Project Report

Predicting NFL Over/Under Outcomes

Alex Bender, Christian Kronfeld

Repository URL: https://github.com/CCody073/Data_Wrangling_Final_Project.git

Introduction

Users of sports betting apps generally face a statistical disadvantage against the house when making bets. We plan to help level the playing field by creating a linear regression model that predicts the likely combined score for NFL games based on key team statistics. By comparing our model's predicted total to the sportsbook maker's over/under line, we can help users make more informed decisions on whether to bet the over or under for a given game.

Data

Our data comes from two main sources that we progressively scrape each week. First, we use The-Odds-API[1] to collect active betting lines from various bookmakers worldwide and the scores from each game since week 8 of the 2024-2025 season. We include only the "over" odds to avoid redundancy, as each game would otherwise have duplicate entries for "under" bets. This dataset provides the total points line, game start time, home and away teams, bookmaker name, and the most recent line update timestamp. We keep only the most recently updated game line from each bookmaker to prevent duplicate entries of the same game at different update times. The timestamps for both the game start time and the update timestamp will need to be formatted to UTC time for consistency across time zones and our datasets.

Next, we web scrape team statistics from TeamRankings.com[2],[3] using Python's selenium package to automate Chrome browser control. This provides each NFL team's offensive and defensive rankings, along with their average points scored and points allowed per game in 2024.

Our script automatically updates game scores by matching unique identifiers (commencement_time, away_team, and home_team) between the CSV file and API data. This eliminates the need for manual updates and ensures accurate score assignment for each game. Additionally, the script automatically calculates the week each NFL game was played based off the commencement_time. We will continue to

[1] Sports Odds API | The Odds API
[2] NFL Football Stats - NFL Team Points per Game | TeamRankings.com
[3] NFL Football Stats - NFL Team Opponent Points per Game | TeamRankings.com

collect entries for each new bookmaker line every two days to enhance the prediction accuracy of our model until the deadline of this project which occurs on Week 13 of the 2024-2025 NFL regular season.

The **union** of offensive and defensive rankings will contain all rows from each dataset since they share consistent formatting from TeamRankings.com. The **intersection** of this unified team data with the bookmaker data requires a custom team name mapping table to standardize naming conventions across datasets. All merges use team names as the key identifier, with the final dataset keeping only the matched records between bookmaker data and team statistics.

With the merges complete, we can then scrape the actual total points for each game every Tuesday (as final games for the week will be complete after Monday). We will then merge this into our main dataset for games that have already occurred, once again using team names as the primary identifier.

Data Dictionary

Column	Type	Source	Description	
type	Text	Sports Odds API	Bookmaker's line label	
odds	Numeric	Sports Odds API	Odds of point value	
point	Numeric	Sports Odds API	Bookmaker's total points for the betting line	
actual_total	Numeric	Sports Odds API	Total points result after game is complete	
commence_time	Date Timestamp	Sports Odds API	Starting time for game	
home_team	Text	All	Specifies the home team for the game	
away_team	Text	All	Specifies the away team for the game	
bookmaker_key	Text	Sports Odds API	Name of sportsbook company	
bookmaker_last_update	Date Timestamp	Sports Odds API	Most recent betting line update from sportsbook maker	
home_offense_rank	Categorical	Team Rankings PPG	Home team offense rank	
home_points_for	Numeric	Team Rankings PPG	Average points scored by home team's offense in 2024	
home_defense_rank	Categorical	Team Rankings Opponent PPG	Home team defense rank	
home_points_against	Numeric	Team Rankings Opponent PPG	Average points allowed by home team's defense in 2024	
away_offense_rank	Categorical	Team Rankings PPG	Away team offense rank	

away_points_for	Numeric	Team Rankings PPG	Average points scored by away team’s offense in 2024	
away_defense_rank	Categorical	Team Rankings Opponent PPG	Away team defense rank	
away_points_against	Numeric	Team Rankings Opponent PPG	Average points allowed by the away team defense	
week	Categorical	Calculated from the commence_time value from **Sports Odds API**	Derived value that determines the week the game was played based of the commence_time value	
Predicted_total	Numeric	Derived Value	Derived value that predicts the total score for each game using our script’s linear regression	
Prediction_correct	Binary	Derived Value	Binary value that states whether the prediction was correct (1) or Incorrect (0)	

3. Analysis

3.1 NFL Rankings and Score Correlation

We wanted to find out whether teams' offensive and defensive rankings correlate with actual combined game scores compared to sportsbooks' over/under lines. We started by calculating correlation coefficients to explore the relationship between variables. The correlation coefficient between offensive rankings and actual scores was -0.200 ($p < 0.001$), while defensive rankings showed a correlation of 0.283 ($p < 0.001$) with actual scores. The negative correlation for offense indicates that better offensive rankings (lower numbers) tend to produce higher-scoring games, while the positive defensive correlation suggests that worse defensive rankings (higher numbers) correlate with higher game totals.

We first thought the correlation strength might vary across different ranking tiers. To prove this, we created a scatter plot of offensive rankings versus actual totals, as shown in Figure 1. This plot revealed considerable variance in scoring across all ranking levels, but with a clear downward trend as offensive ranks increased. Taking a closer look at prediction accuracy by tier showed that middle-tier teams (25-50th percentile) had the highest prediction accuracy at 71.02%, while bottom-tier teams were least predictable at 51.23%. This suggests that games involving middle-tier teams may offer more reliable betting opportunities.

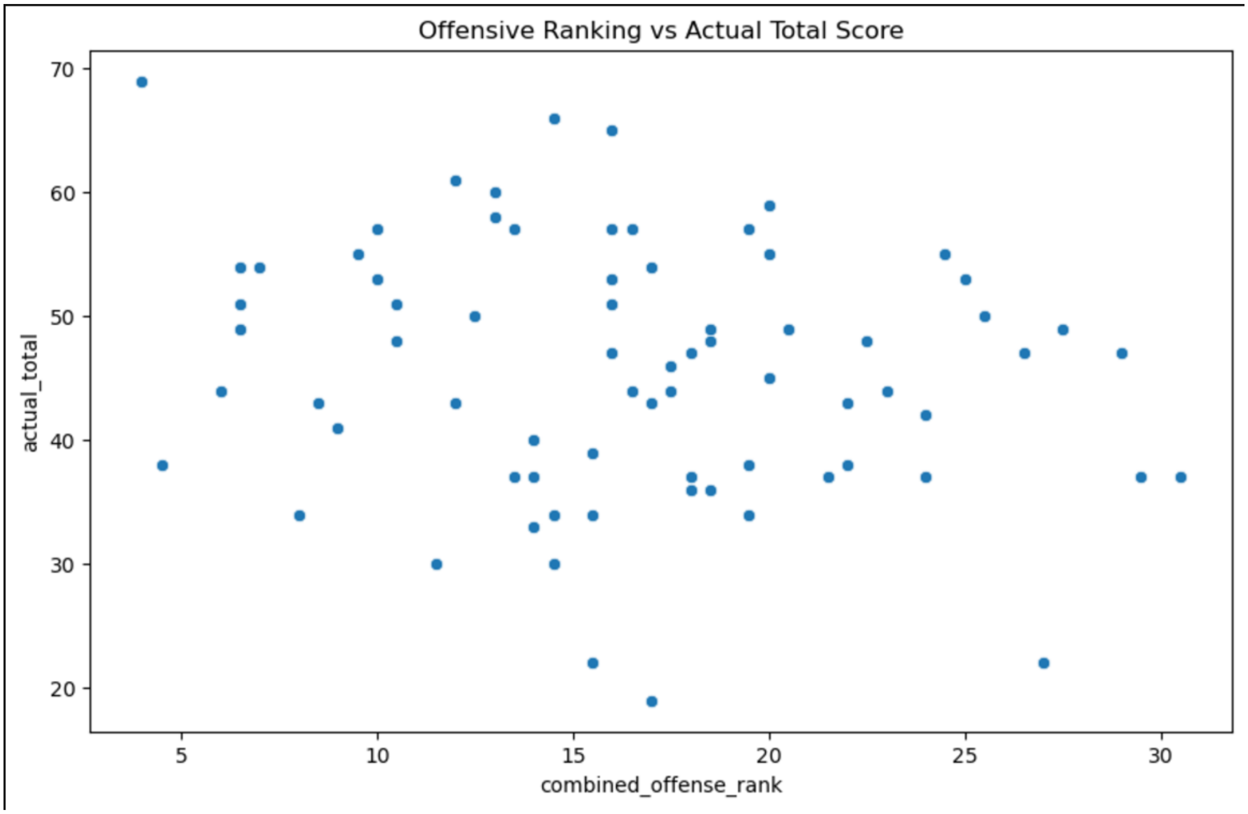

Figure 1: Offensive Ranking vs Actual Total Score

3.2 Home Field Advantage Impact

To determine how home field advantage influences over/under prediction accuracy, we analyzed the scoring patterns and prediction success rates for home teams. The data revealed a statistically significant home field advantage effect (t=2.747, p=0.006). On average, home teams exceeded the predicted total by 1.04 points, and predictions for home teams showed higher accuracy (65.69%) compared to the overall prediction accuracy (60.82%).

A deeper analysis by offensive ranking revealed varying impacts of home field advantage across different offensive tiers, as illustrated in Figure 2. The boxplot demonstrates that top offensive teams (ranks 1-5) showed the strongest home field advantage, averaging between 6-10 points above expectations. For example, teams ranked #1 and #2 offensively averaged over 9 points above the betting line when playing at home, with the #2 ranked teams showing perfect prediction accuracy. Middle-ranked teams (25-50%) showed more variance in their home field impact, while bottom-tier teams demonstrated more consistent but smaller home field advantages.

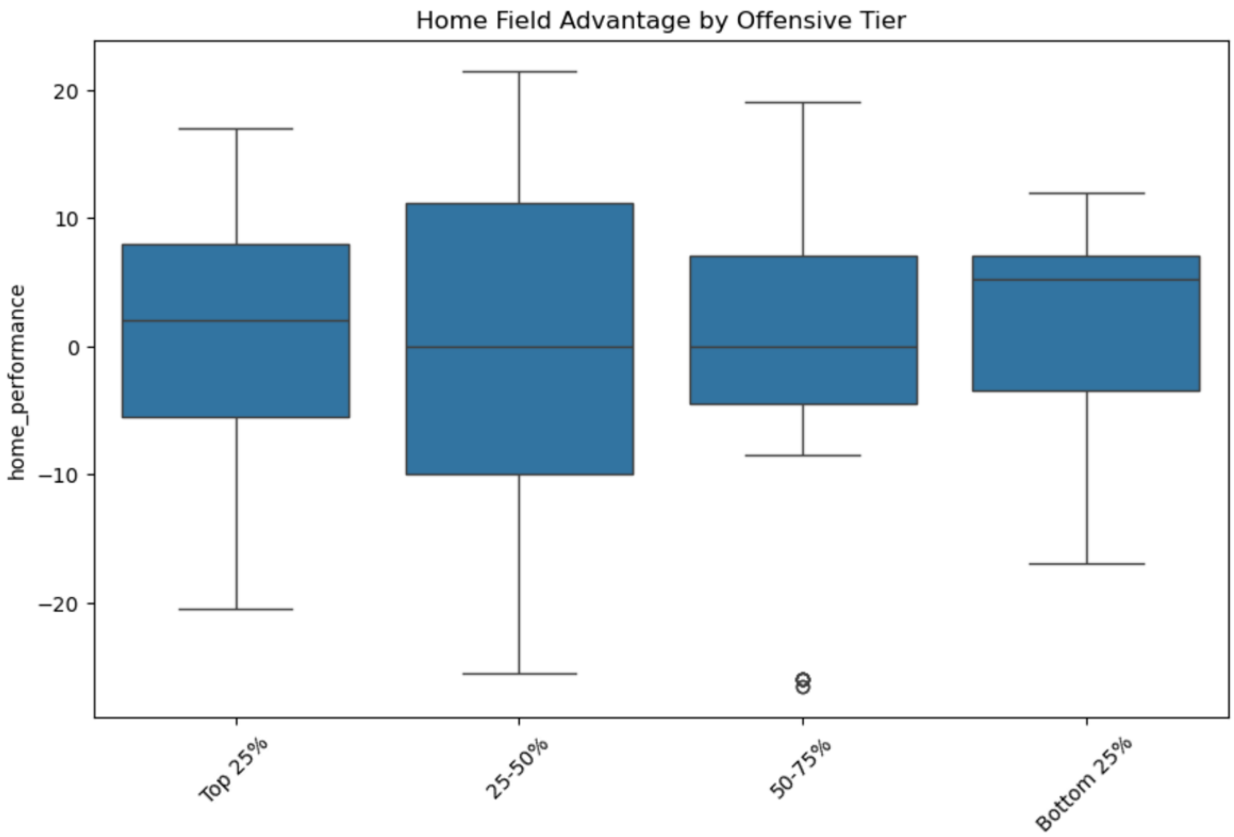

Figure 2: Home Field Advantage by Offensive Tier

3.3 Best Predictive Features

In examining which combination of team statistics serves as the strongest predictor for over/under outcomes, we built a linear regression model incorporating various team metrics. As shown in Figure 3, the model revealed that home team statistics wer

e significantly more predictive than away team statistics. The most important predictive features were home points for (importance score: 4.33), home offense rank (1.61), and home points against (0.99).

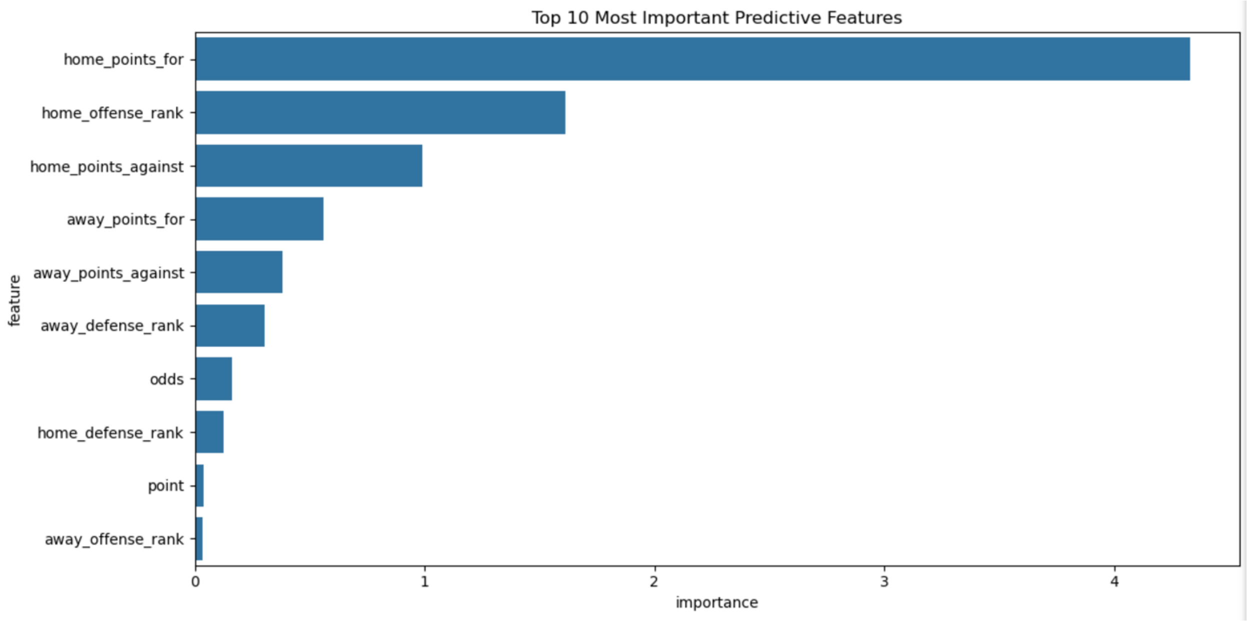

Figure 3: Top 10 Most Important Predictive Features

Our predictive model achieved an R² score of 0.301, indicating it explains about 30% of the variance in game totals. The Root Mean Square Error (RMSE) of 8.427 points suggests moderate predictive accuracy. While away team statistics did show some predictive power, particularly away points for (0.56) and away points against (0.38), their influence was notably less than home team metrics, as clearly demonstrated in Figure 3. These findings suggest that betting strategies might be more effectively based on home team performance metrics, particularly their offensive capabilities.

The relatively modest R² score indicates that while these statistical features provide valuable predictive insight, game totals remain influenced by additional factors not captured in our current model. This aligns with the complex nature of NFL games, where intangible factors and game-day circumstances can significantly impact scoring outcomes.

4. Conclusion

In this project, we analyzed three key aspects of NFL game scoring predictions: team rankings correlation with scores, home field advantage impact, and the most effective predictive features for over/under betting. From our analysis questions, we found the following results:

4.1 What is the relationship between teams' offensive/defensive rankings and actual combined game scores?

- Found moderate correlations: offensive rankings (-0.200) and defensive rankings (0.283) with actual scores
- Middle-tier teams (25-50th percentile) showed highest prediction accuracy at 71.02%

- Bottom-tier teams proved least predictable at 51.23%

4.2 How does home field advantage influence over/under prediction accuracy?

- Discovered statistically significant home field advantage effect (t=2.747, p=0.006)
- Home teams exceeded predicted totals by average of 1.04 points
- Home team predictions showed higher accuracy (65.69%) versus overall accuracy (60.82%)
- Top offensive teams (ranks 1-5) demonstrated strongest home field advantage, averaging 6-10 points above expectations

4.3 Which team statistics serve as the best predictors for over/under outcomes?

- Home team statistics proved significantly more predictive than away team metrics
- Key predictive features: home points for (4.33), home offense rank (1.61), and home points against (0.99)
- Model achieved R^2 score of 0.301 with RMSE of 8.427 points
- Home team offensive capabilities emerged as most reliable predictor

5. Limitations

Our study faced several key limitations:

5.1 Data Collection Constraints

- Historical game data from The-Odds-API limited to three days prior to script execution
- Manual updates required for some actual scores
- Dataset begins from week 8 of the 2024-2025 season, limiting sample size

5.2 Model Performance

- Modest R^2 score of 0.301 indicates significant unexplained variance
- Current model may not fully capture all relevant factors affecting game totals
- Limited ability to account for real-time factors like injuries or weather conditions

5.3 Time Constraints

- Analysis ends at Week 13 of the 2024-2025 NFL regular season
- Unable to analyze full season patterns or playoff game dynamics

6. Future Work

Several opportunities exist to extend and improve this research:

6.1 Data Collection Constraints

- Expand dataset to include multiple seasons for more robust analysis
- Incorporate additional variables such as weather conditions, injury reports, and rest days
- Develop automated solutions for real-time score updates

6.2 Model Refinements

- Explore more sophisticated modeling approaches beyond linear regression
- Develop separate models for different game contexts (divisional games, prime time, etc.)
- Investigate interaction effects between key predictive features

6.3 Application Development

- Create a user-friendly interface for real-time predictions
- Implement automated data collection and model updating
- Develop confidence ratings for predictions based on historical accuracy

6.4 Additional Analysis Areas

- Examine seasonal trends and their impact on prediction accuracy
- Analyze the impact of various bookmakers' line movements
- Study the relationship between prediction accuracy and betting volume