Quarterbacks often get all of the attention, but a key to their success often lay in the linemen in front of them. At the beginning of a play, the offensive and defensive line up against each other. As soon as the play is in motion, the defensive line attempts to break through to press the quarterback while the offensive line works to stave them off. Over the course of the play, both teams are constantly trying to get closer to the quarterback, hit, hurt, or sacked. To do so, 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.

**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?

**Motivation behind using force features:** as the weight of a player increases, there is an inverse relationship with the maximum acceleration that the player can exert. Furthermore, we also see that pass blockers or the offensive linemen are generally heavier and have lower acceleration than pass rushers or the defensive linemen. This relationship motivated us to look at force as a predictive feature since it takes both acceleration and weight into account.

**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 was summed together to get net force
   a. Net force > 0, defense exerted more force
   b. Net force < 0, offense exerted more force

To evaluate how well a team’s defensive linemen worked together to inflict a bad outcome on the QB, we looked at the average maximum net force that the defense exerted along with the total number of hits, hurries, and sacks that were inflicted in the season. We found that defensive lines with a more negative max net force correlate to inflicting a greater number of bad outcomes, giving credit to the force feature approach. The dataset we used included plays from the 2021 season, with the Rams winning. The Rams can be seen at the far left, exerting the highest average max defensive force out of any team and inflicting >120 bad outcomes.

To conclude, our distance and force features provide an actionable technique for players & coaches to control occurrences of hits, hurries, and sacks. We made connections between measurable player-level attributes such as weight and force, which teams can assess, to distance and time attributes, which are result-based measurements that tend to evolve over time. Player-level insights can be helpful in evaluating how defensive linemen should be located on the field to break through and provide a warning to the offensive linemen about which defensive player could hit, hurry, or sack the QB.

**Feature Engineering**

<table>
<thead>
<tr>
<th>1. Distance/Area</th>
<th>2. Forces exerted by pass blockers and rushers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linemen distance to QB</td>
<td>Net X Y Force Distance Weighted Partitioned</td>
</tr>
<tr>
<td>Distance between linemen</td>
<td>Net forces exerted by offense/defense linemen</td>
</tr>
<tr>
<td>Area formed by linemen</td>
<td>Net forces weighted by inverse distance to QB</td>
</tr>
<tr>
<td>Forces exerted by offensive/defense linemen</td>
<td>Partitioning the field into three areas based on position and the forces exerted by linemen in each partition</td>
</tr>
</tbody>
</table>

**Analytical Model**

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

2. 2 fit separate models (logistic regression and random forest) for each framId
   a. **Football intuition:** FramId is most predictive variable but not something that a coach/player can control
      - 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 framId level
      - If we fit model on all frame IDs, we will be adding unnecessary error terms

**Conclusion & Takeaways**

- 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 move closer to the QB, giving more opportunities to hit/hurry/sack
- Top 3 features in terms of information gain: framId, net_x_force_middle and Dist_to QB rush
- The likelihood of a negative outcome occurring rises as frame_ID increases and distance to quarterback decreases, confirming our hypothesis from PCA analysis
- As the net force in the direction exerted in the middle of the field becomes more negative, the defensive linemen exerts more negative force than the offensive, increasing the likelihood for a negative QB outcome

**Next Steps**

There are several future improvements that could be made to our work:

1. **We grouped hits, hurries, and sacks together due to the class imbalance of each negative outcome.** Future work would focus on the impact of our force features on each specific play outcome. For instance, coaches and players are often more interested in the occurrence of sacks and how they can prevent or push for that outcome.
2. **Our work focuses more holistically on the force displayed by a team.** Future research can focus on the occurrence of a bad outcome given when certain players are matched together or face off
3. **We can explore models that better capture the autocorrelation between frames such as modeling the occurrence of a bad outcome in next 50 frames instead of at the end of the entire play