Garrett Francis

3 December 2025

CSC422

Prof. Moseley

Final Technical Report

Problem Definition & Motivation:

There is a lot of information out on the internet that tells us if we are doing healthy things or not and this can be paralyzing at some points when making decisions related to our own health. Is red meat good for us or not? Is diet soda worse than regular soda? How important is sleep? There are so many of these kinds of questions that doctors and "health professionals" online argue over that you can never really know for sure if you are living a healthy lifestyle or not. The question that this project is trying to answer is whether we can notice general trends in people's lifestyles that are indicative of a better quality of life, as well as the length of life. I will be utilizing a neural network to try to analyze this problem due to the high dimensionality of the data and the non-linear relationships in the data. I want to master implementing a neural network and get a better understanding of back propagation and how the hidden layers work. My original metric for success was having an R2 of .90 for my model to be considered accurate. This project advances beyond my midterm work by implementing a Multi-Layer Perceptron network, which is a Deep Learning model, and additionally, it uses a weight ensemble model to increase the accuracy of the model, as well, which is a big improvement on my midterm project.

Related Work & Background:

When looking at another paper by Raoof Napour [2] about the efficacy of different ML models to enhance the prediction capabilities of health in older adults. They looked at a number of different models and their efficacy when analyzing the health of older adults, and they found that a lot of the models used ensemble models and neural networks, which performed well in making predictions on the data.

Another study [1] that researched the probability of someone with diabetes being a high health care utilizer used four different models, including linear regression, boosted trees, and an MLP. The boosted trees performed well with this data when predicting how bad the condition of a person diagnosed with diabetes might be. This is another case where an ML model is able to accurately predict health-related issues.

Lastly, there was a paper [3] that describes how there are big advancements being made with ML and Deep Learning models being used to improve healthcare treatments and health predictions. The things they are using these models for range from predictions on people's health records all the way to medical imaging.

Methodology & Implementation:

The dataset that I worked with was from Kaggle and had a total of 7 features, not including our target feature, which was the health score. The features are age, exercise frequency, sleep hours, diet quality, BMI, alcohol consumption, and smoking status. These features are very general telling of someone’s lifestyle, but they were still enough to make a prediction on their health score. We processed this data by removing the negative values that were found in alcohol consumption so that it could be better understood by the model. The smoking status data was binary, so we had to convert those data types into binary so the model knew how to handle this information as well. Additionally, all of the data features that were continuous were a function of a continuous variable, which was used to get rid of the outliers so that there would be less noise in the data.

The model that we ended up choosing was the MLP + LightGBM weighted ensemble model because it had the best results. We started out with the scikit-learn MLPRegressor model, and this barely performed better than our baseline Linear Regression model, so I knew that I had to improve on it to show that the MLP was recognizing real non-linear relationships between the data. The reason why I chose an MLP model in the first place is that MLPs perform well on tabular data, and especially data that has a lot of non-linear relationships in it. Additionally, the scikit-learn MLPRegressor model already has a backpropagation algorithm built in, so I didn’t have to implement one myself, which made creating and optimizing the MLP model much easier. LightGBM was used because, in the academic papers that I read on ML in healthcare, a lot of studies were done using gradient boosting, and LightGBM is a gradient boosting model, so it made sense to compare the results of the MLP and LightGBM models and see which performed better. After seeing the results, I decided to make it an ensemble model because the MLP and LightGBM models predict in different ways, so where one might be weak, the other will pick up the slack, resulting in a better prediction.

The architecture of the ensemble model is an MLP and a LightGBM model together, and the final output is weighted 85% by the MLP model and 15% by the LightGBM model. The base MLP barely outperformed our baseline with the parameters that it had, so I used a cross-validation grid search to find the best combination of inputs per layer, learning rate, and alpha for the MLP. The MLP had three hidden layers, and the cross-validation search ranged from 256 inputs down to 96 for the first layer. The MLP also used an Adam optimizer and a ReLU activation function, which seemed to work the best for this data. The LightGBM model used the regression objective, which gave it the best results, and also used an n_estimator value of 500.

The models were trained on 80% of the cleaned data from the dataset. This seemed to work great and didn’t have any terrible overfitting as a result of this split. The MLP with the CV search had the longest training time, which was about 2.5 minutes, since it was testing about 90 MLPs per time to see which one performed the best. I also increased the number of iterations to 2,000 and then turned early stopping on to make sure that the model didn’t quit learning before it had plateaued. We also built in feature interactions, some of them linear and some of them non-linear, to help the model notice more trends. The feature interactions included BMI2 and Diet Quality * Exercise frequency. The feature interactions required some tuning because having too many feature interactions created too much noise for the model and brought the performance down.

I used the R2 metric as our main method of evaluation for an individual model's performance, and I also looked at MAE and RMSE to make sure that it wasn’t getting any certain range too wrong. Another metric that was used was the training R2 vs test R2 to see if the model was generalizing well and avoiding overfitting. Our baseline model was a linear regression model that performed a lot better than I expected it to. This told me that the data was a lot more linearly related than I had previously thought, which made me think about adding in non-linear feature interactions to help the MLP grasp onto the relationships going on in the data.

Results & Analysis:

The baseline model had an R2 score of .85, a MAE of 4.6, and a RMSE of 5.6, while our tuned MLP model had an R2 score of .86, a MAE of 4.3, and a RMSE of 5.36. The ensemble model outperformed these models in all categories with an R2 score of .87, an MAE of 4.2, and an RMSE of 5.18. Even though these results are somewhat marginal, the ensemble model does outperform the rest of the models. Before cleaning the data, the baseline model was a lot lower, usually down around .8 for R2, but after doing some more intense data cleaning, that score jumped way up, while the ensemble model barely increased.

Figure 1.

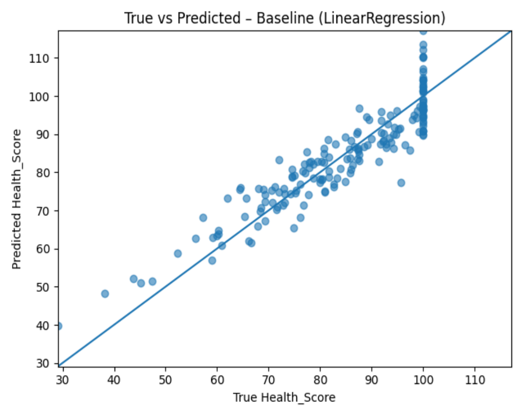

Figure 2.

Figure 1 shows the True vs. Predicted graph for our ensemble model, and Figure 2 shows the same graph for our baseline model. As you can see in both of the graphs, the points follow the center line and are distributed on either side of it. Both overpredict when it comes to the lower health scores, and this is most likely due to the fact that the data set that I used is very top-heavy, with a lot of high health scores included in the data. A difference that we can see in the graphs is that the baseline model overpredicts values a lot more at the high end of things as well, while the ensemble model has a more even distribution when making predictions on the higher health scores.

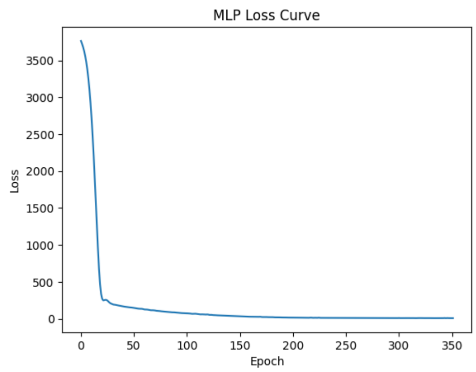

Figure 3.

Figure 3 shows our MLP Loss Curve for the tuned MLP model. It shows that it started out with a ton of loss which is to be expected at the beginning of learning but drops very steeply until about 20-30 epochs. It starts to shallow out there but you can still see that the loss is dropping noticeably until about epoch 200 and this is where it seems to taper off. This is a good graph because it shows that our MLP model has a low loss value which means that it is learning well and is a well trained model. We also compared the train R^2 vs test R^2 and they were only different by about 3 points so there was some mild overfitting happening but it was within a perfectly acceptable range. This means that our model was finding real relationships between the data that it was able to make accurate predictions off of.

Conclusion:

The main findings of my work is that there was a lot of linearity in the data and that this data might have been handled better with a different model. I got great results using the MLP + LightGBM ensemble model but I think that you could get similar results using a Random Forest model or something similar to that. Ensemble models are great when you reach a plateau with one model and you want to improve your model performance. Adding another model in can make the model more well rounded and able to recognize new relationships which is why the ensemble model had the best performance out of all the models we tried. I think that the project did fall short on the accuracy predictions since I was originally shooting for .9 R^2 score and only reached a .87. I think that having more data would have helped and this was a limitation since getting hands on healthcare data is rather difficult because of laws such as HIPAA. I enjoyed this project and I grew a lot in my ability to look at data and figure out an effective way to process it and do feature engineering to strengthen the models that I was working on. Additionally working with these different softwares such as scikit-learn and lightGBM was definitely a learning experience and grew my skillset. In the future I would like to expand this with more data that includes more features so I can find out how much other habits and attributes affect the health of an individual such as blood pressure and stress levels.

5. References

[1] J. K. Tan *et al.*, “Machine learning–based prediction for high health care utilizers by using a multi-institutional diabetes registry: Model Training and Evaluation,” *JMIR AI*, vol. 3, Oct. 2024. doi:10.2196/58463

[2] R. Nopour, “Machine learning models in enhancing prediction of health-related indices among older adults: A scoping review,” *Heliyon*, vol. 11, no. 12, Jul. 2025. doi:10.1016/j.heliyon.2025.e43510

[3] H. Habehh and S. Gohel, “Machine learning in Healthcare,” *Current Genomics*, vol. 22, no. 4, pp. 291–300, Dec. 2021. doi:10.2174/1389202922666210705124359