NFL BIG DATA BOWL: TACKLING BIG DATA

Teagan Milford, Rachana Aithal, Sanjhana Rangaraj, Yolanda Ojeda

17.5M

Approximate number of spectators in each game

*Statistic from Front Office Sports

“ American football is a complex sport, but once an offensive player receives a handoff or catches a pass, all 11 defenders focus on one task -- tackle that ball carrier as soon as possible. Conversely, the ball carrier's role is to advance the ball down the field to gain as much yardage as possible until he is tackled, scores, or runs out of bounds. “

*Excerpt from NFL Big Data Bowl 2024 by The National Football League Analytics Competition on Kaggle

AGENDA

01 THE DATA

02 EDA

03 FEATURE ENGINEERING

04 MODELING

05 CONCLUSION

01.

THE DATA

PLAYER DATA

7 columns x 1684 rows

TRACKING DATA

17 columns x 1048576+ rows

TACKLES DATA

7 columns x 17427 rows

GAME DATA

9 columns x 137 rows

PLAYS DATA

35 columns x 12487 rows

Game data

- `gameId` : Game identifier, unique (numeric)
- `season` : Season of game
- `week` : Week of game
- `gameDate` : Game Date (time, mm/dd/yyyy)
- `gameTimeEastern` : Start time of game (time, HH:MM:SS, EST)
- `homeTeamAbbr` : Home team three-letter code (text)
- `visitorTeamAbbr` : Visiting team three-letter code (text)
- `homeFinalScore` : The total amount of points scored by the home team in the game (numeric)
- `visitorFinalScore` : The total amount of points scored by the visiting team in the game (numeric)

GAMES DATA

PLAYERS DATA

Player data

- `nflId` : Player identification number, unique across players (numeric)
- `height` : Player height (text)
- `weight` : Player weight (numeric)
- `birthDate` : Date of birth (YYYY-MM-DD)
- `collegeName` : Player college (text)
- `position` : Official player position (text)
- `displayName` : Player name (text)

PLAYS DATA

Play data

- `gameId` : Game identifier, unique (numeric)
- `playId` : Play identifier, not unique across games (numeric)
- `ballCarrierId` : The `nflId` of the ball carrier (receiver of the handoff, receiver of pass or the QB scrambling) on the play. This is the player that the defense is attempting to tackle. (numeric)
- `ballCarrierName` : The `displayName` of the ball carrier on the play (text)
- `playDescription` : Description of play (text)
- `quarter` : Game quarter (numeric)
- `down` : Down (numeric)
- `yardsToGo` : Distance needed for a first down (numeric)
- `possessionTeam` : Team abbr of team on offense with possession of ball (text)
- `defensiveTeam` : Team abbr of team on defense (text)
- `yardlineSide` : 3-letter team code corresponding to line-of-scrimmage (text)
- `yardlineNumber` : Yard line at line-of-scrimmage (numeric)
- `gameClock` : Time on clock of play (MM:SS)
- `preSnapHomeScore` : Home score prior to the play (numeric)
- `preSnapVisitorScore` : Visiting team score prior to the play (numeric)
- `passResult` : Dropback outcome of the play (`C` : Complete pass, `I` : Incomplete pass, `S` : Quarterback sack, `IN` : Intercepted pass, `R` : Scramble, text)
- `passLength` : The distance beyond the LOS that the ball traveled not including yards into the endzone. If thrown behind LOS, the value is negative. (numeric)

- `penaltyYards` : yards gained by offense by penalty (numeric)
- `prePenaltyPlayResult` : Net yards gained by the offense, before penalty yardage (numeric)
- `playResult` : Net yards gained by the offense, including penalty yardage (numeric)
- `playNullifiedByPenalty` : Whether or not an accepted penalty on the play cancels the play outcome. Y stands for yes and N stands for no. (text)
- `absoluteYardlineNumber` : Distance from end zone for possession team (numeric)
- `offenseFormation` : Formation used by possession team (text)
- `defendersInTheBox` : Number of defenders in close proximity to line-of-scrimmage (numeric)
- `passProbability` : NGS probability of next play being pass (as opposed to rush) based off model without tracking data inputs (numeric)
- `preSnapHomeTeamWinProbability` : The win probability of the home team before the play (numeric)
- `preSnapVisitorTeamWinProbability` : The win probability of the visiting team before the play (numeric)
- `homeTeamWinProbabilityAdded` : Win probability delta for home team (numeric)
- `visitorTeamWinProbabilityAdded` : Win probability delta for visitor team (numeric)
- `expectedPoints` : Expected points on this play (numeric)
- `expectedPointsAdded` : Delta of expected points on this play (numeric)
- `foulName[i]` : Name of the i-th penalty committed during the play. i ranges between 1 and 2 (text)
- `foulNFLId[i]` : `nflId` of the player who comitted the i-th penalty during the play. i ranges between 1 and 2 (numeric)

TACKLES DATA

Tackles data

- `gameId` : Game identifier, unique (numeric)
- `playId` : Play identifier, not unique across games (numeric)
- `nflId` : Player identification number, unique across players (numeric)
- `tackle` : Indicator for whether the given player made a tackle on the play (binary)
- `assist` : Indicator for whether the given player made an assist tackle on the play (binary)
- `forcedFumble` : Indicator for whether the given player forced a fumble on the play (binary)
- `pff_missedTackle` : Provided by Pro Football Focus (PFF). Indicator for whether the given player missed a tackle on the play (binary)

TRACKING DATA

Tracking data

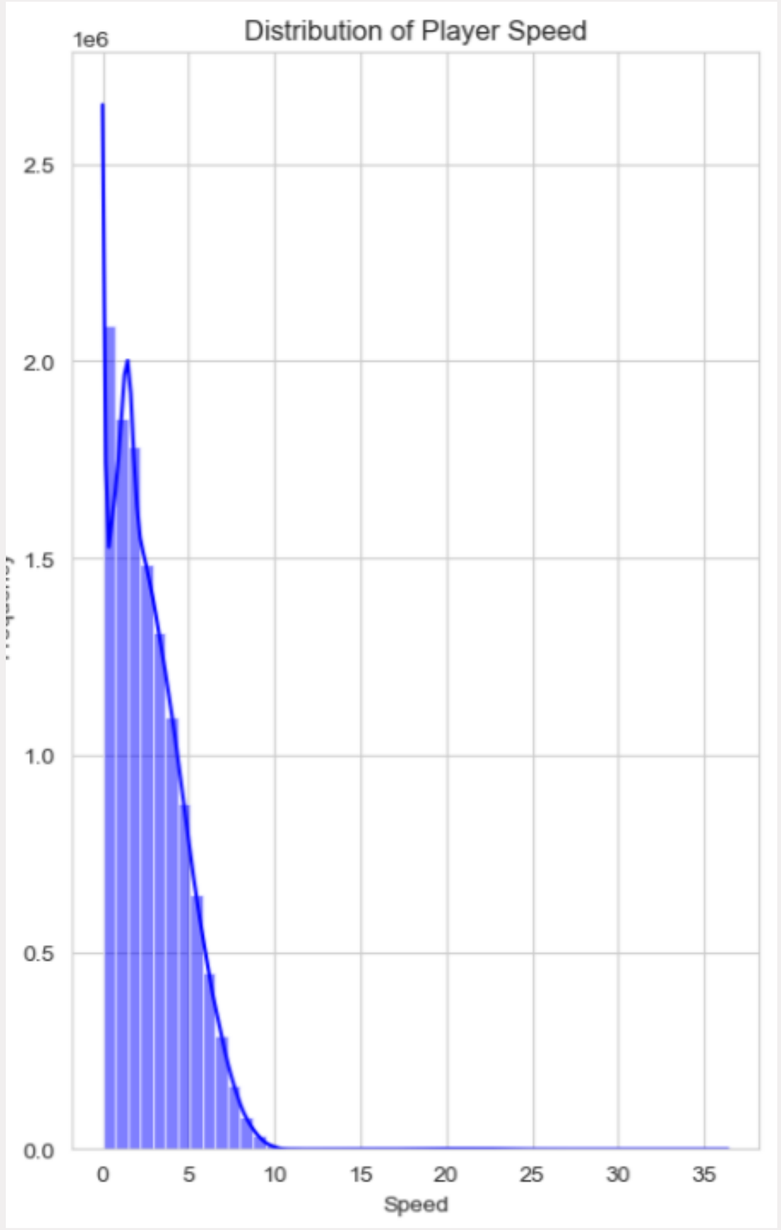
Files `tracking_week_[week].csv` contains player tracking data from week `[week]`.

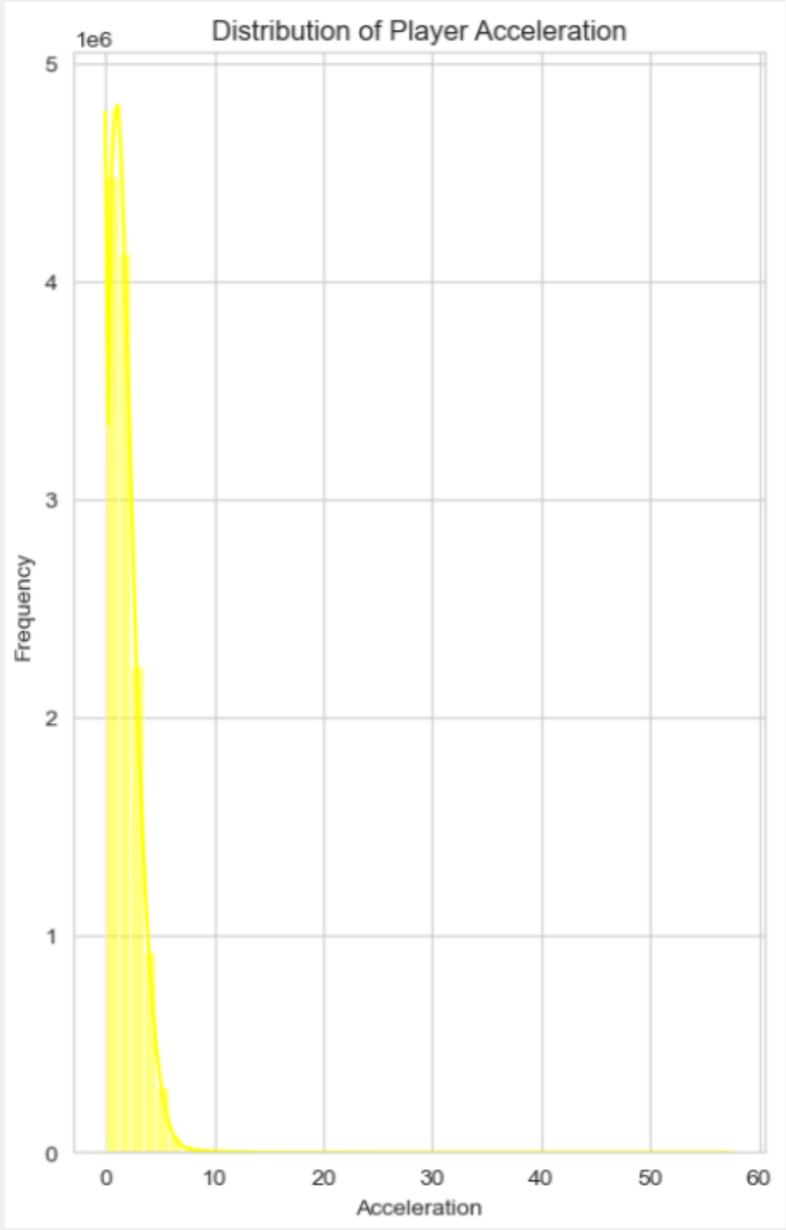
- `gameId` : Game identifier, unique (numeric)
- `playId` : Play identifier, not unique across games (numeric)
- `nflId` : Player identification number, unique across players. When value is NA, row corresponds to ball. (numeric)
- `displayName` : Player name (text)
- `frameId` : Frame identifier for each play, starting at 1 (numeric)
- `time` : Time stamp of play (time, yyyy-mm-dd, hh:mm:ss)
- `jerseyNumber` : Jersey number of player (numeric)
- `club` : Team abbrevation of corresponding player (text)
- `playDirection` : Direction that the offense is moving (left or right)
- `x` : Player position along the long axis of the field, 0 - 120 yards. See Figure 1 below. (numeric)
- `y` : Player position along the short axis of the field, 0 - 53.3 yards. See Figure 1 below. (numeric)
- `s` : Speed in yards/second (numeric)
- `a` : Speed in yards/second^2 (numeric)
- `dis` : Distance traveled from prior time point, in yards (numeric)
- `o` : Player orientation (deg), 0 - 360 degrees (numeric)
- `dir` : Angle of player motion (deg), 0 - 360 degrees (numeric)
- `event` : Tagged play details, including moment of ball snap, pass release, pass catch, tackle, etc (text)

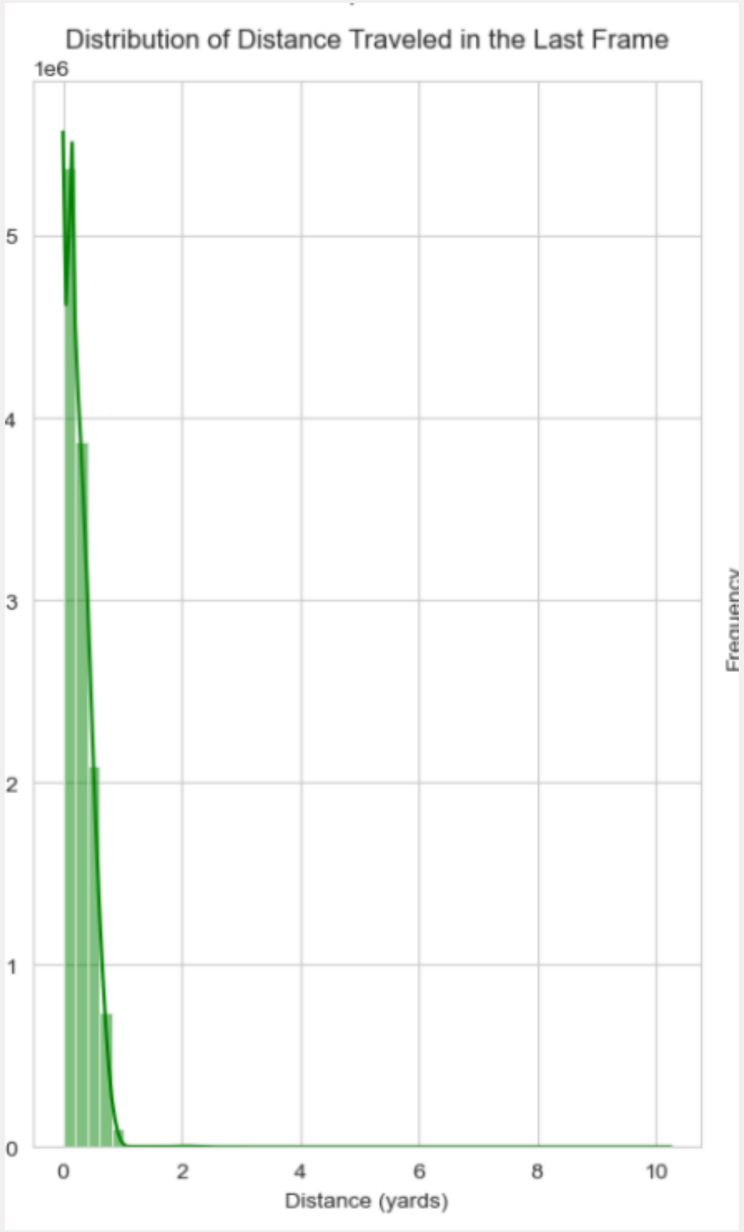
02.

EXPLORATORY DATA ANALYSIS

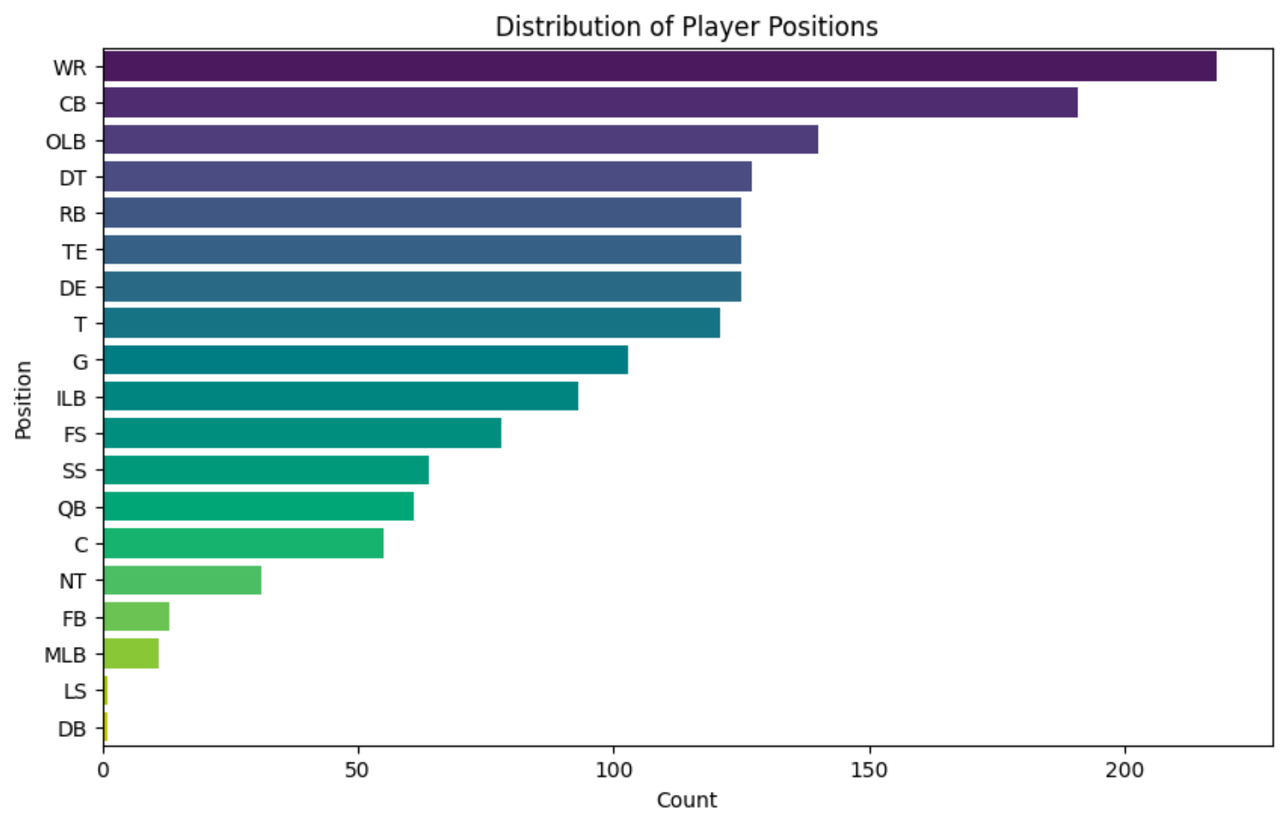
SPEED, ACCELERATION, AND DISTANCE TRAVELED

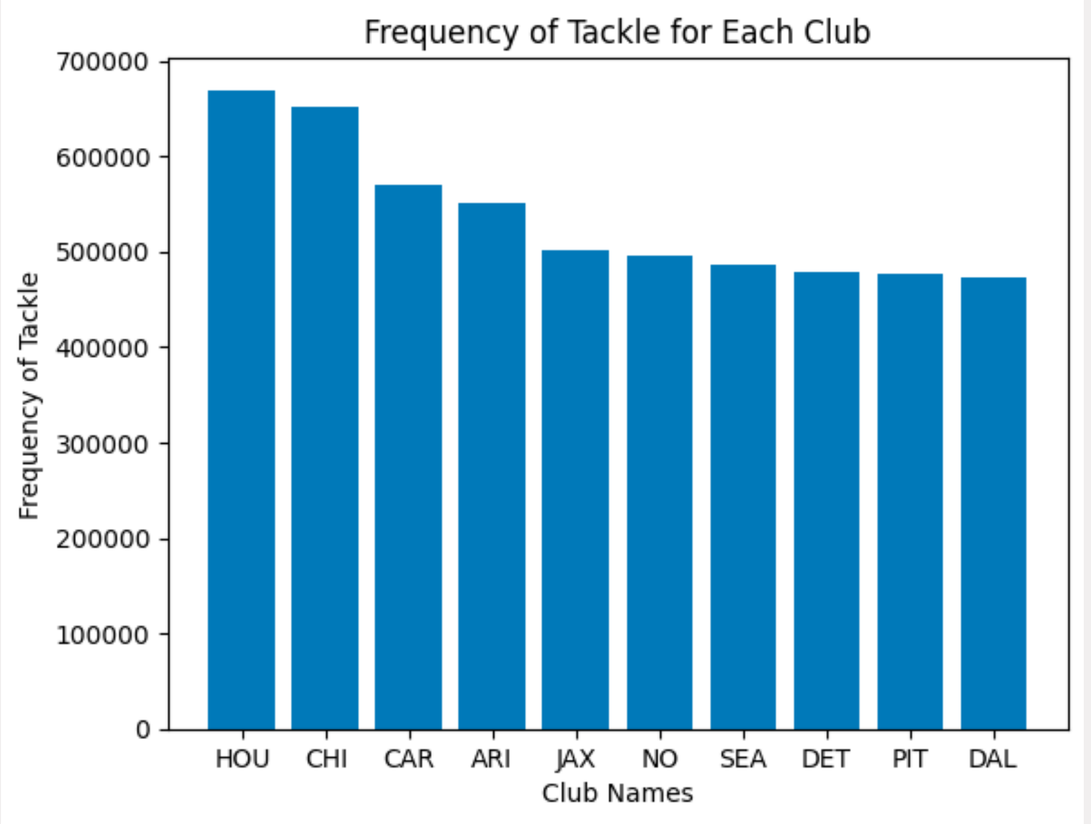




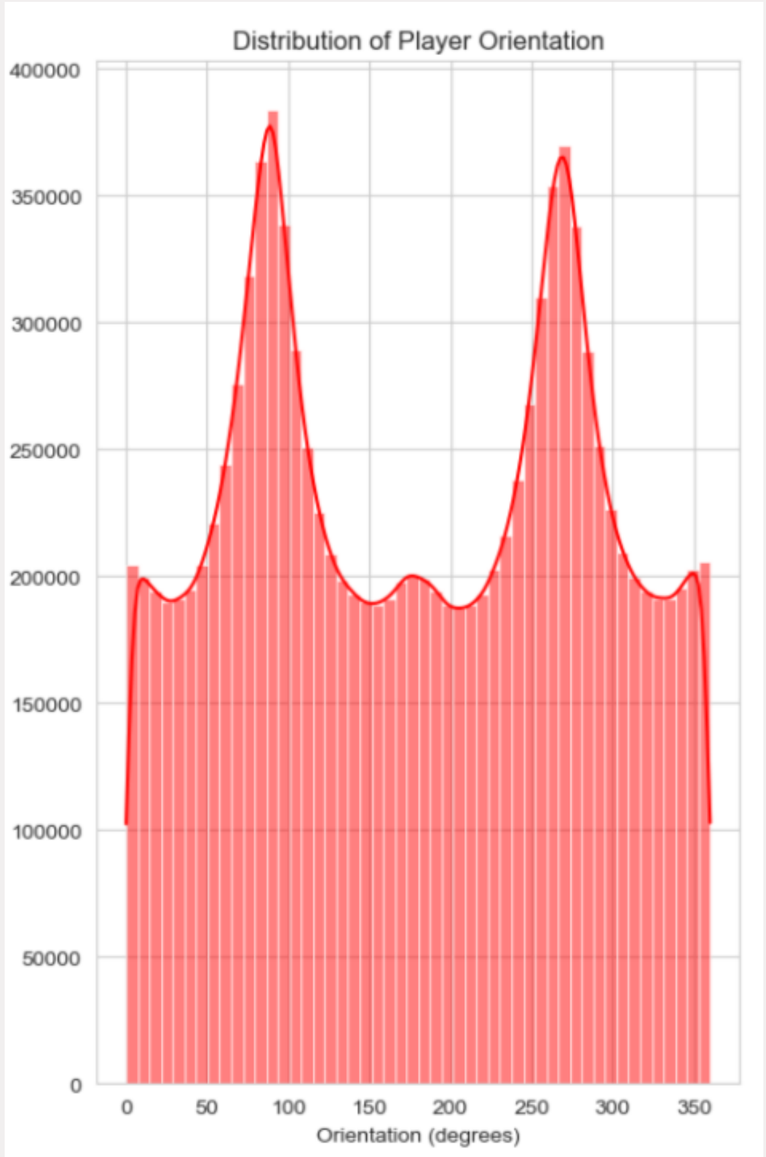


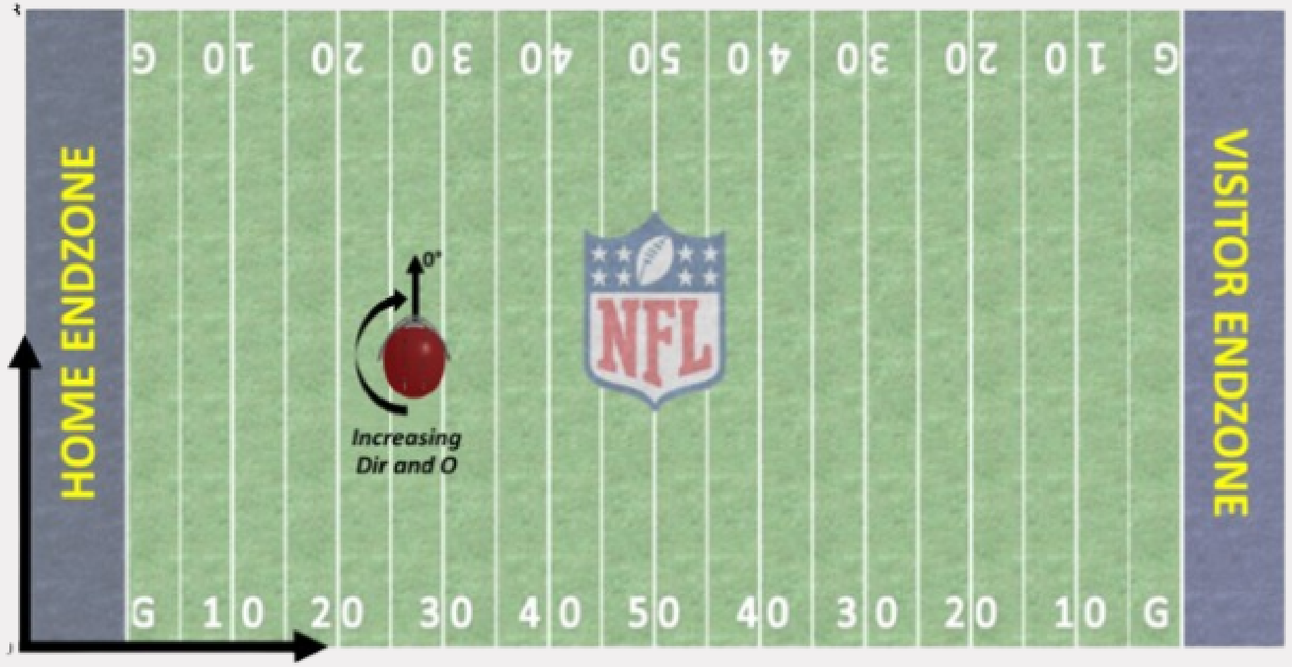
PLAYER POSITIONS &
FREQUENCY OF TACKLES BY CLUB





PLAYER ORIENTATION

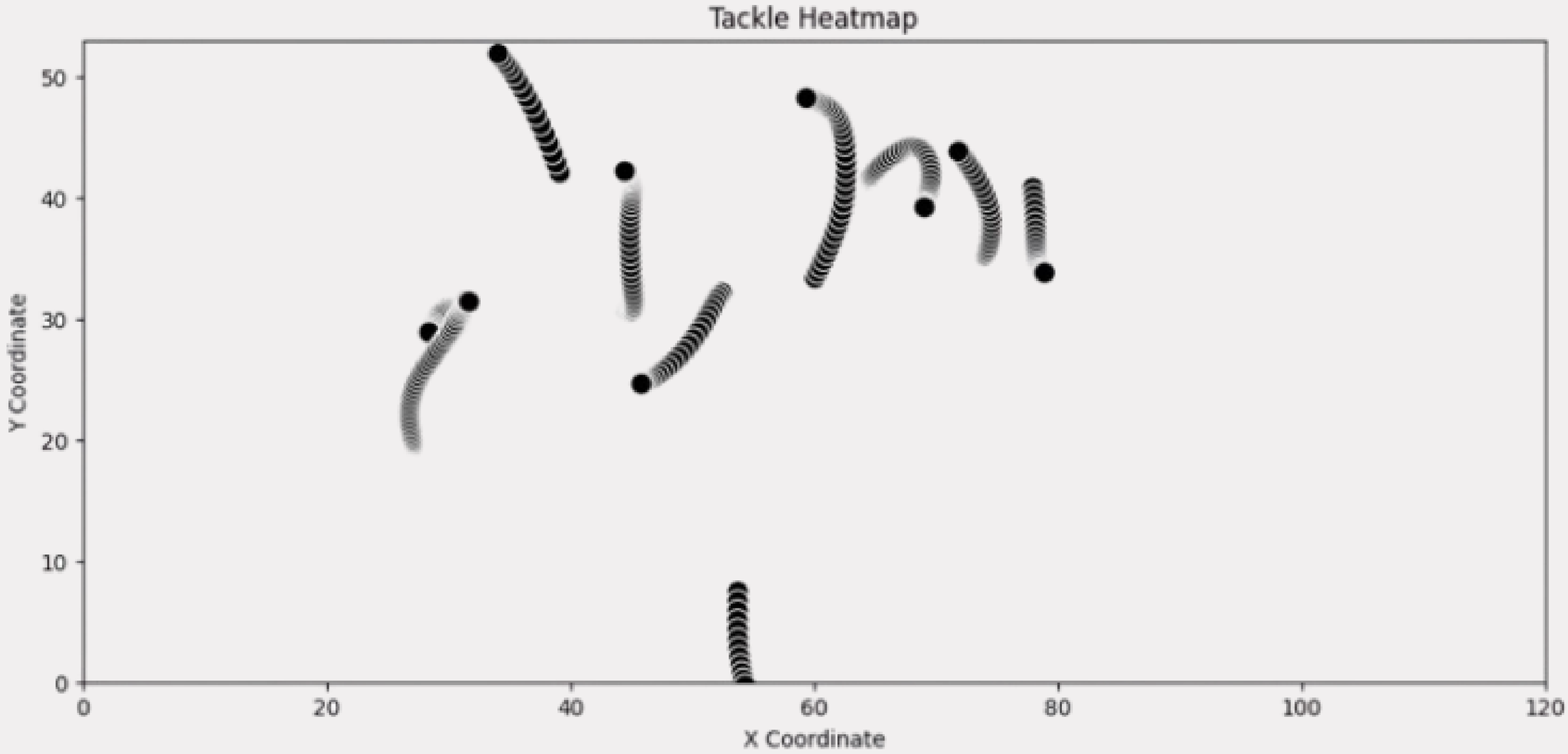




AVERAGE SPEED & ACCELERATION BY EVENT TYPE

	event	s	a
18	penalty_accepted	3.935000	2.213182
22	qb_slide	3.652147	1.941231
12	out_of_bounds	3.422297	1.710925
13	pass_arrived	3.359550	2.255809
15	pass_outcome_caught	3.340374	2.168609
9	lateral	3.300539	2.102525
16	pass_outcome_touchdown	3.281818	1.850909
4	first_contact	3.255991	1.965350
5	fumble	3.127554	1.773823
14	pass_forward	3.117968	2.033018
23	run	2.993274	2.052579
25	safety	2.938636	1.717576
29	touchdown	2.801443	1.534151
8	handoff	2.763073	1.751355
6	fumble_defense_recovered	2.726727	1.579182
17	pass_shovel	2.600833	1.938207
2	autoevent_passinterrupted	2.590189	2.020720
24	run_pass_option	2.581818	1.700000
21	qb_sack	2.540227	1.976955
28	tackle	2.490896	1.711009
20	play_action	2.368564	1.781694
1	autoevent_passforward	2.328994	1.773019
19	penalty_flag	2.314773	1.777955
7	fumble_offense_recovered	1.871212	1.708636
10	line_set	0.574489	0.664527
11	man_in_motion	0.485801	0.616232
26	shift	0.452058	0.654016
27	snap_direct	0.415673	0.596927
3	ball_snap	0.389294	0.671996
0	autoevent_ballsnap	0.368409	0.605227

53.3



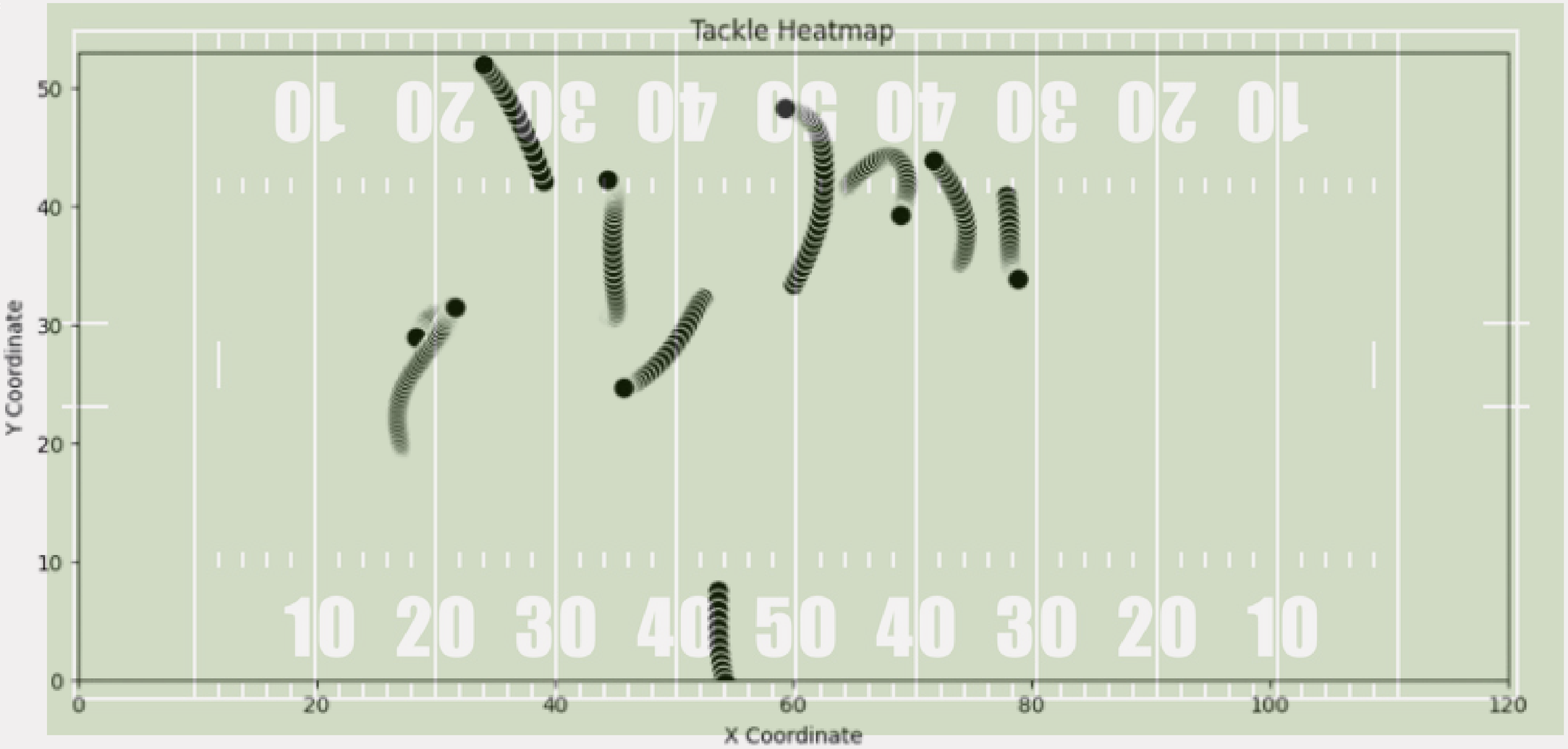
Increasing Y

0

Increasing X

120

53.3


Tackle Heatmap
Y Coordinate
X Coordinate
0
20
40
60
80
100
120
10
20
30
40
50
40
30
20
10
Increasing Y
0
Increasing X

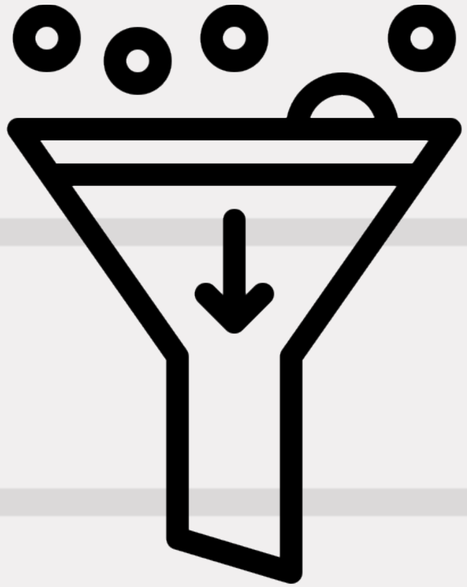
120

03.

FEATURE ENGINEERING

DATA REDUCTION

- Large dataset - 20,000,000 rows
- Computationally expensive
- Random sampling - 20,000,000 to 50,000 rows



ADDED FEATURES

ToRight

IsPlayLeftToRight

IsRusher

IsOnOffense

YardIndexClipped

YardIndex

player_minus_rusher_Sy

player_minus_rusher_Sx

player_minus_rusher_y

player_minus_rusher_x

Y_std

X_std

Sx

Sy

Dir_rad

Dir_std

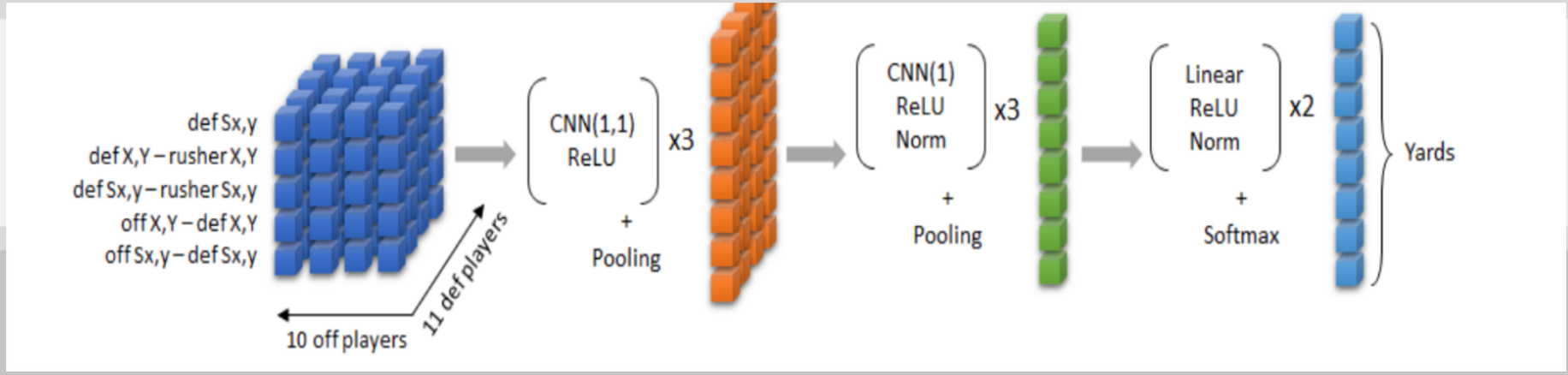
04.
MODELING

CONVOLUTIONAL NEURAL NETWORK

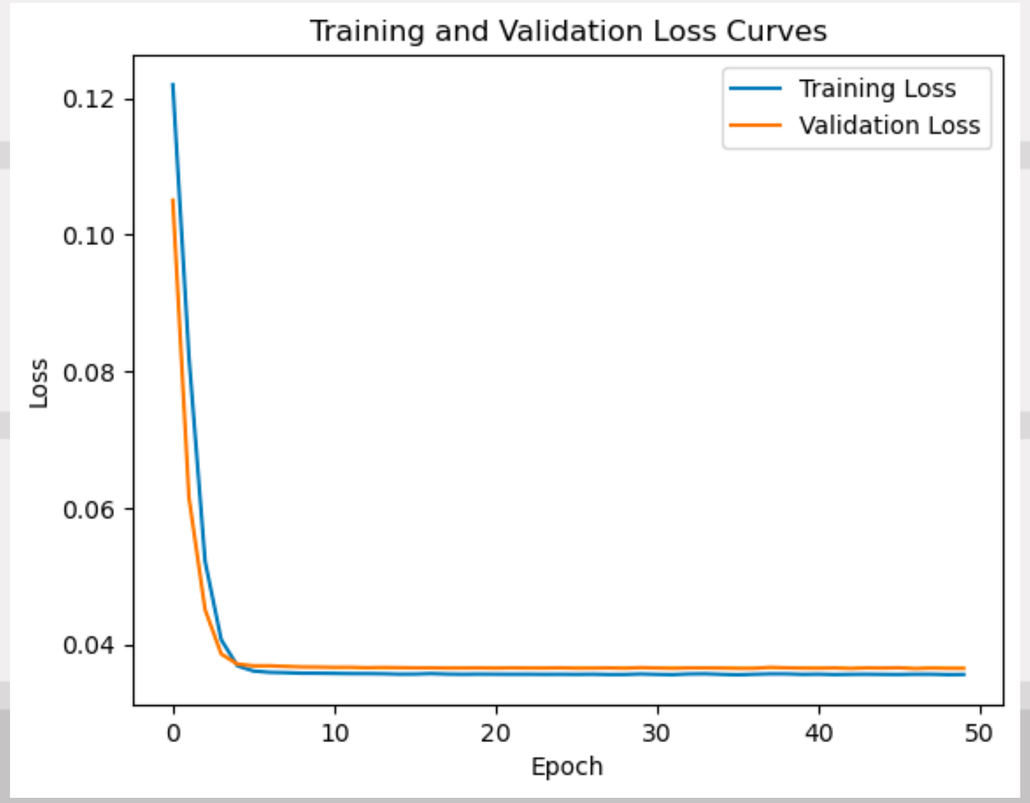
Predicting how many yards a team will gain on given rushing plays as they happen

A rushing play consists on:

- A rusher, whose aim is to run forward as far as possible
- 11 defense players who are trying to stop the rusher
- 10 remaining offense players trying to prevent defenders from blocking or tackling the rusher



CONVOLUTIONAL NEURAL NETWORK



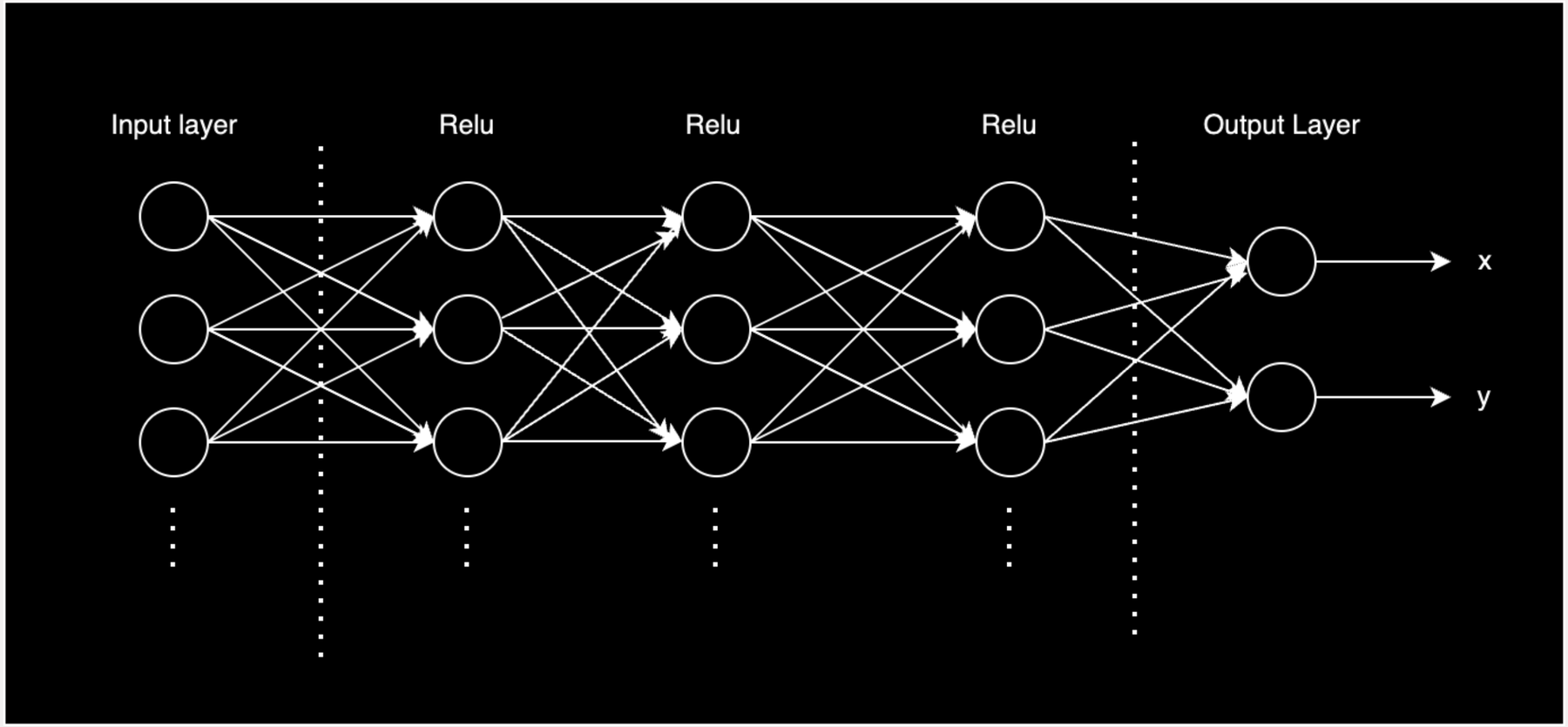
The mean validation loss is 0.0357

LINEAR REGRESSION

- Predict the x,y coordinates where the tackle occurs
- Baseline model

```
Mean Squared Error for y1: 545.3401323800032
Mean Squared Error for y2: 96.49115241388174
```

NEURAL NETWORK



NEURAL NETWORK (KERAS)

Model to predict the x and the y coordinates of the tackles
Better MSE than the baseline

```
Epoch 98/100
20000/20000 [==============================] - 31s 2ms/step - loss: 283.2621 - mean_squared_error: 283.2621 - val_loss: 284.9350 - val_mean_squared_error: 284.9350
Epoch 99/100
20000/20000 [==============================] - 32s 2ms/step - loss: 283.0715 - mean_squared_error: 283.0715 - val_loss: 283.9202 - val_mean_squared_error: 283.9202
Epoch 100/100
20000/20000 [==============================] - 33s 2ms/step - loss: 283.2790 - mean_squared_error: 283.2790 - val_loss: 284.9969 - val_mean_squared_error: 284.9969
6250/6250 [==============================] - 7s 1ms/step - loss: 283.5266 - mean_squared_error: 283.5266
Mean Squared Error on Test Set: [283.5265808105469, 283.5265808105469]
```

```
Epoch 80/100
20000/20000 [==============================] - 32s 2ms/step - loss: 283.1519 - mean_squared_error: 283.1519 - val_loss: 282.4662 - val_mean_squared_error: 282.4662
```

LOGISTIC REGRESSION

- To predict whether or not a tackle occurs
- Baseline model
- High recall

Accuracy	Precision	Recall
0.58	0.60	0.90

DECISION TREE CLASSIFICATION

- To predict whether or not a tackle occurs
- Fairly high Precision and Recall
- Considerable improvement in accuracy over logistic regression

Accuracy	Precision	Recall
0.80	0.84	0.81

RANDOM FOREST CLASSIFICATION

- Used to estimate if a tackle occurs or not
- Number of trees, max depth found using CV
- Better accuracy than decision trees

Accuracy	Precision	Recall
0.88	0.83	1.0

ENSEMBLE METHODS

Adaboost

```
{'learning_rate': 0.01, 'n_estimators': 10, 'random_state': 42}
training time 0.914 s
predict time 0.04 s
Confusion matrix:
[[10707  2763]
 [    0 16326]]
Accuracy: 0.907269432138542, AUC_ROC: 0.8974387527839643
```

AdaBoostClassifier

```
AdaBoostClassifier(learning_rate=0.01, n_estimators=10, random_state=42)
```

Xgboost

```
{'learning_rate': 0.01, 'n_estimators': 10, 'random_state': 42}
training time 0.362 s
predict time 0.017 s
Confusion matrix:
[[10707  2763]
 [    0 16326]]
Accuracy: 0.907269432138542, AUC_ROC: 0.9360249513417278
```

XGBClassifier

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
```

ENSEMBLE METHODS

LightGBM: Light Gradient Boosting Machine. An open-source, distributed, high-performance gradient boosting framework based on decision tree algorithm.

- Overfitting on Small Datasets:
  - its strength lies in handling larger datasets.
- Hyperparameter Tuning:
  - Requires careful tuning, could be complex and time-consuming.

```
Training time for fold: 7.322 s
Training time for fold: 10.245 s
Training time for fold: 10.279 s
Training time for fold: 7.262 s
Training time for fold: 13.385 s
Confusion matrix:
[[10711  2759]
 [   13 16313]]
Accuracy: 0.9069673781715667, AUC_ROC: 0.9290104274806896
```

05.

CONCLUSION

THE BEST MODEL TO PREDICT THE TACKLE DIFFICULTY:

THE ENSEMBLE METHODS

RECOMMENDATIONS

- Based on the predictive models we developed, given the proper data with available computation power, we can recommend utilizing ensemble methods to assist football coaches in optimizing resource allocation. By integrating insights from the ensemble models, coaches can gain a deeper understanding of team dynamics and performance, leading to more effective strategic planning for tackling in future games.

- Neural networks can capture the interactions between various features (such as player positioning, ball movement, and team formations) and how these interactions influence the likelihood and location of a tackle.

- By implementing CNNs, it excels in analyzing spatial data. They effectively extract and learn spatial feature hierarchies, identifying key patterns like player formations, which are essential for accurately predicting yards gained in football.


THANK YOU!!!
CAMPBELL · WILLIAMS FIELD
TEXAS
52
TEXAS
1