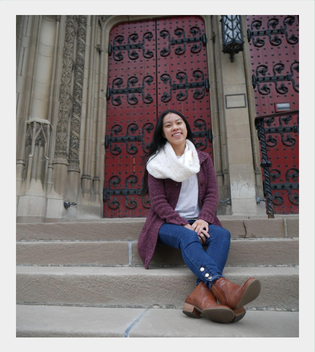
2023 NFL Big Data Bowl

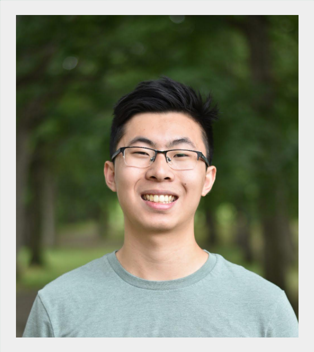
A Disturbance in the Force? Modeling QB Pressure with Force-based Metrics

About Us



Annie Liang

BS Business, Statistics



Harrison Lian

BS Statistics + Machine Learning

Jingchun Quan

BS Computational Finance, Statistics + Machine Learning

Joong Ho Choi

BS Statistics + Machine Learning

Table of Contents

What is the Big Data Bowl?

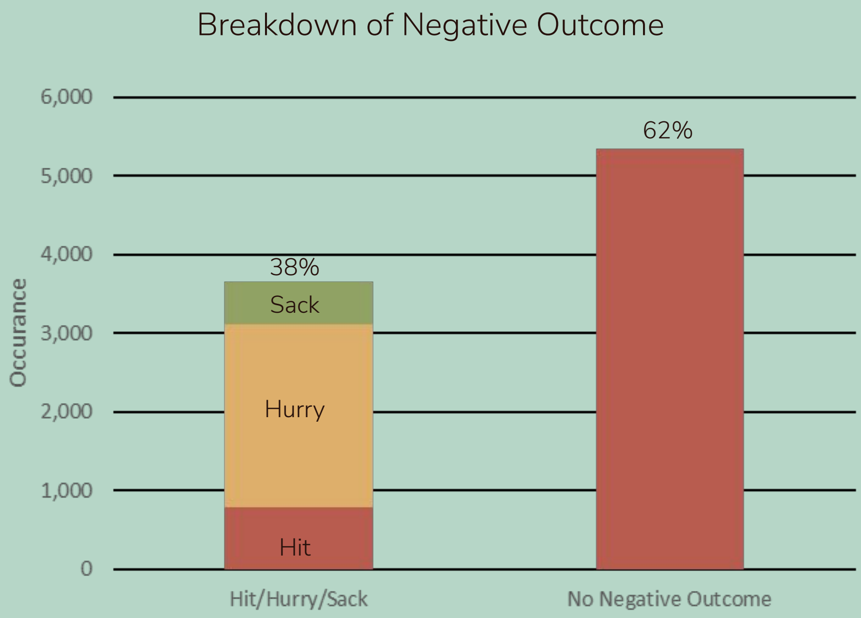
- The premier sports analytics competition hosted by the NFL on Kaggle using player-tracking data with a competition theme
  - This year's theme is linemen on passing plays

- Data for players and ball at each tenth of a second for each play in a game
  - Also includes data from Pro Football Scouting

Features

Goal: Predict whether a hit, hurry, or sack occurs on passing plays

- How does the position and acceleration of offensive and defensive linemen impact the outcome of interest?
- Sack and hit are relatively rare so they are combined to one negative outcome
- Interested in if a play has:
  - Hit
  - Hurry
  - Sack



2 categories of predictive features

Distance/Area

- Linemen distance to QB
- Distance between linemen
- Area formed by linemen

Forces

Distance Weighted

- Forces exerted by offense and defense linemen, weighted by inverse distance to QB

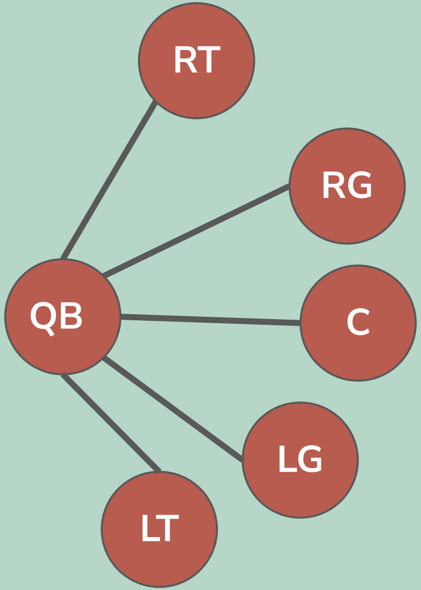
Partitioned

- Partitioning the field into three areas based on position and the forces exerted by linemen in each partition

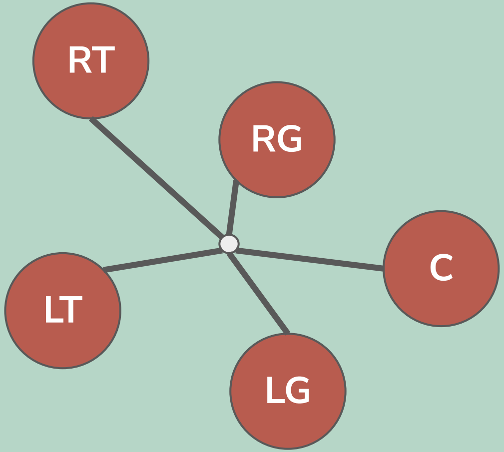
Distance / Area

Spatial relationship between the players

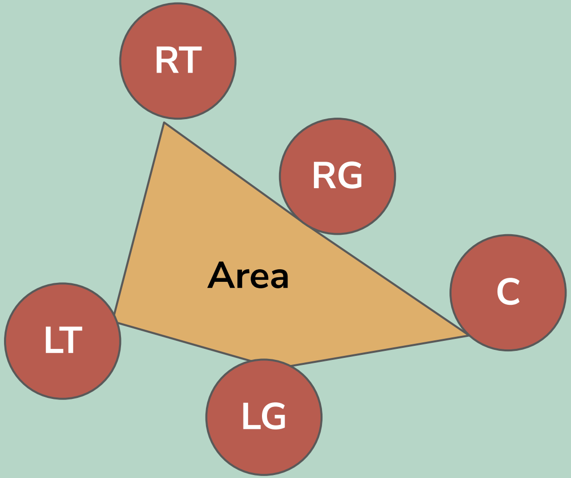
Average distance to the QB:



Average distance to the center point of the linemen:

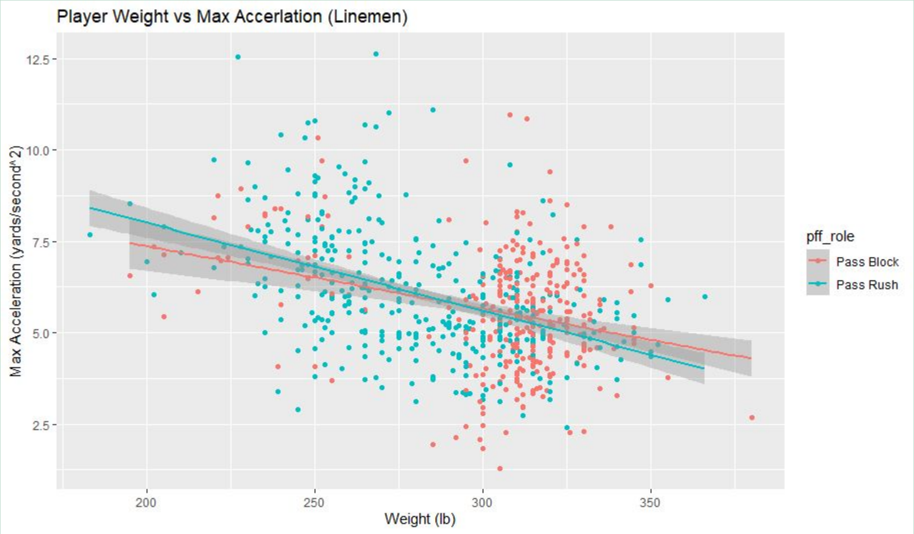


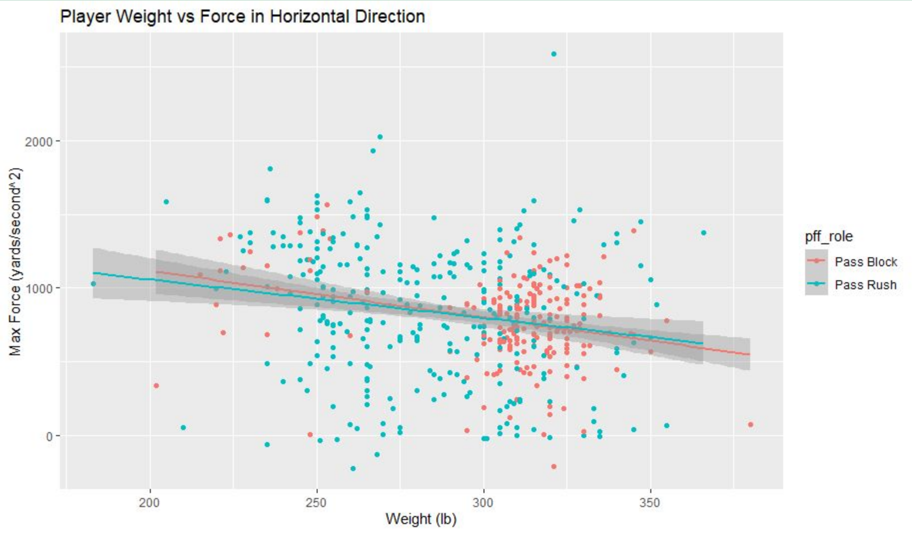
Area enclosed by linemen through shoelace method:



Forces

Force feature motivation:
Correlation between weight and acceleration



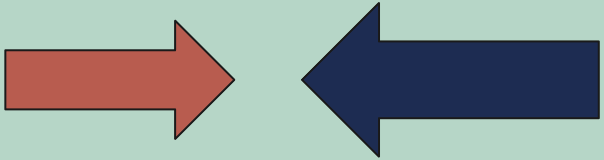


Hypothesis: If defense exerts greater force, higher chance of negative outcome for QB

1. Calculated force exerted by player.
2. Determined x and y forces exerted by direction for pass rushers and pass blockers.
3. Force exerted by was summed together to get net force.
    a. Net force > 0: offense exerted more force
    b. Net force < 0: defense exerted more force

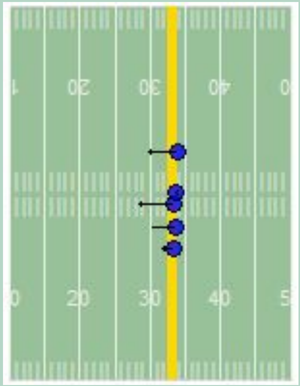
Net force exerted on field

QB



Pass Block

Pass Rush



T. Brady pass incomplete deep right to C. Godwin

Place more weight on the force of players closer to the QB

$$w_i = \frac{1/d_i}{\sum_{j=1}^{n_o} 1/d_i}$$

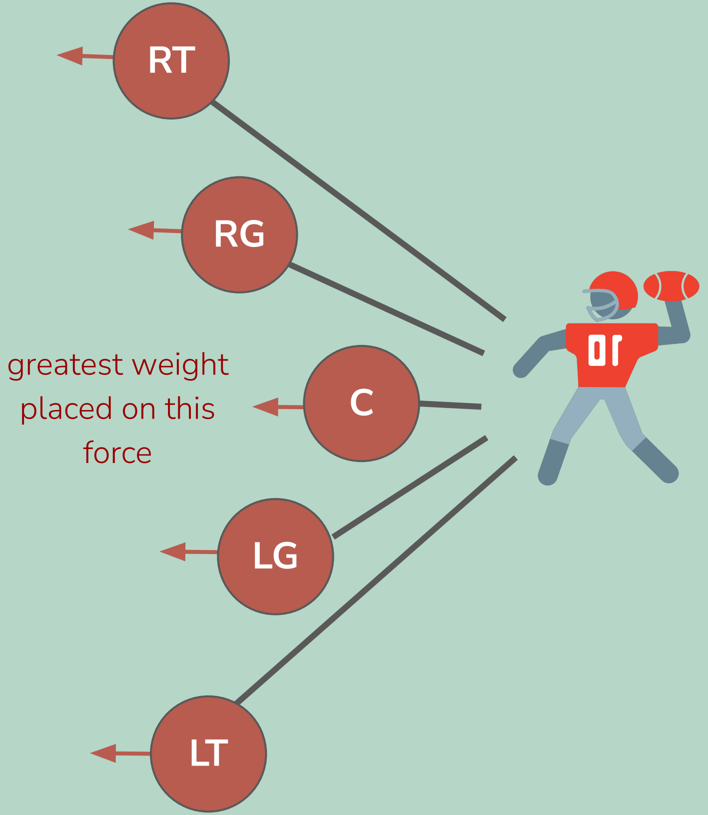
- n_o = number of pass blockers
- n_d = number of pass rushers
- d_i = distance of player i from QB (yards)
- w_i = weight of player i (0-1)
- F_i = force of player i

- Weighted force defense:

$$n_d \sum_{i=1}^{n_d} F_i w_i$$

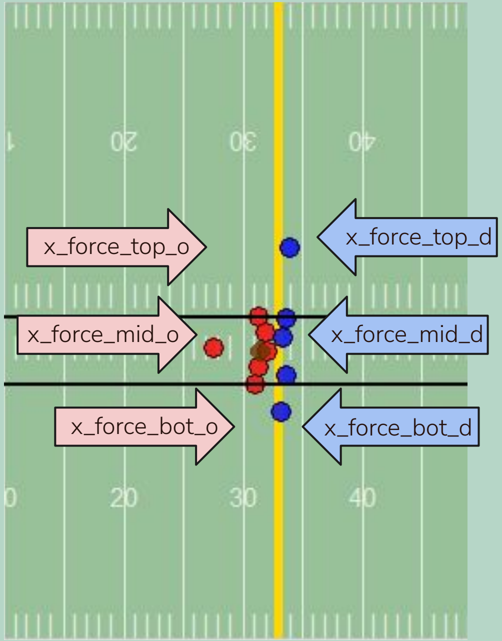
- Weighted force offense:

$$n_o \sum_{i=1}^{n_o} F_i w_i$$

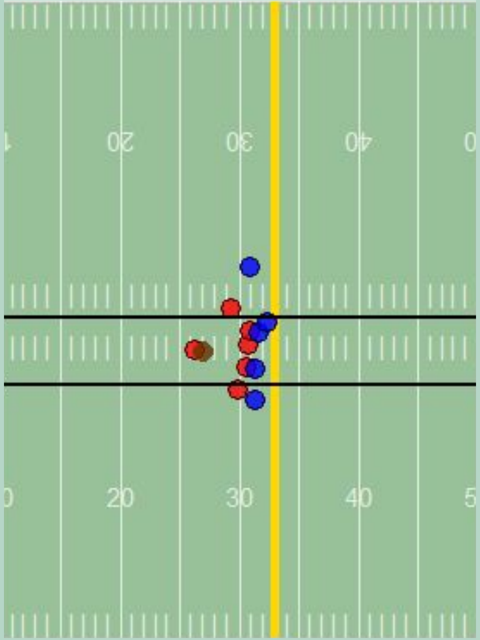


Partition field based on locations of guards at start of play

T. Brady pass incomplete deep right to C. Godwin



Frame 1



Frame 10

Six total partitioned force features

01
x_force_top_net

02
x_force_mid_net

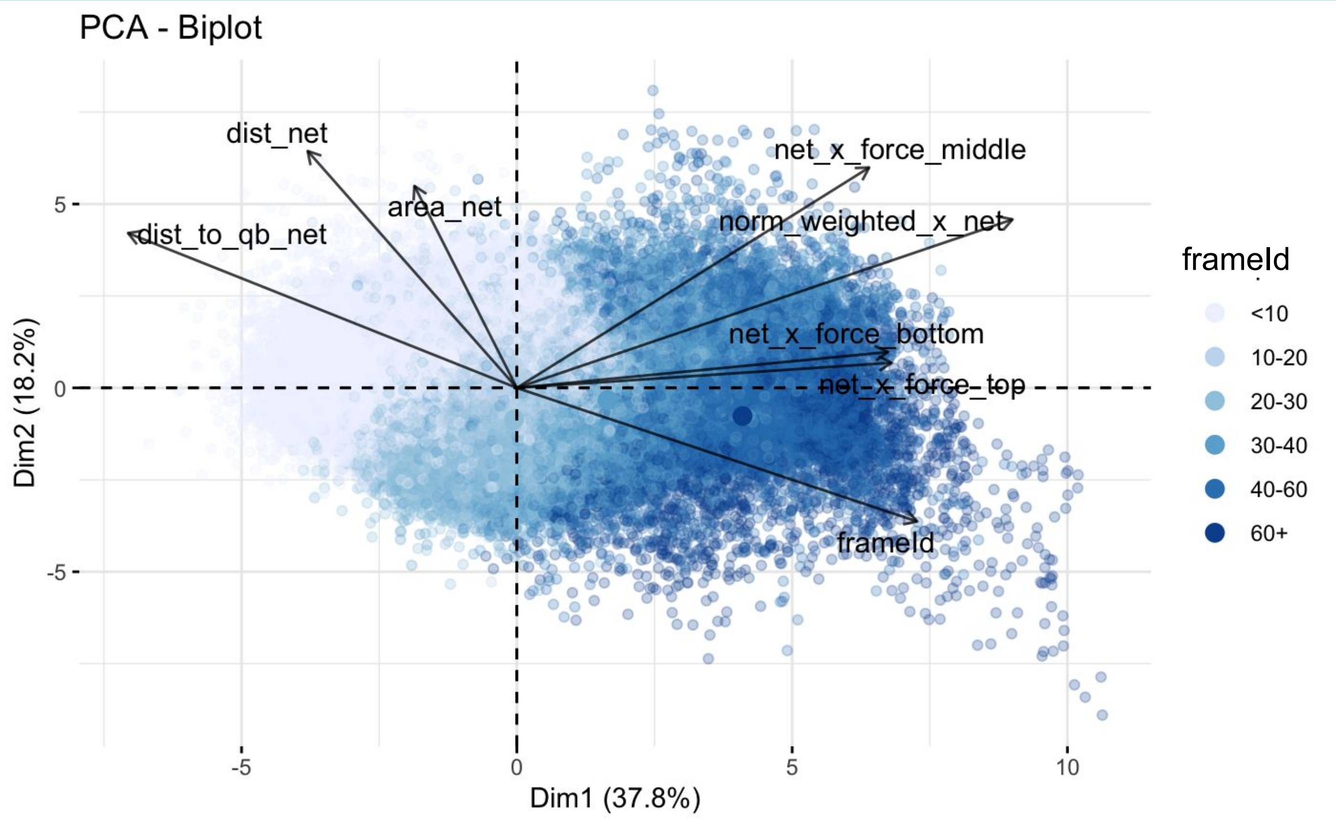
03
x_force_bot_net

04
y_force_top_net

05
y_force_mid_net

06
y_force_bot_net

Principal component analysis of all net features



Modeling

Trimming plays to start at snap and end at pass/negative outcome

FrameId	Event
1	None
:	:
5	None
6	Snap
7	None
:	:
37	None
38	Auto Event Pass Forward
39	None
40	Pass Forward
41	None
42	None
43	None

T. Brady pass incomplete deep right to C. Godwin

FrameId	Event
1	Snap
2	None
:	:
32	None
33	Auto Event Pass Forward

Two Modeling Techniques

1. Fit a model on the entire dataset using frameId + all engineered features as covariates
    a. XGBoost
    b. GLM
2. Fit separate models for each frameId
    a. FrameId is most predictive variable but not something that a coach can control
        i. We can take away the influence of frameId by having separate models trained for each frame
    b. Response is at the play level but our observations are at the frameId level
        i. If we fit model on all frame IDs, we will be adding unnecessary error terms

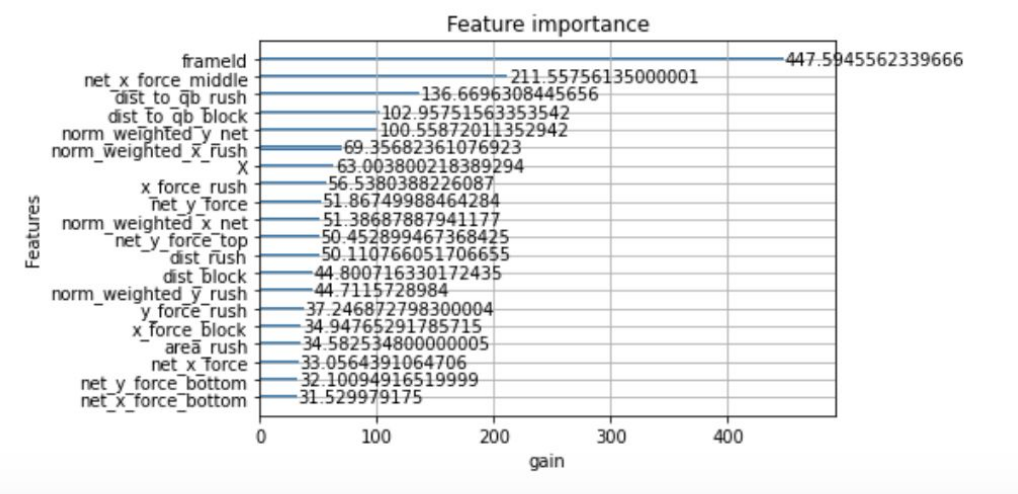
$$\sum_{i=1}^{n} div(y_i, \hat{y_i})$$

$n = \text{num\_games} \times \text{num\_plays} \times \text{num\_frames}$ OR $n = \text{num\_games} \times \text{num\_plays}$

30 40 50 40 30

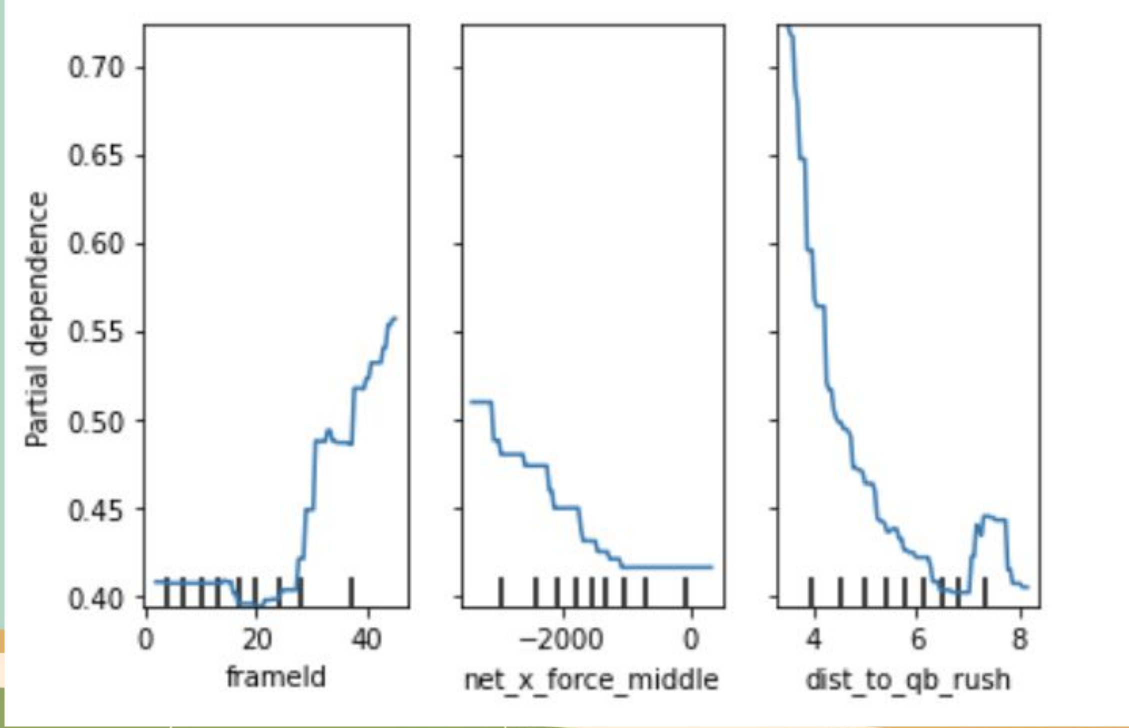
Baseline: XGBoost

- 1: negative outcome (hit, hurry and sack) happens
- 0: otherwise
- Trained on weeks 1-7 and tested on week 8



Precision	Recall
0.73	0.33

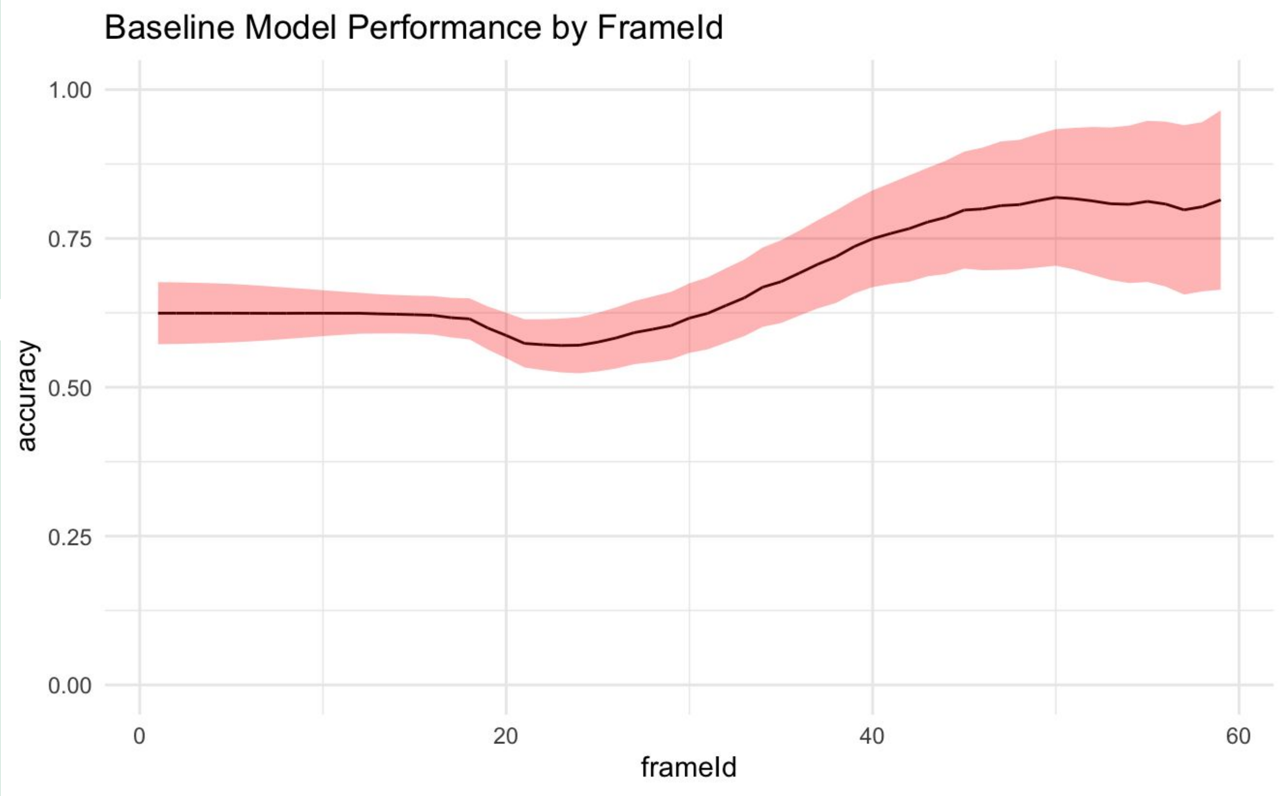
Directional impact from PDP of top 3 variables



Net force > 0: offense exerted more force

Net force < 0: defense exerted more force

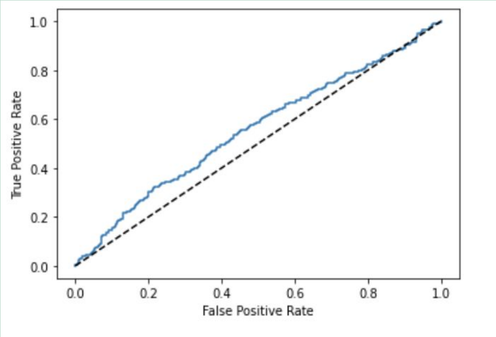
GLM Model accuracy improves as plays last longer

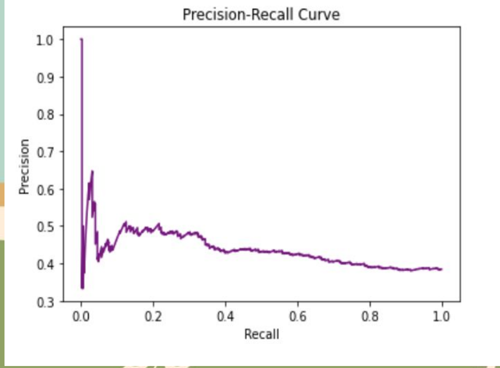


Snapshot Modeling

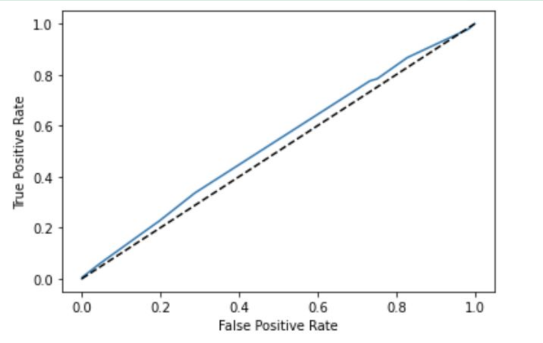
Logistic regression overfits in higher dimensions on small number of datapoints in later frames

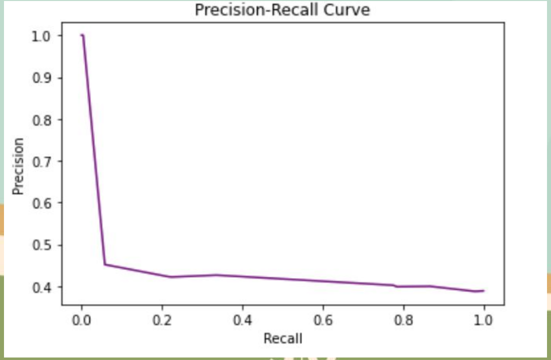
Frame 1



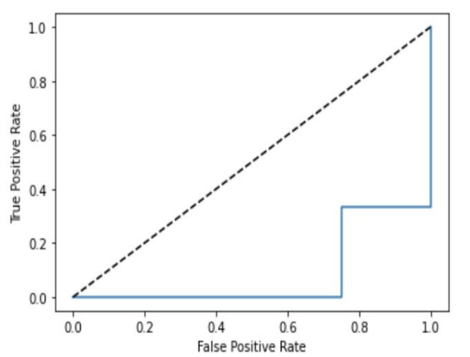


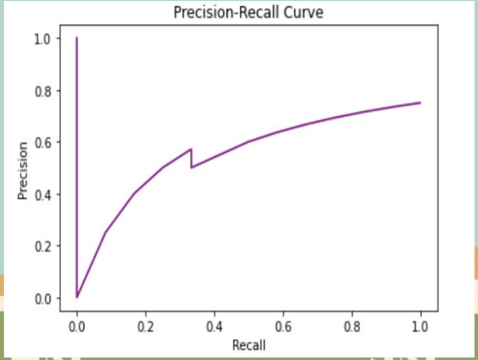
Frame 31





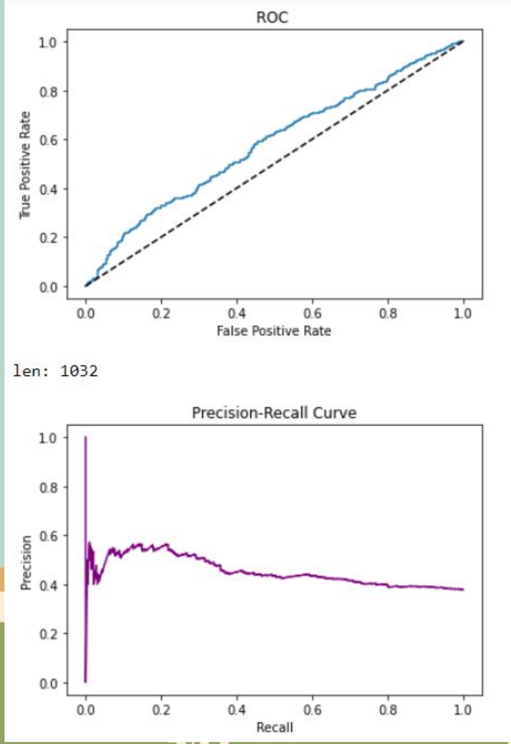
Frame 56



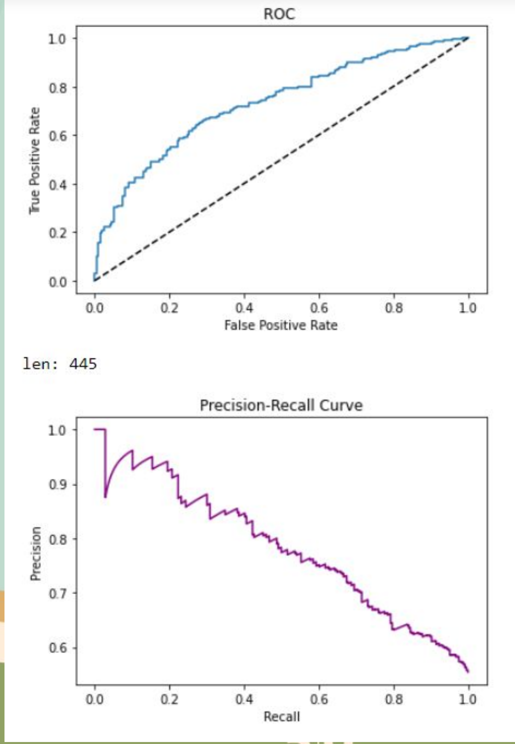


Random forest model improves performance, especially for intermediate frames

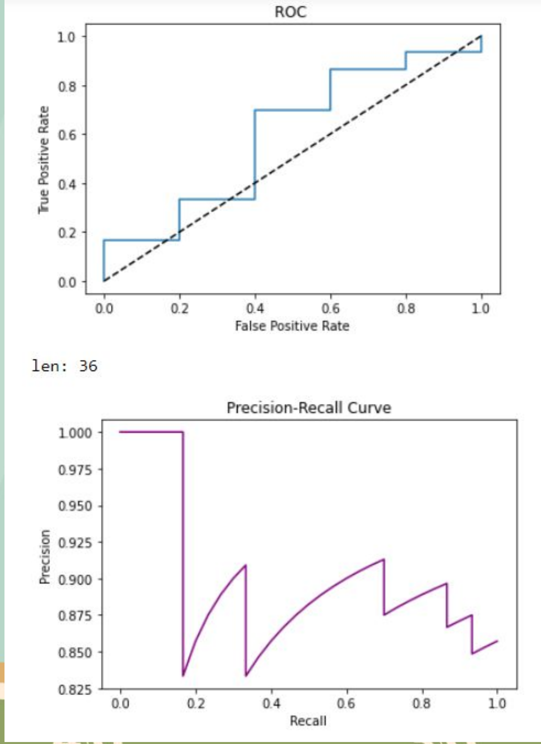
Frame 1



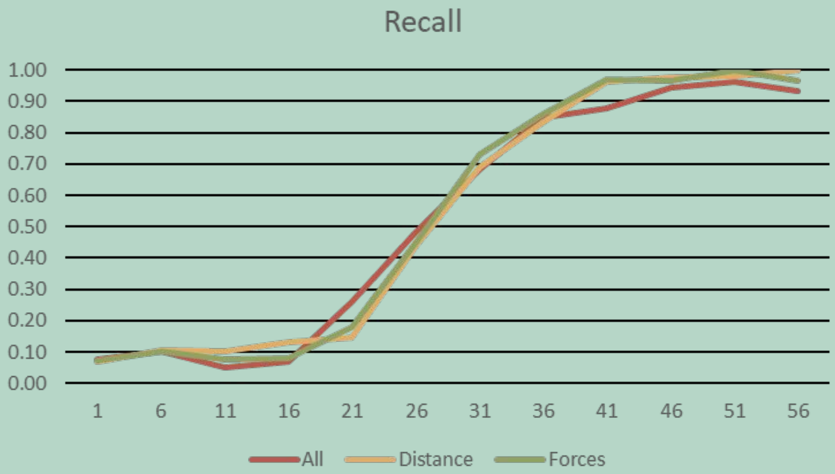
Frame 31



Frame 56



Accuracy and recall metrics for random forest model by frame



Discussion

Conclusions/Takeaways

1

Our models are currently heavily skewed by distance to QB which usually decreases as time passes, making the probability of a negative outcome higher

2

When distance features are dropped, similar performance is observed, which suggests that force features are just as good as distance features

3

Marginal increases in weight decrease a player's force due to the tradeoff between acceleration and weight

Next Steps

1. Model occurrence of bad outcome in next 10 frames instead of at end of entire play
2. Use a deep learning network such as an LSTM to fit spatial temporal data that tunes the number frames stored in memory
3. Explore models that capture autocorrelation between frames
4. Used hazard function to model how factors that players can control can make a play longer and give the QB more time
5. Add new distance feature between pass rusher to the closest blocker
6. Implement a self-updating ELO score between specific pass blocker vs. pass rusher

Questions?
Thank You