-yardage situations, reflecting the need for greater gains. DOWN is a pivotal factor, with later downs (e.g., 3rd or 4th) favouring passing plays due to time constraints and situational urgency. Conversely, running plays are more frequent in short-yardage situations or closer to the end zone, emphasising clock control or short gains. These variables reflect real-world play-calling strategies and are essential for making accurate predictions.

**Model Performance**

The logistic regression model, trained using hyperparameter tuning and evaluated on test data, achieved an accuracy of 60.1%, indicating moderate success in predicting Play Type. For passing plays (Class 0), the model exhibited a precision of 74%, meaning the majority of predicted passes were correct. However, the recall was lower at 52%, indicating some passing plays were misclassified as runs. For running plays (Class 1), the model achieved a precision of 50% and a recall of 72%, demonstrating better identification of actual runs but with more false positives. The model’s balanced F1-scores of 61% (passing plays) and 59% (running plays) highlight its ability to generalise across both classes but reveal room for improvement, particularly in addressing class imbalance.

**Application to Game Strategy and Decision-Making**

The model offers valuable insights for strategic decision-making in football. By analysing predictors such as TO GO, DOWN, and YARD LINE 0-100, coaches can anticipate opponents' likely play types and adjust their strategies accordingly. For example, defensive alignments can focus on defending the pass during long-yardage situations or prioritising the run near the end zone. Offensive strategies can also benefit from identifying patterns in the opponent's play-calling tendencies. Enhancing the model with additional features like player statistics, weather conditions, or team-specific dynamics could further improve its reliability and applicability to real-world game strategies.

**CONCLUSION**

In conclusion, the logistic regression model provides meaningful insights into Play Type and its relationship with situational variables. While the model demonstrated reasonable accuracy on the test data, its performance revealed areas for refinement, including improving predictions for passing plays and addressing class imbalance. The findings underscore the potential of data-driven approaches to inform play-calling strategies, while highlighting the importance of iterative enhancements to maximise the model's effectiveness and generalizability.

Appendix: