COM6012 Assignment Part 1 Report

1. To work with a subset of the larger dataset, the dataset need to be sampled using by ‘randomSample()’. After useing pipelines and cross-validation to find the best configuration of parameters for each model, the results showed as below.

```
****Now start to train the 1% data*****
**********start training rf model***************
Accuracy for best rf model = 0.499768
{
    "bootstrap": true,
    "cacheNodeIds": false,
    "checkpointInterval": 10,
    "featureSubsetStrategy": "auto",
    "featuresCol": "features",
    "impurity": "gini",
    "labelCol": "labels",
    "leafCol": "",
    "maxBins": 2,
    "maxDepth": 5,
    "maxMemoryInMB": 256,
    "minInfoGain": 0.0,
    "minInstancesPerNode": 1,
    "minWeightFractionPerNode": 0.0,
    "numTrees": 20,
    "predictionCol": "prediction",
    "probabilityCol": "probability",
    "rawPredictionCol": "rawPrediction",
    "seed": 42,
    "subsamplingRate": 0.1
}
The running time of Random Forest for 1% dataset is: 279.54109048843384
```

```
Accuracy for best lr model = 0.500895
{
    "aggregationDepth": 2,
    "elasticNetParam": 0.0,
    "family": "auto",
    "featuresCol": "features",
    "fitIntercept": true,
    "labelCol": "labels",
    "maxBlockSizeInMB": 0.0,
    "maxIter": 100,
    "predictionCol": "prediction",
    "probabilityCol": "probability",
    "rawPredictionCol": "rawPrediction",
    "regParam": 0.01,
    "standardization": true,
    "threshold": 0.0,
    "tol": 1e-06
}
The running time of linear regression for 1% dataset is: 279.0384180545807
```

```
Accuracy of MLP = 0.495392
{
    "blockSize": 1,
    "featuresCol": "features",
    "labelCol": "labels",
    "maxIter": 200,
    "predictionCol": "prediction",
    "probabilityCol": "probability",
    "rawPredictionCol": "rawPrediction",
    "seed": 1500,
    "solver": "l-bfgs",
    "stepSize": 0.03,
    "tol": 1e-06,
    "layers": [
        128,
        32,
        2
    ]
}
The running time of MLP for 1% dataset is: 3309.8976109027863
The total using time: 3868.8994357585907
```

The reason of taking a long time is that the cross validation we set has three different values of three different parameters, including the layers containing [[128,32,32,32,2], [128,32,32,2], [128,32,2]]. And for saving time, we just set the best layers in this running time (it is not means the absolute best). Therefore, in the txt file I submit, it shows a less time to train the model.

2. Delivering the best parameter in the best model by '.bestModel' attribution and get the parameters from the '.get_()'. After fit the whole dataset and the transform the whole test data, the running time and the result can be seen as followed.

```
****Now start to train the whole data*****
The trained best maxDepth: 5
The trained best maxBins: 2
The trained best subsamplingRate: 0.1
Accuracy for rf model in whole dataset = 0.49968
AUC for rf model in whole dataset = 0.499833
The running time of Random Forest for 100% dataset is: 102.77491354942322
The trained best elasticNetParam: 0.0
The trained best regParam: 0.01
The trained best maxIter: 100
**********start training lr model on whole dataset***************
Accuracy for lr model on the whole dataset = 0.500793
AUC for lr model on the whole dataset = 0.500627
The running time of linear regression for 100% dataset is: 54.65736651420593
The trained best blockSize of mlp: 512
The trained best maxIter of mlp: 100
**********start training MLP model***************
The trained best blockSize of mlp: 1
The trained best maxIter of mlp: 200
Accuracy for mlp model on the whole dataset = 0.500793
AUC for mlp model on the whole dataset = 0.500627
The running time of MLP for 100% dataset is: 63.89572763442993
```

3. There are several observations from the data and the models above.

a. As for the dataset, this dataset does not in a linear relationship, which makes the accuracy of these three packaged models all equal to around 0.5. These results mean that these models just learn anything. Because of the XOR-PUFs physical model, each x of one response pair need to feed five PUFs model together and get five results which need to be calculate by XOR to get the y of this pair. It means that we need to repeat each input data x five times making it like a 5*128 matrix. What's more, the same structure of W is needed. The 5 results of x*w, like 1 or -1, can be multiplied and then get the y of our response pair.
b. As for the models, these three models need different optimizers during working on this dataset. As for random forest, it just needs a correct structure of dataset, such as the 5*128 matrix we mentioned above. As for the linear regression, it needs the CMAES method to optimize the weight during training model. Otherwise, we can just calculate the tensor product of input data and the size of data becomes power(128,5). As for MLP, it needs to three hidden layers of 32 and the layers of model need set as [128,32,32,32,2].
c. Comparing these three models' running time, the former two, that is random forest and linear regression is quicker than MLP. The reason of this may be there are more weight parameters needed. However, the time cost can exchange the higher perform when using the MLP under the correct input dataset.