Final Report

Ethan, Kilbourne, Michael

12/13/2021

xYAF: A Metric for Evaluating Punt Returns

Part 1: Motivation

Every year, the NFL hosts its annual Big Data bowl [1], in which members of the analytics community are able to submit ideas for new metrics, methods of quantifying performance, and other analytical measures that allow the football community to better understand the game from a data-driven perspective. As part of the 2022 NFL Big Data Bowl, which focused on special teams plays, we designed and implemented a new metric called xYAF, short for expected yards after fielding. Using this metric, we hope to provide a method for quantifying the effectiveness of punt returners against a neutral baseline, as well as examine the circumstances of a specific return.

In the creation of our metric, we took inspiration from the recent trend of "expected" statistics in sports analytics. There are many popular examples of these metrics, including expected goals [2] and expected assists in soccer, expected possession value in basketball [3], and expected yards after catch in football [4]. Though these metrics span across many sports and parts of each sport, they generally share a common goal: providing an objective baseline to evaluate their part of the sport. By making a prediction for an event based on what has happened in the past given similar circumstances, we are able to understand the situation in an objective way.

Our xYAF metric, like other expected metrics, will take into account the current game state to attempt to accurately predict the number of yards gained by a punt returner. We will use all of the relevant information possible to make our prediction, including the location of the punt, location that the punt was fielded, locations of all of the kicking team and receiving team's players at the time of the catch, and the hang time of the punt. By accounting for all of these factors and creating a prediction, we hope to provide a neutral baseline for evaluation of punt returners as well as individual returns.

We hope to create a method for evaluating returns which is better than any naïve method of examining returns or returners, such as looking at raw average return yards for a returner. Though this naïve method may seem like a good way of evaluating a returner, the nature of sports is such that many factors may cause an individual or team to perform better in a specific area (like punt returning). If we were to use raw average yards gained as a proxy for returner ability, we would be ignoring many important factors that affect a punt return. For example, when a team punts the ball from further into their own half of the field, the punter would likely punt the ball as far as possible, as the team would be trying to gain maximal yards and field position with the punt. However, when punting the ball in or near the opponent's half, the punter might try to punt the ball a fewer number of yards to "pin" the team close to their own end zone and avoid a touchback. For this reason, if a punt returner plays on a team with a better defense, who stops the opponent deeper into their own side of the field more often, they will likely receive the ball further away from the location of the punt more often. This will mean that the punting team has further to run to tackle the punt returner, allowing the returner to gain more yards on average as a result. This is an example of an indirect influence that is not in the hands of the returner, and should therefore not influence his reputation as a returner. By providing a baseline that takes into account punt location, fielding location, and defender and blocker locations when the ball is caught, our model will predict more xYAF in the case of a longer punt where the defenders are further away from the returner when the ball is caught, and fewer yards in the case of a short

punt with nearby defenders. We can then evaluate punt returners across teams and systems according to this baseline.

Part 2: Methodology

The creation of our data set requires a large amount of data cleanup. The data originally comes in several files containing the tracking data as well as list of plays over the 2018, 2019, and 2020 NFL seasons. The tracking data originally contains location, direction, and other information for each player at each frame relevant to a special teams play. Since there are about 20,000 plays with 22 players each, and each play contains roughly 50 frames, we are have $20000 * 22 * 50 = 22,000,000$ rows to start with between the three tracking data files.

To transform those rows into usable data for our models we had to perform extensive feature engineering. The most important part of this step is transforming each row into a usable observation. To do this we extracted only the frames in the data where the punt was returned. We then took all player x and y positions and combined them into one row, essentially taking a snapshot of all of the players in the moment the ball was returned. After trimming only the returned punts from the data and combining information about each punt, field, and tackle frame to retain only relevant information, we are left with about 2,000 punts to predict with.

To improve the interpretability and accuracy of the model, we replace the x and y coordinates of each of the players with their distances with respect to the player who caught the ball. We then sort the players of each team in increasing order of their distance to the player who caught the ball and rename the team that received the ball to 'receiving' and rename the team that punted the ball to 'punting'. We removed exactly 60 outliers from the data after observing the strong leverage they had on the models. Doing this increased the Test MSE by roughly 60.

We modeled our data using linear regression, logistic regression, LASSO regression, which was given the values alpha = 1 and nfolds = 10, ridge regression, which was given an alpha value of 0 and a range of 100 lambda values from -2 to 10, GAMs, SVMs, splines which used six degrees of freedom for all predictors, decision trees, random forests, which were generated from 1000 trees, Forward and backward best subset selection, both cases of which were given a maximum number of variables of 20, and boosting, which had a gaussian distribution with a shrinkage value of 0.01 and 1000 trees.

Part 3: Results

When we first ran these models with the whole dataset (which included about 60 statistical outliers), the outcomes produced had much more variance and higher MSEs then when we executed the models with the new dataset excluding the outliers. For each of the models tested, we produced a "predicted vs actual" plot to visualize the accuracy of the models in predicting xYAF. If the data points were close to the center line (which corresponds to the predicted values equaling the actual values), then that shows that the model did a better job at predicting the xYAF that we saw in our test dataset.

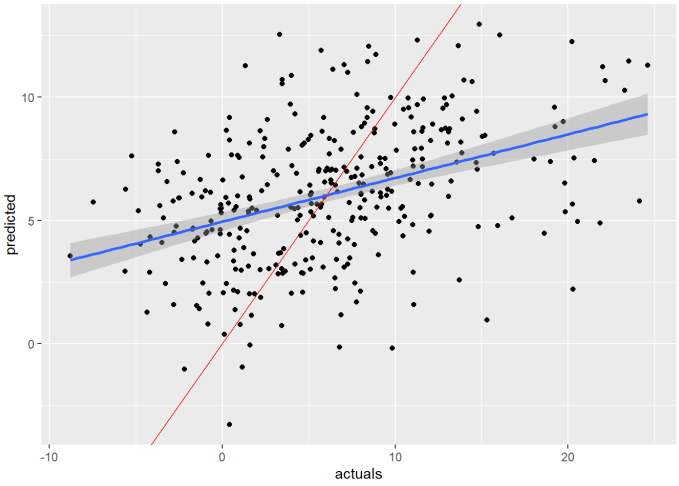

Figure 1: Best Subset Model

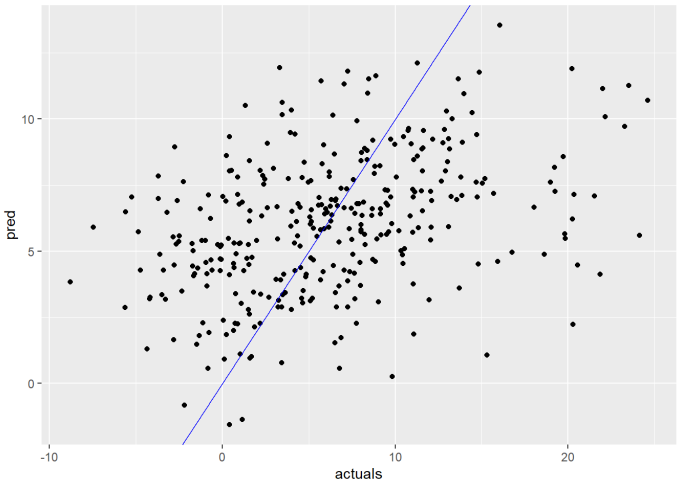

Figure 2: Ridge Model

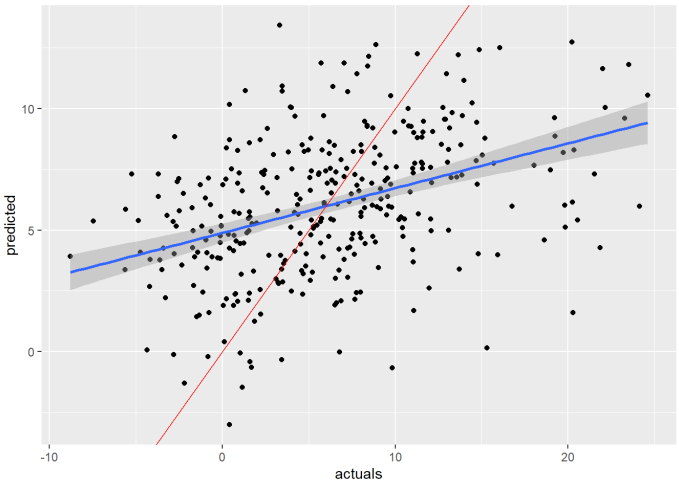

Figure 3: GAM Model

These plots of our best performing models look very similar to each other each model appearing to have a similar spread of data points from the perfect prediction line. These plots show that the models we used had a decent ability to predict xYAF. Based on the mean squared error (MSE) of all the models used, ridge regression has the lowest MSE with 33.52. This means that on average, our predicted xYAF has an error of 5.74 yards.

```
  Linear    Logit    Lasso    Ridge      GAM      SVM   Spline Random_Forest
33.67316 33.67316 33.88527 33.51643 33.67316 35.56712 36.82119      33.85723
Best_Subset Boosting
   33.58888 34.08958
```

Figure 4: MSEs

This error is acceptable in terms of how its played in the context of punts. Having a prediction within 5.74 yards shows that this can be method can be a viable metric for predicting punts. Although this method had the lowest MSE, all the methods we tested had similar MSEs meaning each method had a similar performance in accurately predicting yards gained after fielding the punt. This shows that our methods could be further improved even further to lower the MSE and improve the accuracy of predicting xYAF.

Challenges and Solutions

There are many challenges that we battled when trying to accurately punt return yardage. One of the main issues is with the distribution of punt return yardage in general. We know that for some linear models, we assume a normally distributed response variable. However, punt return yards follow a right-skewed distribution.

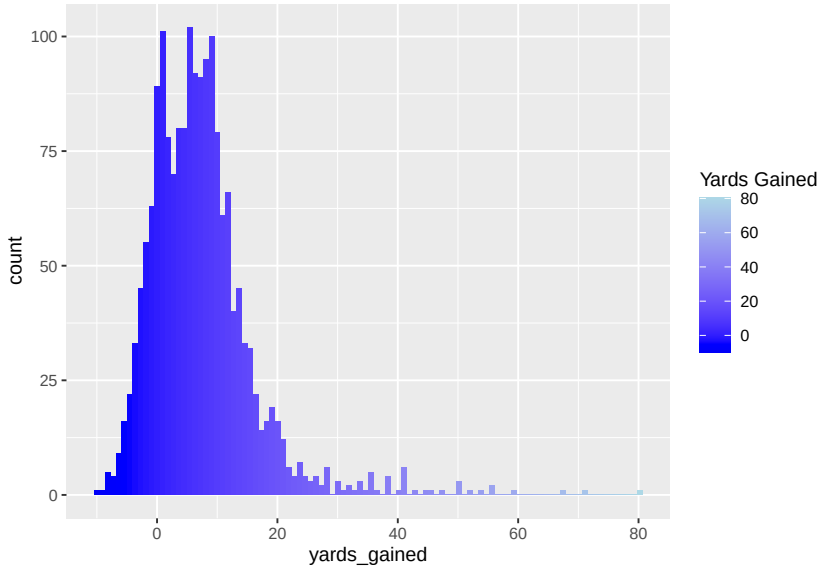

Figure 5: The distribution of all punt returns

As we see in the first distribution, while the vast majority of punt returns are in the 0-15 yard range, we observe punt returns all the way up into the 80-yard range. For certain methods, this will cause poor predictions.

Certain transformations that tend to help this phenomenon, like a square root transform as in the second histogram below, are not viable because we observe negative punt returns about 15% of the time.

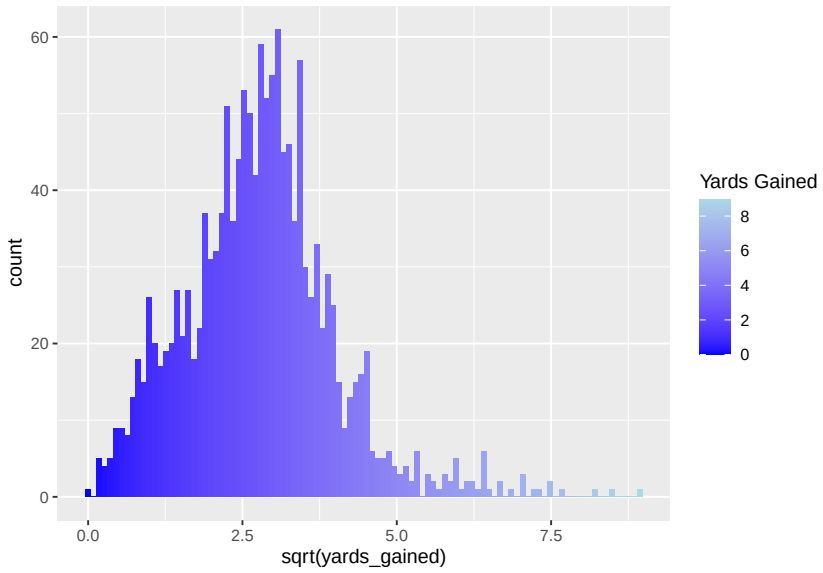

Figure 6: The distribution of all punt returns with a square root transform applied

This leads to another interesting question about our problem, however, that is somewhat unique in prediction cases. When predicting punt return yardage, it may be the case that we do not necessarily want to predict accurately at the right tail of our distribution. If a player makes an outstanding return, evading many nearby defenders and scoring a long touchdown, we wouldn't necessarily want to use this as the standard of what we should expect as these returns are very unlikely. This is the reasoning that led us to dropping the outliers on the right tail, which would not generally be a desirable method for transforming data.

Applications

To evaluate the effectiveness and demonstrate the usefulness of our model, we can apply it to both individual punts as well as evaluate punt returners based on their consistent performance against our predictions. We will first apply the model to individual punt returns, where we will demonstrate how the model would be used in such a case.

The first punt return we will examine takes place between Seattle and San Francisco on November 11th, 2019. Seattle punts the ball from their own 25 yard line. The punt travels all the way to the opponent's 25 yard line before being fielded by Richie James Jr. of the 49ers. When James Jr. catches the ball, he has quite a few defenders close to him, the closest being 6 yards away, and three being fewer 10 yards away. James Jr. does not get very far, losing his balance and succumbing to the defense after only gaining 1.5 yards.

Out model considers all of the relevant information above for this return, and is able to make a very good prediction, predicting 1.52 yards, just 0.06 fewer than the actual return of 1.58 yards. This is an example of a fairly normal punt return, which is likely why our model is able to predict very well in this scenario.

We will also evaluate a second, less common punt return where our model struggles to make an accurate prediction. This return took place between the Pittsburgh Steelers and Arizona Cardinals in Week 14 of the 2019 season. The Cardinals punt from their own 15 yard line, and after a very long punt Diontae Johnson fields the punt at the Steelers 15 yard line. The closest defender is 11 yards away, but there are hardly any

defenders nearby other than him and the Steelers have a blocker near this defender. Johnson is able to evade the entire defense, taking the punt back for an 85 yard touchdown.

Our model makes a very poor prediction of 4.8 yards for this punt return, assuming that Johnson will be tackled by either the first defender or soon after. This contributes $(85 - 4.8)^2 \approx 6400$ units worth of squared error to our results, which is an example of why our MSE is much better after filtering outliers like this punt [all images courtesy NFL]. However, as mentioned previously, we may not even be interested in making an accurate prediction in this case, as this punt return is very unlikely even given its circumstances.

Another use case for our model is evaluating many punts by many different returners and evaluating which returners tend to beat the predictions by the most yards. As mentioned in the Motivation section, this may be preferred to a naïve average return yards metric for measuring returner performance, because our metric is able to account for many factors that go into a punt and evaluate all players on the same baseline. To implement xYAF for evaluating returners, we run our best Ridge model on every punt in the three years worth of data. We make predictions on every punt, take the difference between the actual return and the predicted return, and use this as a measurement for kick returner effectiveness when averaged over every all of the returns. Below are the top 10 punt returners (minimum 10 returns) based on actual - xYAF.

name	avg_actual_minus_pred	punts_returned
DeAndre Carter	6.450089	50
Jabrill Peppers	6.276976	35
Hunter Renfrow	5.987176	23
Dontrell Hilliard	5.671469	12
Marcus Sherels	5.077837	16
David Moore	5.038203	13
Andre Roberts	4.985457	54
De'Anthony Thomas	4.862670	16
Adoree' Jackson	4.850452	13
Braxton Berrios	4.778000	20

These 10 players consistency outperform their predicted yards after catch by the most. The best among them is DeAndre Carter, a Wide Receiver for the Washington Football Team. Carter averages 6.5 more yards per return than would be expected. Examining Carter's statistics, he averages about 9.4 yards per return over his NFL career [5], which is good but not outstanding. However, as mentioned in the Motivation section, we would expect a player who generally receives the ball with less space to average less return yards, and having less space results from the opponent having longer drives and punting the ball from further down the field which would imply that the returner played for a poor defense. In the 2018 and 2019 seasons where Carter made over 75% of his punt returns, Washington's defense ranked 17th and 27th respectively in yards allowed by opponents [6]. These years were also Carter's best two seasons of returning, where he averaged 9.6 and 9.7 yards per return, respectively [5]. This is a great example of the usefulness of xYAF, because while Carter's numbers may not be the highest in raw average return yards (Carter's Football Team ranked 25th and 32nd in average punt return yards in these two seasons, respectively [6]), he is a much better punt returner than those numbers show.

We can also view the returners that performed the worst against our projections.

	name	avg_actual_minus_pred	punts_returned
53	Trevor Davis	0.4436541	20
54	Mecole Hardman	0.2216395	28
55	Golden Tate	0.1102866	15
56	Ryan Switzer	0.0145936	23
57	D.J. Moore	-0.1364440	10
58	Justin Hardy	-0.6882564	14
59	Tyler Lockett	-0.8233248	28
60	Jaydon Mickens	-0.8522292	23
61	Adam Humphries	-1.6317841	18
62	Tavon Austin	-2.4416013	18

These players would be considered under performers in our model.

Future goals

Many factors could go into future improvement of this model, including further feature engineering, experimentation with error metrics, and implementation of likelihood intervals. As with many machine learning problems, the features in this model affect our results considerably. To this end, experimentation with our features would most likely lead to lower error if we were to examine our data closely enough. One transformation that comes to mind that may help would be to include more information about blockers in our model. Our data does include information about how far away the blockers are to the returner when he catches the ball, but we could also measure their distances to the defensive players which may help our models understand whether the player will recieve effective blocking.

There are also alternate error metrics that could be useful in our model. For example, one error metric called SMAPE (Symmetric Mean Absolute Percentage Error) [7]. This metrics has many uses in prediction, as it places a larger focus on accuracy for small prediction than it does for large predictions. This could be useful in our case, as we may want to treat an error of predicting a 1 yard return when the true return was 5 yards more harshly than a prediction of 12 yards when the true return was 16 yards. The further a punt return is, the more difficult it likely is to predict the exact yardage since we have few punts in the higher yard ranges, so this error metric may be desirable.

Finally, there are features that could improve the usefulness of our model further as well. One such feature would be the implementation of likelihood intervals for each punt return, so that we could evaluate the likelihood of the actual result as well as viewing it in context with all of the other possible results. We could implement this using simple confidence intervals, and it may be included once we are able to tune our models to a higher standard of prediction.

Works Cited

[1] https://operations.nfl.com/gameday/analytics/big-data-bowl/

[2] Sam Green. Assessing the performance of Premier League goalscorers. Stats Perform, 2012.

[3] https://grantland.com/features/expected-value-possession-nba-analytics/

[4] https://www.nfl.com/news/next-gen-stats-intro-to-expected-yards-after-catch-0ap3000000983644

[5] https://www.pro-football-reference.com/players/C/CartDe02.htm

[6] https://www.footballdb.com/stats/teamstat.html?lg=NFL&yr=2018&type=reg&cat=T&group=D&conf=

[7] https://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error