Optimal Strategy for Defensive Pressure in American Football: Insights from an Expected Sack Model

Tej Seth, W'23

The School of Information at the University of Michigan, Ann Arbor MI, USA

Abstract

The impact of a defender getting to the quarterback and recording a sack is a critical outcome in American football. However, limited research has been conducted on the positioning of defenders that could potentially increase their chance of success. In th is study, we utilize a combination of Pro Football Focus charting data and NFL Next Gen Stats tracking data to create the probability of a defender recording a sack based on their X and Y coordinate alignment, listed position, offensive and defensive forma tions and game situation information. The data analyzed includes weeks 1 through 8 of the 2021 NFL season, as provided by the league through the 2022 Big Data Bowl competition. An eXtreme Gradient Boosted Classifier was used to predict the probability of a defender documenting a sack on a given play. The findings reveal that distance from the quarterback is the most influential factor, with defenders in closer proximity to the opposition being more likely to penetrate the backfield. Additionally, a player's roster position label emerged as a significant variable, with cornerbacks generally having lower odds of recording sacks compared to outside linebackers. These results suggest that certain teams, such as the Cincinnati Bengals, may strategically position their defenders in more optimal spots for sacking the quarterback compared to teams like the Atlanta Falcons.

Introduction

Sacking the quarterback is one of the most impactful outcomes that can occur on a play from a defensive perspective (Eager, 2018). Furthermore, defenders are incentivized to achieve sacks due to the financial incentives provided by NFL teams through free a gency and contract extensions (Monson, 2021). To evaluate the different types of pass rusher outcomes, we can use Expected Points Added (EPA), which assesses how much closer the offense is to scoring after a play compared to their position before the play, based on expected points model (Baldwin, 2021).

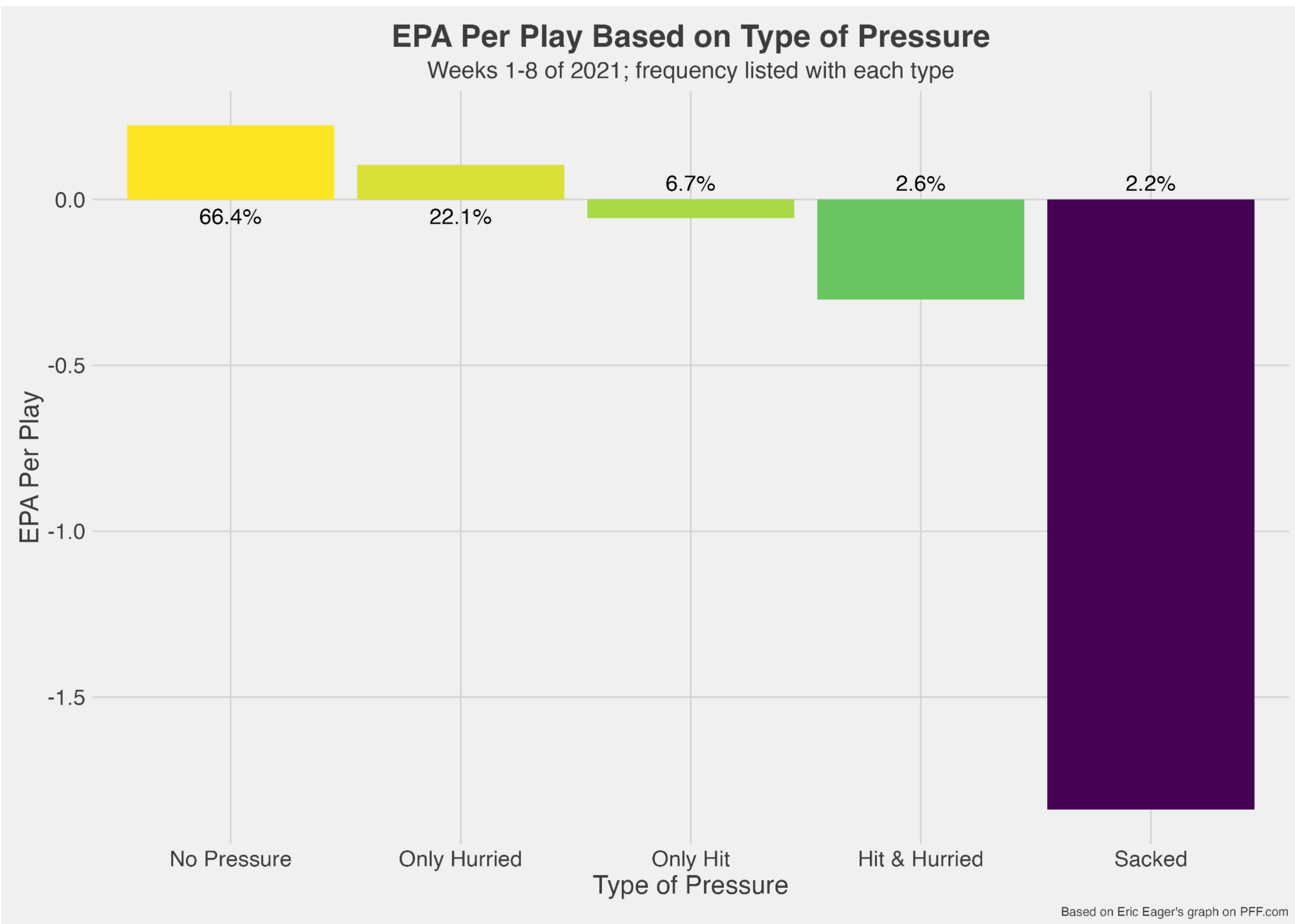


Figure 1: The Expected Points Added (EPA) of each type of pass rusher outcome and the frequency at which they occur

As depicted in Figure 1, pass plays with no pressure generally result in positive EPA for the offense. When the defense applies some pressure, such as a hurry, hit, or both, it results in a decrease in the EPA the offense can generate. However, these differences are marginal compared to achieving a sack, which leads to a significant drop in the EPA of the opposing offense. Sacks are critical to a defense's overall success, and understanding how and why they occur would be beneficial for defensive coaches.

Sacks can occur from various formations, alignments, and positions. Defensive coordinators often face the question of how to position their players optimally to achieve sacks. For example, does an outside linebacker in a 3-4 defensive scheme have a higher chance of getting a sack compared to a defensive end in a 4-3 defensive scheme? What about blitzing a linebacker from his normal alignment five yards behind the line of scrimmage versus a mugged look right at the line of scrimmage?

By developing an expected sack model, we can attempt to answer some of these questions. The model would be able to determine the probability of a player getting a sack based on the game state and the player's pre-snap alignment.

Methods

The National Football League (NFL) hosts an annual competition known as the Big Data Bowl, which provides the public with access to multiple proprietary datasets. In the 2022 edition of the Big Data Bowl, data from weeks 1-8 of the 2021 NFL season was made available. This dataset included several key files, such as "plays.csv" which contains descriptions of each individual play, "players.csv" which provides information on NFL players, "pffScoutingData.csv" which contains manually charted data from Pro Football Focus (PFF), and tracking data that logs the location of each player on the field every tenth of a second, filtered down to the frame at the time of the snap. The tracking data was also adjusted based on the direction the offense was facing and the position of the football on the field, with the football being assigned coordinates of (0,0) and every player receiving new coordinates relative to the football. In the training dataset, each player on each play had their own individual row that included game situation information from "plays.csv", tracking data from the tracking dataset, and player-specific features from "players.csv" and "pffScoutingData.csv".

To model whether or not a player was able to achieve a sack on a play, an eXtreme Gradient Boosted Classifier (XGBClassifier) was made using Python's built-in library, with a 75/25 train and test split. The model incorporated various features that are standard in most football-related models, such as down, yards to go, and yardline number, as well as engineered features aimed at improving predictions. These engineered features included:

- **Offensive Formation**: The formation the offense could be lined up in between Empty, I-Formation, Jumbo, Pistol, Shotgun, Singleback and Wildcat. It's assumed this affects how an offense blocks defenders.
- **Defenders in the Box**: The amount of defenders that are in the box as charted by PFF. It's assumed this influences the number of pass rushers
- **Number of Running Backs, Number of Tight Ends, Number of Wide Receivers, Number of Defensive Linemen, Number of Linebackers, Number of Defensive Backs:** It's assumed the personnel on both sides impacts how difficult it is for an individual defender to get a sack.
- **Relative X Coordinate**: Using the ball as the (0,0) coordinate, this is how far a player is from the ball on the long side of the field (essentially the distance from the line of scrimmage to the player) at the time of snap.

- **Relative Y Coordinate** : Using the ball as the (0, 0) coordinate, this was how far a player was from the ball on the short side of the field (from sideline-to-sideline) at the time of snap.
- **Speed**: The player's yards/ second at the time of the snap. It's assumed speed at the snap could influence how quickly they get into the backfield.
- **Acceleration:** A player's yards/ second2 at time of the snap.
- **Direction** : The angle of a player adjusted for the direction the play is going (0-180 degrees)
- **Orientation** : The angle of a player's motion adjusted for the direction the play is going (0-180 degrees). It's assumed players facing the quarterback will have a higher chance of sacking them.
- **Ball X**: The x coordinate of the ball at the time of the snap (along the long side of the field)
- **Ball Y**: The y coordinate of the ball at the time of the snap (along the short side of the field from sideline to sideline)
- **Official Position** : The position the player is listed as on their team's official depth chart. It's shown that positions have different sack rates.
- **Offensive Line Minimum** : The right tackle's distance from the ball
- **Offensive Line Maximum** : The left tackle's distance from the ball
- **Offensive Line Distance** : The distance between the left tackle and the right tackle on the offensive line.
- **Quarterback Distance From Ball** : The euclidean distance the quarterback is away from the ball at time of the snap.
- **Distance From the Quarterback** : The euclidean distance between the defender and the quarterback at the time of the snap. It's assumed this will be a key feature in the model as it gives the difference between the defender and where the defender is trying to get to.

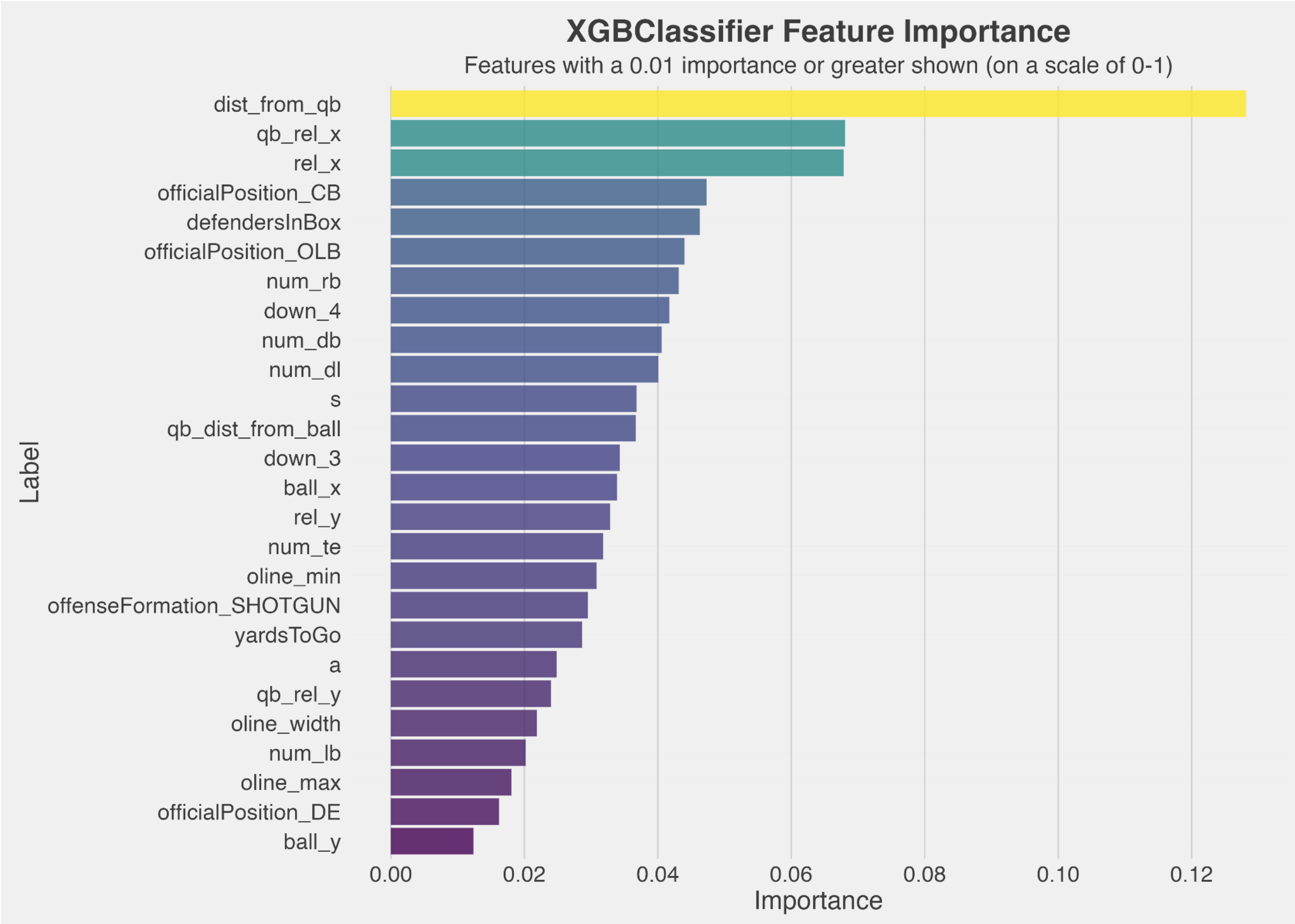


Figure 2: The eXtreme Gradient Boosted Classifier's top features in terms of importance in regards to determining the probability of a defender getting a sack.

Figure 2 presents an analysis of the feature importances derived from the XGBClassifier in the sack prediction model. Notably, the most influential feature in the model was the distance between the defensive player and the quarterback, denoted as "dist_fro m_qb". On average, the distance from the quarterback for players who successfully sacked the quarterback was 7.1 yards, while it was 11.4 yards for those who did not achieve a sack. This finding aligns with intuitive reasoning, as defenders would typically be closer to the quarterback when attempting a sack compared to non-sacking situations.

Furthermore, the quarterback's distance behind the line of scrimmage, referred to as "qb_rel_x", was found to be moderately collinear with the distance from the defender, as defenders tend to be closer to a quarterback who is under center as opposed to in shotgun formation.

The model also identified the significance of late downs, specifically "down_3" and "down_4", as these are often associated with specific pass-rushing packages and creative pass-rushing schemes employed by defenses.

Additionally, the model recognized the importance of specific position designations. When a player was listed as a Cornerback ("officialPosition_CB"), it was noteworthy as cornerbacks tend to achieve sacks the least frequently among all positions, occurring only 0.06% of the time. Conversely, when a player was identified as an Outside Linebacker ("officialPosition_OLB"), it was found to be significant, as Outside Linebackers tend to achieve sacks the second most frequently (after Defensive Ends), occurring 1.36% of the time.

Results

After getting a better understanding of what is influencing the model, we can also look at how the model performed.

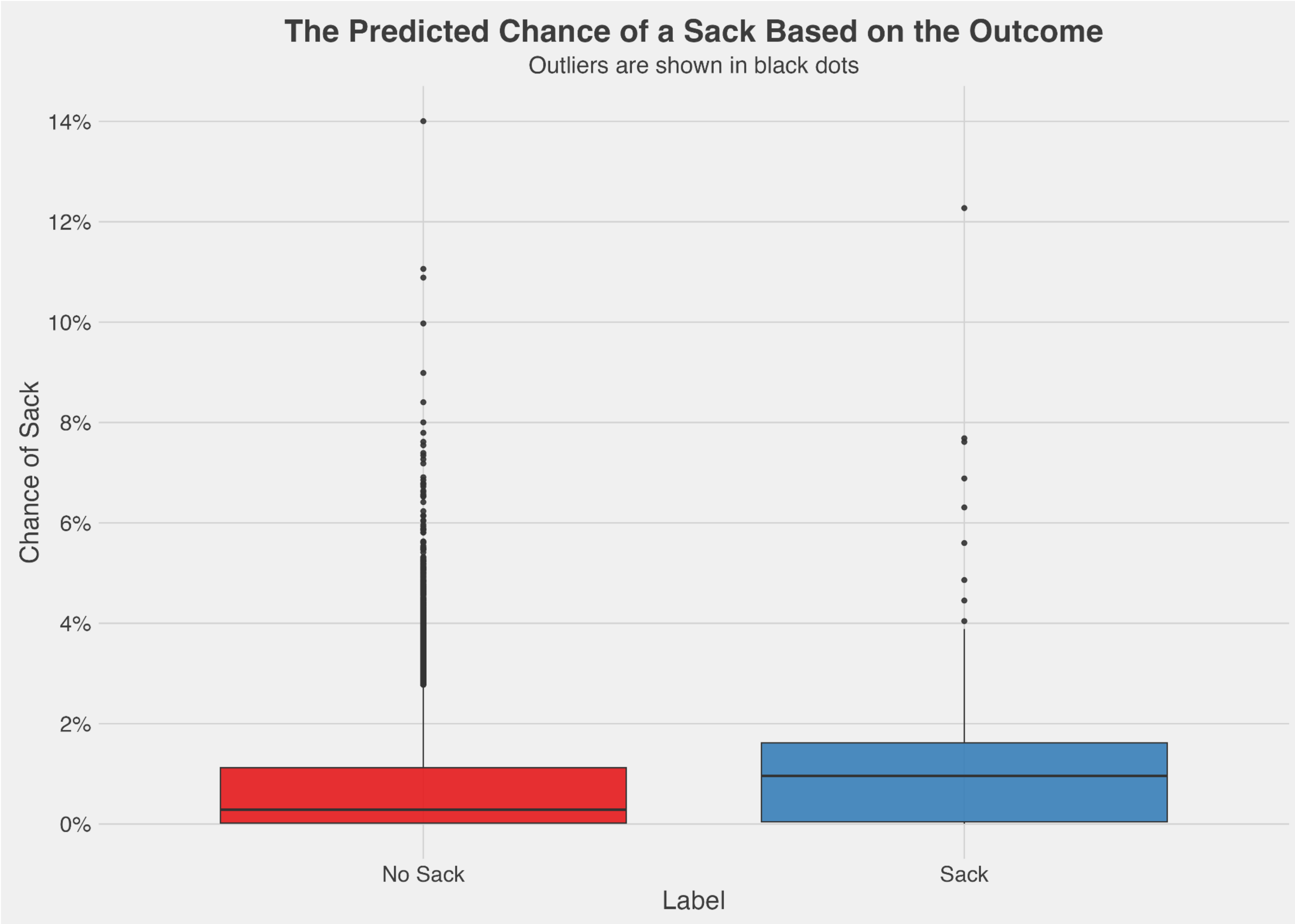


Figure 3: The chance of a sack occurring based on pre snap features and whether or not a sack actually occurred during the play

Due to the rarity of a player obtaining a sack, which occurs only 0.64% of the time, the response variable exhibits zero-inflation. As a result, the model refrains from predicting sack probabilities exceeding 14% for any player. However, as illustrated in Figure 3, the model demonstrates proficiency in differentiating between sack and non-sack scenarios to a certain extent.

For instances where a sack does not occur for an individual defender, the median predicted probability of obtaining a sack is 0.29%, whereas it rises to 0.96% when a sack is achieved. Furthermore, the Brier score for this model, computed on the test dataset, was 0.006, indicating its favorable predictive performance.

In addition to analyzing player -level predictions, it is also possible to evaluate the model's outcomes from a team -level perspective by aggregating the individual probabilities and summing them up: $\sum_{i} sackchance_{i}$

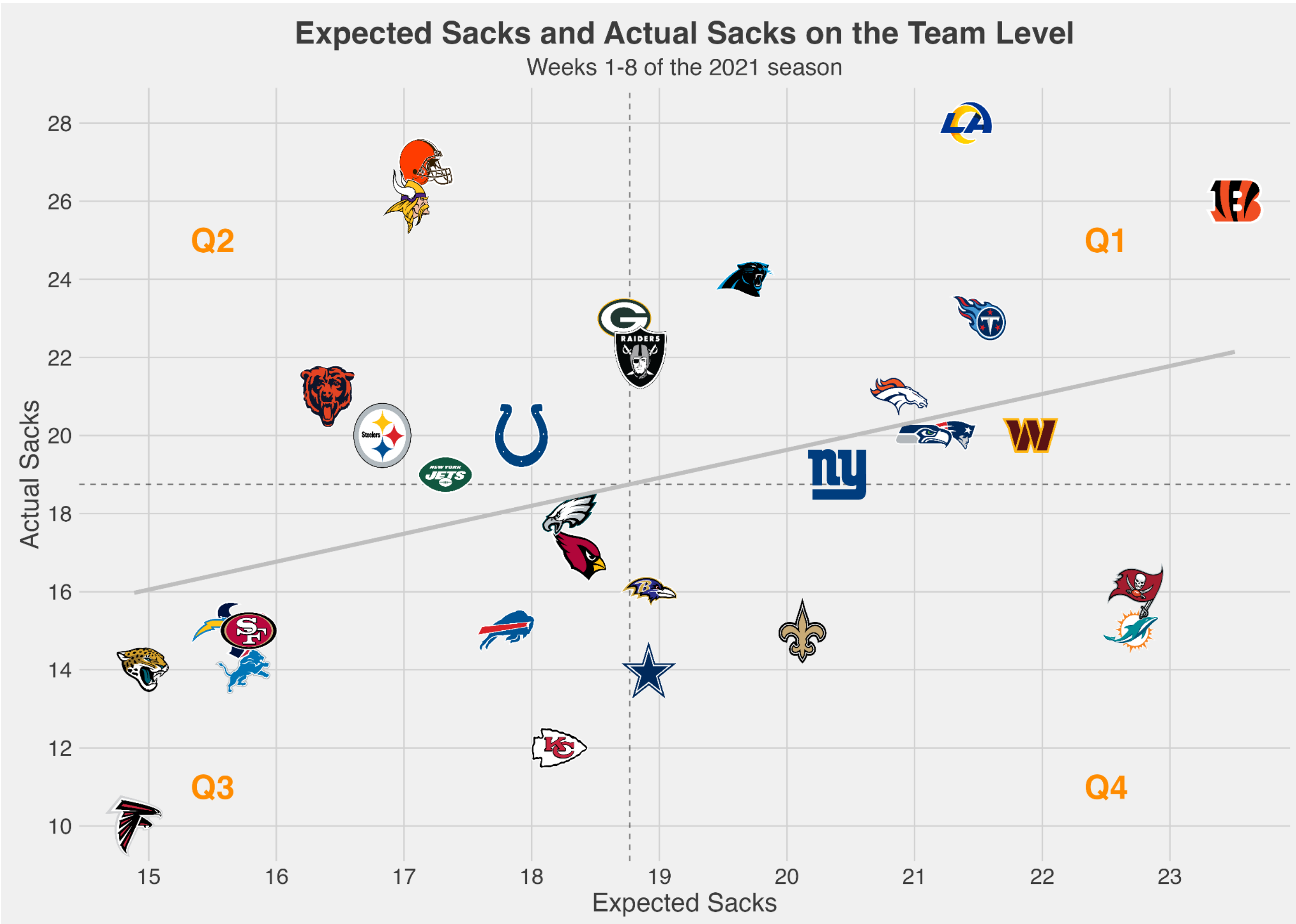
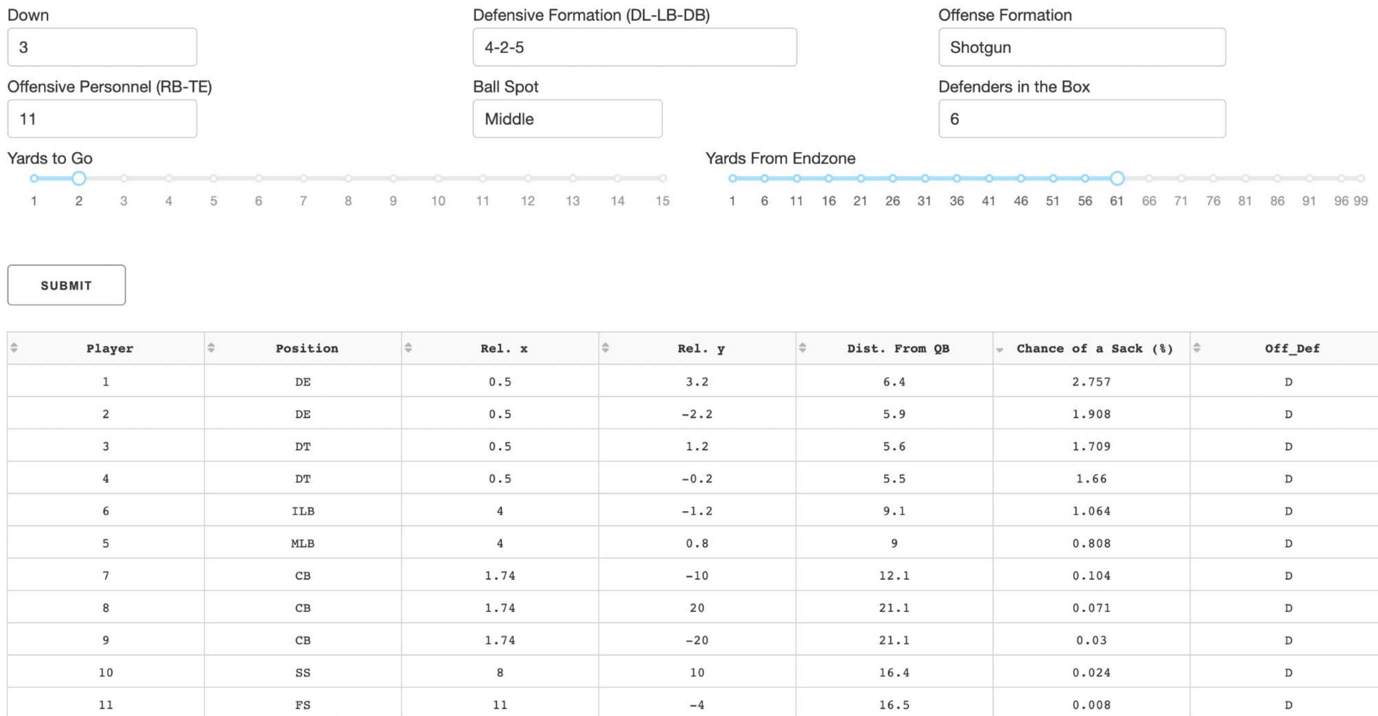


Figure 4: The expected number of sacks and actual number of sacks for each NFL team in weeks 1-8 of the 2021 season.

The teams in Quadrant I are defenses who had players being put in positions that gave them higher chances of getting a sack on average and they took advantage of their opportunities. Quadrant II features teams who were not put in as good of opportunities but overcame it. Quadrant III shows teams that weren't expected to get sacks that often and how that held true for them. Quadrant IV highlights defenses that were put in good positions but underperformed.



Player	Position	Rel. x	Rel. y	Dist. From QB	Chance of a Sack (%)	Off_Def
1	DE	0.5	3.2	6.4	2.757	D
2	DE	0.5	-2.2	5.9	1.908	D
3	DT	0.5	1.2	5.6	1.709	D
4	DT	0.5	-0.2	5.5	1.66	D
6	ILB	4	-1.2	9.1	1.064	D
5	MLB	4	0.8	9	0.808	D
7	CB	1.74	-10	12.1	0.104	D
8	CB	1.74	20	21.1	0.071	D
9	CB	1.74	-20	21.1	0.03	D
10	SS	8	10	16.4	0.024	D
11	FS	11	-4	16.5	0.008	D

Figure 5: The model results based on the parameters that were put in to get the probability of each player getting a sack on an individual play

The model could then be incorporated into a dashboard allowing defensive coaches to experiment with the most optimal positions to put their defenders into to maximize the chance of getting a sack: https://nfl-bdb-front-builder.herokuapp.com/.

Discussion

The primary objective of developing a pre-snap expected sack model, with the goal of providing valuable insights on players and teams and deploying it in a user-friendly dashboard for coaches to strategize for sacks, has been successfully achieved. The app roach to feature selection was intentionally limited to ensure simplicity and ease of use in the dashboard, with careful consideration given to potential features such as nflfastR's expected pass, player technique, and indicators for double mug looks that could have been included if not for the desire to maintain a streamlined user experience.

Looking ahead, the availability of more years of data presents an opportunity to further enhance the model's accuracy through increased training data. Additionally, evaluating the stability of sacks over expected as a predictor of future performance could provide valuable insights for fine-tuning the model and improving its predictive capabilities. With the direction the game of football is moving in general, tracking data can be used to create additional tools that can help coaches in both the pre-game and in-game aspects of gameplanning.

Acknowledgements

I would like to thank Michael Lopez and the rest of his team at the NFL for providing the data for the Big Data Bowl that could also be used in this scope. Additional thank you's to Chris Teplovs, Calvin Smith, Dhruva Krishnamurthy, Eric Eager, Sean Clemert, Zach Drapkin, Sean Sullivan and Meyappan Subbaiah for all the help and support throughout the duration of this project. It was appreciated how generous everyone was throughout the entire process. A special thank you to Michelle Young for contacting me a bout submitting to the Wharton Sports Analytics Student Research Journal.

References

- Big Data Bowl: https://www.kaggle.com/competitions/nfl-big-data-bowl-2023
- Dashboard: https://nfl-bdb-front-builder.herokuapp.com/
- Code: https://github.com/tejseth/hack-a-sack