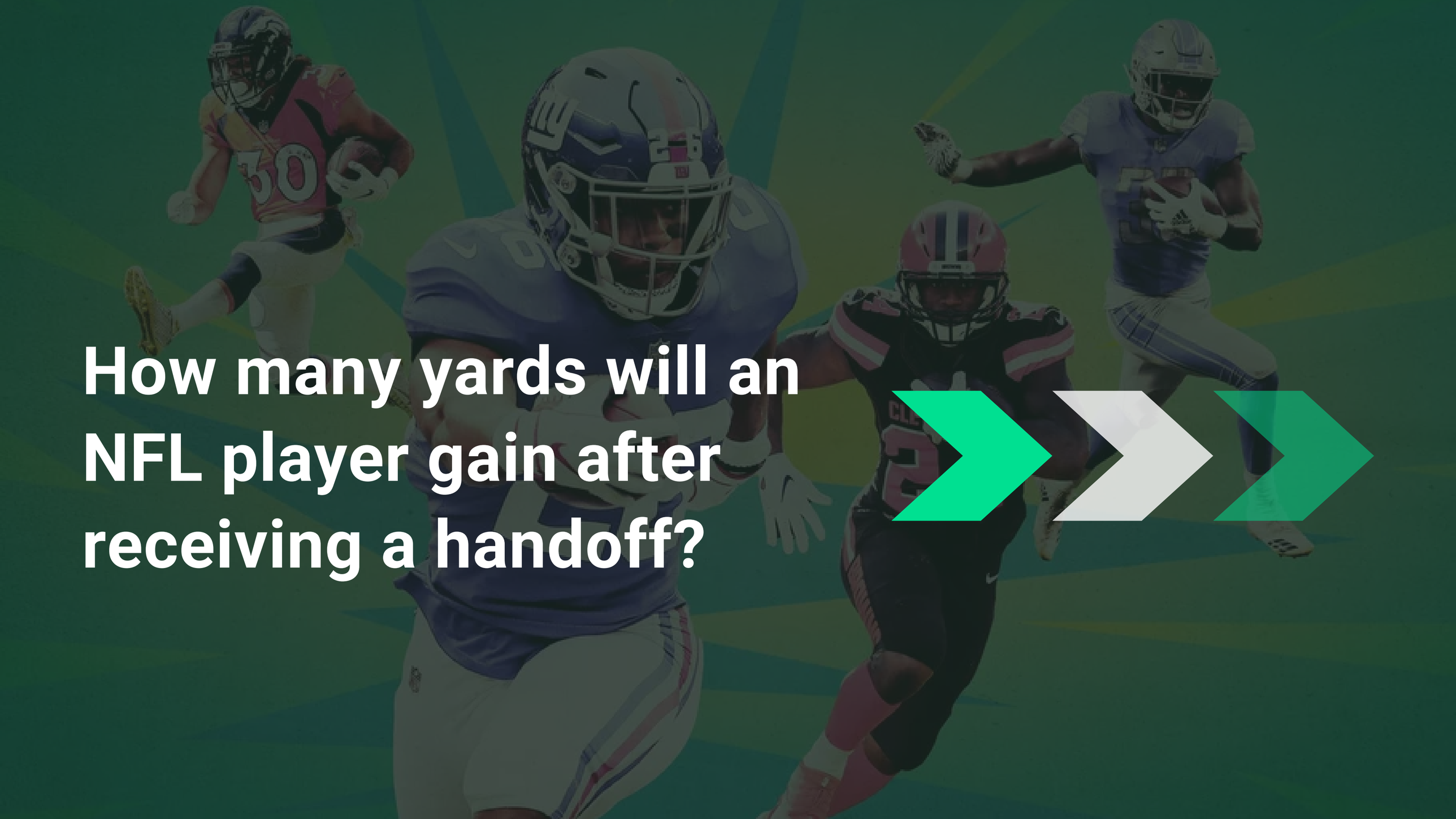
FEATURED KAGGLE CODE COMPETITION

NFL BIG DATA BOWL

By Patrick Ly

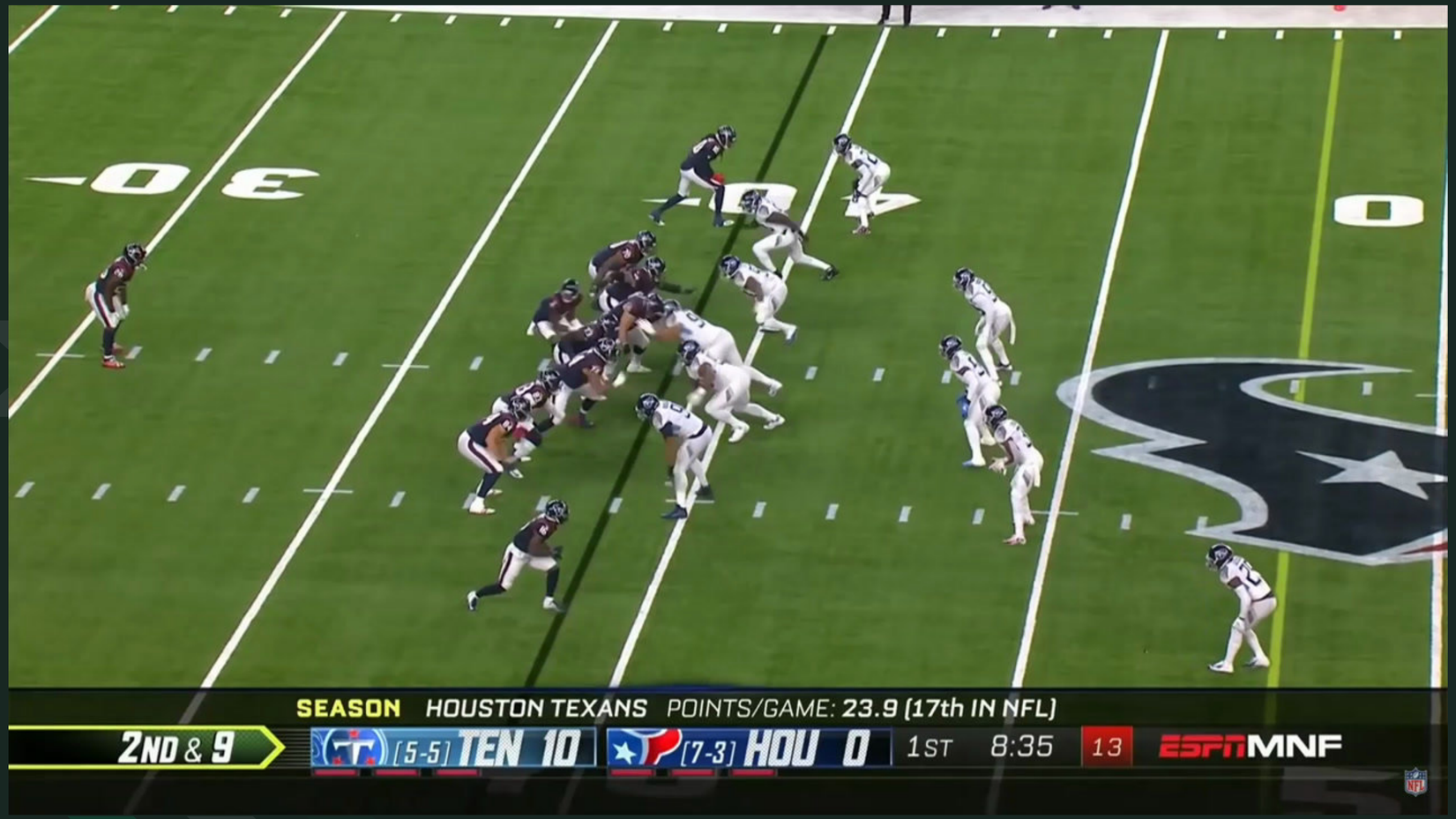


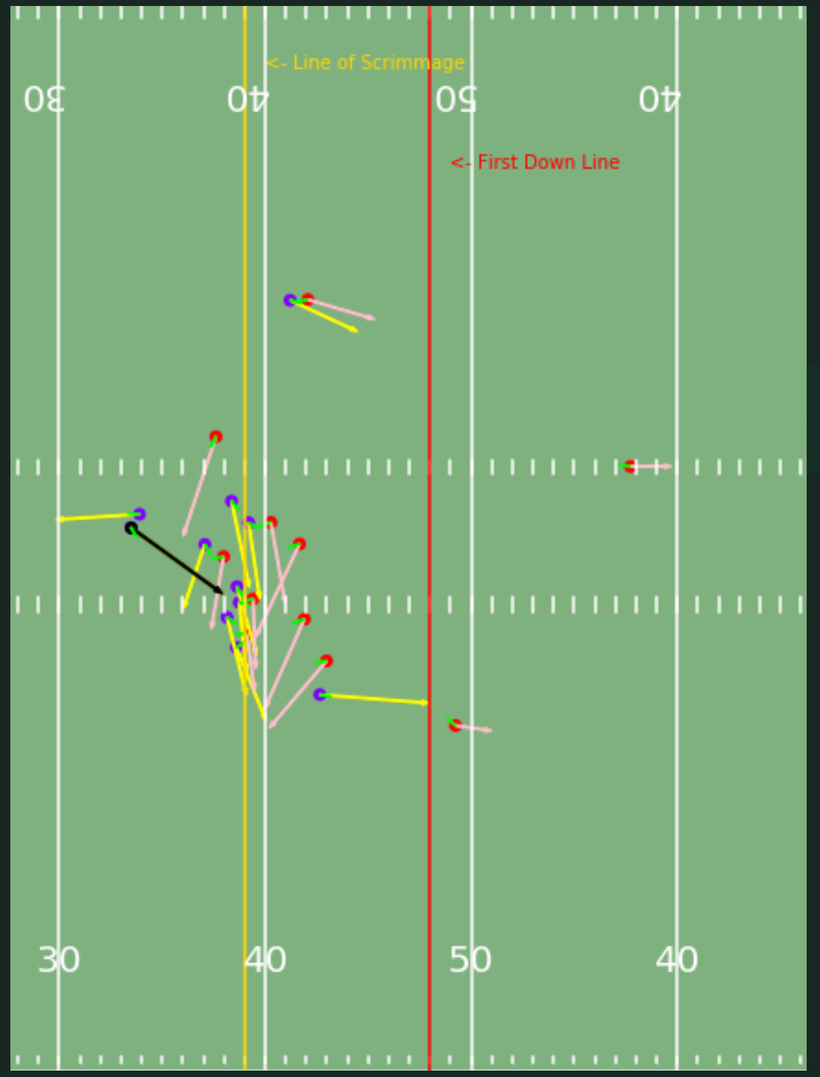
How many yards will an NFL player gain after receiving a handoff?

Player Tracker Data

- **X:** player position along the long axis of field
- **Y:** player position along the short axis of field
- **S:** speed in yards/second
- **A:** acceleration in yard/second^2
- **DIS:** distance traveled prior time points in yards
- **DIR:** angle of player motion (deg)
- **ORIENTATION:** orientation of player (deg)

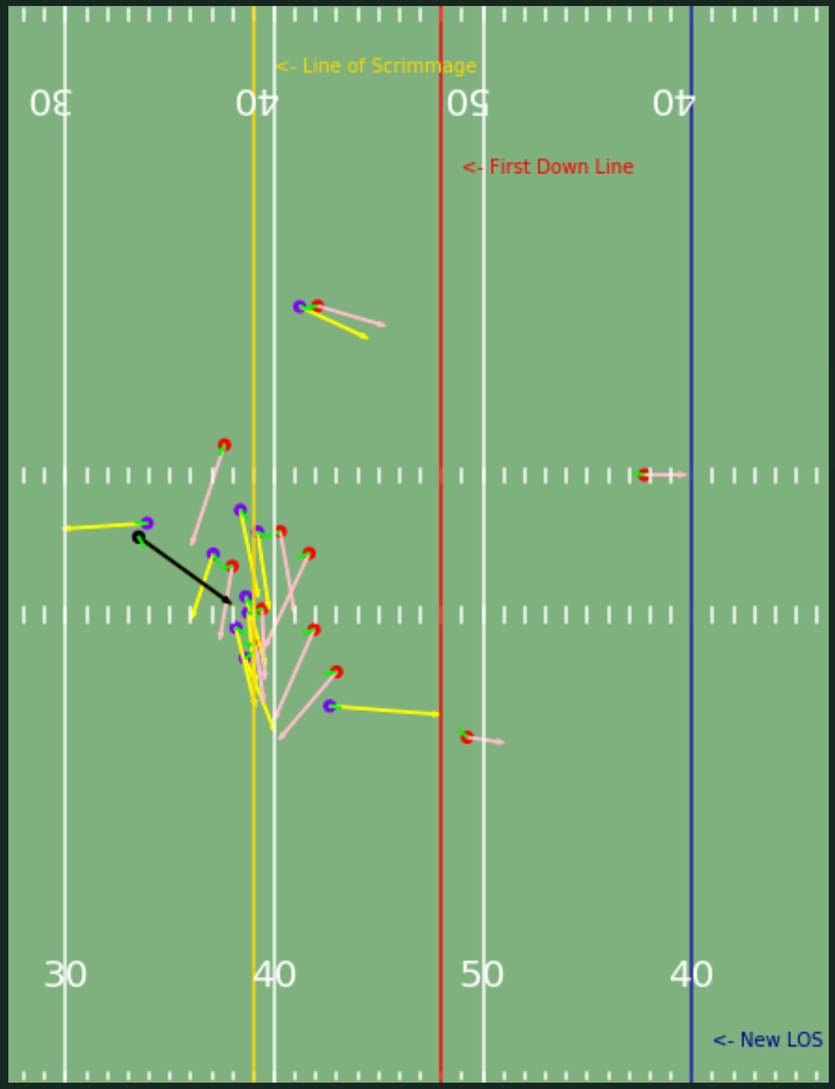
CASE # 1: LAMAR MILLER, WEEK 12 OF 2018





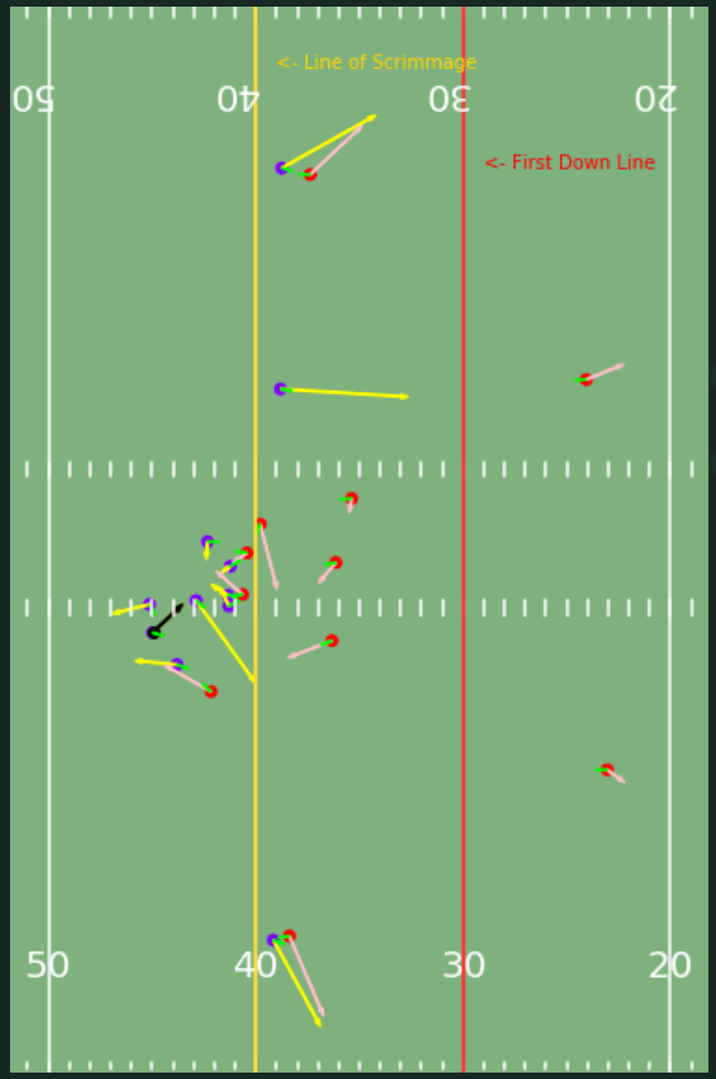
Result: ?

CASE # 1: LAMAR MILLER, WEEK 12 OF 2018



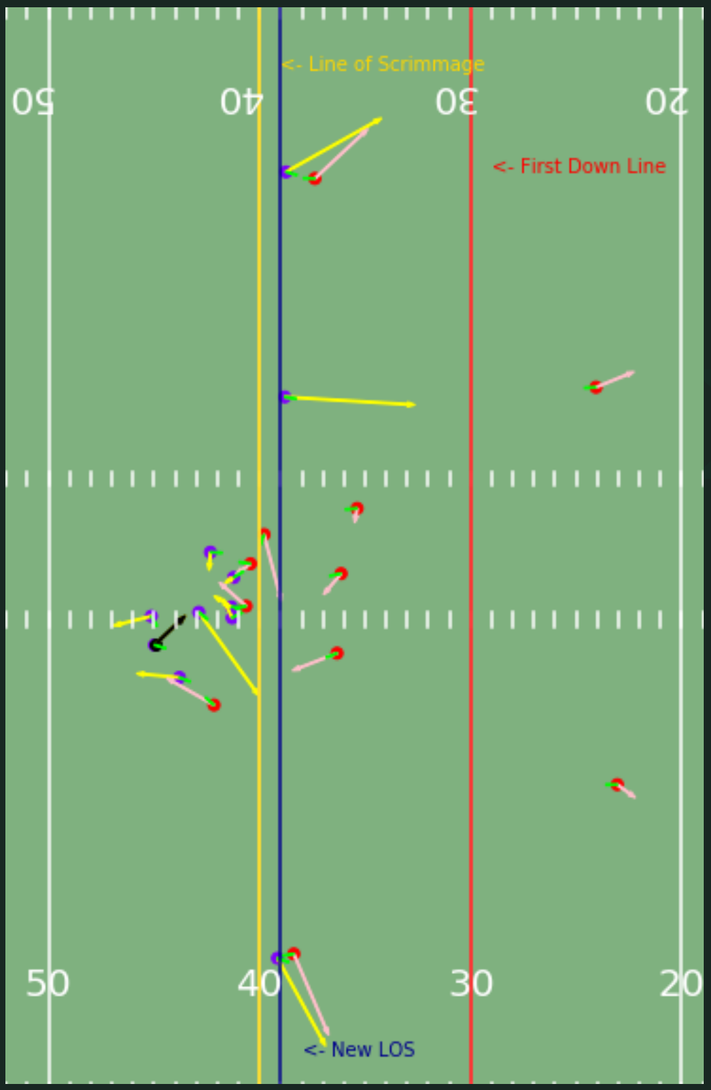
Result: +21 yards

CASE # 2: LAMAR MILLER, WEEK 12 OF 2018



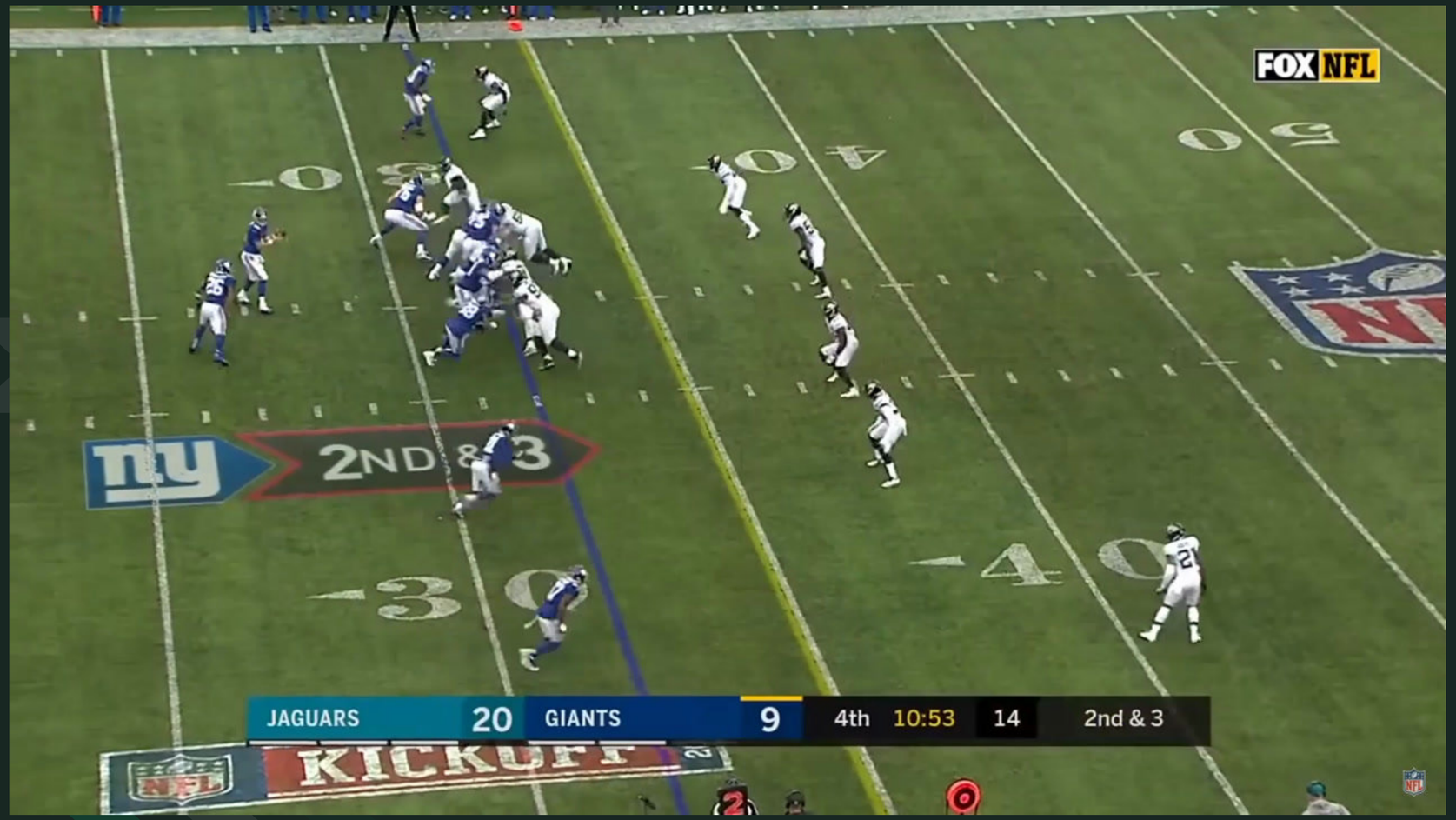
Result: ?

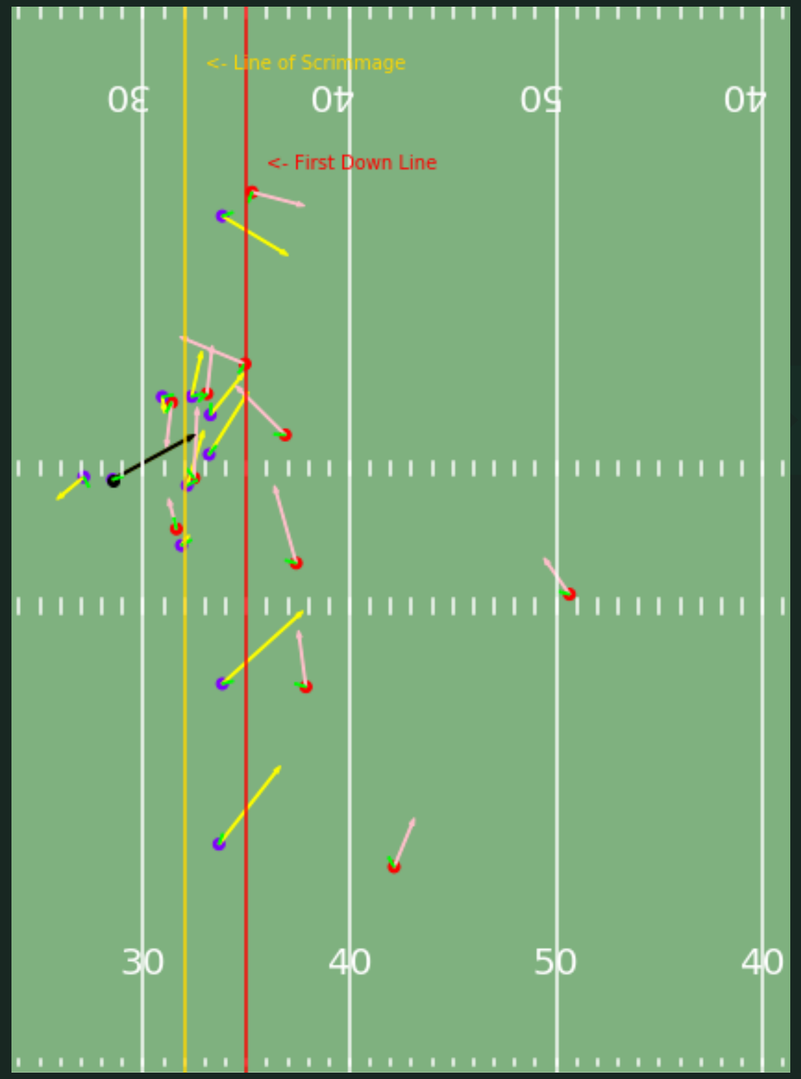
CASE # 2: LAMAR MILLER, WEEK 12 OF 2018



Result: +1 yard

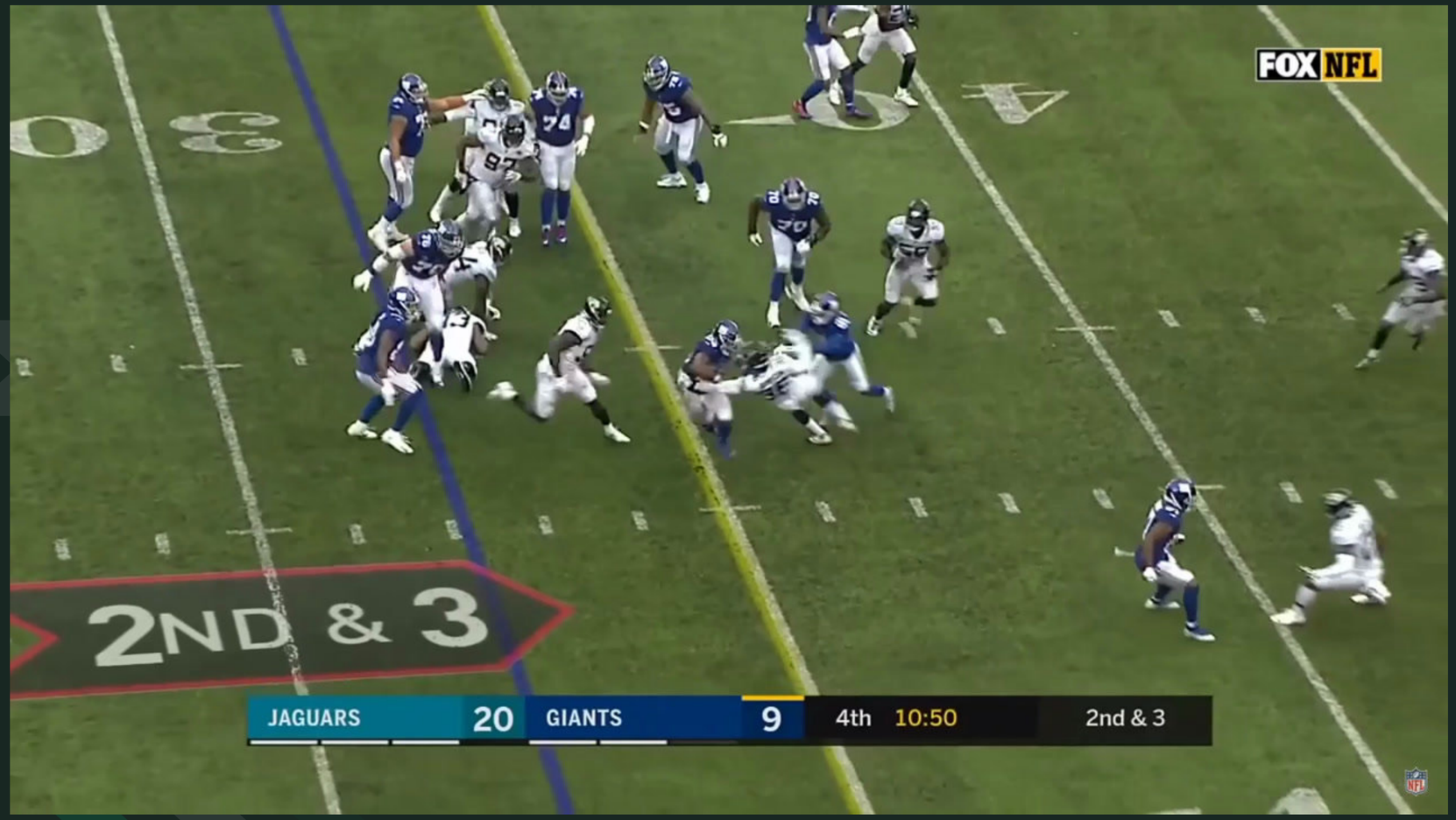
CASE # 3: SAQUON BARKLEY, WEEK 1 OF 2018

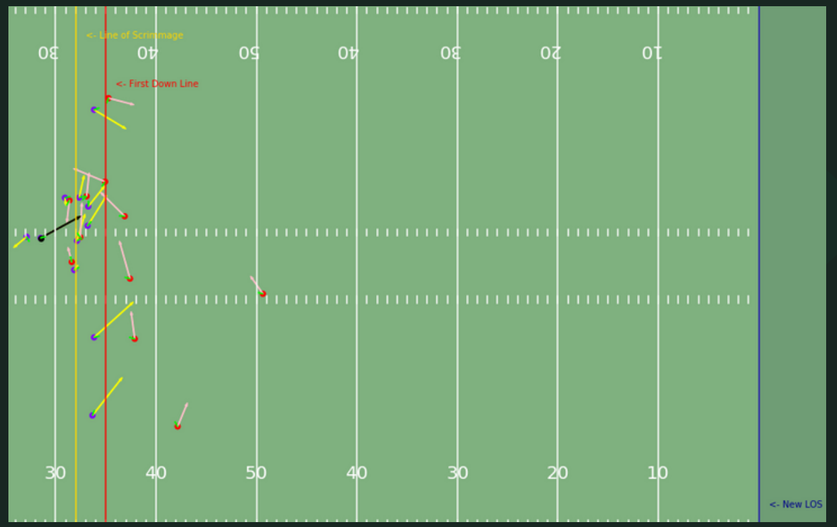




Result: ?

CASE # 3: SAQUON BARKLEY, WEEK 1 OF 2018





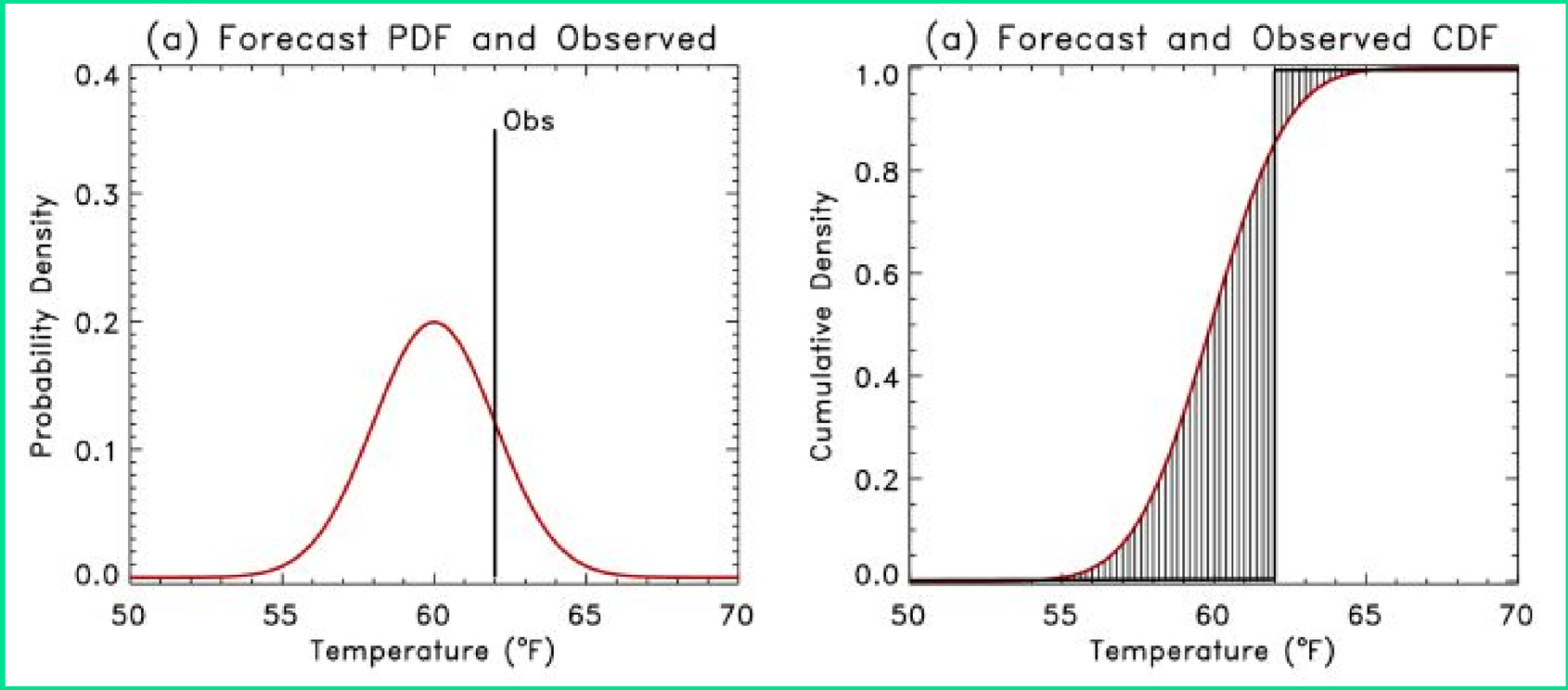
Result: +68 yards

Model Goal & Evaluation Metric

PROBABILITES OF EVERY OUTCOME

In this competition, we will be evaluated on the **Continuous Ranked Probability Score (CRPS).**
In laymen's term, we are creating a probability distribution of every possible outcome for each rush play.

Note**: Taking CRPS and dividing by # of possible outcomes will yield **Mean Absolute Error (MAE)



Model-less Median Benchmark

MEDIAN PREDICTION

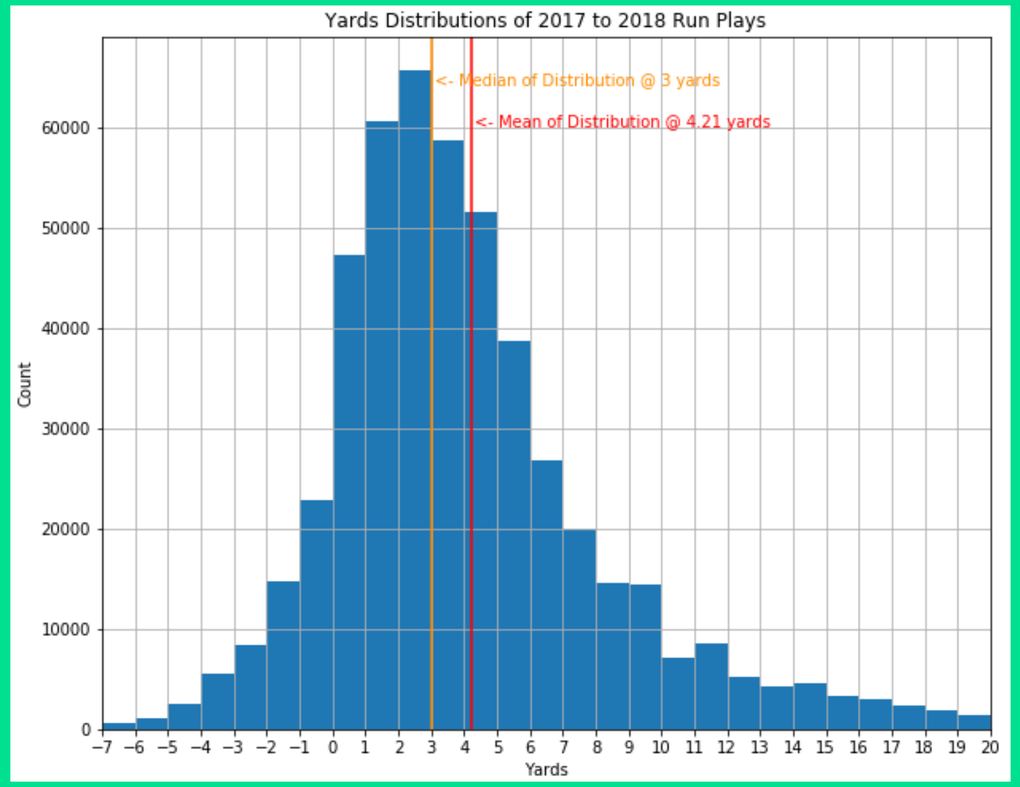
Given no knowledge, what would one predict? How about what usually happens? A median prediction of **3 yards** would be a good benchmark.

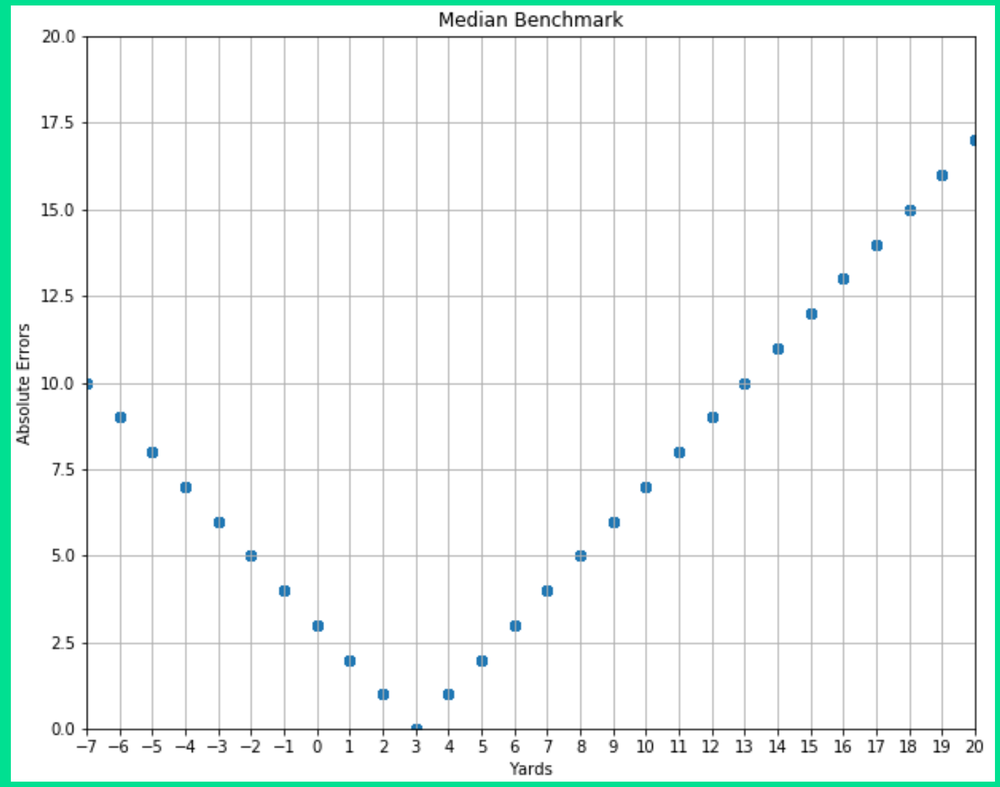
CRPS

0.01845

MAE

3.67 yards





MODEL FRAMEWORK

DECISION TREE

+ HIGH INTERPRETABILITY
+ CAN ATTRIBUTE VARIABLES THAT INFLUENCE RUN PLAYS

- HEAVILY RELIANCE ON FE
- MIGHT NEED SMOOTHING

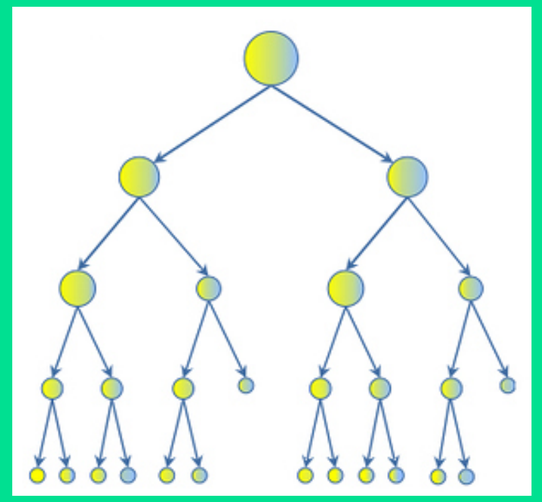


Fig. Decision Tree

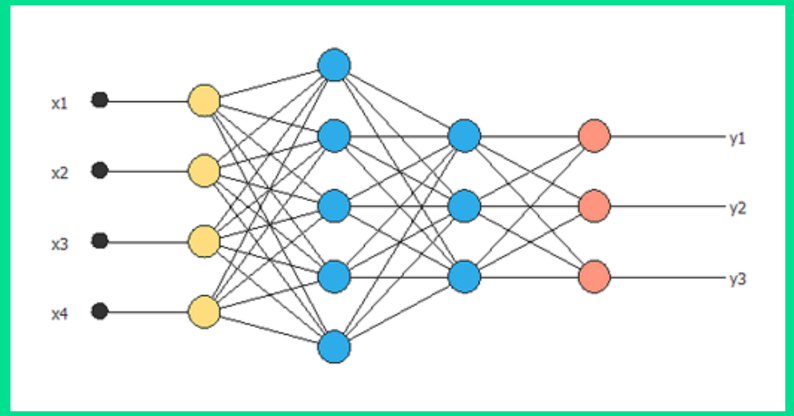


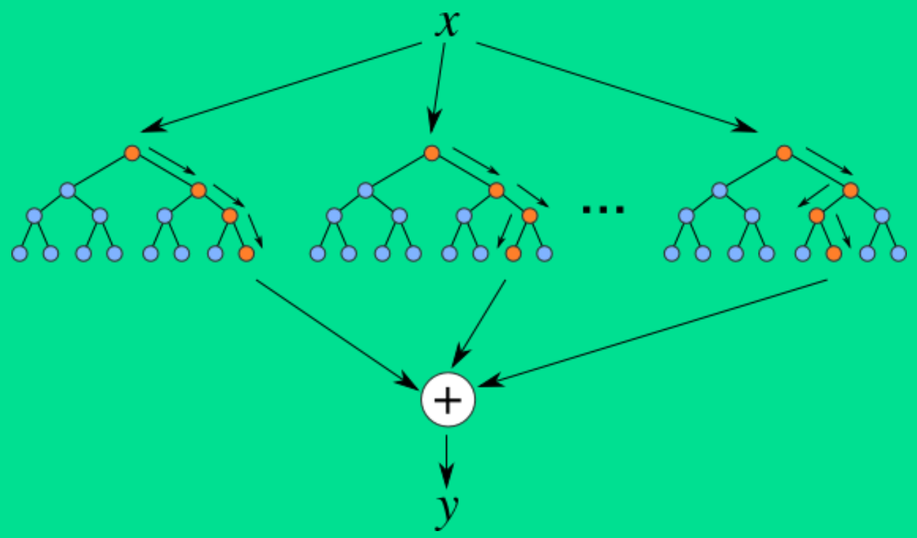
Fig. Neural Network

NEURAL NET

+ PROBABLISTIC OUTPUT
+ LESS FEATURE ENGINEERING

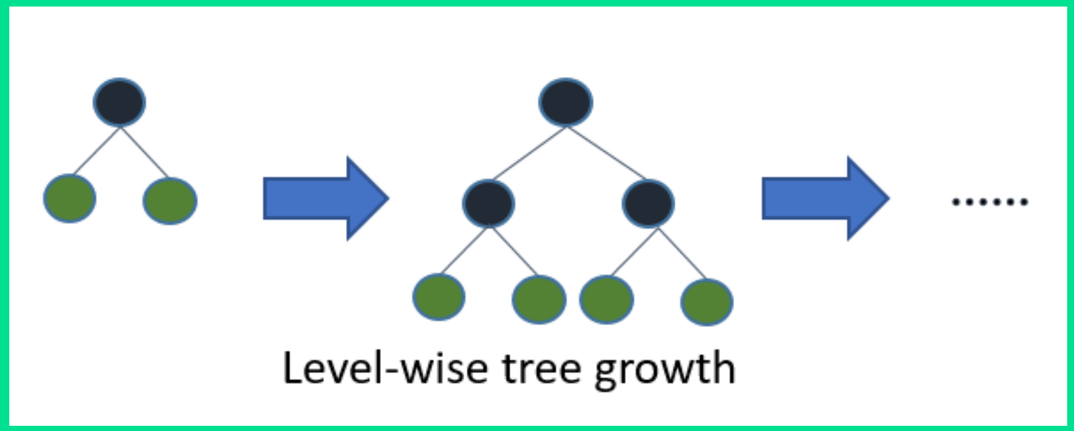
- HARD TO INTERPRET
- CANNOT DIRECTLY IDENTIFY KEY RUN FACTORS

Decision Tree-based Models



RANDOM FOREST

- Ensemble learning method utilizing multiple decision trees
- Independent fully decision trees on random subset of data (bagging)



Microsoft
LightGBM

- Advanced ensemble learning method with a gradient boosting framework.
- Can incorporate bagging techniques
- Implementation of boosting with weak learners (iterative process on the residuals)

1ST ITERATION MODELS

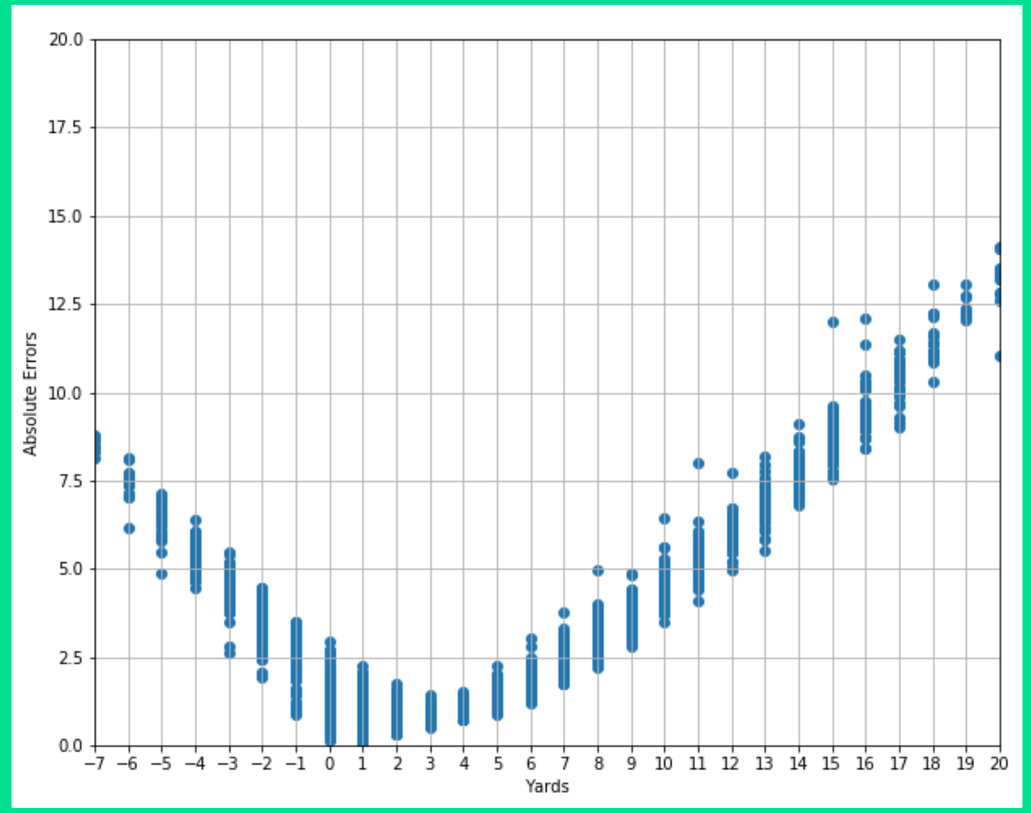
RAW DATA MODELS

What happens if we do some preprocessing (cleaning & standardizing) but leave the data untouched?

What would the models prioritize?

Raw Data Models: Results

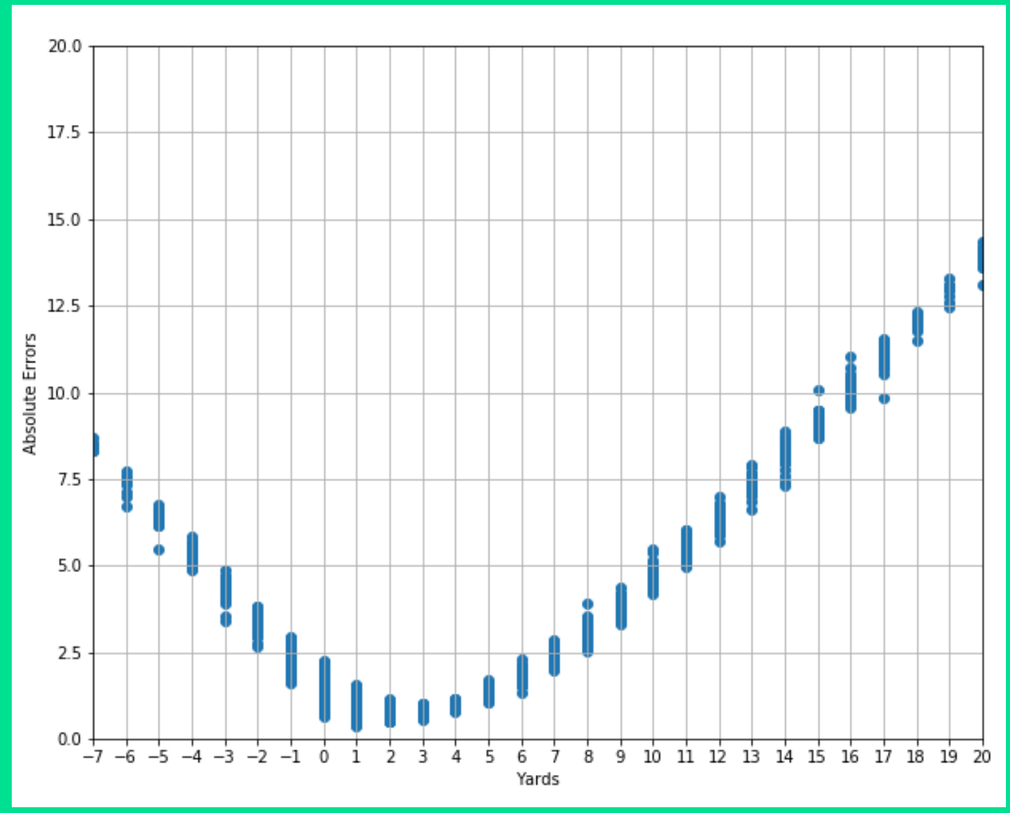
RANDOM FOREST



CRPS
0.01376

MAE
2.74 yards

LIGHTGBM

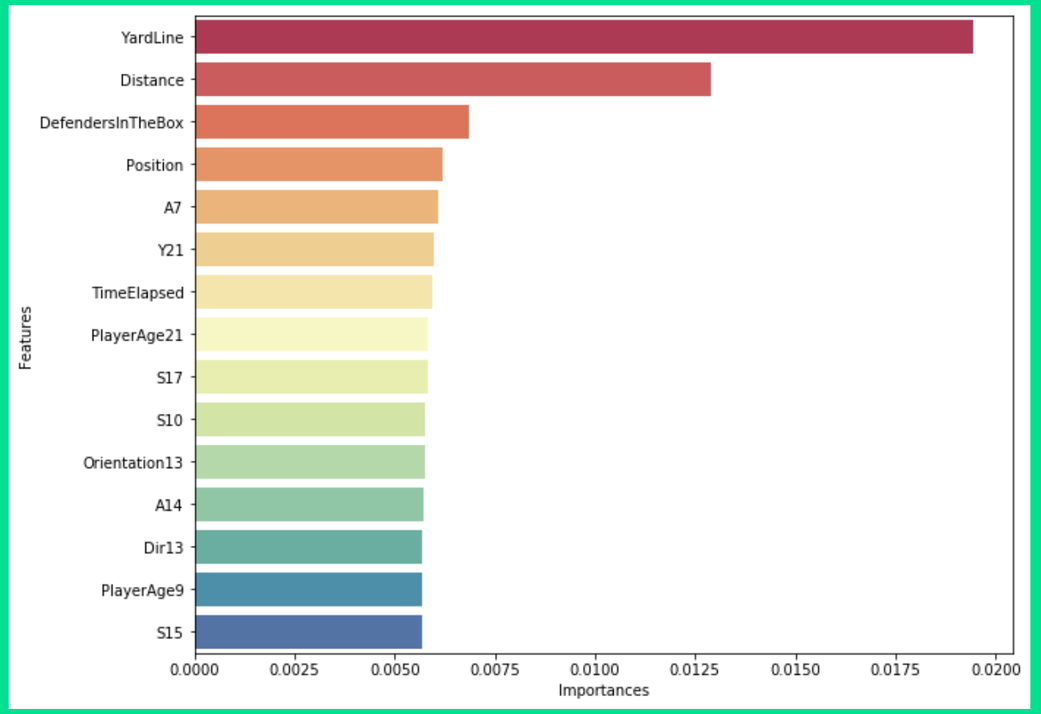


CRPS
0.01388

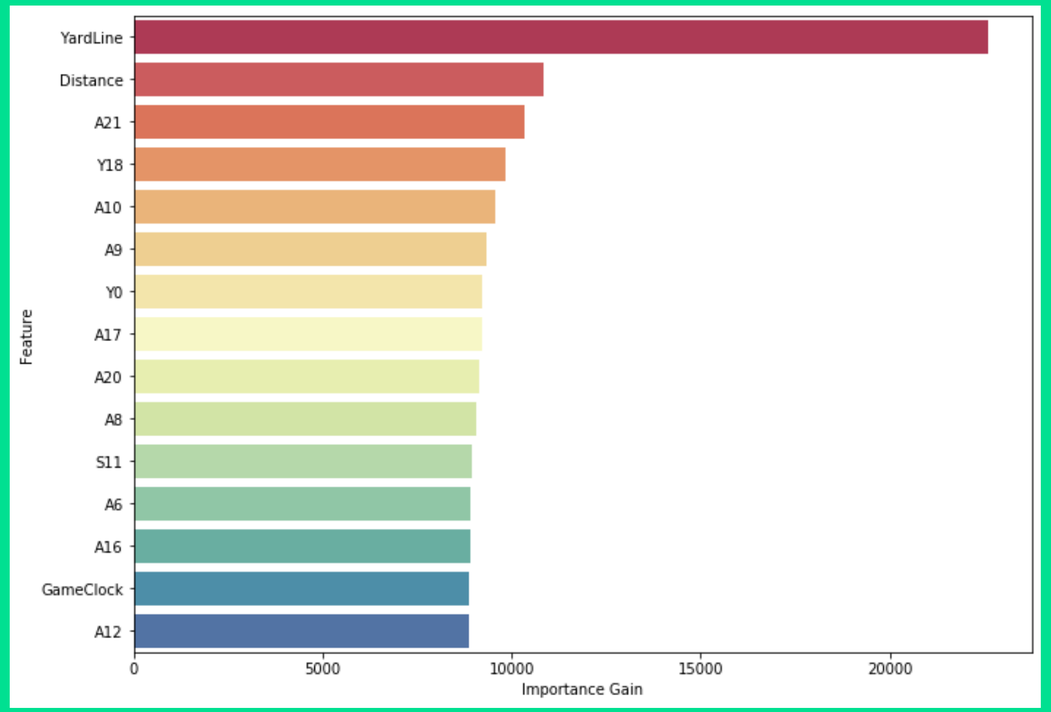
MAE
2.76 yards

Raw Data Models: Feature Importance

RANDOM FOREST



LIGHTGBM



- **YARDLINE:** yard line of the line of scrimmage
- **DISTANCE:** yards needed for a first down
- **DEFENDERS IN THE BOX:** # of defenders lined up near LOS

KEY FINDINGS

- Both models relied on contextual game data.
- Models could not intrepet raw player data.

RUSHER-CENTERED MODELS

Who was the most influence on a run play? A rusher-centered model could capture the main interactions of a running play.

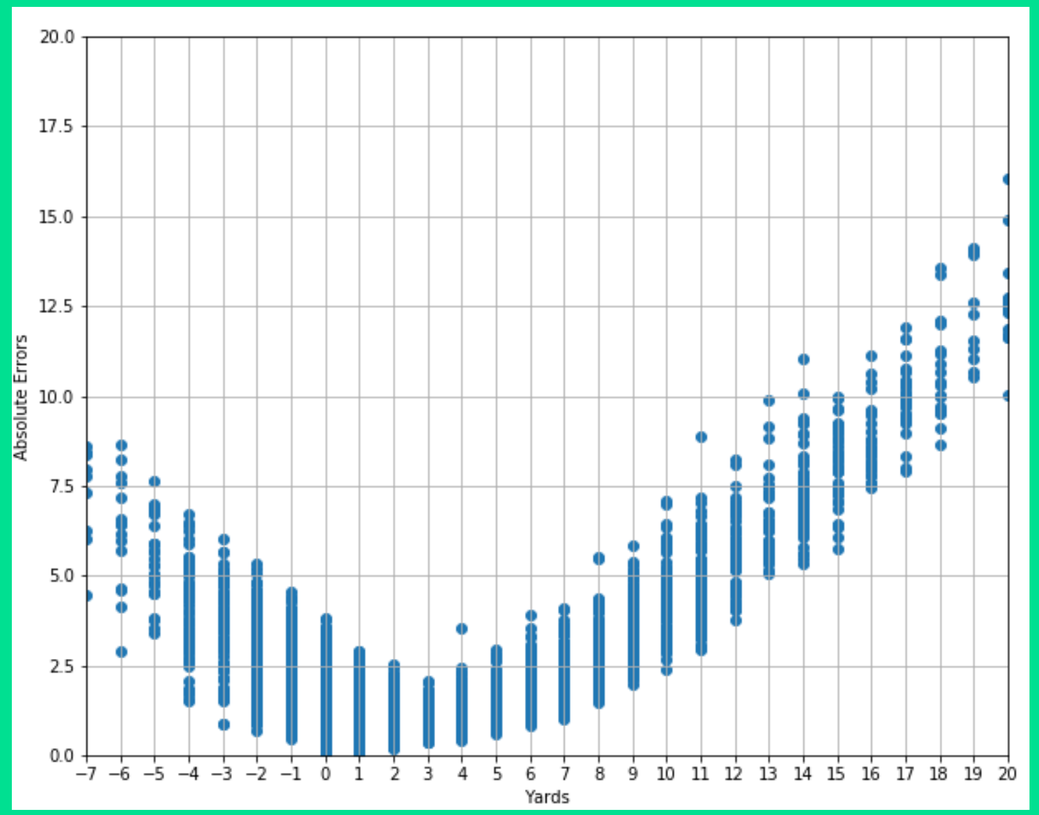
The following models will have their data structure as:

1. **Rusher** - ball carrier
    a. raw rusher player data
2. **Defense** - simplified as one defense unit
    a. general position descriptors
    b. position and time based relations to rusher
    c. penetration measures
3. **Offense** - simplified as one offense unit
    a. general position descriptors
    b. position and time based relations to rusher

FEATURED-ENGINEERED MODELS

Rusher Models: Results

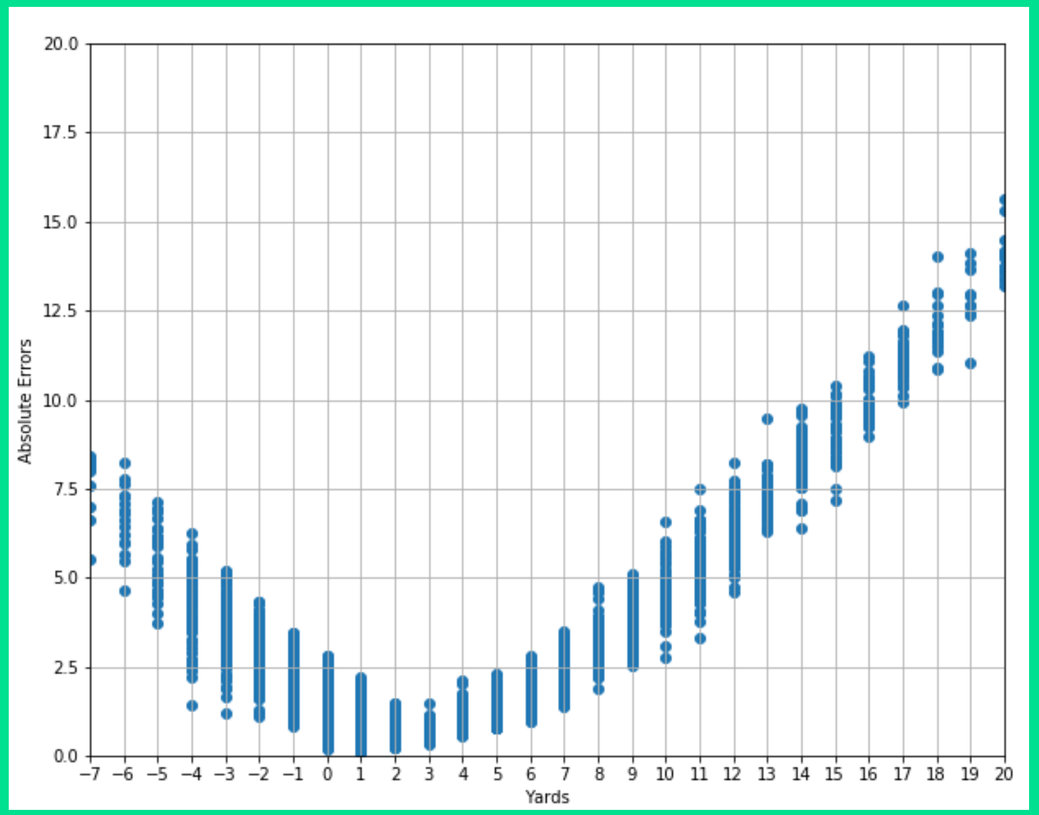
RANDOM FOREST



CRPS
0.01288

MAE
2.56 yards

LightGBM

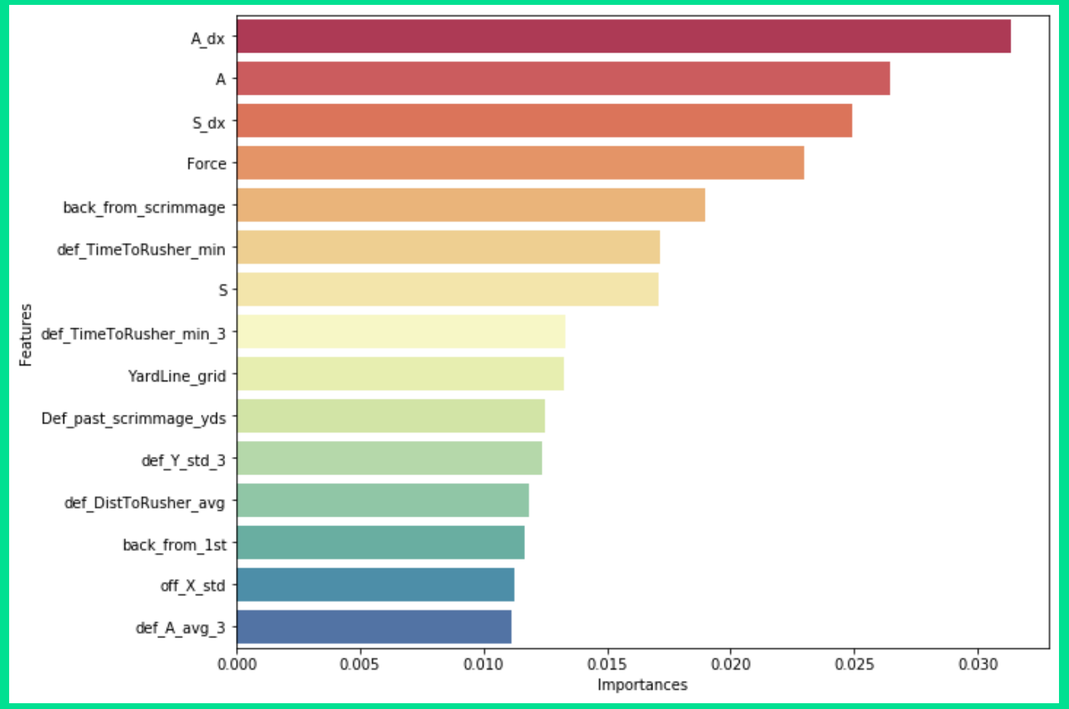


CRPS
0.01314

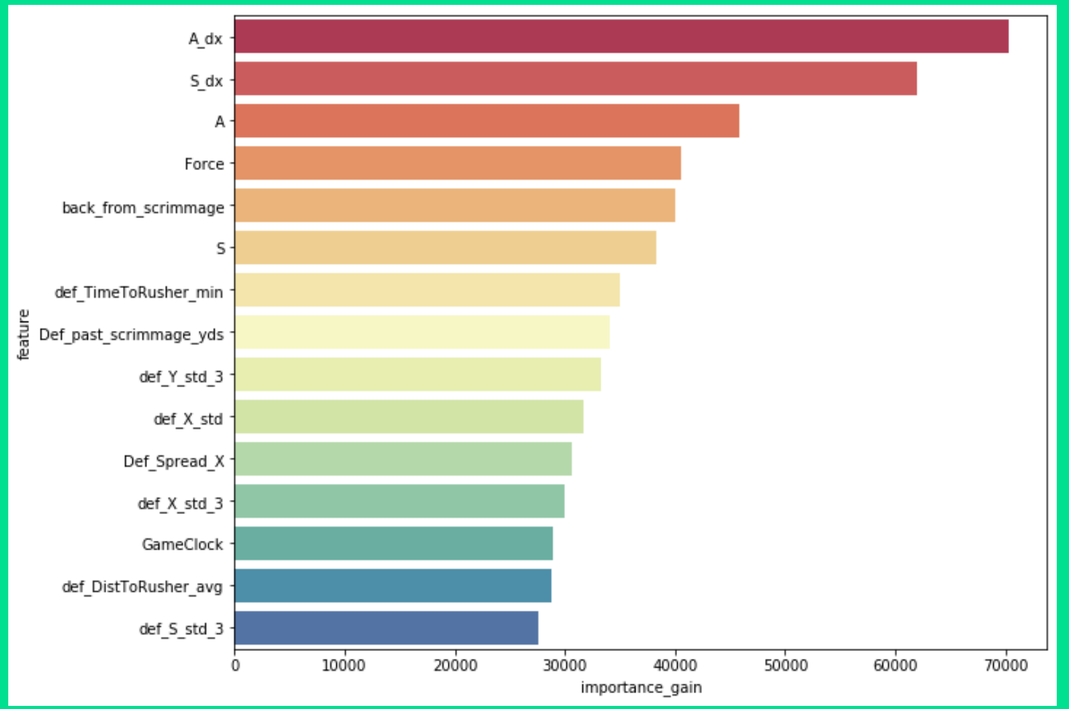
MAE
2.62 yards

Rusher Models: Feature Importance

RANDOM FOREST



LIGHTGBM



- **A:** acceleration of rusher (**dx** = x component downfield)
- **S:** speed of rusher (**dx** = x component downfield)
- **FORCE:** rusher weight multiply with acceleration
- **BACK FROM SCRIMMAGE:** yards rusher is back from scrimmage
- **DEF_TIME TO RUSHER MIN:** smallest time for defender to reach rusher

KEY FINDINGS

- Next Gen data dominate importances
- Rusher data is most important.
- Defense-based data is 2nd in importances.

FINAL RESULTS

MODEL TYPE	CRPS	MAE
Median Benchmark	0.01845	3.67
Raw Data Models	0.01382	2.75
Rusher-Focused Models	0.01301	2.59

CONCLUSION

- Able to create a model to predict most common rush situations.
- Interaction of the rusher with the defense is key to defining a successful run.
- Football is very complex.
- Next Gen Stats data is very valuable.

QUESTIONS?