DSC 680
Benjamin Bartek
February 2, 2025

Analyzing NFL Player Performance Metrics and Game Outcomes

Business Problem

The main goal of this project is to identify the key NFL player and team metrics that influence game outcomes, efficiency, and overall success during the 2023 season. Using statistical and machine learning techniques, the analysis aims to address specific questions, such as determining which metrics are most strongly associated with team victories, assessing whether game outcomes can be accurately predicted using these metrics, and understanding how the resulting insights can translate into actionable strategies for NFL teams.

Background/History

NFL teams generate large amounts of performance data, which coaches and analysts have used for decades to enhance strategies and decision-making. Traditionally, metrics like rushing yards, passing efficiency, and turnover rates have served as benchmarks for performance evaluation. Data from the 2023 NFL season provides a valuable opportunity to apply modern analytical techniques to assess player performance, predict game outcomes, and guide team management strategies more effectively. This project aims to build upon traditional analytical approaches and provide actionable insights for NFL stakeholders.

Data Explanation

This project features two datasets from the publicly available 'NFL 2023 Season Dataset' on Kaggle (Mendoza Chavez, 2023). The first dataset, the game results dataset, contains details of each NFL game, including the week, game ID, participating teams, final scores, and total points. The second dataset, the team statistics dataset, provides detailed metrics such as points per game, yards per play, turnover margins, and rushing statistics. These datasets were prepared and merged to ensure consistency and usability for analysis. Missing values were addressed using median imputation for numeric columns, and categorical variables, such as team names, were encoded for compatibility with machine learning models. Features were also standardized to facilitate clustering and regression analysis. A detailed data dictionary was developed to clarify key variables, including fields such as week, game ID, yards per play average, turnover margin, and rushing play frequency.

Methods

The analysis followed a structured approach that combined Exploratory Data Analysis (EDA), statistical evaluation, and machine learning modeling. EDA included visualizations of data distributions, such as histograms of total points and scatter plots of key metrics, to uncover trends and relationships. Correlation analysis was conducted to identify statistically significant relationships among variables, using heatmaps to visually represent these findings. Statistical modeling was used to explore predictive relationships, with regression models applied to predict total points scored in games. Random Forest classifiers were employed to determine game winners, and feature importance analysis was conducted to highlight the most influential predictors of game outcomes. Additionally, K-Means clustering was utilized to group teams based

on performance metrics, providing insights into common patterns among similarly performing teams.

Analysis

The analysis revealed insights into team performance and game outcomes during the 2023 NFL season. Metrics such as yards per play and turnover margins emerged as some of the strongest predictors of success. Specifically, a higher turnover margin was consistently associated with increased probabilities of winning, highlighting the importance of minimizing turnovers and maximizing defensive takeaways. Additionally, teams with higher average yards per play were more likely to achieve higher scores, emphasizing the value of offensive efficiency.

Figure 1: Enhanced Correlation Heatmap of Team Statistics

DSC 680
Benjamin Bartek
February 2, 2025

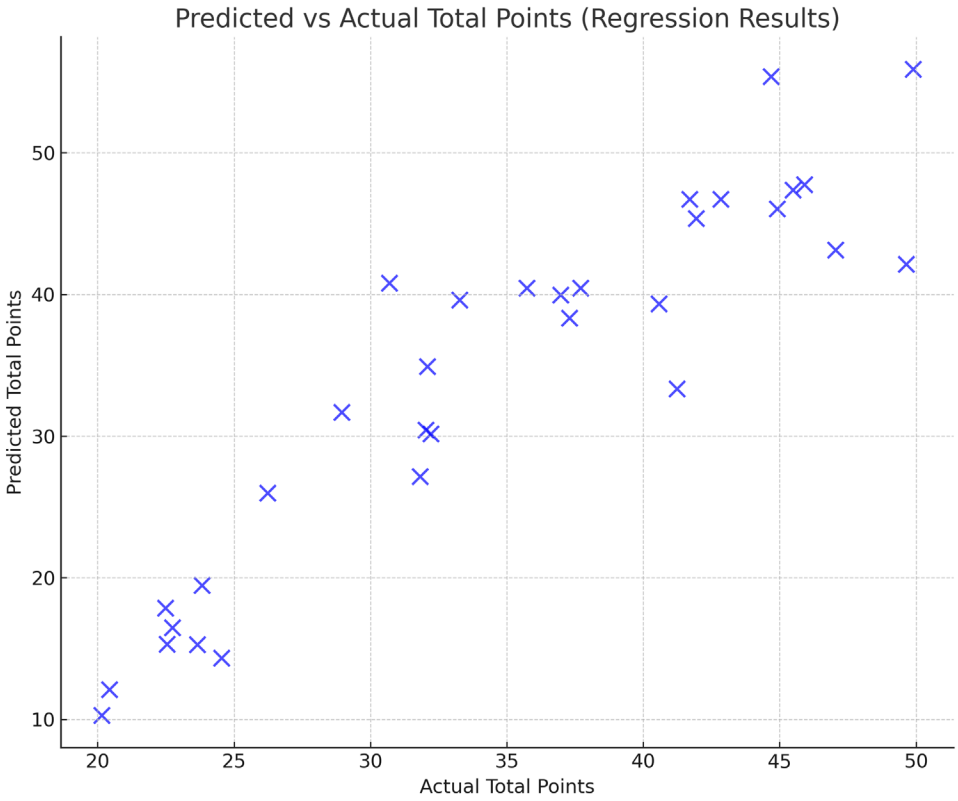

Figure 2: Predicted vs. Actual Total Points

DSC 680
Benjamin Bartek
February 2, 2025

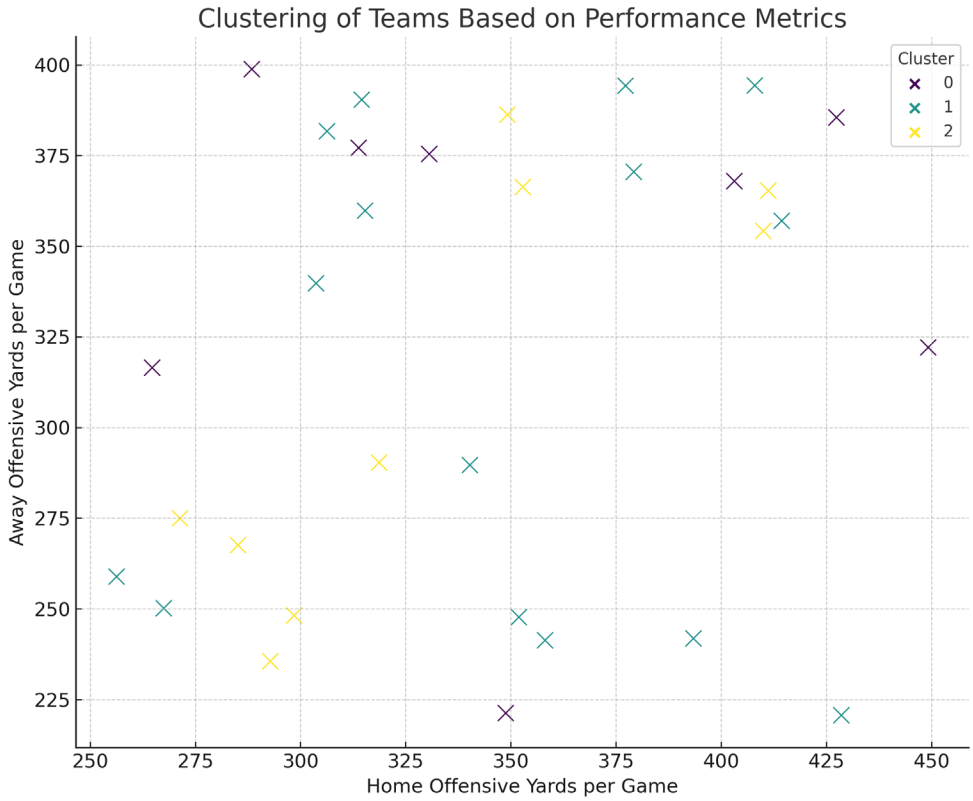

Figure 3: Clustering of Teams Based on Performance Metrics

Challenges

Data cleaning and preprocessing required significant effort to address missing and inconsistent values. Balancing model interpretability with predictive accuracy was another challenge, as complex models often provide limited transparency. Additionally, clustering analysis required careful tuning to ensure meaningful groupings of teams.

Future Uses/Additional Applications

This analysis has the potential for expansion and further application. Future studies could incorporate data from multiple seasons to improve generalizability and assess trends over time. Advanced metrics, such as player tracking data and in-game decision-making variables, could also be included to enhance the analysis. Additionally, real-time game prediction tools could be developed using similar methodologies, offering immediate value to teams and fans alike.

Recommendations

Based on the findings, it is recommended that teams prioritize metrics such as yards per play and turnover margins when developing strategies. Emphasizing these factors during training and

game preparation can improve overall performance. Additionally, teams should consider investing in data analytics infrastructure and expertise to gain a competitive edge in decision-making.

Implementation Plan

The implementation plan involves three key steps. First, data integration efforts should focus on developing a real-time dashboard to track critical metrics during games. Second, training programs should be designed to educate coaching staff and players on the importance of key performance metrics and their implications for game strategy. Finally, a feedback loop should be established to evaluate the effectiveness of analytics-driven decisions and refine strategies based on post-game evaluations.

Ethical Assessment

The data used in this analysis is publicly available and anonymized, minimizing potential privacy concerns. However, it is essential to avoid biases in analysis and ensure that findings are presented transparently and responsibly. This includes acknowledging any limitations or assumptions and avoiding overgeneralizations that may misrepresent player contributions or team dynamics.

References

1. Mendoza Chavez, R. (2023). *NFL* 2023 *Season Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/ruendymendozachavez/nfl-2023-season-dataset

DSC 680
Benjamin Bartek
February 2, 2025

Appendix A

Data Dictionary

This data dictionary provides descriptions of the key variables used in the analysis of NFL team performance and game outcomes during the 2023 season.

Game Results Dataset

Variable Name	Description
week	The week of the NFL season in which the game was played (1-18).
game_id	A unique identifier for each game.
home_team	The team that played at home.
away_team	The team that played away.
home_score	The total points scored by the home team.
away_score	The total points scored by the away team.
total	The combined total points scored by both teams.
winner	The team that won the game.

Team Statistics Dataset

Variable Name	Description
team	The name of the NFL team.
games_played	The total number of games played by the team.
home_ppg	Average points per game scored by the team when playing at home.
away_ppg	Average points per game scored by the team when playing away.
home_ypg	Average yards per game gained by the team when playing at home.
away_ypg	Average yards per game gained by the team when playing away.
home_yards_ppa	Yards per play average for home games.
away_yards_ppa	Yards per play average for away games.
turnover_margin	The difference between takeaways and giveaways for the team.
rushing_play_freq	The percentage of offensive plays that are rushing attempts.
home_oypg	Offensive yards per game for home games.
away_oypg	Offensive yards per game for away games.
home_rpf	Rushing play frequency for home games.
away_rpr	Rushing play frequency for away games.

DSC 680
Benjamin Bartek
February 2, 2025

Audience Questions

1. What are the most important predictors of game outcomes?
2. How can teams operationalize these findings?
3. How accurate are the predictive models?
4. What were the limitations of the analysis?
5. How does this analysis account for player injuries?
6. Can these methods be applied mid-season for adjustments?
7. What external factors were excluded?
8. How does this analysis compare with traditional coaching strategies?
9. How scalable are these techniques to other sports?
10. What ethical concerns might arise with real-time usage?