Heart Disease Prediction

A Feedforward Neural Network

Alexander Garcia

December 2024

Contents

1 Abstract

The goal of this project was to research how to build a feedforward neural network (FNN). To build a proper network, data analysis and preparation, model selection and evaluation, and feature importance and reduction, were all studied and their impacts noted throughout. Within model selection, it is also important to track the metrics, such as accuracy, precision, and recall to aid in determining which model performs the best. These processes are all necessary in understanding how an FNN operates and the impact each has on the final results.

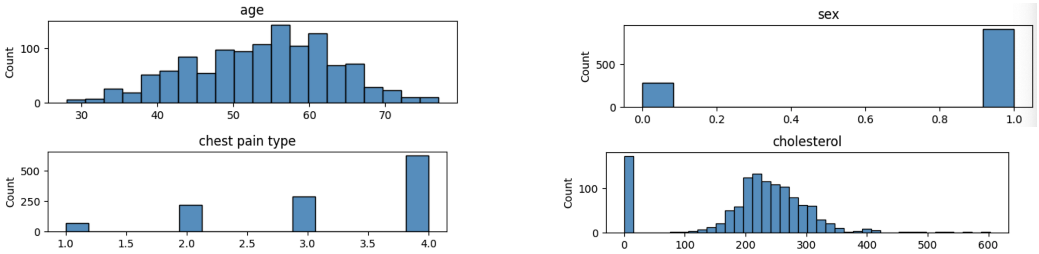
Why heart disease? Heart disease can have an impact directly, through the person dealing with heart disease themselves, or indirectly through family members or friends. Heart disease can lead to death or massive lifestyle changes with surgeries, pacemakers, and more. If we can predict, then we can react and take necessary precautions.

2 Data Analysis

The "Heart Disease" dataset was obtained from Kaggle [Mex] and is a combination of data from Cleveland, Hungary, Switzerland, and Long Beach. The data file contains 1190 samples with 11 features and 1 target variable. The target variable is binary, where 1 represents positive for heart disease and 0 is normal. A couple data points reference, "Angina", which is a type of chest pain caused by reduced blood flow to the heart[Sta]. Please note, descriptions of the features are attached on the last page.

2.1 Data distribution

The data is in a cleaned format in the sense that it is already in numerical representations, but without normalizing the values. In the charts below, the distributions show some imbalance of sex but a well rounded distribution of age. For brevity, a subset of the features' distributions are shown.



3 Model Selection and Evaluation

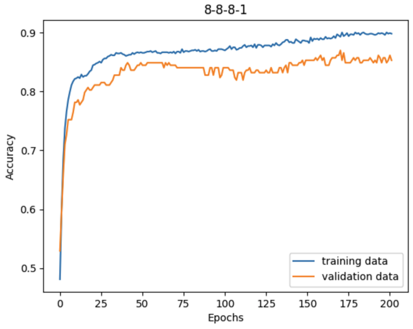
Listed in the table below are the various models and their metrics on the validation and testing datasets. The baseline row is the percentage of the majority class. Meaning, if there are 40 samples with a value of 1 and 60 with a value of 0 then baseline accuracy would be 60%.

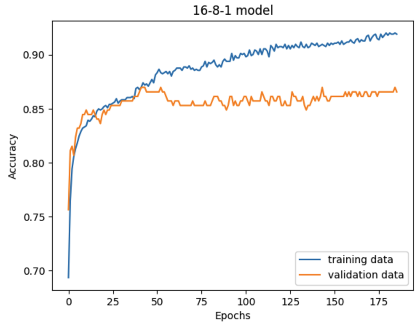
Model	**Acc. Train**	**Acc. Valid**	**Prec Valid**	**Recall Valid**
Baseline	51	56	N/A	N/A
Logistic Regression	83	84	88	82
Neural Network Model (16-8-1)	91	87	88	87
8-8-8-1	91	88	90	88
8-4-1	88	84	87	84
8-1	87	85	88	84
4-1	84	84	88	83
2-1	84	83	84	85

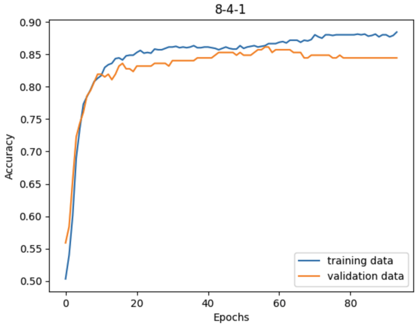
Table 1: Metrics for various models

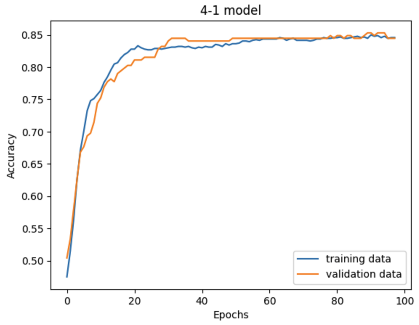
3.1 Learning curves

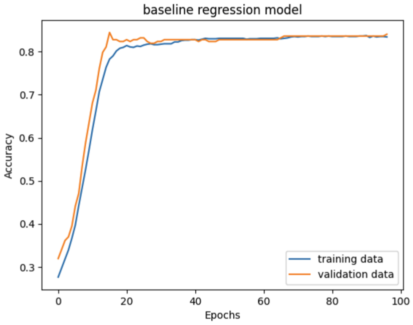
When fitting the models, model checkpointing was used based on validation loss and tracked the accuracy, precision, and recall metrics. An interesting find in the 3 and 4 layer models is that they do result in high validation accuracies, but clear signs of overfitting occur. The widening gap shown in the figure for the 16-8-1 model is a sign of overfitting. The 8-8-8-1 model appears to be heading toward overfitting as well while the cleanest learning curves come from the two layer networks and regression model.





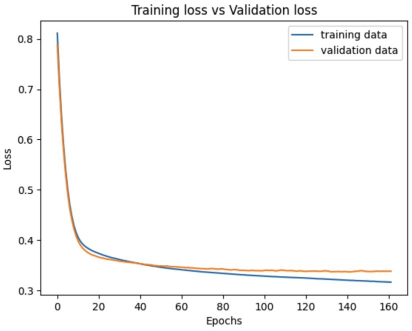




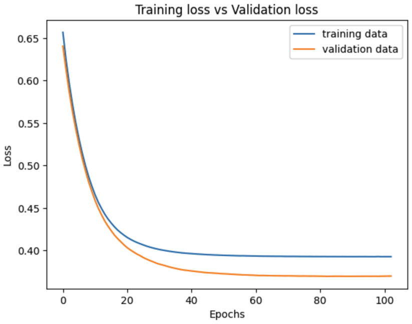


3.2 Training Loss vs Validation Loss

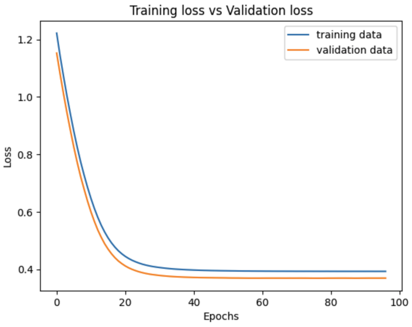
Here I compare the training loss versus the validation loss. An ideal graph would be such that the curve for both training and validation sets flows and decreases together without much separation. As a model trains, this loss becomes important for updating the weights and biases. The model's objective is to minimize this loss and it will alter the weights and biases in a direction that minimizes the loss. It does so through a process known as backpropagation.



(a) 8-1 loss curve



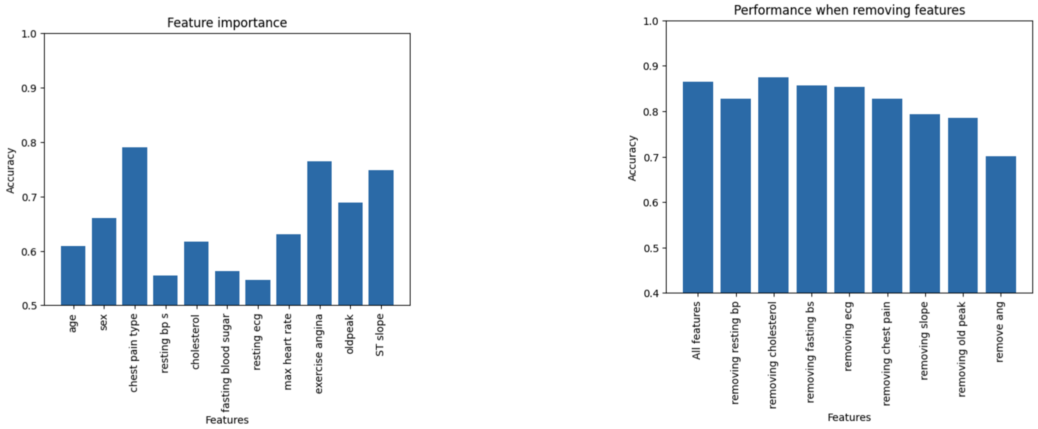
(b) 8-8-8-1 loss curve



(c) 16-8-1 loss curve

4 Feature Importance and Reduction

The goal of feature importance is to run a model for every input feature. This aids in understanding which features contribute most to the model. As seen in the leftmost figure, there are a few features which have little accuracy on their own suggesting they may not have as much impact as a feature like Chest Pain Type. However, when removing features through feature reduction the accuracy did not change as anticipated. I started by removing the lower features which resulted in no significant impact on final accuracy. When removing a higher feature such as Chest Pain Type there was a marginal decrease in accuracy. The biggest drop in accuracy resulted after removal of 3 of the most important features.

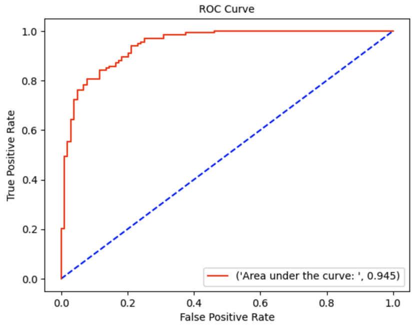


(a) Feature importance

(b) feature reduction

4.1 ROC curve evaluation

Receiver-operating characteristic curve (ROC) calculates the true positive rate and false positive rate by graphing true positives over false positives. An ideal model has a higher area under the curve[Goo].



(a) ROC for 8-8-8-1 model

5 Conclusion

In summary, I learned of the importance for each process of building a neural network. Properly normalizing data and having balance helps the models' predictions by reducing the accuracy a model can achieve by simply memorizing the patterns of the majority class. This can also reduce noise which can have a negative impact on its output. In model evaluation, the learning curves help alert of overfitting by the separation in their graphs. Overfitting results in high performance on training data but low performance on validation data. This is detrimental since a model would not perform up to expectations on unseen data. Through feature importance and reduction I was able to see the impact each feature has and the result of removing the feature from the model. Although the changes were not drastic enough in this project to merit removing a feature, there are cases this process would be more beneficial to the final model. Lastly, the ROC curve illustrates the chosen model did fairly well on its predictions of true and false positive rates.

References

[Goo] Google. *Classification: ROC and AUC.* URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=Receiver%2Doperating%20characteristic%20curve%20(ROC),-The%20ROC%20curve&text=The%20ROC%20curve%20is%20drawn,then%20graphing%20TPR%20over%20FPR..

[Mex] Mexwell. *Heart Disease Dataset.* URL: https://www.kaggle.com/datasets/mexwell/heart-disease-dataset?select=heart_statlog_cleveland_hungary_final.csv.

[Sta] Mayo Clinic Staff. *Angina.* URL: https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373.