2020 NFL Big Data Bowl: Utilizing Pyspark to Develop Models Evaluating the Continuous Ranked Probability Score (CRPS) for Each Play

Anoop Nath, Lauren O'Donnell, OC Ofoma
DS 5110

Abstract:

The National Football League (NFL)'s Next Gen Stats technology collects comprehensive player data using Radio Frequency Identification (RFID) chips embedded in players' shoulder pads. This technology led to the collection of intricate data, encompassing every play of the NFL's pre-, regular, and postseason games.

This technological leap ushered in a new era of data-driven analysis in American football. The utilization of advanced metrics enabled the application of analytics for predicting outcomes like converting fourth downs. The NFL introduced the Big Data Bowl in 2019, encouraging data enthusiasts to explore play-by-play data for trends and solutions to data-related questions. Hosted on Kaggle, the competition promotes data analytics and machine learning applications in NFL decision-making processes.

This study builds upon the NFL Big Data Bowl's premise and redefines the problem as a binary classification task: predicting whether a first down will be achieved in a play. Three models—logistic regression, gradient boosting, and random forest—are employed to solve this problem. Leveraging a dataset encompassing various play attributes and environmental factors, the models are evaluated based on accuracy, precision, recall, F1 score, and Area Under the Curve (AUC) metrics.

Upon analysis, gradient boosting emerges as the most effective model, yielding the highest performance metrics in both data splitting and k-fold cross-validation experiments. Although other models do not significantly lag behind, the superiority of gradient boosting signifies its aptitude for predicting first down success using features such as `DefendersInTheBox`, `Down`, `Distance`, `DistanceToTouchdown`, and `ClosestDefenderDistance`. The study recommends future investigations into expanded datasets, incorporating environmental factors, and exploring advanced techniques like deep learning to further refine prediction accuracy in the context of NFL plays.

Introduction:

The National Football League (NFL) launched Next Gen Stats in 2013, a technology initiative in collaboration with Zebra Technologies. The technology aimed at capturing granular player data during every play by embedding Radio Frequency Identification (RFID) chips discreetly in the shoulder pads of players. This innovation paved the way for comprehensive data acquisition for the sport of American football. In 2014, data was being collected in 17 venues, and by the 2015 NFL season, Next Gen Stats achieved a global rollout across all stadiums, resulting in

substantial data analytics. The real-time data analysis and capabilities were promptly disseminated to sport's viewers, amplifying fan engagement and garnering interest from data scientists. The evolution of Next Gen Stats continues, with the data being publicly distributed in 2016, and the integration of microchips into game footballs by 2017 for further data enrichment.

The integration of this groundbreaking technology enabled a shift in the analysis of American football. These advances facilitated data-driven insights, including assessment of fourth-down conversion probabilities, yardage gains forecasts, individual player metrics and more. In 2019, the NFL introduced the Big Data Bowl, an innovative initiative aimed at motivating emerging talents to analyze play-by-play data for discernible trends and data-driven solutions. The NFL Big Data Bowl has been hosted on Kaggle since its inception in 2019, encouraging data scientists and data enthusiasts to participate and challenge themselves to decipher intricate data-driven challenges. Concurrently, it served as a catalyst for the application of machine learning techniques in the strategic decision-making processes of NFL franchises, spanning player recruitment, practice regiments, and in-game choices.

The focal point of the 2020 NFL Big Data Bowl pertained to estimating the yardage gained by an NFL player subsequent to receiving a handoff. The challenge's evaluation criteria hinged on the Continuous Ranked Probability Score (CRPS), computed by the formula:

$$C = \frac{1}{199N} \sum_{m=1}^{N} \sum_{n=-99}^{99} (P(y \leq n) - H(n - Y_m))^2$$

where P is the predicted distribution, N is the number of plays in the test set, Y is the actual yardage and $H(x)$ is the Heaviside step function ($H(x) = 1 \; for \; x \geq 0$ and zero otherwise). These guidelines came with the caveat that the submission would not score if any of the predicted values have $P(y \leq k) > P(y \leq k + 1)$ for any k (i.e. the CRPS must be non-decreasing)[3].

Aligned with the 2020 NFL Big Data Bowl challenge, our study recontextualized the prompt into a binary classification problem, exploring the prospect of predicting the attainment of a first down. This discrete inquiry set the stage for our analytical pursuit: identifying the optimal model types for predicting the achievement of a first down during a play. To this end, we explored three distinct models: logistic regression, gradient boosting and random forest.

Data and Methods:

The dataset, accessible via Kaggle, consisted of 49 variables, spanning from unique identifiers such as `PlayId`, to game identifier, `GameId`, to player attributes like position, speed, acceleration, defensive presence, down, and more. Furthermore, the data encompassed environmental factors, including the stadium location, turf type, weather conditions, humidity, wind speed, and wind direction. Cumulatively, the dataset encapsulated data from every play throughout the 2018 NFL season, totalling 682,154 rows.

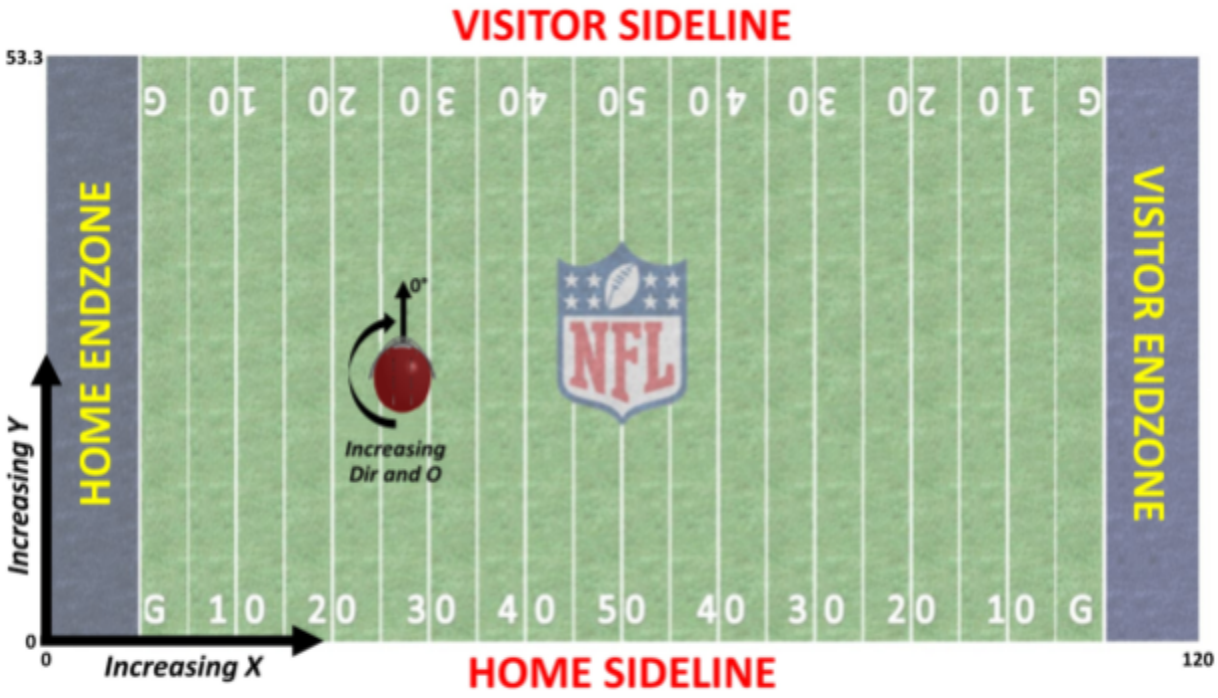

Figure 1: Player Position Along the Field

In preparation for model construction and exploratory data analysis (EDA), we introduced supplementary variables. We generated a feature variable, `DistanceToTouchdown`, delineating the distance to the end zone, computed by evaluating the field preposition relative to the possession team, and `ClosestDefenderDistance`, denoting the distance between the ball carrier (rusher) and the nearest defender. This was calculated by identifying the coordinates of the rusher and finding the shortest Euclidean distance of defending players to the rusher. Subsequent to calculating these variables, the creation of a binary feature, `first_down_gained`, signaled the attainment of a first down at the play's end and served as our target variable.

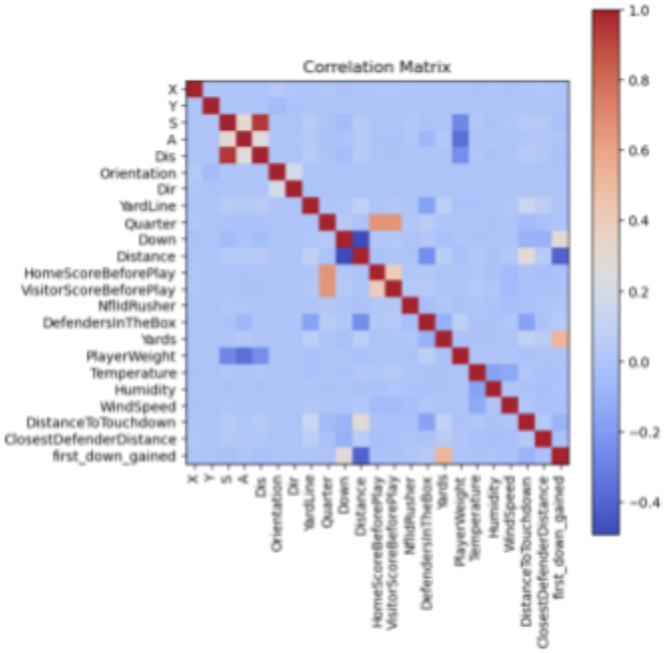

Figure 2: Correlation Matrix

To initiate EDA, we began with a statistical summary of the target variable, `first_down_gained`, followed by an analysis of correlations among all variables within the DataFrame. The correlation matrix, depicted as Figure 2, revealed a strong relationship between `Distance` (to a first down) and `first_down_gained`. Subsequent to this exploration, `DefendersInTheBox`, `Down`, `Distance`, `DistanceToTouchdown`, and

`ClosestDefenderDistance` emerged as the most influential variables. The culminating phase of EDA involved the construction of a violin plot, which visually depicted the density distribution of `first_downs_gained` as a function of `Distance`, highlighting o round out EDA, we constructed a violin plot to represent the density of `first_downs_gained` against `Distance` and a calculation determining 21% of all plays during the 2018 NFL season resulted in a first down.

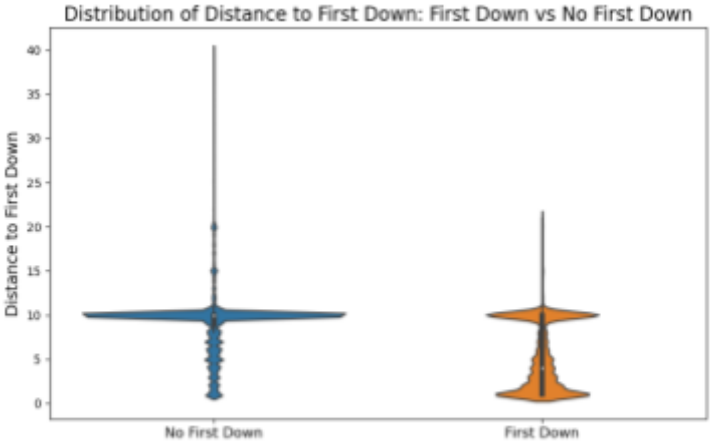

Figure 3: Violin Plot

The concluding steps of data preprocessing encompassed the removal of null values and the assembly of our features into a vector column. The definitive set of features employed for model creation comprised of `DefendersInTheBox`, `Down`, `Distance`, `DistanceToTouchdown`, and `ClosestDefenderDistance`, with `first_down_gained` serving as the target variable.

The data were partitioned into three sets, training, validation, and hold-out, distributed at a ration of 70-20-10. This division was conducted consistently using the seed value of 420. Additionally, our modeling endeavours incorporated k-fold cross-validation, where $k = 5$, to facilitate robust testing and validation across 5 distinct folds of the entire dataset. A dedicated function streamlined the fitting of all k-fold cross-validation models.

Model 1: Logistic Regression

Our benchmark model constituted a straightforward logistic regression. This model was fit utilizing the assembled feature vector containing `DefendersInTheBox`, `Down`, `Distance`, `DistanceToTouchdown`, and `ClosestDefenderDistance` predicting `first_down_gained`.

Figure 4: Logistic Regression Confusion Matrix (Data Splitting)

Model Evaluation

The logistic regression model performed credibly when utilizing the split training and test sets, culminating an accuracy of 82.59%, precision score of 80.93%, recall of 82.59%, F1 Score of 0.8089, and Area Under the Curve (AUC) of 0.7642. These metrics not only facilitated benchmarking but also established a target for subsequent models, encompassing the aspiration of achieving an AUC surpassing 0.8.

A subsequent execution of the logistic regression model entailed k-fold cross-validation, rendering akin results to those procured via data splitting. The k-fold cross-validated logistic regression yielded a test accuracy of 82.63%, precision of 80.95%, recall of 82.63%, F1 score of 0.8096, and AUC of 0.7625.

Model 2: Gradient Boosting

The successive model pursued was a gradient boosted tree, an ensemble model harnessing the strength of multiple individual trees for classification or regression, orchestrated through the mechanism of boosting, iteratively addressing errors through sequential weak learners.

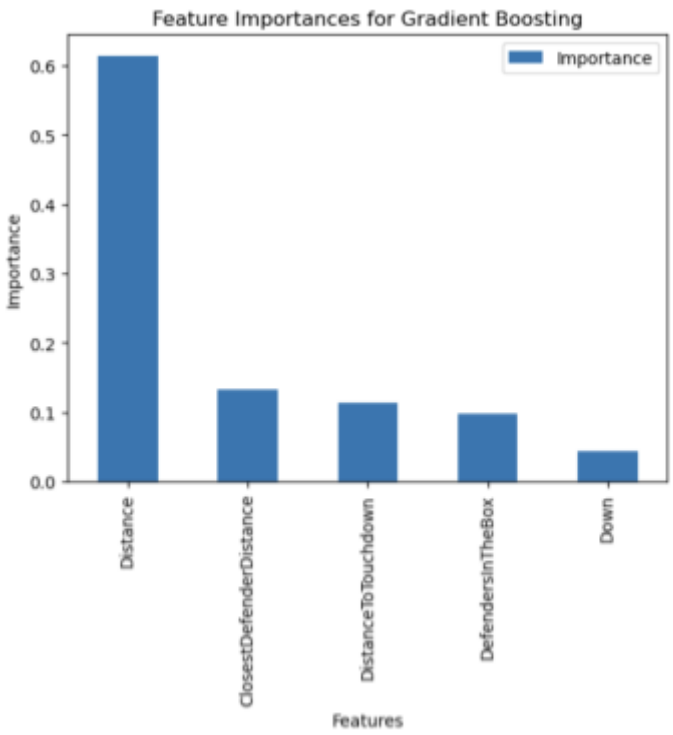

Figure 5: Gradient Boosting Model Feature Importance

Model Evaluation

The gradient boosting model, fitted with the training set and evaluated with the validation set, demonstrated stronger performance compared to its logistic regression counterpart. Delivering elevated metrics across data splitting and k-fold cross-validation, the gradient boosting model with data splitting exhibited an accuracy of 83.15%, precision of 81.66%, recall of 83.15%, F1 score of 0.8158, and AUC of 0.7793. The k-fold cross-validation variant, while trailing the data splitting results, demonstrated an accuracy of 82.98%, precision of 81.41%, recall of 82.98%, F1 score of 0.8143, and AUC of 0.7750.

`Distance` emerged as the predominant variable, as seen in Figure 5, reinforcing the findings from preliminary EDA and subsequently informing the model's prediction capability.

Model 3: Random Forest

The conclusive model implemented was a random forest model, renowned for its ensemble nature that combines outputs from individual trees. Distinctive from gradient boosted trees, random forests adopt the bagging technique to construct individual decision trees.

The random forest models were fit with parameters `numTrees=10` and `maxDepth=5`.

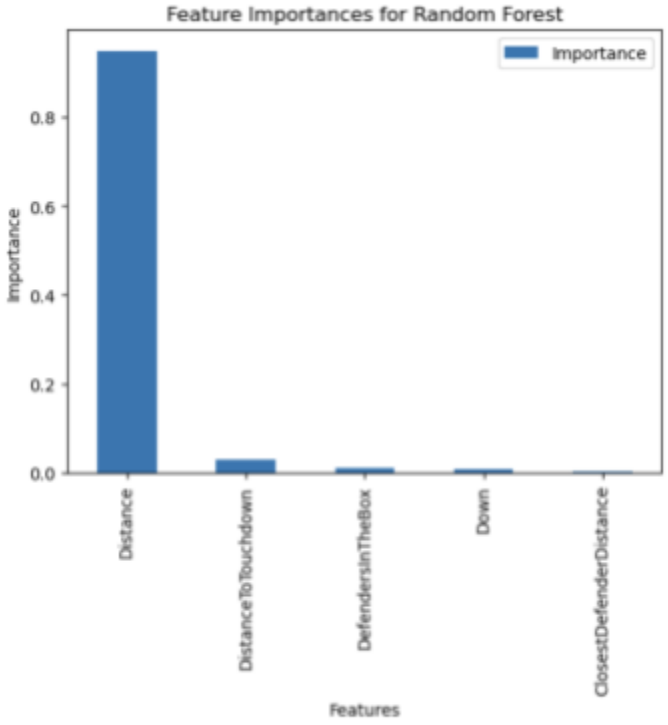

Figure 6: Random Forest Feature Importance

Model Evaluation

In the random forest model, the variable `Distance` retained its prominent standing, aligning with the findings of the gradient boosting model and corroborating EDA observations. Notably, the second-most influential variables exhibited disparity between the random forest model and the gradient boosting model.

Metric analysis of the random forest models revealed poorer performance than that of the gradient boosting models. The random forest model fit and tested utilizing data splitting returned an accuracy of 82.95%, precision of 81.33%, recall of 82.95%, F1 score of 0.8139, and AUC of 0.7604. The random forest model with k-fold cross-validation yielded an accuracy of 82.87%, precision of 81.29%, recall of 82.87%, F1 score of 0.8135, and AUC of 0.7623.

Results:

Collectively, the models performed in close proximity. Nonetheless, the gradient boosting model emerged as the champion model, excelling in terms of performance metrics across both data splitting and k-fold cross-validation. Its elevated accuracy, precision, recall, F1 score, and AUC, albeit marginally, affirmed its heightened predictive capabilities.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression (Data Splitting)	82.59%	80.93%	82.59%	0.8089	0.7642
Logistic Regression (k-Fold CV)	82.63%	80.95%	82.63%	0.8096	0.7625
Gradient Boosting (Data Splitting)	83.15%	81.66%	83.15%	0.8158	0.7793
Gradient Boosting (k-Fold CV)	82.98%	81.41%	82.98%	0.8143	0.7750
Random Forest (Data Splitting)	82.95%	81.33%	82.95%	0.8139	0.7604
Random Forest (k-Fold CV)	82.87%	81.29%	82.87%	0.8135	0.7623

Conclusions:

In summation, the gradient boosting models emerged as the most proficient among the models evaluated for predicting `first_down_gained`. Leveraging features such as `DefendersInTheBox`, `Down`, `Distance`, `DistanceToTouchdown`, and `ClosestDefenderDistance`, the gradient boosting models edged past their counterparts across all metrics.

For future endeavors, we recommend delving into models that encompass an expanded dataset, perhaps drawing from the same dataset and further harnessing variable selection technique and model parameter tuning. Exploring the integration of environmental data in play outcome prediction and considering supplementary data sources, like altitude and sun position, hold promise. Furthermore, the exploration of deep learning techniques, notably convolutional neural networks (CNN), which have shown prowess in predicting NFL play outcomes, could provide enhanced predictions, given the intricacies to American football.

References:

[1] “NFL Big Data Bowl,” NFL Football Operations, https://operations.nfl.com/gameday/analytics/big-data-bowl/#:~:text=Each%20year%2C%20the%20NFL%20Big,of%20players%2C%20plays%20and%20situations.

[2] “NFL announces inaugural Big Data Bowl,” NFL Communications, https://nflcommunications.com/Pages/NFL-ANNOUNCES-INAUGURAL-BIG-DATA-BOWL.aspx.

[3] “NFL Big Data Bowl,” Kaggle, https://www.kaggle.com/competitions/nfl-big-data-bowl-2020/.

[4] “DS 5110 big data systems project tackling the NFL 2020 big data bowl,” GitHub, https://github.com/lauren-odonnell/DS5110_NFL_BigDataBowl.

[5] “NFL next gen stats,” NFL Football Operations, https://operations.nfl.com/gameday/technology/nfl-next-gen-stats/.