Course: *CMPS 620- Machine Learning* – Professor: *Dr. Osei Tweneboah*

Homework #6 Due on 05/04/2021

Instructions: While discussion with classmates are allowed and encouraged, please try to work on the homework independently and direct your questions to me.

We consider a human resource data set concerning employee retention from one Kaggle data analytics competition. The data set contains 14,999 observations and 10 variables. The binary target `left` indicates whether a employee left the company.

1. (Data Preparation) Bring in the data D into Python and name it as `hr`. Change the categorical variable salary in the data set to ordinal.

   Inspect if there is any missing values and, if so, handle them with imputation.

2. (Exploratory Data Analysis) Explore the data with EDA. If you search the keyword 'Human Resources Analytics + Kaggle' on Google, you may find many Python examples posted by other experts with different EDA and supervised learning methods. Please study their approach and feel free to reproduce some of the results in this project. Nevertheless, make sure that you understand what you are doing and interpret the results appropriately. Present at least THREE interesting findings.

3. (Data Partitioning) Randomly split the data D into the training set D_1 and the test set D_2 with a ratio of approximately 2:1 in sample size. Use $np.random.seed(42)$ to fix the random seed so that the results are easily reproducible.

   In the steps to follow, we will train several classifiers with D_1 and then apply each trained model on D_2 to predict whether whether an employee will quit his/her current position or its likelihood. For each approach, obtain the accuracy based on the prediction on D_2: Compare the performance of all these classifiers and summarize the results.

4. (DNN) Train two different deep neural network (DNN) (consider at least three hidden layers) models using epochs=150. Train on both your local machine and on AWS EC2. Report the time it takes to run on both platforms.

5. (Random Forest (RF)) Fit random forests as another baseline for comparison. Train on both your local machine and on AWS EC2. Also, obtain the variable importance ranking from RF. Report the time it takes to run on both platforms.

6. Evaluate the model on the test data.

Summarize the results and compare the above supervised learning approaches in terms of their pros and cons within this application context of employee retention.