Lebanese American University

Computer Science and Math Department, Byblos

CCS 615 Machine Learning

Assignment #3

Joy Chahine

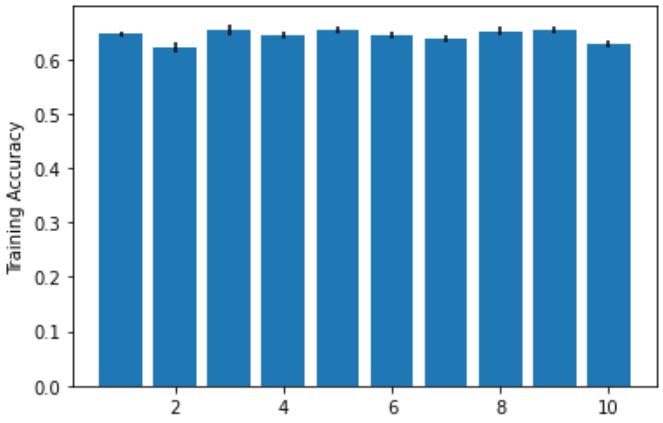
---

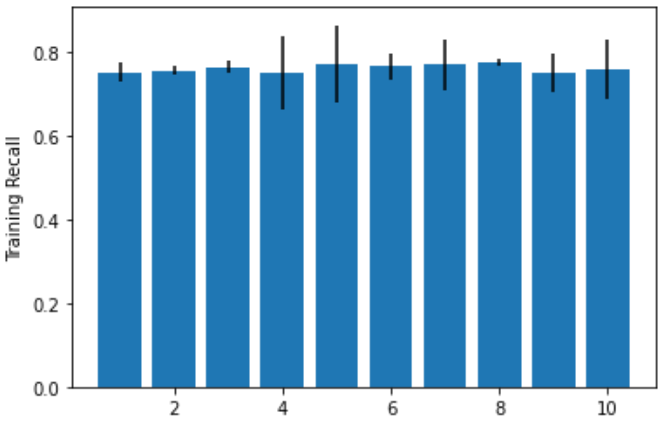
4. Regression

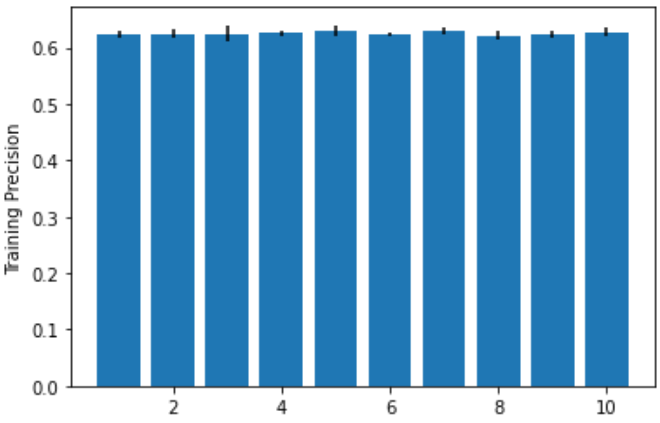
The below table represents the averages of 30 runs of the testing scores of all the experiments done. As seen DT trees still have the highest scores on all measures. In summary overall DT still beats ANN and SVM and LR. As mentioned the aspect of the data representing a HTTP Packet is very easy to interpret by a set of rules and very hard to fit and split the data in a dimensional space. The data is balanced 50 – 50 which means there is no overfitting occurring in the ANN and LR.

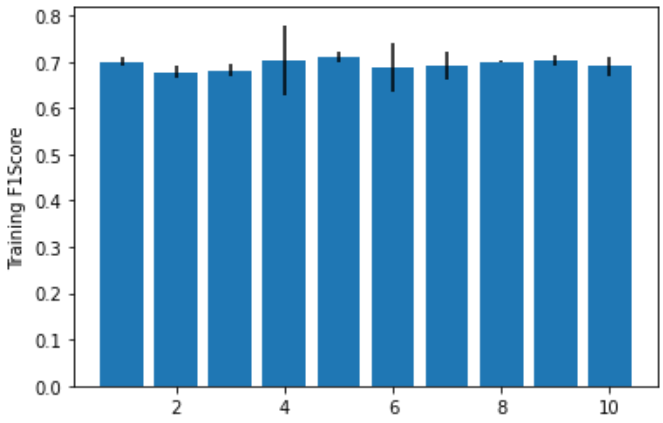
Alg	Accuracy	F1 Score	Precision	Recall	Specification	Sensitivity
DT20	0.95	0.95	0.96	0.94	0.93	0.91
DT40	0.95	0.95	0.96	0.94	0.91	0.93
DT60	0.96	0.95	0.96	0.94	0.91	0.94
DT80	0.95	0.95	0.96	0.94	0.91	0.94
JoyANN	0.62	0.65	0.61	0.80	0.96	0.24
ANN	0.60	0.60	0.64	0.60	0.81	0.45
SVM	0.51	0.70	0.53	0.99	0.60	0.47
LR	0.63	0.65	0.59	0.76	0.75	0.47

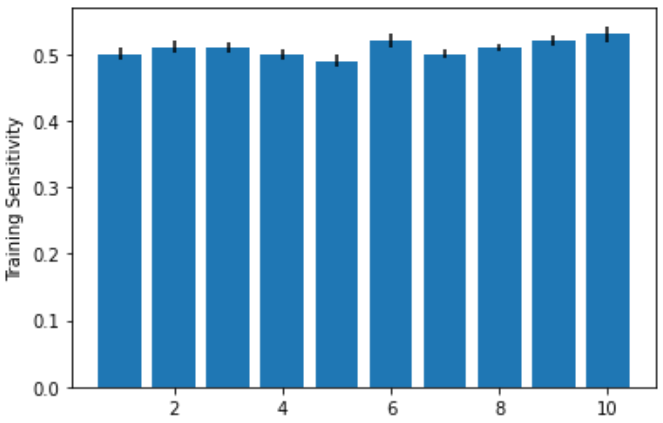
For the Logistic Regression below are plotted the training averages scores with the Standard deviation.

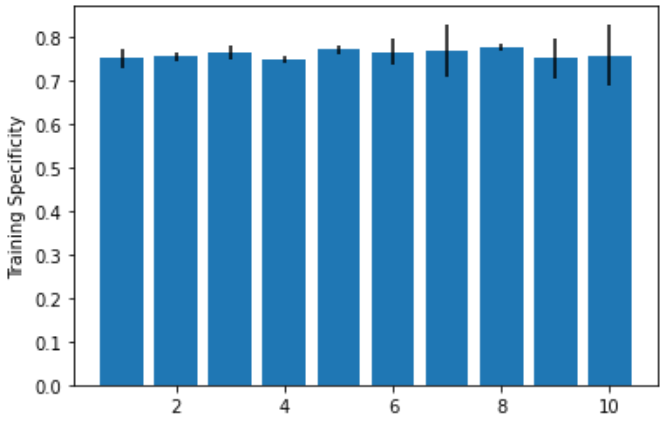












The LR was tunned and trained with a 10-Fold cross validation for 30 runs. The data was split into training, testing, and validation sets. The testing and training scores are very close as seen by the results. Since both SVM and LR gave low scores, this means that the data as is with its features is not represented well and inn other word is very hard to separate in any dimensional space. Knowing that both models were already tunned. Also, the reason behind the low accuracy in ANN could be the same. The dense values of the features need more layers to be processed and learned correctly. One thing that can be done here is feature reduction for sure as well as some data preprocessing.