Predicting and Forecasting NFL Player Performance using Play by Play Statistics and Sentiment

Group 1 members: Shang-Yi Lin, Andy Li, Chen Yang

Abstract

Player performance evaluation is a critical aspect of sports analytics, often relying on historical statistics and advanced modeling techniques. NFL football is one of the most wide-known sports across the U.S. We wanted to create a tool that predicts player rankings based on a variety of stats such as player category, yds, height, weight, etc. We implemented linear models such as OLS, best subset selection, ridge, and lasso regression models to model our tabular data. We also implemented machine learning models such as random forest and XGBoost. Lastly, we were curious on how public sentiment impacted player performance, so we implemented sentiment analysis and combined it with our previous models.

Our findings show that while linear models can serve as a baseline model for NFL predictions, machine learning models such as random forest and XGBoost, through tuning, have significantly higher prediction accuracy on out of sample data. This all followed our initial hypothesis. However, we were surprised to see that our public sentiment feature was not important when predicting player performance.

Introduction

In this project, we utilize three different datasets for our performance prediction. For our first dataset, we combined three tabular datasets—roster, pbp, and next_gen_stats— from nflFastR to construct a comprehensive, player-focused dataset. Our second and third datasets lean more towards public sentiment and player bios consisting of website data scraped off of NFL.com and a Reddit forum repository.

Initially, our hypothesis was that the machine learning models would outperform linear models. This is due to non linear complex relationships in the data. Our modeling required us to find a universal metric for player performance. Since NFL has multiple roles, we decided that we needed to categorize the players and create models for each category separately. We split the roles for QuarterBack (QB), Running Back (RB), and Wide Receiver/Tight End (WR/TE). We then used a metric called average extra points added (Avg_EPA) as our common response for assessing player performance across roles.

For our sentiment addon, we initially hypothesized that sentiment does play a factor in affecting player performance. This is due to the mental stress athletes might run into when facing public criticism. Hence, we included sentiment score obtained from VADER, and Bert base multilingual tokenizer as a feature into our machine learning models.

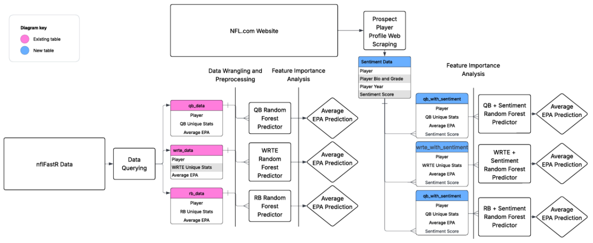


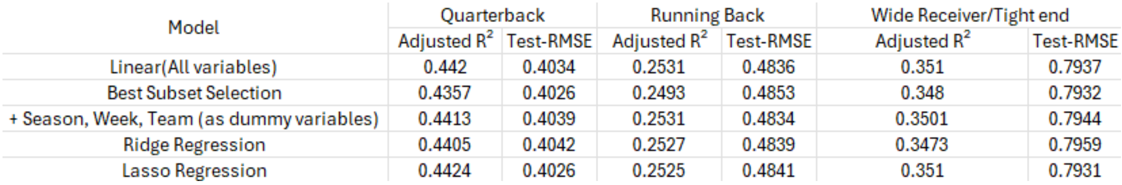
Figure 1: Tentative Project Structure and Initial Workflow

EDA and further data wrangling

The data was prepared by one-hot encoding all categorical variables and transforming the week column into a cumulative career_game_number to better reflect player experience. Rolling average EPA features were created over 1, 3, and 5 games, with missing values filled with neutral zeros.

Methodology Linear Regression

We fit several linear models using all predictors in our dataset and observed that some predictors were highly correlated. To improve model stability and reduce multicollinearity, we computed the correlation matrix and removed variables with an absolute correlation greater than 0.9. However, after removing these highly correlated predictors, the overall model performance—measured by adjusted R-square —decreased.



Model	Quarterback		Running Back		Wide Receiver/Tight end	
	Adjusted R^2	Test-RMSE	Adjusted R^2	Test-RMSE	Adjusted R^2	Test-RMSE
Linear(All variables)	0.442	0.4034	0.2531	0.4836	0.351	0.7937
Best Subset Selection	0.4357	0.4026	0.2493	0.4853	0.348	0.7932
+ Season, Week, Team (as dummy variables)	0.4413	0.4039	0.2531	0.4834	0.3501	0.7944
Ridge Regression	0.4405	0.4042	0.2527	0.4839	0.3473	0.7959
Lasso Regression	0.4424	0.4026	0.2525	0.4841	0.351	0.7931

Table 1: Comparison of linear models

All models in table 1 performed similarly for each position group, indicating no clear advantage of one linear method over another for this dataset.

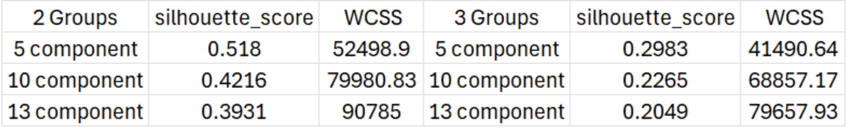
Generalized Additive Model (GAM)

To capture nonlinear effects, a Generalized Additive Model (GAM) was fitted using smooth splines where appropriate For quarterbacks, the model achieved an adjusted R^2 of 0.48 and an RMSE of 0.3925. Among 14 significant predictors, 6 lacked sufficient unique values for spline fitting (e.g., *rush_touchdowns*, *pass_touchdowns*, *interceptions*), and only 2 exhibited linearity (*rush_attempts* and *avg_completed_air_yards*). For running backs, the adjusted R^2 was 0.308 with an RMSE of 0.4798, and only 1 of 18 significant variables showed a linear relationship. For tight ends and wide receivers, the adjusted R^2 was 0.464 and RMSE was 0.7602, with none of the 20 significant predictors displaying linear behavior. These results underscore the relevance of modeling nonlinear effects in player performance analysis.

GAM generally outperformed linear models. Excluding variables unsuitable for smooth splines, most predictors showed nonlinear relationships, suggesting that nonlinear models may better capture the structure of the data.

PCA followed by K-means

Examining the best subset selection, we found that rush_attempts for quarterbacks had a negative coefficient, which seemed counterintuitive since some QBs gain value through rushing. Suspecting different QB types, we applied PCA followed by K-means clustering to segment them.



2 Groups	silhouette_score	WCSS	3 Groups	silhouette_score	WCSS
5 component	0.518	52498.9	5 component	0.2983	41490.64
10 component	0.4216	79980.83	10 component	0.2265	68857.17
13 component	0.3931	90785	13 component	0.2049	79657.93

Table 2: K-means Clustering Evaluation: PCA Components vs. Cluster Count

We selected 2 clusters with 5 principal components, as it provided the highest silhouette score (0.518) and a relatively low WCSS, indicating well-separated and compact clusters. The QB dataset was split accordingly: Group 1 (~500 observations) exhibited low variability across most predictors, resulting in poor GAM performance (adjusted $R^2 = 0.356$), while Group 2 (~3,000 observations) showed strong model performance with an adjusted R^2 of 0.879.

The poor performance of Group 1 suggests the need for further investigation into the effectiveness of this segmentation, potentially through refined clustering criteria or additional data collection.

Random Forest Regressor

We first utilized a Random Forest Regression model for predicting player EPA. This model utilizes bootstrapping because each tree sees a random subset of samples from the training set. The predictions from all of the trees are then averaged across the entire forest.

The Random Forest model is then tuned using grid search cross validation. The best parameters found through grid search include a max_depth of 50, max_features sqrt, min_samples_leaf of 1, and min_samples_split of 5. The minimum cross validation score for these selected features is 0.2409.

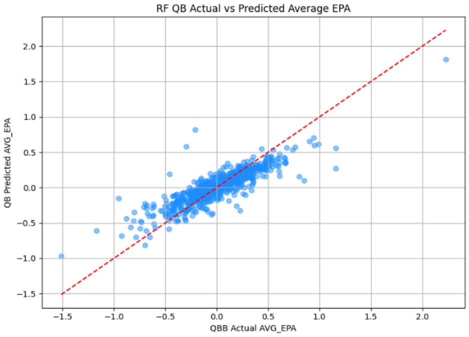
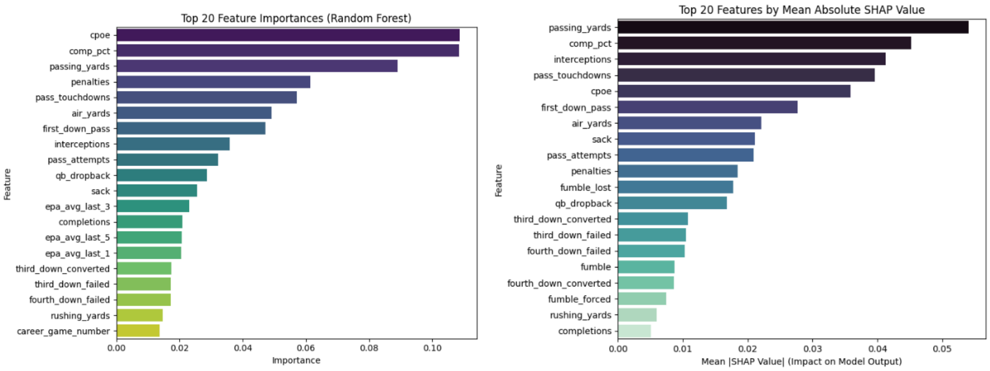


Figure 4: QB Predicted vs Actual RF

By fitting the testing data into the tuned random forest model, we obtain a root mean squared error of 0.1787, and R-squared of 0.7316. From figure 4, we observe that the model prediction vs actual spread generally resembles the 45 degree line. Compared to the previous linear models, the random forest model has a significantly higher prediction accuracy.

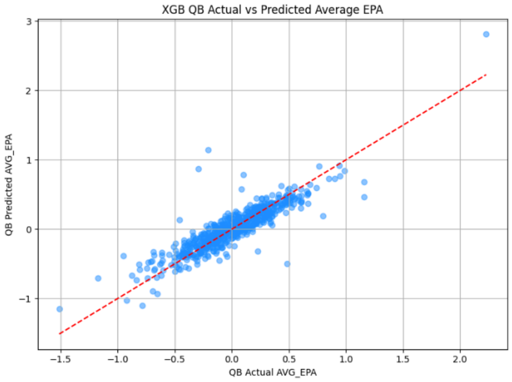


Figures 5 6: Tuned Random Forest QB Feature importance and gain plots

From the importance and gain plots in figures 5 and 6, we observe that the model uses features such as cpoe, comp_pct, passing_yards, the most. The features that resulted in the most gain in the model were passing_yards, comp_pct, interceptions, and pass_touchdowns. This makes sense because those features are the player's in game statistics which directly impacts average EPA. We also see that in the feature importances plot, the rolling EPAs are within the top 20 most important features which shows that players' previous game performances do impact the next game's performance.

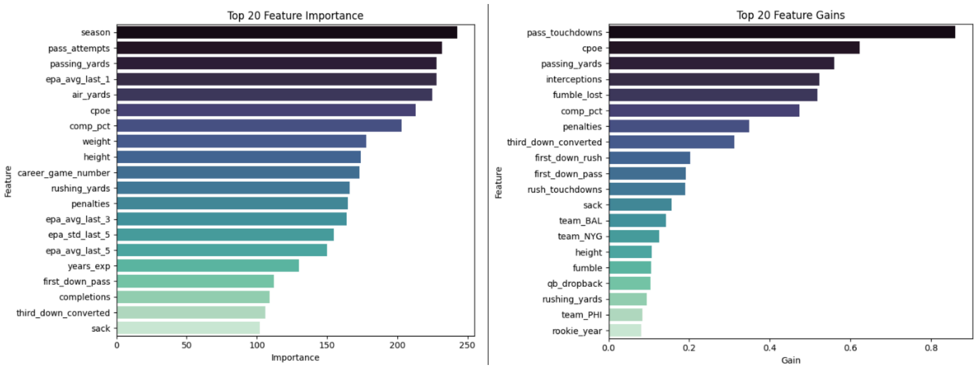
XGBoost

The next model we used was an XGBoost regression model. We utilized gridsearch with 5 fold cross validation to select parameters. The best parameters found were n_estimators of 200, Dropout Additive Regression Tree booser (Dart), learning_rate of 0.1, max_depth of 3, colsample_bytree of 0.7, and subsample of 0.7.



By fitting the testing data into the tuned random forest model, we obtain a root mean squared error of 0.1617, and R-squared of 0.7801. By comparing the RMSE and R-Squared metrics, we find that the XGBoost model has higher prediction accuracy than the random forest model.

In the XGB actual vs predicted plot, we observe that the spread of predictions closely resembles the 45 degree line. Compared to our random forest model, the XGBoost model has better prediction accuracy. However a drawback to this model is that it requires tuning of more hyperparameters, and is more susceptible to overfitting noisy data. However, parameters such as the Dart booster, and built in L1/L2 regularization helps prevent overfitting.



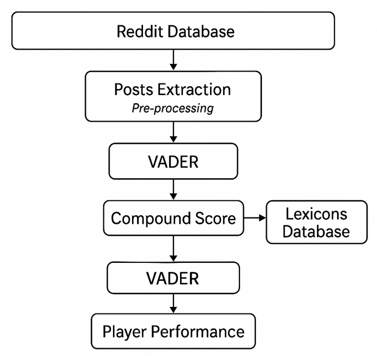
Figures 8 9: Tuned XGBoost QB Feature importance and gain plots.

From the importance and gain plots in figures 8 and 9, we observe that the model uses features such as season, pass_attempts, passing_yards, and epa_avg_last_1 the most. The features that resulted in the most gain in the model were pass_touchdowns, cpoe, passing_yards, and interceptions. Once again, the feature importance plot includes the rolling average EPAs in the top 20 indicating that players' previous games are useful in predicting next game average EPA.

Sentiment Addition

To prepare text for sentiment analysis, we combined the title and selftext fields into a unified full_text column, and removed entries that were empty or just URLs. We then applied VADER (Valence Aware Dictionary and sEntiment Reasoner)—a rule-based sentiment analysis tool tailored for social media and short texts.

Figures 12: Reddit-Based Sentiment Analysis Pipeline Using VADER



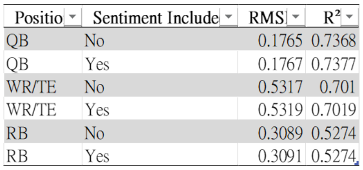
VADER is a sentiment analysis tool trained on social media data that uses a built-in dictionary and a set of rules to understand how positive or negative a piece of text is. Each word or phrase in the dictionary has a score. For example, the word "great" has a positive score of +3.1, while "terrible" has a negative score of –3.4.

When we run VADER, it gives four scores: pos (how much of the text is positive), neg (how much is negative), neu (how much is neutral), and compound, (number from –1 to +1 that shows the overall feeling). In our project, we used the compound score to measure sentiment in each Reddit post.

Model Integration Result

After generating sentiment scores for 688,324 Reddit posts, we aggregated them by player and week, calculating both the average sentiment score and number of mentions. These two features were then added into our Random Forest regression model, along with player performance statistics, to predict weekly EPA (Expected Points Added).

To evaluate model performance, we compared results across three player positions: Quarterback (QB), Wide Receiver/Tight End (WR/TE), and Running Back (RB).



Positio	Sentiment Include	RMS	R^2
QB	No	0.1765	0.7368
QB	Yes	0.1767	0.7377
WR/TE	No	0.5317	0.701
WR/TE	Yes	0.5319	0.7019
RB	No	0.3089	0.5274
RB	Yes	0.3091	0.5274

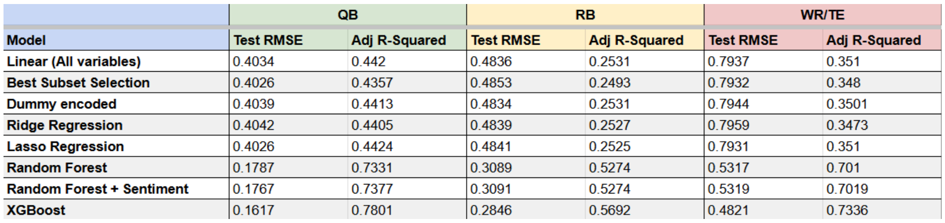
Figure 16: Model Performance With and Without Sentiment Features

Through our implementation of random forest with sentiment features, we noticed that the addition of these features did not improve the prediction accuracy. Additionally, their feature importance scores were among the lowest across all variables, which means they did not play a meaningful role in predicting EPA. Among all player positions, the model performed best when predicting quarterback (QB) performance, reaching an R^2 of 0.7377. However, this high level of accuracy came mainly from traditional performance statistics, not from sentiment data.

These results suggest that public sentiment on Reddit does not have a direct impact on player performance. This goes against our initial hypothesis that player performance is impacted by public sentiment. One possible reason is that players are mentally strong and are not affected by what people say online. Another explanation could be that Reddit posts do not reflect the entire public sentiment. Lastly, our dataset may not contain the full scope of public sentiment. With more resources available for web scraping, we can query more reliable sentiment data to better understand this relationship.

Conclusion

Through our comparison of all implemented models, we found that linear models have the worst prediction accuracy. However, they can serve as a baseline model for us to compare to. Machine learning models such as random forest and XGBoost, when tuned, have significantly improved evaluation metrics. XGboost with Dart booster has even higher prediction accuracy than random forest. Lastly, we observed that sentiment-related features such as avg_sentiment_score and num_mentions contributed very little to the random forest model.



	QB		RB		WR/TE	
Model	**Test RMSE**	**Adj R-Squared**	**Test RMSE**	**Adj R-Squared**	**Test RMSE**	**Adj R-Squared**
Linear (All variables)	0.4034	0.442	0.4836	0.2531	0.7937	0.351
Best Subset Selection	0.4026	0.4357	0.4853	0.2493	0.7932	0.348
Dummy encoded	0.4039	0.4413	0.4834	0.2531	0.7944	0.3501
Ridge Regression	0.4042	0.4405	0.4839	0.2527	0.7959	0.3473
Lasso Regression	0.4026	0.4424	0.4841	0.2525	0.7931	0.351
Random Forest	0.1787	0.7331	0.3089	0.5274	0.5317	0.701
Random Forest + Sentiment	0.1767	0.7377	0.3091	0.5274	0.5319	0.7019
XGBoost	0.1617	0.7801	0.2846	0.5692	0.4821	0.7336

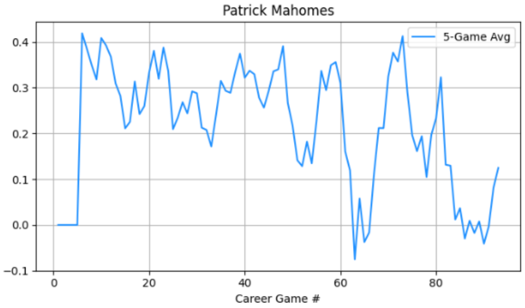
Table 3: Combined evaluation of all models

Future Work

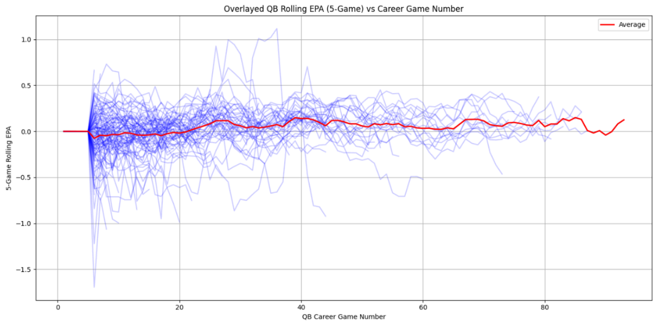
With more time, we can introduce more complex models using news sentiment scraped off NFL.com. If we are able to overcome the computing limitations, we can scrape professional written articles and retrieve individual players' sentiment through named entity recognition (NER). Additionally, graph neural networks may distinguish different chemistries between players when they are positioned together. Hence, there is much more potential for this topic given more time and resources.

Appendix (Addition to our report)

For our appendix, we will cover some of our additional exploratory data analysis we performed to better understand the different features and relations of our datasets.



The figure shows an individual's 5 game rolling average EPA across their career game number. We can observe that coming into his professional career, Patrick Mahomes had a strong performance at the start and throughout his career. The frequent spikes may be attributed to factors such as different strategies that led to Patrick Mahomes not earning as many points or potential injuries that may have inhibited performance. Around his 80th game Patrick Mahomes suffered an ankle injury which matched up with the extreme decrease in the 5 game rolling EPA.

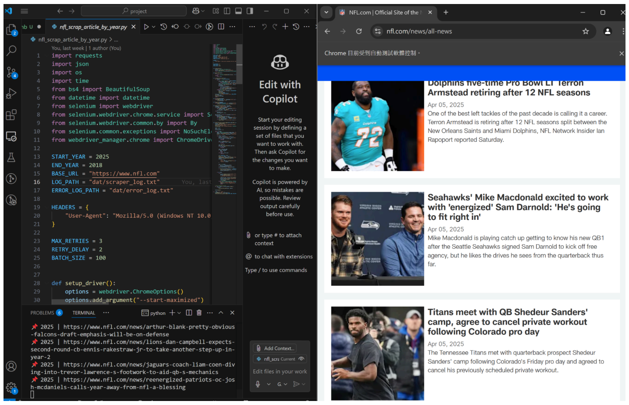


In the QB rolling EPA plot, we observed that many players had very volatile rolling EPAs early on in their career. However, as the players progressed through more and more games, their 5 game rolling EPAs stabilized.
This serial analysis may benefit our prediction models due to it being a measure of consistency and potential injury detection which is why we decided it would be useful to include into our models.

Data Collection and Preprocessing

We initially attempted to extract sentiment data from two official sources: NFL.com scout reports and news articles. Although we successfully developed a Selenium-based web crawler, the lack of an API and the use of infinite scrolling made large-scale historical scraping (2018–2023) highly inefficient. The crawler became slow and unstable beyond recent pages, so we ultimately abandoned this approach.

Instead, we turned to a public Reddit dataset from Academic Torrents, which includes all Reddit submissions and comments from 2005 to 2024. Due to the size of the comment files, we focused on submission data. After decompressing and filtering posts between 2018 and 2023, we kept 15 key columns (e.g., title, selftext, created_utc, author, num_comments, etc.).


project
nfl_scrap_article_by_year.py
nfl_scrap_article_by_year.py > ...
You, last week | 1 author (You)
Edit with Copilot
PROBLEMS 6
TERMINAL
python
Add Context...
nfl_scra Current
Edit files in your work
NFL.com | Official Site of the
nfl.com/news/all-news

```
import requests
import json
import os
import time
from bs4 import BeautifulSoup
from datetime import datetime
from selenium import webdriver
from selenium.webdriver.chrome.service import S
from selenium.webdriver.common.by import By
from selenium.common.exceptions import NoSuchEl
from webdriver_manager.chrome import ChromeDriv

START_YEAR = 2025
END_YEAR = 2018
BASE_URL = "https://www.nfl.com"
LOG_PATH = "dat/scraper_log.txt"      You, last
ERROR_LOG_PATH = "dat/error_log.txt"

HEADERS = {
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0
}

MAX_RETRIES = 3
RETRY_DELAY = 2
BATCH_SIZE = 100


def setup_driver():
    options = webdriver.ChromeOptions()
    options.add_argument("--start-maximized")
```

Start your editing session by defining a set of files that you want to work with. Then ask Copilot for the changes you want to make.

Copilot is powered by AI, so mistakes are possible. Review output carefully before use.

or type # to attach context

@ to chat with extensions

Type / to use commands

```
📌 2025 | https://www.nfl.com/news/arthur-blank-pretty-obvious
-falcons-draft-emphasis-will-be-on-defense
📌 2025 | https://www.nfl.com/news/lions-dan-campbell-expects-
second-round-cb-ennis-rakestraw-jr-to-take-another-step-up-in-
year-2
📌 2025 | https://www.nfl.com/news/jaguars-coach-liam-coen-div
ing-into-trevor-lawrence-s-footwork-to-aid-qb-s-mechanics
📌 2025 | https://www.nfl.com/news/reenergized-patriots-oc-jos
h-mcdaniels-calls-year-away-from-nfl-a-blessing
```

Chrome 目前受到自動測試軟體控制。

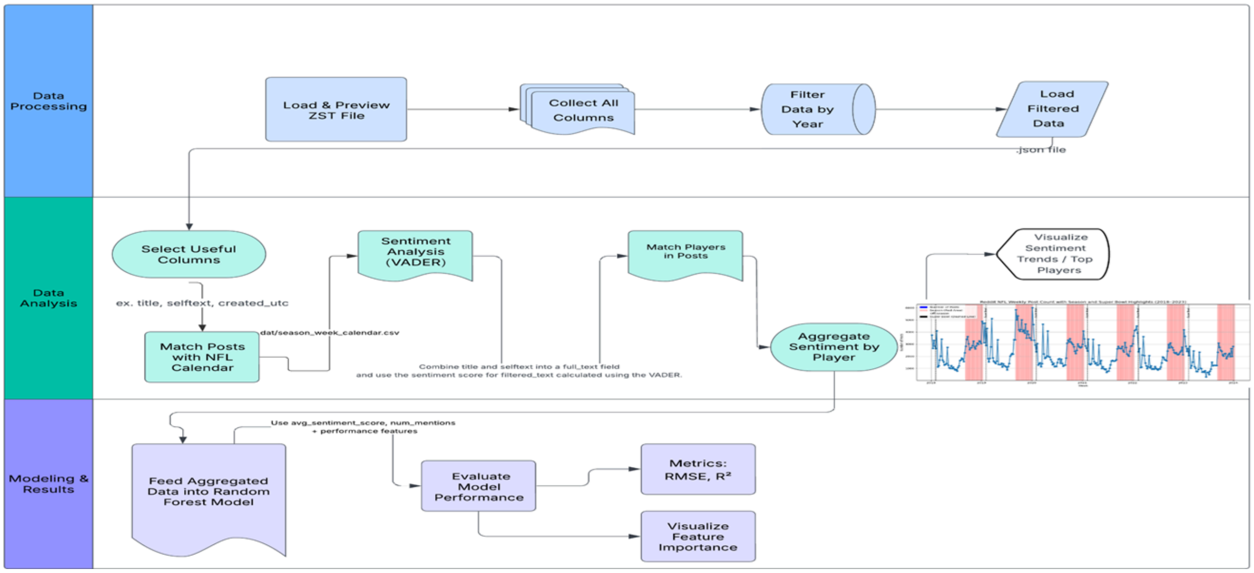
Dolphins five-time Pro Bowl LT Terron Armstead retiring after 12 NFL seasons
Apr 05, 2025
One of the best left tackles of the past decade is calling it a career. Terron Armstead is retiring after 12 NFL seasons split between the New Orleans Saints and Miami Dolphins, NFL Network Insider Ian Rapoport reported Saturday.

Seahawks' Mike Macdonald excited to work with 'energized' Sam Darnold: 'He's going to fit right in'
Apr 05, 2025
Mike Macdonald is playing catch up getting to know his new QB1 after the Seattle Seahawks signed Sam Darnold to kick off free agency, but he likes the drives he sees from the quarterback thus far.

Titans meet with QB Shedeur Sanders' camp, agree to cancel private workout following Colorado pro day
Apr 05, 2025
The Tennessee Titans met with quarterback prospect Shedeur Sanders' camp following Colorado's Friday pro day and agreed to cancel his previously scheduled private workout.

Figure 10: NFL Article Scraper Chrome Scrolling via Selenium

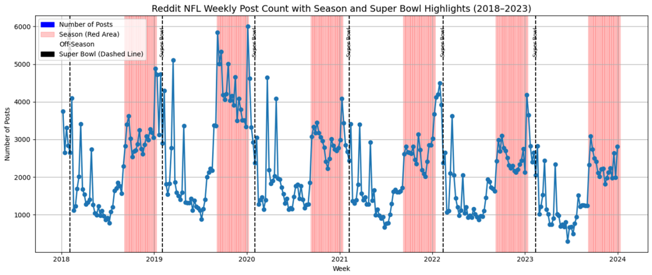
This diagram visualizes the full pipeline from ZST file ingestion and pre-processing, through VADER-based sentiment analysis and player-matching, to final integration in Random Forest modeling.


Data Processing
Load & Preview ZST File
Collect All Columns
Filter Data by Year
Load Filtered Data
.json file
Data Analysis
Select Useful Columns
ex. title, selftext, created_utc
Match Posts with NFL Calendar
dat/season_week_calendar.csv
Sentiment Analysis (VADER)
Combine title and selftext into a full_text field and use the sentiment score for filtered_text calculated using the VADER.
Match Players in Posts
Aggregate Sentiment by Player
Visualize Sentiment Trends / Top Players
Modeling & Results
Use avg_sentiment_score, num_mentions + performance features
Feed Aggregated Data into Random Forest Model
Evaluate Model Performance
Metrics: RMSE, R²
Visualize Feature Importance

Figures 11: End-to-End Sentiment Analysis Workflow

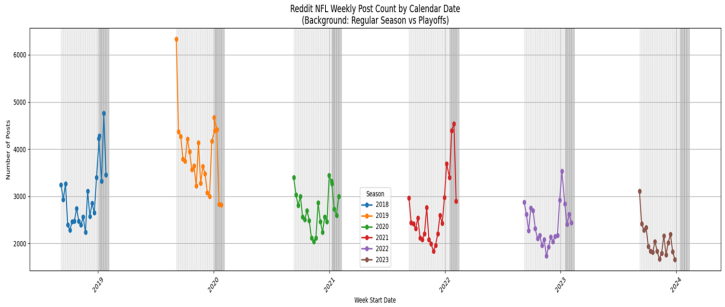
Aggregation and Integration into Modeling

Once sentiment scores were generated for 688,324 Reddit submissions, we matched them to NFL players by searching for name mentions in each post. This enabled us to compute the average sentiment score and number of mentions per player per week.



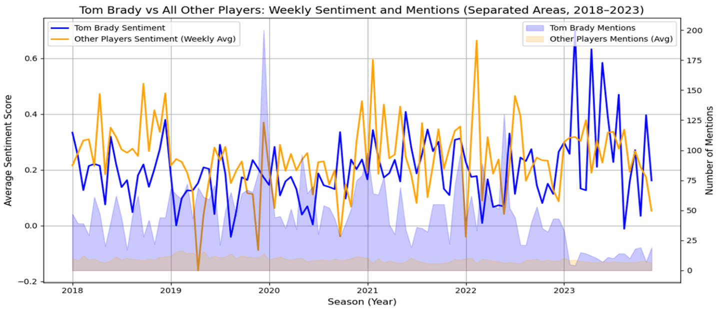
Figures 13: Weekly Reddit NFL Post Count with Season & Super Bowl Highlights (2018–2023)

This visualization shows the volume of NFL-related Reddit posts, overlaid with seasonal highlights. Spikes near the Super Bowl and season start reflect heightened fan activity.



This plot compares individual seasons, allowing clearer alignment with real-world events such as Tom Brady's championship or Antonio Brown's team switch.

Figure 14: Per-Season Reddit Post Counts (2018–2023)



The case study shows weekly sentiment scores and mention volumes for Tom Brady vs. all other players. Sentiment peaks coincide with key career moments.

Figure 15: Tom Brady vs. All Other Players: Weekly Sentiment and Mention