Outcome Prediction

Pradnya Tolnur
Football Analytics Intern
(Offensive Analytics)
UNC Charlotte Football Team

UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

Objectives

The main objective of this project is to assist the coaches and football staff to better devise the improvement strategies for the team by analyzing the available data from the football matches of season 2023. Below are the approaches taken to achieve this objective.

- Build Machine Learning Models to predict the outcome of a play based on previous Five Plays in the game.
- Performance Analysis.
- Strategic Insights.

Data Sources

- Catapult
- PFF Ultimate

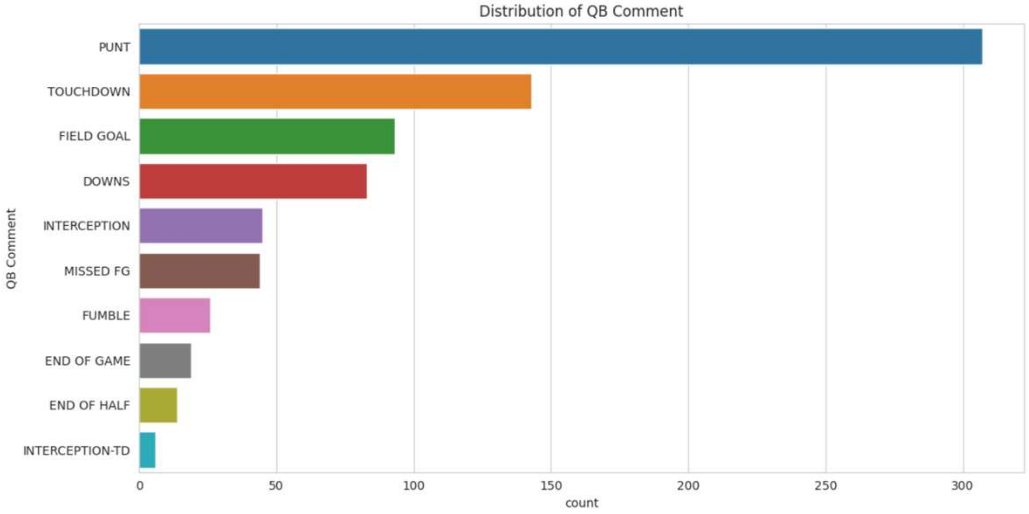
Feature Selection

Based on the research, game understanding and discussions with the subject matter experts from the University football team (Coaches and analytics team), the features most relevant in predicting whether the outcome of a play would be a touchdown, punt or an interception are:

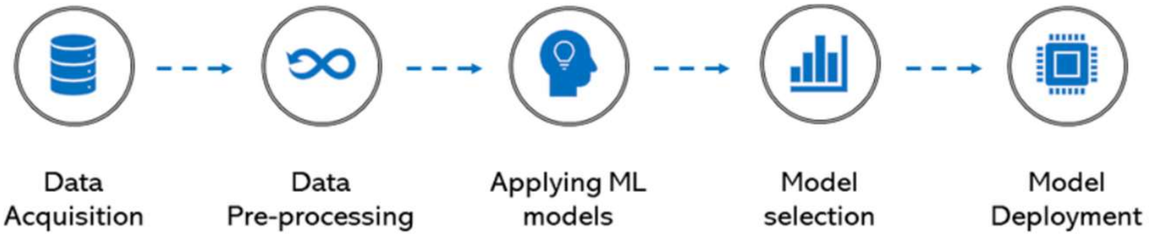
Down, Distance, Field Position, Formation, Gain, Run Concept, Run/Pass, Pass Result, The Play.

Distribution of Target Variable in the Dataset

From the count graph of the Target distribution, it is visible that the most number the outcomes are PUNT, TOUCHDOWN, FIELD GOAL and DOWNS.



Process Overview



Data Acquisition → Data Pre-processing → Applying ML models → Model selection → Model Deployment

- Preprocessed the integrated data from both the sources, i.e., Catapult and PFF Ultimate.
- Encoding was done for the Categorical feature and Lag features were created to account for the dependency of the prediction based on previous five plays' outcome.
- Given the multivariate nature of the target variable and time-sequential prediction task, the models chosen were:

1) Gradient Boosting Machine

2) Long Short-Term Memory (LSTM) Model.

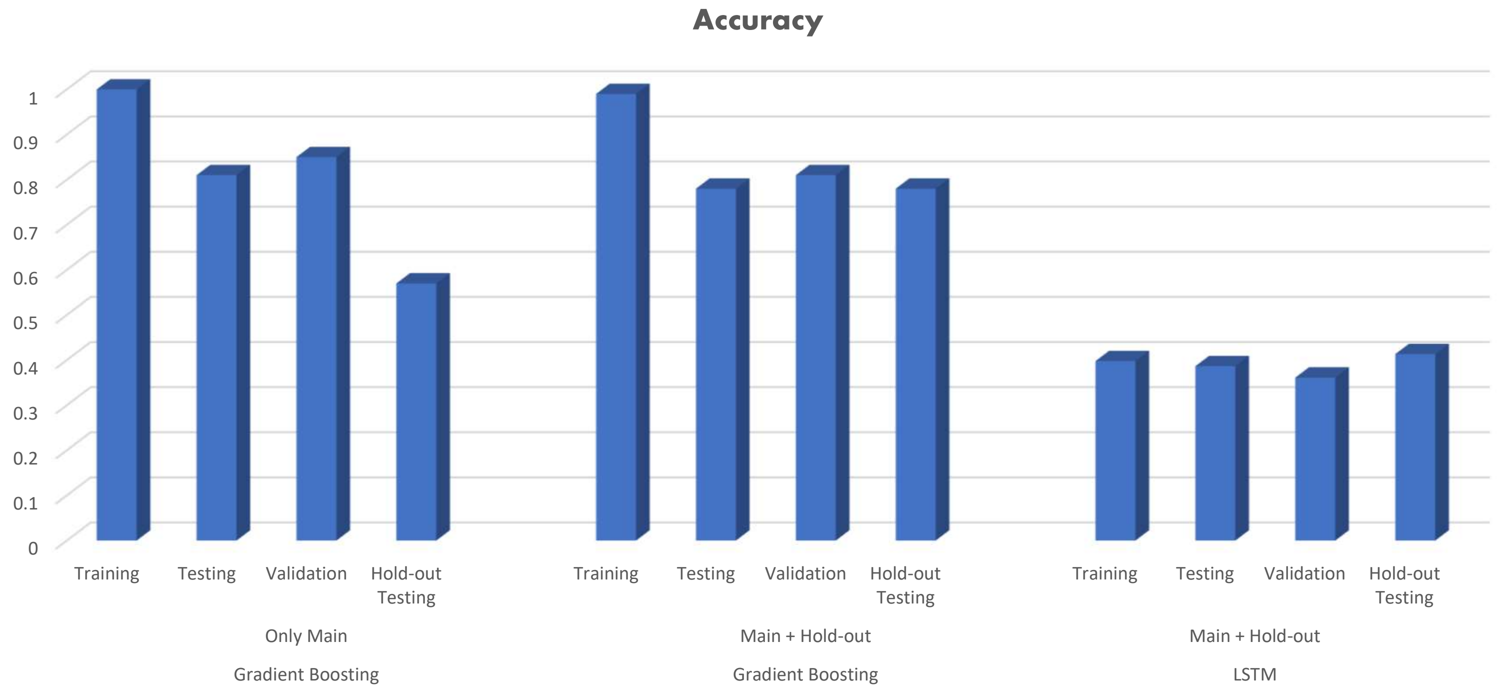
Model Performance Analysis

- Models were built using a 'Stratified Splitting'. Stratified data splitting was important to ensure that the split datasets are equal representatives of the entire dataset and that the model wouldn't be biased towards any specific category of the target variable

- Furthermore, the model performance was also tested on a hold-out dataset to observe the performance on completely unseen data.

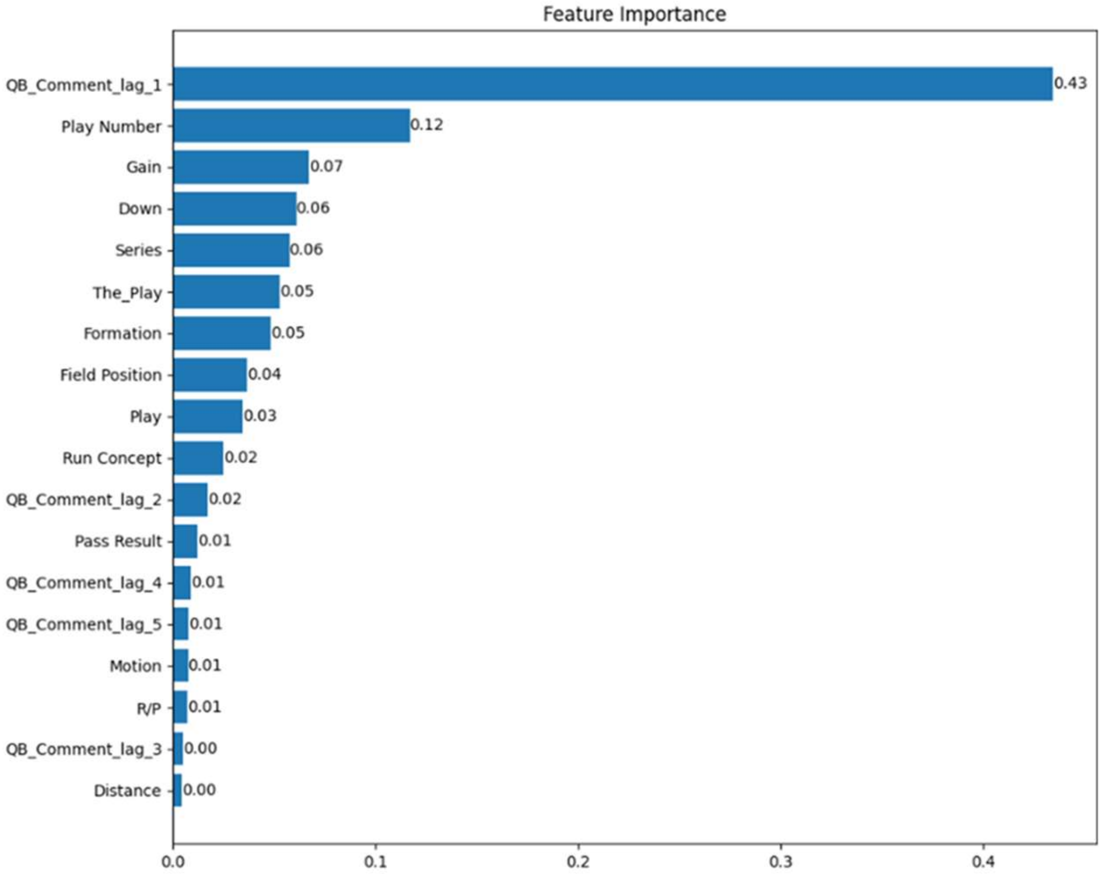
Model Performance Analysis

Here are the model performance comparisons based on the dataset that splitting method was applied to:



Feature Importance

The feature importance plot shows that the features immediate previous play, play number, gain, down, formation, field position are some of the predictors contributing most to the model predictions.



Observations

- Based on the model comparison, it is evident that the ‘Gradient Boosting Machine’ model performs better with higher accuracy than the LSTM model.

- Upon further, deeper analysis of the GB model performance, it is important to note that the hold-out data being created with the stratified splitting approach vs it being taken out before the splitting makes a noticeable difference in the accuracy scores on testing, validation and hold-out datasets.

- To improve the model performances, more and more data points are needed to train the model on.

Team Performance Analysis

To gauge the offensive effectiveness of a team, analysis like below can prove to be helpful to the coaches:

- Average gain = 4.9 yards gained per play,
- Pass Completion Rate = 76.28%
  This is the percentage of completed passes which effectively gives the proportion of pass attempts that were successful.
- Quarterback Performance Analysis: For QB#11, the frequency of the most probable outcome = 'PUNT' is 41%.
  Grouping the data by 'QB#' and counting how frequently different possible game outcomes like TOUCHDOWN, PUNT, FIELD GOAL, etc. occur for each quarterback.
- Formation Personnel Analysis: How different formations and personnel groupings affect the gain in yards per play.
  Maximum yards of gain was observed against the formations 'BONE' and 'DAFFY' whereas the players 11 and 12 scored the longest yards this season.
- Such performance metrics can be customized based on the needs of the coaching staff through user interface application integrated with Streamlit.

Strategic Insights

- In extension to the team performance analysis, strategic insights based on different features influencing the game outcome can be immensely helpful as well. To demonstrate a few:
- Given the 'DOWN' (1st, 2nd, 3rd) and 'DISTANCE', which play strategy was chosen most of the times –

Down	Distance	Play	Percentage
0	10	LION OLY	14.28%

- Given the 'FORMATION', 'PERSONNEL' and 'Run/Pass', which play strategy was chosen most of the times –

Formation	Personnel	R/P	Play	Percentage
BONE	21	R	BEETLE H-TRAIL	100

- Such performance metrics can be customized based on the needs of the coaching staff through user interface application integrated with Streamlit

THANK YOU !