Predicting NFL Standings

Albion Shoshi, Alex Bechakas, and Gavin Crowley

1 Introduction

The NFL is renowned for its unpredictability, a hallmark that brings both excitement and intrigue to the sport. The thrill of an underdog victory or an unexpected twist in a season often stems from countless factors—injuries, preseason preparation, and evolving team dynamics—that make predicting outcomes a challenging yet rewarding endeavor.

The goal of this project is to predict the winning percentage of NFL teams by analyzing historical offensive and defensive statistics from the modern era (2003–2023). We harness machine learning techniques, specifically multiple linear regression and Random Forest models, to uncover patterns that influence team success.

Our motivation for this project lies not only in a shared passion for the NFL but also in the curiosity to explore how key metrics—such as point differentials, total yards, and defensive statistics—contribute to a team's performance over a season. While the randomness inherent in sports, influenced by injuries and preseason variability, adds an extra layer of complexity, it also highlights the versatility of machine learning.

By integrating two extensive data sets, one focused on offensive performance and the other on defensive metrics—we trained and evaluated various models to predict team outcomes for the 2023 NFL season. Model accuracy was assessed using RMSE, where the AIC-based regression model emerged as the top performer, with predictions falling within one or two wins of the actual results.

Machine learning proves to be an ideal tool for NFL predictions because of its ability to capture complex, non-linear relationships between variables. Its adaptability to dynamic and uncertain environments, like professional sports, allows us to generate robust insights while embracing the unpredictable nature that makes the NFL so captivating.

2 Background & Related Work

A relevant study by Gifford and Bayrak (2023) constructed predictive models to forecast NFL game outcomes using decision trees and binary logistic regression. Their analysis spanned 16 NFL seasons (2002–2018) and evaluated 4,096 games, demonstrating that offensive turnovers, defensive turnovers, and total yardage were the most significant predictors of a team's success. Their logistic regression model achieved a high accuracy of 83.1%, highlighting the importance of limiting turnovers on offense and forcing them on defense as key contributors to wins.

Our project builds on their work by incorporating a broader range of metrics, including both offensive and defensive statistics, to predict team success. While Gifford and Bayrak focused on binary win/loss outcomes, our analysis expands the methodology by evaluating multiple models, such as AIC-based regression, BIC-based regression, and Random Forest, to assess their predictive accuracy. Furthermore, our study focuses on predictions for the 2023 NFL season, integrating more recent data to capture evolving team performance trends and league dynamics.

Additionally, platforms like Pro Football Reference provide standardized metrics, such as Adjusted Yards per Attempt and Points Allowed, which serve as benchmarks for evaluating team performance. These

metrics form the foundation of our data set, enabling us to identify patterns and relationships between key performance indicators and overall team success.

Machine learning techniques, such as Random Forests and ensemble methods, have proven effective in capturing complex, non-linear relationships in sports analytics. Our project builds on these approaches, utilizing these methods to produce robust predictions while addressing the inherent unpredictability of the NFL.

3 Exploratory Data Analysis

To create the final data set for our analysis, we combined two comprehensive sources of NFL team statistics spanning the 2003 to 2023 seasons. The first data set, sourced from Kaggle, includes offensive metrics such as passing yards, first downs, wins, and point differentials, while the second data set, obtained from Pro Football Reference, focuses on defensive statistics like points allowed, yards allowed, and interceptions. To ensure consistency, the data sets were merged on common keys, such as team name and year, creating a unified structure. During the data cleaning process, variable names were standardized to clearly differentiate between offensive and defensive metrics (e.g., Yards was renamed to Offensive_Yards or Defensive_Yards_Allowed). Additionally, duplicate features, such as Margin of Victory, were removed due to excessive missing values, and any inconsistencies were resolved. This cleaned and structured data set now forms a reliable foundation for our machine learning models, enabling a comprehensive analysis of key performance indicators to predict NFL team success.

Table 1: Summary of NFL Team Performance (2003)

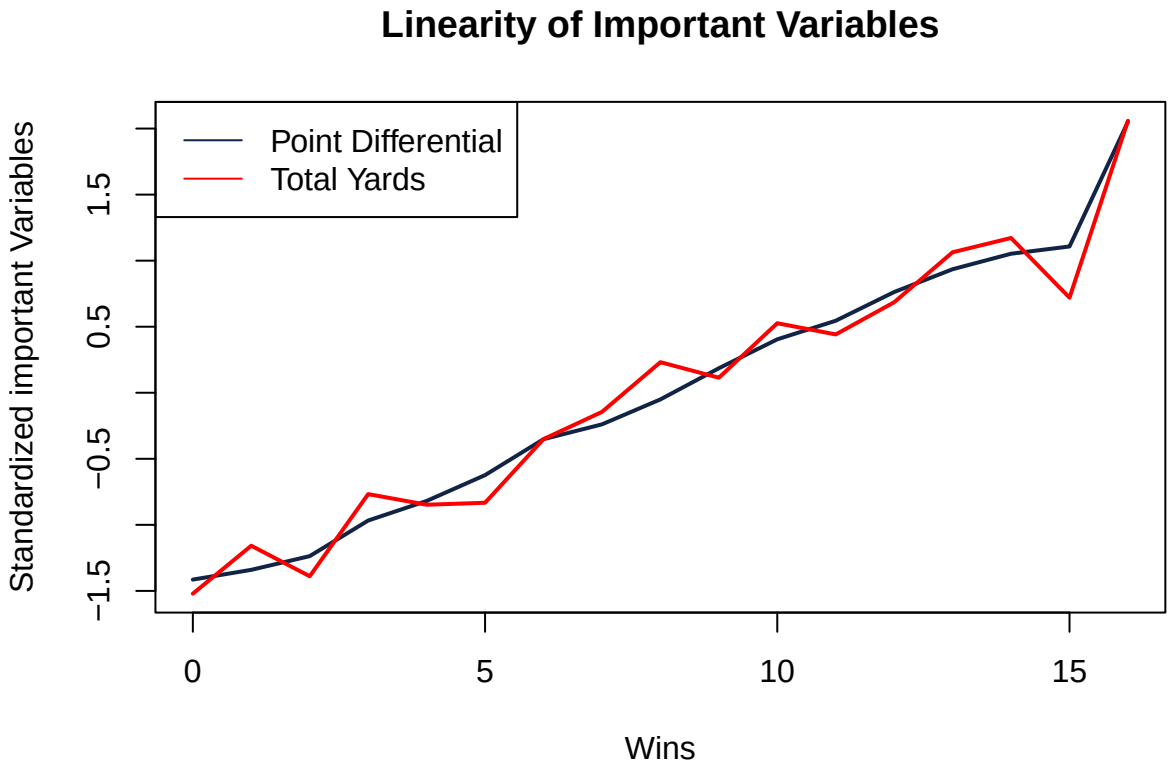
Of_year	Of_team	Of_wins	Of_losses	Of_win_loss_perc	Of_points	Of_points_diff
2003	New England Patriots	14	2	0.875	348	110
2003	Miami Dolphins	10	6	0.625	311	50
2003	Buffalo Bills	6	10	0.375	243	-36
2003	New York Jets	6	10	0.375	283	-16
2003	Baltimore Ravens	10	6	0.625	391	110
2003	Cincinnati Bengals	8	8	0.500	346	-38
2003	Pittsburgh Steelers	6	10	0.375	300	-27
2003	Cleveland Browns	5	11	0.313	254	-68
2003	Indianapolis Colts	12	4	0.750	447	111
2003	Tennessee Titans	12	4	0.750	435	111
2003	Jacksonville Jaguars	5	11	0.313	276	-55
2003	Houston Texans	5	11	0.313	255	-125
2003	Kansas City Chiefs	13	3	0.813	484	152
2003	Denver Broncos	10	6	0.625	381	80
2003	Oakland Raiders	4	12	0.250	270	-109
2003	San Diego Chargers	4	12	0.250	313	-128

4 Modeling

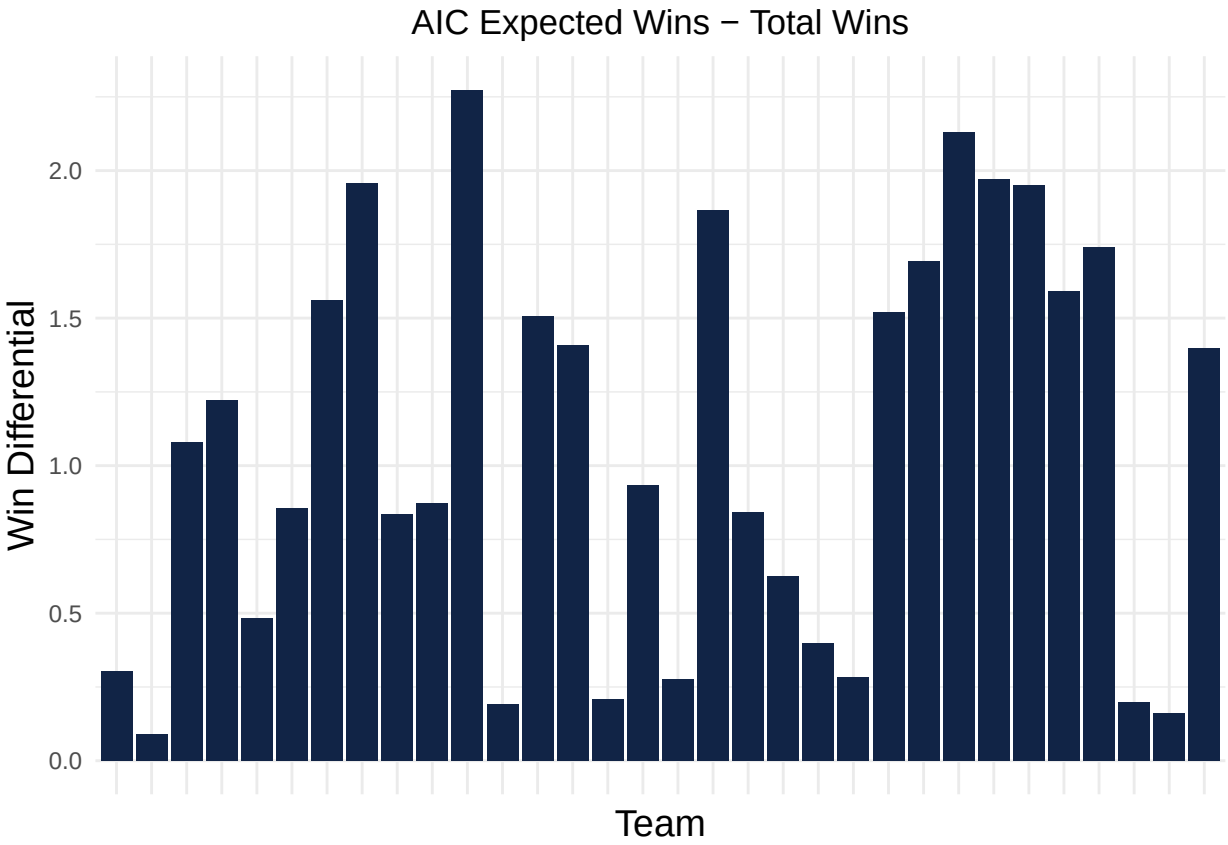
For predicting the final NFL standings we used regression modeling. We used bi-directional selection using AIC and BIC as criteria for the models along with Random Forest for regression. We compared the three models using AIC and and BIC and went with the model with the highest accuracy based on the RMSE.

4.1 Relevance of Regression

We began by showing the relevance of linear regression in this comparison by demonstrating the correlation between variables and wins. We took two of the more important variables in Point Differential, and Total Offensive Yards and compared them to the wins. After standardizing the variables, and plotting them on a line chart it is very clear that they share a linear relationship with the wins stat, which is what we are looking for. To further this argument, the correlations of Point Differential was 0.989 and for Total Offensive Yards it was 0.969. Both of these variable demonstrated strong linear correlation meaning linear regression has relevance in this case.

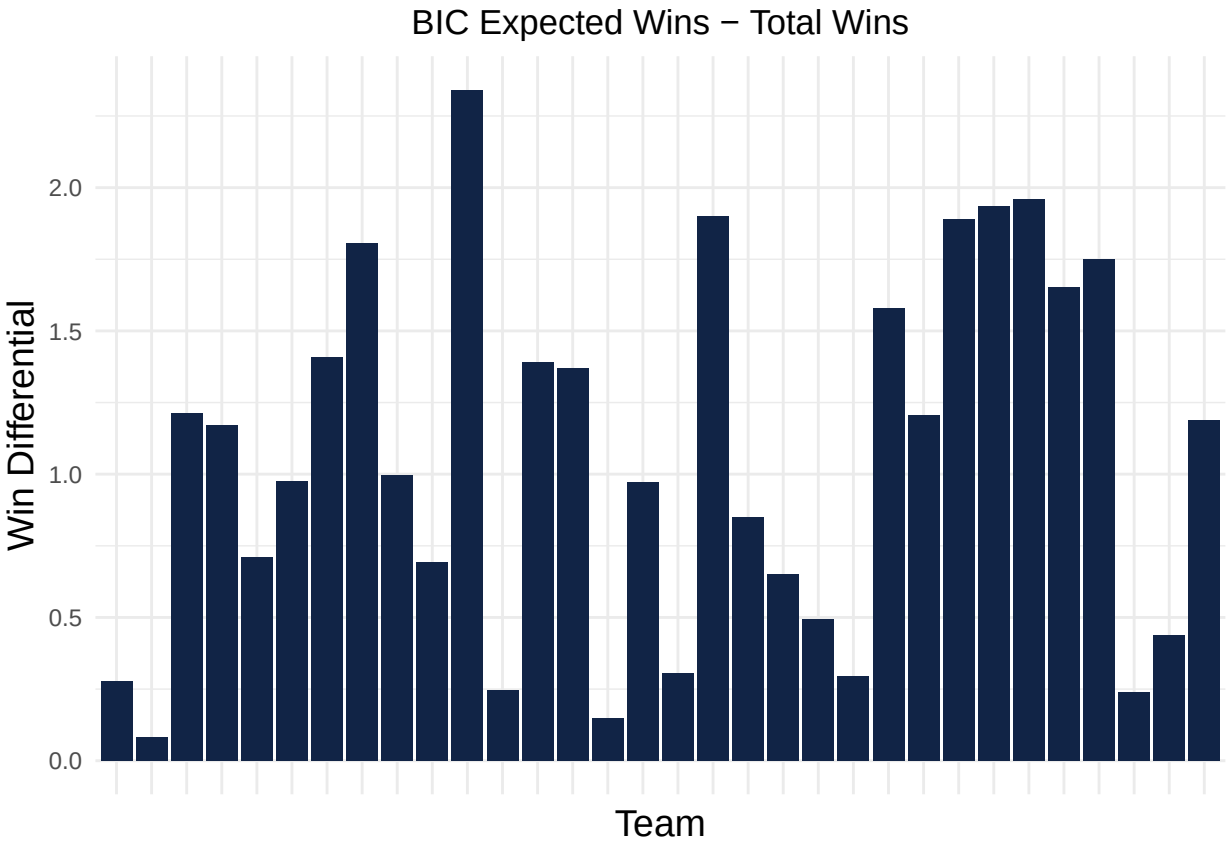


4.2 AIC Linear Regression



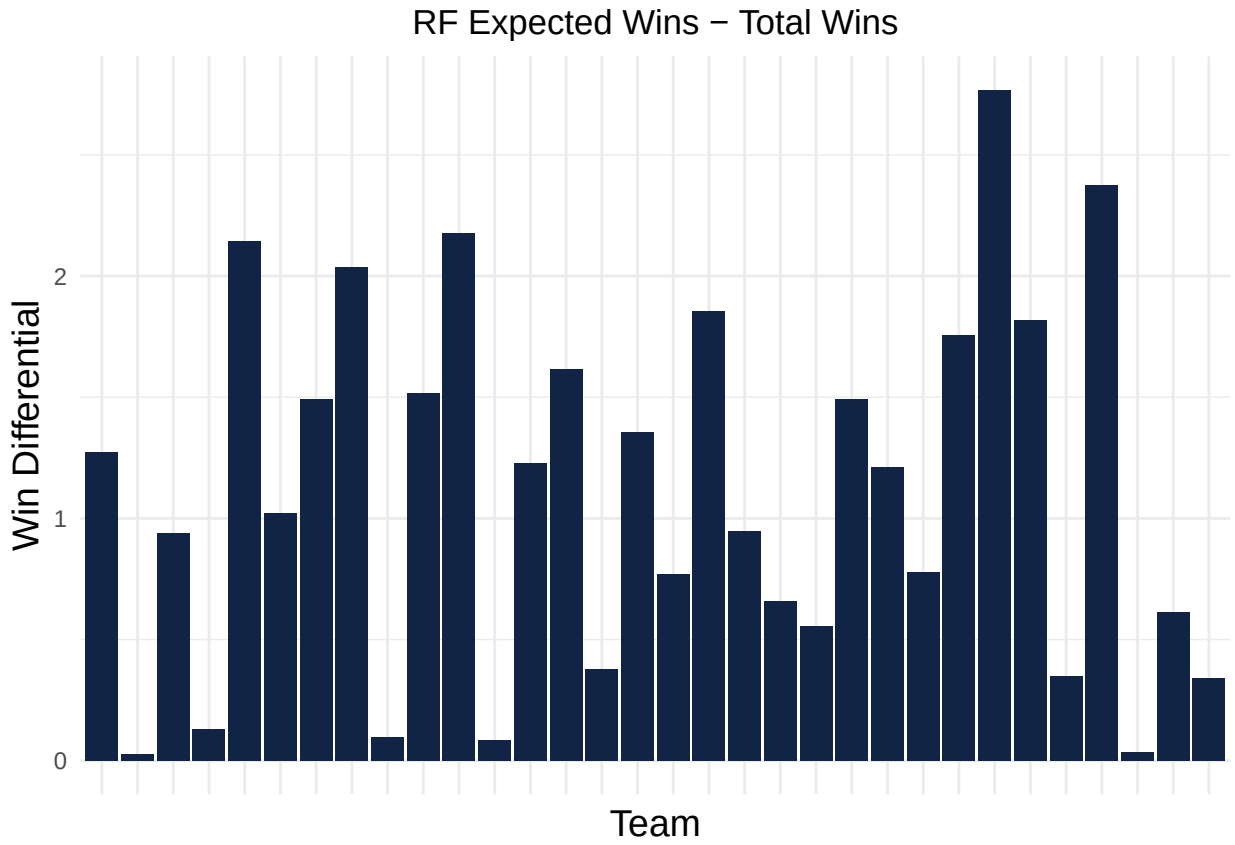
	Of_team	Of_wins	predictedAmodelWin	predictedAmodelWinDiff
641	Buffalo Bills	11	12.2204	1.2204
642	Miami Dolphins	11	10.3742	0.6258
643	New York Jets	7	4.8703	2.1297
644	New England Patriots	4	4.2827	0.2827
645	Baltimore Ravens	13	14.0808	1.0808
646	Cleveland Browns	11	9.0413	1.9587
647	Pittsburgh Steelers	10	8.0485	1.9515
648	Cincinnati Bengals	9	7.4387	1.5613
649	Houston Texans	10	8.4938	1.5062
650	Jacksonville Jaguars	9	8.7926	0.2074

4.3 BIC Linear Regression



	Of_team	Of_wins	predictedBmodelWin	predictedBmodelWinDiff
641	Buffalo Bills	11	12.1713	1.1713
642	Miami Dolphins	11	10.3475	0.6525
643	New York Jets	7	5.1095	1.8905
644	New England Patriots	4	4.2947	0.2947
645	Baltimore Ravens	13	14.2115	1.2115
646	Cleveland Browns	11	9.1935	1.8065
647	Pittsburgh Steelers	10	8.0388	1.9612
648	Cincinnati Bengals	9	7.5913	1.4087
649	Houston Texans	10	8.6095	1.3905
650	Jacksonville Jaguars	9	8.8522	0.1478

4.4 Random Forest Regression



	Of_team	Of_wins	predictedRFWin	predictedRFWinDiff
641	Buffalo Bills	11	11.1299	0.1299
642	Miami Dolphins	11	10.3422	0.6578
643	New York Jets	7	5.2444	1.7556
644	New England Patriots	4	5.4919	1.4919
645	Baltimore Ravens	13	12.0618	0.9382
646	Cleveland Browns	11	8.9631	2.0369
647	Pittsburgh Steelers	10	8.1811	1.8189
648	Cincinnati Bengals	9	7.5065	1.4935
649	Houston Texans	10	8.7712	1.2288
650	Jacksonville Jaguars	9	8.6220	0.3780

4.5 Regression Results

The following table compares the RMSE's of the different models to demonstrate the accuracy of the models. As seen below, both AIC and BIC are very close, but AIC is slightly more accurate. Although AIC is the most accurate, all three of the models are very accurate, and viable in useage

Regression Type	RMSE
AIC	1.232074
BIC	1.236749
Random Forest	1.345836

5 Conclusion/Discussion

In conclusion, our study successfully utilized machine learning techniques to predict NFL team success for the 2023 season by analyzing both offensive and defensive statistics. Using data from Kaggle and Pro Football Reference, spanning from 2003 to 2023, we created a comprehensive and reliable data set through rigorous cleaning and preprocessing. The AIC-based regression model emerged as the most effective, achieving the lowest Root Mean Squared Error (RMSE) among the models tested. The model demonstrated strong predictive performance, with 50% of predictions falling within a 1-game difference and 94% within a 2-game difference from actual results. These findings build on prior work, such as Gifford and Bayrak (2023), by emphasizing the predictive power of key metrics like point differentials, turnovers, and yardage, while expanding the analysis to include modern machine learning approaches and recent data. Overall, our study highlights the effectiveness of machine learning in addressing the inherent unpredictability of the NFL. The AIC-based model's results confirm that historical performance metrics remain reliable indicators of success despite evolving league dynamics. Future research could further enhance accuracy by incorporating additional variables such as injuries, coaching decisions, and player-specific performance, capturing the complexities of professional sports even more comprehensively.

References

Gifford, M., & Bayrak, T. (2023). Predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression. Decision Analytics Journal, 8, 100296.

NFL team statistics and historical data. Pro Football Reference.

Kaggle Link: https://www.kaggle.com/datasets/nickcantalupa/nfl-team-data-2003-2023