1 Phase 3: Data Split

In Phase 3, the objective was to prepare the dataset for training and evaluation of the machine learning models by splitting it into distinct subsets. This phase ensures that model training is conducted on one portion of the data, while validation and testing are performed on separate, unseen subsets to avoid overfitting and provide an unbiased assessment of performance. Additionally, this phase involves examining class distributions and addressing imbalances that could affect model accuracy.

First, the loan data was randomly shuffled, then split into training, validation, and testing periods. For the training and testing split, the data was split 80 percent for training, and 20 percent for testing. Once this was finished, I then divided it into 64 percent for training, 16 percent for validation, and 20 percent for testing.

1.1 Data Normalization

When determining the balance of the data, I noticed that the data was heavily imbalanced. By taking the numbers of 0's and 1's, which dictate the loan status, this was confirmed.

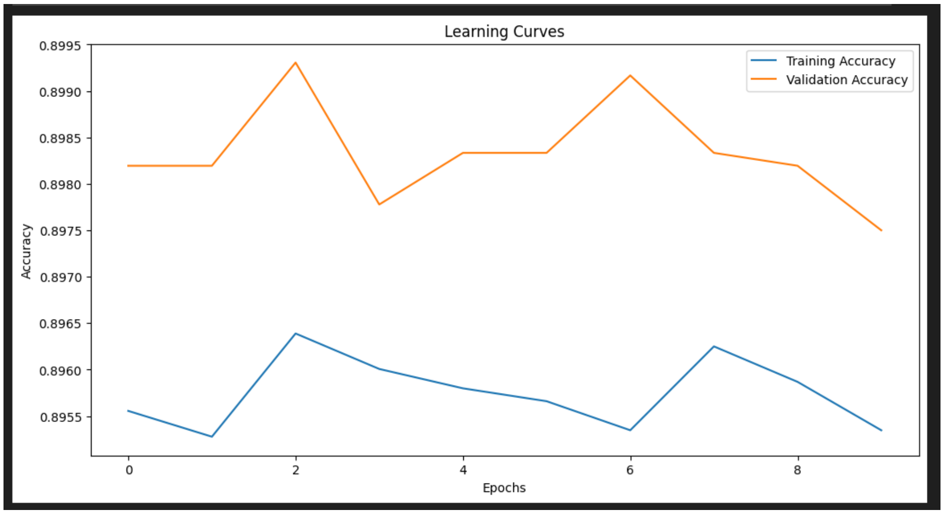
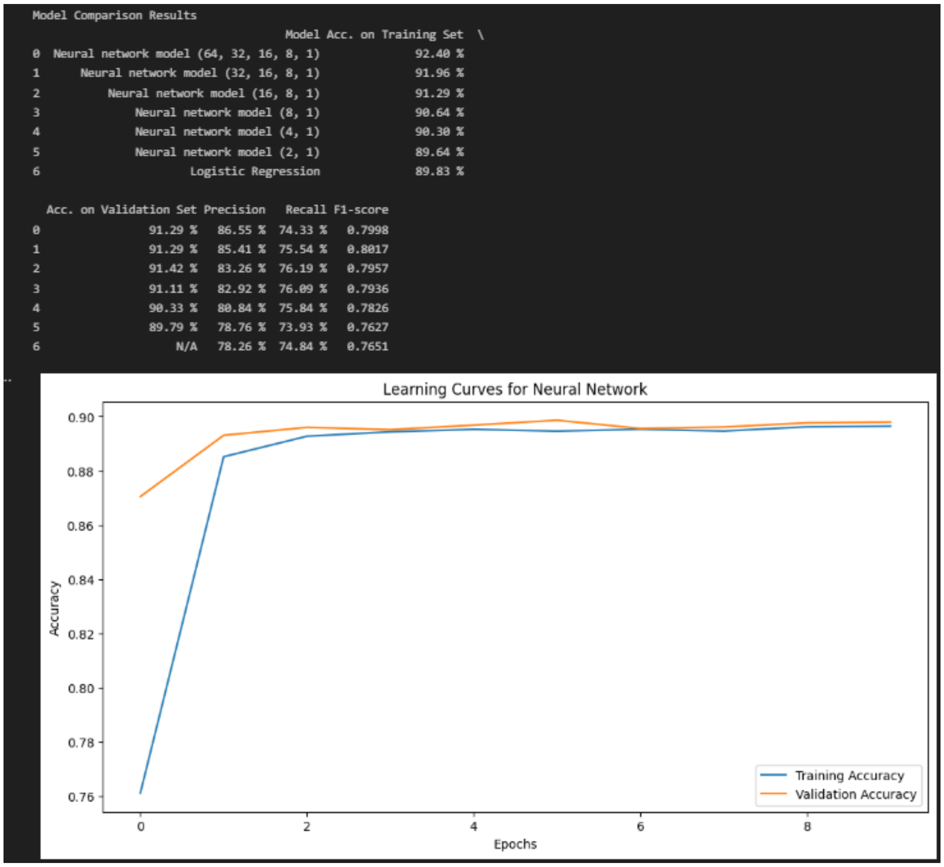


Figure 1: Distribution of approved vs. denied loans

From the chart above, we can observe that the dataset is noticeably imbalanced, with a significantly larger portion of loans being denied (class 0) compared to those approved (class 1).

To account for this imbalance during model evaluation, I will use performance metrics beyond accuracy, including precision, recall, and F1-score. These metrics provide a more complete view of how well the model handles both classes.

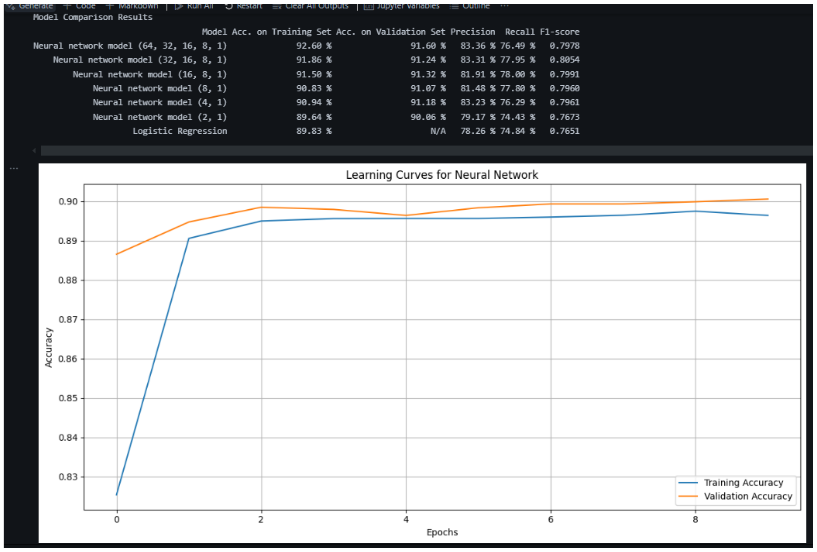
1.2 Model Comparison



```
Model Comparison Results
                                     Model Acc. on Training Set  \
0  Neural network model (64, 32, 16, 8, 1)                 92.40 %
1      Neural network model (32, 16, 8, 1)                 91.96 %
2          Neural network model (16, 8, 1)                 91.29 %
3              Neural network model (8, 1)                 90.64 %
4              Neural network model (4, 1)                 90.30 %
5              Neural network model (2, 1)                 89.64 %
6                      Logistic Regression                 89.83 %

  Acc. on Validation Set Precision   Recall F1-score
0                91.29 %   86.55 %  74.33 %   0.7998
1                91.29 %   85.41 %  75.54 %   0.8017
2                91.42 %   83.26 %  76.19 %   0.7957
3                91.11 %   82.92 %  76.09 %   0.7936
4                90.33 %   80.84 %  75.84 %   0.7826
5                89.79 %   78.76 %  73.93 %   0.7627
6                    N/A   78.26 %  74.84 %   0.7651
```

Figure 2: Model Comparison Table



```
Model Comparison Results
                                     Model Acc. on Training Set Acc. on Validation Set Precision  Recall F1-score
Neural network model (64, 32, 16, 8, 1)              92.60 %                  91.60 %   83.36 % 76.49 %   0.7978
    Neural network model (32, 16, 8, 1)              91.86 %                  91.24 %   83.31 % 77.95 %   0.8054
        Neural network model (16, 8, 1)              91.50 %                  91.32 %   81.91 % 78.00 %   0.7991
            Neural network model (8, 1)              90.83 %                  91.07 %   81.48 % 77.80 %   0.7960
            Neural network model (4, 1)              90.94 %                  91.18 %   83.23 % 76.29 %   0.7961
            Neural network model (2, 1)              89.64 %                  90.06 %   79.17 % 74.43 %   0.7673
                    Logistic Regression              89.83 %                      N/A   78.26 % 74.84 %   0.7651
```

Figure 3: Model Comparison Table 2

In this phase, multiple models were evaluated to determine which performed best on the loan classification task. The models included a logistic regression model and various neural network architectures with differing numbers of layers and neurons, ranging from shallow to deep networks.

The performance of these models was assessed using several key metrics: accuracy, which represents the percentage of correctly classified samples; precision, which measures the proportion of true positive predictions among all positive predictions; recall, which evaluates the proportion of actual positives that were correctly identified; and F1-score, which is the harmonic mean of precision and recall, providing a balance between the two.

The logistic regression model served as a baseline for comparison. It achieved moderate accuracy, precision, recall, and F1-score values. Although it was simple and quick to train, it lacked the capacity to capture complex patterns in the dataset, which limited its overall effectiveness. Nevertheless, it provided a useful benchmark against which the neural networks could be measured.

In contrast, the neural network models demonstrated varied performance depending on their architecture. Deeper networks generally achieved higher accuracy and F1-scores. Larger architectures, such as those with layers structured as (64, 32, 16, 8, 1), performed better on both training and validation sets. These networks were more capable of modeling complex relationships within the data. On the other hand, smaller architectures, such as (8, 1) or (4, 1), tended to underperform, likely due to underfitting, as they lacked sufficient capacity to model the data effectively.

The learning curves of the most complex neural network showed steady improvement in both training and validation accuracy over time. This suggests that the model not only learned from the training data effectively but also generalized well to unseen validation data, without clear signs of overfitting.

From all the models tested, the best-performing architecture appeared to be the deeper neural network with a structure like (64, 32, 16, 8, 1). It achieved the highest F1-score, indicating strong performance in both precision and recall, and therefore providing balanced and reliable predictions.

This analysis highlights the strengths and weaknesses of each model and provides practical guidance for selecting the most suitable one for the task at hand.

2 Conclusion

This phase accomplished the task of model evaluation by properly splitting the dataset and accounting for class imbalance. Using additional metrics such as precision, recall, and F1-score allowed for a deeper analysis of model performance, particularly on the underrepresented class. These insights will guide model tuning in the next phases and ensure that final performance evaluations are meaningful and relidelmar.umsl.edable.