Andy Giorgio
12/18/2024
Stat 5905
Paul Lupinacci

Predicting NFL Touchdown Scorers

Introduction:

The NFL is America's largest professional football league as well as the highest grossing sports league in the world. The players and teams are not the only ones making money off the league though. Sports betting has started to be legalized in more and more states and companies as well as certain individuals stand to make huge profits from this development. In 2017, Nevada was the only state with legal mobile sports betting. Today, it is legal in 38 states as well as Washington DC. The NFL is by far the most bet on sport across these sports books and one of the most popular bets are anytime touchdown scorers. This means people are trying to accurately predict which players are going to score in each week. With this rapid growth of sports gambling and interest in predicting game outcomes, finding the stats that are influencing these results is crucial and could be very beneficial. Additionally, as of 2022, 29.2 million people in the United States participated in fantasy football. Fantasy Football is where people make a league where they draft players from all different NFL teams and then tally up a point total each week depending on how the players they draft perform. Points are given based on things like yards and touchdowns with touchdowns being much more heavily weighted. Since it is on a weekly basis and you can only start a set number of players, it would be very beneficial to know which of the players on your team are more likely to score so you can start those players. Both scenarios point to the value of being able to predict who will score touchdowns each week.

The NFL and sport of football also have some important background before moving onto the model building. First, defining what it means to be a touchdown scorer. In both fantasy football and in sports books, a touchdown scorer is whatever player possesses the ball in the endzone that results in a score for the team. This specifies that it is not counted for the player throwing the touchdown, passing touchdowns are counted as their own stat. Additionally, there are 4 primary offensive positions that score majority of the touchdowns and are given the odds to score by sportsbooks. The first position in quarterback, these are usually less likely to score since they throw the ball, but some quarterbacks are known to run more and score more because of this. Second is running back, these are usually the players that are most likely to score because they run the ball frequently and usually get the ball when the team is close to the endzone. Third is wide receiver, these are the widest

range for odds because they are the ones catching the ball and some get the ball very often while others rarely get it, but it is much more variable than how much the running backs get it. Finally, tight ends are basically bigger receivers that must block more which leads them to generally have the least likelihood of scoring.

Goals and Research Questions:

The goal of this project is to determine which variables have the most significant impact on scoring a touchdown. Then using those variables, a model will be created that will give an expected value of touchdowns scored for a given player. It will be determined using these expected values whether the model is statistically significant. This will be helpful for the fantasy football managers in knowing who they should start that week. However, determining who to bet on is a bit more complex. When betting on an anytime touchdown scorer the book gives you set odds to bet on for each player. No matter the odds, there is always an implied probability that the odds have. For example, a line of +100 means you bet 100 dollars you win 100 dollars. It makes sense that this implies a 50 percent chance since it is even payout on both sides. When the odds are something like –300, you must bet 300 dollars to win 100. This implies a 75 percent chance of what you are betting on happening since you put up 3 times the money of what you are betting against. Finally, odds of +300, you pay 100 dollars to win 300 dollars. This implies a 25 percent chance of an event occurring. To have the model be useful to sports bettors, the model must not only accurately predict who will score touchdowns, but it must do so more accurately than the models of the sports books. To make the probability conversion even more complicated, the books aren't going to give even odds to the public because they need to profit over time rather than just breaking even. This is called a house edge, and a prime example is that when a bet has 50/50 odds, rather than having both sides be +100, they make both sides –110. This means you must bet 110 dollars to win 100. It also means if there are equal bets on both sides then the book is profiting 10 dollars on every 220 dollars bet. This is standard in gambling and can be seen with things like the greens on a roulette wheel. What this means for the model is that it must be that much more accurate to not only beat the book, but also its edge. There are some other important questions outside the betting world to answer as well. Questions like what variables are actually having the most influence on touchdown scorers? Can a model be made to accurately predict touchdowns in general, unrelated to betting odds? How can predictions be made using statistics of players who get injured, replace an injured player, or breakout mid-season? All of these will be important in the completion of the project.

Potential Variables:

For each position there are different variables that will have significant impact on whether they score or not. How much they are in the game is obviously crucial, snaps/game and snap percentage will give insight for how often a player is on the field for any position. Where it might differ by position is the amount a player has the ball, also known as their volume, which is extremely important. For quarterbacks and running backs the number of rushes, they attempt gives us their volume. For wide receivers and tight ends, their volume is the number of times they are thrown the ball (targets) or number of catches. Some running backs also act as receivers on some plays, so their volume might also take targets and catches into account. Touchdown scoring history will also likely be a major factor, both from the previous year and from the current season up to the given week. Another variable likely to be significant is the team's scoring rate for rushing versus receiving touchdowns and, the opposing team's defensive allowance of rushing scores versus receiving scores. There could be some value in using the sports books to help with the prediction, for example, looking at the expected number of touchdowns in the game could influence how many touchdowns the model should be predicting for the players in that game. It could also be interesting to look at the players' odds in the previous weeks of the season to see if that has any correlation with their chance to score in a new week. Other variables that might have an impact that are worth exploring include, age, whether the team is home or away, weather, and things like usage in the redzone (within 20 yards of endzone) and player efficiency might also come into play. A final hurdle will be to deal with player injuries. They can be very random, but after an injury something must be done to account for that player's volume as well as the rates of usage by position and possibly the offense's success in general.

Data Sources:

There are a couple websites that provide exactly the data needed for the model. The first is the Fantasy Pros website. This website has more advanced statistics that the model will require such as snaps/game, snap percentage, rush percentage, target percentage, and touch percentage. Each of these provides a different insight into a player's volume and it will be crucial to the model's success. The individual advanced statistics are helpful, but having the more basic statistics for all players and teams is also important. For that, Pro Football Reference will be used. This will round out the player and team data needed for the model. The variables from Pro Football Reference include, touchdowns scored last year, touchdowns scored this year, teams' points scored per game, teams' points allowed per game, position, age, and any other variables on players or teams that might need to be

used. Finally, the last data source will be Betting Pros website. Betting Pros gives us the required data for variables more related to betting. For example, the players odds in previous weeks to score a touchdown. Also, the number of touchdowns expected by each team in each week. Another perk of Betting Pros is that it has all the touchdown scorer odds for every player in every book. It also provides a way to look at the consensus odds and the best available odds for every player. This will be helpful because we can use the consensus odds to give a best estimate of the books implied chance for a player to score, but then when betting using the model we can bet on the book that is giving us the best odds for the player we find value in. The reason this is helpful is because each book has its own odds and if the model finds value against the average, betting on the book that gives even more value increases the chances of profit.

Initial Model Building:

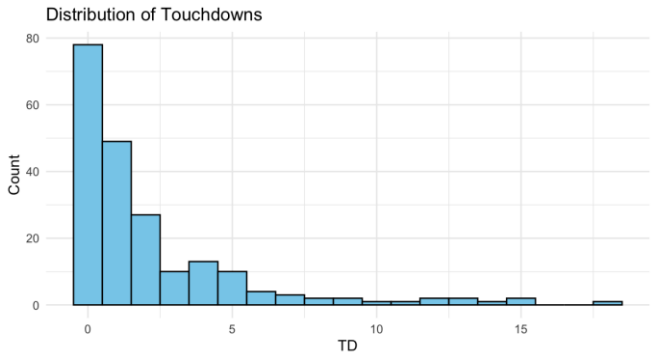
Now that the goal of the project has been determined, the variables have been determined, and the data is collected, the model can begin to be constructed. The data from the primary sources will be joined and the first step will be determining the most significant variables and how much each should be weighed. Once the weight is determined, the model will be run on a given week for the NFL. The odds given by the model will then be compared to the implied odds given by the sports books. If significantly different it can then be used to bet. A margin of some value must be found to decide how much confidence the model needs to place a bet on a given player. Once that margin is determined and theoretical bets are placed the winnings for that week will be tracked. The model will be run again each week for the remainder of the season and the same process will take place. The significance of the model over that period should become evident and at that point conclusions will be made regarding the significance of the variables as well as the overall model. The overall profit of the model will also be of great interest although random variation could be at play so it might not be as important as other conclusions.

First to determine which variables are going to be used in the prediction model for the 2024 season, the data from the 2023 season will be used. By compiling all the data from 2023 and doing an analysis to determine which variables are most significant at predicting touchdowns, the variables for the 2024 prediction model will be determined. When starting the initial analysis both receiving and rushing data were combined and a model based on total touchdowns scored was made. It became quite clear that this would not be effective, as having the rushing and receiving statistics together as well as rushing and receiving touchdowns made it so the variables either lacked significance or had negative coefficients that did not make any sense. For this reason, it was quickly

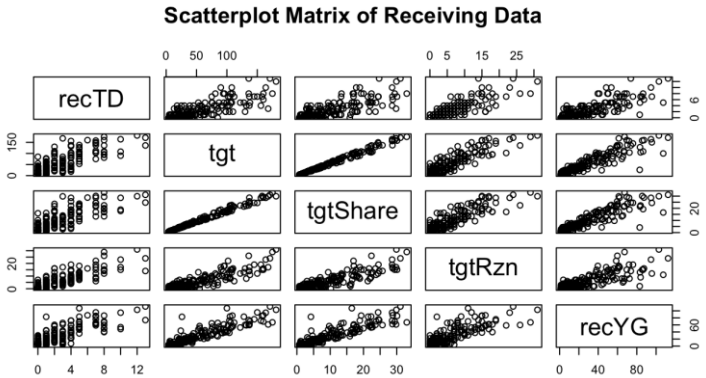
determined that making two separate models, one for rushing touchdowns and one for receiving touchdowns would be far more effective.

2023 Receiving Touchdown Model:

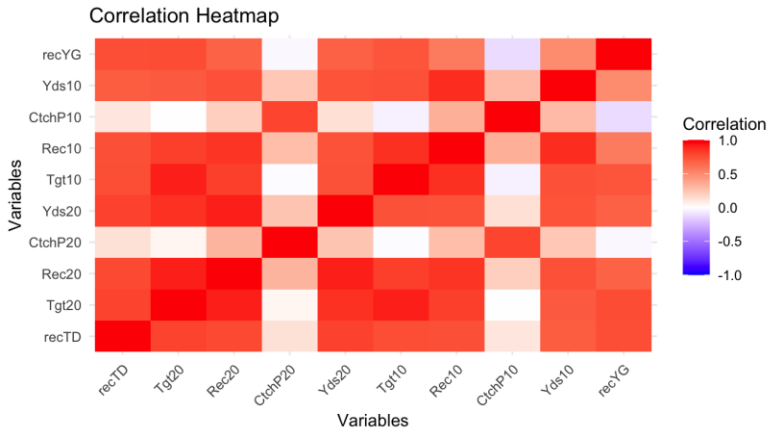
To start with the receiving model for the 2023 season, all the 2023 receiving statistics for all receivers, tight ends, and running backs were compiled. Before running any tests, the normality of the touchdown scored variable must be checked. Below is a histogram of the variable. It is clear the data is not normally distributed. Transforming the variable did not solve the problem so instead, a generalized linear model with a Poisson distribution will be used rather than a basic linear model. This makes sense because the touchdowns scored are count data which should follow a Poisson distribution. This will be applied to all future tests and models as both types of touchdowns have similar distributions.



The first test was to be run using the following variables; targets in 2023 season, percentage of team targets in 2023, targets in the redzone in 2023, and yards per game in 2023. After looking at the scatter plot matrix below it is clear there is high multicollinearity between variables. This makes sense because players with a high number of targets will have a higher percentage of targets and also likely more targets in the redzone and more yards.

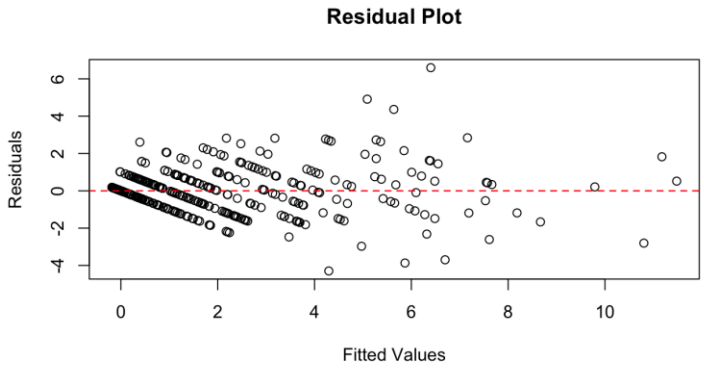


Despite the multicollinearity, all these variables seem like they should have an impact on the receiving touchdowns based on logic as well as the plots above. A forward selection stepwise AIC was run with these variables and the resulting model found that only the receiving yards per game in the 2023 season and the targets in the redzone should be included in the model. The model had an R-squared value of 0.7458 which indicates that nearly 75% of the variability in receiving touchdowns scored in 2023 can be explained by these variables. Redzone targets had a coefficient of 0.253 and receiving yards per game had a coefficient of 0.039 and both had p-values far less than 0.0001. This also solved the multicollinearity issue because the model with just these two variables had a vif value of 2.689 which is well below a worrying vif value. Due to the extreme significance of the redzone target variable it seemed like digging deeper into the redzone statistics could enhance the model. Statistics on targets, receptions, yards and catch percentage within the 10- and 20-yard line were found and added into the model. Catch percentage was clearly not significant so that was removed. Like the previous model there was an extreme issue of multicollinearity for the same reasons.



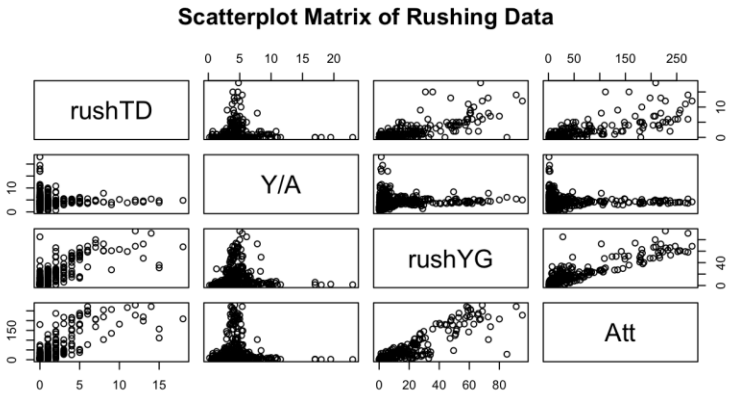
Again, each variable seemed important, so a stepwise AIC was run, but this time it was decided that a bidirectional was to be used as when just a forward selection was used some of the variables left in the model were insignificant probably due to the high multicollinearity. The final model ended up including the same variables as before, being receiving yards per game and targets in the redzone (20-yard line), but receptions within the 10-yard line were added. With the addition of receptions within the 10 the VIF for redzone targets did increase to 5.45 which is a bit worrisome, but due to its extreme significance this will be ignored for now. The new R-squared came out to be 0.761 which is a bit higher than before, but nothing too extreme. The new coefficients were, 0.102 for redzone targets, 0.043 for receiving yards per game and 0.526 for receptions within the 10. Again, all these variables had p-values well below 0.0001. With that, the variables that were to be used for the 2024 predictions were set; receiving yards per game, targets within

the redzone, and receptions within the 10-yard line. The residual plot of this model can be seen below. It appears to have a bit of an opening megaphone shape and thinning out, but it isn't too bad and after trying some transformations on the variables nothing was really able to improve the plot so this model will be used.



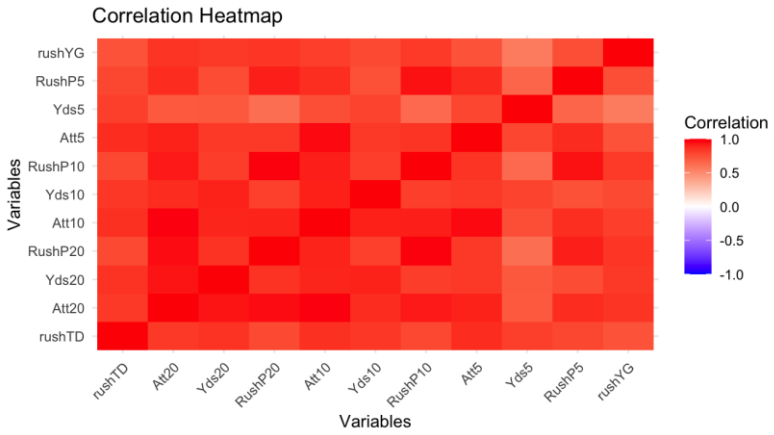
2023 Rushing Touchdown Model:

For the 2023 rushing touchdown model nearly the same structure was used as for the receiving touchdown model. The main difference being that the statistics used are slightly different. The initial variables in this model were; Rushing attempts, yards per attempt, and yards per game. The original data had no information on redzone statistics, but those were to be added later. The scatter plot for this model showed that there was likely some multicollinearity, but also that the yards per game and total attempts were likely correlated with rushing touchdowns and yards per attempt might not be.

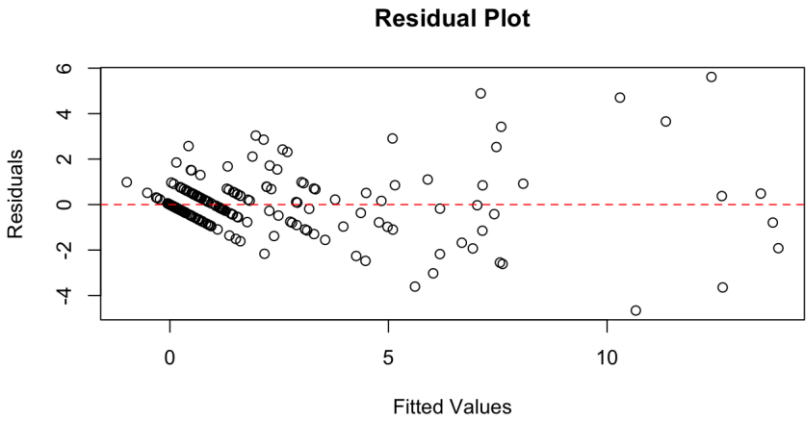


After running forward selection on this data, it was confirmed that rushing yards per game and total attempts were significant in predicting rushing touchdowns. The model produced an R-squared value of 0.6322 which is lower than the original receiver model, but

once redzone statistics are added in it should rise a decent amount. The coefficients for the variables were 0.022 for rushing attempts and 0.041 for rushing yards per game. Both had p-values less than 0.001. The VIF for the model was 5.075 which isn't ideal being above 5, but once the redzone variables are added one of these variables could be dropped so they will be left alone for now. Since the redzone statistics clearly have an influence on touchdowns the redzone rushing statistics were added to the model. The rushing redzone statistics include statistics on attempts, yards and rushing percentage withing the 5-, 10- and 20-yard line. Not surprisingly, we see extremely high multicollinearity within the redzone data here as well.



Once again, a stepwise AIC was run, and both previously significant variables were dropped from the model. The new variables that took their place were attempts within the 5-yard line, yards within the 5-yard line and yards within the 20-yard line. At first this might seem weird that both the previous variables were dropped, but for running backs it makes sense that redzone stats dominate because they score nearly all touchdowns from close range whereas receivers are much more likely to score touchdowns longer than 20 yards. The final model for the running backs gave an R squared of 0.8488 which seems very high but makes sense because so many rushing touchdowns are scored from the 1- or 2-yard line. The coefficients for the variables were 0.296 for attempts within the 5, 0.167 for yards within the 5 and 0.029 for yards within the 20. Each had a p-value of less than 0.0001. The residual plot of this model can be seen below. Like the receiving touchdowns, it appears to have a bit of an opening megaphone shape and thinning out, but it isn't too bad and after trying some transformations on the variables nothing was really able to improve the plot so this model will be used.

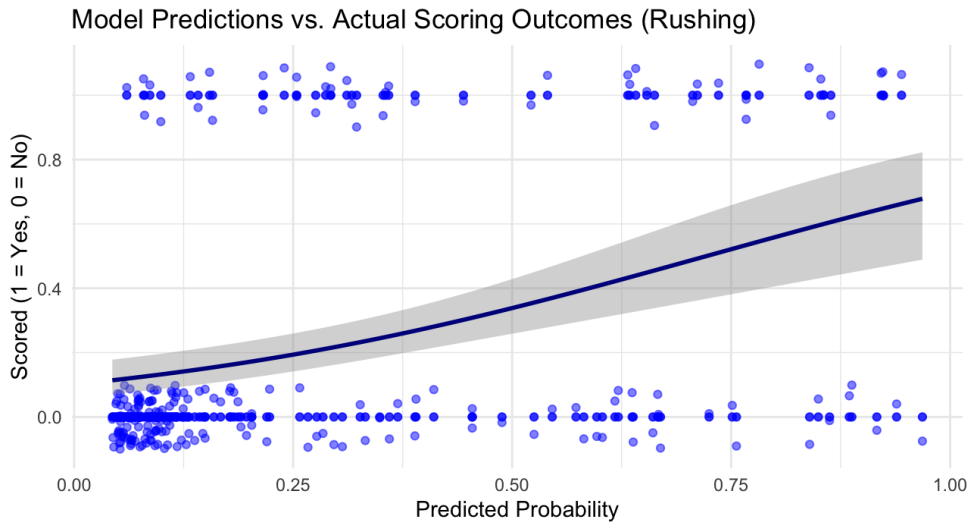


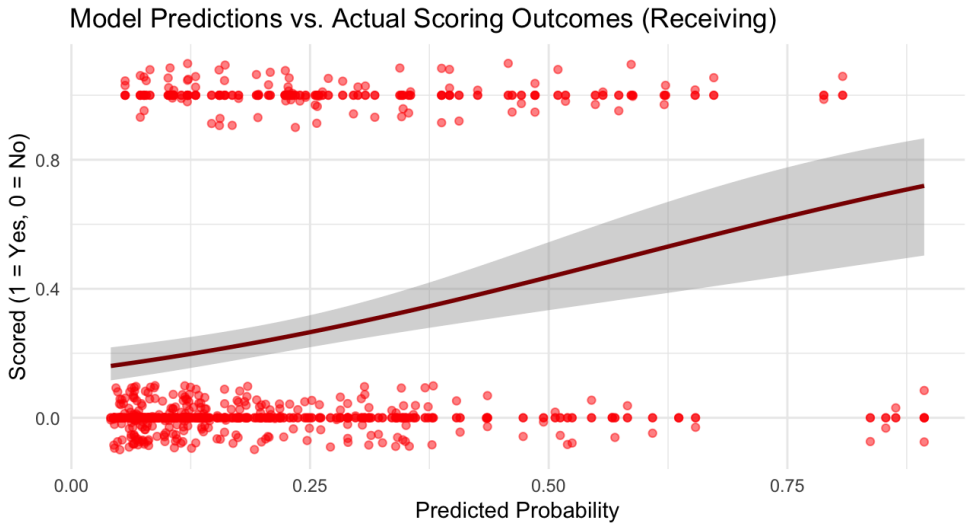
2024 Prediction Models:

Now that the variables for the prediction model have been selected the predictions for 2024 can be made. All the same data that was used to determine the variables from 2023 was collected for 2024. The rushing and receiving data were again separated for the two models. Both models were to function in the same way, where the selected variables from 2023 were used in a binary GLM (generalized linear model). The GLM will be used instead of a normal linear regression since the response is whether a player will score a touchdown or not and this is not normally distributed but rather binary. Also, to improve the model since a player scoring in a previous week is strongly correlated with a player scoring again in a future week, the baseline for players in the model is the decimal percentage of games the player has scored in so far throughout the season. Each stat that was selected from the 2023 is converted into a per game stat for the 2024 season so far and then depending on the stat it moves the probability to score up or down. A big issue that had to be addressed was player injuries as then people filling in that only played a few games would have their predictions impacted by too small a sample of games to have accurate averages. To avoid this, only players who had played more than 7 games are included and given predictions rather than everyone who has stats so far this year. After making the predictions and the week passes the true outcome of each player scoring or not is known, and tests to see how effective the model is will then be run. After the probabilities are made, they are also converted to the equivalent American betting odds to determine which players the model sees value in. If the odds generated by the model are lower than the odds of the sports books, then the model sees value and would place a bet on that player. The reason this makes sense is because in the long run if the model probabilities are outperforming the book probabilities, then profit will be made.

2024 Prediction Models' Success:

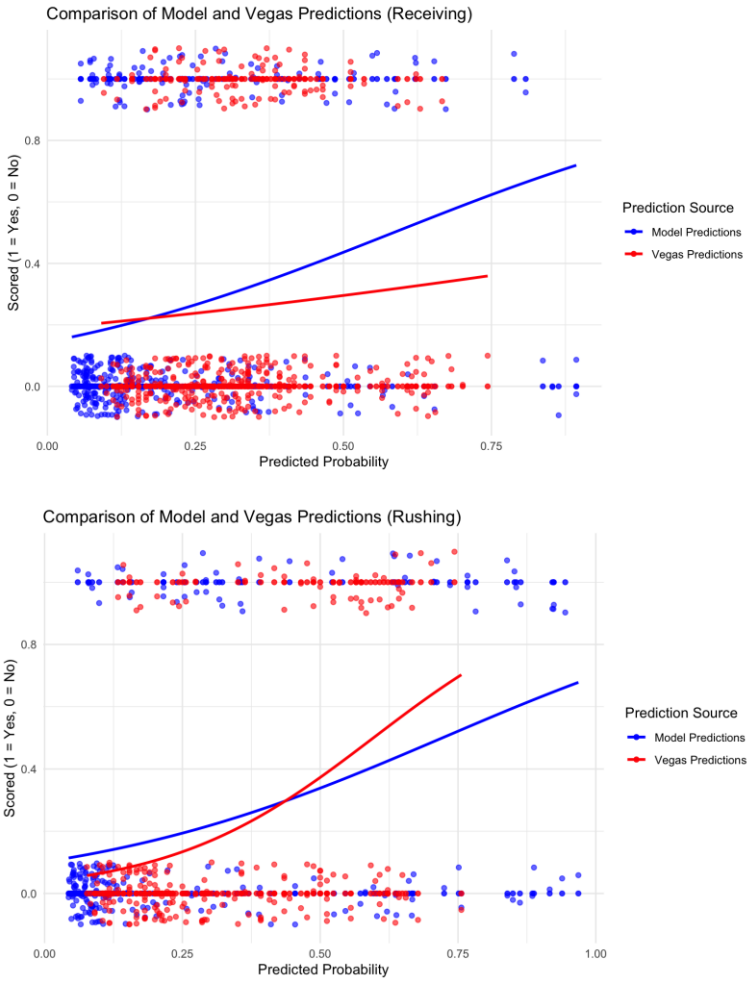
First to test the success of the models in their entirety unrelated to the betting edges, plots of both the receiving and rushing touchdown model success are below. This data is after tracking weeks 13, 14, and 15 of the 2024 NFL season.





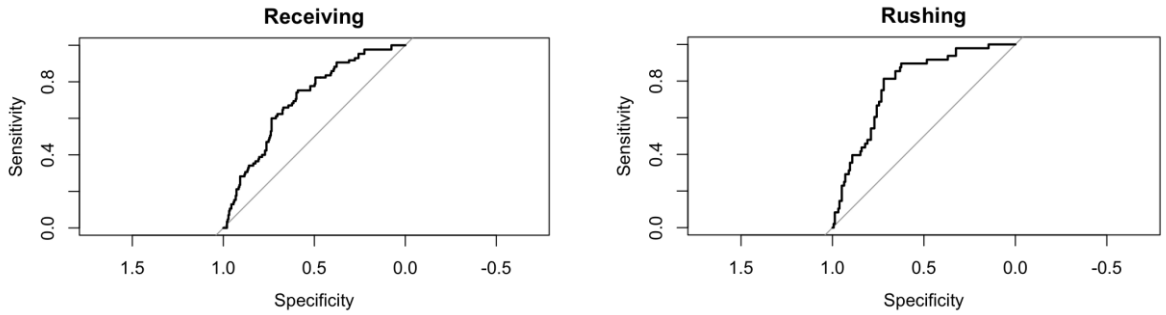
Both models clearly have predictive power on who will or will not score touchdowns. After running logistic regressions on both models' positive coefficients for the predicted probability on touchdown scored are seen for both with p-values > 0.0001. The coefficient is 3.0225 for the rushing model and 3.0438 for the receiving model. Both indicate that when the predicted probability is increased by 0.1 the change in odds will be

approximately $e^{0.3}$ = 1.35. This means that as the predicted probability increases by 0.1 the chance of a player scoring increases by a factor of approximately 1.35. This is reassuring as it means that the higher predicted probability given the higher the likelihood that the player will score a touchdown. Rather than only comparing on an individual player level how the sports books predictions relate to the model, a plot comparing the sports books implied odds with the models' predicted odds and their success are below.



From these plots, though the models were both accurate individually, when adding the Vegas probabilities on it appears that rushing touchdowns are far easier to predict than receiving touchdowns for Vegas. It can be hypothesized that this will lead to more profits from receiving touchdowns since the model appears to be superior to that of Vegas. These are also just the general predictions and have no indication of the odds comparisons of individual players, which could alter the outcome of profits as well.

After these comparisons, a brier score was calculated for both models. This measures the accuracy of probabilistic predictions, with values ranging between 0 and 1. A lower Brier score indicates better predictions, for a binary outcome we would expect a 0.25 score for random guessing. The receiving model had a value of 0.183 and the rushing model had a value of 0.167. Both indicate a decent amount better than random guessing. To further investigate the accuracy of each model, an ROC (receiver operator characteristic) curve and the area under curve was found. The ROC uses the true positive rate (the proportion of actual positives correctly predicted as positive) and false positive rate (the proportion of actual negatives incorrectly predicted as positive) to see how the model is working. An ideal model would have a curve with the maximum area under the curve as it would have 100% true positive and 0% false positive and a bad model would have 0.5 area under the curve as it would be random guessing and hit half the time. The curve for receiving can be seen below on the left and rushing on the right.



Again, the rushing model appears a bit better, but both have above average predictive accuracy. Between 0.7 and 0.8 are generally good for AUC (area under curve) and for these models the AUC measurement was 0.7013 for the receiving model and 0.7820 for the rushing model. This indicates that both are good predictors but could still be better.

2024 Bet Tracking

Now that the general models' accuracy have been measured, the economic success of implementing the models will be tracked. As stated previously, whichever players have predicted probabilities from the model higher than that of the sports books implied odds will be bet on. Determining how much to bet is the next step in the process. One method is to just bet a standard unit on each player the model sees value in. The next is to use whatever the model has as an edge for each player as well as the player's predicted probability. This

will lead to players with larger gaps between model and book as well as players with generally higher odds to be bet on more. The reason being that it is more logical to bet on things that are more likely to happen even if the payout is lower. Also, if the model sees a more extreme difference in odds it will want to capitalize on how much the book is undervaluing that player. This type of bet is called using the Kelly criterion and the formula for it is $f = \frac{bp-q}{b}$, where f is the fraction of a chosen bankroll you will bet on a player. The other variables are; the predicted probability of winning (p), the predicted probability of losing (q), and the decimal odds received on the bet (b). This considers the predicted probability of winning along with the size of the edge the model sees over the book's odds. Both this method and a standard unit bet were tracked over weeks 13, 14, and 15 of the 2024 NFL seasons and the results can be seen below. Rushing profits will be analyzed first and then receiving profits analyzed below that.

Rushing	Kelly Bet Amount	Kelly Profit	Kelly Return	Standard Bet Amount	Standard Profit	Standard Return
Week 13	$901	$341.80	37.93%	$1,050.00	$345.56	32.91%
Week 14	$489	$376.07	76.98%	$650.00	$246.49	37.92%
Week 15	$867	-$22.32	-2.57%	$1,000	-$73.28	-7.33%
Total	$2,257	$695.55	30.82%	$2,700.00	$518.77	19.21%

After 3 weeks of tracking the rushing model showed quite good returns. The results were a 30.82% return over the 3 weeks using the Kelly Criterion to bet while the standard bet model gave a 19.21% return. It can be seen both had ups and downs with the first two weeks being more profitable and a down week during week 15. Something else noticeable is far fewer players were bet on during week 14, there were more bye weeks which is likely the cause, but the drop off seems quite significant. Also, something to note is in some cases the extreme randomness of the scorers. For example, in week 15, the only week with negative profit, Johnathon Taylor had one of the highest edges and was bet on the 2nd highest amount of the week. This is noted because he scored in week 15, but after booth review, it was found that he dropped the ball inches before entering the endzone. Not only did this cost the colts a touchdown, but had his touchdown counted the week's profit would have been positive for both the standard betting and the Kelly criterion betting. All that to say that no matter how much numbers play a role in who scores, they are just people playing a game at the end of the day and there will always be random events impacting the results.

Recieving	Kelly Bet Amount	Kelly Profit	Kelly Return	Standard Bet Amount	Standard Profit	Standard Return
Week 13	$962	-$261.02	-27.13%	$1,250.00	$668.48	53.48%
Week 14	$946	-$343.36	-36.31%	$1,200.00	-$185.40	-15.45%
Week 15	$1,472	$774.03	52.59%	$1,750.00	$662.62	37.86%
Total	$3,380	$169.65	5.02%	$4,200.00	$1,145.70	27.28%

The receiving model has some more notable fluctuations and differences between the standard bet versus the Kelly bet. Again, both after the 3 weeks showed profits, however on the contrary to the rushing model, the receiving model had higher profits with the standard bets at 27.28% return compared to the Kelly return of 5.02%. The reasoning behind this is not known for certain but might be due to the higher variation in receiving touchdown scorers. It makes sense as running backs there are usually only 1 or maybe 2 potential scorers on a team. There are usually 4 or 5 potential receiving touchdown scorers on each team which causes them to be much harder to predict. This can also be seen by the earlier plots showing Vegas predictions as the running backs appeared far easier for them to predict. The reasoning behind the higher profits for the standard model versus Kelly model is likely due to the lower odds players being bet on significantly less. Receivers generally have lower chances to score than running backs and their odds reflect that. When the model likes a big underdog, it will bet close to nothing on them unless the model sees an extreme edge. This leads to little profit even if the player scores. On the contrary, the standard model could hit on a singular big underdog and pay off losses on numerous other players that failed to score. This can be seen clearly in week 13. A significant loss for the Kelly model saw the largest percentage return by the standard model. This was due to numerous big favorites not scoring while just a couple big underdogs did score. The losses were more than made up for in the standard model, but the Kelly model barely saw any winnings from the underdogs. This wasn't really an issue for the running backs since many of the bigger favorites scored and it is less likely for huge underdogs to be bet on or to score. Based on these results, it appears that for the running backs betting using the Kelly Criterion is smartest while betting using the standard unit bet is better for betting on the receiving touchdowns. Future weeks will continue to be tracked, and this conclusion may or may not be modified depending on the results.

Conclusions:

The findings of this paper can be summarized in three main points. First, based on a large variety of variables for both rushing and receiving categories in the 2023 NFL season, there are a few that stand out in terms of best predicting touchdowns. For rushing touchdowns, the most influential variables were found to be attempts within the 5-yard line, rushing yards gained within the 5-yard line and rushing yards gained within the 20-yard line.

For receiving touchdowns, the most influential variables are receiving yards per game, targets within the 20 yard-line and receptions within the 10 yard-line. Since these variables are the most significantly related to touchdowns, they are the variables that will be used to predict touchdowns for the 2024 NFL season.

When using these variables, it was found that over a 3-week span both the rushing and receiving models were significantly accurate in predicting touchdown scorers. With significant positive slopes in the logistic regression along with significantly low brier scores and high areas under curve with the ROC curve, the success of the models cannot be ignored. Additionally, when looking at the plots with the models along with Vegas odds, Vegas as well as the Model are worse at predicting receiving touchdowns than rushing. Relatively however, the model is superior in terms of receiving touchdown predictions and Vegas was superior in terms of predicting rushing touchdowns. The ability to simply predict players to score is not the only goal of this model, however.

When using the model to track profits over time when betting on players to score it is again clear the model was significantly effective. When tracking for betting standard units over time as well as an amount using the Kelly criterion it was found that both methods for both models brought a profit. For rushing touchdowns, the Kelly bets resulted in a 30.82% return while the standard bet resulted in a 19.21% return. This makes sense due to the higher likelihood of big favorites scoring as opposed to the players with far longer odds. On the other hand, the receiving touchdowns saw only a 5.02% profit using the Kelly criterion bets and the standard unit betting gave a 27.28% return. Having a positive return at all is a good sign, but percentages close to 30 are incredible. With only 3 weeks tracked, it could see negative correction in future weeks, but the current results are extremely promising. The fact that every model resulted in profit in itself is a great sign. Though based on current results it might point to focusing bets on the standard bets on receivers and Kelly criterion bets on running backs.

Future Work

One of the biggest factors not accounted for in the model is the opposing defenses. The defenses faced are clearly a factor in the NFL when it comes to touchdown scorers and if you look at sports books odds for touchdown scorers it is clear they are adjusting for them which indicates it is probably something that should be accounted for. To do this there are a few options. First you could do it based on the ranking of the defense based on type of touchdown allowed and then scale the predicted probabilities based on the opposition rank that the player is facing that week. Another option would be to take the percentage above or

below the mean touchdowns allowed for the league and use that value to then scale the predicted probability. There are probably more ways that defense could be accounted for, and some might be better or worse than others but incorporating some factor for opposing defense would more than likely increase the model's accuracy. On top of that adding indicators for weather would likely help the model as well. A game in week 15 had lots of rain which commonly leads to games being more run heavy due to the ball being harder to catch. This leads to less receiving touchdowns. In this game the model bet 5 receivers to score touchdowns and not only did none of them score, but no one scored in the entire game. Obviously, this could be the result of random variation, and the weather could have been a non-factor, but had someone done an analysis I am sure games in rain and snow would show that less touchdowns will be scored and adding a scalar to reduce the predicted probabilities in those games would likely be helpful. Additionally, the models currently keep the receiving and rushing touchdowns completely separate, but when you bet on a touchdown scorer, they can cash the bet by doing either. To correct this the probabilities must be added together in a way that combines the probabilities but does not double count the probability of the player scoring a rushing touchdown and receiving touchdown in the same game. This would be easy enough to do but was simply overlooked during the project to this point. This has likely led to certain running backs being under bet or not bet at all since they might receive a large boost in their chance of scoring if the receiving touchdown odds were added in. Injuries are still a concern, but dealing with these is a bit more complex. Future work could include accounting for players who were injured in predictions or even altering player predictions accounting for current players who are injured. For example, David Montgomery is likely not going to play the rest of the season. His teammate Jahmyr Gibbs will now surely be given much worse odds by the books due to his increased likelihood of scoring. The model will still predict his probability based on the stats so far this season and will likely not see value in him even though there very well could be. Finally, accounting for player development or aging through the season. Certain young players, especially rookies start off slower and bloom later in the season. As the model is based on the entire season statistics it will weigh the earlier games in the season the same as the more recent ones. This would cause the model to overlook value in the young players who have broken out. The same could happen with older players being replaced by the new players as they would have reduced roles, but the model would not account for it. This would potentially lead to the model seeing value in these older players who have been replaced and are not actually as likely to score as the model thinks. To aid these weighting recent games more heavily might help. Some of these might not actually result in model improvement and could just be misconceptions sports fans might have, but it is at least worth looking into them all in case they could provide model enhancement.

Appendix:

Data Sources:

https://www.fantasypros.com/nfl/reports/snap-count-analysis/

Data for 2023 and 2024 redzone statistics and other weekly and seasonal data

https://www.pro-football-reference.com/

More data for 2023 and 2024 statistics and other weekly and seasonal data

https://www.pro-football-reference.com/

Data on individual player odds for scoring a touchdown in any given week

Link to Bet Tracking Spreadsheet:

https://docs.google.com/spreadsheets/d/1D16bG69Occc4LaBUGH5Wl0Sk4EHtWmK0Q2CbgnXk7dg/edit?usp=sharing

R-Output for Models:

Original 2023 Receiving Model and VIF:

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.8892 -0.7480  0.0167  0.5943  3.5947

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.36475    0.32770  -1.113    0.270
tgtRzn       0.29928    0.05329   5.616 6.35e-07 ***
recYG        0.03483    0.01430   2.436    0.018 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.362 on 56 degrees of freedom
Multiple R-squared:    0.7,     Adjusted R-squared:  0.6893
F-statistic: 65.33 on 2 and 56 DF,  p-value: 2.291e-15

> vif(selected_model)
  tgtRzn    recYG
2.256658 2.256658
```

Original 2023 Rushing Model and VIF:

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.7903 -0.9997 -0.3578  0.6241 11.5316

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.328682   0.266580  -1.233   0.2192
Att          0.020550   0.004394   4.677 5.76e-06 ***
rushYG       0.049221   0.015609   3.153   0.0019 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.269 on 177 degrees of freedom
Multiple R-squared:  0.5551,    Adjusted R-squared:  0.5501
F-statistic: 110.4 on 2 and 177 DF,  p-value: < 2.2e-16

> vif(selected_model)
     Att   rushYG
3.860431 3.860431
```

Updated 2023 Receiving Model with Redzone Statistics and VIF:

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.2947 -0.6671 -0.0410  0.5959  6.5981

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.287844   0.098027  -2.936 0.003542 **
Tgt20        0.101765   0.028715   3.544 0.000448 ***
recYG        0.042853   0.004699   9.120  < 2e-16 ***
Rec10        0.525636   0.069925   7.517 4.79e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.221 on 348 degrees of freedom
Multiple R-squared:  0.7607,    Adjusted R-squared:  0.7586
F-statistic: 368.8 on 3 and 348 DF,  p-value: < 2.2e-16

> vif(selected_model_rec)
   Tgt20    recYG    Rec10
5.446400 2.575151 3.425071
```

Updated 2023 Rushing Model with Redzone Statistics and VIF:

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.5896 -0.6591 -0.1731  0.5490  5.6632

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.145294   0.158505   0.917    0.361
Att5        0.292186   0.052538   5.561 1.15e-07 ***
Yds20       0.027789   0.004763   5.834 3.06e-08 ***
Yds5        0.174044   0.032561   5.345 3.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.442 on 155 degrees of freedom
Multiple R-squared:  0.8318,    Adjusted R-squared:  0.8285
F-statistic: 255.4 on 3 and 155 DF,  p-value: < 2.2e-16

> vif(selected_model)
    Att5    Yds20     Yds5
3.954949 3.097334 2.663254
```

Heads of the Week 12 predicted probabilities for both types of touchdowns:

	Player	Tgt20G	recYG	Rec10G	recTDG	predicted_prob
1	Amon-Ra St. Brown	1.4000000	68.50000	0.90000000	0.90000000	0.8753232
2	George Kittle	1.6250000	70.00000	0.87500000	0.87500000	0.8715062
3	Ja'Marr Chase	1.6363636	96.00000	0.45454545	1.00000000	0.7845006
4	Justin Jefferson	1.3000000	91.20000	0.30000000	0.50000000	0.6519052
5	Travis Kelce	1.3000000	50.70000	0.60000000	0.20000000	0.6020486
6	Stefon Diggs	0.6250000	62.00000	0.50000000	0.37500000	0.6005261
7	Malik Nabers	0.6250000	75.87500	0.37500000	0.37500000	0.6002979
8	Ladd McConkey	0.6000000	61.50000	0.50000000	0.40000000	0.5968242
9	Garrett Wilson	1.4545455	65.63636	0.45454545	0.45454545	0.5937118

	Player	Yds20G	Att5G	Yds5G	rushTDG	predicted_prob
1	Saquon Barkley	12.70000000	1.10000000	1.60000000	0.80000000	0.94496405
2	David Montgomery	11.30000000	1.20000000	1.50000000	1.00000000	0.93575967
3	Kyren Williams	10.80000000	1.20000000	1.60000000	0.80000000	0.93475855
4	Joe Mixon	11.87500000	1.12500000	1.37500000	1.00000000	0.92770176
5	Derrick Henry	8.45454545	1.27272727	1.45454545	1.00000000	0.90290827
6	Jalen Hurts	10.70000000	1.20000000	0.90000000	1.00000000	0.88960454
7	Rhamondre Stevenson	5.20000000	1.00000000	1.50000000	0.60000000	0.75764216
8	Jonathan Taylor	6.25000000	1.25000000	0.62500000	0.62500000	0.75131929
9	James Cook	10.30000000	0.70000000	0.90000000	1.00000000	0.74850932

Logistic Regression Tracking success of both models through 3 weeks:

```
glm(formula = Scored ~ predicted_prob, family = "binomial"
415rec)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.7786     0.2145  -8.290  < 2e-16
predicted_prob   3.0438     0.7009   4.343 1.41e-05

(Intercept)    ***
predicted_prob ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 378.90  on 333  degrees of freedom
Residual deviance: 359.26  on 332  degrees of freedom
AIC: 363.26

Number of Fisher Scoring iterations: 4
```

```
glm(formula = Scored ~ predicted_prob, family = "binomial"
415rush)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.1815     0.2853  -7.647 2.06e-14
predicted_prob   3.0225     0.6028   5.014 5.32e-07

(Intercept)    ***
predicted_prob ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 223.14  on 204  degrees of freedom
Residual deviance: 196.22  on 203  degrees of freedom
AIC: 200.22

Number of Fisher Scoring iterations: 4
```