NFL Player Performance Modeling

Andy Li, Chen Yang, Joe Lin

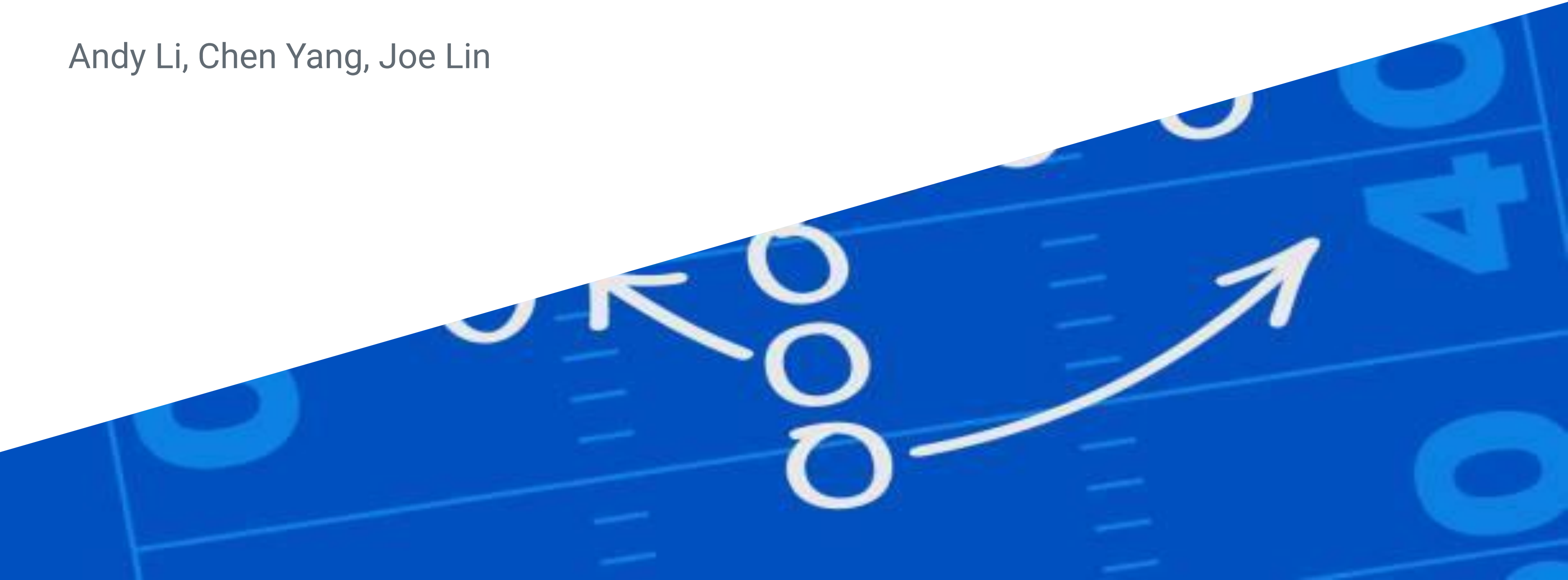


Table of Contents

1. Objective and Introduction
2. Our Datasets
3. Linear and GAM models
4. Random Forest and XGBoost models
5. Sentiment addon analysis
6. Future work

Objective

- Predict player performance due to various factors. (e.g. player age, season, experience, team, etc.)

- Implement multiple regression and machine learning models.

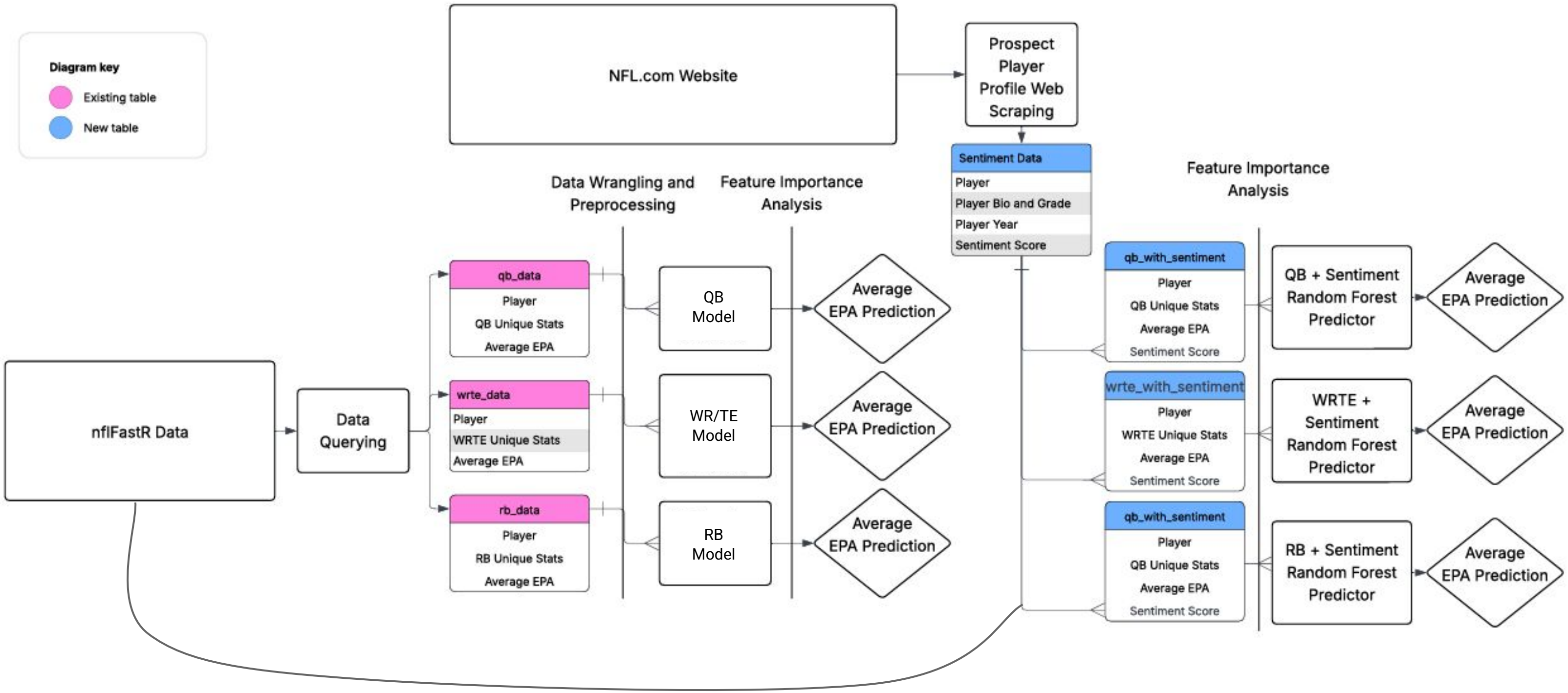
- Many people like us do not have a background in American football
  - This analysis will help people find players to follow and watch.
  - May help with sports betting.

- Useful for teams and coaches to identify players who are over performing or underperforming.
  - May help with injury detection, or other performance inhibiting factors.

Known Methods

- There are many NFL predictive models that teams and coaches already use.
    - Rithmm is an example of an AI powered tool which takes historical data and predicts game results
- Costs money

General Workflow



Our Data

Tabular Data - nflfastR

- Roster - Who the players are:
  - Name, team, position, height, weight
- Pbp - what happened on the field:
  - Tracking every play in every game
  - Play type, players involved, yards gain, etc
- Next_gen_stats - How the play happened (player tracking stats)
  - aggressiveness(tight window throws)
  - Air distance metrics (intended and completed)
- We combined all three datasets into one to maximize available information.
  - Quarterback position, we obtained a total of 3,995 player-week observations covering each week from the 2018 to 2023 NFL seasons.

Response:

Avg_EPA (Average Extra Points Added) measures the average impact of each play on a team's likelihood of scoring during a given game.

Variables:

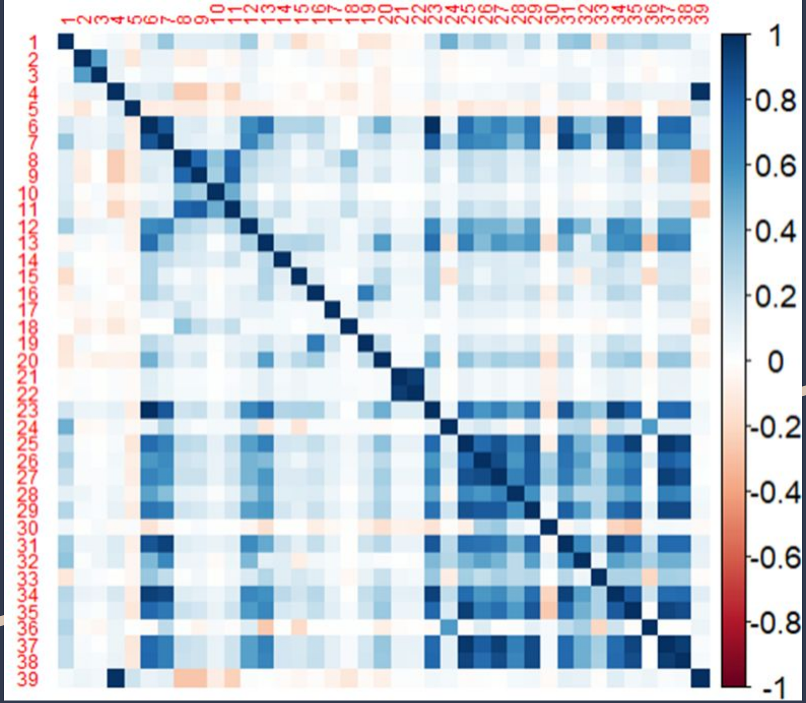
40+ variables across physical attributes, performance, and advanced passing stats

Player Info: weight, height, years_exp

Game context: season, week, team

Passing Metrics, Rushing Metrics, Situational Performance.

Linear Models



Before Removing Highly Correlated Variables

Model	Adjusted R^2	MSE
Linear (All variables)	0.4567	0.1593
Best Subset Selection	0.4507	0.1597
+ Season, Week, Team (as dummy variables)	0.4565	0.1601
Ridge Regression	0.4476	0.1615
Lasso Regression	0.4569	0.1590

After Removing Highly Correlated Variables

Model	Adjusted R^2	MSE
Linear (All variables)	0.4420	0.1627
Best Subset Selection	0.4357	0.1621
+ Season, Week, Team (as dummy variables)	0.4413	0.1601
Ridge Regression	0.4405	0.1633
Lasso Regression	0.4424	0.1621

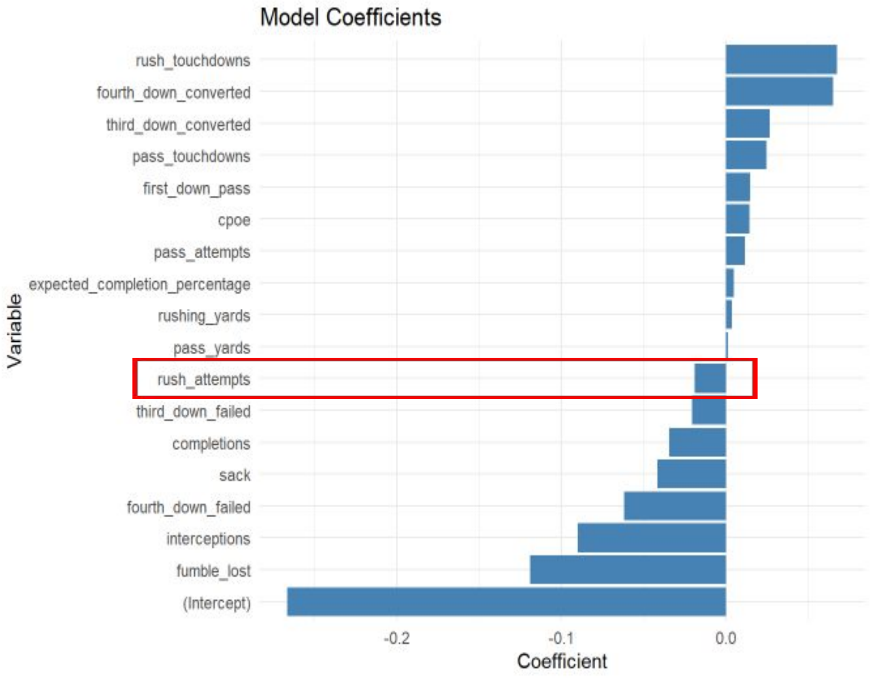
Beyond Linear - GAM

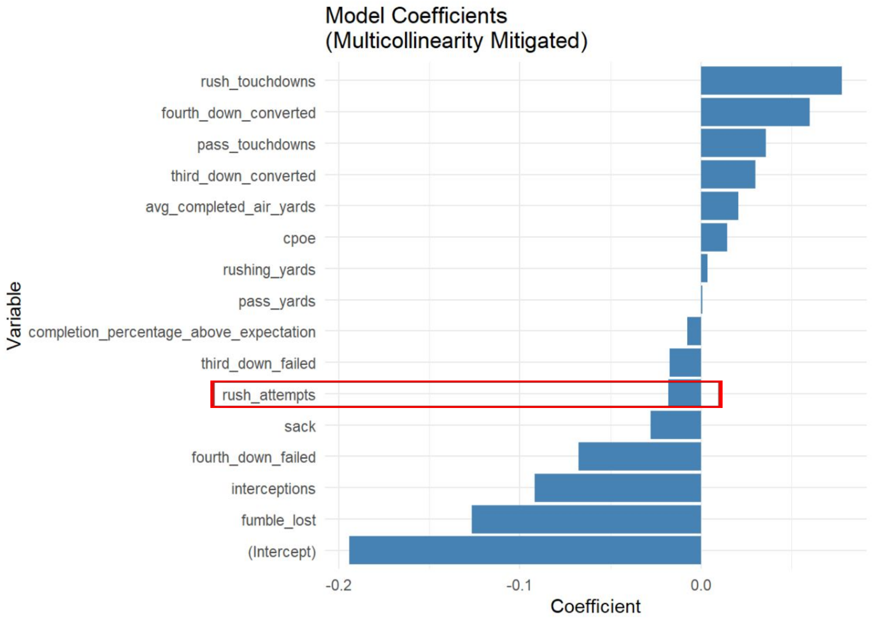
Blue variables cannot be used in smooth splines due to insufficient unique values

With all variables:
R-sq.(adj) = 0.506 Deviance explained = 52.7%
After Removing Non-Significant Variables (16 retained)
R-sq.(adj) = 0.5 Deviance explained = 50.7%
MSE: 0.1484801
Linear terms retained:
rush_attempts, rushing_yards, rush_touchdowns, fourth_down_converted, fourth_down_failed, fumble_lost, pass_yards, pass_touchdowns, interceptions, completions
Non-linear terms retained:
first_down_pass, first_down_rush, third_down_converted, third_down_failed, sack, cpoe

After Removing Highly Correlated Variables:
With all variables:
R-sq.(adj) = 0.487 Deviance explained = 50.5%
After Removing Non-Significant Variables(16 retained)
R-sq.(adj) = 0.48 Deviance explained = 48.6%
MSE: 0.1540506
Linear terms retained:
rush_attempts, rush_touchdowns, fourth_down_converted, fourth_down_failed, fumble_lost, avg_completed_air_yards, pass_touchdowns, interceptions
Non-linear terms retained:
rushing_yards, first_down_rush, third_down_converted, third_down_failed, sack, cope, pass_yards, completion_percentage_above_expectation

Coefficients of Linear models – best subset selection





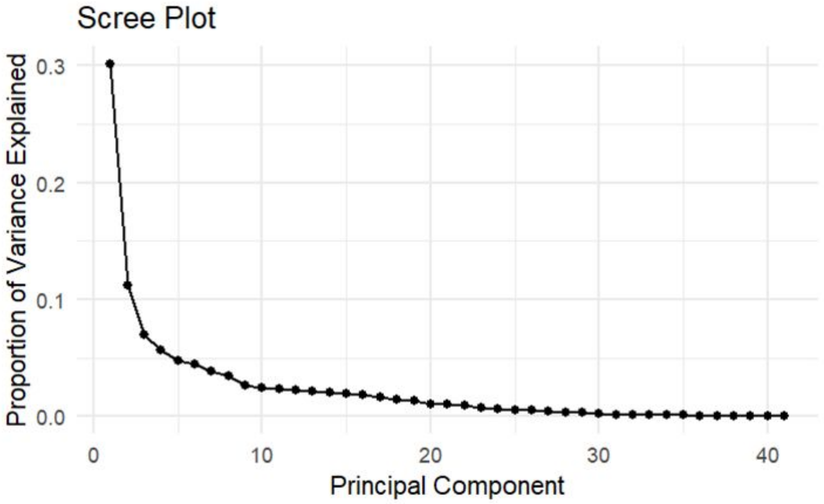
Clustering –
PCA & K - Means

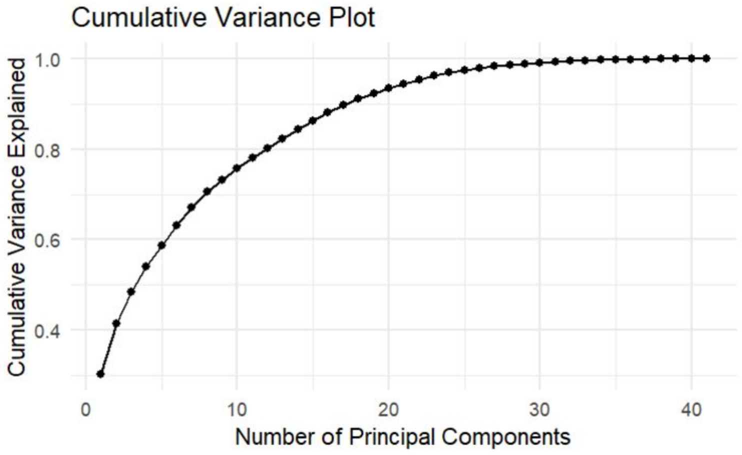
We selected 5, 10, and 13 components and applied clustering into 2 and 3 groups respectively -to find the best balance between compression and clustering quality

5 components with 2 groups has highest silhouette score and lowest within-cluster sum of squares (WCSS).

When including clustering as dummy variables in the GAM model, the cluster indicators were not statistically significant

GAM: R-sq.(adj) = 0.489 Deviance explained = 50.8%





GAM- Segmenting the Dataset via Clustering

If we choose to segment the dataset based on clustering results, we can fit separate models for each group to capture group-specific patterns.

Group 1 (~500 observations)

- Many variables lacked sufficient unique values
- Resulted in limited use of spline terms and poorer model performance

Group 2:(~3000 observations)

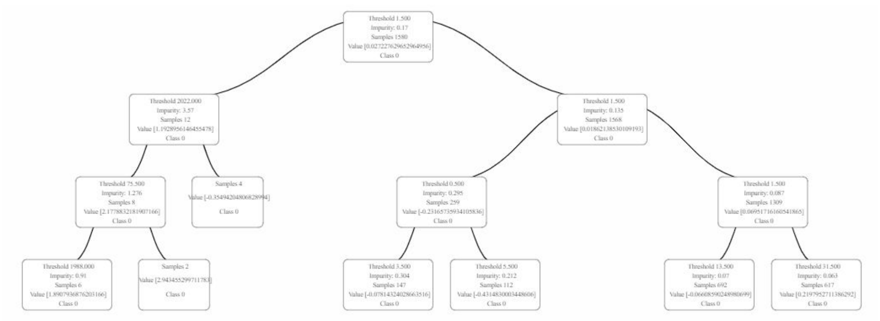
- Adjusted R^2 = 0.879
- Deviance Explained = 88.1%
- MSE = 0.0114

Further data collection and testing are needed to evaluate its robustness.

Random Forest Model

Hyperparameters used:
n_estimators = 100
max_depth=50
max_features='sqrt'
min_samples_leaf = 1
min_samples_split = 5
bootstrap=True

Tuned using GridSearchCV with 5 folds. (Minimum RMSE scoring)

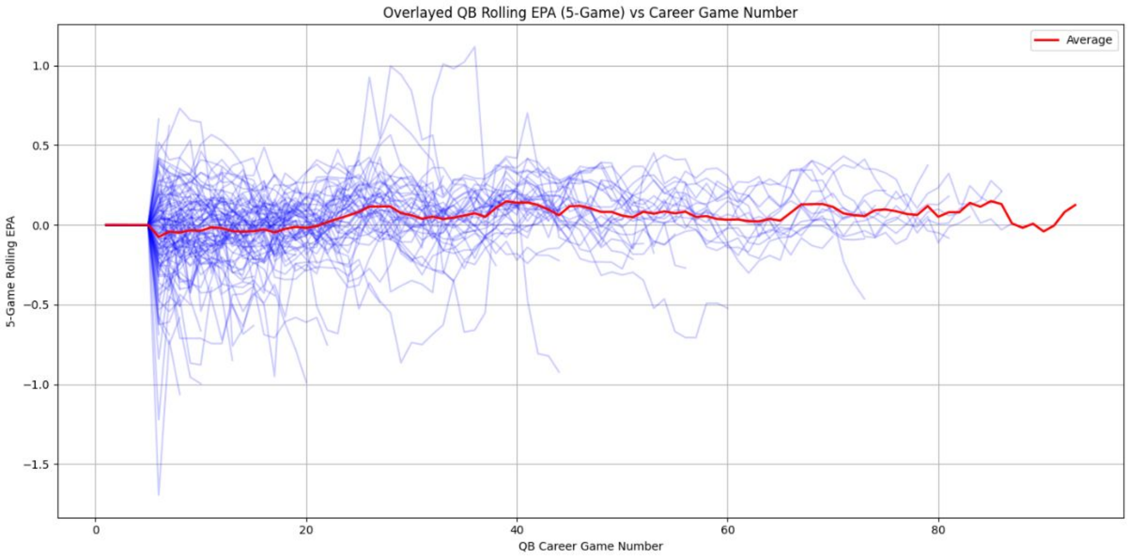


Preprocessing for Random Forest

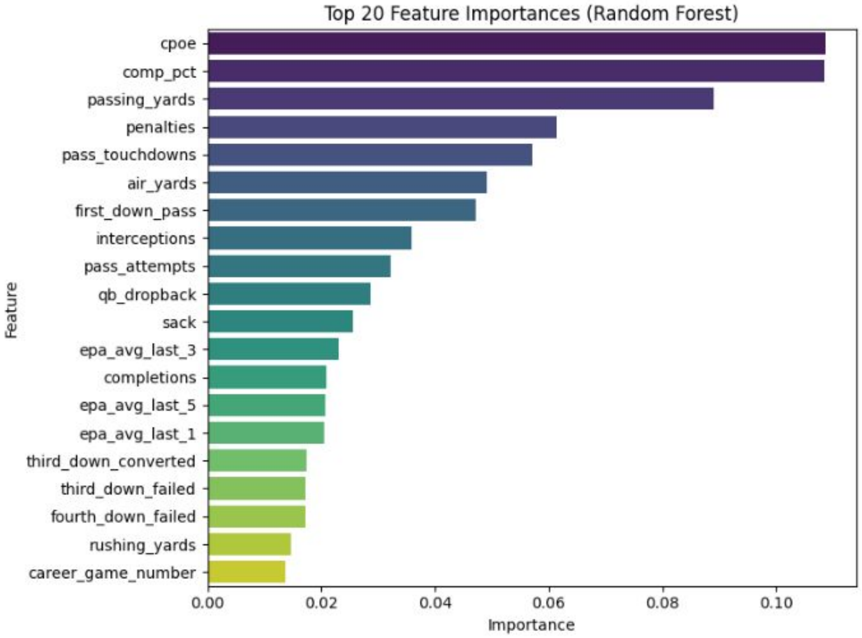
- Removed irrelevant and duplicate features.
  - Gsis-id, birth month, etc.
- Converted weeks to player career game number.
- Computed rolling EPA for player's past 1, 3, 5 games.
- Created dummy variables for team.

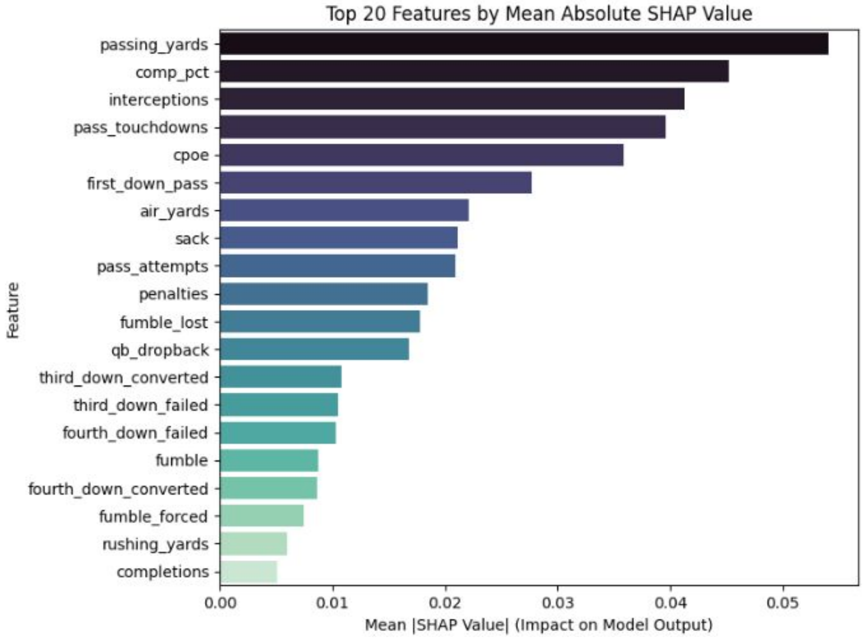
Features (QB): Season, height, weight, birth year, pass attempts, epa_avg_last, etc.

Response: avg_epa (Expected extra points added)



Feature Importance and Gain (RF)





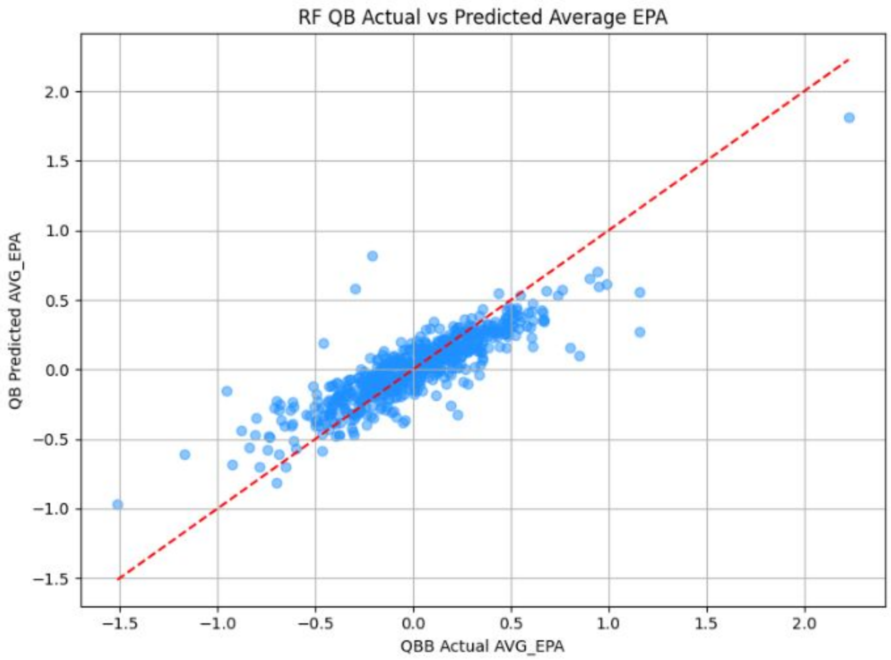
Random Forest Evaluation

RMSE: 0.1787

R-Squared: 0.7377

Adj. R-Squared: 0.7331

Predicts the test set quite more accurately than previous linear models



XGBoost Model

Hyperparameters used:
n_estimators = 200
objective='reg:squarederror'
booster = 'dart'
colsample_bytree = 0.7
learning_rate = 0.1
max_depth = 3
subsample = 0.7

Tuned using GridSearchCV with 5 folds. (Minimum RMSE scoring)

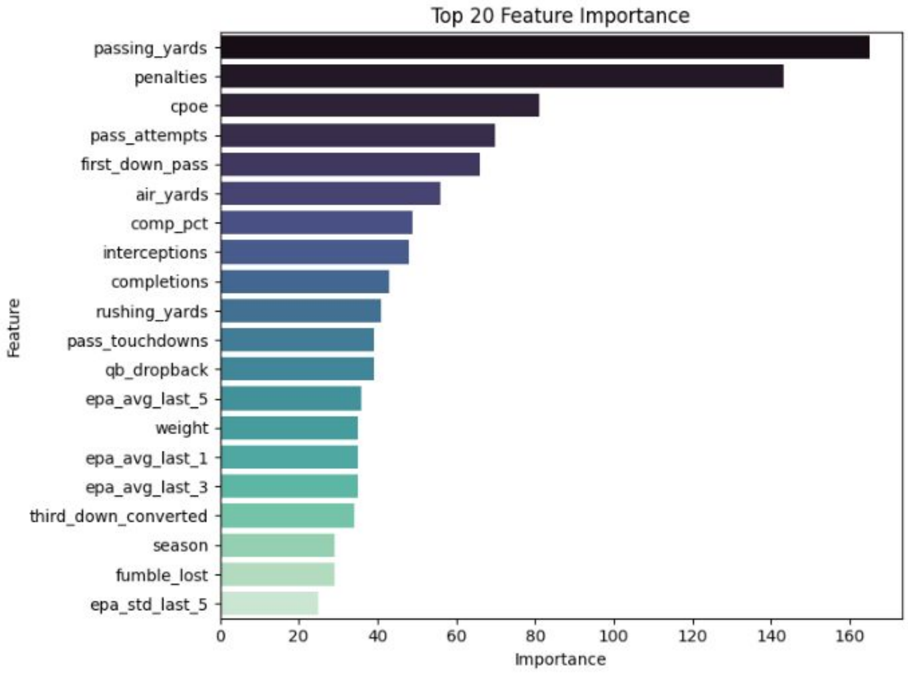
Dart booster (Dropout meet multiple Additive Regression Trees), is an extension of the default GBTree booster.

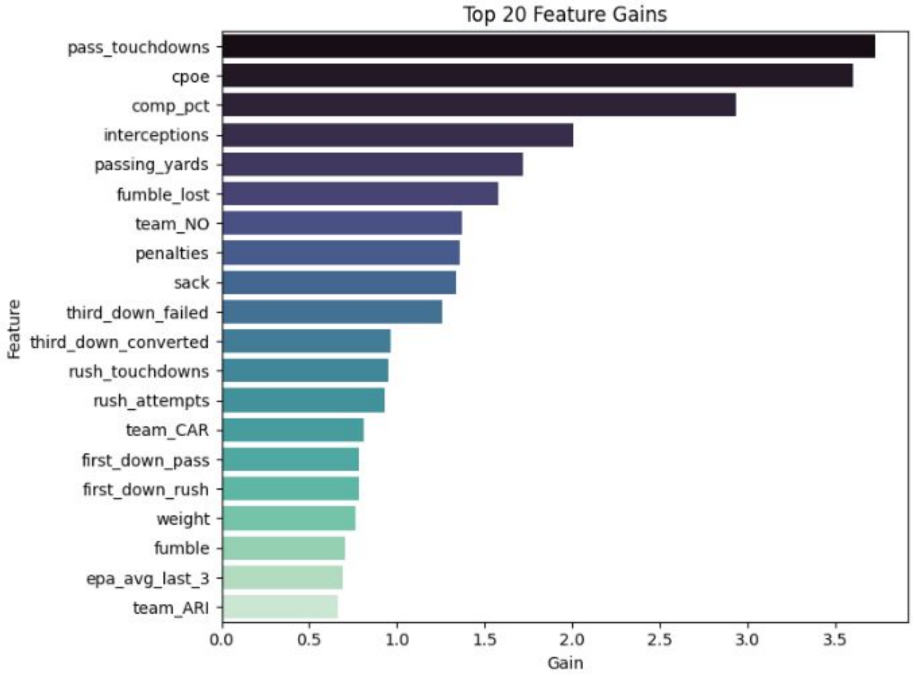
However through GridSearch CV, the best performing dropout rate is 0. (No dropout)

Besides dropout, DART still adjusts how predictions are aggregated during boosting rounds leading to much better performance using Dart over GBTree.

Future analysis required.

Feature Importance and Gain (XGB)





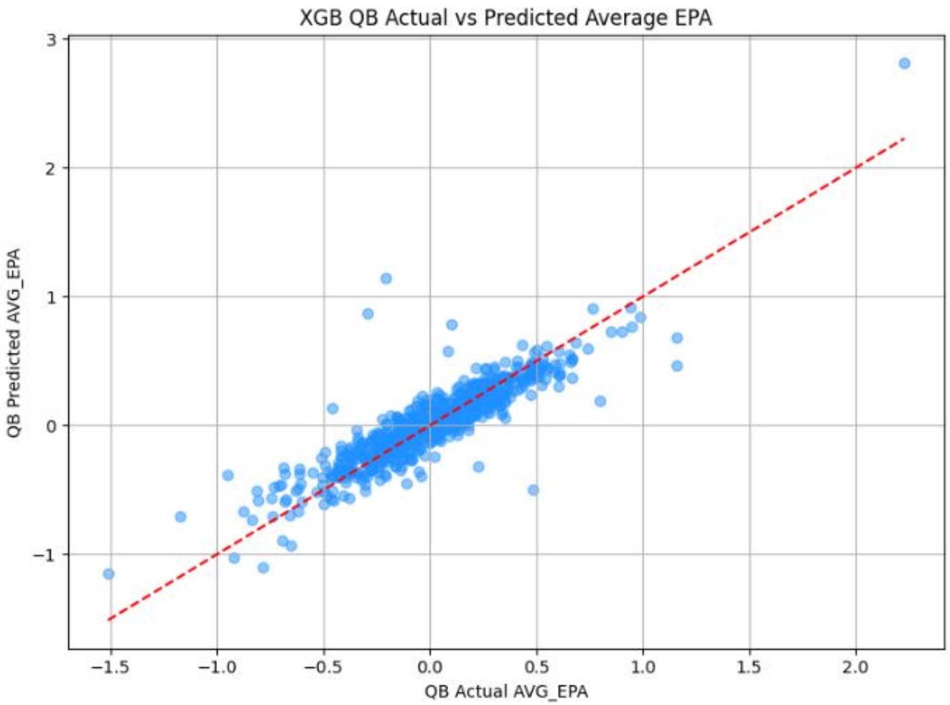
XGBoost Evaluation

RMSE: 0.1617

R-Squared: 0.7801

Adj. R-Squared: 0.7754

Improvement in prediction accuracy over standard Random Forest Regressor



Pros and Cons (RF vs XGB)

Random Forest:

- **Pros**
  - Easier to tune (Less Hyperparameters)
  - Low variance trees (Bootstrapping and random feature selection)
  - Faster training
- **Cons**
  - Lower prediction accuracy
  - Each tree is built independently (No boosting)

XGBoost:

- **Pros**
  - Better prediction accuracy (In this case)
  - Better at modeling complex nonlinear relationships
  - Built in L1/L2 regularization and Dart to prevent overfitting
- **Cons**
  - Much easier to overfit noisy data
  - More effort to tune hyperparameters
  - Slower training

Sentiment Addon

We wanted to see if player performance is impacted by outside factors such as public sentiment.

News articles, Reddit posts, comments, and expert written notes.

Athletes often struggle with not only physical stress, but also mental stress.

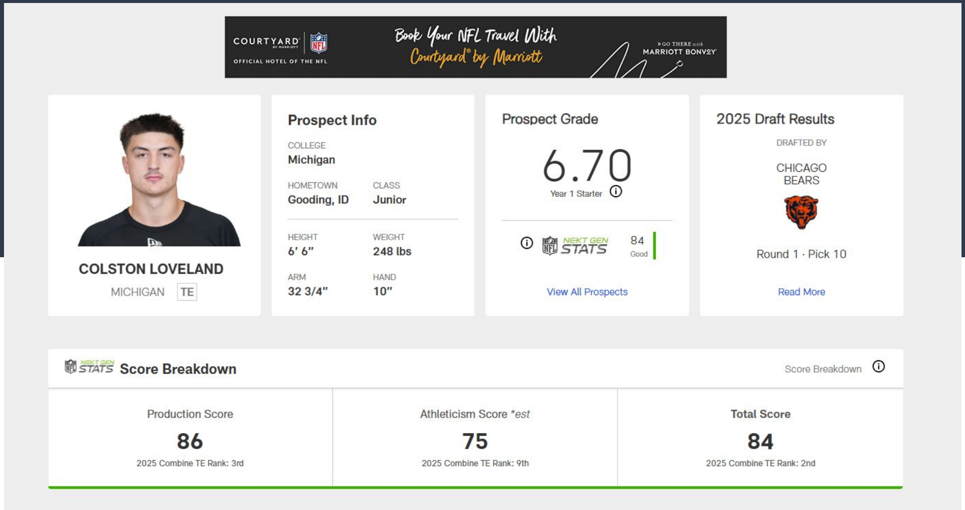
Public criticism found in news articles, forum posts, and comments might contribute to the mental stress of the players.

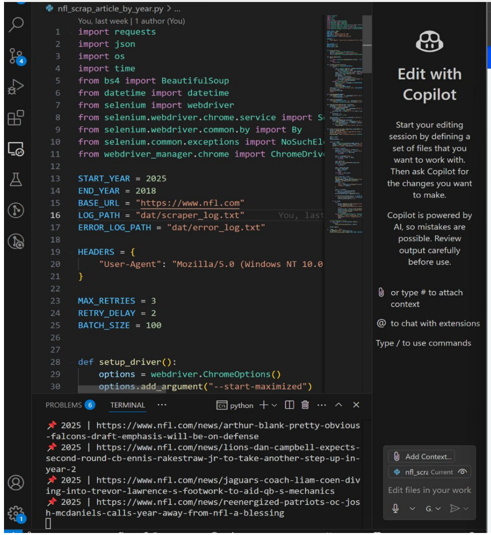
We attempt to scrape public articles, label articles by their publication, and utilize LLM to embed article text for a sentiment score.

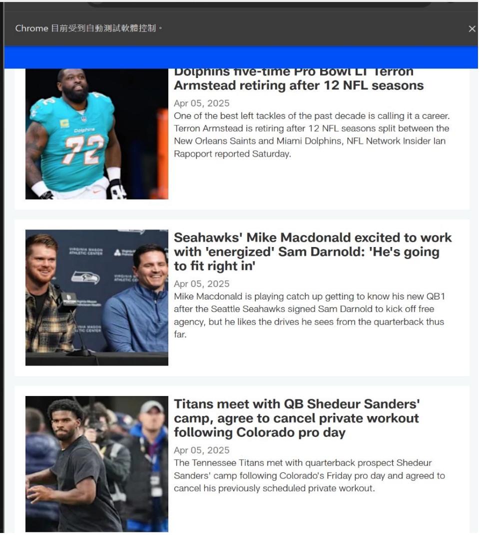
The sentiment score is used as an additional feature in the models and analyzed if significant.

Our first attempts for sentiment data

- We explored two official data sources:
  - [NFL Official Scout Reports]() → Only provide pre-draft evaluations, not dynamic or updated during player careers.
  - [NFL.com News Articles]() → They offer rich, dynamic content, but collecting large amounts of data is very challenging
- We built a Selenium crawler to dynamically scrape news articles.
  - Successfully scraped recent articles.
  - **BUT**: Required infinite scrolling from the most recent (2025) back to older years.
  - The process became very slow when we tried to collect data from all years (2018–2025).
- Ultimately, we decided to **abandon these two sources for sentiment analysis.**



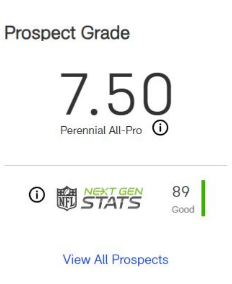


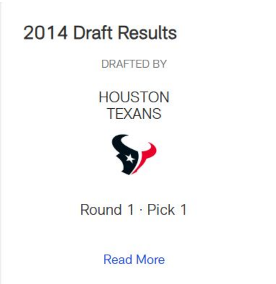


Problems with Prospect bio data

Successful in scraping player bio data. However, we ran into a few problems through EDA.

- **Biased sentiment**
  - Most of the expert written bios hold prospects in a different standing. Often use strong positive language.
  - **Survivorship bias:** Players who make it through the draft are more likely to have 4, 5 star ratings.
- **Outdated:** Since this is prospect data, the data is retrieved during the draft of each player. Might not be accurate for the current player.





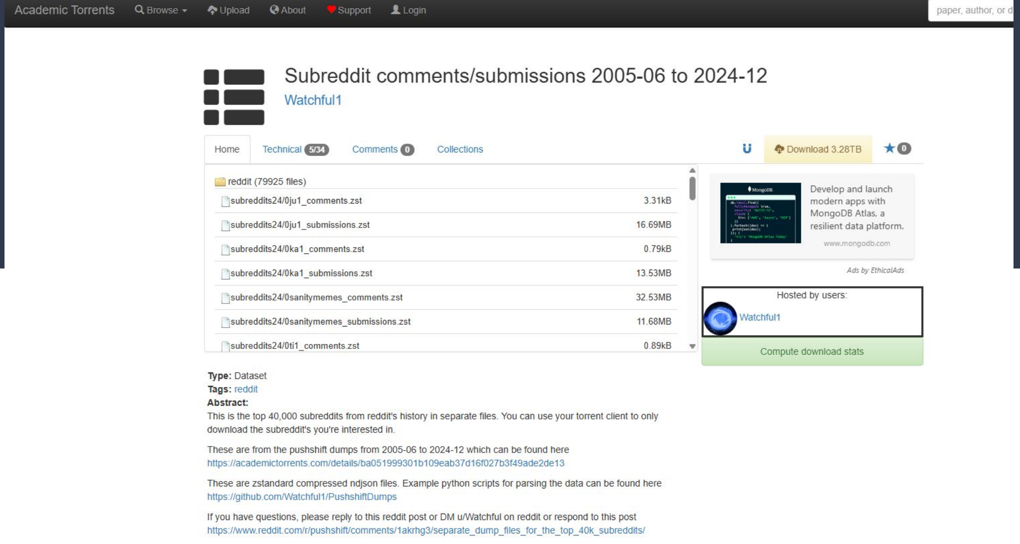
star_rating	
4.0	1586
5.0	320
3.0	114
2.0	26

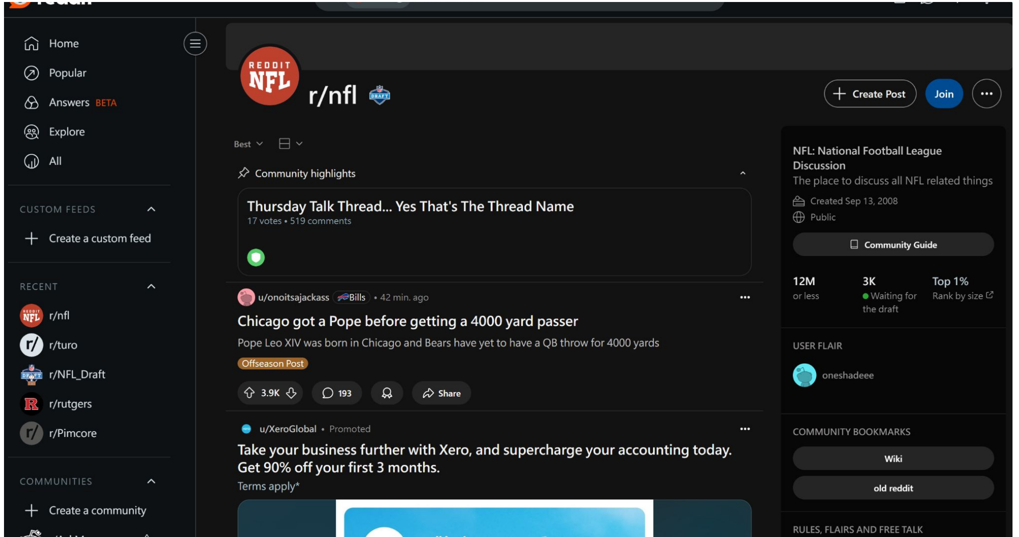
Player Bio

A physical specimen with a rare size-speed combination, Clowney was not as impactful as a junior while playing through injuries and being forced to deal with opposing offenses that fully accounted for him with extra chip protection. Was a 20-year-old junior affected by turnover on the defensive coaching staff. Could benefit tremendously from a stable positional coach and strong, veteran mentor on the defensive line who will hold him accountable, show him the way and serve as a fatherly figure. Is one of the most unique talents in the draft and could easily be a double-digit sack producer in the pros from either end. Is every bit worthy of the first overall pick -- will immediately upgrade a defensive line and improve the production of those around him.

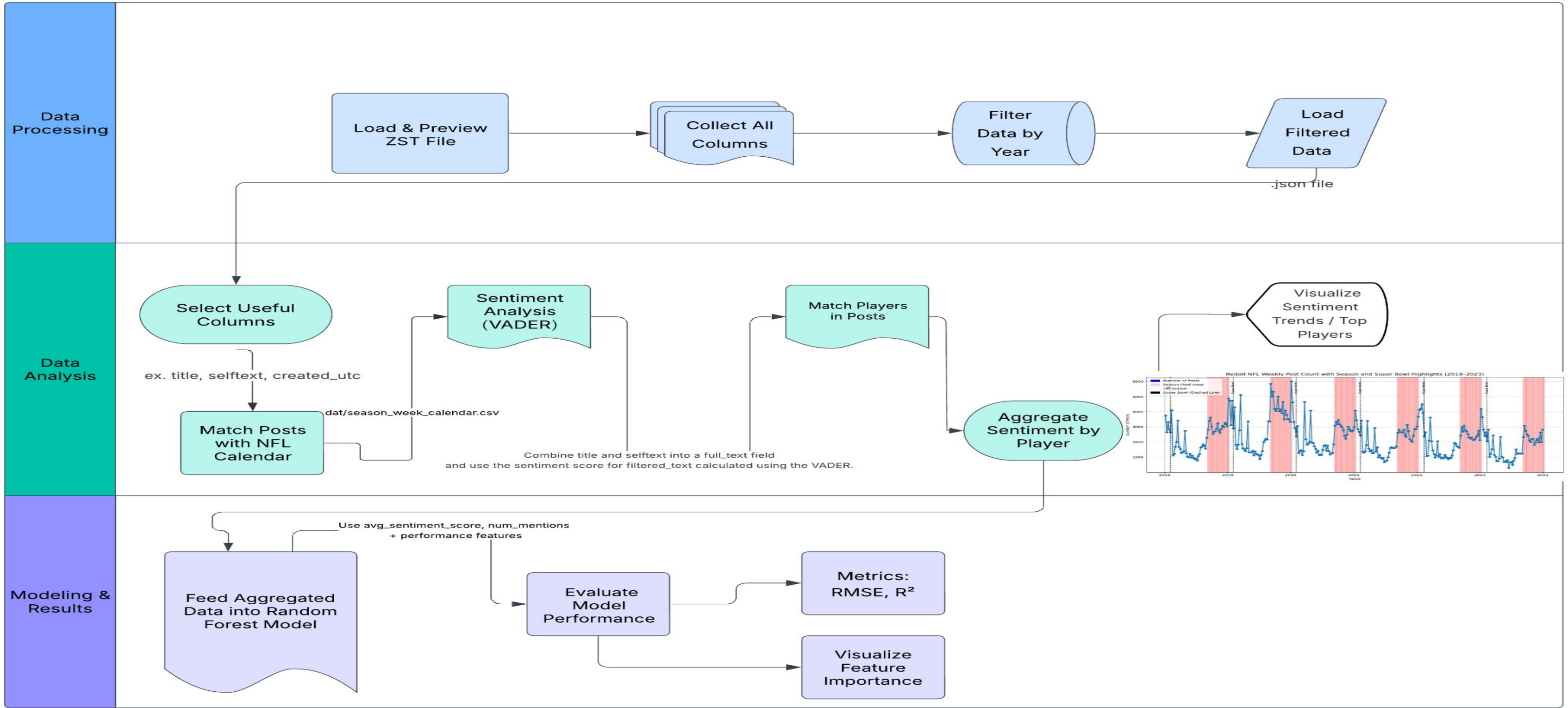
Reddit Sentiment data

- Data source:
  - Reddit full posts dataset (2018–2023), downloaded from Academic Torrents.
- Advantages:
  - Public, large-scale, and free for academic use.
  - Includes weekly fan reactions, opinions, and emotional swings during the season.
- We processed:
  - Weekly post content.
  - Player name matching.
  - Aggregated average sentiment score and mention count per player-week.



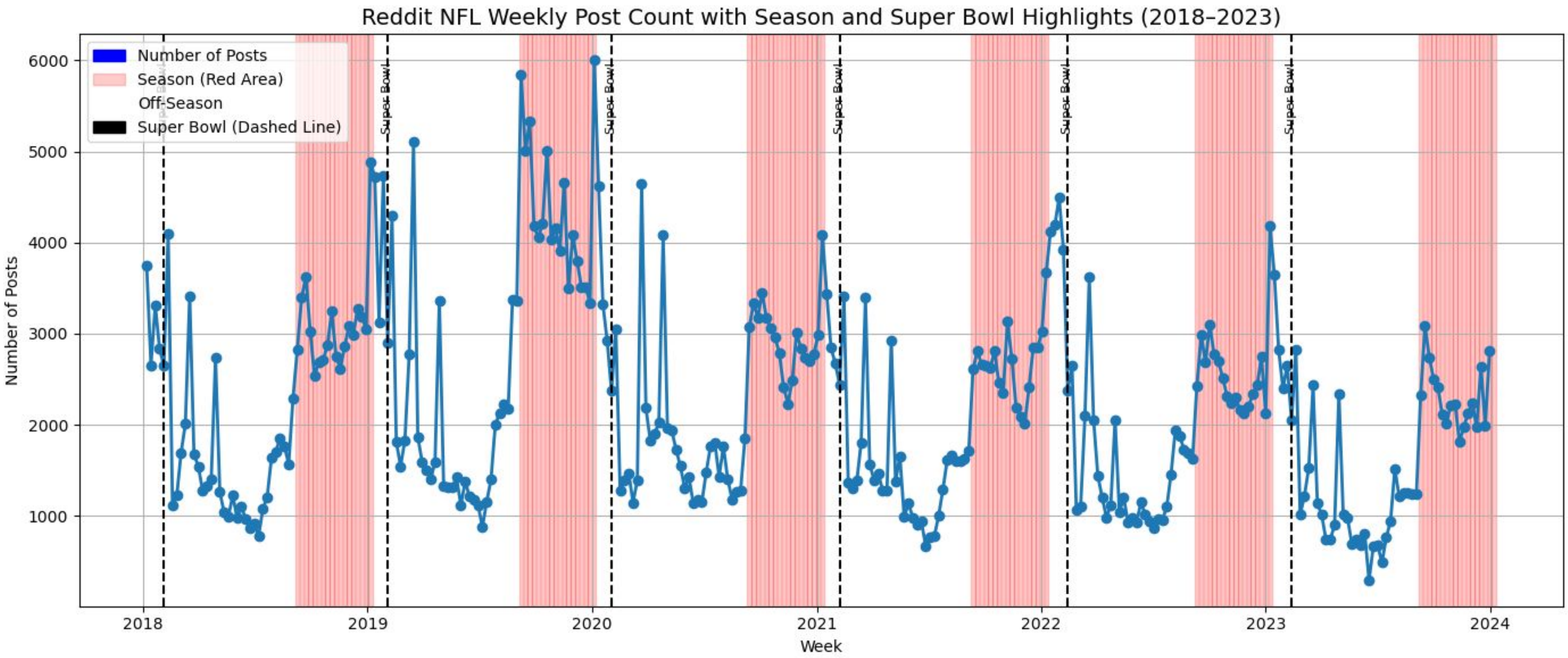


Processing Reddit Sentiment

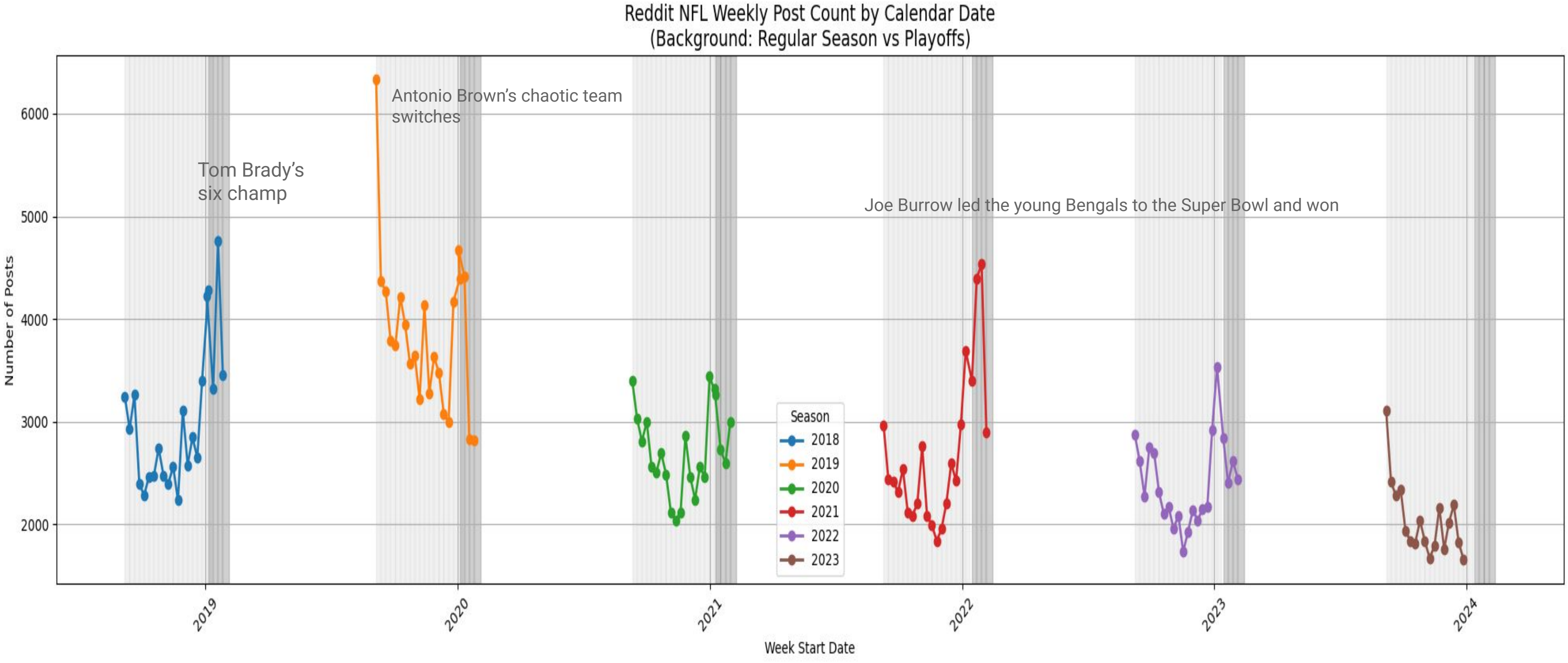


Overall Reddit Post Trends

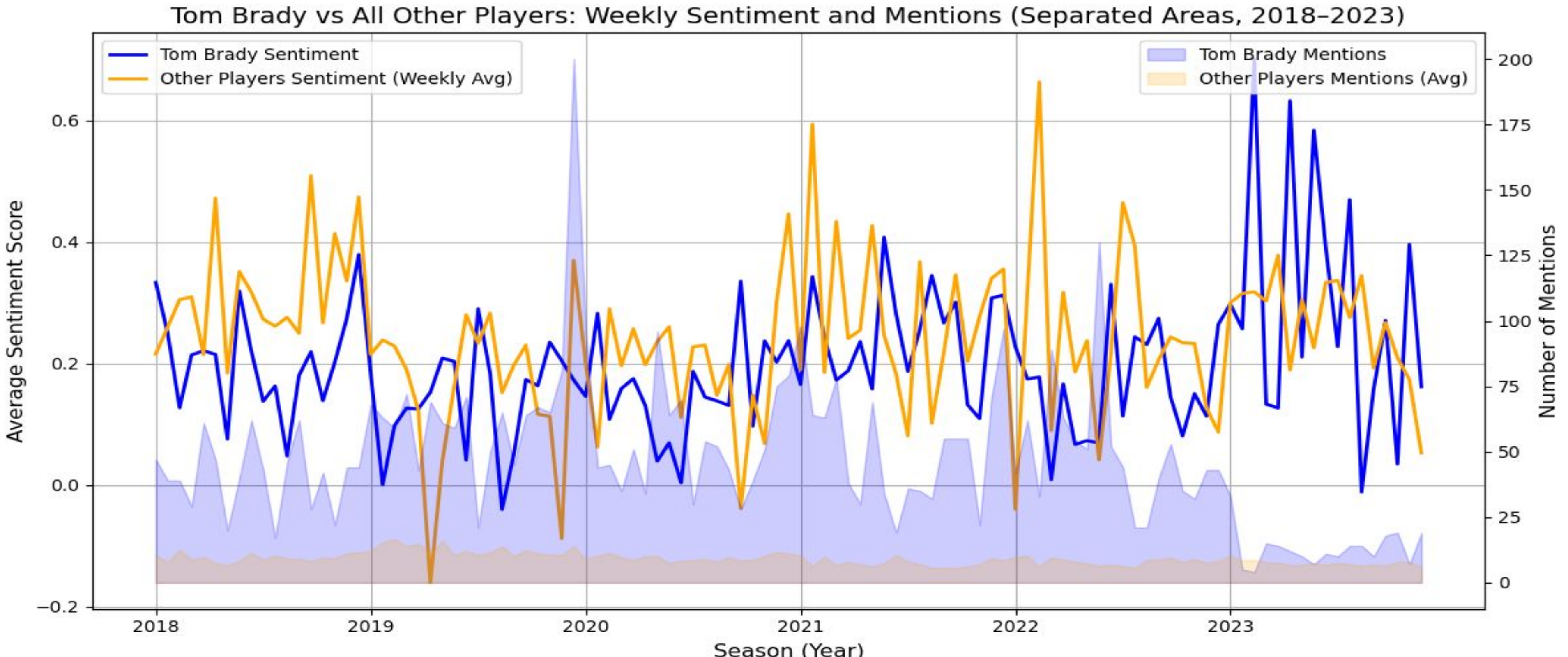
Reddit has seen a significant increase in posts during the NFL season.



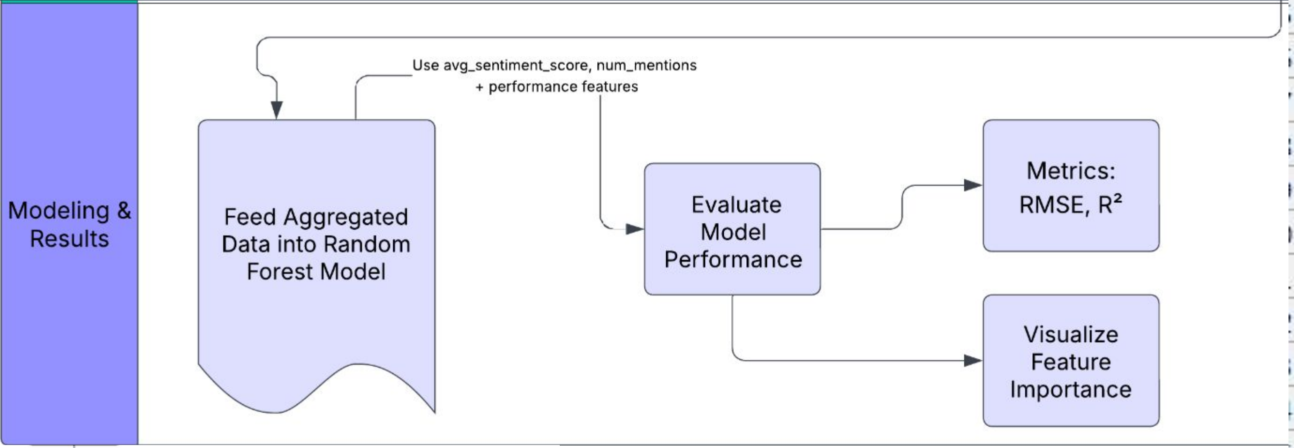
Per-Season Comparison

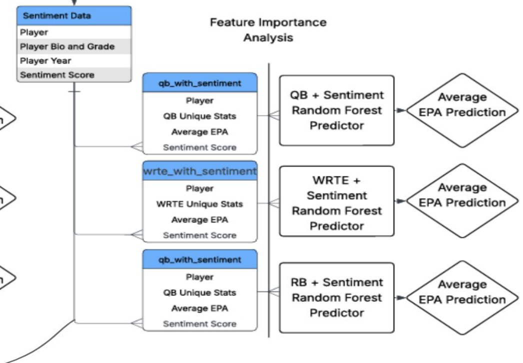


Player-Specific Sentiment



Adding Sentiment Does Not Improve Model Performance

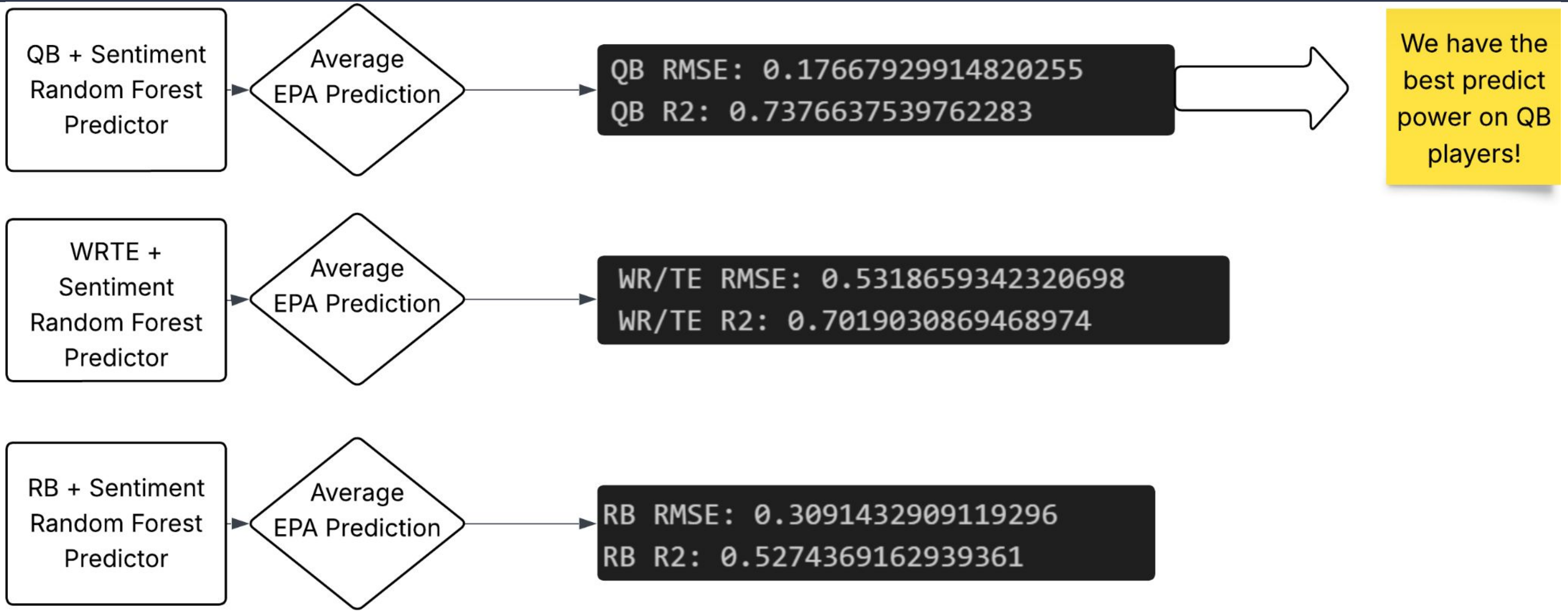




Sentiment score and number of mentions are obviously not variables that play an important role in the random forest model.

A	B	C
variables	importance	
team_KC	0.000833	
penalty_ya	0.000815	
team_NYJ	0.000808	
team_PHI	0.000797	
team_HOU	0.000787	
team_NO	0.000753	
team_LV	0.000747	
first_down_	0.000712	
team_LA	0.000557	
team_IND	0.000556	
team_LAC	0.000486	
rush_attem	0.000482	
team_GB	0.000419	
fumble_no	0.000329	
team_OAK	0.000234	
tackled_foı	0.000216	
rush_tds	0.000105	
avg_sentim	7.18E-06	
num_menti	3.89E-06	

Summary



QB + Sentiment Random Forest Predictor
Average EPA Prediction
QB RMSE: 0.17667929914820255
QB R2: 0.7376637539762283
We have the best predict power on QB players!
WRTE + Sentiment Random Forest Predictor
Average EPA Prediction
WR/TE RMSE: 0.5318659342320698
WR/TE R2: 0.7019030869468974
RB + Sentiment Random Forest Predictor
Average EPA Prediction
RB RMSE: 0.3091432909119296
RB R2: 0.5274369162939361

Explanation of Findings

- Our analysis suggests that **public sentiment from Reddit does not significantly impact player performance.**
- This may reflect **players' strong mental resilience or simply that they do not follow online sentiment about themselves.**

★ Overall, while public sentiment is an interesting social signal, it does not directly translate into predictive power for on-field performance in our models.

Thank you!