Phase 3

Kel Gruber

November 26, 2023

1 Phase Goal

In this phase, I built several neural network models using the "Bank Marketing Data Set" from Kaggle Data Science that I cleaned in Phase 1. The goal is to determine which model produces the best prediction results so each model will be evaluated on its accuracy on the training dataset, the accuracy on the validation (testing) dataset, precision, recall, and f1 score.

2 Introduction

I decided to begin this phase by building the simplest model, the Logistic Regression model with one neuron first. To prevent overfitting, it was logical to begin this phase of the project by building smaller models first. I also tested various training schedules and total epochs with this first model and settled on training my models for 100 epochs. The early stopping flag was also turned on to prevent overfitting the model, and all models were set to stop training if no improvements were made to the val_loss metric after 20 epochs. Once I felt comfortable that I built the best logistic regression models I decided to try some different architectures to see how the size and shape of the model were related to the accuracy and other evaluation metrics.

In all, I built and tested 7 different neural network models. I also tried adding more training epochs to some of the models after the initial 100 epochs to see if I could improve the models with more training. Ultimately while additional training did slightly improve accuracy in these models, the other evaluation metrics fell and often the early stopping flag stopped the training early at 20-50 additional epochs so I did not feel that additional training made any considerable improvement to the models. Below are the results of these model experiments.

Model Type	Model Training Acc.	Validation Acc.	Precision	Recall	F1-Score
Logistic regression model	72.42%	71.64%	0.73	0.67	0.70
Neural network model (2-1)	73.39%	72.46%	0.77	0.63	0.69
Neural network model (4-1)	73.73%	72.05%	0.76	0.63	0.69
Neural network model (8-1)	73.40%	71.99%	0.77	0.62	0.68
Neural network model (8-4-1)	73.90%	72.17%	0.77	0.61	0.68
Neural network model (16-8-1)	73.90%	72.58%	0.78	0.62	0.69
Neural network model (32-16-8-1)	73.87%	72.40%	0.77	0.62	0.69

Table 1: Neural Network Models and Results.

3 Evaluation

Since none of my models had very high accuracy I decided to try training one of the lower accuracy models again to see if I could improve it with additional training. Unfortunately, this did not make a large difference in the overall accuracy of the model and the early stopping flag stopped the training early at 51 epochs indicating that the model did not improve for 20 of those additional epochs so it did not make sense to continue training these models. Many of the other models, such as the 32-16-8-1 model, had stopped early in the initial training session so it did make sense to attempt additional training for those models.

Epochs	Model Training Acc.	Validation Acc.	Precision	Recall	F1-Score
After 100 Epochs	72.42%	71.64%	0.73	0.67	0.70
Additional 51 epochs	72.54%	71.87%	0.74	0.66	0.70

Table 2: Logistic Regression Model Training Comparison

On initial review, all models appear to have similar results of about 71%-73% accuracy but when you look at the additional evaluation metrics such as precision, recall and f1 score you can see that some models are better than others. The Logistic Regression model performed better in some ways than the larger architectures. It might have the lowest accuracy overall but the Training and Validation Accuracy values are quite close together which shows that the model learned to make decent generalizations about the data during training so it performs similarly on the testing data.

The logistic regression model also has the highest recall and F1 scores of all the models, and the precision is close to the recall score as well. This indicates that the balance and completeness of the true positives predicted are good. Precision measures the ratio of correct positive predictions out of all positive predictions that were made. Recall is the measure of the total true positives divided by the total number of true positives and false negatives. The recall in our training data and testing data is quite different though, in the training data it is quite low indicating that there is a high number of false negatives. Typically the closer the higher precision gets the lower recall becomes which can be seen in some of the larger models that as the accuracy and precision creep up the recall and F1 score drop.

Finally, the f1 score combines precision and recall into one measure. This score is also best in the Logistic Regression model which is why I decided that this is the best model built overall because it is the most consistent in identifying the correct true positives. While the accuracy overall is the lowest I think in this problem identifying the people who are most likely to make term deposits is more useful to the bank marketing callers than correctly identifying everyone's final decision. In the other models, with the recall dropping along with F1-Score this means that the model is less likely to miss potential positives and gives the marketers a better list of customers they should contact about their term deposit accounts.

One important note on these results is that the other models are not technically worse than the Logistic Regression model, the precision values in those models are significantly better in those models but the drop in the recall and F1 scores is why I decided to not choose them. I do believe that with more time and training adjustments I could improve the accuracy and other evaluation metrics of these models. I do not believe that the logistic regression model is the absolute best model that could be produced for this problem, especially since it still has a rather low accuracy, precision, recall, and F1-score compared to other models built. However, I do believe that out of these 7 models this one is closest to being the most useful to the bank in determining which customers they should contact first in their next term deposit marketing campaign.

4 Challenges

Some of the challenges of this phase included determining how much training each model needed, the sizes and structure of the different networks to build, and deciding which model was best for solving the initial project problem. Please see the Phase 3 Jupyter Notebook for more details on the Phase 3 execution and results.

References

[Herbas, 2020] Herbas, J. (2020). Deep learning neural network: Complex vs. simple model.