ASSIGNMENT 4 OPTIMAL PLAY TYPE PREDICTOR

ALY 6020 PROFESSOR: JUSTIN GROSZ 24/11/24

Akhila Singaraju

**INTRODUCTION:**

In this analysis, I developed a logistic regression model to predict play types (pass or run) and assessed its performance on training and testing datasets. I examined the model’s strengths, highlighted areas for improvement, and explored how the results could be used to inform game strategies.

**DATA CLEANING:**

For cleaning the training dataset, I started by ensuring the accuracy of data interpretation, focusing on handling missing values effectively. I removed columns that were not critical for my analysis. Specifically, I removed the 'points scored by either team' because it was not important for the analysis, and Yards Gained', which was also irrelevant. Additionally, the Down column, crucial for my analysis but with less than 1% missing values, led me to remove only the rows containing these missing entries, maintaining the robustness of my dataset without significant data loss.

Outliers in numerical columns were addressed using the interquartile range (IQR) method, capping them at the nearest valid boundary. I converted categorical variables into numerical codes to ensure compatibility with analytical models. Additionally, I standardized key numerical columns like to ensure uniformity and avoid skewed results from differing scales.

I created the 'MOMENTUM' feature by calculating the score differential between the home and road teams. This feature effectively captures the dynamic state of the game by quantifying the current lead, which is instrumental in predicting how teams might adjust their strategies. By assessing which team is ahead, the feature helps model decisions around whether to adopt a more defensive posture or to press the advantage with aggressive plays, providing a clear numerical value that reflects the game's balance of power.

Another feature I introduced is 'TIME\_PRESSURE,' computed as the inverse of the remaining game time in seconds. This feature highlights the increasing urgency as the game progresses, particularly as it approaches the final moments. The lesser the time left, the higher the pressure, influencing teams to alter their tactics—either securing a lead or catching up. Additionally, the 'TEAM\_MISMATCH' feature, measuring the absolute difference between offensive and defensive team IDs, offers insights into potential team mismatches. This metric aids in predicting outcomes by evaluating the relative strengths and weaknesses of the opposing teams, adding a layer of strategic depth to the analysis.

For the testing dataset, I applied the same cleaning process as used for the training dataset to ensure consistency and compatibility. I removed several unnecessary columns that were irrelevant to the analysis and not present in the training dataset. These included 'points scored by either team,' 'road team,' 'date,' 'week#,' 'play-id,' 'drive-id,' 'home team,' 'yards gained,' 'pass outcome complete, incomplete, sack,' 'pass length short, deep,' 'pass location left, middle, right,' 'air yards,' 'yards after catch (YAC),' 'run location left, middle, right,' and 'touchdown pass/offensive TD, rush/offensive TD, return/defensive TD.' This step ensured that both datasets were aligned in structure and content. Additionally, rows with missing values in the 'down' column were dropped, as they accounted for only a minimal portion of the data.

I also addressed outliers in numerical columns like 'to go,' and 'score differential' using the interquartile range (IQR) method, capping values outside the IQR range to maintain data consistency. Categorical variables, such as 'offensive team venue' and 'offensive team,' were transformed into numeric codes using label encoding, while the 'remaining time in the quarter' column was converted into seconds for uniformity and ease of analysis. Feature engineering enhanced the dataset by incorporating 'Momentum,' which measures the score differential between home and road teams, Team Mismatch and 'Time Pressure,' calculated as the inverse of remaining game time. These metrics capture game dynamics effectively, reflecting scoring opportunities and strategic urgency. Additionally, I standardized critical numerical features like 'TO GO' and 'Yard Line 0-100' to ensure consistency across the dataset, enhancing the model’s ability to perform robust predictions while aligning precisely with the training dataset's structure.

**EDA**

For this exploratory data analysis, I investigated the relationship between various game features and play type (pass or run). The goal was to determine how factors such as overall distribution, score margin, and game downs influence play-calling decisions.

PLAY TYPE DISTRIBUTION:

The first graph shows the overall distribution of play types, where passing plays are labeled as 0 and running plays as 1. From the chart, it is evident that passing plays are more frequent than running plays. The difference in frequency also highlights the versatility of passing plays compared to the more situational nature of running plays. (\*\* see chart 1 in appendix for visualization)

HYPOTHESIS 1: Teams trailing in score are more likely to favor passing plays.

I hypothesized that trailing teams would favor passing plays to cover ground quickly, while leading teams might prefer running plays to manage the clock.

A graph of different colored squares

Description automatically generated

However, the results showed no clear relationship between score margin and play type. The counts for passing and running plays remained relatively balanced across all categories, suggesting that score margin alone does not strongly dictate play-calling decisions.

HYPOTHESIS 2: Early downs favor running plays, while later downs favor passing due to higher pressure.

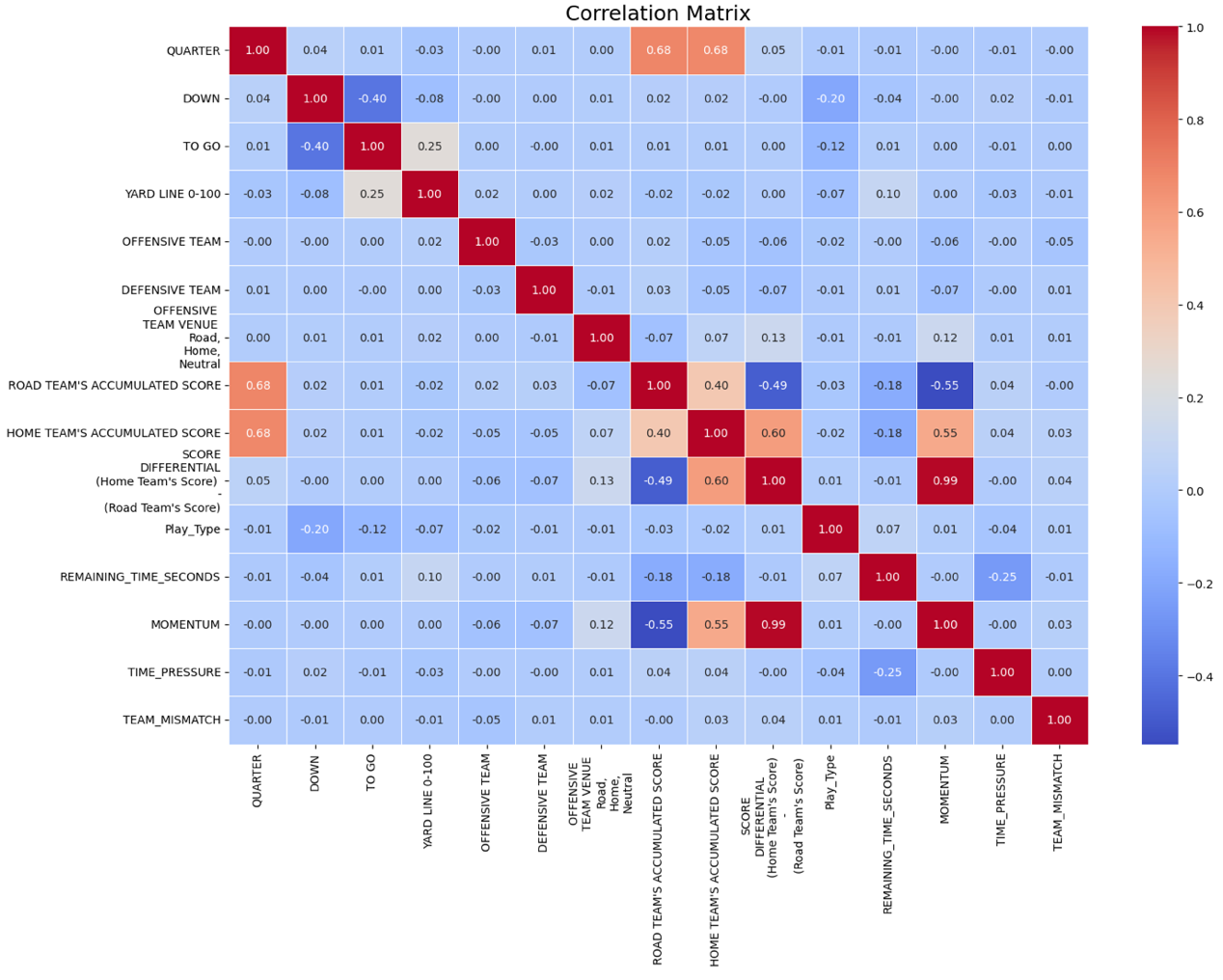
I hypothesized that the first downs favor passing running plays, while later downs favor passing

A graph of different colored bars

Description automatically generated

As expected, running plays were most frequent on first downs, reflecting a safer strategy to establish the drive early in the sequence. On later downs, particularly third and fourth, passing plays dominated, aligning with the hypothesis that teams face greater pressure to gain yardage in these situations. This pattern highlights the strategic adjustments teams make as they progress through a drive.

**CORRELATION MATRIX:**



In the correlation matrix I examined, 'Play\_Type' (pass or run) shows the most notable correlation with 'Down' at -0.20. This correlation suggests that there is a slight tendency for teams to favor passing plays as they progress to later downs. The negative sign of the correlation indicates that as the down number increases (moving from first to fourth), the likelihood of choosing a pass over a run increase.

Regarding other factors like 'Momentum' and 'Time\_Pressure', the correlation coefficients are very close to zero with 'Play\_Type', indicating these aspects of the game scenario have minimal influence on the decision between a pass or a run. Additionally, 'Remaining\_Time\_Seconds' shows a similarly negligible correlation of -0.01 with 'Play\_Type', underscoring that the amount of time left in the game has little direct impact on play-calling. This suggests that coaches and players may prioritize strategies based on down and distance situations over the game clock or the current score dynamics, focusing more on immediate tactical needs rather than the broader game context.

The final variables selected for the model are 'quarter', 'yard line 0-100', 'down', 'to go', 'road team's accumulated score', 'home team's accumulated score', 'offensive team', 'defensive team', 'score differential (home team's score - road team's score)', 'remaining time seconds', 'offensive team venue (road, home, neutral)', 'momentum', 'time pressure', and 'team mismatch'.

**MODEL EVALUATION**

**MODEL 1: RANDOM FOREST: \*see appendix chart 2 and 3 for confusion matrix and feature importance.**

Random forests are a popular choice in machine learning for several reasons, particularly their ability to handle a variety of data types and complexities without extensive data preparation. One of the primary advantages of random forests is their robustness against overfitting, which is common in models that rely on a single decision tree.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** |
| 0 | 0.70 | 0.73 | 0.72 |
| 1 | 0.60 | 0.57 | 0.58 |
| **Accuracy** | **0.6615** |  |  |

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0** | **Predicted 1** |
| True 0 | 4333 | 1618 |
| True 1 | 1818 | 2382 |

In evaluating the performance of my Random Forest model, I closely analyzed its capacity to accurately classify two different outcomes using a given dataset. The model achieved an overall accuracy of approximately 66.15%. This level of accuracy reflects the model's general effectiveness in making correct predictions across all test instances.

The top 3 important features are To go, Down and Yard line 0-100 followed by time pressure and momentum.

Delving into the specifics presented in the classification report, the model demonstrated a precision of 0.70 for class 0 and 0.60 for class 1. This suggests that the model is relatively more reliable in predicting non-events (class 0) compared to events (class 1). The recall scores—0.73 for class 0 and 0.57 for class 1—indicate that the model is better at identifying all relevant instances of class 0 than class 1. The F1-scores, which balance precision and recall, stood at 0.72 for class 0 and 0.58 for class 1.

**MODEL 2: GRADIENT BOOST: \*see appendix chart 4 and 5 for confusion matrix and feature importance.**

Gradient boosting is another powerful ensemble technique used for both classification and regression problems. Unlike random forests, which build trees independently, gradient boosting builds one tree at a time in a sequential manner, where each new tree helps to correct errors made by the previous trees.

| **Class** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- |
| 0 | 0.71 | 0.72 | 0.72 |
| 1 | 0.60 | 0.58 | 0.59 |
| **Accuracy** | **0.6647** |  |  |

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0** | **Predicted 1** |
| True 0 | 4292 | 1659 |
| True 1 | 1744 | 2456 |

In reviewing the performance of my Gradient Boosting model, I analyzed how it performed in classifying two different outcomes within the dataset. The model achieved an overall accuracy of 66.47%, indicating a moderate level of precision in its predictions across the test instances.

The top 3 important features are to go, Down and Yard line 0-100 followed by time pressure and momentum. From this breakdown, it's evident that the model demonstrates higher precision and recall for class 0 (non-events), with scores of 0.71 and 0.72, respectively. For class 1 (events), the precision and recall are slightly lower, at 0.60 and 0.58, respectively. This indicates that the model is somewhat better at identifying non-events than event.

**MODEL COMPARISION:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Class** | **Precision** | **Recall** | **F1-Score** |
| Random Forest | 0 | 0.70 | 0.73 | 0.72 |
| Random Forest | 1 | 0.60 | 0.57 | 0.58 |
| Gradient Boosting | 0 | 0.71 | 0.72 | 0.72 |
| Gradient Boosting | 1 | 0.60 | 0.58 | 0.59 |

From this table, we can observe that both models perform comparably, with Gradient Boosting showing a slight improvement in recall for class 1, suggesting it is marginally better at identifying true positives in the event class than the Random Forest model. Both models have the same precision for class 1, indicating a similar capability in accurately predicting event occurrences. For class 0, the Gradient Boosting model also displays a marginal improvement in both precision and recall, which indicates a slightly better overall performance in correctly identifying non-events compared to the Random Forest model.

After careful evaluation, I have selected Gradient Boosting as the final model for this classification task, primarily due to its slightly superior performance in precision and recall compared to the Random Forest model. This edge is particularly evident in the handling of class 1 predictions, where Gradient Boosting shows improved sensitivity, crucial for minimizing false negatives in many practical scenarios. Additionally, the inherent mechanism of Gradient Boosting, which builds trees sequentially to correct previous errors, effectively manages the bias-variance trade-off, reducing overfitting—a common challenge in complex models. These factors, combined with the model’s ability to adaptively focus on difficult areas of the data, make Gradient Boosting particularly suited for achieving the best overall performance on this dataset.

RESULTS ON TEST DATASET:

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** |
| 0 | 0.75 | 0.67 | 0.71 |
| 1 | 0.57 | 0.65 | 0.61 |
| **Accuracy** | **0.6648** |  |  |

|  | **Predicted 0** | **Predicted 1** |
| --- | --- | --- |
| True 0 | 663 | 323 |
| True 1 | 226 | 426 |

After applying my Gradient Boosting model to the test dataset, it achieved an accuracy of approximately 66.48%. This data indicates that the model has higher precision for class 0 (0.75) compared to class 1 (0.57), suggesting it is more reliable in predicting non-events. However, the recall rates show that the model is somewhat better at identifying true positives for class 1 (0.65) than for class 0 (0.67), highlighting its capability to capture a significant portion of the positive class instances.

The top 3 important features are to go, Down and Yard line 0-100 followed by time pressure and momentum.

**SIGNIFICANT VARIABLES:**

The feature importance graph for the Gradient Boosting model clearly indicates that the variables "TO GO", "DOWN", and "YARD LINE 0-100" are the most significant in predicting whether a play will be a run or a pass. "TO GO" likely measures the distance needed for a first down, which crucially influences play-calling as teams often prefer passing in longer yardage situations to quickly gain large chunks of yards. "DOWN" is also highly influential because the down number (e.g., 1st down, 2nd down, etc.) typically dictates the urgency and type of play; for instance, teams might be more likely to pass on later downs when needing to recover from lost yardage. The "YARD LINE 0-100" reflects field position, which is pivotal as strategic choices vary significantly depending on how close a team is to scoring zones.

**MODEL EXPECTATIONS AND IMPROVEMENTS**

The model's predictions show a reasonable alignment with these expectations. For example, the model effectively uses these key variables to distinguish between run and pass plays based on situational context dictated by game rules and common strategies. However, the model could still benefit from improvement in reducing false positives and false negatives, as indicated by the confusion matrices. Specifically, enhancing its sensitivity (recall) for pass plays could prevent overlooking potential passing scenarios, and reducing false positives in run predictions would avoid incorrect anticipations of run plays.

**GAME STRATEGY**

These analytical insights and model results could be extremely valuable for enhancing game strategy or decision-making in football coaching and analysis. By understanding which factors most significantly influence play-calling decisions, coaches could develop more nuanced offensive strategies that better exploit defensive weaknesses. For instance, knowing that "TO GO" is a critical predictor, offensive coordinators might design plays that create manageable second and third downs to keep the defense guessing. Additionally, incorporating predictive modeling like this into real-time decision-making tools could provide coaches with data-driven recommendations during games, potentially increasing the effectiveness of play-calling under various game conditions.

**CONCLUSION:**

In conclusion, the analysis of feature importance and model performance in predicting play types underscores crucial insights for football strategy, particularly identifying key variables like distance to go, down number, and field position as significant predictors. While the model performs adequately, enhancements in sensitivity to pass plays and reductions in false positives for runs could further refine its effectiveness. These insights offer valuable opportunities for coaches to leverage data-driven tools to optimize play-calling, enhancing tactical decisions during games. By incorporating predictive modeling into real-time decision-making, teams can dynamically adapt to game situations, potentially transforming traditional approaches and providing a competitive edge in football strategy.

APPENDIX:

CHART 1 – PLAY TYPE DISTRIBUTION:

**A graph of a number of different colored squares

Description automatically generated**

CHART 2: RANDOM FOREST CONFUSION MATRIX

**A chart of a blue yellow and purple color

Description automatically generated with medium confidence**

CHART 3: RANDOM FOREST FEATURE IMPORTANCE  
A graph of blue bars with white text

Description automatically generated

CHART 4: GRADIENT BOOSTING CONFUSION MATRIX:

A chart of a color chart

Description automatically generated with medium confidence

CHART 5: GRADIENT BOOSTING FEATURE IMPORTANCE  
A graph of blue rectangular bars with white text

Description automatically generated