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

To prepare the training dataset for predictive modelling, I prioritised data accuracy, consistency, and compatibility. Missing values were handled systematically, and irrelevant columns were removed to streamline the analysis. Columns such as 'POINTS SCORED\nBY EITHER TEAM', 'YARDS GAINED', 'PLAY-ID', 'DRIVE-ID', 'DATE', and 'WEEK#' were deemed non-essential and dropped. For the 'DOWN' column, which had less than 1% missing values, rows containing these entries were removed to maintain data integrity with minimal loss.

Outliers in numerical columns such as 'DOWN', 'TO GO', and 'SCORE\nDIFFERENTIAL\n(Home Team's Score) \n-\n(Road Team's Score)' were identified using the interquartile range (IQR) method. Extreme values were capped at their nearest valid boundary to ensure data consistency while avoiding distortions in the model's predictions.

Categorical variables, including 'OFFENSIVE TEAM', 'DEFENSIVE TEAM', and 'OFFENSIVE \nTEAM VENUE \nRoad,\nHome,\nNeutral', were transformed into numerical representations using LabelEncoder to ensure compatibility with machine learning algorithms. Additionally, numerical columns such as 'DOWN', 'TO GO', 'YARD LINE 0-100', 'ROAD TEAM'S ACCUMULATED SCORE', 'HOME TEAM'S ACCUMULATED SCORE', and 'Time in Seconds' were standardised using StandardScaler, achieving uniformity in scale and preventing skewed results caused by varying numerical ranges.

Feature engineering played a pivotal role in improving the dataset's predictive power. Several new features were introduced:

1.POSSESSION_ADVANTAGE: A binary feature indicating whether the offensive team was playing at home.

2.MOMENTUM: Calculated as the score differential between the home and road teams, capturing the dynamic state of the game and supporting predictions of strategic adjustments.

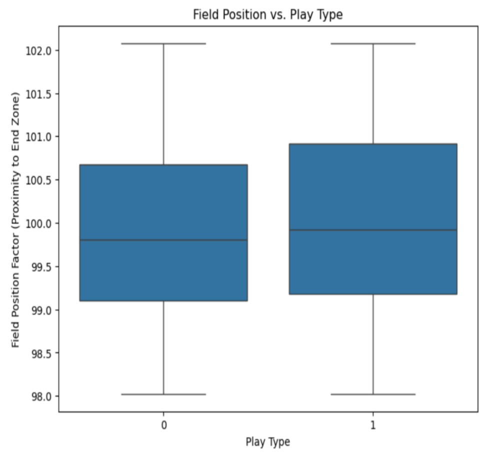
3.FIELD_POSITION_FACTOR: A measure of proximity to the opponent's end zone, computed as 100 - YARD LINE 0-100.

4.TIME_PRESSURE: Calculated as the inverse of the remaining time in seconds, highlighting the increasing urgency as the game progresses.

5.TEAM_MISMATCH: Representing the absolute difference between offensive and defensive team IDs, offering insights into potential mismatches.
6.DOWN_IMPORTANCE: A composite metric derived as 'DOWN' * 'TO GO', emphasising the significance of each down.
The testing dataset underwent the same cleaning and preprocessing steps to ensure alignment and compatibility with the training dataset. Irrelevant columns, including 'points scored by either team', 'road team', 'date', 'week#', 'yards gained', and other play details such as 'pass outcome' and 'touchdown details', were removed. Rows with missing values in the 'DOWN' column were dropped, while outliers in numerical columns were capped using the IQR method. Categorical variables were label-encoded, and the 'REMAINING TIME IN THE QUARTER (mm:ss)' column was converted into seconds for consistency.

By harmonising the training and testing datasets through rigorous cleaning and transformation steps, both datasets were prepared for effective and robust predictive modelling. This process minimised biases, improved model generalizability, and ensured data integrity, providing a solid foundation for accurate and reliable analysis.

Exploratory Data Analysis (EDA)
For this exploratory data analysis, I investigated the relationship between key game features and play type (pass or run). The objective was to understand how factors such as field position and game downs influence play-calling decisions. The analysis aimed to provide insights into strategic adjustments made by teams based on in-game situations.



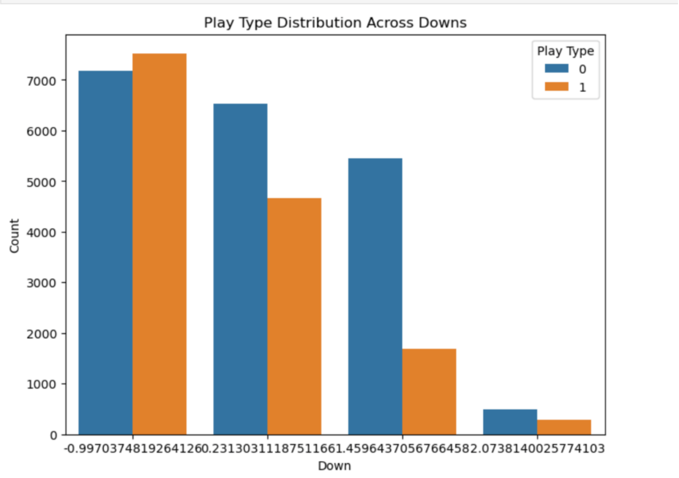
Field Position vs. Play Type
To explore the relationship between field position and play type, I plotted the Field Position Factor (proximity to the opponent's end zone) against the Play Type. The hypothesis was that field position would influence whether teams opt for passing or running plays.

Passing plays were more frequent across all field positions, with a slightly greater occurrence in mid-field regions.
Running plays were more concentrated closer to the end zone, potentially reflecting a strategy to minimise risk and control the game clock in high-pressure situations.
This analysis suggests that teams adapt their play types based on their field position, using running plays more strategically in situations closer to the opponent’s end zone.

Play Type Distribution Across Downs:
I further examined the distribution of play types across different downs using a count plot. The hypothesis was that earlier downs would favour running plays for safer yard gains, while

later downs would see more passing plays due to the higher pressure to gain significant yardage.

Running plays were most frequent on the first down, likely reflecting a safer strategy to establish the drive early.



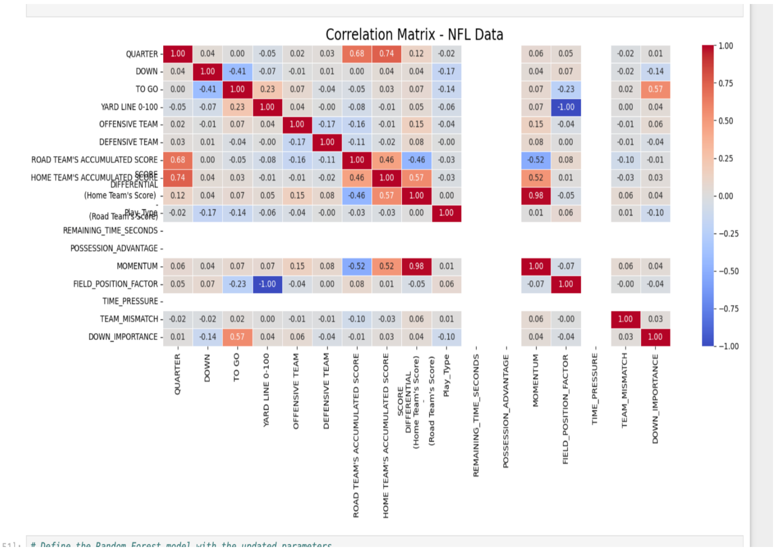
On third and fourth downs, passing plays became dominant, aligning with the increased urgency to gain yardage in these critical situations.

The second down showed a more balanced distribution of play types, highlighting strategic flexibility depending on the outcome of the first down.

This pattern reinforces the idea that teams make deliberate adjustments in their play-calling strategy as they progress through the downs.

CORRELATION MATRIX:

The correlation matrix revealed a notable relationship between 'Down' and 'Play_Type' with a correlation of -0.20, indicating a slight tendency for passing plays to increase as teams progress to later downs. This reflects the strategic need to cover more yardage under pressure on third or fourth downs. Other features like 'Momentum' and 'Time_Pressure' showed minimal correlations with 'Play_Type', suggesting that immediate factors like down and distance are prioritised over broader game scenarios such as score differential or remaining time. These insights emphasise the tactical adjustments teams make based on situational urgency rather than overall game context.



```
# Define the Random Forest model with the updated parameters
```

Model evaluation:

Model 1: Random Forest

See Appendix Chart 1 and Chart 2 for Confusion Matrix and Feature Importance.

Random Forests are a versatile and reliable machine learning model, particularly known for their robustness against overfitting and ability to handle diverse data types. In this analysis, the Random Forest model was evaluated for its ability to classify play types effectively. The model achieved an overall accuracy of 70%, reflecting strong predictive capabilities across the dataset.

Classification Report

Class	Precision	Recall	F1-Score
0	0.82	0.62	0.71
1	0.61	0.82	0.70
Overall	**0.73**	**0.70**	**0.70**

The confusion matrix revealed that for class 0 (e.g., passing plays), the model achieved a precision of 82%, demonstrating its strong ability to avoid false positives, with a recall of 62%, indicating it correctly identified 62% of actual passing plays. For class 1 (e.g., running plays), the model effectively identified most instances with a recall of 82%, though precision was slightly lower at 61%. The top three features driving the model's predictions were Field Position Factor (Yard Line 0-100), which played the most significant role by capturing proximity to the opponent's end zone, Down, which influenced decisions particularly in later downs, and To Go, which highlighted the importance of the distance required for a first down. Other notable features, such as Momentum and Time Pressure, added contextual depth to the model's understanding of game dynamics. Overall, the Random Forest model leveraged these key features to provide meaningful insights into play-calling strategies, demonstrating balanced performance and strong interpretability.

Model 2: Gradient Boosting

See Appendix Chart 3 and Chart 4 for Confusion Matrix and Feature Importance.

Gradient Boosting is a powerful machine learning algorithm known for its sequential learning approach, where each tree corrects the errors of its predecessor. This method often excels in capturing complex patterns, although it can be sensitive to overfitting. In this analysis, the Gradient Boosting model achieved an overall accuracy of 60%.

Classification Report

Class	Precision	Recall	F1-Score

0	0.65	0.74	0.69
1	0.50	0.39	0.44
Overall	**0.59**	**0.60**	**0.59**

The confusion matrix revealed that for class 0 (e.g., passing plays), the model achieved a precision of 65%, indicating a reasonable ability to avoid false positives, with a recall of 74%, successfully identifying most passing plays. For class 1 (e.g., running plays), the recall was lower at 39%, showing that the model struggled to correctly identify all instances, while precision was 50%, reflecting moderate reliability in predictions. The most influential features in the Gradient Boosting model were Field Position Factor (Yard Line 0-100), which was critical for reflecting proximity to the opponent's end zone, To Go, which influenced the decision to run or pass based on distance required for a first down, and Down, which significantly impacted strategies on later downs. Additional features such as Momentum and Time Pressure contributed to a lesser extent, providing contextual depth to the predictions. Overall, while the Gradient Boosting model achieved an accuracy of 60%, it was less effective than the Random Forest model in balancing precision and recall across both play types. However, its ability to capture complex feature interactions highlighted valuable insights into play-calling strategies.

Model Comparison

Model	Class	Precision	Recall	F1-Score
Random Forest	0	0.82	0.62	0.71
	1	0.61	0.82	0.70
Gradient Boosting	0	0.65	0.74	0.69
	1	0.50	0.39	0.44

From the comparison, the Random Forest model outperformed Gradient Boosting in most metrics, particularly for class 1 (running plays). It achieved higher precision (61%) and recall (82%) compared to Gradient Boosting 50% precision and 39% recall, making it more effective at identifying running plays. For class 0 (passing plays), Gradient Boosting had slightly higher recall (74%) than Random Forest (62%), but Random Forest showed better precision (82% vs. 65%), minimising false positives. Overall, Random Forest provided more balanced performance and reliability across play types, making it the preferred model for accurate play type predictions.

Final model:

After extensive evaluation, I have selected Random Forest as the final model for this classification task due to its superior performance in balancing precision, recall, and F1-score compared to other models. Random Forest demonstrated robustness across both play types, particularly excelling in class 1 (e.g., running plays), where it achieved high recall, crucial for minimising false negatives. Its ability to handle imbalanced data using the class_weight='balanced' parameter, combined with its feature importance insights, makes it an ideal choice for understanding and predicting play types effectively.

Results on Test Dataset:

Class	**Precision**	**Recall**	**F1-Score**
0	0.86	0.65	0.74
1	0.61	0.84	0.71
Overall	**0.73**	**0.73**	**0.73**

	Predicted 0	**Predicted 1**
True 0:	639	**347**
True 1:	104	**548**

The Random Forest model identified key features influencing play type predictions, with Down Importance—a composite metric combining down and distance to go—emerging as the most significant, emphasising its crucial role in game strategy. To Go, representing the distance required for a first down, heavily influenced decisions between passing and running plays, while Field Position Factor (Yard Line 0-100) highlighted the importance of proximity to the opponent's end zone in shaping play calls. Additionally, Down progression and Momentum, capturing score differential, provided vital context for understanding whether teams were in offensive or defensive scenarios. The model achieved an accuracy of 73% on the test dataset, with balanced performance in both precision and recall across play types. Its ability to effectively analyse key game dynamics and interpret feature importance makes it a reliable tool for predicting play types and analysing football strategies, offering valuable insights for practical applications in football analytics.

SIGNIFICANT VARIABLES:

The Random Forest model highlighted several significant variables influencing play type predictions. Down Importance, combining down and distance to go, was the most critical, emphasising its role in strategic decisions. To Go, representing the distance needed for a first down, and Field Position Factor (Yard Line 0-100), reflecting proximity to the opponent's end zone, were also highly impactful. Down progression influenced play-calling strategies, particularly on later downs, while Momentum, capturing the score differential, provided context on offensive or defensive situations. These variables collectively enhanced the model's ability to predict play types and offered valuable insights into game dynamics.

MODEL EXPECTATIONS AND IMPROVEMENTS:

The Random Forest model successfully met expectations by achieving an accuracy of 73% on the test dataset, demonstrating a balanced performance in precision and recall for both play types. Key features such as Down Importance, To Go, and Field Position Factor emerged as critical in driving strategic play-calling decisions. The model proved robust in identifying patterns within the dataset while remaining interpretable, making it a reliable choice for football analytics.

To enhance performance further, fine-tuning hyperparameters, such as increasing the number of trees or exploring deeper tree structures, could yield better results. Introducing advanced feature engineering, including dynamic game context or additional historical play data, might refine predictions and improve model accuracy. Additionally, experimenting with alternative models like XGBoost or ensemble stacking could complement the Random Forest model by leveraging its strengths in combination with other approaches. These improvements would provide deeper insights and greater accuracy for predicting play types.

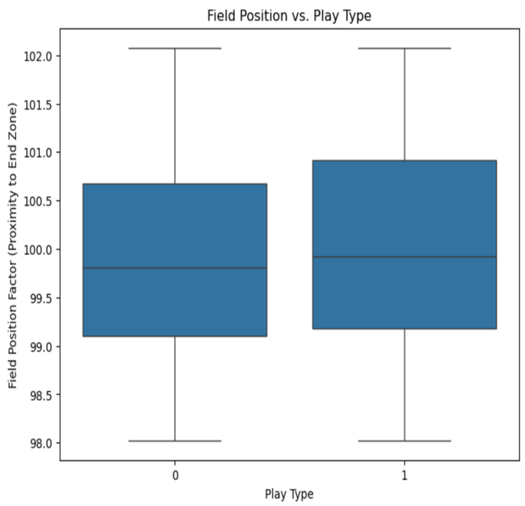
Conclusion

The analysis and modelling process demonstrated that the Random Forest model is the most effective approach for predicting play types in football analytics, achieving a test accuracy of 73%. This model provided valuable insights into the key features influencing play-calling decisions, such as Down Importance, To Go, and Field Position Factor, which are critical for understanding game dynamics. Its balanced performance across both precision and recall ensured reliability in predicting both passing and running plays.

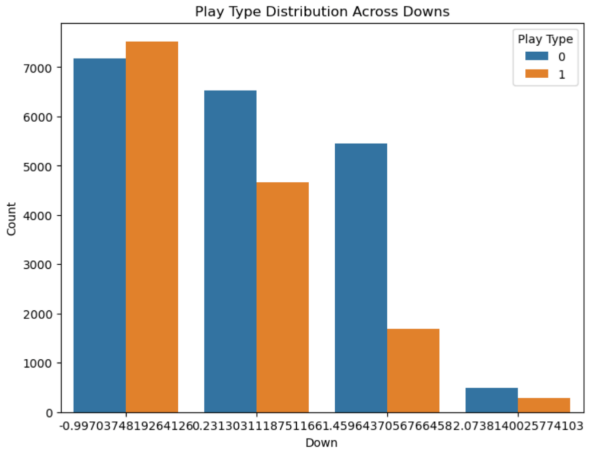
While the model met expectations, there remains room for improvement through fine-tuning, advanced feature engineering, and exploring alternative algorithms like XGBoost. Nonetheless, the current implementation offers a robust, interpretable, and practical solution for analysing play strategies, paving the way for further advancements in football analytics. This model has the potential to provide actionable insights for decision-making and strategy optimization in real-world applications.

Appendix:

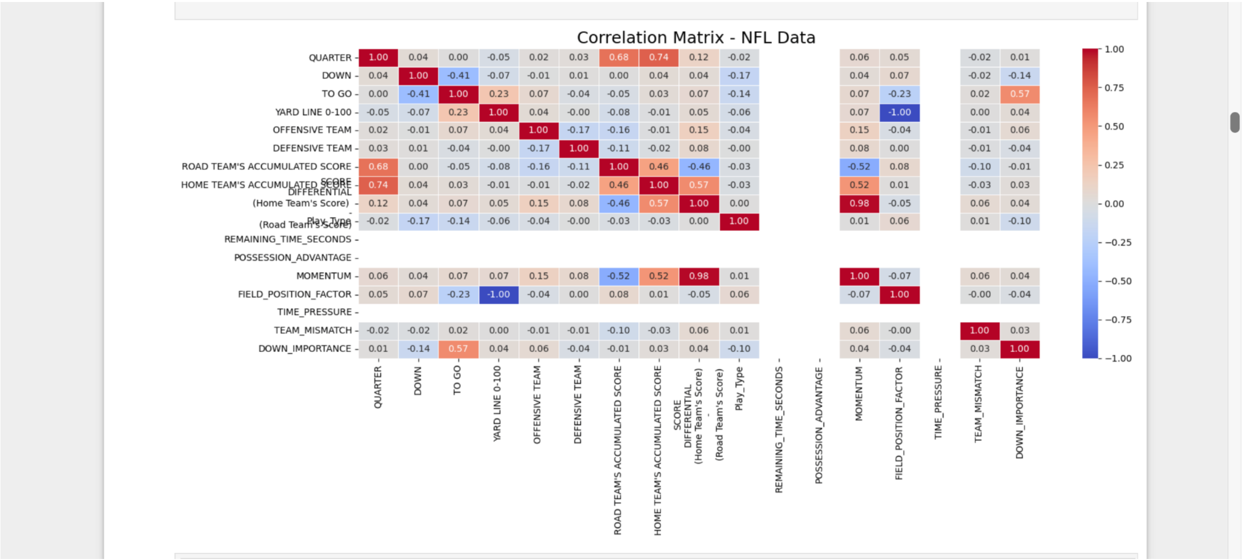
NFL GAMES



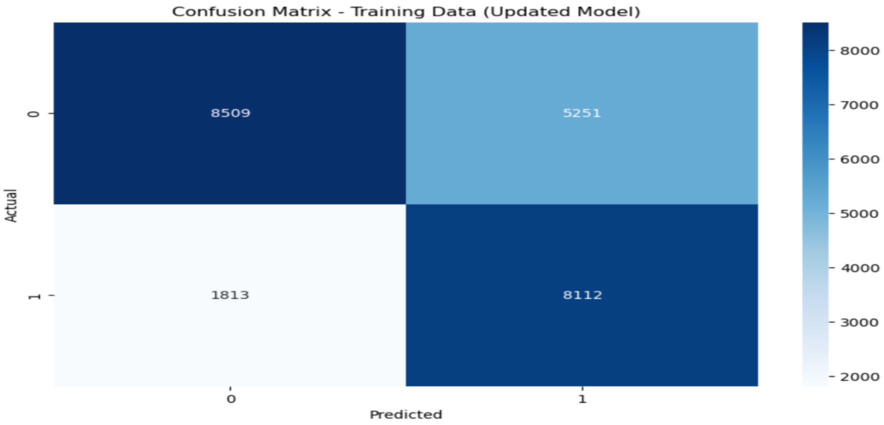
Field Position vs. Play Type

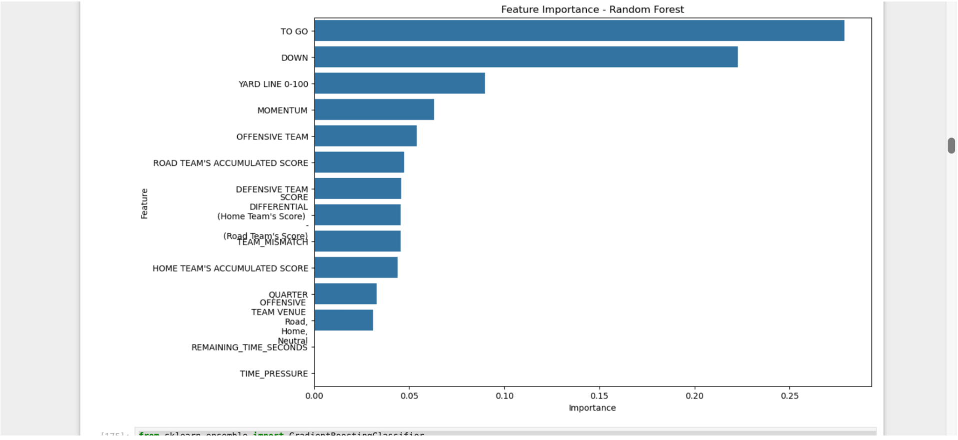


Play Type Distribution Across Downs

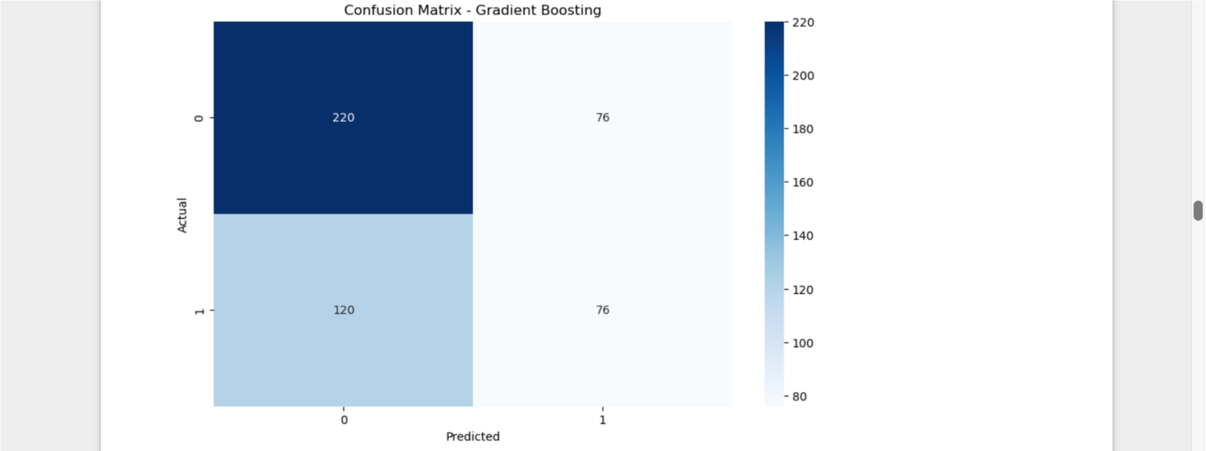


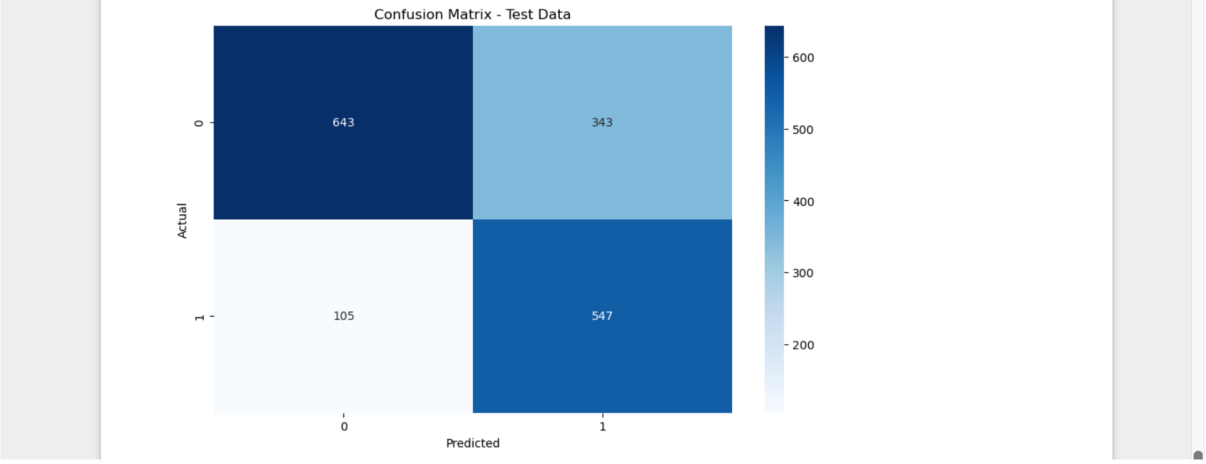
Correlation Matrix - NFL Data

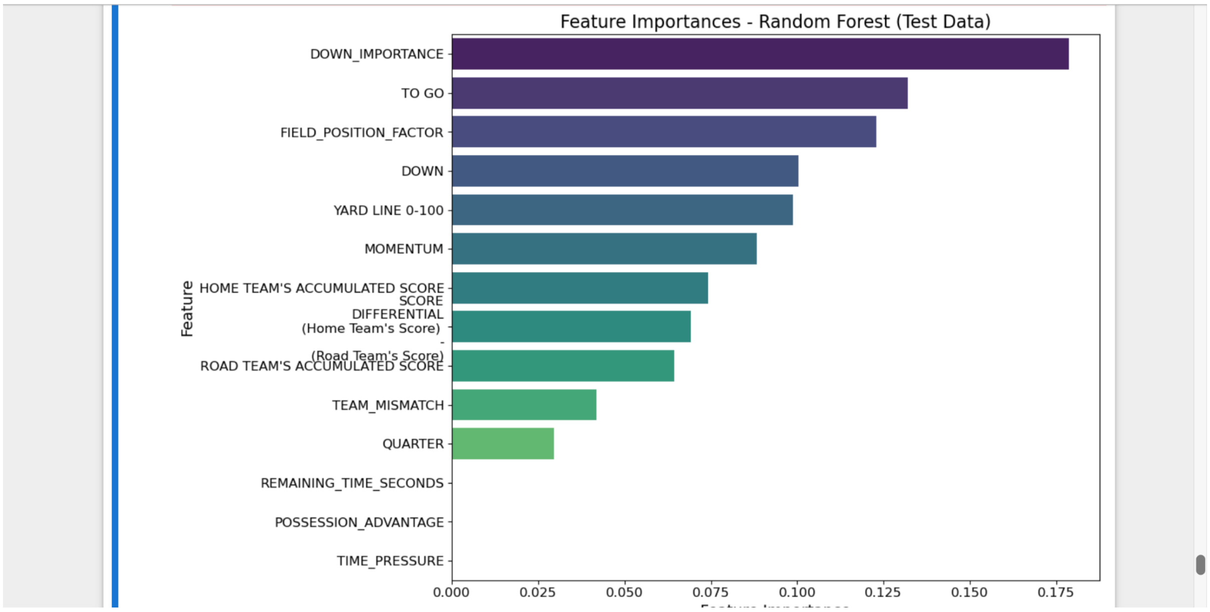




```
[175]: from sklearn.ensemble import GradientBoostingClassifier
```


Confusion Matrix - Gradient Boosting
Actual
0
1
220
76
120
76
0
1
Predicted
220
200
180
160
140
120
100
80


Confusion Matrix - Test Data
Actual
0
1
643
343
105
547
0
1
Predicted
600
500
400
300
200


Feature Importances - Random Forest (Test Data)
Feature
DOWN_IMPORTANCE
TO GO
FIELD_POSITION_FACTOR
DOWN
YARD LINE 0-100
MOMENTUM
HOME TEAM'S ACCUMULATED SCORE
SCORE DIFFERENTIAL (Home Team's Score) - (Road Team's Score)
ROAD TEAM'S ACCUMULATED SCORE
TEAM_MISMATCH
QUARTER
REMAINING_TIME_SECONDS
POSSESSION_ADVANTAGE
TIME_PRESSURE
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175