Progress Report I

JAROENSIRIPONG, Punyawat

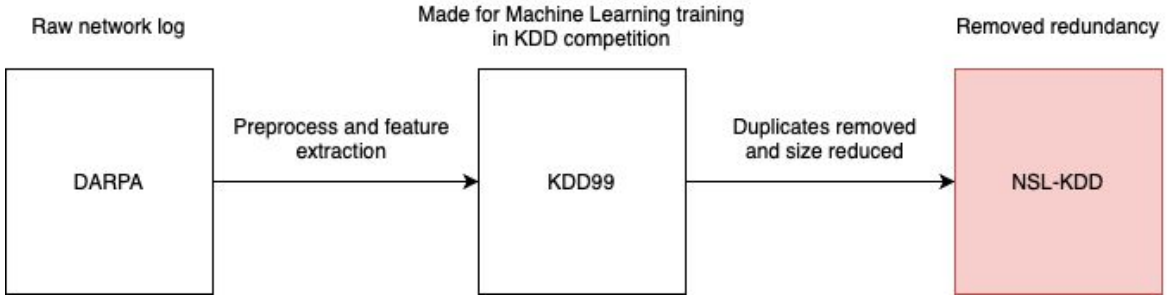
Introduction

- Data Analysis
  - NSL-KDD is the dataset that improve from KDD99 by reduce redundancy.
- Data Preprocessing
  - Change String data in to integer data using One-Hot-Encoder
  - Scaling data using RobustScaler
- Feature Importance
  - feature_importance_
  - L1 Regularization (Lasso)
  - Correlation Analysis
  - Permutation Importance
- Future works

Data Analysis

NSL - KDD dataset

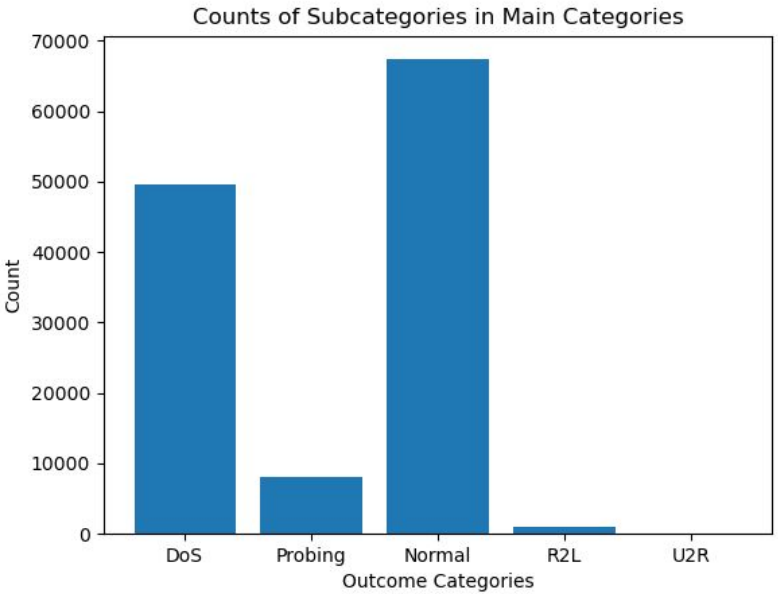
- Originated from DARPA and KDD Cup 1999 Data which used for 3rd International Knowledge Discovery and Data Mining Tools Competition.
- The dataset was prepared by MIT Lincoln Labs to simulate the cyberattack in U.S. Air Force LAN (Local Area Network)
- The data split in to 2 set :
  - KDDTrain+.txt : 125972 sample ; Use to train the model.
  - KDDTest+.txt : 22543 sample ; Use to test the model.
- The dataset have 38 features.



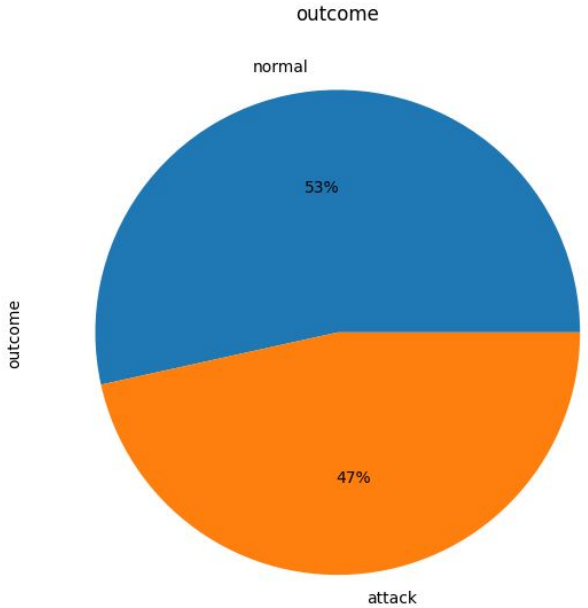
NSL - KDD dataset

- The dataset result contain 5 major categories:
  - probing is the initial state of attacks by the attacker to gathers information from victim.
  - DoS (Denial-Of-Service) is the attacks accomplish this by flooding the target with traffic.
  - Remote to Local (R2L) is the unauthorized access from a remote machine.
  - User to Root (U2R) is the unauthorized access to local superuser (root) privileges
  - Normal is the normal flow that did not contain any attack.
- 4 major attack can split in to 23 attack types.
- For now, the goal is to train the model in binary classification.

NSL - KDD dataset



Count of outcome categories



scale of outcome

Data Preprocessing

Data Preprocessing

- Change 23 types of attack into normal or attack (0, 1)
- There’re some columns that contain string data, thus these data must be changed into integer
  - ‘protocol_type’ (3 types), ‘service’ (70 values) and ‘flag’ (11 values)
  - Using One-Hot-Encoder
  - The features will increase from 38 features to 122 features
- Scaling the data by using RobustScaler()

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

Data Preprocessing (Dataset that preprocessed)

	land	logged_in	is_host_login	is_guest_login	duration	src_bytes	dst_bytes	wrong_fragment	urgent	hot	...	flag_REJ	flag_RSTO	flag_RSTOS0	flag_RSTR
0	0	0	0	0	0.0	0.369565	0.000000	0.0	0.0	0.0	...	0	0	0	0
1	0	0	0	0	0.0	-0.159420	0.000000	0.0	0.0	0.0	...	0	0	0	0
2	0	1	0	0	0.0	0.681159	15.800388	0.0	0.0	0.0	...	0	0	0	0
3	0	1	0	0	0.0	0.561594	0.813953	0.0	0.0	0.0	...	0	0	0	0
4	0	0	0	0	0.0	-0.159420	0.000000	0.0	0.0	0.0	...	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
125967	0	0	0	0	0.0	-0.159420	0.000000	0.0	0.0	0.0	...	0	0	0	0
125968	0	0	0	0	8.0	0.221014	0.281008	0.0	0.0	0.0	...	0	0	0	0
125969	0	1	0	0	0.0	7.923913	0.744186	0.0	0.0	0.0	...	0	0	0	0
125970	0	0	0	0	0.0	-0.159420	0.000000	0.0	0.0	0.0	...	0	0	0	0
125971	0	1	0	0	0.0	0.387681	0.000000	0.0	0.0	0.0	...	0	0	0	0

125972 rows × 122 columns

Feature Importance

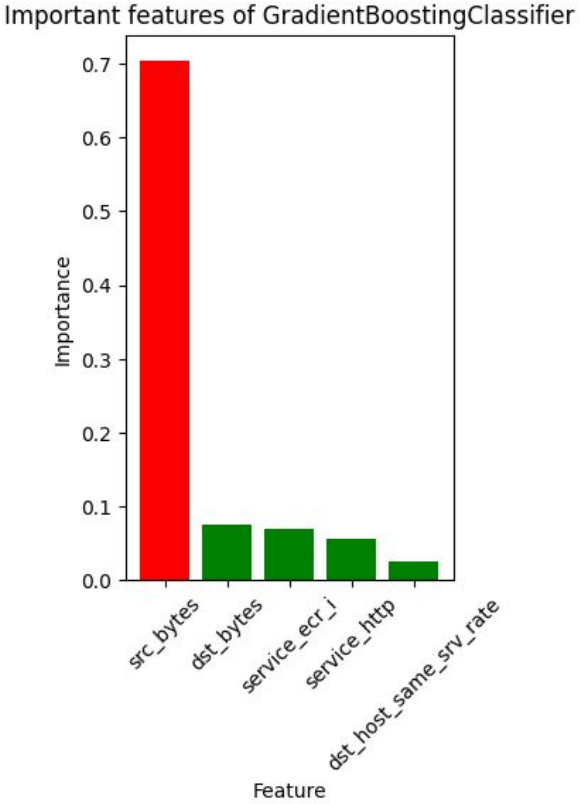
Feature Importance

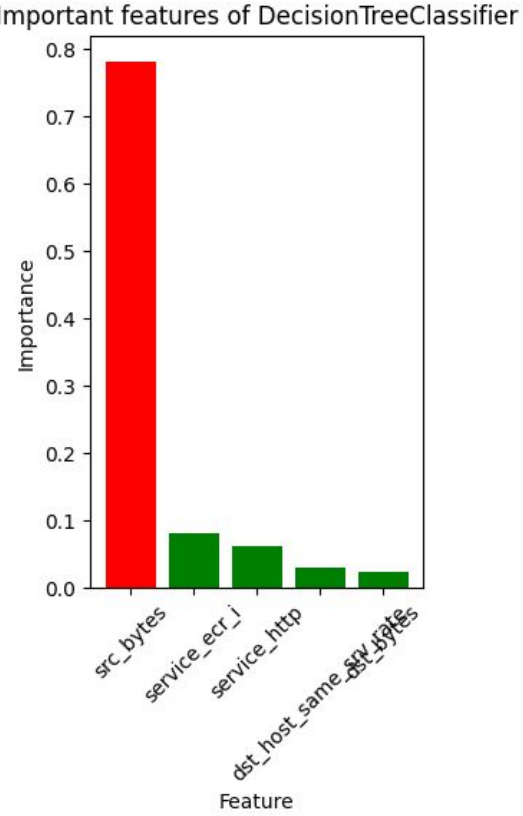
- Finding the features that have massive impact in the dataset to predict outcome
  - feature_importances_ form scikit-learn library.
  - L1 Regularization (Lasso)
  - Correlation Analysis
  - Permutation Importance

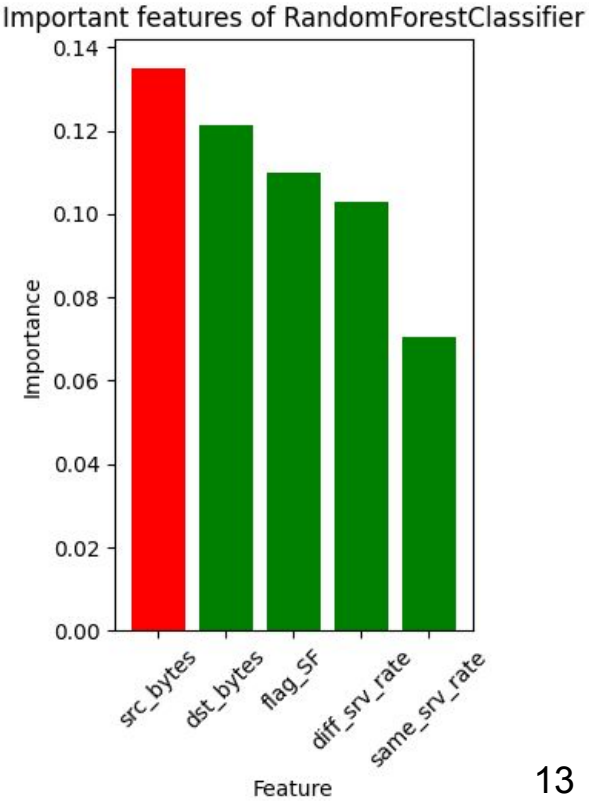
feature_importances_

- The build-in feature from scikit-learn trained model that computed as the mean and standard deviation of accumulation of the impurity decrease within model
- Comparison by training 3 models:
  - Decision Tree Classifier
  - Gradient Boosting Classifier
  - Random Forest Classifier

feature_importances_



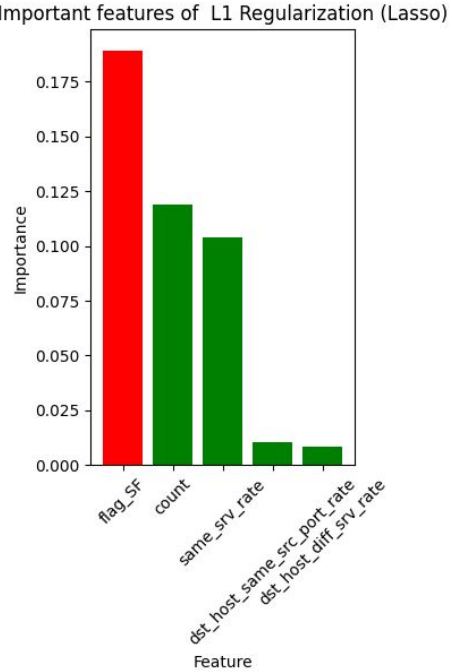




L1 Regularization (Lasso)

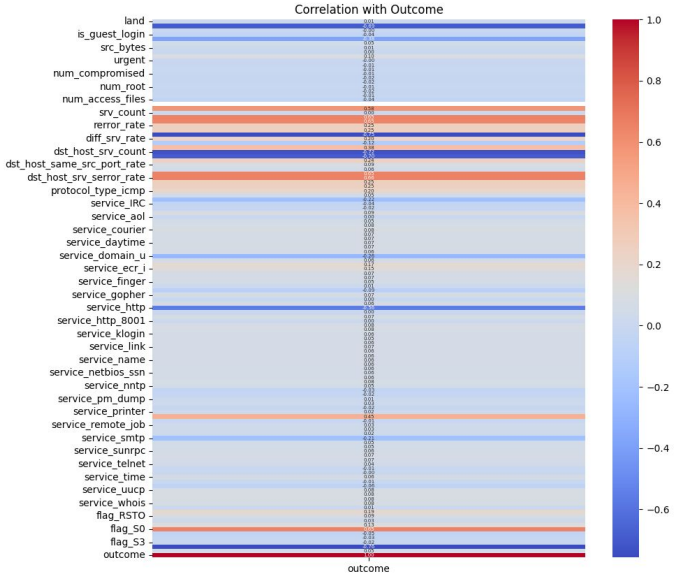
- Find feature importance by sorting the coefficient of Lasso

	Feature	Importance
0	flag_SF	0.189268
1	count	0.118863
2	same_srv_rate	0.104020
3	dst_host_same_src_port_rate	0.010237
4	dst_host_diff_srv_rate	0.008419
...	...	...
117	service_other	0.000000
118	service_pm_dump	0.000000
119	service_pop_2	0.000000
120	service_pop_3	0.000000
121	land	0.000000



Correlation Analysis

- Finding feature importance by sorting correlation between outcome and other features.



	outcome
flag_SF	0.757102
same_srv_rate	0.752253
dst_host_srv_count	0.723395
dst_host_same_srv_rate	0.695029
logged_in	0.691363
...	...
service_http_2784	0.003382
is_host_login	0.002934
srv_count	0.001190
service_tim_i	0.000912
num_outbound_cmds	NaN

123 rows × 1 columns

Permutation Importance

- Finding feature importance by randomly permuted or shuffled the values of a specific feature of the trained model.
  - Random Forest Classifier
  - Gradient Boosting Classifier
  - Decision Tree Classifier

Weight	Feature
0.0205 ± 0.0013	dst_bytes
0.0180 ± 0.0007	src_bytes
0.0126 ± 0.0004	protocol_type_icmp
0.0125 ± 0.0002	service_ecr_i
0.0118 ± 0.0008	count
0.0098 ± 0.0002	dst_host_same_src_port_rate
0.0058 ± 0.0003	logged_in
0.0046 ± 0.0004	srv_count
0.0039 ± 0.0002	protocol_type_tcp
0.0036 ± 0.0003	diff_srv_rate
0.0034 ± 0.0002	service_http
0.0017 ± 0.0005	service_private
0.0017 ± 0.0005	dst_host_count
0.0014 ± 0.0004	dst_host_srv_diff_host_rate
0.0011 ± 0.0005	dst_host_same_srv_rate
0.0010 ± 0.0009	dst_host_srv_count
0.0009 ± 0.0003	dst_host_diff_srv_rate
0.0006 ± 0.0002	dst_host_rerror_rate
0.0005 ± 0.0003	dst_host_srv_serror_rate
0.0004 ± 0.0000	wrong_fragment
… 102 more …	

Random Forest Classifier

Weight	Feature
0.1618 ± 0.0001	src_bytes
0.0165 ± 0.0011	dst_bytes
0.0108 ± 0.0004	hot
0.0079 ± 0.0002	service_ecr_i
0.0056 ± 0.0007	service_http
0.0053 ± 0.0003	service_private
0.0046 ± 0.0004	dst_host_same_src_port_rate
0.0030 ± 0.0001	service_ftp_data
0.0018 ± 0.0005	count
0.0009 ± 0.0005	dst_host_srv_count
0.0009 ± 0.0003	dst_host_same_srv_rate
0.0006 ± 0.0000	dst_host_srv_diff_host_rate
0.0005 ± 0.0001	dst_host_diff_srv_rate
0.0005 ± 0.0001	num_failed_logins
0.0004 ± 0.0003	dst_host_serror_rate
0.0003 ± 0.0001	duration
0.0002 ± 0.0001	num_compromised
0.0002 ± 0.0000	flag_RSTR
0.0002 ± 0.0001	dst_host_srv_serror_rate
0.0002 ± 0.0000	flag_S1
… 102 more …	

Gradient Boosting Classifier

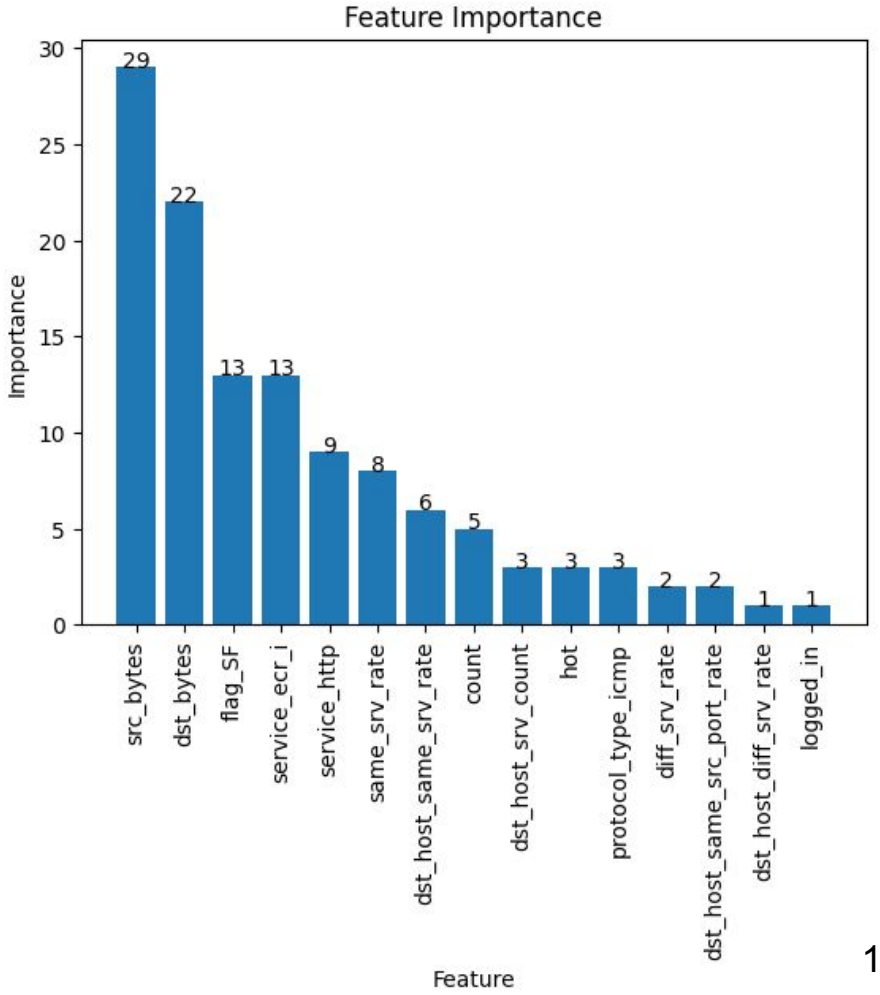
Weight	Feature
0.2555 ± 0.0036	src_bytes
0.1893 ± 0.0019	dst_bytes
0.0272 ± 0.0011	service_http
0.0271 ± 0.0004	service_ecr_i
0.0161 ± 0.0005	dst_host_same_srv_rate
0.0120 ± 0.0007	hot
0 ± 0.0000	dst_host_serror_rate
0 ± 0.0000	dst_host_srv_serror_rate
0 ± 0.0000	dst_host_rerror_rate
0 ± 0.0000	dst_host_srv_rerror_rate
0 ± 0.0000	protocol_type_icmp
0 ± 0.0000	protocol_type_udp
0 ± 0.0000	service_IRC
0 ± 0.0000	service_X11
0 ± 0.0000	service_Z39_50
0 ± 0.0000	service_aol
0 ± 0.0000	service_auth
0 ± 0.0000	dst_host_srv_diff_host_rate
0 ± 0.0000	protocol_type_tcp
0 ± 0.0000	flag_SH
… 102 more …	

Decision Tree Classifier

Conclusion of Feature Importance

- Select the top 5 of each method and assign weight respectively with the order.
- Make a Summation of weight from each method (For same feature that appear in other method)
- Sorting the features.

	Feature	Importance
14	src_bytes	29
2	dst_bytes	22
7	flag_SF	13
12	service_ecr_i	13
13	service_http	9
11	same_srv_rate	8
5	dst_host_same_srv_rate	6
0	count	5
6	dst_host_srv_count	3
8	hot	3
10	protocol_type_icmp	3
1	diff_srv_rate	2
4	dst_host_same_src_port_rate	2
3	dst_host_diff_srv_rate	1
9	logged_in	1



	Feature	Importance
14	src_bytes	29
2	dst_bytes	22
7	flag_SF	13
12	service_ecr_i	13
13	service_http	9
11	same_srv_rate	8
5	dst_host_same_srv_rate	6
0	count	5
6	dst_host_srv_count	3
8	hot	3
10	protocol_type_icmp	3
1	diff_srv_rate	2
4	dst_host_same_src_port_rate	2
3	dst_host_diff_srv_rate	1
9	logged_in	1

- src_bytes
  - Number of data bytes transferred from source to destination in single connection
- dst_bytes
  - Number of data bytes transferred from destination to source in single connection
- flag_SF
  - Normal establishment and termination.
- service_ecr_i
  - Internet Control Message Protocol (ICMP)
- service_http
  - http Protocol

Future Study Plan

- Implementation the model from paper
  - Using a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to Classify Network Attacks (Muhuri, 2020) (Currnetly working)
    - Using LSTM - RNN to detected the attack
    - High accuracy model
- Finding the method to change Network log data into image then find the model to detected attack incident

Thank you