Predictive Modeling of NFL Total Game Scores

By Dylan Dietrich, Conor Gibbons, and Ben Berinsky

In the past few years, the industry of sports gambling has exploded, with legal NFL betting estimated to reach a high of $30B in 2025. Sports betting companies advertise virtually everywhere, and it has never been easier to place bets with the prevalence of mobile gambling apps. Football betting lines, including over/unders and point spreads, are calculated through both statistical modeling and market adjustment based on the behavior of bettors. A record number of Americans wagered on last year's Superbowl, with over 68 million placing wagers [2]. Odds are made to heavily favor sports books, and a vast majority of gamblers lose money in the long run. We wanted to see how well various models perform at predicting total score between two teams based on their scoring performance in the past, both in the previous season and throughout the current season, or if we needed more updated information such as injuries, specific players, and coaches.

Research question: Can we accurately predict total points scored in a game between two teams based on past scoring data?

Methods:

The dataset we used is from Kaggle and consists of around 4,500 NFL games spanning multiple seasons, containing game level information, team names, final scores, betting lines, stadium information, and weather information. Data preprocessing involved getting rid of variables we didn't want to use, such as weather (there were too many missing values), and also adding some new variables which will help us to train our model and predict the accurate points scored. Specifically, we computed each team's win-loss-tie record and calculated rolling averages of points scored and allowed throughout the current season and incorporated each team's average scoring from the previous season as a proxy for team strength.

We implemented four models to predict total game scores: Ridge Regression, Gradient Boosting, SVR, and a Neural Network. All models used preprocessing pipelines with StandardScaler for numeric features and OneHotEncoder for categorical variables like playoff indicator. Hyperparameter tuning was performed using GridSearchCv with a 5-fold cross-validation, optimizing for negative mean squared error.

Findings:

Unfortunately, all of our models performed relatively poorly in their abilities to predict final NFL scores. As shown in the table, none of the R^2 values were above 0.1 with the best performing model being our ridge regression that had the lowest R^2 and mean squared error. These scores indicate that the models are not a good fit for the data and that there is a lot of unexplained variance. The MSE values of all models are in the high 100s into low 200s which is pretty high given our model is predicting NFL scores. These large values show that the models' predictions are generally far from the actual values. Overall, our results were worse than we had hoped with our best model being the ridge regression by a small factor.

Model	**Ridge Regression**	**SVM**	**Gradient Boosting**	**Neural Network**
Best Hyperparameters	Alpha: 100 Solver: auto	C: 10, Epsilon:0.01 Gamma: 0.001 Kernel: rbf	LR: 0.01 Max depth:3 Min samples l:2 Min samples s:5 N estimators:200 Subsamples: 0.8	LR: 0.0005 Hidden units: 16 Epochs: 150
R^2	0.0844	0.03	0.031	0.02
MSE	172.98	208.42	208.462	188.854

Conclusion:

For our best model, we got an R^2 of 0.08 and an MSE of 186.7 on the test set. This means that our best model, ridge regression, had a weak positive correlation for predicted score to actual score. Our model was approximately 13 points off, on average, when predicting points. Our penalized linear regression model performed better on this data than our other models because the structure of our data was very linear. Our predictors included past information about team scoring, including the past season's scoring average of both teams as well as the current season's rolling scoring average. Regularization stabilized our linear regression model. The relationship between our predictors and target was roughly linear, so our other models were unable to make accurate predictions.

Throughout our process, we found that overall, our models were limited in their predictive power. From this, we conclude that more information is needed to predict the final score of NFL games than just past scoring data. Although models such as SVRs, neural networks, and gradient boosted decision trees can capture a great deal of information from data, there are just too many extraneous factors that affect the final score of NFL games. With the impact of injuries, roster turnover from year to year, and many other team specific factors, more than just a model is needed to accurately predict the final score. This highlights the necessity of human actors in data science. Models can be used as a good baseline for predictions, but humans must factor in elements that cannot be captured by models and implement this to make accurate predictions, at least in NFL scoring.