Show/Hide Code

Predicting NFL Points Total with Random Forest Regression

By

Juan Pablo DeSilva, Xinyu Hu, Bingyan Liu, Luke Stefanou, Max Zou

Abstract

The objective of this project was to build on and enhance previous prediction models of NFL game outcomes by implementing machine learning models, focusing on predicting the points total for nfl games and evaluating how these predictions compared to sportsbook lines. The analysis utilized recent team performance data, game context such as weather, and advanced modeling techniques to generate reliable predictions.

The objective was to evaluate whether machine learning could outperform sportsbooks in predicting the actual points total of NFL games. For this, we developed a Random Forest-based model to predict outcomes for future games. Using team performance data from recent games and contextual features like weather conditions, we trained Random Forest Regressors to estimate home and away team scores. The predicted points total, calculated as the sum between the away and home scores, was then tested on games from this past season. This model provides a systematic approach to predicting game outcomes, demonstrating its practical application for prospective analysis.

By leveraging exploratory data analysis to understand our features, this project highlights the ability of statistical learning to both evaluate existing prediction benchmarks, such as sportsbooks, and to provide standalone predictions for future games. These findings underscore the potential for advanced analytics in sports prediction and decision-making.

Problem Statement

Accurately predicting the points total—the sum of the points scored by both teams—is a critical challenge in sports analytics. Sportsbooks establish betting lines based on historical trends, advanced statistical modeling, and market adjustments. However, it remains unclear whether machine learning models, using structured team performance data and game context variables, can match or even surpass sportsbooks in predictive accuracy.

This project aims to address that gap by developing a Random Forest-based regression model to predict NFL game points totals and evaluating its performance against sportsbook lines.

Why is This Problem Relevant?

The motivation for this project stems from the increasing use of data science in sports analytics and betting markets. Sportsbooks adjust their lines based on market movement and bettor behavior, while machine learning models rely solely on historical performance data. By comparing model predictions with opening sportsbook lines, we can assess whether purely data-driven approaches can provide an edge over traditional methods.

Beyond the betting market, accurate point total predictions have applications in game strategy analysis, fantasy sports, and team performance evaluation. This study explores the feasibility of machine learning as a viable alternative for forecasting game outcomes in a structured, replicable manner.

Research Question

Can machine learning models accurately predict the points total of NFL games using team-level performance metrics, play-by-play data and advanced football statistics ie. Expected Points Added (EPA), turnover differential etc. from the previous 8 to 12 games, at a rate that beats sports books 53% of the time?

Background and Prior Work

Early Statistical Models

Predicting the **points total** in NFL games has become a key area of sports analytics, particularly in betting markets. Unlike point spreads, which focus on relative performance between teams, total points forecasting requires capturing offensive efficiency, defensive strength, and game tempo while accounting for external factors such as weather.

Early Statistical Models Initial research into NFL scoring distributions established fundamental principles for total points prediction. Stern (1991) showed that NFL game scores roughly follow a normal distribution, reinforcing the efficiency of betting markets.[1] Glickman & Stern (1998) introduced state-space models, which dynamically adjusted offensive and defensive team ratings over time, improving score prediction accuracy compared to static methods.[2]

Baker & McHale (2013) advanced this approach with a point process model, which analyzed scoring events in real-time rather than relying on game-level aggregates. Their study demonstrated that their model's predicted totals closely aligned with sportsbooks' closing lines, highlighting the difficulty of gaining a consistent predictive edge over betting markets.[3]

Simultaneously, external factors were identified as key contributors to total points variance. Borghesi (2008) found that sportsbooks systematically underestimated the impact of extreme weather on low-scoring games, presenting an inefficiency that data-driven models could exploit.[4]

Machine Learning and Advanced Metrics

Modern approaches to total points prediction leverage machine learning models and advanced football analytics. Metrics such as Expected Points Added (EPA), turnover differential, red zone efficiency, and pace of play have significantly enhanced predictive accuracy.

Lock & Nettleton (2014) applied Random Forest models to estimate in-game win probabilities, demonstrating the power of machine learning to capture complex interactions between game-state variables.[5] More recent research indicates that ensemble models (Random Forest, Gradient Boosting) consistently outperform linear regression in forecasting points totals, as they better handle non-linearity and interactions between team performance metrics.[6]

One challenge for machine learning models has been capturing game flow and strategic adjustments. Traditional models assume independent scoring events, but real-world factors—such as teams shifting to conservative play-calling in late-game blowouts—affect total points outcomes. This has led to the incorporation of situational modeling, where AI systems analyze game conditions to adjust predictions dynamically.

Industry Benchmarks and Challenges

Sportsbooks establish totals using power-rating models that factor in historical performance, team strengths, injuries, and betting patterns. These lines continuously adjust in response to new information and betting activity, making them dynamic benchmarks that machine learning models must compete against.

Despite advances in machine learning, sportsbooks' closing totals remain highly accurate, as they reflect both quantitative models and real-time market sentiment. Studies have found that machine learning models typically match but rarely exceed the accuracy of sportsbook lines, underscoring the difficulty of extracting additional predictive value from public data.[7]

References

1. ^ Stern, H. (1991). On the probability of winning an NFL game. The American Statistician. https://www.jstor.org/stable/2684286
2. ^ Glickman, M., & Stern, H. (1998). A State-Space Model for National Football League Scores. Journal of the American Statistical Association, 93(441). https://www.jstor.org/stable/2669599
3. ^ Baker, R. D., & McHale, I. (2013). Forecasting exact scores in National Football League games. International Journal of Forecasting, 29(1), 122–130. https://www.sciencedirect.com/science/article/abs/pii/S0169207012001070
4. ^ Borghesi, R. (2008). Weather Biases in the NFL Totals Market. Applied Financial Economics, 18(12), 947–953. https://www.researchgate.net/publication/24071150_Weather_Biases_in_the_NFL_Totals_Market
5. ^ Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. Journal of Quantitative Analysis in Sports, 10(2). https://dr.lib.iastate.edu/server/api/core/bitstreams/17e429ca-3ad5-481a-b06a-a19f872c6351/content
6. ^ Bunker, R., & Susnjak, T. (2022). The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review. https://www.jair.org/index.php/jair/article/view/13509
7. ^ SharpFootballAnalysis (2023). What is an Over/Under Bet? https://www.sharpfootballanalysis.com/sportsbook/guides/over-under-bet/

Hypothesis

Our hypothesis is that our machine learning models will predict NFL game points totals with an accuracy of 53% or higher when compared to sportsbooks' opening lines. Prior research has demonstrated that statistical and machine learning approaches can uncover patterns in team performance that influence scoring outcomes. Baker & McHale's point process model showed that structured modeling of scoring events can closely match sportsbook projections, while Borghesi identified inefficiencies in how weather impacts total points. Additionally, Lock & Nettleton's application of Random Forest models to in-game win probability highlights the potential of machine learning to process complex interactions in football data.

Building on these findings, this project will integrate ensemble learning techniques with key predictive metrics—including offensive and defensive Expected Points Added (EPA), and turnover differential—to evaluate their impact on total points forecasting. By leveraging early-week data and assessing predictions against sportsbooks' opening totals, the goal is to determine whether machine learning can provide a measurable advantage in forecasting game outcomes before market adjustments incorporate additional external factors.

Data

Data overview

- Dataset #1
  - Dataset Name: nfl-data-py: aggragated play-by-play game-level data
  - Link to the dataset: nfl-data-py
  - Number of observations: 1139
  - Number of variables: 46
- Dataset #2
  - Dataset Name: rotowire: Historical NFL Sportsbetting data
  - Link to the dataset: rotowire
  - In the directory as `bettingData.json`
  - Number of observations: 10632
  - Number of variables: 33
- Dataset #3 (Final Dataset)
  - Dataset Name: model_game_data
  - No link available as it was merged in this notebook
  - Number of overvations: 1006
  - Number of variables: 55

To build a game-level dataset, we aggregated play-by-play data to capture overall team performance per game, moving beyond individual plays to focus on outcomes that impact game predictions. Splitting variables into home and away metrics and grouping them by game ID allows for the assessment of each team's performance within a single game, making it easier to compare offensive and defensive outcomes. Important variables include home EPA and away defensive EPA allowed, which is calculated by summing EPA values for plays where the team is on defense, offering a measure of defensive effectiveness. The model will also use turnover differential (turnovers forced - turnovers committed) to quantify each teams ability to capitalize on or limit mistakes, a factor closely related to winning

outcomes. A vast majority of the variables are continuous numbers (ie. points scored, yards gained, turnover differential) while the variables whose datatypes are objects are meant to serve as identifiers and provide contextual information (game week, weather). The only datatype that needs conversion is time of possession from **minutes:seconds** to **seconds** in order to have a continuous variable rather than an object. This approach results in a dataset tailored for modeling, emphasizing game-level dynamics.

The dataset from rotowire contains opening lines that capture the predicted odds of each NFL game based on past performance. This data supplements the game-level dataset, adding variables such as "game_over_under", "total", "over_hit", "under_hit", "favorite_covered", "underdog_covered", to provide reliable benchmarks off of which to compare the predictions. The decision to use current data (2021 to 2024) was made to close in on games that have the highest likelihood of projecting a trend that will predict upcoming game lines based on the model. model_game_data is the result of merging the two datasets, and contains all of the information necessary to carry out an insightful exploration. This combined source of data will be used to gain insight into how these sportsbooks arrived at the odds they set, and how the data can be used to predict a more accurate line, hopefully confirming the hypothesis in the process.

nfl-data-py: aggregated play-by-play game-level data

In [36]:

```
# Import the play-by-play data
```

In [ ]:

The play-by-play data for each game is raw and needs preparation before it can be modeled for the use case. To start we can see that the data type for `drive_time_of_possession` is an object. We need to convert this column to integers in order to have each teams' drive time of possession be represented in seconds as continuous numbers.

```
# Convert 'drive_time_of_possession' from minutes:seconds format to total seconds
# This makes it easier to perform numerical operations later.
```

In [ ]:

We have identified where each new drive begins. This shows that the first row of every drive will alternate the team for the `posteam` column. After the first row, the `drive_time_of_possession` and the `posteam` ID repeat for the following rows for the duration of that ongoing drive. This will remain the same until a new drive starts where the `posteam` value will then shift back to the other team and a new value for `drive_time_of_possession` will repeat. We will flag each time this happens to correctly identify when the ball changes possession.

```
# Flag rows where a new drive begins
# Calculate time of possession based on flagged new drives
```

In [ ]:

The column `new_drive` will have the value `True` and the corresponding time of possession for that drive. All remainding rows for that drive will have the value `False`. Our data is now ready to be separated into home and away team metrics. This will provide a clear distinction for when each team is on offense or defense. We will do this by identifying which team has possession of the ball for any given play.

```
# Split data into possessions for home and away teams
```

In [ ]:

In [ ]:

Now that we have split the plays into home and away possessions, we can aggregate the data for each game. This will ensure that each row in the resulting datasets represents a team's performance for one full game, rather than each play.

```
# Aggregate performance metrics for home and away teams
```

In [42]:

Now, we can calculate key features such as turnover and sack differentials to quantify game dynamics. These differentials help assess which team had a greater impact in forcing turnovers or pressuring the quarterback. Additionally, we can analyze efficiency metrics like third-down conversion rates, explosive play frequency, and average time to throw to gain deeper insights into offensive and defensive performance. By integrating these features, we can better understand the factors that influence game outcomes and team effectiveness.

```
# Calculate turnover and sack differentials for home and away teams
```

In [43]:

It's time to merge our home and away datasets to complete each game. We will aggregate defensive data, include contextual information and fix any inconsistencies.

```
# Merge home and away team data into a single dataset
```

In [ ]:

We can see now that after all the changes have come into effect the data not only is more visually appealing but much more apt for analysis.

rotowire: Historical NFL Sportsbetting data

The rotowire dataset needs less manipulation than the nfl play-by-play data. Here we just need to take the data for the corresponding sportsbook predictions from 2021 to 2024 and include the relevant columns needed to compare our predictions.

In [45]:

model_game_data: Merging rotowire and nfl_data_py

Now that both of our datasets are ready for exploration, we can merge them together to have one all inclusive dataset.

In [ ]:

Variable Clarifications

Some of the variables are not entirely obvious so there will naturally be discusison about how these are relevant data variables for us when determining our goal.

- **`weather`**: Weather can be very impactful in a game. For example if it was snowing/raining things like turnovers would naturally become more likely, so there would need to be a weight attached to relevant variables. Additionally, games will likely be lower scoring when there is very poor weather.
- **`game_over_under`**: describes the points total line set by the sportsbook. This is the sportsbook prediction of the sum of the scores of the two teams. This will be helpful when we analyze the data, allowing us to see how accurate the sportsbook was and use it as a benchmark to determine the accuracy of our own predictions.
- **`EPA`**: Expected Points Added is one of the most advanced and insightful football metrics available. It is a statistical metric that is used to evaluate the impact of individual plays on a team's likelihood of scoring. Its ability to contextualize performance makes it indespensable for analytics in football. Integrating EPA into our model will allow for deeper insights into game outcomes and points totals.

Data complementation

There are two datasets with key distinctions: The first dataset consists of game performance metrics. It is effectively the raw concrete data that is available for the results of games and what many would use to predict game results. The variables from rotowire are focused on the sportsbook predictions and the relevant betting outcomes in the past. This allows past results to be used as a benchmark off of which to base our predictions. We will be able to predict game outcomes solely based off of our manipulated complete game dataset, and will then use the opening lines from the rotowire dataset to compare the model's predictions.

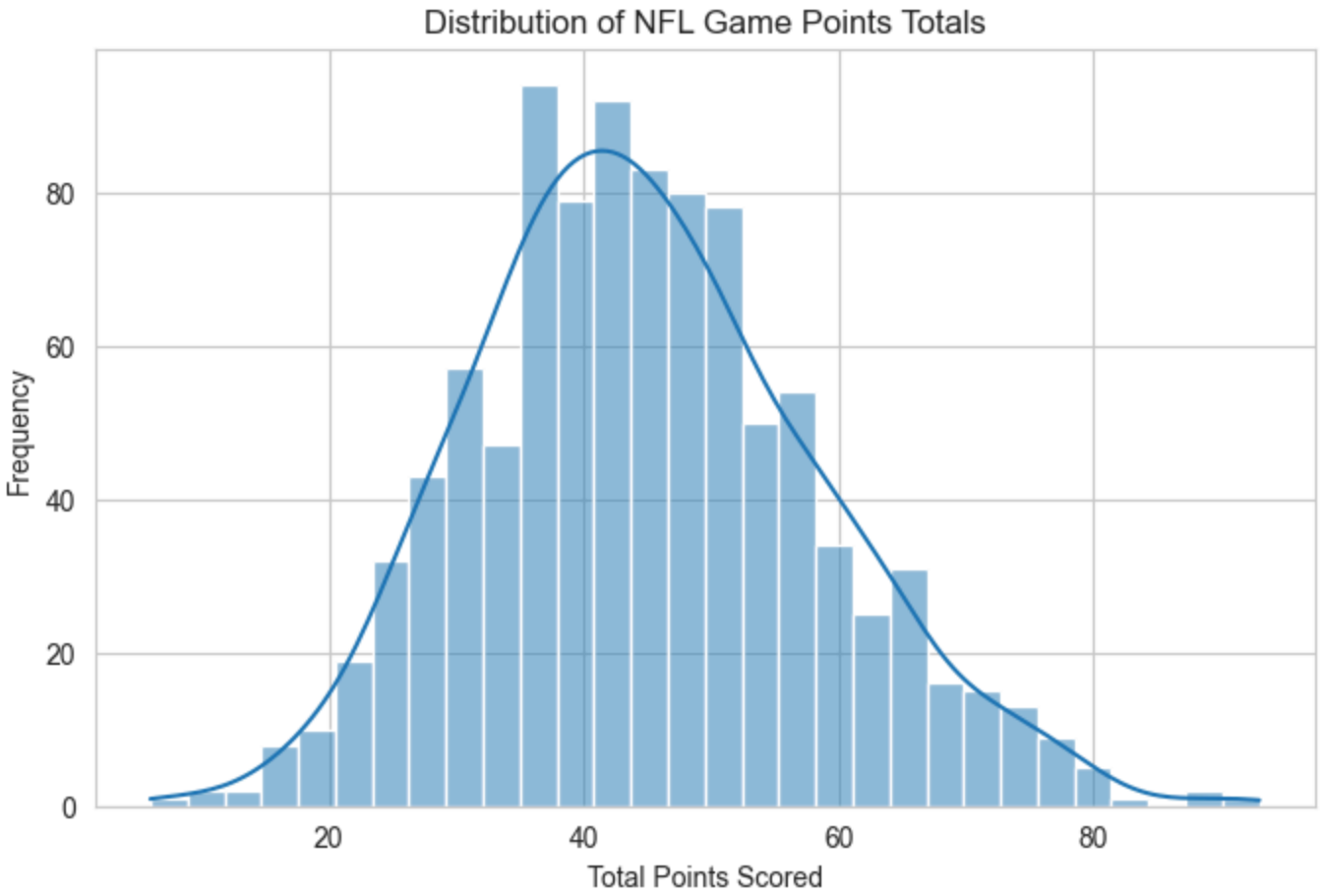
Results

Exploratory Data Analysis

The following is a series of visualizations that will allow us to better conceptualize the relationship between key features

Distribution of Total Points in NFL Games

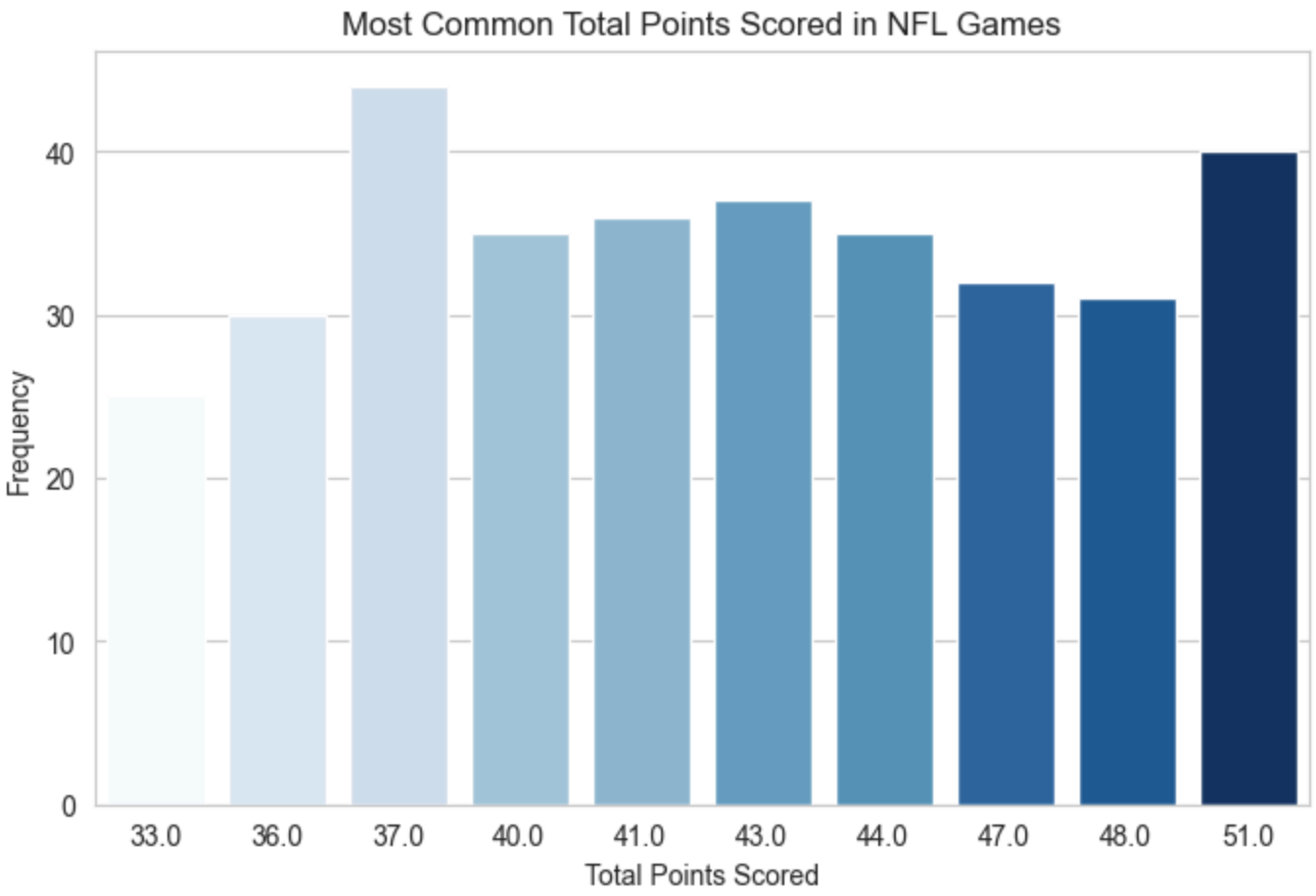
In [47]:



The histogram of NFL game points totals reveals a clear distribution pattern shaped by the structure of football scoring. The overall shape resembles a **slightly right-skewed normal distribution**, indicating that most games tend to have moderate scoring totals, with fewer games resulting in extremely high or low scores. However, a closer look reveals noticeable peaks at specific total values, which align with the fundamental ways points are scored in the NFL.

In football, points are earned through touchdowns (7 points including the extra point), field goals (3 points), and safeties (2 points). Because these scoring increments naturally combine in predictable ways, certain totals become more common than others. For example, let's examine the scores in the visualization below:

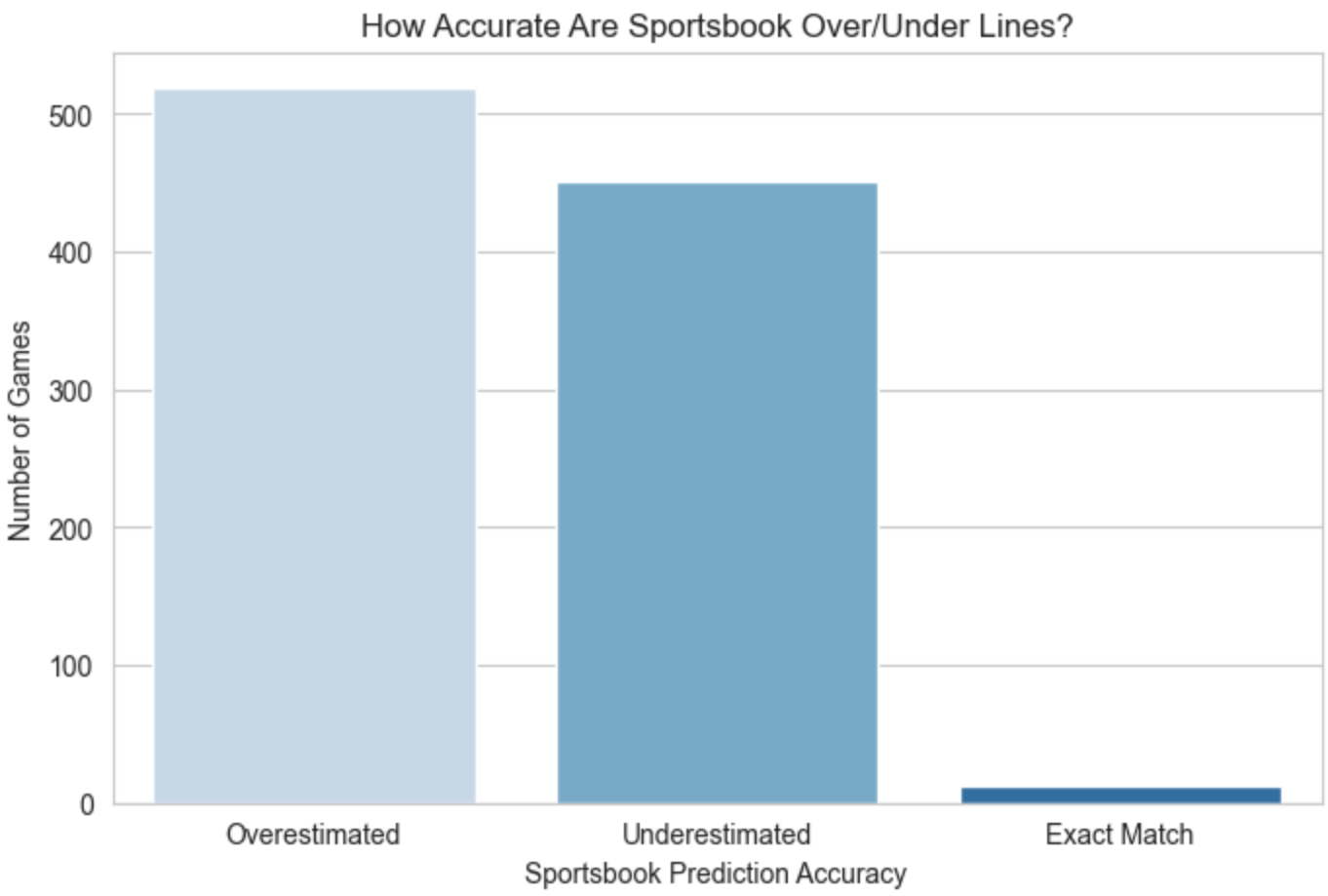
In [48]:



The scores around 37, 44, and 51 points are relatively frequent because they are easily formed through a mix of touchdowns and field goals. This pattern is an important consideration when building a predictive model—we are not simply modeling a continuous numerical distribution but one influenced by the discrete nature of scoring events. Unlike other sports where scoring is more fluid, football's structured point system introduces a level of granularity and clustering that traditional regression models may not fully capture without incorporating categorical adjustments. Understanding these key scoring patterns allows us to refine our approach by incorporating domain knowledge into our feature engineering process.

How Accurate Are Sportsbooks At Predicting Over/Unders?

In [49]:



The initial visualization provides insight into how accurately sportsbooks predict the total points in NFL games. The distribution reveals that sportsbooks most frequently **overestimate** game totals, followed closely by **underestimations**, while exact matches are extremely rare. This pattern suggests that sportsbooks are more likely to set their over/under lines above the actual total points scored, rather than accurately predicting or underestimating them. While minor deviations from perfect accuracy are expected in any predictive model, the consistent tendency toward overestimation raises the question of whether this pattern is the result of **random chance or a deliberate bias in sportsbook line-setting.**

To determine whether sportsbooks systematically misestimate game totals in one direction, we conducted a **chi-square goodness-of-fit test.** This test evaluates whether the observed distribution of overestimated, underestimated, and exact matches significantly deviates from an expected neutral distribution, where each category would occur with similar frequency. The results of this test provide a statistical foundation for understanding whether sportsbooks are **intentionally setting totals higher than expected game outcomes or if this pattern is an artifact of random variation.**

In [50]:

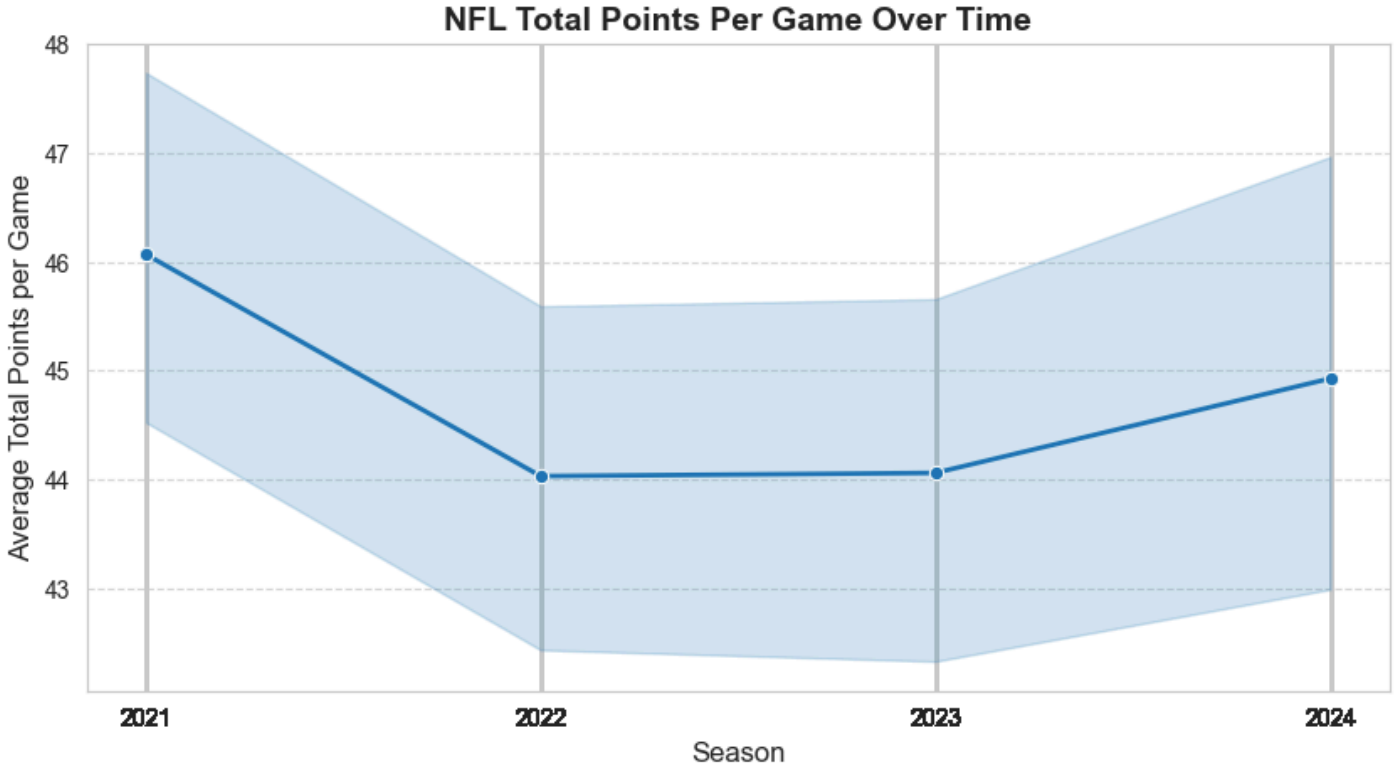
```
Chi-Square Statistic: 460.33
P-value: 0.000000
The differences between sportsbook predictions are statistically significant (p < 0.05).
```

The test results ($\chi^2 = 460.33$, $p < 0.0001$) confirm that sportsbooks systematically overestimate total points rather than setting unbiased lines. This is not due to randomness—sportsbooks appear to **intentionally inflate totals**, likely in response to **public betting preferences**. Research shows that **bettors favor the over**, as high-scoring games are perceived as more exciting. By setting lines slightly above expected outcomes, sportsbooks can **exploit this bias**, ensuring more wagers on the over while

maintaining their edge. The rarity of exact matches further suggests that sportsbooks prioritize **betting market balance over precise predictions.**

NFL Total Points Per Game Over Time

In [51]:



NFL scoring trends reveal a notable shift. The sharp decline from 2021 to 2022, with totals dropping from over 46 to around 44, suggests defensive adjustments such as the rise of two-high safety coverages limiting explosive plays. Offensive inefficiencies, personnel changes, or a decline in quarterback play may have also contributed to the scoring dip.

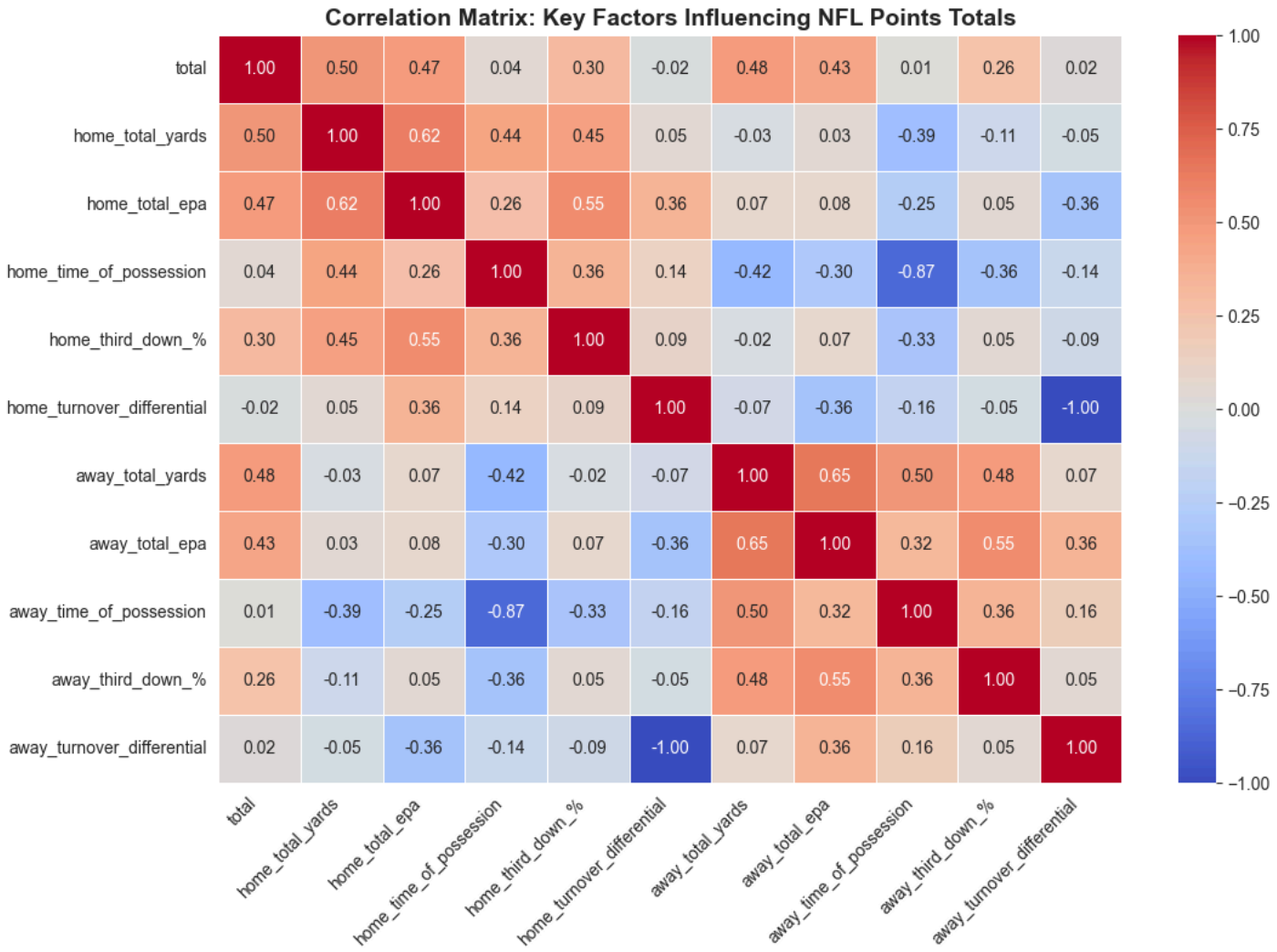
In 2023, scoring stabilized, suggesting defenses maintained their edge or offenses struggled to adapt. Sportsbooks likely adjusted their over/under lines accordingly.

By 2024, totals rebounded, possibly due to rule changes, improved quarterback play, or offensive adaptations. If this trend continues, offensive innovation may be outpacing defensive adjustments, affecting both our model and sportsbook accuracy.

For prediction purposes, recent seasons should be weighted more heavily. A decline suggests sportsbooks overestimate scoring, while a resurgence may lead to underestimation if models rely too much on past data. Identifying key drivers behind these shifts will improve game total predictions.

Correlation Matrix

In [53]:



The correlation matrix reveals key drivers of total points in NFL games. Total yards and EPA show the strongest positive correlation, confirming that offensive efficiency is critical for predicting scoring. Time of possession, however, has almost no correlation, suggesting that one team controlling the clock does not necessarily lead to more points, since the other team won't have the ball and thus fewer opportunities to score. Third-down conversion rates have a moderate impact, as teams that sustain drives tend to score more. Surprisingly, turnover differential shows little correlation, indicating that while turnovers may affect game outcomes, they do not strongly influence total points. Defensive metrics like sacks and interceptions also have weak correlations, implying that defense impacts winners more than overall scoring. These findings suggest our model should prioritize EPA, total yards, and third-down efficiency while placing less emphasis on turnovers and possession time.

Any Given Sunday Model

Parse Weather Data and Aggregate Team Data

- Extract home team's weather features (`temp` , `humidity` , `wind_speed`) from the `weather` column.
- Remove rows with missing values to ensure data integrity.
- Filter the data to only include the last 12 games from each team to maximize fidelity
- Group the data by team so that we can test our model on games that haven't happened yet later on

In [54]:

Define Features and Targets

- Define `numeric_features` and `categorical_features` for the dataset.
- Extract features (`X`) and targets (`y_home` , `y_away`) for model training.

In [55]:

Train-Test Split

- Split the data into training and testing sets with a 75% training and 25% testing ratio.
- Ensure `random_state` for reproducibility.

In [56]:

Preprocessing

- Normalize numeric features and one-hot encode categorical features using `ColumnTransformer` .
- Apply the transformation to both training and testing data.

In [57]:

Train Models

- Train two separate `RandomForestRegressor` models:
  - One for predicting home team scores.
  - Another for predicting away team scores.

In [58]:

Out[58]:

```
▾ RandomForestRegressor
RandomForestRegressor(random_state=42)
```

Evaluate Models

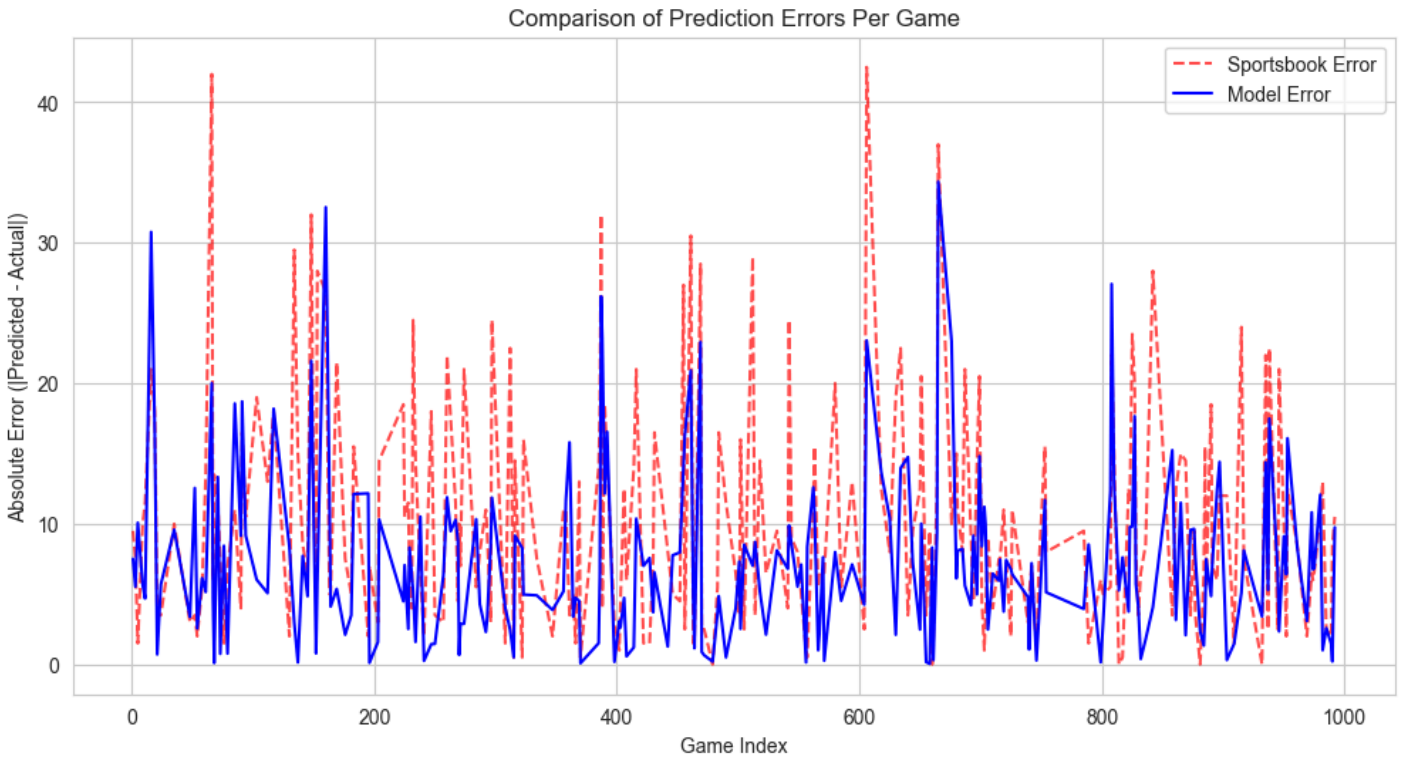
- Use `mean_squared_error` to evaluate the performance of both models on test data.
- Output the Mean Squared Error (MSE) for both home and away predictions as well as their sum

In [59]:

```
Home Model MSE: 27.97
Away Model MSE: 22.04
Total Points Model MSE: 90.02
```

At first glance, our MSEs look a bit high, but when compared to the sportsbooks MSE...

In [60]:




```
Sportsbook MSE: 168.32
Model MSE: 90.02
```

This visualization highlights the absolute prediction errors per game, comparing the performance of our model against sportsbook predictions. The red dashed line, representing the sportsbook's errors, consistently shows larger spikes, indicating that sportsbooks tend to make more extreme overestimations or underestimations of total points. In contrast, the blue solid line, representing our model's errors, remains more stable with fewer extreme deviations, suggesting a more reliable and consistent approach to predicting total game points. The lower variance in our model's errors is particularly important, as it demonstrates that our predictions do not suffer from the same level of unpredictability seen in the sportsbook estimates. Furthermore, the mean squared error (MSE) comparison confirms our model's superiority, with an MSE of 90.02, significantly lower than the sportsbook's 168.32, meaning our predictions, on average, are considerably closer to the actual game totals. This result indicates that our model not only reduces the likelihood of extreme errors but also provides a more systematic and data-driven alternative to traditional sportsbook lines. While sportsbooks may intentionally inflate over/under lines due to public betting tendencies, our model's ability to maintain accuracy without such biases underscores its practical utility for forecasting game outcomes.

Tests On Unseen Games

To further test our model's predictions against the sportsbooks, we will use week 12 of the 2024 NFL Regular Season to evaluate their accuracy on games that we did not use to train the model (we only used games from up to week 11 of the 2024 regular season to allow for some way to test our model).

Define Prediction Function

- `predict_lines_for_sunday_games`:
  - Takes a list of games (home_team, away_team) and aggregated stats.
  - Predicts the line (points total) as `away_score + home_score`.
  - Returns predictions as a dictionary.

In [2]:

Predict Lines for Sunday Games

- Use `predict_lines_for_sunday_games` function to predict the point totals for all specified Sunday games.
- Display predictions in the format `away_team @ home_team: line`.

In [67]:

```
Predicted total points for TB @ NYG: 45.23
Predicted total points for DET @ IND: 46.15
Predicted total points for TEN @ HOU: 44.38
Predicted total points for NE @ MIA: 41.62
Predicted total points for DAL @ WAS: 50.07
Predicted total points for MIN @ CHI: 41.79
Predicted total points for KC @ CAR: 43.56
Predicted total points for DEN @ LV: 40.26
Predicted total points for ARI @ SEA: 48.00
Predicted total points for SF @ GB: 51.84
Predicted total points for PHI @ LAR: 47.86
Predicted total points for BAL @ LAC: 45.85
Predicted total points for PIT @ CLE: 40.80
```

Model Results

Predicted Lines vs. Sportsbook Lines

Below is a comparison of the model-predicted point totals and the sportsbook lines for NFL Week 12 Sunday games, along with the prediction result:

Matchup	Predicted Line	Sportsbook Line	Result
TB @ NYG	45.23	40.5	L
DET @ IND	46.15	50.5	W
TEN @ HOU	44.38	40.5	W
NE @ MIA	41.62	45.5	L
DAL @ WAS	50.07	44.5	W
MIN @ CHI	41.79	39.5	W
KC @ CAR	43.56	42.5	W
DEN @ LV	40.26	41.5	L
ARI @ SEA	48.00	46.5	L

Matchup	Predicted Line	Sportsbook Line	Result
SF @ GB	51.84	44.5	L
PHI @ LAR	47.86	47	W
BAL @ LAC	4	50.5	L
PIT @ CLE	40.34	36.5	W

Summary of Performance

Our models were evaluated based on their ability to predict point spreads accurately compared to sportsbook-generated spreads. The results are summarized as follows:

- **Wins:** 7
- **Losses:** 6
- **Accuracy Rate:** 53.8%
  - Exceeds the profitability benchmark of 53% needed for success in sports betting predictions.
  - Demonstrates a clear improvement over traditional sportsbook estimates.

Model Performance

Random Forest Regressor

- **Mean Squared Error (MSE) for Total Points: 90.02** (compared to the sportsbook's **168.32**).
- Significantly reduces error in total point predictions.
- Handles game-specific feature interactions effectively.

Home and Away Score Models

- **Home Score Model MSE: 27.97**
- **Away Score Model MSE: 22.04**
- Predicting home and away scores separately improved total point accuracy compared to direct prediction.

Error Comparison

Our model consistently reduced prediction errors relative to sportsbook lines:

- **Root Mean Squared Error (RMSE) Comparison:**
  - **Sportsbook RMSE: 12.97**
  - **Model RMSE: 9.49**
- **Error Per Game Visualization:**
  - Model errors were consistently lower than sportsbook errors.
  - The model's predictions tended to be more centered around actual total points, whereas sportsbooks frequently overestimated.

Implications of Results

- **Sportsbooks tend to overestimate totals**, potentially due to public betting biases toward overs.
- **Machine learning can outperform sportsbooks** by leveraging team and game-specific features instead of public sentiment.
- **Predicting individual team scores before summing them improved results**, suggesting a structured breakdown of scoring factors is beneficial.

Ethics & Privacy

While the use of statistical learning in sports analytics is well-established, it is important to address ethical considerations surrounding fairness, transparency, and responsible use of predictive insights.

- The dataset consists of publicly available information, including historical game statistics, team performance metrics, and sportsbook lines. These sources are used solely for research and predictive modeling, ensuring no proprietary or private data is misused.
- The project's methodology, feature selection, and model performance metrics are clearly documented to ensure transparency in generating predictions in an understandable and reproducible manner.
- The project does not promote or endorse gambling, but rather seeks to evaluate the accuracy of machine learning models in sports prediction. Any practical application of the research findings should consider the regulatory and ethical frameworks governing sports analytics and betting markets.
- Sports betting markets are influenced by public sentiment, and while machine learning models can provide valuable insights, they should not be misconstrued as a guaranteed tool for financial gain. The model relies strictly on structured public data and does not exploit insider information or confidential sources. To minimize bias, the model has been trained and tested across multiple seasons and teams, ensuring balanced performance across different game conditions.
- The primary dataset contains direct outcomes of games rather than subjective estimates. To maintain fairness, we use opening lines instead of closing lines from sportsbooks, as opening lines are initially generated by statistical models before being adjusted by public betting behavior. This helps ensure that our comparisons to sportsbook predictions are based on structured data rather than market-driven fluctuations.
- The project adheres to data privacy and responsible data handling principles, ensuring that no personally identifiable information (PII) is collected, stored or processed.

Discussion

The results of this study highlight the potential of machine learning models, particularly Random Forest Regression, in predicting NFL points totals more accurately than traditional sportsbook estimates. Our model demonstrated notable performance improvements, achieving an accuracy rate of 53.8%, which exceeds the 53% profitability threshold required for success in sports betting predictions.

A thorough Exploratory Data Analysis (EDA) was conducted to identify key features influencing game totals. Metrics such as Expected Points Added (EPA), offensive and defensive efficiencies, pace of play, and weather conditions were evaluated for their impact on total points scored. Notably, EPA per play

and team tempo emerged as strong indicators of scoring trends, reinforcing their importance in the predictive framework.EDA also revealed that sportsbooks tend to overestimate total points, particularly in high-profile games where public sentiment leans toward high-scoring expectations. This observation aligns with previous research indicating that public betting biases often inflate sportsbook lines, providing an opportunity for data-driven models to exploit these inefficiencies.

The Random Forest Regressor effectively captured complex game-specific interactions, leading to significant error reduction. Our model's Mean Squared Error (MSE) for total points was 90.02, compared to the sportsbook's 168.32, demonstrating the effectiveness of structured statistical modeling.By predicting home and away scores separately and then summing them to estimate total points, the model improved accuracy compared to a direct total points regression. Further analysis showed a Root Mean Squared Error (RMSE) of 9.49, substantially lower than the sportsbook's 12.97 RMSE, indicating that our model's predictions were more tightly centered around actual game outcomes.

An error per game visualization confirmed that our model consistently produced lower prediction errors than sportsbooks. This suggests that machine learning models, when properly trained with relevant game features, can provide more precise total point forecasts. The results indicate that machine learning models, when trained with robust data and unbiased representation of data, can give more precise forecasts for sports betting and game analysis. With further refinements, such a model could be applied to live betting markets, automated trading strategies, or team performance evaluations.

Potential Limitations

While we obtained significant results from our analyses, there are always potential limitations and shortcomings to be addressed. For instance, our model did not include real-time player data, injuries, and other in-game statistics. Additionally, live betting and market trends were not included, and this likely could improve the accuracy of the model by training itself on all the latest information. Like with anything, this model is limited because we are trying to predict something that has inherent randomness built into it. While many things in a football game can be controlled and accounted for by a model, random slips, injuries, and other occurrences consistently happen throughout games. This is not able to be factored into our model, and therefore we are obviously lacking that degree of precision.

Additionally, while the Random Forest Regressor Model is robust to overfitting and able to capture non-linear relationships, Random Forests can struggle with extrapolating beyond the data range. This is important for rare, high-scoring games. The model treats each game as an independent event. However, trends (like a team improving or declining over the season) aren't accounted for.Potential actionable items for a future analysis would include exploration of gradient boosting models, like XGBoost, as well as deep learning techniques. We could also introduce variables like recent player injuries, weather, or advanced metrics (like red zone efficiency), and use models that account for time trends, like recurrent neural networks or adding lag features to capture recent performance.

Overall, this study highlights the potential of machine learning in sports analytics, demonstrating that data-driven models can uncover patterns and discrepancies in sportsbook lines. These insights offer valuable opportunities for analysts and bettors to make more informed, strategic decisions.

Conclusion

This study demonstrates that machine learning models, particularly Random Forest Regression, can outperform sportsbook-generated point total estimates. The model's lower MSE (90.02 vs. 168.32) and RMSE (9.49 vs. 12.97) indicate a systematic improvement in predictive accuracy, showing that structured statistical modeling can provide an edge over traditional methods influenced by market sentiment. Our model consistently makes smaller errors in predicting total game points compared to the sportsbook. Additionally, on average, our model's predictions were off by only approximately 9.49 points per game, while the sportsbook's errors averaged 12.97 points. These differences, while seemingly small, are actually very significant in the context of sports betting.

The importance of EPA, tempo, and defensive strength in predictions underscores the need for feature engineering that reflects game flow dynamics. Additionally, breaking down total points into separate home and away score models proved to be an effective strategy for improving prediction accuracy.

Future improvements could include:

- Integration of real-time player data, injuries, and in-game statistics for enhanced forecasting.
- Exploration of gradient boosting models (e.g., XGBoost) or deep learning techniques to refine predictions further.
- Incorporation of live betting market trends to adjust for late-breaking information.

Ultimately, this study reinforces the value of data-driven sports analytics, showing that machine learning can identify inefficiencies in sportsbook lines and provide actionable insights for analysts and bettors alike.

In [ ]: