2020 NFL Big Data Bowl Kaggle Competition

Gabrielle Rappaport[1], Hamza Tazi Bouardi[1], Pierre-Henri Ramirez[1] and Danial Mirza[1]

[1]Master of Business Analytics, Operations Research Center, Massachusetts Institute of Technology

I. Context

Given a dataset with play-by-play information from regular season NFL games, we want to know:

- Can we leverage analytics techniques to predict a **potentially game-making play?**
- How can players, coaches and managers use these insights to **optimize playing strategies** and change the course of key games?
- What are the implications for estimating the odds and making **real-time wagers** on granular, play level outcomes?

II. Problem

Whether it's winning yards to maintain possession or crossing the goal line to score the final touchdown, the distance travelled by a player before they are tackled is a crucial metric in any NFL game.

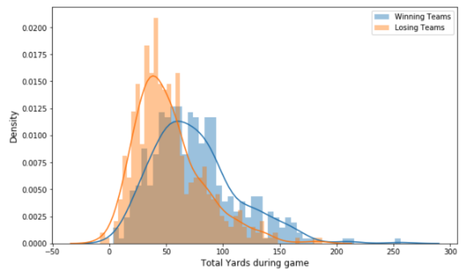


Figure 1: Distribution of yards travelled, winning teams vs. losing teams

Our goal will be to predict the number of yards the attacking team will gain in a single down.

American football is played between two teams of 11 players each. The goal of the game is to score points by advancing the ball to the opponent's endzone. The attacking team must travel at least 10 yards with the ball within 4 attempts (known as "downs") in order to maintain possession. Meanwhile, the defending team try to intercept the ball. If the attacking team is successful, they receive 4 more attempts to continue advancing towards the scoring zone. Otherwise, the possession of the ball changes sides, and the defending team goes on the attack.

We develop and implement robust ensemble models for prediction. A key consideration is whether to predict the distance travelled by each offensive player in a given down, or to predict for the entire team at once. We investigate and compare these approaches across a range of different model specifications.

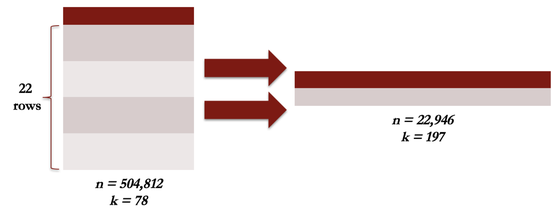
III. Data

Our dataset captures **real time location data, speed and acceleration** for every player on the field as well as additional **game play and game environment metrics**. Sensors track tags charting individual movements within inches. Each row corresponds to a player's activity in a single down.

- 512 games over two NFL regular seasons (2017-2018)
- 23,171 plays or downs

IV. Feature Engineering

1. **Global pre-processing**
   These steps are applied on the full dataset, where we predict the yardage for each offensive player separately in a given down.
   - Standardise play direction so as to compare offense and defense.
   - Standardizing categorical variables into smaller categories: Turf, Wind direction, Stadium type, Weather.
   - Treat and remove some outliers
2. **Table Aggregation**
   These steps are applied on the aggregate dataset, where we produce one prediction for the yardage travelled by the offensive team as a whole.
   - Filter key playing positions for each team
     Offense Strategic Positions: OT, C TE, T, G, FB, RB, HB, OG.
     Defense Strategic Positions: NT, MLB, OLB, DL, LB, DE, ILB, DT.
   - Aggregate player level features by position for each down



V. Modeling

1. **General Fitting Methodology**
   - The models' hyperparameters were tuned using 5-Fold Cross-Validation using either the **CRPS** or **RMSE** as a scoring metric
   - To produce the CDF prediction, we did a yard by yard weight prediction which we then normalised to get a CDF of 1.
2. **Data Aggregation approach**
   - We fitted Feedforward Neural Networks, Sparse Linear Regression, Random Forest, XGBoost and CART to predict a distribution or single value for the whole team.
   - We tried using features selected by Sparse LR in other models however the results were not conclusive and were discarded.
3. **Ensemble Learning approach :**
   - We produce individual (player by player) yard predictions, for both the offensive team and the defensive team.
   - From these outputs, we created a second model, taking as inputs the ordered predictions of the first model. To fit such a model on unseen data, we trained the initial model on the first half of the training data, making predictions on the other half, and then vice versa, training on the second portion and predicting on the first.
   - The first stage model was either Random Forest or XGBoost, and the subsequent model was either Linear Regression, Random Forest or XGBoost.

VI. Results

We evaluate our models on different metrics, for the full probability distribution or single value prediction, for both the full dataset (Individual) and the aggregate dataset (Whole Team).

Dataset	**FNN**	**RF**	**XGB**	**XGB^2**
Individual	0.0130	0.0107	0.0092	**0.0089**
Whole Team	0.0134	0.0123	**0.0102**	NA

Table 1: Probability Distribution on validation set (**CRPS**)

Dataset	**FNN**	**SparseLR**	**RF**	**XGB**	**CART**
Individual	6.038	6.204	6.097	**5.947**	6.173
Whole Team	6.185	6.311	6.117	**5.963**	6.194

Table 2: Single Value Prediction on validation set for Classic models (**RMSE**)

Dataset	**RF+LR**	**RF^2**	**RF+XGB**	**XGB+LR**	**XGB+RF**	**XGB^2**
Individual	6.204	6.211	6.340	5.250	4.145	**3.978**

Table 3: Single Value Prediction on validation set for Ensemble (**RMSE**)

The best performance was obtained on individual predictions with an Ensemble method using 2 XGBoost models, leading to an $OSR^2 = 0.59$

Feature Importance

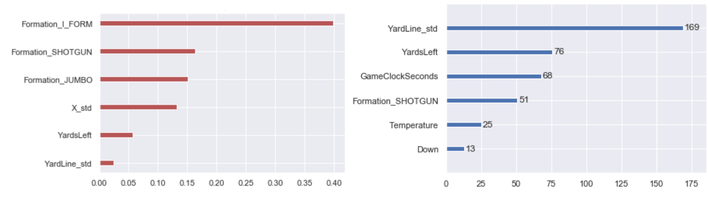


Figure 2: Feature Importance for Random Forest (**Left**) and XGBoost (**Right**)

- **Yards Left:** Yards remaining to opponent's endzone is a key motivating factor for distance travelled in a given play. If the rusher is nearing the goal line, we expect a shorter distance traveled. Meanwhile, if there is a lot of ground to cover, offense team is likely to make a long-yardage play.
- **Jumbo Formation:** Used exclusively in short-yardage plays, especially near the goal line. Designed to score points by brute force, in situations where distance required is small. The model therefore confirms our intuition.

VII. Takeaways

Towards Better Playing Strategies

Our best performing model could be leveraged by coaches to investigate the sensitivity of yardage to key play decisions such as the offense formation. If model predicts an improvement under alternate formation, this can inform decisions in the next down, motivating real-time playing strategies at a granular level.

Revolutionising real-time betting

Real time tracking and analysis from video data allows us to move from bets made on the outcome of whole games, to more specific bets made at a play-by-play level. By computing a probability distribution of the predicted yardage before the end of a down, we can accurately compute the odds of key events and make bets accordingly.