Andreas Demosthenous

**Subject: Classification/Detection of DDoS traffic using machine learning algorithms**

**Background Information:**

A distributed denial-of-service (DDoS) attack is a malicious attempt to disrupt the normal traffic of a targeted server, service or network by overwhelming the target or its surrounding infrastructure with a flood of Internet traffic.

DDoS attacks achieve effectiveness by utilizing multiple compromised computer systems as sources of attack traffic.

From a high level, a DDoS attack is like an unexpected traffic jam clogging up the highway, preventing regular traffic from arriving at its destination.

DDoS attacks affect the security of a system, specifically the compromise of their availability.

**Motivation:**

Currently, Distributed Denial of Service Attacks are the most dangerous cyber danger. By inhibiting the server's ability to provide resources to genuine customers, the affected server's resources, such as bandwidth and buffer size, are slowed down. Therefore, developing effective intrusion detection systems using the power of AI is a very important task.

**Methodology:**

I used python to analyze existing DDoS datasets and evaluate different network traffic features and algorithms regarding 2 problems. DDoS types of classification among 10 different DDoS attack approaches and DDoS attack detection (2 – classes => DDoS traffic or Normal traffic)

I used the CIC-2019 DDoS dataset to experiment with. This dataset was created from the Canadian Institute of Technology using network traffic during a DDoS attack simulation. They used an open-source tool (CICFlowMeter) to convert the captured traffic to the dataset format.

Each dataset’s tuple represents a flow (Source to destination) with features(columns) such as Source IP, Destination IP, Protocol, Flow Duration, Total Forwarded Packets, Total Length of Forwarded Packets, Forwarded Packets/s, etc.

About the dataset:

* Initial size: 20+ GB
* Different .csv files for each DDoS kind mixed with normal traffic
* 87 network related features(columns)
* Millions of tuples (each row representing information about a source-destination flow)
* 1 target column labeling the kind of malicious DDoS traffic the packet(row) corresponds to. (BENIGN for normal traffic)

**Pros:** The dataset is relatively clean and big enough.

**Cons:** The normal traffic samples are relatively low compared to the rest of the DDoS traffic samples. (Less than 1%)

1. Sample(rows) Reduction:

To be able to load the dataset I had to reduce its size so for each DDoS kind (.csv file) I kept all the normal traffic sample as it was very small and extracted a fixed number X of random instances per DDoS kind to evaluate the performance for each X.

X = [5000, 10000, 50000, 100000]

\*Tuples containing null/Nan values were removed

\*Fields containing negative values were converted to 0

1. Feature(columns) Reduction:

* **Irrelevant features:** Imanually removed the features F that serve identification purposes and would be irrelevant to a real-world scenario.

F = ['FlowID', 'SourceIP', 'DestinationIP', 'Timestamp', 'SimillarHTTP', 'SourcePort', 'DestinationPort']

* **Constant features:** I used VarianceThreshold method from sklearn library to identify and remove the features G that remained constant (threshold = 0) and therefore are useless.

G = ['BwdPSHFlags', 'FwdURGFlags', 'BwdURGFlags', 'FINFlagCount', 'PSHFlagCount', 'ECEFlagCount', 'FwdAvgBytes/Bulk', 'FwdAvgPackets/Bulk', 'FwdAvgBulkRate', 'BwdAvgBytes/Bulk', 'BwdAvgPackets/Bulk', 'BwdAvgBulkRate']

* **Highly correlated features:** Iused both Pearson and Spearman algorithms to generate the correlation matrix of the features to identify and remove the features with at least 95% correlation with another feature. H is the union of the highly correlated features of the 2 algorithms.

H = ['ActiveMin', 'AveragePacketSize', 'FwdHeaderLength.1', 'MinPacketLength', 'SubflowBwdBytes', 'BwdIATMax', 'ACKFlagCount', 'ActiveMax', 'BwdPacketLengthMean', 'BwdIATMean', 'RSTFlagCount', 'SubflowBwdPackets', 'AvgFwdSegmentSize', 'FwdPacketLengthMean', 'MaxPacketLength', 'PacketLengthVariance', 'FwdIATStd', 'BwdPacketLengthMax', 'Init\_Win\_bytes\_forward', 'PacketLengthMean', 'BwdIATMin', 'FwdPackets/s', 'FwdIATMin', 'IdleMin', 'BwdPackets/s', 'BwdIATStd', 'IdleMean', 'FwdIATMax', 'FlowIATMax', 'AvgBwdSegmentSize', 'IdleMax', 'SubflowFwdPackets', 'Init\_Win\_bytes\_backward', 'FlowIATStd', 'FwdIATTotal', 'TotalLengthofFwdPackets', 'FwdIATMean', 'SubflowFwdBytes', 'PacketLengthStd', 'act\_data\_pkt\_fwd']

\* Target DDoS Labels are encoded numerical values to be fed into the ML algorithms.

1. Feature Selection:

\*Previously mentioned features are dropped before the best feature selection

To select the K most relevant features I used Chi-Square Tests and F-Tests and Mutual Info Classification tests.

K = [5, 10, 15, 20, 25]

K-Best features:

* X: Amount of malicious traffic per DDoS kind
* N: Amount of tuples
* Y: Amount of features

|  |  |  |  |
| --- | --- | --- | --- |
| X=5,000  N=79,526  Y=28 | X=10,000  N=127,653  Y=28 | X=50,000  N= 512,589  Y=28 | X=100,000  N= 993,847  Y=28 |

For each of the following 5 k values, 4 sample sizes and 3 selection algorithms there are 60 K-Best feature subsets generated.

|  |  |  |
| --- | --- | --- |
| chi2 | f\_classif | mutual\_info\_classif |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| K=5 | K=10 | K=15 | K=20 | K=25 |

\*I added another subset with the 22 best features according to a related study (Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy) on the same dataset.

1. Machine Learning:

For the dataset training I will try the following algorithms:

* 1. Decision Tree Classifier
  2. Random Forest Classifier
  3. K-Neighbors Classifier
  4. Gaussian Naïve Bayes
  5. Extreme Gradient Boosting Classifier
  6. C-Support Vector Classifier
  7. Neural Network -- Multi-layer Perceptron classifier
  8. Logistic Regression
  9. Linear Discriminant Analysis

I split the dataset into a training (75%) and a testing (25%) set and measured each algorithm’s accuracy on both sets.

**Regarding preprocessing:**

* **Tuple reduction** by keeping a fixed number of tuples for each DDoS type and keeping all normal traffic.
* **Negative values** converted to 0
* Tuples with **null/nan values** removed
* **Feature reduction**: Irrelevant, constant, and highly correlated features removed
* k-best **feature selection** with **chi2**, **f-tests** and **mutual info classification** algorithms
* **Scaling** with pipeline to [0,1] scale before training to avoid info leakage

Regarding target classes: 11 possible classes (10 DDoS types + benign - normal)

**7. Results**

**Graphs:**

**Chart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Graphical user interface, application

Description automatically generatedMachine Specs:**

**Graphical user interface, application

Description automatically generated**

Os: Windows 10 pro

**8. Conclusions:**

1. Most accurate algorithms are tree-based (Decision Trees, Random Forest, and Extreme Gradient Boost) with XGBC being the most precise in every case.

I am assuming this is the case since DDoS attacks have simplified structure (High packets/s, small time interval between consequent packets, high flow duration, small packet lengths etc.), therefore a tree-based algorithm is able to detect that by traversing a tree and checking the status of a network feature in each iteration to learn and later detect the flows where their features are matching with DDoS flows features.

1. Accuracy decreases when the sample size increases.

I am assuming this happened because in every sample size I kept the normal traffic samples the same and therefore the testing would contain more normal traffic samples which are easier to detect than DDoS specific types. The accuracy seems more stable from the 500k sample to 1mil sample sizes since normal traffic takes a very small portion of the set.

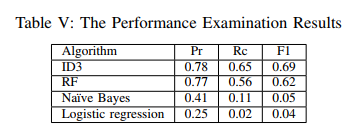
1. Decision Trees and Random Forest are slightly overfitted. Overfitting is generally more common in these tree-based algorithms, but XGBC is able to handle overfitting much better.
2. SVM is too slow (5.5 hours of training for the ~500k sample)
3. Models with more features (higher K improved the accuracy).

I assume this is the case since having to classify between different ddos kinds is a much harder problem than simply detecting any ddos traffic. Therefore, more features are necessary to detect that distinction.



1. Scaling significantly improved accuracy especially for MLP and SVM. For some algorithms like MLP and SVM it is necessary to perform scaling to remove any bias created from the numerical size of the features (higher value -> more important -> incorrect assumption in our case). Tree based algorithms don’t have such a requirement.
2. The experimental results coincide with the study’s results on the same dataset. The k-best subset of the study (S = 16, k = 22) was the best one selected using the XGBC algorithm and the algorithms accuracies are similar but for Logistic regression I managed to get a higher accuracy of around 50-60%. For Random Forest our accuracy was 70-80% and for Naïve Bays around 50-60% depending on the sample size. In our experiments I used Gaussian Naïve Bayes instead of the Multinomial Naïve Bays the study used, since I faced some technical difficulties running it.

Accuracy results of the study:





1. Best subsets for XGBC are:
2. (S = 16, k = 22) the subset I extracted from the study)
3. (S = 12, k = 20)
4. (S = 13, k = 25)

**(S = 16, k = 22)** = ['Protocol', 'FlowDuration', 'TotalFwdPackets', 'TotalLengthofFwdPackets', 'FwdPacketLengthMax', 'FwdPacketLengthMin', 'FwdPacketLengthStd', 'FlowIATMin', 'FwdIATTotal', 'FwdIATMean', 'FwdIATMax', 'FwdIATMin', 'FwdHeaderLength', 'FwdPackets/s', 'MinPacketLength', 'MaxPacketLength', 'PacketLengthStd', 'ACKFlagCount', 'AveragePacketSize', 'SubflowFwdBytes', 'Init\_Win\_bytes\_forward', 'min\_seg\_size\_forward']

**(S = 12, k = 20)** = ['Protocol', 'FlowDuration', 'TotalFwdPackets', 'TotalBackwardPackets', 'TotalLengthofBwdPackets', 'FwdPacketLengthMax', 'FwdPacketLengthMin', 'FwdPacketLengthStd', 'BwdPacketLengthMin', 'FlowBytes/s', 'FlowPackets/s', 'FlowIATMean', 'FlowIATMin', 'BwdIATTotal', 'FwdHeaderLength', 'BwdHeaderLength', 'URGFlagCount', 'Down/UpRatio', 'min\_seg\_size\_forward', 'Inbound']

**(S = 13, k = 25)** = ['Protocol', 'FlowDuration', 'TotalFwdPackets', 'TotalBackwardPackets', 'TotalLengthofBwdPackets', 'FwdPacketLengthMax', 'FwdPacketLengthMin', 'FwdPacketLengthStd', 'BwdPacketLengthMin', 'BwdPacketLengthStd', 'FlowBytes/s', 'FlowPackets/s', 'FlowIATMean', 'FlowIATMin', 'BwdIATTotal', 'FwdPSHFlags', 'FwdHeaderLength', 'BwdHeaderLength', 'URGFlagCount', 'CWEFlagCount', 'Down/UpRatio', 'min\_seg\_size\_forward', 'ActiveMean', 'ActiveStd', 'IdleStd']

**9. DDoS detection**

I repeated the experiment reducing the target classes to 2 (DDoS or Not DDoS) and the accuracy for most algorithms was 100%.

**10. K-best features I generated**

Explanation for K\_best\_5000: 5000 stands for 5000 samples per ddos type-> smallest data sample. Subsets 1-3 contain the 5-best features generated from each of the 3 subset selection algorithms (chi2, f-tests, mutual info tests, 4-7 the 10-best features etc., and subset 16 is the subset from the study.

**K\_best\_5000** = [  
 [**'FlowDuration'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'BwdIATTotal'**],  
 [**'Protocol'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'Inbound'**],  
 [**'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**],  
  
 [**'FlowDuration'**, **'TotalLengthofBwdPackets'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'BwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'Inbound'**],  
 [**'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FwdHeaderLength'**, **'Inbound'**],  
  
 [**'FlowDuration'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FwdPSHFlags'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'Inbound'**],  
  
 [**'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FlowDuration'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**,**'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**,**'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'Inbound'**],  
  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**,**'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**,**'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**,**'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**,**'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'SYNFlagCount'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'IdleStd'**, **'Inbound'**],  
  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalLengthofFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'FlowIATMin'**, **'FwdIATTotal'**, **'FwdIATMean'**, **'FwdIATMax'**, **'FwdIATMin'**, **'FwdHeaderLength'**, **'FwdPackets/s'**, **'MinPacketLength'**, **'MaxPacketLength'**, **'PacketLengthStd'**, **'ACKFlagCount'**, **'AveragePacketSize'**, **'SubflowFwdBytes'**, **'Init\_Win\_bytes\_forward'**, **'min\_seg\_size\_forward'**],  
  
]

**K\_best\_10000** = [  
  
 [**'FlowDuration'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'BwdIATTotal'**],  
 [**'Protocol'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'Inbound'**],  
 [**'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**],  
  
 [**'FlowDuration'**, **'TotalLengthofBwdPackets'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'TotalFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'BwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'URGFlagCount'**, **'Down/UpRatio'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FwdHeaderLength'**, **'Inbound'**],  
  
 [**'FlowDuration'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'TotalFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FwdPSHFlags'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'min\_seg\_size\_forward'**, **'Inbound'**],  
  
 [**'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'Inbound'**],  
  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'SYNFlagCount'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'IdleStd'**, **'Inbound'**]  
  
 ,[**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalLengthofFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'FlowIATMin'**, **'FwdIATTotal'**, **'FwdIATMean'**, **'FwdIATMax'**, **'FwdIATMin'**, **'FwdHeaderLength'**, **'FwdPackets/s'**, **'MinPacketLength'**, **'MaxPacketLength'**, **'PacketLengthStd'**, **'ACKFlagCount'**, **'AveragePacketSize'**, **'SubflowFwdBytes'**, **'Init\_Win\_bytes\_forward'**, **'min\_seg\_size\_forward'**],  
  
]

**K\_best\_50000** = [  
  
 [**'FlowDuration'**, **'FlowBytes/s'**, **'FlowIATMean'**, **'BwdIATTotal'**, **'IdleStd'**],  
 [**'Protocol'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'Inbound'**],  
 [**'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**],  
  
 [**'FlowDuration'**, **'TotalLengthofBwdPackets'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'BwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FwdHeaderLength'**, **'min\_seg\_size\_forward'**],  
  
 [**'FlowDuration'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FlowDuration'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FwdPSHFlags'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'min\_seg\_size\_forward'**, **'Inbound'**],  
  
 [**'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'Inbound'**],  
  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**, **'Inbound'**]  
  
 ,[**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalLengthofFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'FlowIATMin'**, **'FwdIATTotal'**, **'FwdIATMean'**, **'FwdIATMax'**, **'FwdIATMin'**, **'FwdHeaderLength'**, **'FwdPackets/s'**, **'MinPacketLength'**, **'MaxPacketLength'**, **'PacketLengthStd'**, **'ACKFlagCount'**, **'AveragePacketSize'**, **'SubflowFwdBytes'**, **'Init\_Win\_bytes\_forward'**, **'min\_seg\_size\_forward'**],  
  
]

**K\_best\_100000** = [  
  
 [**'FlowDuration'**, **'FlowBytes/s'**, **'FlowIATMean'**, **'BwdIATTotal'**, **'IdleStd'**],  
 [**'Protocol'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'Inbound'**],  
 [**'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**],  
  
 [**'FlowDuration'**, **'TotalLengthofBwdPackets'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'BwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FwdHeaderLength'**, **'min\_seg\_size\_forward'**],  
  
 [**'FlowDuration'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FlowDuration'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FwdPSHFlags'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'min\_seg\_size\_forward'**, **'Inbound'**],  
  
 [**'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'Inbound'**],  
  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdPSHFlags'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**, **'Inbound'**],  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalBackwardPackets'**, **'TotalLengthofBwdPackets'**, **'FwdPacketLengthMax'**, **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'BwdPacketLengthMin'**, **'BwdPacketLengthStd'**, **'FlowBytes/s'**, **'FlowPackets/s'**, **'FlowIATMean'**, **'FlowIATMin'**, **'BwdIATTotal'**, **'FwdHeaderLength'**, **'BwdHeaderLength'**, **'URGFlagCount'**, **'CWEFlagCount'**, **'Down/UpRatio'**, **'min\_seg\_size\_forward'**, **'ActiveMean'**, **'ActiveStd'**, **'IdleStd'**, **'Inbound'**],  
  
 [**'Protocol'**, **'FlowDuration'**, **'TotalFwdPackets'**, **'TotalLengthofFwdPackets'**, **'FwdPacketLengthMax'**,  
 **'FwdPacketLengthMin'**, **'FwdPacketLengthStd'**, **'FlowIATMin'**, **'FwdIATTotal'**, **'FwdIATMean'**, **'FwdIATMax'**, **'FwdIATMin'**,  
 **'FwdHeaderLength'**, **'FwdPackets/s'**, **'MinPacketLength'**, **'MaxPacketLength'**, **'PacketLengthStd'**, **'ACKFlagCount'**,  
 **'AveragePacketSize'**, **'SubflowFwdBytes'**, **'Init\_Win\_bytes\_forward'**, **'min\_seg\_size\_forward'**]  
  
]

***Features Descriptions:***

Feature Name Description

Flow duration Duration of the flow in Microsecond

total Fwd Packet Total packets in the forward direction

total Bwd packets Total packets in the backward direction

total Length of Fwd Packet Total size of packet in forward direction

total Length of Bwd Packet Total size of packet in backward direction

Fwd Packet Length Min Minimum size of packet in forward direction

Fwd Packet Length Max Maximum size of packet in forward direction

Fwd Packet Length Mean Mean size of packet in forward direction

Fwd Packet Length Std Standard deviation size of packet in forward direction

Bwd Packet Length Min Minimum size of packet in backward direction

Bwd Packet Length Max Maximum size of packet in backward direction

Bwd Packet Length Mean Mean size of packet in backward direction

Bwd Packet Length Std Standard deviation size of packet in backward direction

Flow Bytes/s Number of flow bytes per second

Flow Packets/s Number of flow packets per second

Flow IAT Mean Mean time between two packets sent in the flow

Flow IAT Std Standard deviation time between two packets sent in the flow

Flow IAT Max Maximum time between two packets sent in the flow

Flow IAT Min Minimum time between two packets sent in the flow

Fwd IAT Min Minimum time between two packets sent in the forward direction

Fwd IAT Max Maximum time between two packets sent in the forward direction

Fwd IAT Mean Mean time between two packets sent in the forward direction

Fwd IAT Std Standard deviation time between two packets sent in the forward direction

Fwd IAT Total Total time between two packets sent in the forward direction

Bwd IAT Min Minimum time between two packets sent in the backward direction

Bwd IAT Max Maximum time between two packets sent in the backward direction

Bwd IAT Mean Mean time between two packets sent in the backward direction

Bwd IAT Std Standard deviation time between two packets sent in the backward direction

Bwd IAT Total Total time between two packets sent in the backward direction

Fwd PSH flags Number of times the PSH flag was set in packets travelling in the forward direction (0 for UDP)

Bwd PSH Flags Number of times the PSH flag was set in packets travelling in the backward direction (0 for UDP)

Fwd URG Flags Number of times the URG flag was set in packets travelling in the forward direction (0 for UDP)

Bwd URG Flags Number of times the URG flag was set in packets travelling in the backward direction (0 for UDP)

Fwd Header Length Total bytes used for headers in the forward direction

Bwd Header Length Total bytes used for headers in the backward direction

FWD Packets/s Number of forward packets per second

Bwd Packets/s Number of backward packets per second

Packet Length Min Minimum length of a packet

Packet Length Max Maximum length of a packet

Packet Length Mean Mean length of a packet

Packet Length Std Standard deviation length of a packet

Packet Length Variance Variance length of a packet

FIN Flag Count Number of packets with FIN

SYN Flag Count Number of packets with SYN

RST Flag Count Number of packets with RST

PSH Flag Count Number of packets with PUSH

ACK Flag Count Number of packets with ACK

URG Flag Count Number of packets with URG

CWR Flag Count Number of packets with CWR

ECE Flag Count Number of packets with ECE

down/Up Ratio Download and upload ratio

Average Packet Size Average size of packet

Fwd Segment Size Avg Average size observed in the forward direction

Bwd Segment Size Avg Average size observed in the backward direction

Fwd Bytes/Bulk Avg Average number of bytes bulk rate in the forward direction

Fwd Packet/Bulk Avg Average number of packets bulk rate in the forward direction

Fwd Bulk Rate Avg Average number of bulk rate in the forward direction

Bwd Bytes/Bulk Avg Average number of bytes bulk rate in the backward direction

Bwd Packet/Bulk Avg Average number of packets bulk rate in the backward direction

Bwd Bulk Rate Avg Average number of bulk rate in the backward direction

Subflow Fwd Packets The average number of packets in a sub flow in the forward direction

Subflow Fwd Bytes The average number of bytes in a sub flow in the forward direction

Subflow Bwd Packets The average number of packets in a sub flow in the backward direction

Subflow Bwd Bytes The average number of bytes in a sub flow in the backward direction

Fwd Init Win bytes The total number of bytes sent in initial window in the forward direction

Bwd Init Win bytes The total number of bytes sent in initial window in the backward direction

Fwd Act Data Pkts Count of packets with at least 1 byte of TCP data payload in the forward direction

Fwd Seg Size Min Minimum segment size observed in the forward direction

Active Min Minimum time a flow was active before becoming idle

Active Mean Mean time a flow was active before becoming idle

Active Max Maximum time a flow was active before becoming idle

Active Std Standard deviation time a flow was active before becoming idle

Idle Min Minimum time a flow was idle before becoming active

Idle Mean Mean time a flow was idle before becoming active

Idle Max Maximum time a flow was idle before becoming active

Idle Std Standard deviation time a flow was idle before becoming active