Exercise 4 Report

Akash Shingha Bappy [2307938]

1. My Learnings:

In this exercise, I have learned about deep learning with a Multilayer Perceptron Classifier using Spark MLlib. First, a spark session was created along with importing all the necessary libraries. Then the data was loaded and preprocessed such as labeling and numerical conversion. Additionally, the data was split to training, validation and test data with a data size of 70%, 20% and 10% respectively. Then, A pipeline was constructed along with a classifier to fit with the training data. Lastly, the data was evaluated by computing matrics like weighted precision, recall, and accuracy for each set along with the confusion matrix which was plotted at the end. Overall this exercise covers the implementation of a deep learning classifier using PySpark Mlib that includes data preprocessing, model configuration, training, result evaluation and visualization.

2. Result:

From steps 6 and 7, the evaluation matrices(figure 1) and confusion matrix(figure 2) were obtained which gives an insight about the performance of the model. Here, the precision and recall were quite well as a result an accuracy of over 97% was achieved for all training, validation and test sets. Along with that, from the confusion matrix, it can be said that the model is much robust as the number of misclassifications was very low (FP=104, FN=50) compared to the true classifications (TN=3760, TP=3532).

```
Train weightedPrecision = 0.9779044413197855
Validation weightedPrecision = 0.9784763194570193
Test weightedPrecision = 0.9794300151810532
Train weightedRecall = 0.9777107847904308
Validation weightedRecall = 0.9783262682893477
Test weightedRecall = 0.9793177544990599
Train accuracy = 0.9777107847904308
Validation accuracy = 0.9783262682893478
Test accuracy = 0.9793177544990599
```

Figure: 1 Evaluation Matrices (Precision, Recall, Accuracy) For Train, Validation And Test Set.

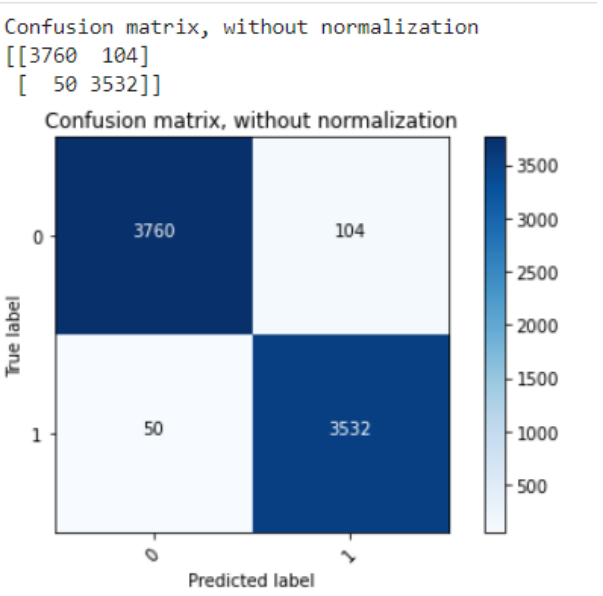

Figure 2: Confusion Matrix Comparing Predicted And True Labels

Reference:

1. BIg Data Processing and Application: Exercise 4.pdf
2. Evaluation Metrics - RDD-based API - Spark 3.5.1 Documentation (apache.org)