Joseph Russoniello, Zayad Almazrouei, Sebastian Mendoza, Tongai Zha

Using NFL Player Statistics to Predict Game Winners

Boston University College of Computing and Data Science

DS110: Introduction to Data Science with Python

Professor Kevin Gold

May 1, 2024

Introduction

Winning a given NFL game has become a significant achievement in American sports culture. Since the inception of the NFL, individual games throughout the season have drawn huge attention from fans across the nation. Each game is not just a display of athletic prowess but also a crucial part of each team's journey towards playoff contention and, ultimately, the Super Bowl. The importance of each game is underscored by its potential impact on a team's season, making every matchup a high-stakes event. Like the Super Bowl, regular season and playoff games also feature exciting moments, tactical gameplay, and can influence betting decisions. The ability to accurately predict the outcomes of these games is highly valued in the sports betting industry, as it can significantly affect profitability. As the popularity of NFL games continues to grow, so does the importance of predictive accuracy for fans, analysts, and bettors alike.

This paper will aim to use historical player data to predict which team will win in a given game. The data we sourced was from the official NFL website that provides player statistics from all the way back in 1970 to the more recent 2023. This paper aims to transform competitive team statistics, such as offensive Yards per Attempt, to predict which team will win a game, and whether or not that team could win the super bowl. It is our goal to determine how well a team's statistics correspond with their performance, and whether winning and losing teams have statistically different performance metrics.

We will begin our paper with evaluating different sources pertaining to our central question of "Can we predict which NFL team will win a given game?" Then we will explain the methods we used in our code, our analysis of various statistical methods, our application of machine learning methods to achieve accurate predictions alongside visualizations of our data. Lastly, we will synthesize our results from the data to reach a conclusion.

Previous works

Analyzing the Odds: How to Predict the Outcomes of NFL Games

Better insider went with a statistical approach to be able to predict a game's outcome (Stone, 2023). They examined the different factors that work together to result in a win. These include: Team quality, Game conditions, Analyzing strengths and weaknesses, Points per game, Yards per game, Mental and physical conditions. Betters insider believe that taking into account the different statistics, you are able to make a more accurate prediction which ultimately results in winning a bet.

Here's the 10-step program to creating a championship culture

The NFL official website published an article that talks about a comprehensive top-down approach that leads to creating a team of champions (NFL, 2009). It starts with a committed owner willing to support the vision of the coach and front office. It involves hiring an elite quarterback and coaching staff that can develop players and foster a locker room of selfless team-first mentality. The article mentioned different off the field components that foster a winning team.

NFL Fact or Fiction: What Makes a Winning Team?

Bleacher report identifies crucial elements that separate winning teams from losing teams (Sussman, 2011). They drew the conclusion that teams that can marry physical talent with intelligent decision-making, technical mastery and an overall commitment to fundamental football seem to have the best chance of emerging victorious.

NFL and Data: How Analytics is Used in the Game Prediction

Logan Data studies how data driven analysis techniques are used to predict games (Stone, 2023). By identifying patterns in historical data, analysts are able to come to conclusions in order to result in more accurate predictions of games.

NFL Betting Market: Using Adjusted Statistics to Test Market Efficiency and Build a Betting Model

James Donelly used advanced statistical models to develop a profitable betting strategy in the NFL point bracket model (Donnelly, 2013). The author uses a variety of methods and datasets to eventually test the model on the 2010-2012 seasons. The model is not able to be profitable in those three seasons but this analysis demonstrates just how unpredictable the NFL is from year to year.

Data Collection and Manipulation

This paper's methodology gathered professional team information from publicly available team statistics through the National Football League website and similar websites to train Random Forest Classification models on previous game history.

The NFL website mentioned above was parsed using the Beautiful Soup object from the bs4 python library to find all HTML tables on the website, and then a combination of io.StringIO and pandas were used to transform the HTML table into a pandas dataFrame. Each url was accessed using the f-string format:

f"https://www.nfl.com/stats/team-stats/offense/{option}/{season}/reg/all

In this link {option} represented the website's 3 different modes (offense, defense, and special-teams), and {season} represented the year of data that was being pulled. Each "pull" of year's statistics would result in a list of dataframes that could be merged using a tailor-made .to_one_df() command resulting in a 150 column dataframe, with each column representing a different team statistic,further referred to as features.

A similar process was performed to access a list of all previous football, their winners, and the games' score. The websites used to access of history of played games, followed a format of "Winning Team, Losing Team, Score", so to combat this for item in the finalized dataframe, which team was labeled as"Team 1" and which was labeled as "Team 2" were randomly shuffled to ensure that the winning team's location was not always in the first index, avoiding future bias in the Machine Learning process.

Once the team and matchup dataframe were established, we began the arduous process of transforming the data to a format that a Machine Learning model could read for an individual game. The Team dataframe was transformed into a dictionary with indices of the team's name and the season year (Ex: “2023 Cardinals”). Using this dictionary, we iterated over every game played since 1970, and the 2 opposing teams were searched in the team dictionary, as we stored the matchup’s data in a 2D list for faster iteration than appending row-by-row on a pandas DataFrame. During this iteration, the winners of each game labeled as “Team 1” or “Team 2” were collected as well as the game’s difference in score (a game with a final score of 24 to 10, would output 14). At the end of iteration the 2D list was transformed into a pandas dataFrame with 300 features corresponding to each of the two team’s stats. For each row in this DataFrame, we also had collected a binary classification of which team won, and numerical variable of the score difference for each game. These would be used as the features and labels for our first two machine learning models.

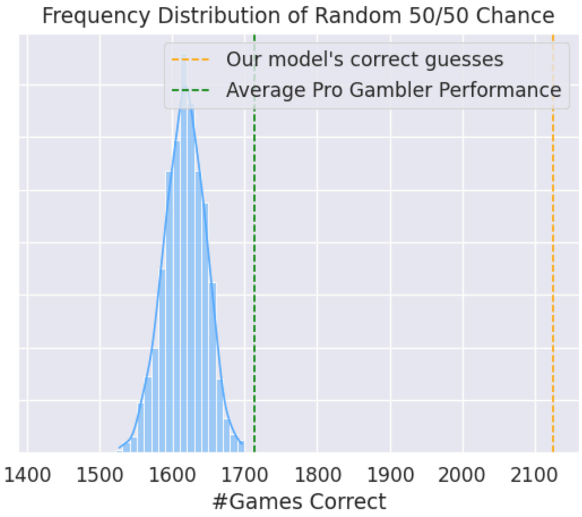
To gather the data for our third machine learning model, we iterated through the Individual Team dataset, and using a website of all superbowl winners we created one more boolean target list of whether each team won the superbowl in that year.

For the case of sports-betting, using live-updated websites of professional team data has a distinct advantage over using kaggle datasets. Each web parsing model is set up to take an input .csv file of historical data of previously pulled data, then search the internet to see if the csv data is up to date. If it is not, this method supports the automatic addition of new years’ data into the file, future-proofing our experiment and ensuring that the model will grow with time, without requiring future modifications for data sourcing/manipulation.

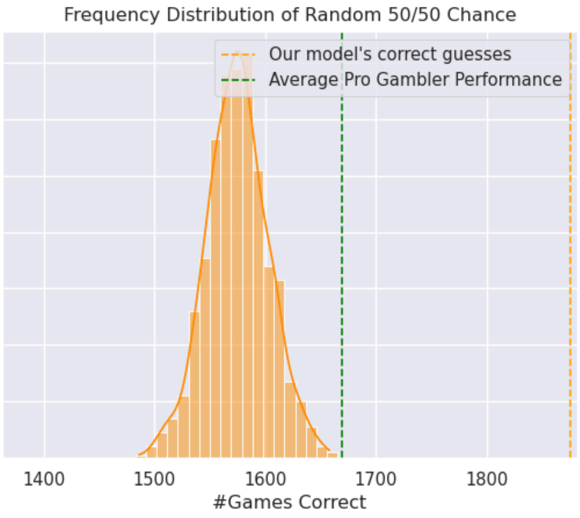
Machine Learning Models

Model 1: Random Forest Classification to Determine Game Winners

With the two-team features dataframe, and the binary list of game winners, we first separated the data into a training and test set (80-20% ratio) and then plugged the training set into a scikit-learn RandomForestClassifier model. This model showed arbitrarily high accuracy in the training set, indicating some tendencies of overfitting. To combat this, we used a cross-valuation method using both the training and test set with different parameters such as limiting the max number of features, limiting the depth of each individual Decision Tree in the Random Forest, and modifying the number of trees in the forest, to optimize the model and prevent overfitting when possible. The most effective model on this dataset was found with 100 estimators, a maximum tree depth of 10, and "sqrt" max features. On our test set of approximately 3,000 games, the model correctly guessed the winning team 65.76 percent of the time, far above our initial goal of 54%. This success rate is plotted against the distribution of success with a random chance model and our goal performance above.



This model, though incredibly effective, is not useful in the context of sports betting, as it requires end-of-season data to make its predictions. To allow the model to predict outside of its internal data scope, our group retrained the model with the two teams previous year data as the features.



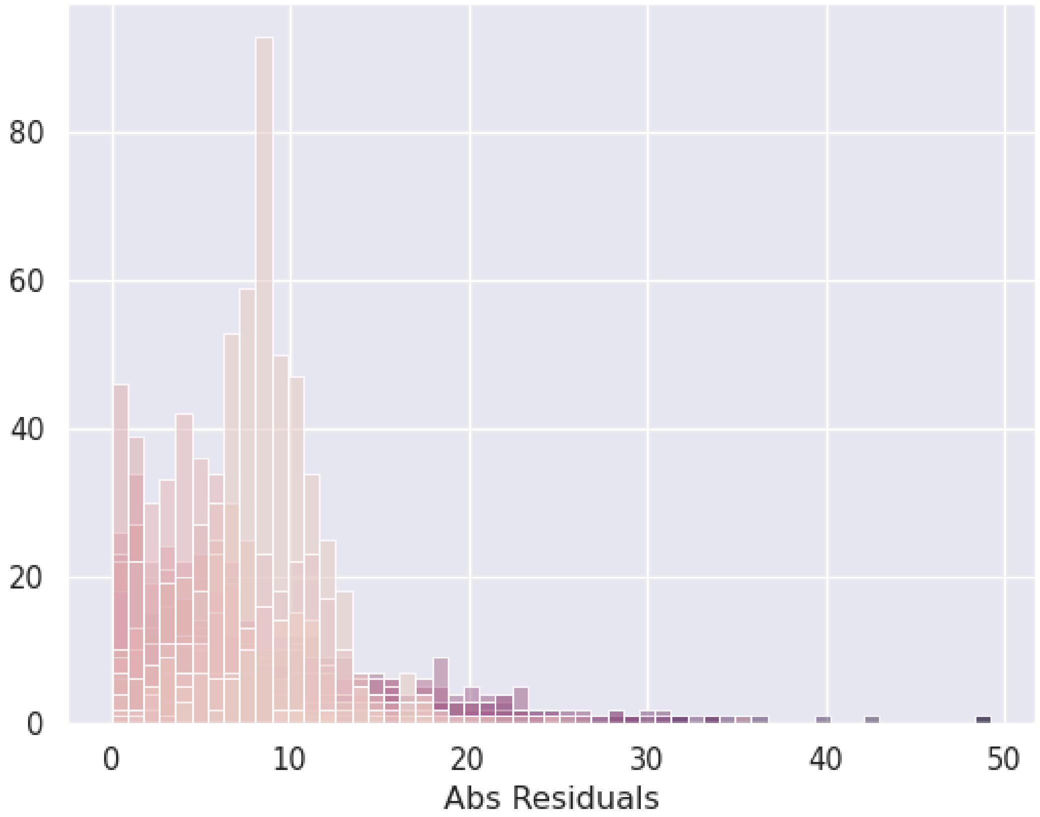
We trained a Random Forest Classifier with the same cross-valuation method as previously described, resulting in optimal settings of 200 estimators, a maximum depth of 20, and "sqrt" max features. On the same test set, this model showed an effectiveness of approximately 59.38%, markedly worse than the current year model, but now with the ability to predict outside of its scope 1 year in advance. This effectiveness is also plotted against the distribution of random chance above.

Model 2: MultiRegression on Score Difference

Using the two-team features dataframe, and the numerical list of score differences in the games, our team also attempted to predict the difference in score for each game. For this purpose, Multiple Regression was initially used; however, even after cross-valuation and optimization of settings, we could not push the regression's Coefficient of Determination (a key measure in a evaluating the strength of a Regression Model) past 0.05, showing no significant correlation between the multiple variables and the predicted variable.

We also attempted to use a RandomForestRegressor to do the same, but the ML model's tendency to overfit the training data hurt its effectiveness as well. This model showed an effectiveness of 0.8079 when used on the training set, but still was scored at -0.05 when used on the test set. This matter could be revisited with more robust statistical/Machine Learning tools in future studies, but our analysis did not manage to find a successful predictor of Score Difference given team stats.

In the following figure, the residuals from the linear regression prediction model are depicted. Games with a higher actual score difference are shown with darker red hues, and shows with lower score difference are shown in peach. As depicted below, the model could be described as too timid, with its difference in accuracy from the target value increasing as the difference in score between the two teams grows. Future investigations of this dataset may want to take this timid nature of the established multi-regression into account when training future models.

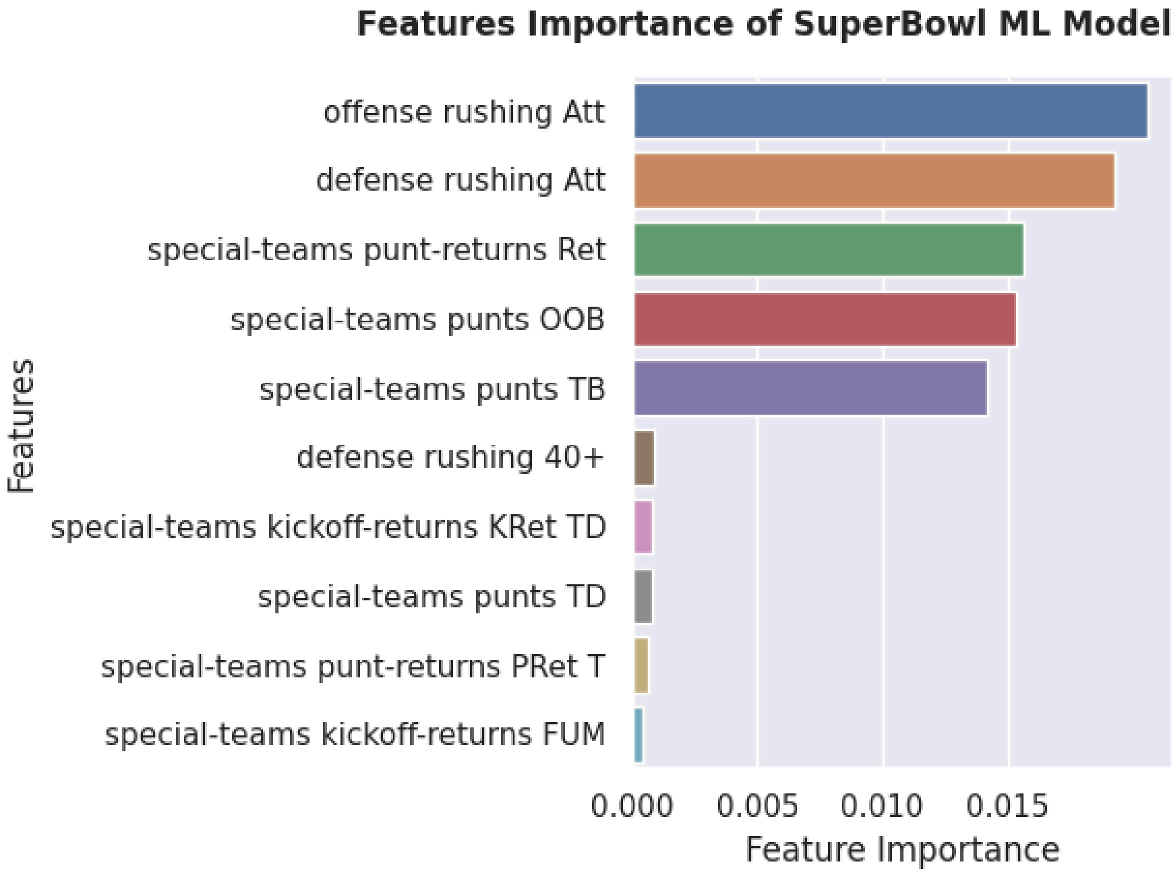


Machine Learning Model 3: Super Bowl Winners

Using the single-team statistic dataset for features, and the boolean list of whether each team won the superbowl as labels, we also trained a RandomForestClassifier to predict whether a team would win the superbowl. We pursued this optional ML model to tell us more about the underlying datasets for further exploration in statistical methods and visualizations. This Random Forest model, with default parameters, was able to predict whether a team would win the superbowl with a staggering 96% accuracy.

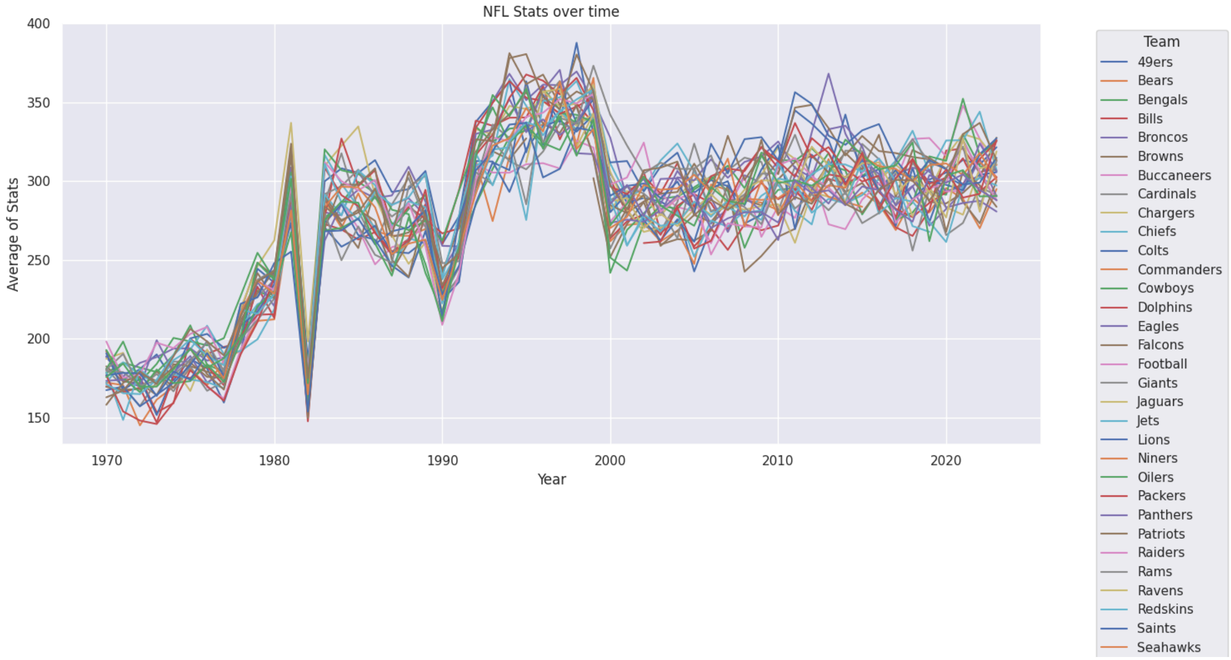
This measure is far less impressive when given adequate context. A model which we'll call a "hater model" that would claim every single team will lose the superbowl would pass this same test with around a 96% accuracy, as only one team each season is able to win the Superbowl, skewing the odds in its favor. Notably, our model correctly identified the superbowl-winning team 68.05% of the time, indicating that it had far more strategy than the aforementioned hater model, and was moderately successful at identifying extremely high performing teams when given a list of all team stats.

Examining the models feature priority reveals interesting insights about what makes a team a champion. Pictured below are the 5 most and least important features for superbowl prediction. The two most important attributes were, unsurprisingly, the key measures of offensive and defensive performance; however, the third, fourth, and fifth most important features were from the special-teams, with third being average yards converted per punt return, and percentage of punts OOB. At face value, these stats seem far more niche than other metrics of offensive/defensive prowess, but when put in context these stats determine exactly how many yards the opposing teams are forced to play, clearly impacting whether or not the team is a champion.

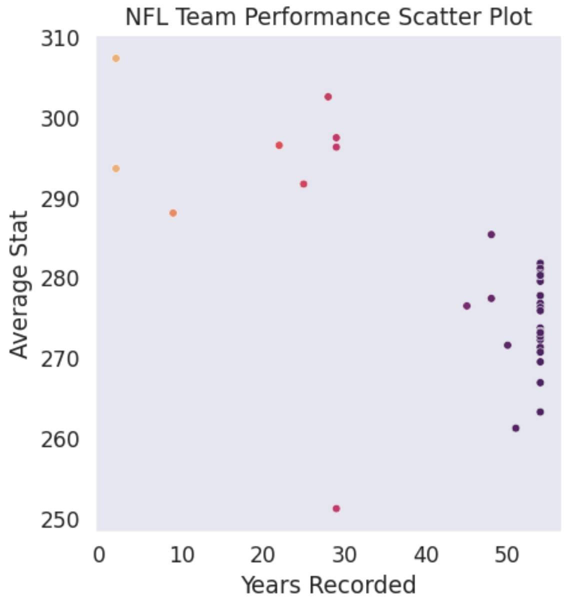

Features Importance of SuperBowl ML Model
Features
offense rushing Att
defense rushing Att
special-teams punt-returns Ret
special-teams punts OOB
special-teams punts TB
defense rushing 40+
special-teams kickoff-returns KRet TD
special-teams punts TD
special-teams punt-returns PRet T
special-teams kickoff-returns FUM
0.000
0.005
0.010
0.015
Feature Importance

Other Visualizations

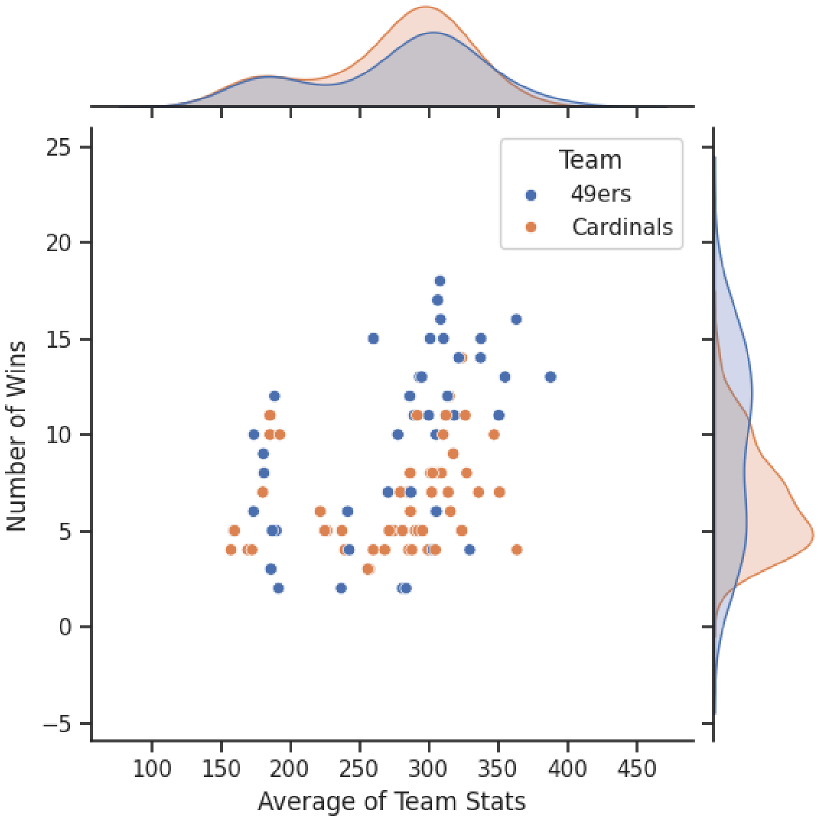
To deepen our understanding of the statistical landscape, we used multiple visualizations to examine both the trend in overall player statistics over the course of around 60 years and the impact that the Law of Large Numbers has on the statistics of individual NFL teams. Our examination of NFL stats over time found a large section of outliers in 1981 that should be eliminated, to guarantee that trends are accurately represented in ML modeling without being skewed by extreme results. This visualization can be used to inform further data preprocessing steps, such as outlier removal, to improve the accuracy and efficacy of future model training efforts. Furthermore, these visualizations show trends in the performances and reliability of team averages as more data is collected, which in turn offers more valuable insights into the dynamics of team stats in the NFL.



Another useful visualization shows the impact of the Law of Large Numbers on team performance over time. Teams' average statistics become less prone to significant outliers as they



play more years in the league. We used this analysis to show how team statistics stabilize over time as additional data becomes available, in addition to offering insights into the evolution of player performance. Despite some variance in team performance, the average stats of all teams appear to approach a constant centered around 270 as the number of years recorded increases, drastically decreasing the possible variance in the data. This general downward trend also inverses the general upward trend shown in the previous visualization.



Further examination in player statistic trends shows another interesting hidden pattern. Using a joint plot distribution for team average statistics shows that a teams number of wins is not entirely dependent on their average statistics. For a high-performing team, such as the San Francisco 49ers, an increase in average team stats appears to roughly correlate to an increase in the number of wins; however, for a low performing team, such as the Cardinals, an increase in average statistics appears to have little to no impact on the distribution of the number of wins. This indicates that there is likely a certain “X-factor” that makes certain teams high-performing outside of their team’s statistics.

Statistical Tests

Chi-squared Test

To evaluate the performance of our predictive model against established standards of professional gamblers, we conducted a Chi-squared test. Professional gamblers typically achieve a success rate of 54% in predicting game outcomes. Our model, which utilizes last year's player statistics to predict game outcomes, achieved a different success rate.

The Chi-squared test yielded a p-value of approximately 9.6338e-06, indicating that the differences in success rates between our model and the standard were highly significant. This result suggests that our model performs significantly better than the typical pro gambler standard, as the p-value is well below the conventional threshold of 0.05.

The statistical evidence from the Chi-squared test substantiates the effectiveness of our model. By significantly surpassing the success rate of seasoned gamblers, our model demonstrates its potential as a valuable tool for bettors and sports analysts alike, potentially reshaping betting strategies in the sports community.

T-test

We conducted T-tests to compare the performance statistics of Super Bowl-winning teams against those of the losing teams across various categories including passing, rushing, receiving, scoring, defense, special teams, and other relevant metrics. Our hypothesis was that winning teams would show statistically significant better performance in several key statistics compared to losing teams.

The T-tests revealed significant differences between the performance statistics of Super Bowl winners and losers. Specifically, winning teams demonstrated superior performance in numerous areas such as passing accuracy, yards per attempt, total touchdowns, and defensive metrics like sacks and interceptions. These results suggest that not only offensive capabilities but also robust defense and special teams play a critical role in achieving Super Bowl victories.

The T test conclusively shows that Super Bowl-winning teams outperform losing teams in critical statistical categories. These results support the notion that a combination of strong offense, resilient defense, and effective special teams contribute significantly to Super Bowl success. These insights could guide teams in strategic planning and preparation to enhance their competitive edge in pursuit of the championship.

Conclusion

Our investigation into utilizing NFL player statistics for predicting game outcomes leveraged sophisticated data manipulation and machine learning techniques to achieve promising results. The application of statistical tests, such as the Chi-squared and T-tests, provided insights into the significant factors influencing game outcomes, and the differences between high performing and low-performing teams. For instance, the Chi-squared test elucidated a statistical difference between our models' performance and professional gambler averages, showing high levels of statistical confidence that our model outperforms professional gamblers on the training set. The machine learning models, particularly the Random Forest Classifier, demonstrated considerable efficacy in forecasting game winners, reflecting the intricate interplay between diverse team statistics and game results. This model's performance, while variable under different parameters, underscores the complex dynamics at play, suggesting that a multifaceted approach to predictive modeling is crucial for higher accuracy .Furthermore, our visual analytics facilitated a deeper understanding of the relationships between team statistics and game outcomes. Through various visual representations, we observed how certain team characteristics correlate strongly with winning, providing tangible insights that could inform team strategies and betting approaches. These visualizations also helped in distilling complex statistical information into digestible formats, making it accessible for strategic decision-making.

Despite achieving substantial insights, this study opens avenues for further research by incorporating additional variables such as player-specific performance data and real-time game dynamics, which were outside the scope of our current dataset. Expanding our analytical framework to include these factors could enhance the predictive accuracy and offer a more granular understanding of game dynamics. Our research not only contributes to the theoretical framework of sports analytics but also has practical implications for teams, analysts, and bettors, offering enhanced methods for predicting NFL game outcomes. The integration of advanced data analytics in sports is proving to be a game-changer, and as data availability and analytical tools evolve, so too will our ability to accurately predict and influence sports outcomes.

Credits

Research and Data Analysis:

Joseph Russoniello: Led the data collection efforts, performed detailed statistical analysis, and interpreted data results.

Joseph Russoniello: Manipulated input data, creating data structures fit for machine learning processing

Joseph Russoniello: Focused on machine learning modeling, algorithm selection, and optimization.

Zayad Almazrouei: Researched previous sources pertaining to the topic of the project and interpreted their findings.

Managing statistical findings/visualizations and analysis

Sebastion Mendoza: Responsible for creating and handling parts of the visualizations made.

Tongcai Zha: Managed the compilation and analysis of statistical data.

Writing and Documentation

Zayad Almazrouei: Responsible for handling the paper, incorporating team inputs and aligning the narrative with research findings.

Zayad Almazrouei, Joseph Russoniello: Focused on editing, finalizing the paper, and ensuring that the documentation met all academic standards.

Sebastian Mendoza, Zayad Almazrouei: Wrote the comprehensive description of the machine learning models, including detailing the process of the random forest classifier.

Tongcai Zha: Interpreted the statistical significance of the findings in the context of NFL game predictions.

Visualization and Presentation

Sebastian Mendoza, Joseph Russoniello : Created graphical representations of data and results, ensuring clear communication of complex information through visual means.

Joseph Russoniello: Prepared the final presentation, highlighting key findings and conclusions to present in class.

Special Acknowledgments

Joseph Russoniello: Provided critical support in literature review and validation of sources as well as ensuring adherence to project objectives.

Zayad Almazrouei: Offered logistical support and facilitated team communications.

References

Donnelly, J. P. (2013). NFL betting market: Using adjusted statistics to test market efficiency and build a betting model. *Claremont Colleges*, *721*.

NFL. (2009, January 25). Here's the 10-step program to creating a championship culture. *NFL*. https://www.nfl.com/news/here-s-the-10-step-program-to-creating-a-championship-culture-09000d5d80e57f03

Sathish, S. (2023, September 7). *Artificial intelligence*. Logan Data Inc. https://logandata.com/nfl-and-data-how-analytics-is-used-in-the-game-prediction/

Stone, J. (2023, February 3). Analyzing the odds: How to predict the outcomes of NFL games. *Bettors Insider*. https://www.bettorsinsider.com/nfl/2023/02/03/analyzing-the-odds-how-to-predict-the-outcomes-of-nfl-games

Sussman, E. (2011, October 18). NFL fact or fiction: What makes a winning team? *Bleacher Report*. https://bleacherreport.com/articles/897119-nfl-fact-or-fiction-what-makes-a-winning-team