ML Project Proposal

Introduction/Background

In this project, we will utilize ML algorithms to predict the final score, spread, and total points of NFL games by training on historic data and testing on the 2023 season outcomes.

Literature Review

Our main literature review comes from the University of Michigan, where students used neural networks to predict the outcomes of college football games by running 5 different regression models. This study is trained on years of historical data on various team metrics and tested on live results. Their output shows that Lasso Regression was the most effective predictor; however, there is room for continuous improvement in this space, which we hope to capture by analyzing the NFL rather than NCAA play [2].

Dataset Description

The dataset we have chosen contains team data for each game of the NFL season from 2002-2023. Each row is an NFL game and each column is information about the game statistics such as location, box scores, or time of possession.

Dataset Links

- https://www.kaggle.com/datasets/cviaxmiwnptr/nfl-team-stats-20022019-espn

Problem Definition

Problem

It is currently difficult to predict the outcome of NFL games due to the complexity of the game and randomness involved. We aim to solve this problem by creating an ML model using pre-game statistics to predict the outcome of matchups.

Motivation

In this study we aim to reduce uncertainty in the field of sports forecasting and fill in the gaps of existing research on predicting outcomes as a whole through the contained sample size of the NFL. This study has practical and fiscal applications in the fields of sports forecasting, betting markets, and team strategy development.

Methods

Data preprocessing methods:

1. Data Cleaning: We plan to remove seasons prior to the 2007 seasons due to inconsistencies. We also plan on saving data from the 2023 season for testing.
2. Dimensionality Reduction: Of the 61 features present in our dataset, there are numerous which we do not think are statistically significant (date, time, etc.).
3. Normalization of Features: Many features in our dataset encompass vastly different ranges of possible values. By normalizing the data we ensure that "all features contribute equally to the final prediction" [1].

ML Algorithms

1. Neural Net: This would handle the roughly 13,000 data points with ease, and it would be able to account for the diversity of both quantitative and categorical features [3].
2. K-Means: This would be useful to cluster teams into groups based on their relative strength. This could be helpful for assessing trends in these clusters.
3. Random Forests: This would reduce overfitting and be more resilient to noise in the datasets, useful for predicting the complex non-linear aspects of NFL games.

Results/Discussion

Quantitative Metrics

1. Root Mean-Squared Error: RMSE is ideal for neural networks and random forests because they predict continuous values. A low RMSE value would demonstrate the model is effectively predicting scores with minimal error.
2. Mean Absolute Error: This would work well for our algorithms because it's not highly sensitive to outliers, which could give us a better understanding of the difference between actual and predicted values.
3. Coefficient of Determination: This is useful for neural networks and random forests because it indicates how well the model fits the data by assessing the variance relationship between results.

Project Goals/Expected Results

The goal is to predict the final score, spread, and total points for NFL games. We are aiming to predict the winner of the spread correctly 53% of the time, as this metric constitutes a model as "profitable" through attaining benchmarks of an RMSE of <5, MAE of <6, and an R^2 of >0.80 [4].

References

[1]S. Jaiswal, “What is Normalization in Machine Learning? A Comprehensive Guide to Data Rescaling,” *Datacamp.com*, Jan. 04, 2024. https://www.datacamp.com/tutorial/normalization-in-machine-learning (accessed Oct. 02, 2024).

[2]L. Boll, “Gridiron Genius: Using Neural Networks to Predict College Football,” *Umich.edu*, 2023, doi: https://hdl.handle.net/2027.42/176935.

[3]B. Radjewski, “Talking Tech: Building an Artifical Neural Network to Predict Games,” *CFBD Blog*, Nov. 12, 2021. https://blog.collegefootballdata.com/talking-tech-building-an-artifical-neural-network-to/ (accessed Oct. 02, 2024).

[4]“Predicting NFL Total Score and Point Spread Bets,” *Qu.edu*, May 17, 2024. https://iq.qu.edu/experiential-learning/course-projects-and-capstones/student-projects/predicting-nfl-total-score-and-point-spread-bets/ (accessed Oct. 02, 2024).

Contribution Table

Name	Proposal Contributions
Reuben Covey	Created the video presentation. Defined the problem and motivation in the proposal.
Sebastian Stephens	Created the GitHub repository and pages as well as made the project timeline gantt chart.
Sofia Varmeziar	Researched quantitative metrics we could use to test the outputs of our model. Defined our goals and expectations for the project.
Greyson McReynolds	Researched data pre-processing techniques and ML algorithms that we could feasibly use to complete the project.
Sam Deckbar	Researched past literature on the subject, did the introductory part of the proposal, and found a good dataset to use for our project.