COLLEGE OF ENGINEERING
HONORS PROGRAM
UNIVERSITY OF MICHIGAN

Gridiron Genius: Using Neural Networks to Predict College Football Games

Luke Boll (Honors Capstone) [1]   Jason McCormick (Advisor) [1]

[1]University of Michigan

Research Question

How well does a neural network perform on a regression task in comparison to other machine learning algorithms in a nondeterministic environment like college football?

Introduction

- A neural network is a machine learning method inspired by the human brain. It consists of a large number of interconnected processing nodes, or ”neurons”, that work together to learn and recognize complex patterns in data.
- The perceptron is the fundamental building block of a neural network. It accepts inputs, assigns weights to each input, adds them up, and then applies an activation function. By altering the weights and biases of the perceptrons, the neural network can learn to recognize patterns in the input data and generate predictions based on that information.
- Because of their ability to learn and improve their performance over time, neural networks are a powerful tool for predictive modelling in complex problems.
- College football is a highly dynamic and unpredictable environment with many factors affecting the outcome of a game, making it difficult to build accurate predictive models. However, neural networks can help account for the complexity of the sport and have been shown to perform well in nondeterministic environments before.

Data Collection

- To investigate the research question and train the neural network, data was acquired from scraping the site teamrankings.com using the Beautiful Soup library in Python
- Data was collected for team statistics from 7369 games over the span from the 2007 season to the 2018 season in the following categories:

Scoring Offense	Scoring Defense
Offense by Quarter	Defense by Quarter
Total Offense	Total Defense
Rushing Offense	Rushing Defense
Passing Offense	Passing Defense
Special Teams Offense	Special Teams Defense
Turnovers	Penalties

- For all of these statistics, the following averages were collected:

Season	Previous Season
Last 3 Games	Last 1 Game
Home	Away

- Overall, this data provided a comprehensive assessment of team performance and momentum in all phases of a football game

Data Preprocessing

1. **Removing games played before October**. During the first month of the season, many teams play out of conference games that are not necessarily reflective of a team’s performance over the course of a season. Therefore, games played before October were removed from the dataset based on expert knowledge of college football.
2. **Creating target variable from game scoreline**. A neural network or any machine learning algorithm needs to know what is predicting. In this case, it is attempting to predict the difference between the home and away final score.
3. **Removal of highly correlated input features**. Highly correlated input features provide redundant information to the model, which can lead to overfitting and make it difficult for the model to distinguish between them. Therefore, any features with a correlation ≥ 0.8 were removed from the feature space. The choice of which feature to remove was based on the correlation with the target variable and which provided the greatest information to the model.

After these steps, there were 5097 games with 976 input features each. These were then broken up in training, validation, and testing splits with the respective ratio of $64 : 16 : 20$.

Model Training and Implementation

- Neural networks are intended to minimize a loss function. For most regression, the standard is to use mean squared error and this is what the model was trained to minimize. Mean squared error is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the true point differential, $\hat{y}_i$ is the predicted point differential, and n is the number of data points.
- The hyperparameters of the model were selected based on using a Grid Search and finding the model corresponding to the lowest validation error.
- For implementing the model, the Pytorch library was used and training was continued until the validation loss had not reached a new minimum for 5 epoches, or passes, over the dataset. This was done to prevent overfitting and speed up the training process. An example of the training plot can be seen below:

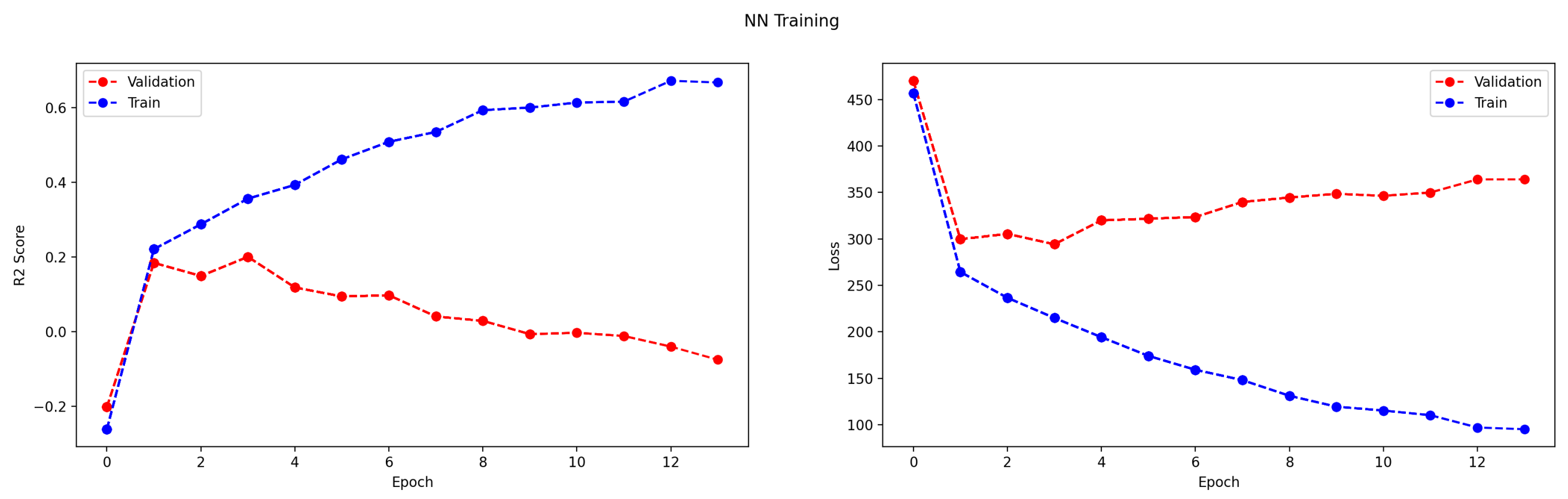

Figure 1. Image of the training plot for a neural network

Results

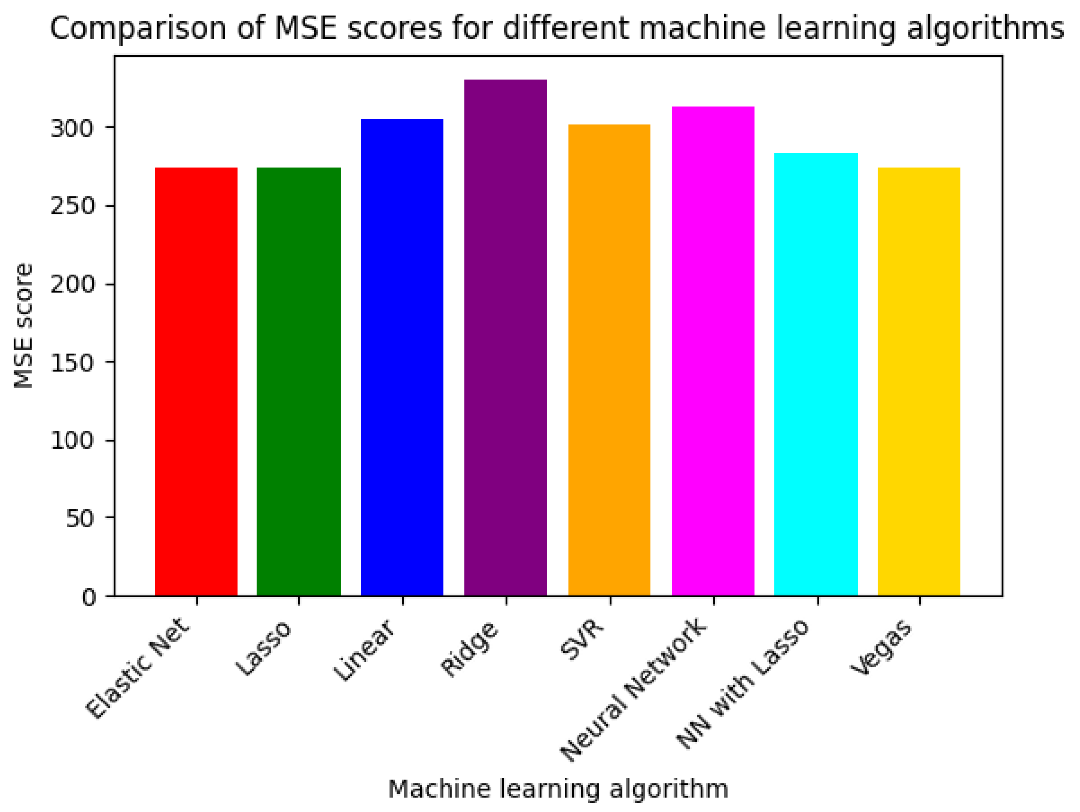

Figure 2. Comparison of the MSE on the testing set for several machine learning algorithms

- Overall, the Vegas point spread had the strongest modelling of the final score differential for college football games.
- The neural network trained on the lasso regression inputs corresponded to a lower testing error, demonstrating the importance of data when training these networks.

Conclusion and Further Research

- The study examined the use of neural networks for predicting college football outcomes, and compared their performance to several other commonly used regression methods as well Vegas point spreads.
- As shown by the results, the neural network was able to outperform many of the other means of regression. However, it was still not as accurate as the Vegas point spread predictions or the lasso regression models
- Another interesting finding is that the lasso regression only included two features as its input and one of them was the Vegas spread for the game. This is part of the reason that those models correspond so favorably with the Vegas spread. One potential extension would be to look at how the different algorithms work without this feature.
- Overall, neural networks are a promising approach for predicting college football outcomes, and may have applications in other domains as well. However, this is contingent upon having a strong dataset to be used in the modeling.

References

[1] Team rankings, 2005-2023.

[2] Michael E. Akintunde, Elena Botoeva, Panagiotis Kouvaros, and Alessio Lomuscio. Formal verification of neural agents in non-deterministic environments. *Autonomous Agents and Multi-Agent Systems*, 2021.

[3] Andrew Blaikie, Gabriel Abud, and John A. David. Nfl & ncaa football prediction using artificial neural networks. 2011.

[4] Charles South and Edward Egros. Forecasting college football game outcomes using modern modeling techniques. *Journal of Sports Analytics*, 2020.

[5] Cecilia Summers and Michael J. Dinneen. Nondeterminism and instability in neural network optimization. 2021.

lukeboll@umich.edu