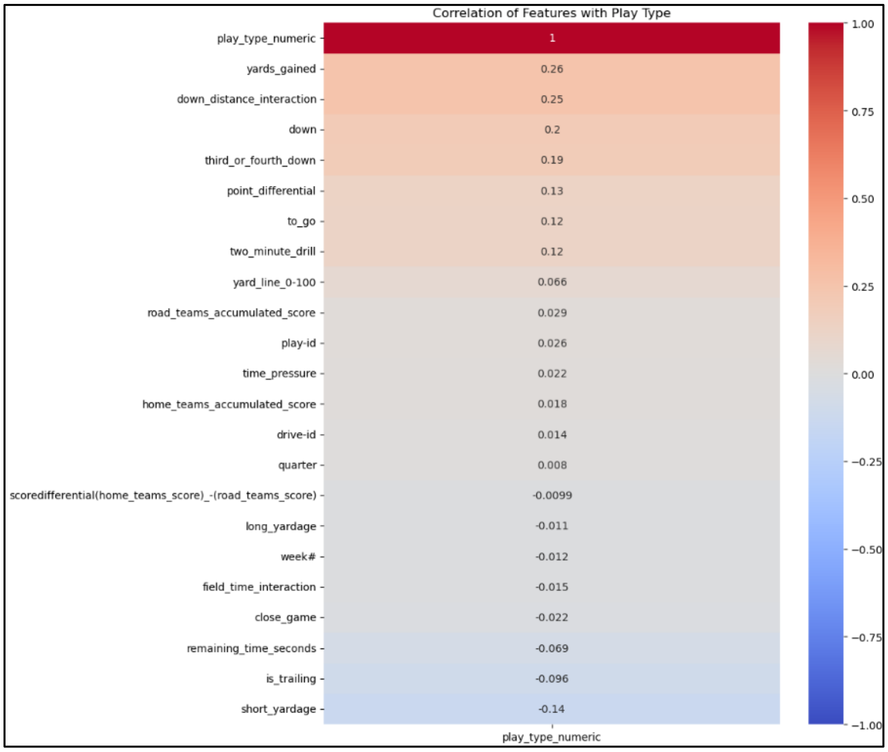
Predicting NFL Optimal Play Type

INTRODUCTION

This initiative offers a great chance to improve the American football players ability to make smart decisions. Predicting whether a play will be a run or a pass is the goal of the study, which can provide teams a significant competitive edge. This ability to foresee has broad ramifications for many football business parties. The football industry could benefit from this predictive model for play-calling in a number of areas, including team performance, player development, fan interaction, and business operations. All parties involved may enjoy a more strategic, effective, and entertaining football experience if it is well implemented.

DATA CLEANING AND PREPARATION

For the football analytics team, the dataset's null values offer both opportunities and obstacles. 95.97% of the data in the 'points_scoredby_either_team' column is missing, which is the biggest problem. A useful predictor for play type prediction may be lost due to the high proportion of missing data, which effectively makes the column useless for analysis and model building. We chose to remove this column completely as a result, which would affect the model's capacity to account for the influence of score on play-calling choices. The other variable is the 'yards_gained' column, which has 23.60% missing values. Although there is still a significant quantity of missing data, it is not severe enough to justify removing the column completely. We decided to plot the distribution of this variable where we found that it is slightly skewed. So we replaced the null values with median yards gained for each play type by using the group by central tendency method. This method minimizes the effect of the missing data on the analysis as a whole while maintaining the column's potential predictive power. Finally, as the 'down' column had 0.36% missing values in it we decided delete the records having missing values. With a very low percentage, removing the records will not affect our data. Thus after the thorough check the data is good to go for next process.



The Interquartile Range (IQR) approach was used to identify and manage outliers, especially in the 'home_teams_accumulated_score' and 'road_teams_accumulated_score' columns, after time-based data were converted to seconds. In order to capture intricate game dynamics, this football play prediction project's feature engineering process required the creation of multiple additional variables. Important features 'time_pressure' and 'two_minute_drill' to represent important time periods, and 'point_differential' to capture the score situation. Other binary characteristics were developed to represent particular game conditions, such as 'close_game', 'is_trailing', 'long_yardage','short_yardage', and 'third_or_fourth_down'. Further, to capture more complex relationships, interaction concepts like 'down_distance_interaction' and 'field_time_interaction' were introduced. StandardScaler was used to normalize numerical features, while one-hot encoding was used for categorical data. The Variance Inflation Factor (VIF) was used to evaluate multicollinearity and normalize numerical features. The 'play_type' target variable was transformed into a binary numeric variable. The integrity and applicability of the dataset for forecasting American football play types were guaranteed by this thorough data cleaning and preparation procedure, which also provided a strong basis for further research and model development.

DATA MODELLING AND EVALUATION

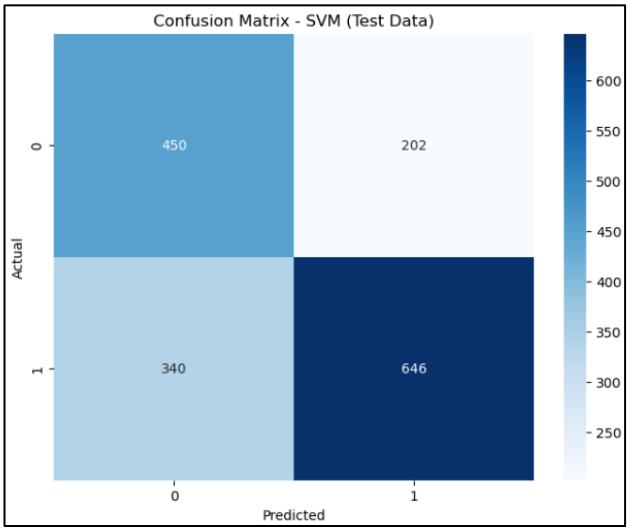
Using the supplied training dataset, several machine learning methods were implemented and evaluated as part of the modeling process for predicting American football play types. 'down', 'to_go', 'yard_line_0-100', 'point_differential', 'two_minute_drill', 'time_pressure', 'long_yardage', 'short_yardage', 'is_trailing', 'remaining_time_seconds' and 'field_time_interaction' were among the fifteen carefully chosen variables within the feature set. To guarantee reliable model evaluation, the dataset was divided into training (80%) and validation (20%) sets. K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest and X Gradient Boosting (XGBoost) were the six models that were used. The scaled training data was used to train each model, and the validation set was used to assess each model. To evaluate each model's predictive power, performance metrics such as accuracy, precision, recall, and F1-score were computed. To further visualize the models' performance in terms of true positives, true negatives, false positives, and false negatives, confusion matrices were created. This all-encompassing strategy made it possible to compare many algorithms in detail and ascertain which one performed best in predicting American football play types.

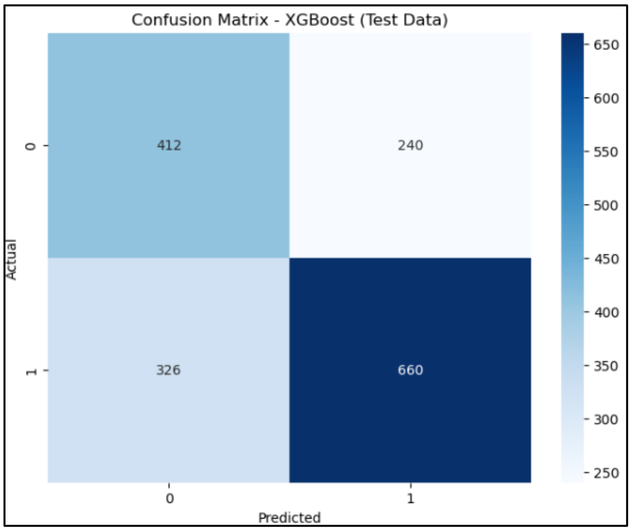
After the initial Training of models on the training data, the models that performed better than others were, SVM Model and XGBoost. Evaluations of both the model's ability to forecast different American football play types revealed unique advantages and disadvantages for both of them. With precision scores of 72.98% and 76.18%, respectively, the SVM model showed an accuracy of 66.68% on the training data and 66.91% on the test data. On the test data, its recall was marginally lower at 65.52%, yielding an F1-score of 70.45%. As seen by its recall score, this suggests that even while SVM is accurate, it might overlook some favorable examples.

The XGBoost model, on the other hand, plummeted to 65.45% on the test data after achieving a little better accuracy of 66.95% on the training data. It achieved an F1-score of 71.65% by maintaining a balanced precision and recall on the training set (71.71% and 71.59%, respectively). With precision at 73.33% and recall at 66.94%, it performed worse on the test set, yielding an F1-score of 69.99%. The decline in measures from training to test data points to possible overfitting.

Because of its exceptional generalization skills and reliable performance on both datasets, the SVM model is selected as the optimal model for this assignment. Although SVM's recall is marginally lower than XGBoost's, its higher precision and consistent F1-scores indicate that it is less likely to overfit and more dependable for real-world applications where reducing erroneous predictions is crucial.

COMPARING BOTH MODELS





Metrix/Model Name	SVM (Train)	SVM (Test)	XGBoost(Train)	XGBoost(Test)
Accuracy	0.6668	0.6691	0.6864	0.6545
Precision	0.7298	0.7618	0.6906	0.7333
Recall	0.6812	0.6552	0.6864	0.6694
F1 Score	0.7046	0.7045	0.6878	0.6999

Several important technical factors show that the Support Vector Machine (SVM) model performs better than other models and is the best option for forecasting the different kinds of American football plays. First off, the SVM model consistently maintains F1-scores of 70.45% and 70.46% on the test and training sets, respectively, demonstrating its exceptional generalization skills. This constancy suggests strong performance across several datasets, which is an essential quality for practical uses. Second, the SVM model outperforms XGBoost in terms of precision on the test set (76.18%), which is essential for reducing false positives in play type predictions. During games, this improved accuracy is especially helpful for making strategic decisions. Furthermore, the SVM model exhibits exceptional stability, indicating dependable performance in a variety of settings with little performance decline between training and test data. Additionally, the SVM model performs consistently across datasets, demonstrating superior resistance to overfitting for this specific task in contrast to XGBoost, which exhibits overfitting. Together, these technological advantages make the SVM model a more reliable and appropriate option for forecasting different kinds of American football plays, providing insightful information for in-game strategy and decision-making.

MODEL DISCUSSION AND FINAL THOUGHTS

The effectiveness of the Support Vector Machine (SVM) model in forecasting American football run/pass plays provides important information about the variables affecting play-calling choices. The three factors that have the most effects on the model's predictive power are down (0.195), down-distance interaction (0.252), and yards gained (0.256). According to this, teams are more inclined to pass when there is more yards and a later down, which is consistent with standard football strategy. The two-minute drill (0.119) and point difference (0.126), two aspects of the game context, also have a moderate impact on play-calling decisions. It's interesting to note that teams tend to run more in short yardage situations (-0.137) and trailing game status (-0.097), which have a negative association.

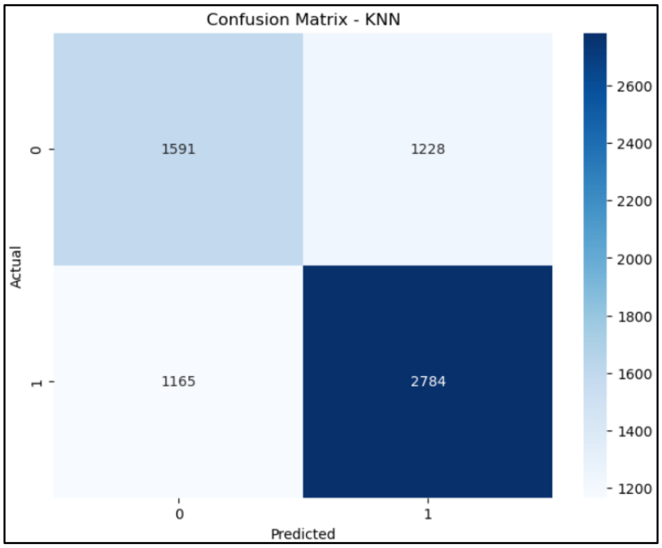
With 67.80% accuracy on the test set, the SVM model's performance shows a reasonable level of predictive power, but it also identifies areas that need work. The model might have trouble with more complex circumstances, but it seems to be very good at spotting obvious passing situations (like long yards on third down). The model is a dependable tool for defensive game preparation because of its balanced precision (69.02%) and recall (67.80%), which indicate that it is equally good at anticipating run and pass plays.

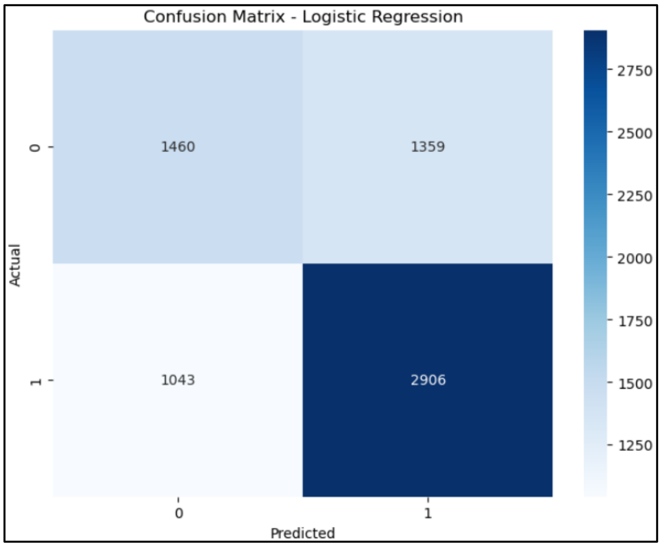
In real-world applications, these insights could greatly improve game strategy. Based on game circumstances, defensive coordinators could utilize this model to more precisely predict offensive play-calling trends. For example, defensive formations and player choices may be influenced by the understanding that teams are more likely to pass on later downs with more yardage. The model's 67.80% accuracy rate, however, also suggests that offensive teams continue to gain from a certain amount of play-calling uncertainty. Teams could use these insights to gain a competitive edge by either deviating from or conforming to projected trends.

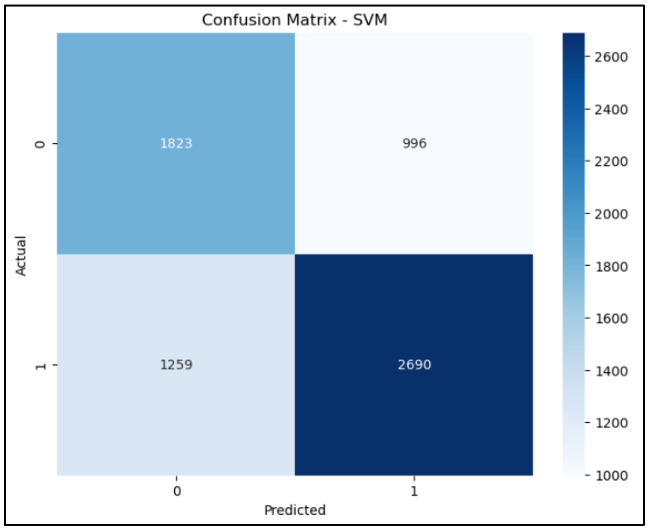
Additional contextual factors, including particular player matchups, weather, or more in-depth history trends of particular teams and coaches, could be added to the model to make it even better. As the model adjusts to changing team tactics and game conditions, real-time model updates during gameplay may also improve the model's prediction ability.

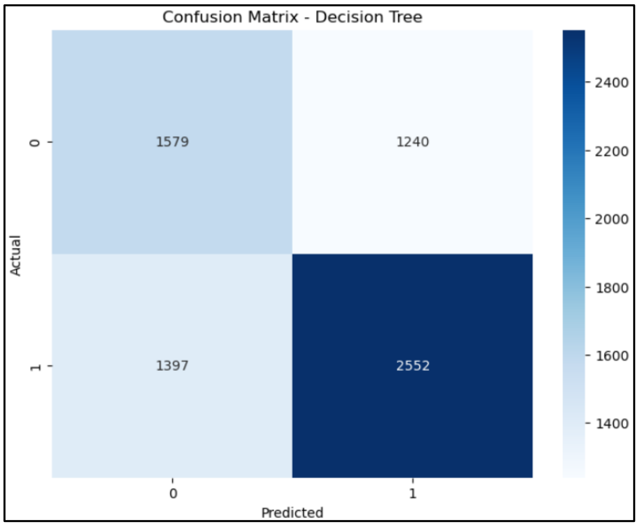
In conclusion, even though the SVM model offers insightful information for football strategic decision-making, it should only be utilized as one tool in a holistic game planning approach. Although its forecasts provide a data-driven basis for strategic choices, coaches' and players' experience is still essential for deciphering and implementing these insights in the fast-paced environment of a football game.

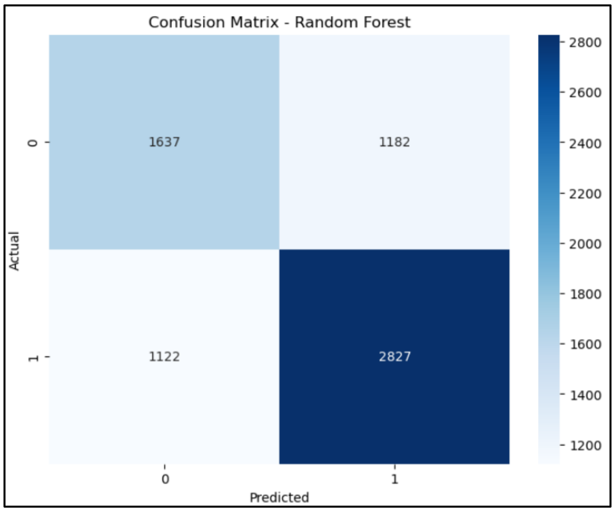
APPENDIX

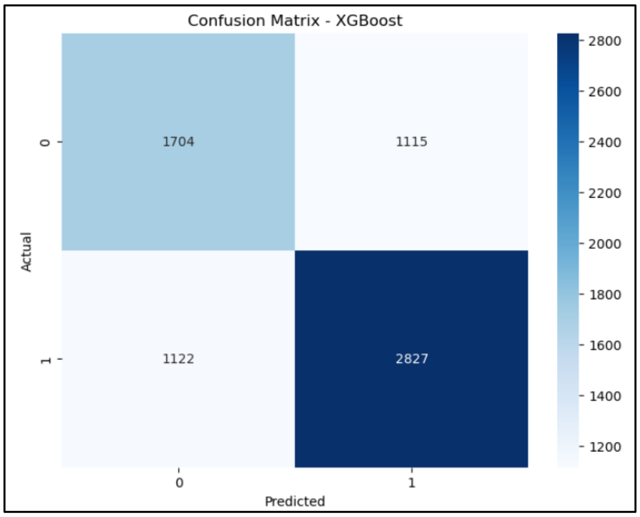

Confusion Matrix - KNN
1591
1228
1165
2784
Actual
0
1
Predicted
0
1
2600
2400
2200
2000
1800
1600
1400
1200


Confusion Matrix - Logistic Regression
1460
1359
1043
2906
Actual
0
1
Predicted
0
1
2750
2500
2250
2000
1750
1500
1250


Confusion Matrix - SVM
1823
996
1259
2690
Actual
0
1
Predicted
0
1
2600
2400
2200
2000
1800
1600
1400
1200
1000


Confusion Matrix - Decision Tree
1579
1240
1397
2552
Actual
0
1
Predicted
0
1
2400
2200
2000
1800
1600
1400


Confusion Matrix - Random Forest
1637
1182
1122
2827
Actual
0
1
Predicted
0
1
2800
2600
2400
2200
2000
1800
1600
1400
1200


Confusion Matrix - XGBoost
1704
1115
1122
2827
Actual
0
1
Predicted
0
1
2800
2600
2400
2200
2000
1800
1600
1400
1200