COM 6012

Assignment 1

Username – acr21mp
Name – Mahim Patil

Discuss at least four observations (e.g., anything interesting), with two to three sentences for each observation. If you need to, you can run additional experiments that help you to provide these observations.

1. All three models have accuracies close to each other. An interesting insight is that neural network also has the same accuracy as other models. Neural network is better than Random Forests and Logistic regression as it can learn more features because of the architecture. But in this case, all three have performed same. It may be because the neural network is a shallow neural network and the case may have been different if it was a deep network.
2. I tried casting the data as integer just to check what difference it would make but, the result was negligible difference in the accuracy. Although the accuracy of the Logistic Regression and Neural network increased by 4%, the Random forest model's accuracy decreased by 6%. Finally, I casted the data to double as it gave good accuracy for all the models.
3. The dataset has too many features, so performing dimensionality reduction or performing feature engineering could have increased the accuracy of the models.
4. The training time of neural network increases with the size of layers. This is same with random forest model where if we increase number of trees. The training time also decreases with increasing the number of cores which is obvious because the processing power increases.

References

1. https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/
2. https://spark.apache.org/docs/2.4.0/api/python/index.html