# Deep Learning Forecasting of NFL Game Outcomes Relative to Betting Odds

## Report for the Computational Project Managerial Forecasting and Decision Analysis (Fall 2024-EIND 468-588)

Quinlin Gregg
Mechanical & Industrial Engineering Department
Montana State University
Bozeman, Montana, 59717
Email: quinlin.gregg@msu.montana.edu

*This paper presents a comprehensive analysis of different predictive modeling techniques applied to time series forecasting. We explore several models, including Linear Regression (LSR), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Multilayer Perceptrons (MPL), and various transformation techniques like the Holt method. Each model's performance is evaluated on a set of real-world data, with an emphasis on their ability to capture temporal dependencies and provide accurate forecasts. By comparing these models, this work aims to offer valuable insights for practitioners in choosing the best-suited approach for their specific forecasting needs.*

Keywords: National Football League; Deep Learning; Sports Analytics; Forecasting

## Literature review

Many studies have been done examining the efficiency or lack thereof of the betting market, usually looking for arbitrage opportunities. Golec and Tamarkin (1991) [1] used linear regression to look for biases in betting odds. Their formulation included the betting odds themselves, as well as dummy variables representing whether a team was at home or whether a team was favored. If the odds were unbiased, the regression coefficients of the dummy variables should be zero and that of the betting odds should be one. They ultimately found that bets on home teams and underdogs did better than those on favorites or visiting teams.

Gifford and Bayrak (2023) [2] investigated whether binary logistic regression or decision trees could be used to predict which team won an NFL game. They used data including the number of turnovers on offense and defense, passing and rushing yards, and whether the game went to overtime to predict the winner. They evaluated their models using misclassification rate; that is, how often the model predicted that the losing team won. They found that the binary logistic regression had a 16.9% misclassification rate, and the decision tree had a 21.6% misclassification rate.

## Problem description

There are multiple types of bets that can be made on sports. In this paper, two will be focused on. The first is a spread bet. A spread bet is betting on whether a team (usually the home team) will beat another team by at least *n* points. That is, $\textit{Home Score - Away Score} >^{?} n$. If the home team beats the away team by at least $n$ points, it is said that they "covered." The second is an Over/Under bet, where the bet is on whether the combined point total of both teams will exceed $n$ points. That is, $\textit{Away Score} + \textit{Home Score} >^{?} n$. These bets are typically formulated as "50-50" bets; they should have an equal chance of being higher or lower than the $n$ point level.

The existing research tends to focus on evaluating market efficiency or on predicting the winner of the games based on the statistics from the game itself. This project aims to predict the distribution of possible scores for both teams to evaluate how likely a given bet is to hit. For example, given how the teams have performed in recent games, how likely is it that the total score of a game will exceed the over/under?

A working model of this sort could be used for multiple things. It could be used to build a portfolio of profitable sports bets to "beat the house." It could also be used by the sportsbooks themselves to more accurately price their bets according to the payouts they want to offer. This could potentially allow for more creative and varied payout structures, offering players more options.

This model assumes that the outcome of each drive is independent of all others, including both those of the other teams and previous drives of the same team. It then attempts to predict the proportions of drives that will end in a touchdown, a field goal, and a punt (defined here as any drive that does not end in a touchdown or a field goal). Once these values are forecasted, the outcomes of the game can be modeled with a trinomial distribution to determine the distribution of final scores. The advantage of using this approach is that once the drive outcomes are forecasted, simulating each game with a trinomial distribution is computationally quick, and many Monte Carlo simulations can be done to find the distribution of values. The advantage of this is that once the model for forecasting the drive-level outcomes is built, calculating the game-level outcomes is a fast and simple Monte Carlo simulation. This potentially allows one built model to forecast many games efficiently.

## Mathematical Modeling

The forecasting model is built in two parts. First, the proportion of drives that will end in a touchdown, a field goal, and a punt (defined here as any drive that ends in neither a touchdown nor a field goal) are predicted. To do this, the previous 16 games for each team are pulled to see what percentage of drives they scored on and allowed scores on. These produce eight time series in total: touchdown and field goal; home and away; and offense and defense. The respective offensive and defensive forecasts are then combined by averaging them. For example, the percentage of touchdowns scored by the away team is the average of the touchdowns scored by the away team's offense and the touchdowns allowed by the home team's defense. These time series were used to build multiple different models, including a Least-Squares Regression, Holt exponential smoothing, a Recurrent Neural Network, a Long Short-Term Memory Model, a Multi-Layer Perceptron, and a Transformer Model.

Second, these percentages are used to build the distribution of possible point totals each team can score. Under the assumption that drives are independent of one another, the outcome of the game can be modeled with a trinomial distribution. Random samples were pulled from the trinomial distribution in a Monte Carlo simulation to build a distribution of possible point totals for each team, as well as distributions for the sum and differences of their scores.

## Solution Method

The main benefit of this model is its simplicity. It does not rely on proprietary formulations of advanced stats like EPA (Expected Points Added) or DVOA (Defense-Adjusted Value Over Average), instead opting for simple drive outcomes that are indisputable. Furthermore, once the model for forecasting the drive-level outcomes is built, forecasting the outcome of the game only requires a computationally quick and simple Monte Carlo simulation.

To train the model, data from the 2020-2023 NFL seasons were pulled, only including regular season games. Starting with the 2021 season, every game was pulled and the previous 16 games for both teams were looked up. The known drive-level outcomes of these games were used to build the eight time series: away/home, touchdown/field goal, and offense/defense. In all, there were 1,706 time series across the offensive and defensive touchdowns and field goals. (The away and home time series were merged by the side of the ball and score type). The true known value was added to the end of the time series, with the goal being to use the 16 previous values to predict the 17th. The data were split into 80% training data and 20% testing data, and each of the models was run.

The 16-game window was chosen for two seasons. Firstly, the 2020 season had 16 games, so week 1 of the 2021 season can use all of the 2020 season to build its forecast. Secondly, starting in the 2021 season, the NFL moved to a 17-game schedule, so the final week of the 2021 season and beyond exclusively use data from earlier in the season for their forecast.

Once each of the models was built, data from the 2023 and 2024 NFL seasons were used to test its ability to forecast real games. (The 2023 season must be included here as well to give something for the model to use to forecast the early parts of the 2024 season.) Each model was run to forecast the mean score and score distribution of each game. The mean score can be calculated as $7 * p_{touchdown} + 3 * p_{field\ goal} + 0 * p_{punt}$ where $p_{punt} = 1 - p_{touchdown} - p_{field\ goal}$. The distribution is modeled as random samples of a trinomial distribution such that $X \sim trinomial(p_{touchdown}, p_{field\ goal}, p_{punt})$.

Many algorithms were tested to build the model for forecasting drive-level outcomes, including Least-Squares Regression, Holt exponential smoothing, a Recurrent Neural Network, a Long Short-Term Memory Model, a Multi-Layer Perceptron, and a Transformer Model.

## Results and conclusion

The results of the models were mixed. Some models have the potential to be used for forecasting games and for making profitable bets, but most do not.

The models did not match especially well with the testing data. They, on average, had about a 0.100 MAE, which means that the average prediction is 10 percentage points of the true proportion. The models typically perform better predicting field goals than they do touchdowns, likely because modern kickers consistently make their kicks.

| Model | TD Off | TD Def | FG Off | FG Def |
|---|---|---|---|---|
| LSR | 0.105 | 0.104 | 0.091 | 0.095 |
| RNN | 0.113 | 0.113 | 0.103 | 0.108 |
| LSTM | 0.110 | 0.104 | 0.091 | 0.095 |
| MPL | 0.115 | 0.116 | 0.101 | 0.108 |
| Transform | 0.121 | 0.105 | 0.094 | 0.095 |
| Holt | 0.135 | 0.140 | 0.118 | 0.120 |

Fig 1: MAE the test data for each model when training on the 2020-2023 NFL seasons.

When forecasting the 2024 NFL season, the models fared all right. In terms of expected final score, the models were typically off by about 7 points per team. (That means that in theory, the over/under total could be off by as much as 14 points). This means that the models are not especially good at predicting how many points the teams will score; however, considering how erratic and unpredictable sports are themselves, this is to be expected.

| Model | TD MAE | FG MAE | Score MAE |
|---|---|---|---|
| LSR | 0.108 | 0.099 | 7.378 |
| RNN | 0.115 | 0.106 | 8.214 |
| LSTM | 0.108 | 0.098 | 7.367 |
| MPL | 0.116 | 0.105 | 8.480 |
| Transform | 0.107 | 0.101 | 7.581 |
| Holt | 0.132 | 0.120 | 9.512 |

To make money on “50-50” bets, the bets need to win about 52-55% of the time to overcome the edge the sportsbook has. To assess whether these models can reach that threshold, the percentage of games where the model correctly predicts whether the over or under will hit as well as whether the away or home team will cover was measured for the 2024 season. The models generally do not do better than the 50/50 chance we would assume by default given the nature and formulation of the bet. The models broadly do not outperform a random coin flip. This suggests both that there is room for improvement in this model and that the sportsbooks are good at making sure their betting lines are true 50/50s. The LSR, MPL, and Transform models are potential candidates for models that beat the sportsbooks in over/under if they can maintain their accuracy long-term.

| Model | Spread Accuracy | O/U Accuracy |
|---|---|---|
| LSR | 49.5% | 55.2% |
| RNN | 49.4% | 49.0% |
| LSTM | 49.0% | 50.0% |
| MPL | 45.4% | 51.6% |
| Transform | 49.5% | 51.6% |
| Holt | 46.9% | 48.5% |

Future potential areas of inquiry include evaluating different window lengths and separating the raw data before forecasting. The window length of 16 was chosen to align with the NFL schedule, but there is no reason to think that is the optimal window length. It is possible that a shorter window length would be more appropriate to account for player injuries that could dramatically affect a team’s performance in game.

Separating the data into offense and defense before forecasting could also potentially improve performance. Currently, the effects of the offense and defense are all wrapped up in one number. For example, if the away team scored touchdowns on 85% of drives, is that because they are a good offense or because they are playing a bad defense? The difference could be important when forecasting how the teams may perform later in the season. If that 85% could be separated into offensive and defensive components, then the accuracy of the model might be improved. Potential methods for doing so could involve including other measures, like the aforementioned EPA and DVOA to isolate offensive and defensive effects.

## References

[1] J. Golec, “The degree of inefficiency in the football betting market Statistical tests,” Journal of Financial Economics,

vol. 30, no. 2, pp. 311–323, Dec. 1991, Doi: https://doi.org/10.1016/0304-405x(91)90034-h.

[2] M. Gifford and Tuncay Bayrak, “A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression,” Decision Analytics Journal, vol. 8, pp. 100296–100296, Aug. 2023, doi: https://doi.org/10.1016/j.dajour.2023.100296.