NFL Big Data Bowl 2025 Project Report

**Executive Summary**

Our report analyzes player tracking data from the NFL Big Data Bowl 2025, focusing on third-down plays to understand team and player tendencies. Using advanced data science techniques, including machine learning and PySpark, we predict whether a play will result in a run or pass. Key findings show that teams tend to pass more often on third down compared to first and second downs. We highlight specific team tendencies with detailed visualizations, aiding strategic decision-making for coaches and analysts.

In partnership with Kaggle, the NFL Big Data Bowl 2025 encourages innovation in football analytics by using data science to generate practical solutions. Our team competes in the Metric Track, leveraging pre-snap data to enhance game planning and in-game decision-making during the 2024 NFL season's second half.

The competition evaluates submissions based on four criteria: Football Score, Data Science Score, Report Score, and Data Visualization Score. These assess applicability to the NFL, football data complexity, uniqueness of ideas, data analysis techniques' accuracy and innovation, report clarity, and data visualization effectiveness.

Our team will deliver a comprehensive project report, Python code, and presentation addressing these scoring components, contributing to the advancement of football analytics in the NFL.

**Project Overview**

The NFL Big Data Bowl has fostered innovation in football analytics since 2019. This year's theme focuses on understanding player and team behavior during the critical 40-second window before a play is executed. Participants analyze player tracking data captured by Radio Frequency Identification (RFID) technology. Our project, part of the metric track, explores third-down plays in the first nine weeks of the 2024 NFL season to discern patterns crucial for strategic planning.

Third-down plays significantly impact a game's outcome, with play-calling decisions influenced by factors like yards-to-go and team preferences. We will analyze data from all 32 NFL teams, examining the relationship between play-calling decisions and variables such as yards-to-go, game situations, and team-specific tendencies. Our goal is to extract actionable insights to inform coaching staff and analysts in optimizing game plans and in-game decision-making during the second half of the 2024 NFL season.

By focusing on this key aspect of the game and employing advanced data analytics techniques, we aim to contribute valuable insights to strategic play-calling discussions in the NFL.

**Data Collection and Preparation**

Our primary goal was to analyze third-down play tendencies in the NFL using machine learning techniques. To achieve this, we utilized data from the plays.csv dataset, which encompasses detailed information on each play from the first nine weeks of the 2024 NFL season. Our focus was specifically on third-down plays, as they are critical situations that can significantly impact a game's outcome.

Feature Selection:

From the 50 columns provided in the dataset, we identified four key features relevant to predicting run vs. pass plays on third down using pre-snap information. These four features were chosen because they capture important pre-snap information that can be used to predict the team’s decision to run or pass the ball on third down. The data types of these features (integer and categorical) were also suitable for the machine learning models used in the analysis.

|  |  |  |
| --- | --- | --- |
| Attributes | Description | Data Type |
| absoluteYardlineNumber | The distance from the end zone for the possession team. | Integer |
| offenseFormation | The formation used by the possession team (e.g., Shotgun, I-Formation). | Categorical (String indexing and one-hot encoding) |
| receiverAlignment | The alignment of receivers, such as 0x0, 1x0, 1x1, 2x0, 2x1, 2x2, 3x0, 3x1, 3x2. | Categorical (String indexing and one-hot encoding) |
| yardsToGo | The number of yards needed to achieve a first down. | Integer |
| isRun | A new target column indicating if the play was a run (1) or pass (0). | Binary (0 or 1) |

**Data Preprocessing**

To facilitate our analysis, we created a new binary column called "isRun." This new target was crucial, as it provided a clear binary label (0 for pass, 1 for run) for the machine leaning models to learn from.

Analyzing Third-Down Tendencies:

Our objective was to determine if machine learning could help predict run vs. pass plays on third down, both for the entire NFL as well as specific some specific teams, such as the Philadelphia Eagles (PHI). Due to the limited number of third-down plays available for a single team, we opted to use a training set that includes third-down plays for all teams rather than just the upcoming opponent. The Philadelphia Eagles had a total of 102 third-down plays (41 runs and 61 passes), while league-wide, there were 3270 third-down plays (655 runs to 2624 passes).

We used PySpark, a powerful tool for efficient data analysis and processing, to work with this large dataset. We began by organizing the data and calculating the overall percentages of run and pass plays on third down across all teams:

* Run plays: 19.98%
* Pass plays: 80.02%

Next, we examined individual team strategies on third down and found:

* Philadelphia Eagles: Highest run rate (~40%)
* Pittsburgh Steelers: Lowest run rate (7.8%)
* Most teams: Run rate between 30% and 15%

A graph of a number of people

Description automatically generated with medium confidence

These insights into each team's strategies and preferences during third-down situations can be valuable for coaches and analysts looking to improve their game plans and make better decisions during the second half of the 2024 NFL season.

**Methodology**

To address the imbalance distribution of run and pass plays in the dataset, we employed a technique called class weights. Class weights help ensure that our analysis accounts for the relative importance of each type of play, especially when one type occurs much less frequently than the other.

By using class weights, we can maintain the accuracy of our analysis and reduce the possibility of over-adjusting our model to the more common play type. In this case, we calculated the following class weights:

* **Label 1 (Run): Weight = 2.5031**
* **Label 0 (Pass): Weight = 0.6248**

Since run plays are less frequent in the dataset, they are assigned a higher weight. This allows our model to give more importance to correctly predicting run plays, ultimately leading to a more balanced and accurate analysis.

We ensured no missing values in selected feature columns and then prepared the data considering teams' tendencies in 'third and long' situations. We removed plays representing 'third and 10' or longer based on the assumption that teams are less likely to run the ball in such situations, eliminating the upper bound (80th percentile) from the dataset:

* **Lower Bound (10th Percentile): 1**
* **Upper Bound (80th Percentile): 10**

After filtering, we had 2684 third-down plays remaining. To ensure unbiased data splitting, we shuffled the dataset and applied the **randomSplit** method to create an 80% training set and a 20% testing set.

Feature Selection and Model Tuning:

Given the limited number of features, we decided to retain all features for modeling is due to their importance and impact on model performance. Previous trials indicated that removing features like **absoluteYardlineNumber** negatively impacted model performance. We performed a grid search which helps find the best parameter combination for the Random Forest model. Various combinations of 'Max Depth' and 'Num trees' were tested to find the best configuration.

**Feature Importance Across Models**

A black and white text on a white background

Description automatically generated A black and white text with black text

Description automatically generated

A close-up of a number

Description automatically generated A white and black text on a white background

Description automatically generated

**Grid Search Performance Evaluation**

A screenshot of a computer program

Description automatically generated

\*Due to the size of the code snippet, a screenshot of the grid search is provided below for better readability.

For example, the best model parameters for the Random Forest Classifier are:

* **RandomForestClassifier\_a672aaad4ae4\_\_numTrees: 50**
* **RandomForestClassifier\_a672aaad4ae4\_\_maxDepth: 20**

Analytical Framework:

Our framework combines statistical analysis and machine learning techniques using PySpark to handle the large dataset efficiently. The process involves:

1. **Feature Selection:** The features were chosen because they directly impact the decision-making process of teams on third down. These factors capture critical pre-snap information that can help predict whether a team will choose to run or pass the ball, providing valuable insights for coaches, analyst, and fans. Identifying pertinent features relevant to predicting play outcomes:

* **absoluteYardlineNumber**
  + The field position can influence the team’s risk tolerance and the types of plays they are more likely to call.
* **offenseFormation**
  + Different formations are associated with specific play types (run or pass), and this information can help predict the team’s decision-making on third down.
* **receiverAlignment**
  + The alignment can impact defensive coverage and offensive play execution.
* **yardsToGo**
  + It directly influences the urgency and type of play a team might choose, affecting overall play outcomes.

1. **Machine Learning Models:** Implementing various models (Decision Trees, Random Forests, Logistic Regression, Gradient Boosted Trees and MLP) help analyze run vs. pass plays more accurately, particularly for the Philadelphia Eagles.
2. **Encoding Categorical Variables:** One-Hot Encoding is a technique that transforms text-based information into a numerical format that machine learning models can understand and work with.
3. **Model Training and Evaluation:** We split the dataset, with 80% used to train the models and 20% to test their accuracy. Class weights are applied to balance the data, ensuring fair evaluation.
4. **Ensemble Prediction:** We create a more accurate prediction model by combining multiple models and considering their predictions, using majority voting to determine the final outcome.

**Results**

Model Evaluation:

We trained and evaluated multiple machine learning models, including Decision Trees, Random Forests, Logistic Regression, and Gradient Boosted Trees. Each model had a classification report generated. The classification report provided a detailed assessment of how well each machine learning model performed in predicting whether a play would be a run or a pass. This report included important metrics like accuracy, precision, and recall, which helped the researcher understand the strengths and weaknesses of each model.

By comparing performance of different models, they could identify the most reliable and accurate model to use for predicting play outcomes. This approach ensured that the final model would be robust and able to make accurate predictions, even on new, unseen data. Evaluating multiple models also helped the researchers gain a deeper understanding of the key factors that influence play decisions, such as field position, formation, and the distance can be valuable for coaches and analysts who want to make informed decisions during the game.

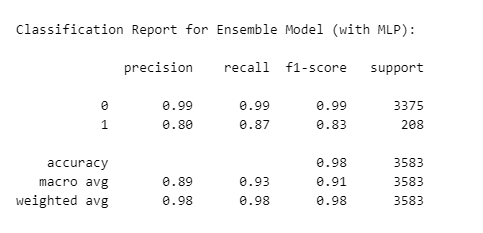
Performance Summary:

* All models showed strong performance in predicting pass plays (Class 0), with high precision, recall, and F1-scores.
* Predicting run plays (Class 1) was more challenging, characterized by lower precision and recall due to class imbalance and the complexity of identifying run plays.
  + This class imbalance made it more difficult for machine learning models to accurately predict when a run play would occur, as there were fewer examples for the models to learn from, and predicting run plays is inherently more complex than predicting pass plays.
* We selected the ensemble of models as a baseline, as the combined predictions were able to perform better than all of the models individually.

Model Performance Metrics:

* **Best F1 Score for Run Category:** **0.83.**
* **Challenges:** Even after tuning the model, predicting run plays remained difficult, reflecting the dataset's class imbalance (teams pass 80% of the time on third downs). An example of this can be found by attempting to predict runs vs. passes for individual teams.

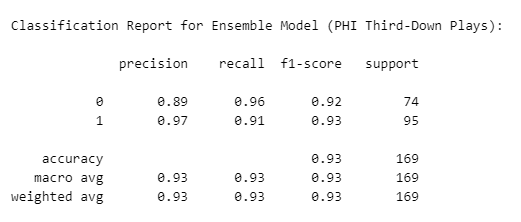
Improved Model Performance Through Ensembling



Philadelphia Eagles Analysis:

The ensemble model was able to predict runs vs. Passes well for the Piladelphia Eagles (PHI), yielding a decent F1 score of 0.93.

Philadelphia Eagles Third Down Prediction Performance

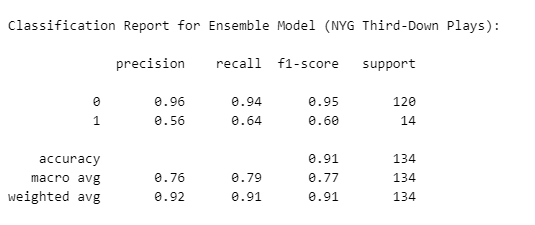


A screenshot of a report

Description automatically generated

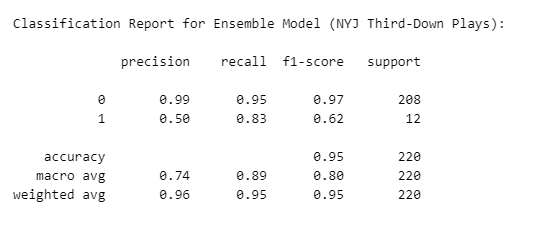
New York Giants Analysis:

The ensemble model struggled however to predict runs for the New York Giants, with an F1 score of .60 for runs.



New York Jets Analysis:

The ensemble model also struggled when predicting runs vs. Passes on third down for the New York Jets. Because of the model’s inconsistent performance when predicting runs for different teams, it may not be best to move forward to using the ensemble model in production.



Rush and Pass Location Analysis:

An attempt was also made to predict the location of runs and passes on third down.

* The analysis focused on **rushLocationType** and **passLocationType** features to explore rushing and passing play locations.
* Most third-down plays were pass plays, leading to many 'rushLocationType' values labeled as 'NA.'
* To refine insights, 'NA' and 'UNKNOWN' categories were filtered out.
  + Filtering out these non-informative categories allowed us to focus on the actual running plays and gain more valuable insights into where teams tend to run the ball on third down.

Importance of Traditional Statistics:

Analysis showed that 85% of passes were located "INSIDE\_BOX," suggesting limited value in using machine learning for predicting pass locations. For this reason, our focus was shifted to predicting rush locations.

A close-up of a number

Description automatically generated

*\*Pass Locations Heavily Concentrated in the “INISIDE\_BOX” Area*

Class Weights for Rush Location Prediction:

To address class imbalance in the **rushLocationType** feature, class weights were computed and subsequently recalculated after filtering the dataset to include only third downs with 3 or fewer yards to go. However, ongoing class imbalance hindered accurate predictions for categories 2-4. Oversampling techniques were tested but did not improve results. Class weights give more importance to the less common rush location categories during the training process. This helps the models better learn the patterns for these minority classes, even though there are fewer examples available. This technique aims to improve the model’s ability to accurately predict the less common rush locations.

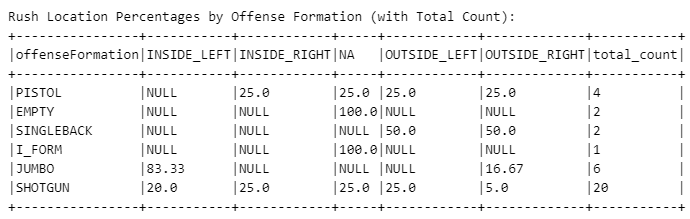
League Insights:

* Teams used run plays approximately 20% of the time on third downs, with the Philadelphia Eagles showing the highest run rate (around 40%) and the Pittsburgh Steelers the lowest (7.8%).
* The likelihood of a pass play significantly increased when four or more yards were needed for a first down.

**Discussion**

We found that using traditional statistics and data analysis provides better insight when compared to the use of machine learning. For example, our analysis of the Philadelphia Eagles' (PHI) running habits by formation revealed valuable insights:

* **Shotgun Formation:** PHI predominately uses the Shotgun formation on third down with three or less yards to go. PHI runs the ball 75% in this scenario, which could be considered surprising as the Shotgun formation is traditionally known as a passing formation.
* **Jumbo Formation:** This is the second most frequent formation for PHI on third down and three or less yards to go. We found that PHI typically runs to the Inside left when using this formation.



These findings could potentially offer useful team strategies during pivotal plays. Philidelphia’s clear preference to run the ball from a Shotgun formation on third down with 3 or less yards to go does not align with conventional football wisdom. This could offer coaches critical data for game preparation. Leveraging traditional data science and statistics techniques could assist coaches with game strategy and allow coaches to adjust their strategies dynamically for each opponent.

**Conclusion**

This project effectively utilized Pyspark and data analytics techniques to analyze third-down play tendencies in the NFL. Our findings provide valuable insights for teams, coaches, and analysts seeking to refine game strategies and improve decision-making processes. By employing PySpark, we efficiently managed large datasets, which proved crucial in deriving meaningful conclusions. Further exploration of player tracking data holds promise for continued innovation in football analytics, ultimately leading to a deeper understanding of the game and the development of more sophisticated strategies.

While machine learning has demonstrated its power in numerous use cases, our analysis highlights the continued importance of traditional data science and statistics. In the context of third-down play tendencies, well-established football knowledge and basic analytics still play a significant role in gaining insights.

As Large Language Models (LLMs), like ChatGPT, gain prominence in various industries, it remains vital to recognize the importance of traditional data analytics tools such as NumPy, Pandas and Scikit-learn (sklearn). Combining these conventional methods with emerging technologies will prove most beneficial for individuals and companies leveraging data analytics in their decision-making processes.