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

Annie Liang, Harrison Lian, Jingchun Quan, Joshua Choi
Project Advisor: Ron Yurko

Background & Introduction

Quarterbacks often get all of the attention, but a key to their success often lie in the linemen in front of them. The offensive line protects the QB, giving time for the team to carry out the play. At the same time, the opposing team's defensive line attempts to break through and hit, hurry, or sack the QB. We are interested in seeing how the position and force of offensive and defensive linemen lead to a negative outcome of a hit, hurry, or sack. We analyzed player, play, game, scouting, and tracking data from NFL and Pro Football Focus in order to create features that would be predictive of a negative outcome and reveal new insights into how football coaches can integrate this information into plays.

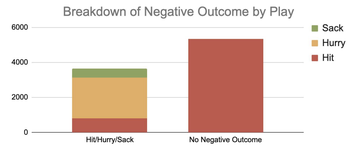


Figure 1: 38% of plays have a negative outcome, divided among hit, hurry, and sack

Overall Goal: How does the position and acceleration of offensive and defensive linemen in a play impact the outcome of the quarterback getting hit, hurried, or sacked?

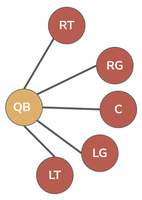
Figure 2: Negative correlation between weight and acceleration of pass blockers and rushers

Motivation behind using force features: as the weight of the players increases, there is an inverse relationship with the maximum acceleration that the player exerts, This relationship motivated us to look at force since it takes both acceleration and weight into account

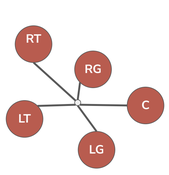
Feature Engineering

1. Distance/Area	2. Forces exerted by pass blockers and rushers		
	Net X & Y Force	Distance Weighted	Partitioned
• Linemen distance to QB • Distance between linemen • Area formed by linemen	• Forces exerted by offense and defense linemen	• Net forces weighted by inverse distance to QB	• Partitioning the field into three areas based on position and the forces exerted by linemen in each partition

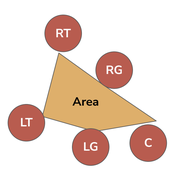
Average distance to the QB:



Average distance to the center point of the linemen:



Area enclosed by linemen through shoelace method:

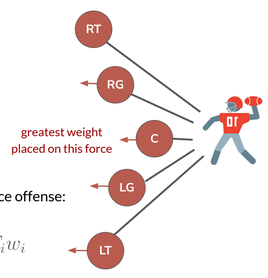


Distance weighted net forces

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

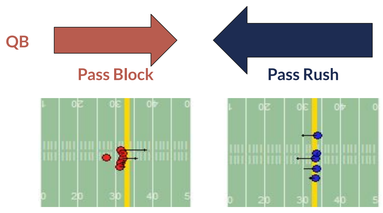
- n_o = number of pass blockers
- n_d = number of pass rushers
- d_i = distance of player i from QB (yards)
- w_i = weight of player i (0-1)
- F_i = force of player i

- Weighted force defense: $n_d \sum_{i=1}^{n_d} F_i w_i$
- Weighted force offense: $n_o \sum_{i=1}^{n_o} F_i w_i$



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

Net force exerted on field



T. Brady pass incomplete deep right to C. Godwin

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

Partitioned net forces

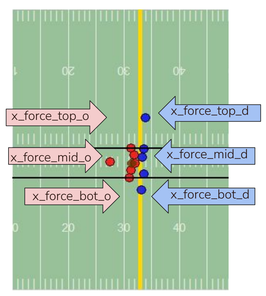


Figure 3: Net force features partitioned by guard locations at start of play

Analysis & Modeling

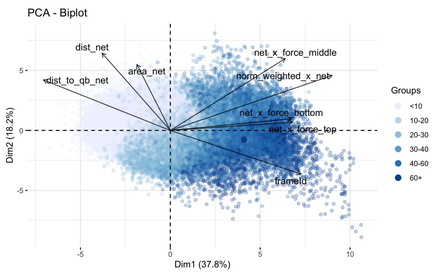


Figure 4: PCA analysis of engineered features

1. Fit XGBoost and GLM as baseline models on the entire dataset using frameId + all engineered features as covariates

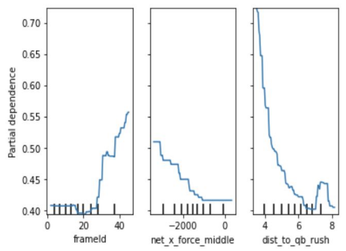


Figure 5: Partial dependence plots of top 3 variables from XGBoost

- Predictive accuracy of binary negative outcome variable ranges between 57% and 83%
- Accuracy generally increases with time elapsed in a play, along with the variance in predictions

- Force features and distance features are orthogonal, revealing different sources of variability in the data
- Dist_to_qb and frame_Id have strong negative correlation, since as plays go on, players tend to move closer to the QB, giving more opportunities to hit/hurry/sack

- Top 3 features in terms of information gain: frameId, net_x_force_middle and dist_to_qb_rush
- The likelihood of a negative outcome occurring rises as frame ID increases and defensive distance to quarterback decreases, confirming our hypothesis from PCA analysis
- As the net force in the x direction exerted in the middle of the field becomes more negative, the defensive linemen exerts more net force than the offensive, increasing the likelihood for a negative QB outcome

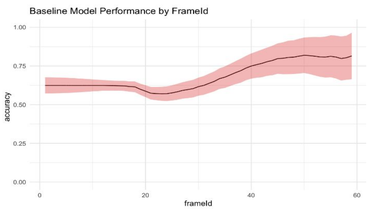


Figure 6: Relationship between GLM's accuracy and Frame ID

2. Fit separate models (logistic regression and random forest) for each frameId
   a. **Football intuition:** FrameId is most predictive variable but not something that a coach/player can control
      i. We can take away the influence of frame ID by having separate models trained for each frame to better observe other features
   b. **Statistical intuition:** 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

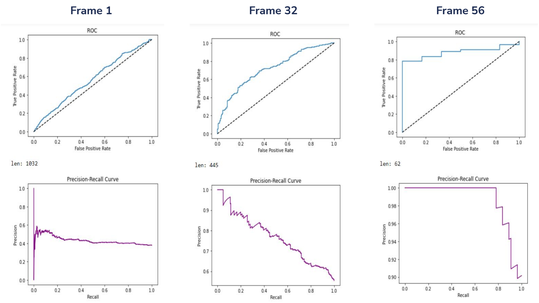


Figure 7: ROC and Precision-Recall Curves for selected frames from the random forest model

- Random forest performed better than logistic regression at nearly all frames due to high dimensionality of data, but both models improved in later frames
- Significant reduction in number of plays that last until the later frames (most plays were already hit, hurried, or sacked within 6 seconds or 60 frames)

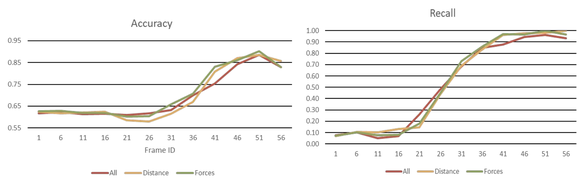


Figure 8: Tradeoff of accuracy and recall between models with all, distance, and just force features

Conclusion & Takeaways

We successfully modeled the likelihoods of the QB getting hit, hurried, or sacked using the tracking data of linemen, specifically focusing on the distance features as well as feature engineered net force calculations. We also made connections between player attributes, which coaches can control, to distance and time attributes, which tend to evolve over time. Throughout our modeling and analysis, we also observed:

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 gives us confidence in the predictive ability of the net force features
3. Marginal increases in weight decrease a player's force due to the uneven tradeoff between lower acceleration and higher weight

Next Steps

Figure 9 illustrates the relative importance of the current features used in our 2 modeling approaches. We note that many of our features are heavily correlated with each other (distance features and force features) as confirmed by the PCA. Besides exploring additional features to re trends on the spatial temporal data of linemen, other possible directions to expand our research include:

1. Model occurrence of bad outcome in next 10 frames instead of at end of entire play
2. Explore models that capture autocorrelation between frames
3. Used survival models and hazard functions to model how factors that players can control can make a play last longer and give the QB more time
4. Implement a self-updating ELO score between specific pass blocker vs. pass rusher that could attribute a score to a team of defensive or offensive linemen

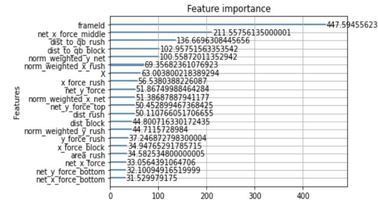


Figure 9: Features in order of significance from current modeling