**DNS Question 2 Part 1**

Our work and results for part 1 is Part1.ipynb.

It includes:

1. Data cleaning
2. Base model, modified MIT model, End Game model, Invincea model, NYU model,Xgboost stack model
3. Benign domain name filter
4. Metric: precision, recall, score 1, ROC

**DNS Question 2 Part 2**

**Unsupervised Learning**

Our work for part 2 feature extraction is in “Extract Entropy, MCR, and ngramscore.ipynb” and “IP\_Matrix\_Creation.ipynb”. These notebooks have the feature extraction process

Our results for part 2 are in Results Part 2.ipynb

**Supervised Learning**

Our work for part 2 Supervised learning approach is in Part2\_Supervised\_learning\_with\_TLDs.ipynb and Part2\_Supervised\_learning\_without\_TLDs.ipynb. These results show grouping DGAs by only their string without TLD still gets an 80% accuracy on test data when using supervised learning. This means it is possible to use only string features to get this high accuracy.