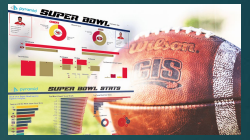
# Predicting NFL Outcomes from Previous Performance Statistics

Sai Dandu, Max Friel, Dorian Tisdale

# Introduction and Problem Statement

The National Football League (NFL) is one of the most data-rich sports leagues and provides massive data source of structured and unstructured data from various aspects from performance metrics to game time weather information.

Given the availability of this data, the primary challenge we are attempting to tackle is can the outcome of future NFL games be predicted within a certain level of confidence for which team will win the game and if the teams will cover the spread.




This project will attempt to address this challenge by leveraging AI/ML techniques that we have learned in our coursework while considering how AI is leveraged in related sports like soccer as well as AI applications in the sports video game industry.

# Prior or Related Work

- Decoding NFL Team Offenses: A comprehensive Data-Driven Analysis of the 2022 Season https://www.slideshare.net/slideshow/nfl-2022-stats-report/265537944
- Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics
- Sports Results Prediction Model Using Machine Learning; Alden Obradovic, Dino Keco

# NFL Play-By-Play Data (NFLFastR)

The nflfastR data is a dataset used for analyzing NFL play. The data is used in sports analytics, containing detailed information about every play in the NFL season.

## Dataset Features:

- Game Information
- Play Information
- Team and Player Stats
- Advanced Metrics (Expected Points Added, Win Probability, etc.)
- Penalty Information
- Special Teams
- Game Outcomes

## Applications:

- Exploratory football analytics
- Fantasy football projections
- Team strategy evaluations
- Player performance impact and analysis
- Gambling and predictive modeling

# Data manipulation

- Calculated Last 4 and last year for chosen player stats
- Calculated Time Of Possession For Teams
- Dataset did not include 2024 data

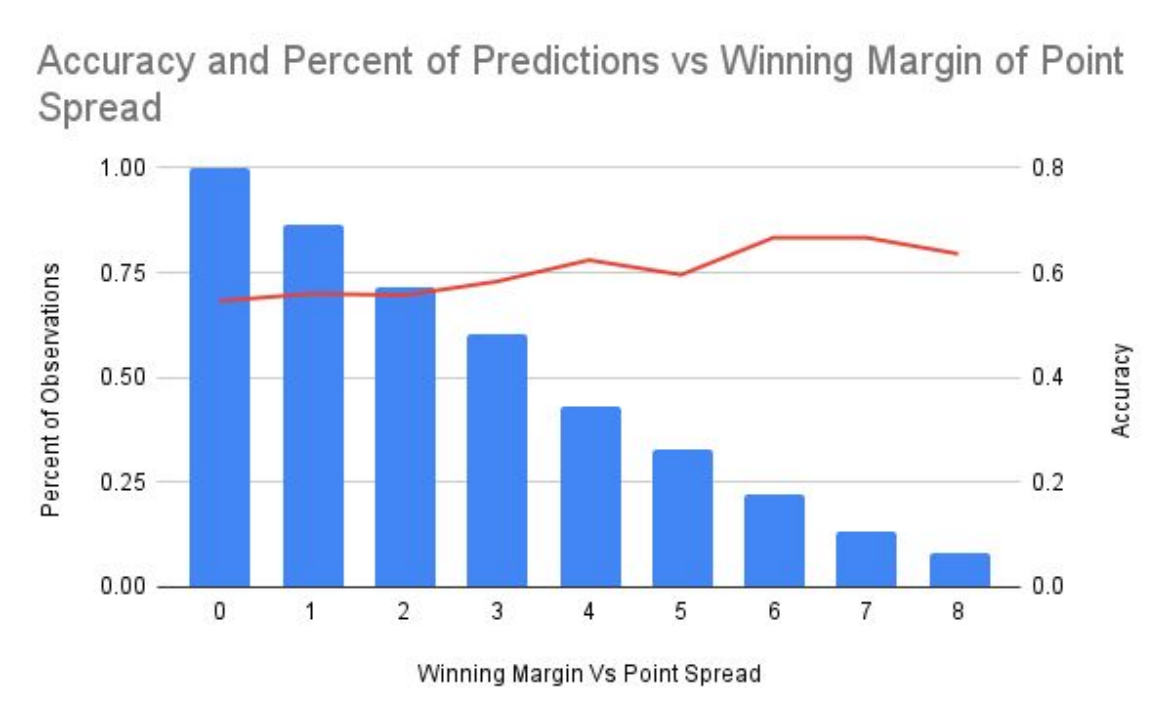
# Support Vector Machines(SVM)

- Allows us to predict future outcomes by separating into groups based on past performance
- After experimenting with different kernel tricks, decided on polynomial of degree 47
- 55.7% Accuracy

# Linear Regression

- Easy to implement and computationally efficient, making it well-suited for analyzing large datasets like NFL game statistics.
- Allows us to predict numeric outcomes - the final score between the two teams
- Better results when treating one game as two observations rather than running a regression twice on a one observation per game dataset
- Can Leverage the predicted scores to then determine who will win against the spread
- 54.6% overall accuracy

# Linear Regression

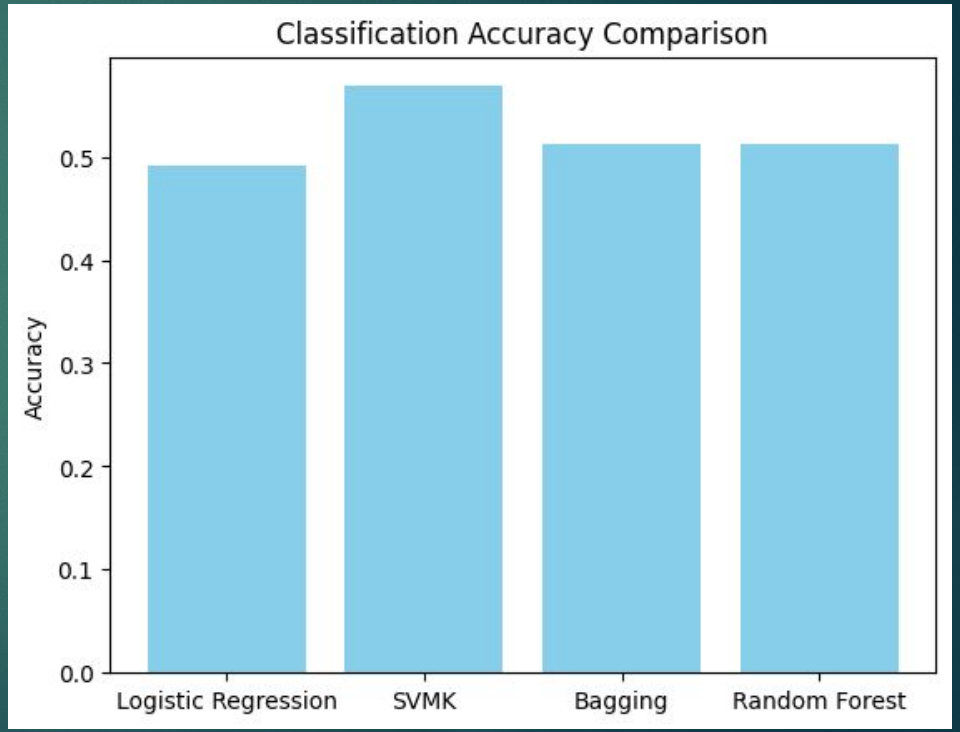


# Logistic Regression

- Logistic regression is used to predict binary outcomes, making it ideal for scenarios like predicting whether a team will win against the spread.
- Uses a sigmoid function to estimate probabilities, providing a likelihood of an outcome.
- Highlights the influence of different game metrics on the outcome, offering insights into which metrics are most impactful.
- 54.98% overall accuracy

# Ensemble Bagging and Random Forests

- By using Ensemble methods, we can show the advantages over a single method approach for analysis.
  - Some future work could levage ensemble methods to test the reduction of noise cause in the results created by single model weaknesses or to reduce the risk of overfitting by leveraging multiple models trained.
- While Bagging and RF achieved similar accuracy, they both independently outperformed the Logistics Regression model but underperformed compared to SVMK.
  - SVMK overperformance could be due to its ability to handle non-linear relationships.



# Future Work

- Performance data over the past 4 games/season does not take into account a weighting for the opponents. Adding a weight may give more accurate results.
- Ensemble methods were performed based on binary results. Using the probabilities rather than a simple win/loss may improve accuracy.
- Due to the ever changing landscape of the NFL strategy and rules there is only so much data that is relevant to current outcomes. Given that, using s-fold cross validation may be useful to improve accuracy.
- Research further into why the weights for linear regression were different when calculating the home and away score separately

# Conclusion

- Linear Regression produced interpretable results with an accuracy of 54.6%
- Logistic Regression achieved a validation accuracy of 54.98%, modeling binary outcomes like win/loss with feature transformations.
- SVM and Ensemble Methods showed the most potential by achieving an accuracy of 55.7%.

# References


- Moustakidis, S.P., Plakias, S., Kokkotis, C., Tsatalas, T., & Tsaopoulos, D. (2023). Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics. *Future Internet, 15*, 174.
- Mozgovoy, M., Preuss, M., & Bidarra, R. (2021). Team Sports for Game AI Benchmarking Revisited. *Int. J. Comput. Games Technol., 2021*, 5521877:1-5521877:9.
- Sports Illustrated. (n.d.). *Madden NFL 23 Super Bowl 57 simulation predicts winner.* Retrieved November 17, 2024, from https://www.si.com/videogames/features/madden-nfl-23-super-bowl-57-simulations-prediction
- Wang, Z., Veličković, P., Hennes, D. *et al.* TacticAI: an AI assistant for football tactics. *Nat Commun* 15, 1906 (2024). https://doi.org/10.1038/s41467-024-45965-x
- Obradović, A., & Kečo, D. (2024). Sports results prediction model using machine learning. *SAR Journal, 7*(3), 184–189. https://doi.org/10.18421/SAR73-03