Network Intrusion Detection

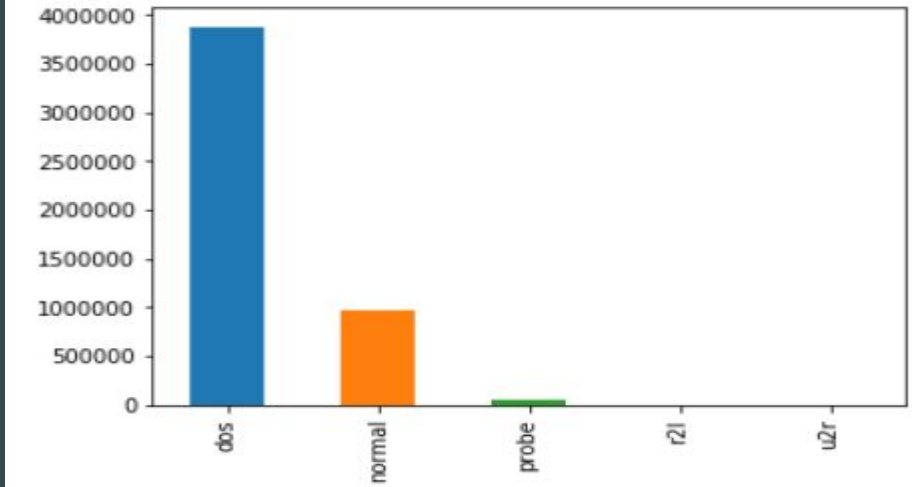
CMPE 255 - Spring 2019

Team: Yang Chen, Fulbert Jong, David Tang

Introduction/Data Sets

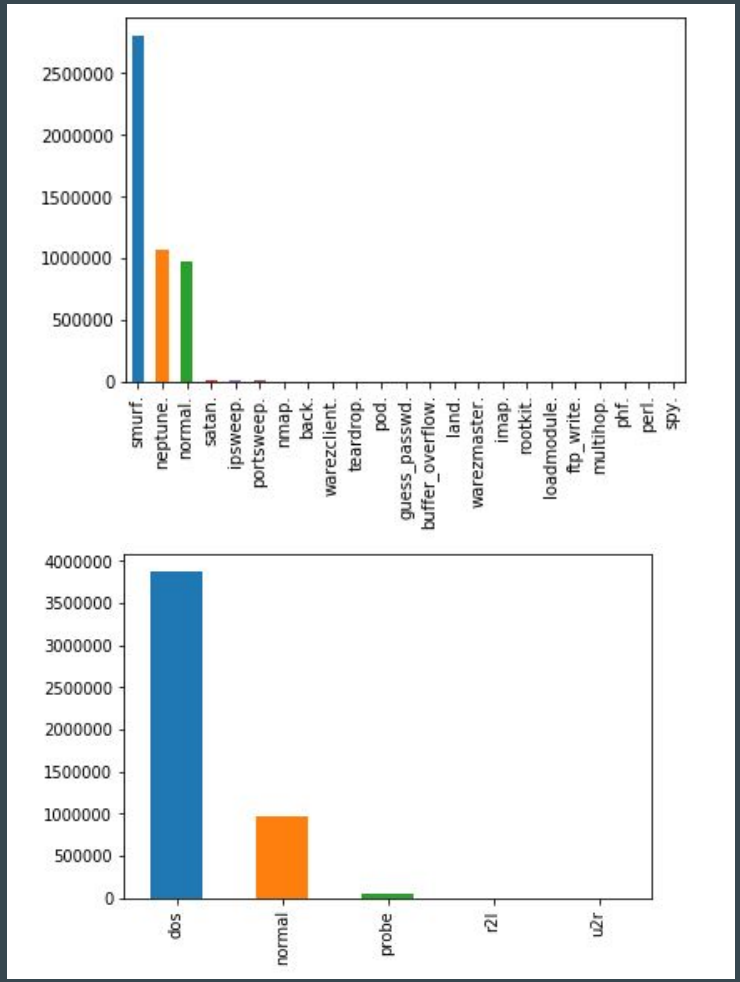
- Datasets retrieved from KDD 199 challenge
  - Almost 5 millions rows
  - 40 features
  - Label (originally 20 -> 5)
- TCP Dump Data
- Heavily imbalanced
  - 52 u2r attack out of almost 5 millions connections
- F-1 Score is used as a metric because of imbalanced dataset

```
dos        3883370
normal      972781
probe        41102
r2l           1126
u2r             52
Name: 41, dtype: int64
```



Classifications

- Normal
- DOS (denial of service)
  - back, land, pod, teardrop, smurf
- R2L (remote to user)
  - ftp_write, guess_passwd, imap, multihop
- U2R (user to root)
  - Buffer_overflow, loadmodule, perl, rootkit
- Probe
  - ipsweep, nmap, portsweep, satan



Data preprocessing

- Encode strings like (‘tcp’, ‘udp’, ‘icmp’) into numbers
  - (0, 1, 2)
- Double encode strings like ‘smurf’ => 5 => 1 (DOS)
- Remove features where the values never change

Tools & Libraries

- Numpy
- Scipy
- Matplotlib
- Keras (with Tensorflow)
- Sklearn
- Docker

Training models

- Naive Bayes
- Decision Trees
- Artificial Neural Networks

```
yang@yang-Latitude-E6420:~/Documents/github/pyPJ/kdd99ml$ python randomForestCla
ssifier.py
Loading raw data
Transforming data
X_train, y_train: (395216, 41) (395216,)
X_test, y_test: (98805, 41) (98805,)
Training model
Score:  0.9999949394761346
Predicting
Computing performance metrics
Confusion matrix:
 [[19589     0     0     0     0     0     0     0     0     0     1     0
      0     0     1     0     0     0     0     0     0]
 [    0     6     0     0     0     0     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    1     0     1     0     0     0     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0 21467     0     0     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0     0 56000     0     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0     0     0     7     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
```

5]:
```
## Bernoulli Naive Bayes Classifier
clf_BernoulliNB = BernoulliNB(alpha=0.01, class_prior=None, fit_prior=True)
clf_BernoulliNB = test_classifier(clf_BernoulliNB)
```

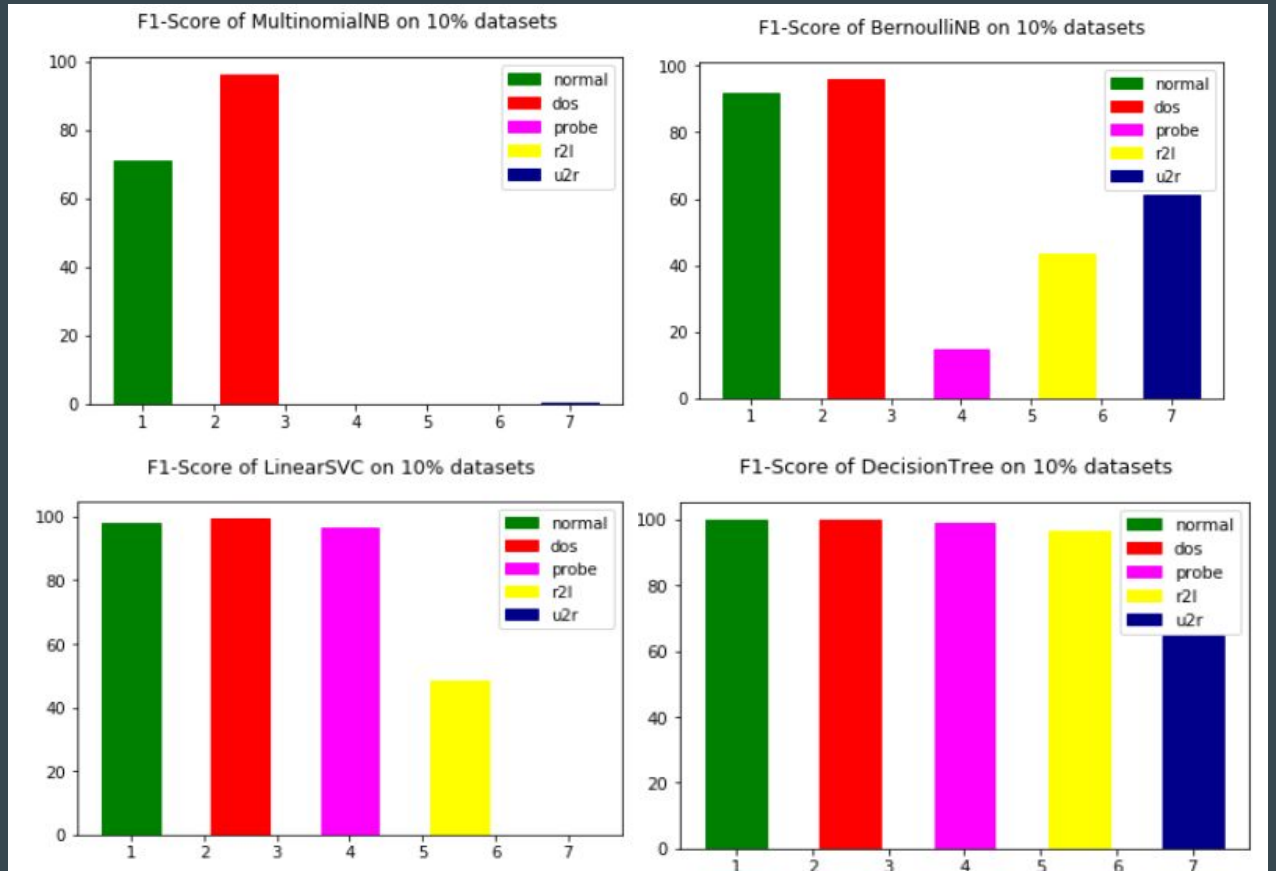
	precision	recall	f1-score	support
0	0.98	0.89	0.94	320433
1	0.99	0.94	0.96	1282047
2	0.09	0.35	0.15	374
3	0.08	0.54	0.14	13
4	0.08	0.61	0.15	13616
avg / total	0.98	0.93	0.95	1616483

```
Accuracy Score: 0.9297926424218504
Classifier Training time = 8.84793472290039
Classifier Prediction time = 1.9096009731292725
```

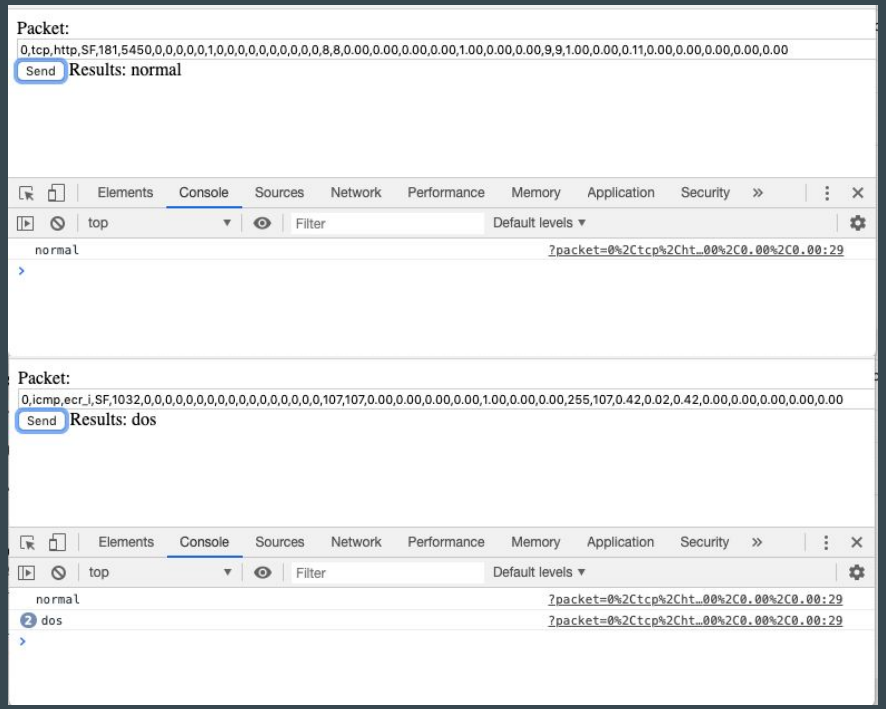
	precision	recall	f1-score	support
0	0.92	0.96	0.94	23686
1	0.99	0.96	0.98	67053
2	0.22	0.98	0.36	715
3	0.00	0.00	0.00	1843
4	0.00	0.00	0.00	12
micro avg	0.94	0.94	0.94	93309
macro avg	0.43	0.58	0.46	93309
weighted avg	0.95	0.94	0.94	93309

accuracy 0.9447105852597284

Algorithm Comparison



Application



Challenges

- Getting imbalanced data to classify
  - A specific type of attack happened around 60 times total in the training set
- Getting Keras to predict to a class
  - Default predictions are 0 or 1

Questions?

...