Data analysis of a cybersecurity dataset from lanl

1. Downloaded 7GB file auth.txt.gz
2. Created small subsample of the data using bash, 'sample2v.csv'
3. Loaded this subsample into jupyter notebook and started exploratory analysis
    a. Analysis is in exploration.ipynb
    b. Fails constitute only 1% of all data.
    c. Machine learning on such data can have 99% accuracy if the algorithm always predicts a "success". Need to resample the data from auth.txt.gz to generate files that contain equal number of successes and fails.
    d. All data is categorical
    e. Auth.txt.gz contains 9 columns
    f. Columns 5-7 contain on the order of 10 different labels. Labels don't seem to have hierarchical structure, so I will expand the columns so that every new label becomes a column (feaure) with entries 1/0 (True/False) depending on whether the label applies.

Data analysis of a cybersecurity dataset from lanl

1. Further feature engineering
   a. Columns 1-4 contain on the order of 40,000 different labels. I do not want to convert every label into a feature as the number of features will explode making machine learning hard.
   b. I expand columns 1-2 into 4 columns that separately keep track of source user, source domain, destination user, and destination domain. All users that start with C-labels and U-labels are converted to just ‘C’ or ‘U’ labels to reduce the number of labels in this column.
   c. I also do a number of comparisons between data in columns 1-4 and it’s derivatives to see when labels are the same/different. I used the resulting data as features for machine learning.
2. Machine learning
   a. Tried logistic regression, logistic regression with L1 penalty, logistic regression with L2 penalty, Gradient Boosting
   b. Logistic regression with L1 penalty worked best
   c. Used logistic regression with L1 penalty on 15 files with 400,000 data points, equal sampling of successes and fails. Files are non-overlapping sampling of auth.txt.gz
   d. Mean accuracy score is 0.94 and standard deviation is 0.0004.