Binary Classification of Internet Traffic

Identifying Distributed Denial of Service Attacks

Chris Soyars

Faculty Advisor: Professor Michael Frankel

‡ Distributed Denial of Service (DDOS) cant variables.

Traffic patterns, attack methods, and security practices can evolve rapidly.
- ‡ Model likely to benefit from periodic re-evaluation.
- ‡ Unknown success rate against novel attack vectors of the same general class.

‡ Limitations & Improvements
- ‡ No analysis of trends by time or location (IP address)
- ‡ Lack of precision on some variables
- ‡ Possible errors from traffic logging software used

METHODS

- ‡ Examine data set to gather comprehensive information on variable properties. Eliminate variables that cannot be used due to poor or lacking information. Identify and impute missing or erroneous data if possible.
  - ‡ ~50000 observations had values NA, infinity, or implausibly negative for certain variables due to lack of precision and accuracy in time-related information. Values were recalculated to restore information.
- ‡ Cluster variables to identify variables of greatest interest.
- ‡ Discretize variables and calculate odds of benign vs ddos traffic for each bin to extract additional information and trends from selected variables. Natural log of odds also calculated for each bin.
- ‡ Eliminate non -significant variables from selection for logistic model.
- ‡ Create logistic regression model using selected variables. Model is trained on 80% of the data set.
- ‡ Refine logistic regression model to eliminate redundancy and select for highly significant variables.

RESULTS

‡ Logistic regression model with 17 selected variables
[redacted]9% concordant pair rate (C=0.999). Model
[redacted]s a lower probability of possible DDOS traffic for
[redacted] cases than actual DDOS attacks in most
[redacted]ations.
[redacted]al KS statistic of .954
[redacted] can be simplified depending on needs and
[redacted]es. C=.995 with as few as 4 variables.
[redacted]l measurements only: C=.965 with 3 variables,
[redacted] marginal gains for additional initial statistics.
[redacted]vity and specificity are maximized with probability
[redacted]ld of 0.188.
[redacted]ariable model correctly identifies 98.7% of benign
[redacted]c and 98.2% of DDOS traffic.

Table 2. Confusion Matrix for Logistic Regression Model with 17 Variables

Label	Prediction		
Frequency Percent	Benign	DDOS	Total
Benign	1248228	16168	1264396
	81.94	1.06	83.00
DDOS	4410	254495	258905
	0.29	16.71	17.00
Total	1252638	270663	1523301
	82.23	17.77	100.00

SAS CODE

```
%IMPV5 (DSN=class.test, VARS=&varlist, EXCLUDE=Label,
PCTREM=1,MSTD=);

PROC SQL;
        SELECT NAME INTO: VARNAME SEPARATED BY ' '
        FROM DICTIONARY.COLUMNS
        WHERE UPCASE(LIBNAME)="DDOS" AND
UPCASE(MEMNAME)="DDOS" AND NAME NOT IN("Label");
QUIT;
PROC VARCLUS DATA=import OUTTREE=tree MAXCLUSTERS=71;
        VAR &varname;
RUN;

PROC GLMSELECT DATA=ddos.disc2;
        MODEL label=&mvar / DETAILS=all SELECTION=lasso
STATS=all;
RUN;

PROC LOGISTIC DATA=train DESC OUTEST=betas
        OUTMODEL=scoringdata;
        MODEL label=&mvarsn /SELECTION=BACKWARD
        CTABLE pprob=(0.16 to 0.21 by 0.001)
        LACKFIT RISKLIMITS;
        OUTPUT OUT=output p=predicted;
        SCORE DATA=valid OUT=ddos.score;
RUN;
```