Examining the Imapct of Weather on NFL Decision-Making

Jeff Du, Sahil Tilak

2023-12-17

Introduction

NFL head coaches are constantly under scrutiny, whether warranted or unwarranted, for the decisions made over the course of the game that ultimately lead to a win or loss. Coaches have a variety of options at their disposal depending on the current condition and momentum of the game, and most of their decision-making stems from a primarily objective examination of their team's needs. Third and 10 down 20 points in the 4th quarter? In all likelihood a pass play to try and mount a comeback. Fourth and 20 up 10? Most likely a punt to minimize the risk of giving good field position to the opponent. Coaches, players, and fans can on the whole agree on the right play-call in many hypothetical game scenarios.

However, every coach suffers from inherent biases that have the potential to impair their judgment and decision-making when baselined against a baseline of rationality. For example, the hot-hand fallacy has been extensively studied across varying sports (though primarily in basketball), with little statistical evidence to support the notion that athletes experiencing positive outcomes are more likely to experience it repeatedly. Yet in practice, coaches will often apply decision-making that has worked in the past in new situations despite adverse conditions for success. Entrusted with managing numerous responsibilities throughout the game, coaches must rely on heuristics and past anecdotal experiences to guide them throughout the game, resulting in these observed biases.

Key considerations that consistently affect game planning and subsequently decision-making are the game's weather conditions. For stadiums that have an open roof (there are 22 outdoor stadiums and an additional 5 domed stadiums with retractable roofs), games are subject to the effects of the surrounding climate: temperature, humidity, wind speed, and precipitation. In contrast, closed stadiums fine-tune and control for climate, reducing the influence of the aforementioned conditions to essentially 0.

All of these weather measures have the potential to affect the probability of success for any given play. High wind speeds can affect both the trajectory and velocity of passes, potentially lowering the probability of a completion, while high humidity results in less dense air that allows a football to travel farther, potentially increasing the probability of a field goal conversion. Unfortunately, there has been minimal existing literature exploring the effects of temperature on NFL games, and most of that has been amateur-led and conducted with a primary focus on exploiting sports betting inefficiencies.

Entering the game, coaches have generally prepared for the climate conditions and have adjusted their game plans accordingly. Yet over the course of the game, numerous scenarios emerge that warrant a particular play whose success rate may be hindered as a result of the weather. In these situations, how does a coach assess the gravity of weather? We investigate this phenomenon through three main research questions:

1. For non fourth down plays in which the coach predominantly decides to run or pass, do the weather conditions impact the coach's ultimate decision-making?
2. Similarly, for fourth down plays in which the decision tree expands to include punting and kicking a field goal, do the weather conditions play any role in influencing the coach's ultimate decision-making?
3. For both non fourth down and fourth down plays, do weather conditions have any impact on the probability of a successful outcome (first down, touchdown, field goal made, punt that pins opponent deep in their territory)?

We hope that our findings will be informative in uncovering any systematic bias in either the overestimation or underestimation of weather conditions by NFL coaches, allowing them to tailor their play calling in a way that best positions their team for success.

Proposed Methodology

We divide our analysis into non fourth down and fourth down situations because each scenario considers a different set of options (run and pass on non fourth downs, with the added optionality of kicking a field goal or punting on fourth downs). Because there are so many potential variables that affect the decision to run pass, punt, or kick, we will use gradient boosting to fit decision trees predicting the end play call of a particular situation in order to gauge feature importance and see if weather conditions appear to be "important" in predicting a coach's decision-making. From there, we will fit two logistic regression models with the features we deemed relevant, with the response variables being outcomes of success for the particular down (first down and touchdown for non fourth down plays, and if the team in possession eventually won the game for fourth down plays). We examine if the weather predictors are statistically significant.

Data Description

We use two datasets in our analysis. The first dataset, obtained from Kaggle, consists of hourly weather observations recorded from Meteostat and Weather Underground for games from the 2000-2001 NFL season to the 2020-2021 NFL season. In total, there are 37,155 hourly observations. Each observation relevant weather metrics such as temperature (in Fahrenheit), dew point, humidity (0 - 100), precipitation (mm), wind speed (MPH), and estimated condition. The estimated condition is qualitatively described, with levels such as clear, heavy rain, moderate rain, light rain, light snow, and moderate snow.

The second dataset, obtained from the nflverse R package, consists of play-by-play data for every NFL game since the 1999-2000 NFL season. For our analysis, we will only use data from the 2018-2019 to 2020-2021 seasons, in order to align our timeframes with the weather observations dataset. In total, there were 144,422 plays over these three NFL seasons. In addition to play-specific characteristics such as down, yard line, time remaining in quarter, score of game, play type, and result of play, each play also records the home and away team and the type of stadium in which the game is being played in (outdoors, dome, closed stadium, and open stadium).

Although the play-by-play dataset does contain variables describing the weather conditions such as temperature and wind speed, it is play-agnostic and therefore does not change over the course of the game. This is problematic because we cannot assume that weather conditions remain constant as the day progresses. Therefore, we cannot rely solely on the play-by-play dataset and must incorporate the hourly observations dataset. Both datasets share a common game ID variable and similar time of day measurements, so after initial data cleaning (rounding the time of day measurements in the play-by-play dataset to the nearest hour to match the format of the hourly weather observations, which occur at the start of each hour), we are able to merge the two datasets, and after filtering out plays that do not have a specified down (kickoffs, extra points, timeouts), we are left with 112,381 plays with hourly weather observations.

EDA

We first explore the significant of the 112,381 with respect to the current game situation. Our dataset contains a win probability variable that describes the estimated win probability for the team in possession given the current game conditions at the start of the play. Here we plot the density plot of win probabilities.

We observe a trimodal distribution with peaks occurring near a win probability of 0, 50, and 100. We are not particularly interested in plays occurring in game situations where the outcome is essentially decided, and so we filter out plays that occurred in situations where the win probability was less than .10 or greater than .90. We are now left with 84,890 plays.

From these plays, we consider how playcalling on first, second, and third downs vary versus fourth downs. We display the number of occurrences of each play type among non fourth down plays versus fourth down plays from our training datasets.

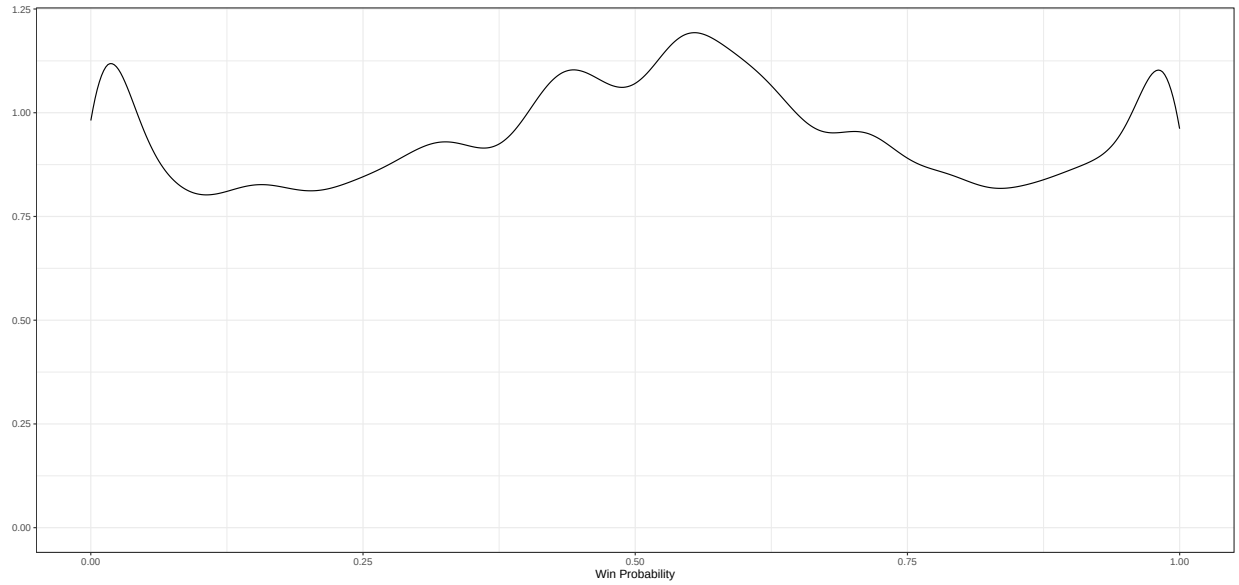


Figure 1: Distribution of Win Probability on Plays

First, for non fourth down plays:

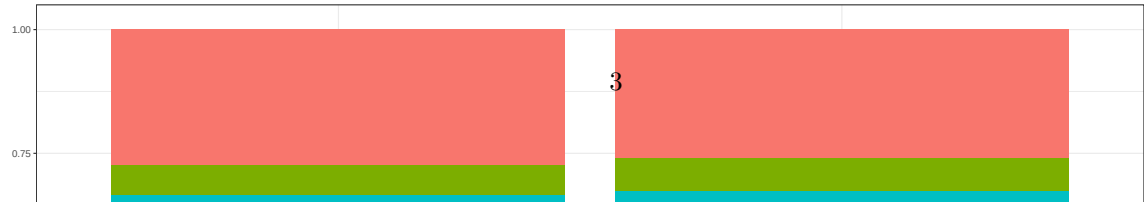
Play Type	Frequency
Field Goal	185
No Play	4904
Pass	42054
QB Kneel	217
QB Spike	147
Run	29149

Then, for fourth down plays:

Play Type	Frequency
Field Goal	2067
No Play	367
Pass	509
Punt	4825
Run	450

As expected, coaches will never punt and only kick field goals near the end of a half on non fourth downs, while prioritizing punts and field goals on fourth down. Therefore, for our non-fourth-down analysis, we will only be concerned with predicting whether a play is a run or pass (filtering out non-run and non-pass plays, or a binary categorical response), and for our fourth down analysis, we will focus on predicting four outcomes: punt, kick, run, and pass.

Finally, let's consider whether a stadium having an open versus closed roof affects the distribution of 4th down play outcomes. The play-by-play dataset has four possible outcomes for the roof variable: outdoors, open, closed, and dome. We will consider a domed stadium that has an open roof for the particular game to be equivalent to an outdoors stadium and a domed stadium that has a closed roof for the particular game to be equivalent to a non-retractable dome stadium.



We see that although the divisions of punt, field goal, run, or pass are relatively similar for both outdoor and indoor stadiums, a higher percentage of 4th down plays are field goals in domed stadiums than outdoor stadiums, suggesting that coaches are cognizant of the effects of weather on effective field-goal kicking.

Train Test Split

We will use plays from the 2018-2019 and 2019-2020 seasons as our training set (51226 non-fourth-down plays and 5616 fourth down plays) and plays from the 2020-2021 season (25430 non-fourth-down plays and 2618 fourth down plays) as our test set. This roughly corresponds to a 2/3 train-test split for both non-fourth-down and fourth down plays.

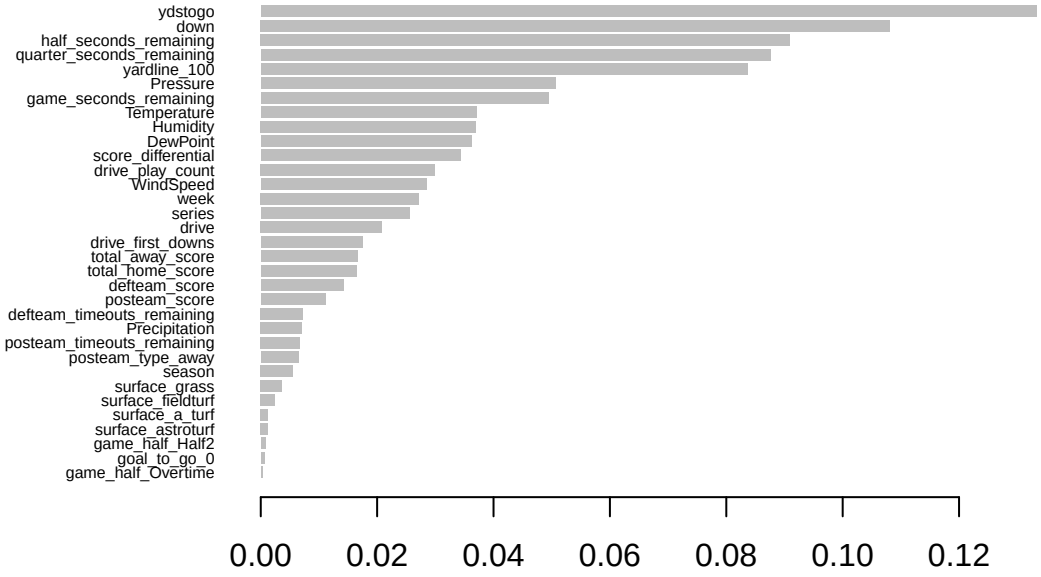
Non-Fourth Down Analysis

Training

We first analyze non fourth down playcalling, with a particular emphasis on deciding whether to run or pass. As we are interested in examining the weather bias, we are only interested in examining plays occurring in an open roof setting. Applying these filters, we are left with 36,036 observations.

With 383 variables in the training dataset, including many that describe post-play statistics (yards gained, turnover, first down conversion, etc.), we perform an initial qualitative variable selection process, selecting 30 variables that we believe coaches could potentially consider before deciding which play to call. These range from numeric variables such as down, yard line, yards to go, and seconds remaining in half to categorical variables like type of stadium, type of surface, and game half, but all are game conditions pre-determined before the play. See appendix for full list of variables.

To gauge whether such variables appear to be the most influential in predicting run or pass plays, we use XgBoost to fit gradient-boosted decision trees with our label being the play type. Out of these 30 variables, 5 are categorical, so we perform preprocessing through one-hot-encoding. We use a max tree depth of 10 and 100 rounds. Here we plot the importance feature graph from the model.



Our training log loss was .22. From the feature importance graph, we find that Pressure, Humidity, DewPoint, and Temperature have some of the highest feature scores (as quantified by Gain), indicating that such conditions are influential for predicting a coach's decision to run or pass. But do such weather conditions have a statistically significant relationship to the outcome of the particular play?

We consider a successful outcome on non-fourth down plays to be either a first down conversion or a touchdown. Therefore, we proceed with fitting a logistic regression model with the response being a binary variable that records whether or not the play satisfied the aforementioned conditions. We fit the model with all predictors that recorded a Gain metric above .01 in our gradient-boosted model, with an exception for precipitation as we want to conduct inference on all weather predictors. We now include play type as a predictor to control for differences in run and pass plays. Here we display the full model:

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \sum_{j=1}^{6} \beta_j X_j + \beta_7 \text{Run} + \sum_{j=8}^{23} \beta_j X_j$$

In our regression, the pass play type is the base group.

$\beta_1 ... \beta_6$ indicate the 6 weather related variables included in the model: Pressure, Temperature, Wind Speed, Humidity, Precipitation, and Dew Point

$\sum_{j=8}^{23} \beta_j X_j$ indicates all predictor variables not related to play type or weather that were significant from the tree-based model (i.e. Yard Line, Yards to Go for a First Down, Time Remaining in the Half, etc).

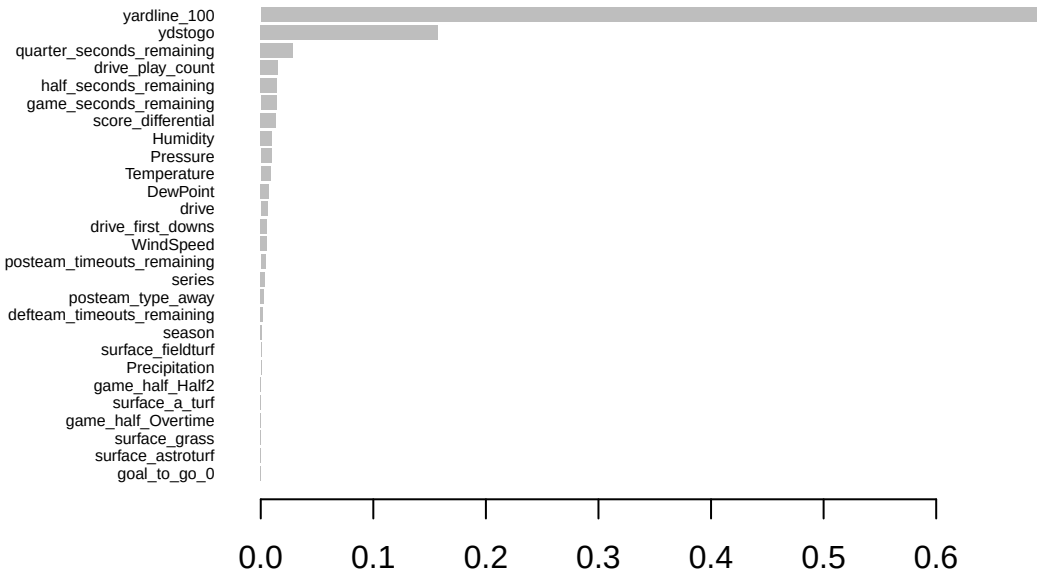
	Estimate	Standard Error	Z Value	Pr(>|z|)
Pressure	0.151	0.0868	1.74	0.0822
Humidity	-0.00551	0.00372	-1.48	0.139
DewPoint	0.00749	0.00723	1.04	0.3

	Estimate	Standard Error	Z Value	Pr(>|z|)
Temperature	-0.00795	0.007	-1.14	0.256
WindSpeed	-0.00756	0.00343	-2.2	0.0277
Precipitation	1.46	0.779	1.88	0.06

At an alpha level of .05, we observe that pressure, humidity, dewpoint, temperature, and precipitation are not statistically significant, though wind speed is, even after controlling for the type of play in the particular game scenario. There is a disconnect between the perceived influence of weather on a coach's decision to run or pass and the impact of weather on successful play outcomes.

Fourth Down

Now, we want to look at the other aspect of fourth down play-calling: fourth down decision-making. At each fourth down event, an offense has three decisions that can be made: go for it (i.e. call a run or pass play to try and get a first down), punt, or kick a field goal. First, we run a similar model where we determine the features that are important, and then use those features to implement a multinomial logistic regression model. Fourth down decision-making is more clear cut than run/pass decision-making, so we will filter the dataset to only include plays where outcomes may be more ambiguous (i.e. not going for fourth down on fourth and 20).



From the feature importance graph, we find that Humidity, Pressure, and Temperature are among the 10 have some of the highest feature scores (as quantified by Gain), but the predominant variable in terms of feature importance is the current yard line position and yards to go. This intuitively makes sense, because teams generally only go for it when they are in the opponent's territory.

Now, we run a logistic regression model to determine if weather has a significant effect on play outcomes.

Instead of using series success rate as the response variable for this model, we choose to use the result of the game (1 if the possession team wins and 0 if the possession team loses) as a proxy for measuring success of plays. We fit the model with all predictors that recorded a Gain metric above .005 in our gradient-boosted model, with an exception for precipitation as we want to conduct inference on all weather predictors. Below is the full model:

$$Pr(\text{Outcome}_i = \text{Win}) = \pi_i$$

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \sum_{j=1}^{6} \beta_j X_j + \beta_7 \text{Run} + \beta_8 \text{Pass} + \beta_9 \text{Punt} + \sum_{j=10}^{18} \beta_j X_j$$

$\beta_1 ... \beta_6$ indicate the 6 weather related variables included in the model: Pressure, Temperature, Wind Speed, Humidity, Precipitation, and Dew Point

$\beta_7 ... \beta_9$ indicate the the play types for which we are predicting outcomes. In our regression, the field goal play type is the base group.

$\sum_{j=10}^{18} \beta_j X_j$ indicates all predictor variables not related to play type or weather that were significant from the tree-based model (i.e. Yard Line, Yards to Go for a First Down, Time Remaining in the Half, etc).

	Estimate	Standard Error	Z Value	Pr(>|z|)
Pressure	-0.129	0.177	-0.731	0.465
Humidity	0.00294	0.0075	0.393	0.695
DewPoint	-0.0082	0.0144	-0.567	0.57
Temperature	0.0078	0.0137	0.568	0.57
WindSpeed	-0.00711	0.00711	-1	0.317
Precipitation	-0.709	1.66	-0.428	0.669

At an alpha level of .05, we observe that none of the weather variables: pressure, humidity, dewpoint, temperature, and precipitation are statistically significant, even after controlling for the type of play in the particular game scenario. There is a disconnect between the perceived influence of weather on a coach's decision to run or pass and the impact of weather on successful play outcomes. We determine that weather does not particularly drive a coach's decision-making on fourth down, and that the current game circumstances (yards to go and yardline) predominantly drive decision-making.

Conclusion

Through the lens of two different game scenarios (runs/passes in early down situations, fourth downs), we were able to assess the validity of the two hypothesis that we made. First, we see that our hypothesis about weather impacting coaches was true. Based on the results of our tree models, there are weather variables that are predictors of play outcomes, both in the run/pass model and the fourth-down model. However, our second hypothesis, that weather would be a predictor of play success was not true. Temperature predictors were not significant predictors of success in both models. These takeaways have noteable impacts from different perspectives. The first is from the perspective of an NFL coach. It seems that coaches do take into account weather in decision-making. Based on the results of our logistic regression models, this is mis-informed, as weather has no significant impact on how impactful a play is. Therefore, coaches are likely making sub-optimal decision-making in the face of different weather conditions. As the NFL is the highest level of football on the world, coaches need to find every opportunity they can to maximize win probability, and weather-related decision-making is one area where they are failing. The second impact is from the perspective of a sports bettor. These models can allow a sports bettor to find the scenarios in which weather will have an impact on predicted play probabilities. Since we have established that these are the scenarios in which a coach is likely to make a sub-optimal decision, it opens opportunities for a bettor to bet against the coach on the team that they are playing. Of course, determining whether or not this is a legitimate strategy would take back-testing,

but the results of our models would certaintly catch the eye of any bettor looking to take advantage of mispricing of betting odds.

There are also avenues for future research in this area. In both logistic regression models, we defined the idea of play efficiency using different parameters - "Sucess", or getting a first down/touchdown on a play for our non-fourth down model, and whether or not the team ends up winning or losing for the fourth down model. These are both proxies of efficiency. In reality, there is no perfect way to determine how efficient a play is, but there are avenues a researcher could take. One way is to create a win probability model that measures probability of winning before a play and after a play, and use the difference in win probability before and after as a measurement of the efficiency of a play.

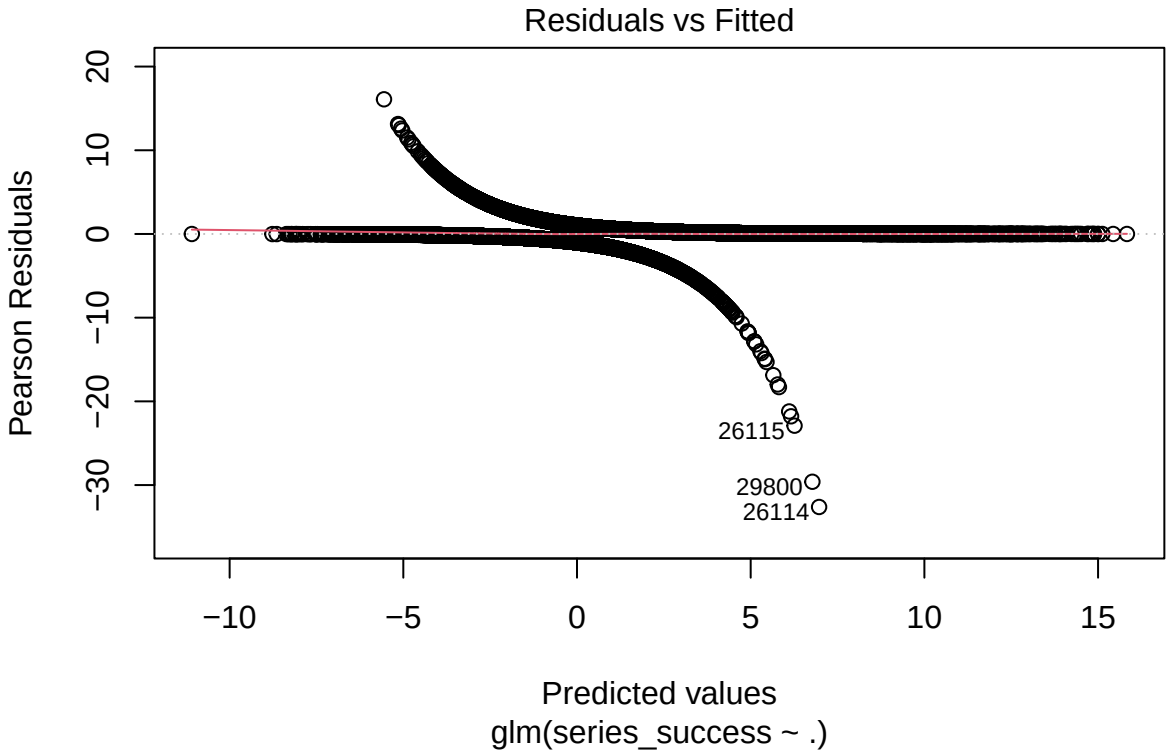
Appendix

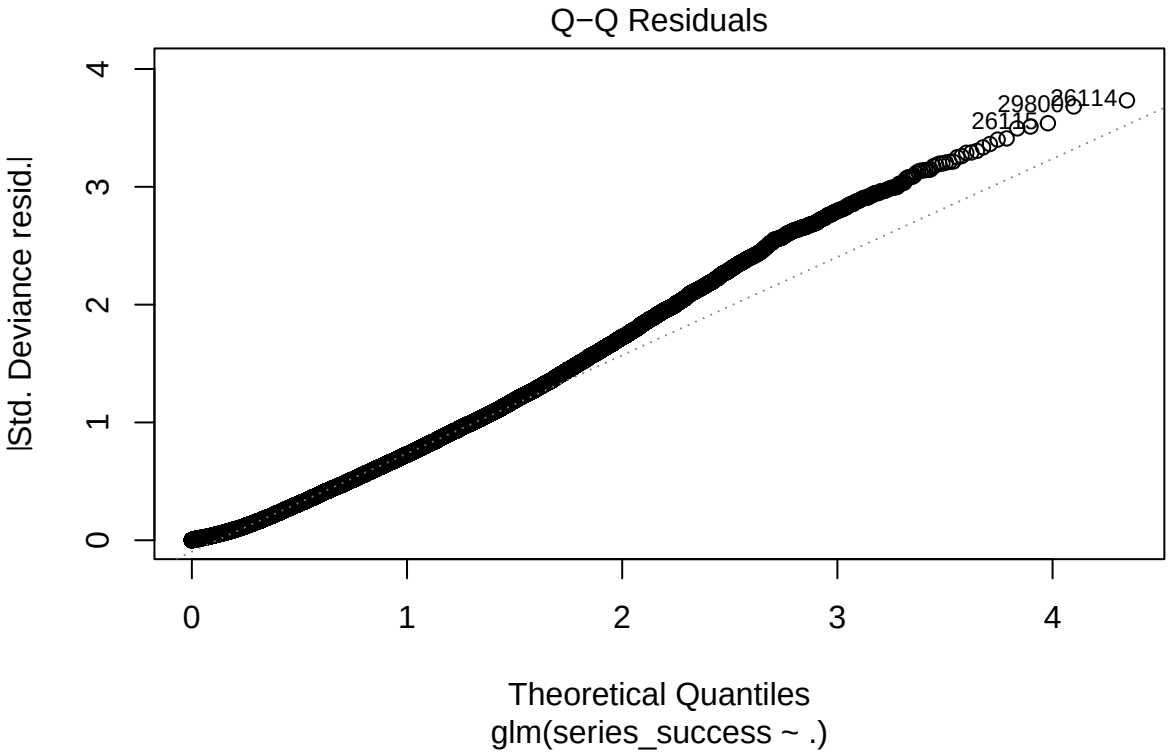
Testing

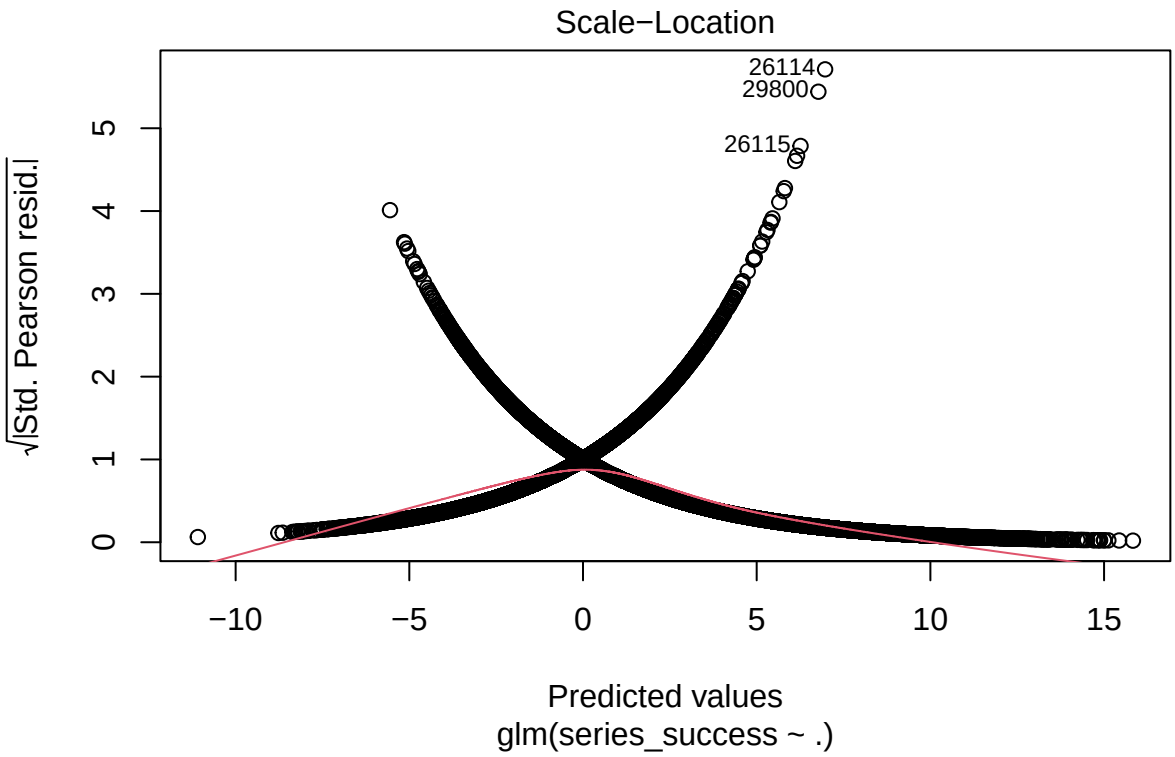
We use the same gradient-boosted model on our test dataset of 2020 plays. First, we plot the confusion matrix for non fourth down plays, where the error rate is 35%

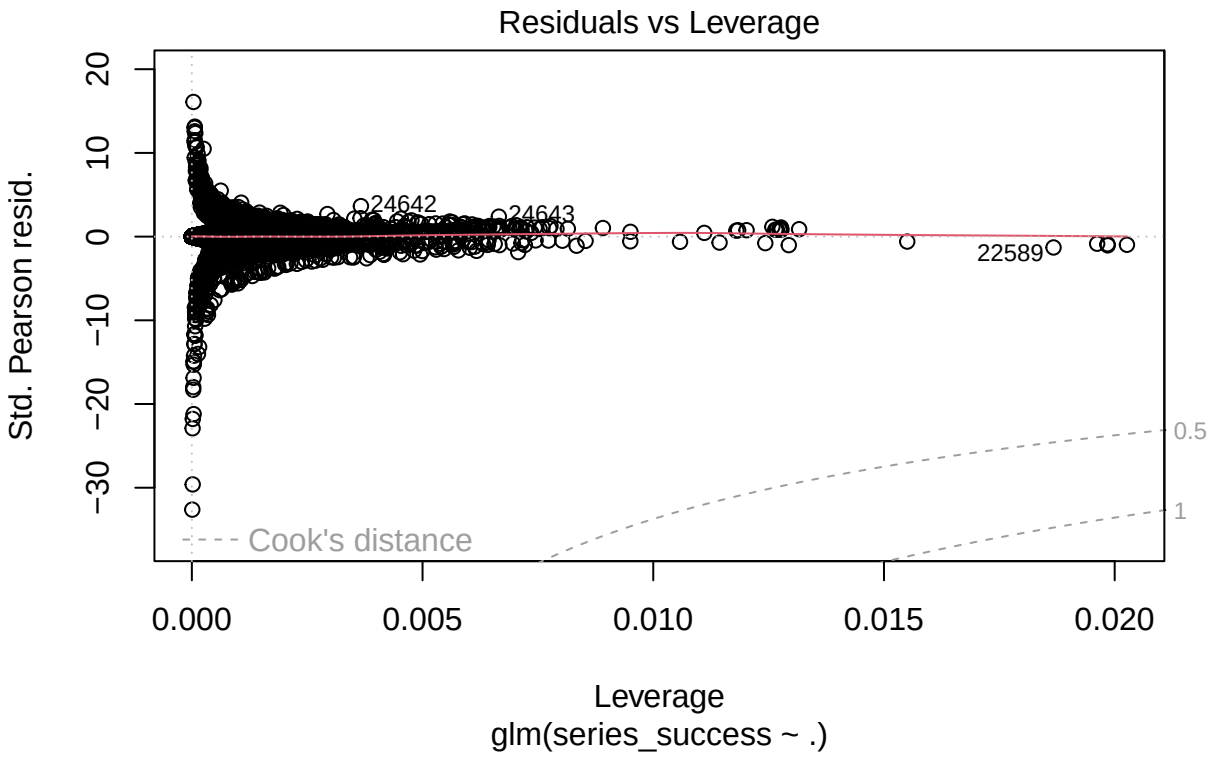
Here.

Model Diagnostics: Run/Pass Model on Non-Fourth Down Plays

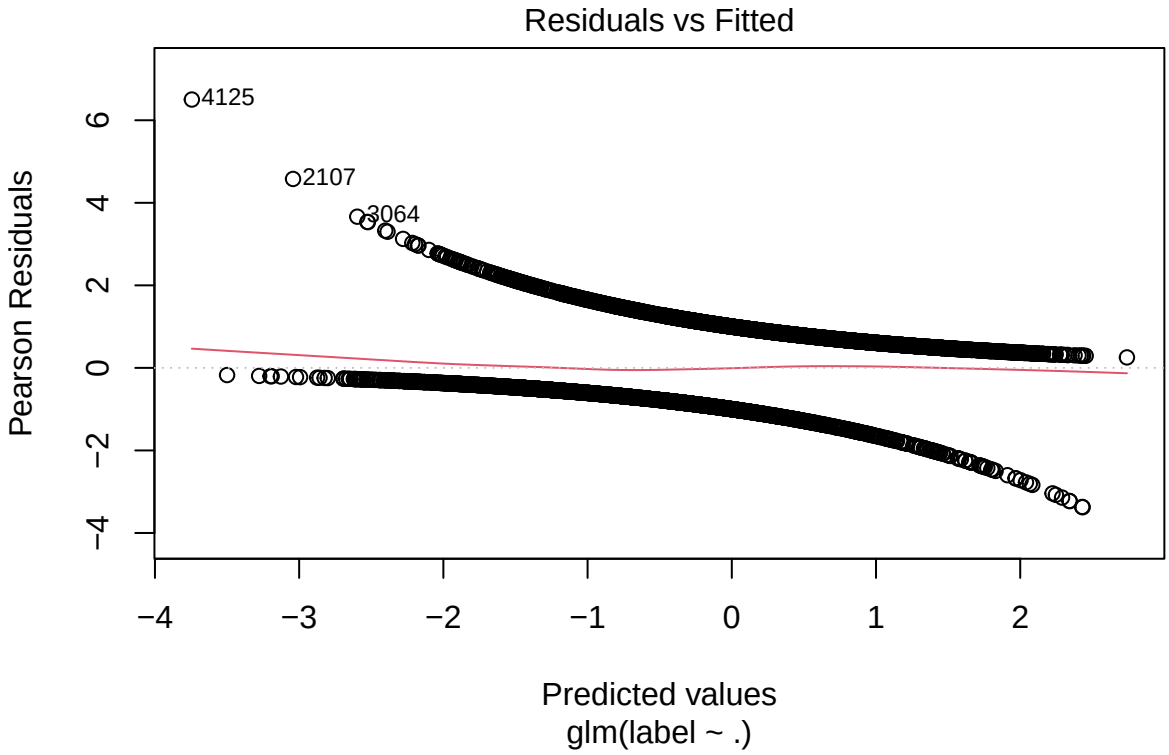


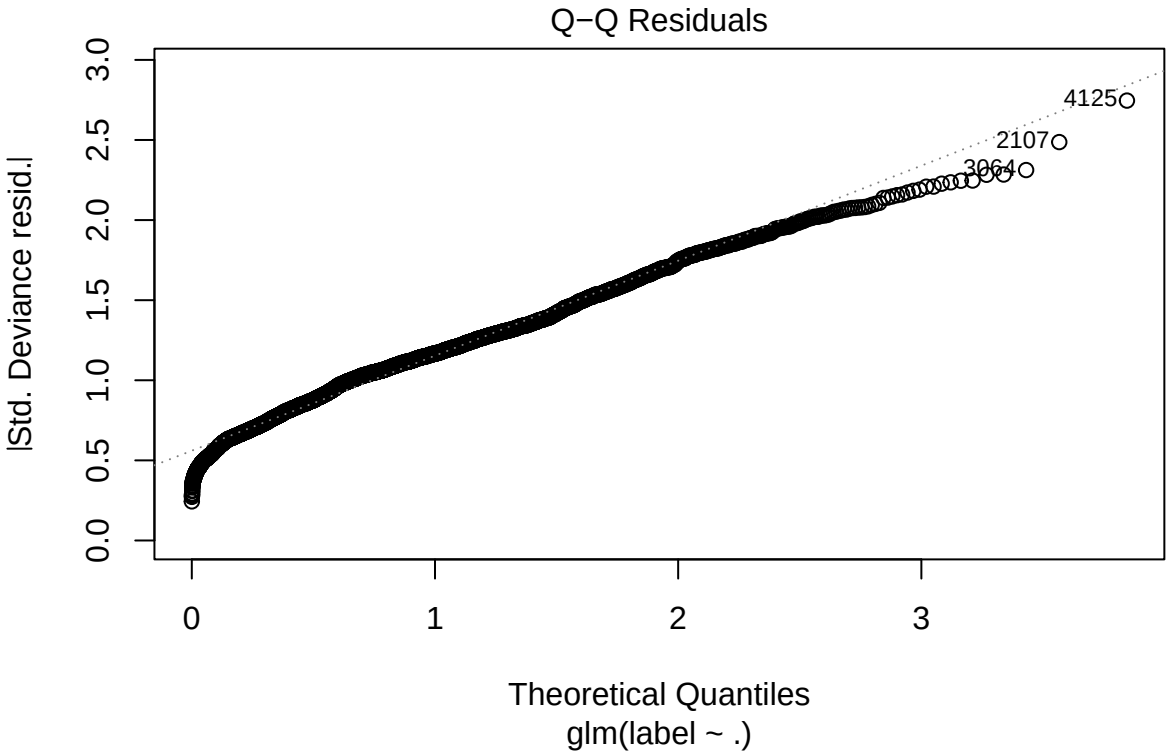


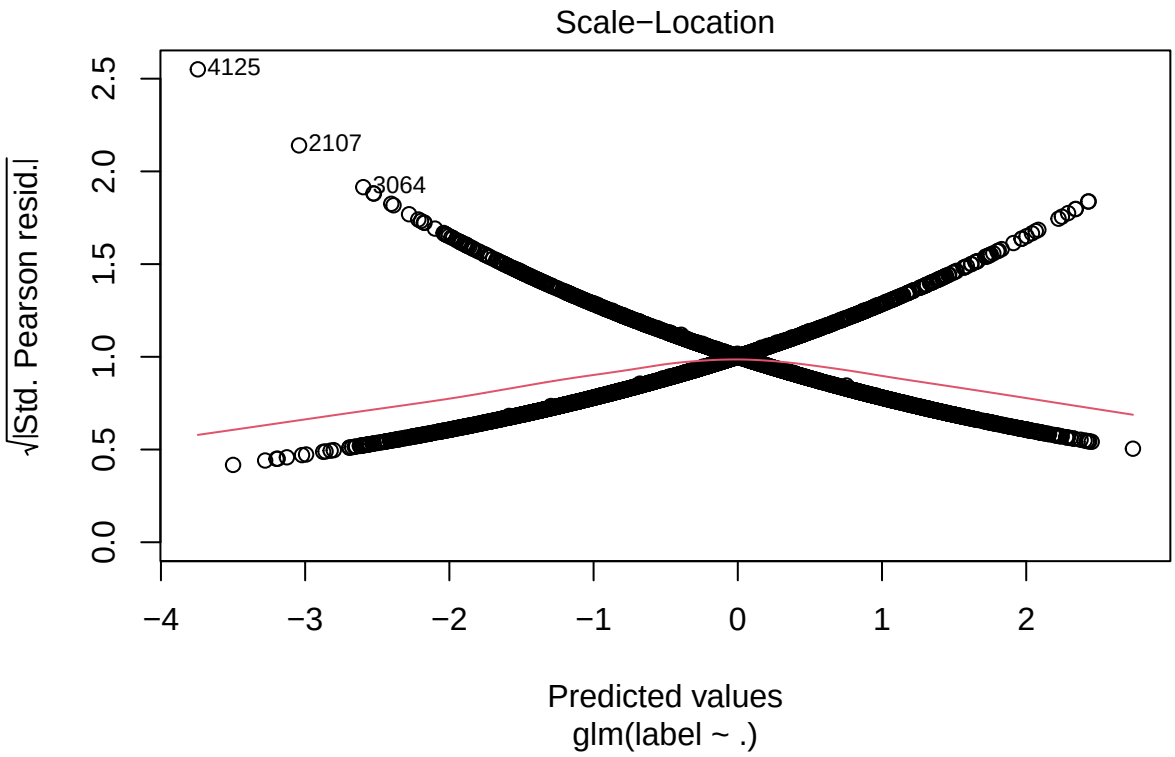

Scale−Location
26114
29800
26115
√|Std. Pearson resid.|
0
1
2
3
4
5
−10
−5
0
5
10
15
Predicted values
glm(series_success ~ .)

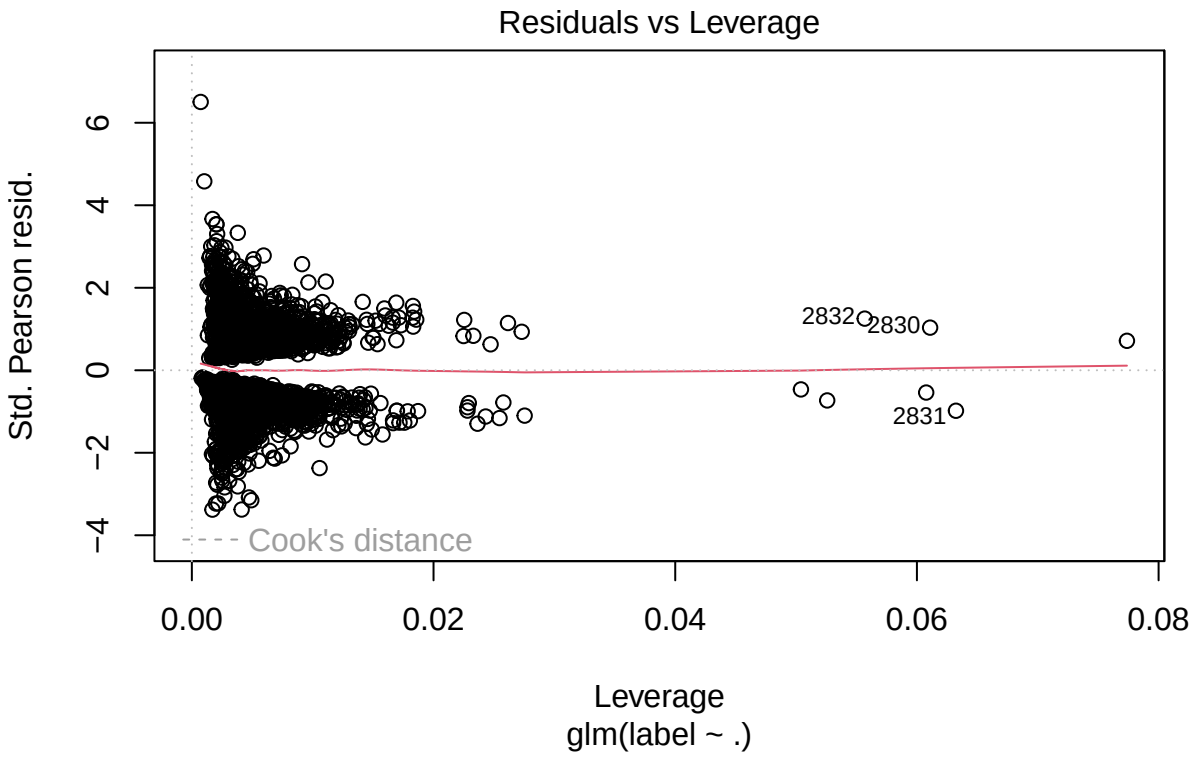

Residuals vs Leverage
Std. Pearson resid.
20
10
0
−10
−20
−30
24642
24643
22589
0.5
1
Cook's distance
0.000
0.005
0.010
0.015
0.020
Leverage
glm(series_success ~ .)

Model Diagnostics: Fourth Down Model






Scale−Location
4125
2107
3064
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
−4
−3
−2
−1
0
1
2
Predicted values
glm(label ~ .)


Residuals vs Leverage
Std. Pearson resid.
6
4
2
0
−2
−4
2832
2830
2831
Cook's distance
0.00
0.02
0.04
0.06
0.08
Leverage
glm(label ~ .)