Using Machine Learning to Analyze NFL Penalty Data

Leo DiPerna
Virginia Tech
Blacksburg, Virginia
dipernalz@vt.edu

Eric Uehling
Virginia Tech
Blacksburg, Virginia
uehlingeric@vt.edu

Figure 1: Players for the Kansas City Chiefs and Houston Texans lining up for a play during an NFL game [7].

1 INTRODUCTION

Which NFL penalties have the biggest impact on the results of drives, or on the results of games or on the outcomes of seasons? Which referees tend to call which penalties? Is it possible to predict the outcome of drives given the penalties that occurred during that drive? Is it possible to predict the penalties that a specific referee crew will call?

As avid NFL fans, these are the types of questions that we aimed to answer to gain a deeper insight into the effect that NFL penalties have on the league. NFL fans know that penalties have a big impact on the outcome of games, but are less sure about the specifics of their impact.

We were able to create a regression model that predicts the points scored during a drive given the the number of offensive and defensive penalties and the number of penalty yards from those penalties. This model used a gradient boosting regressor. We were also able to create a generalized linear model that predicted the number of each type of penalty that a referee crew called during a game.

We believe that we were able to uncover insights that fans, teams, and analysts of the game will all find useful and interesting.

2 PREVIOUS WORK

Before beginning to analyze our data, we looked at a number of previous papers that analyzed NFL data so that we had the necessary background and context to develop our own methods. We will provide a brief overview of the most important works that we looked at.

Researchers at University North in Croatia [4] looked at a wide range of previous research in sport outcome prediction and provided a broad overview of techniques that are being looked at. The papers that they looked at encompassed a wide range of sports, including NFL, NBA, and various international soccer leagues. They found that the most of these papers used some sort of feature selection or feature extraction algorithm prior to using an ML algorithm, which was most often neural networks. They also found that most researchers decided to treat the problem of outcome prediction as a classification problem.

Researchers at the University of Port Harcourt in Nigeria [1] attempted to use neural networks in combination with linear regression to predict the results of NFL games. They used linear regression to choose features from their dataset and then used those features to train a neural network. This gave us some insight into creative ways that we could select features when training our models.

Students at Denison University [2] attempted to use neural networks to predict the results of NFL games. They made a number of different models which selected different sets of features and trained their models based on these sets of features. For example, one set of features that they used comprised only of basic statistics such as passing yards, rushing yards, points, and turnovers. They derived another set of features using principal component analysis to reduce a large set of features to a smaller set. This gave us some more insight into how we could select features when training our models.

Researchers at Texas Tech University [3] investigated how the ambient temperature during NFL games resulted in more aggression from players. They measured aggression by classifying certain penalties as being agressive penalties and looking at how many of those penalties were committed during a given game. This gave us some preliminary ideas on ways that we could analyze and gain insights from penalty data.

A master's student at the University of Iowa [5] examined NFL penalties over the last 20 years in an effort to gain a greater understanding of the effect of individual referees, the difference between penalty calls against different NFL teams, and changes in penalties over time. He found that that there were some major changes in the frequency of penalty calls over time, and certain seasons where there were noticeably less and noticeably more penalties called. However, he was also able to determine that these variations were more due to rule changes and organizational changes rather than the decisions of individual referees.

Researchers at Skidmore College [6] examined how consistently certain penalties are called during different phases of an NFL game. They examined the frequency of penalties at different times during the game and also looked at the status of the game of the game as it related to the frequency of penalty calls. They found that judgement penalty calls varied considerably depending on the status of the game. These penalties were called less frequently than expected during the first five and last five minutes of the game.

3 FEATURE PREPARATION

This project consisted of three steps: data collection, data cleaning, and data analysis. We will discuss the data collection and data cleaning processes in this section as well as some of the preliminary data exploration that we did. We will go into the specifics of our data analysis in the next section.

3.1 Collection

To be able to conduct a full analysis of the impact of penalties on different facets of the NFL, we needed to collect a wide range of data from NFL games. The data that we collected includes a dataset of individual penalties, a dataset of drive results, a dataset of game results, a dataset of season results, and a couple of miscellaneous dataset to link the others together.

We were unable to find any easily accessible datasets that satisfied all of our requirements, so we decided to collect our data via web scraping. We scraped data from Pro Football Reference, a website that provides a massive amount of NFL statistics dating back to the start of the league. We decided to collect data starting from the 2009 season and continuing to the present.

Pro Football Reference has a webpage for every NFL game. This webpage contains tons of information about every game, including the results of each drive, statistics for each player, and the results of each play. Each webpage has the same format, and the URL of each webpage had a predictable format. This allowed us to create a web scraping script that could systematically generate a link for each game from 2009 to the present, read the data from each table in the webpage, and output the important data to CSV files. Using this script, we were able to collect data on drive results and game results.

However, we still needed to collect data on penalties. To collect penalty data, we used a different website, nflpenalties.com. This site contains a page for each season-team combination that lists all penalties that the team committed in that season in a table. By scraping this site, we were able to create a dataset that contains every penalty committed by every team from 2009 to the present.

3.2 Data Cleaning

After scraping the data, we needed to do some data cleaning to prepare the features that we wanted for model training. The raw data files that we used were:

- `drives.csv` – Contains a row for every drive that occurred during our study period.
- `penalties.csv` – Contains a row for every penalty that occurred during our study period.
- `game_detail.csv` – Contains a row for every game that occurred during our study period.

To analyze the effect of penalties on the outcome of drives, we needed to know which penalties occurred on which drives. This meant that we needed to merge the data in `drives.csv` with the data in `penalties.csv`. Since each row in `penalties.csv` contained the time during the game that the penalty occurred, and each row in `drives.csv` contained the time during the game that the drive occurred, we were able to get a count of each penalty during each drive by iterating over each penalty and matching it with its drive.

In the different data files that we collected, there were a number of different formats for team and game IDs. This made determining which game a drive occurred in and determining which game a penalty occurred in difficult, so we had to standardize the format of the game ID column across our dataframes.

There were also some annoying quirks with the data that we had to solve. For example, the NFL season changed from 16 regular season games to 17 beginning during the 2021 season. This was in the middle of our study period, so we had to account for this when determining the week that a game occurred. There are also a number of teams that changed their names or cities during our study period, so we also had to account for that and make sure that teams were matched to the same team ID regardless of whether their name had changed.

We decided to ignore special teams penalties and focus solely on offensive and defensive penalties, so we had to filter out penalties based on the phase of the game that they occurred in.

3.3 Exploratory Data Analysis

Before deciding which models to make, we did some exploratory data analysis to gain some insight into the data. Here are some of the most interesting insights that we were able to gain from looking at the data.

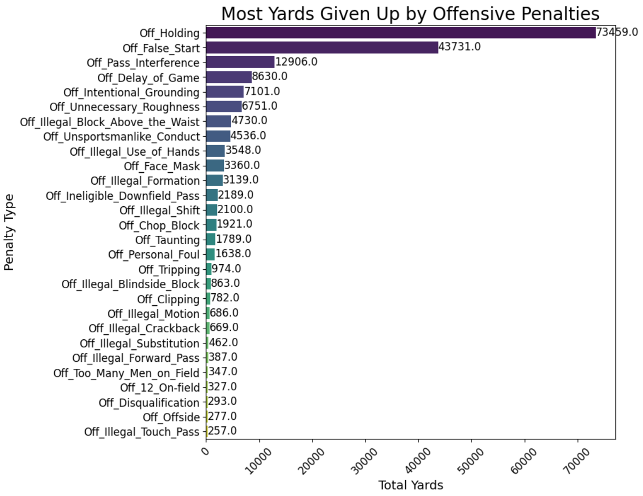

Figure 2: Penalty yardage by offensive penalties.

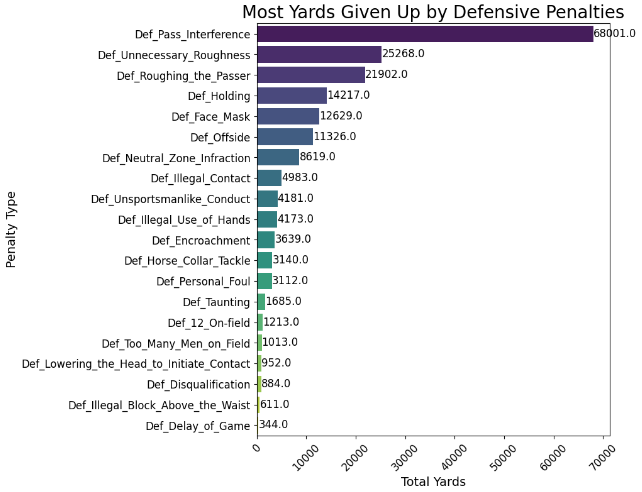

Figure 3: Penalty yardage by defensive penalties.

From these graphs, we can see that certain penalties have a much larger impact on the outcomes of games than others. On the offensive side, holding and false starts seem to have a much bigger impact than all of the other penalties combined. On the defensive side, defensive pass interference has a much larger impact than any other penalty, which makes sense because it is a spot foul that can cost a team an unlimited number of yards.

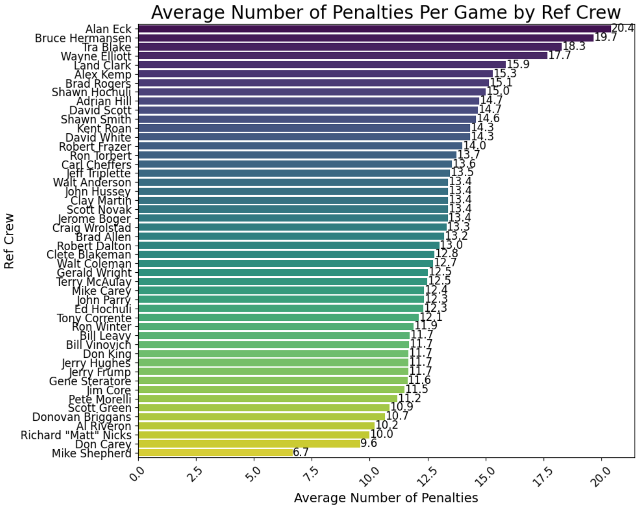

Figure 4: Average number of penalties called per game by referee crew.

We observed that there was a wide range of penalties that each referee crew called per game. The lowest was Mike Shepherd's crew, which only called an average of 6.7 penalties per game. The highest was Alan Eck's crew, which called 20.4 penalties per game. This means that Alan Eck's crew called over three times as many penalties per game when compared to Mike Shepherd's crew. Such a drastic discrepancy led us to think that we might be able to predict a different number of penalties in a game based on the referee crew that was doing that game.

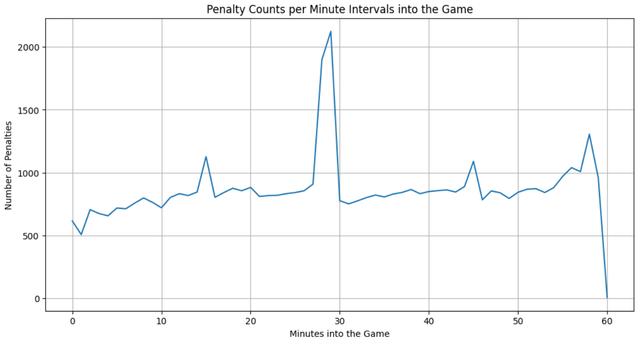

Figure 5: Distribution of penalties at different times during a game.

We also found it interesting that there seemed to be a noticeable difference in the number of penalties called at different times during a game. It seems like the number of penalties increases from the beginning of each half to the end of each half, with a large spike near the end of the half. This spike is larger at the end of the first half than it is at the end of the second half. This might indicate that more penalties are called during tense situations than during normal situations.

4 RESULTS

4.1 Gradient Boosting Classifier for Drive Outcomes

The Gradient Boosting Classifier was designed to predict the outcome of a drive (Touchdown, Field Goal, or No Score) based on penalties. Here are the detailed results:

Classification Metrics:

- Accuracy: 67%
- Precision, Recall, and F1-Score for each class:
  - Field Goal: Precision = 43%, Recall = 2%, F1-Score = 3%
  - Touchdown: Precision = 54%, Recall = 21%, F1-Score = 30%
  - No Score: Precision = 68%, Recall = 97%, F1-Score = 80%

Feature Importance:

Feature	Importance
Total Defensive Penalties	0.4298
Line of Scrimmage (los)	0.2426
Time Left in Seconds	0.1873
Total Defensive Penalty Yards	0.0804
Total Offensive Penalty Yards	0.0311
Total Offensive Penalties	0.0289

Table 1: Feature importance for the Gradient Boosting Classifier model.

4.2 Gradient Boosting Regressor for Drive Points

This model aimed to predict the number of points scored during a drive.

Performance Metrics:

- Mean Squared Error (MSE): 6.885
- R-Squared (R^2): 0.136

Despite its relatively low explanatory power, this model highlights the complexities of directly linking penalties to scoring outcomes without considering more contextual game factors.

4.3 Negative Binomial Model for Penalty Prediction

An ensemble of Negative Binomial models was used to predict the frequency of specific penalties based on game context and referee crews.

Sample Model Performance - Offensive Holding:

- Mean Squared Error (MSE): 1.137
- R-Squared (R^2): 0.053

The weights for different predictors like team ID, opponent ID, and referee crew suggest nuanced influences on penalty calls. However, the low R^2 values across all models indicate limited predictability based on the available features.

4.4 Neural Network Model

Our neural network model aimed to refine predictions by incorporating a broader set of input features and more complex interactions.

Performance Metrics:

- Best Validation Loss: 0.0070
- Test Mean Absolute Error (MAE): 0.0396

Despite detailed tuning, the neural network did not significantly outperform simpler models, indicating that the existing feature set may not capture all relevant predictors of NFL penalties or drive outcomes.

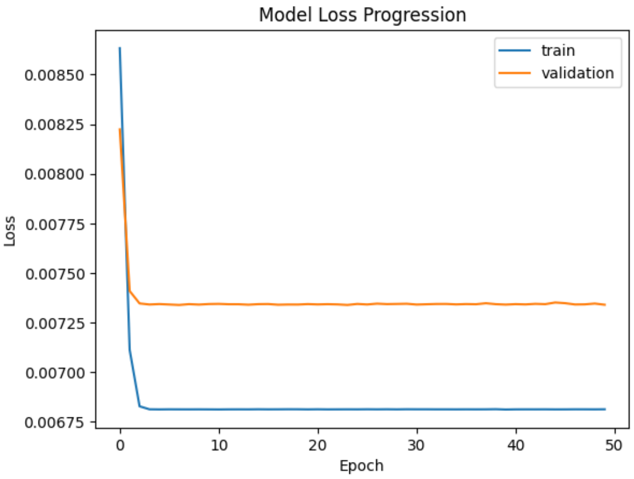

Figure 6: Validation loss and MAE across training epochs for the neural network model.

5 DISCUSSION

The results indicate a challenging landscape for directly predicting NFL game dynamics from penalties and drive data. While defensive behaviors and game situations like line of scrimmage and time left play pivotal roles, the predictive power remains moderate. Future work should consider integrating more dynamic and contextual game data, including player-specific actions and more detailed situational variables.

The variability in referee behavior also presents an intriguing avenue for further exploration. As shown, referee crews significantly influence penalty frequencies, suggesting a potential for models that can adapt to referee-specific patterns.

6 CONCLUSION

This project explored various machine learning approaches to understand and predict the impacts of NFL penalties on game outcomes. While we achieved moderate success in some areas, the results underscore the complexity of the task and the potential need for richer datasets and more refined models. Future research could explore the integration of more granular data, such as player tracking information, to enhance model accuracy and applicability.

REFERENCES

[1] Oscar Uzoma Anyama and Chinwe Peace Igiri. 2015. An Application of Linear Regression & Artificial Neural Network Model in the NFL Result Prediction. *International Journal of Engineering Research & Technology* 4, 1 (Jan. 2015), 457–461. https://www.ijert.org/research/an-application-of-linear-regression-artificial-neural-network-model-in-the-nfl-result-prediction-IJERTV4IS010426.pdf

[2] Andrew D. Blaikie et al. 2011. NFL & NCAA Football Prediction using Artificial Neural Networks. In *Proceedings of the 2011 Midstates Conference on Undergraduate Research in Computer Science and Mathematics*. https://personal.denison.edu/~lalla/MCURCSM2011/4.pdf

[3] Curtis Craig et al. 2016. A relationship between temperature and aggression in NFL football penalties. *Journal of Sport and Health Science* 5, 2 (2016), 205–210. https://doi.org/10.1016/j.jshs.2015.01.001

[4] Tomislav Horvat and Josip Job. 2020. The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery* 10, 5 (2020). https://doi.org/10.1002/widm.1380

[5] Zachary McDaniel. 2021. *NFL Penalty Analysis, Referee Influence and Penalty Trends Over Time*. Master's thesis. University of Iowa. https://iro.uiowa.edu/esploro/outputs/9984112118002771

[6] Kevin Snyder and Michael J. Lopez. 2015. Consistency, Accuracy, and Fairness: A Study of Discretionary Penalties in the NFL. *Journal of Quantitative Analysis in Sports* 11, 4 (2015), 219–230. https://doi.org/10.1515/jqas-2015-0039

[7] New York Times. 2022. Kansas City Chiefs vs. Houston Texans. Image. https://www.nytimes.com/2022/12/22/business/youtube-nfl-sunday-ticket.html