Comprehensive Analysis of Botnet Behavior in IoT Networks:

Patterns, Impacts, and Detection of type of attack Using Machine Learning

Team Members

Ankur Kaushal

Dyuti Dasmahapatra

Yash Verma

1. Discovery

Business Understanding

IoT Networks: Networks of devices that can connect to the internet and talk to each other.

Botnets: Groups of infected devices controlled by hackers without the owners knowing.

Threats from Botnets: Botnets can cause a lot of trouble in IoT networks.

Types of Attack

Denial of Service (DoS) Attacks

The botnet floods a target device or network with an overwhelming amount of traffic, rendering it unable to respond to legitimate requests.

Distributed Denial of Service (DDoS) Attacks

Similar to DoS, involve multiple devices in the botnet collectively targeting a single victim.

Data Theft and Privacy Breaches:

Used to steal sensitive information from IoT devices, such as personal data, login credentials, or financial information.

Pain points

- Vulnerabilities in IoT Devices
- Large Attack Surface
- Lack of Security Standards
- Limited Resources for Security Updates
- Privacy Concerns
- Potential for Large-Scale Disruptions

Problem Statement

Despite advancements in IoT security, botnets continue to pose significant threats to IoT networks, compromising data integrity and system performance.

This study aims to comprehensively analyze botnet behavior in IoT networks, including attack patterns, impacts, and detection mechanisms using machine learning.

Analytic problem type

Classification problem: Identifying the type of botnet attacks within IoT network traffic data.

Objectives

Gain insights into botnet behavior and characteristics within IoT networks through an in-depth analysis of attack patterns, features, and impacts.

Develop accurate detection mechanisms using machine learning to identify various types of botnet attacks in IoT networks.

Success Criteria

Successfully identify and analyze botnet attack patterns in IoT network traffic data.

Develop machine learning models with high detection accuracy for different types of botnet attacks.

Demonstrate the effectiveness of the detection mechanisms by reducing the false positive rate and response time

Key Risks

Limited availability of comprehensive and diverse datasets for training machine learning models

Ensuring that machine learning models generalize well to new and unseen botnet attack instances.

Limited understanding of IoT network architecture, protocols, and security vulnerabilities

Initial Hypothesis

Hypothesis 1

Botnet attacks in IoT networks exhibit distinct patterns in network traffic data, including abnormal packet and byte counts, unusual traffic rates, and atypical protocol usage.

Hypothesis 2

Machine learning models trained on diverse datasets containing features indicative of botnet activities can accurately detect and classify different types of botnet attacks in IoT networks.

Use of the 3 V’s

Volume

Analyzing large volumes of IoT network traffic data to identify patterns and anomalies indicative of botnet attacks.

Variety

Handling diverse types of network traffic features, including source and destination IP addresses, ports, protocols, packet counts, and traffic rates.

Velocity

Detecting botnet attacks in real-time or near real-time to enable proactive defense measures and minimize the impact on IoT network infrastructure.

Resource Requirements

Data: Access to comprehensive and labeled datasets containing IoT network traffic data with instances of botnet attacks.

Computing Resources: Sufficient computational power and storage capacity to preprocess, analyze, and train machine learning models on large datasets.

Expertise: Domain knowledge in IoT security, data preprocessing, machine learning, and cybersecurity to effectively conduct the analysis and develop detection mechanisms.

Software Tools: Utilization of programming languages (e.g., Python), libraries (e.g., scikit-learn, TensorFlow), and data visualization tools for data analysis and model development.

2. Data Perparation

pkSeqID	stime	flgs	flgs_number	proto	proto_number	saddr	sport	daddr	dport	pkts	bytes	state	state_number	ltime	seq	dur
1000001	1528084952.58206	e	1	udp	3	192.168.100.148	37153	192.168.100.6	80	8	480	INT	4	1528084977.58371	120567	25.00
1000002	1528084952.58206	e	1	udp	3	192.168.100.148	37154	192.168.100.6	80	8	480	INT	4	1528084977.58371	120568	25.00
1000003	1528084952.58206	e	1	udp	3	192.168.100.148	37155	192.168.100.6	80	8	480	INT	4	1528084977.58371	120569	25.00
1000004	1528084952.58206	e	1	udp	3	192.168.100.148	37156	192.168.100.6	80	8	480	INT	4	1528084977.58371	120570	25.00
1000005	1528084952.58207	e	1	udp	3	192.168.100.148	37157	192.168.100.6	80	8	480	INT	4	1528084977.58371	120571	25.00
1000006	1528084952.58207	e	1	udp	3	192.168.100.148	37158	192.168.100.6	80	8	480	INT	4	1528084977.58371	120572	25.00
1000007	1528084952.58207	e	1	udp	3	192.168.100.148	37159	192.168.100.6	80	8	480	INT	4	1528084977.58371	120573	25.00
1000008	1528084952.58207	e	1	udp	3	192.168.100.148	37160	192.168.100.6	80	8	480	INT	4	1528084977.58371	120574	25.00
1000009	1528084952.58207	e	1	udp	3	192.168.100.148	37161	192.168.100.6	80	8	480	INT	4	1528084977.58371	120575	25.00
1000010	1528084952.58207	e	1	udp	3	192.168.100.148	37162	192.168.100.6	80	8	480	INT	4	1528084977.58371	120576	25.00
1000011	1528084952.58207	eU	6	udp	3	192.168.100.148	37163	192.168.100.6	80	8	480	INT	4	1528084977.58371	120577	25.00
1000012	1528084952.58207	eU	6	udp	3	192.168.100.148	37164	192.168.100.6	80	8	480	INT	4	1528084977.58371	120578	25.00
1000013	1528084952.58207	eU	6	udp	3	192.168.100.148	37165	192.168.100.6	80	8	480	INT	4	1528084977.58371	120579	25.00
1000014	1528084952.58207	eU	6	udp	3	192.168.100.148	37166	192.168.100.6	80	8	480	INT	4	1528084977.58371	120580	25.00
1000015	1528084952.58207	eU	6	udp	3	192.168.100.148	37167	192.168.100.6	80	8	480	INT	4	1528084977.58371	120581	25.00
1000016	1528084952.58207	eU	6	udp	3	192.168.100.148	37168	192.168.100.6	80	8	480	INT	4	1528084977.58371	120582	25.00
1000017	1528084952.58207	eU	6	udp	3	192.168.100.148	37169	192.168.100.6	80	8	480	INT	4	1528084977.58371	120583	25.00
1000018	1528084952.58207	eU	6	udp	3	192.168.100.148	37170	192.168.100.6	80	8	480	INT	4			
1000019	1528084952.58207	eU	6	udp	3	192.168.100.148	37171	192.168.100.6	80	8	480	INT	4			
1000020	1528084952.58207	eU	6	udp	3	192.168.100.148	37172	192.168.100.6	80	8	480	INT	4			
1000021	1528084952.58207	eU	6	udp	3	192.168.100.148	37173	192.168.100.6	80	8	480	INT	4			
1000022	1528084952.58207	eU	6	udp	3	192.168.100.148	37174	192.168.100.6	80	8	480	INT	4			
1000023	1528084952.58207	eU	6	udp	3	192.168.100.148	37175	192.168.100.6	80	8	480	INT	4			

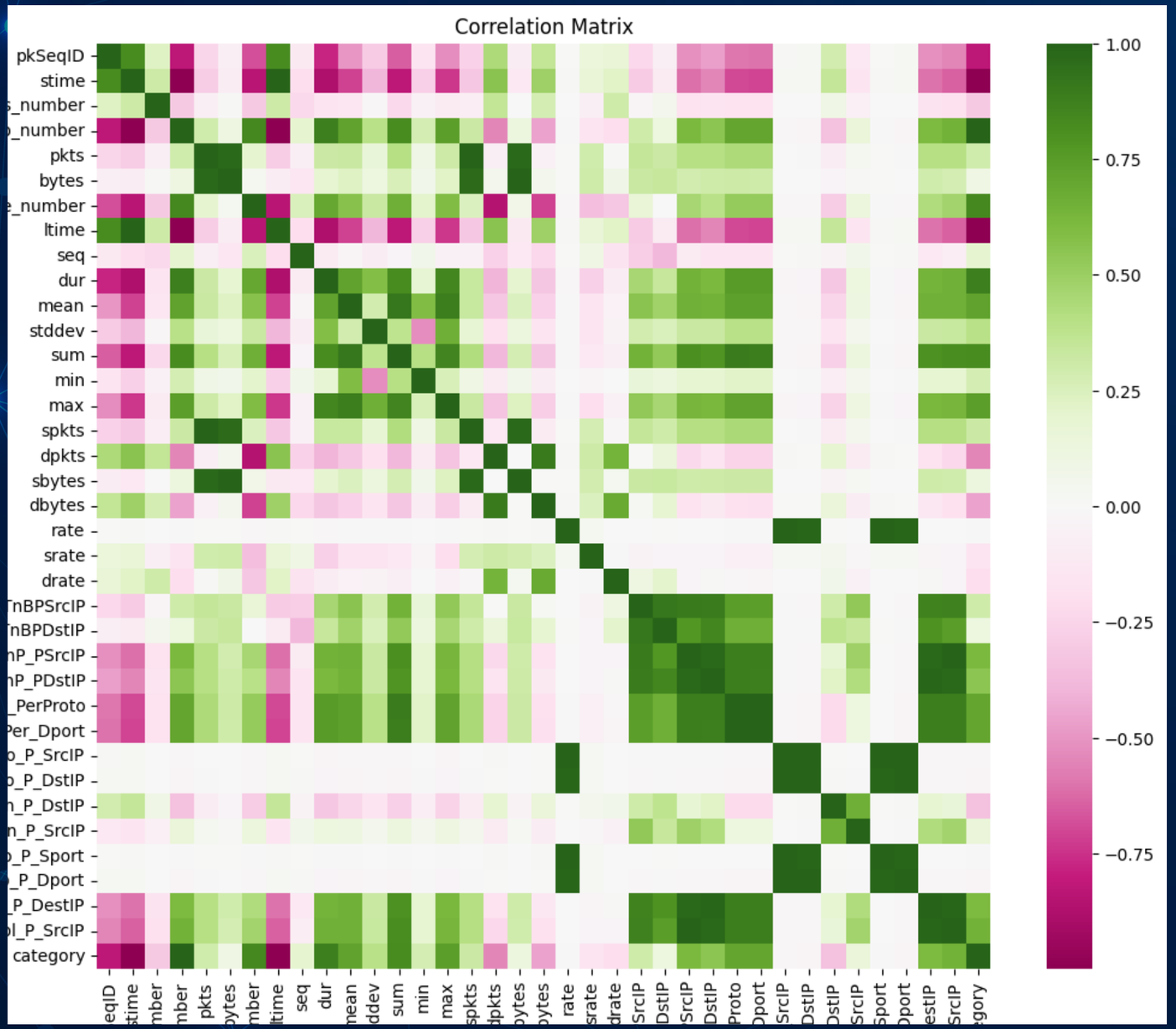
DATASET

Dataset Description

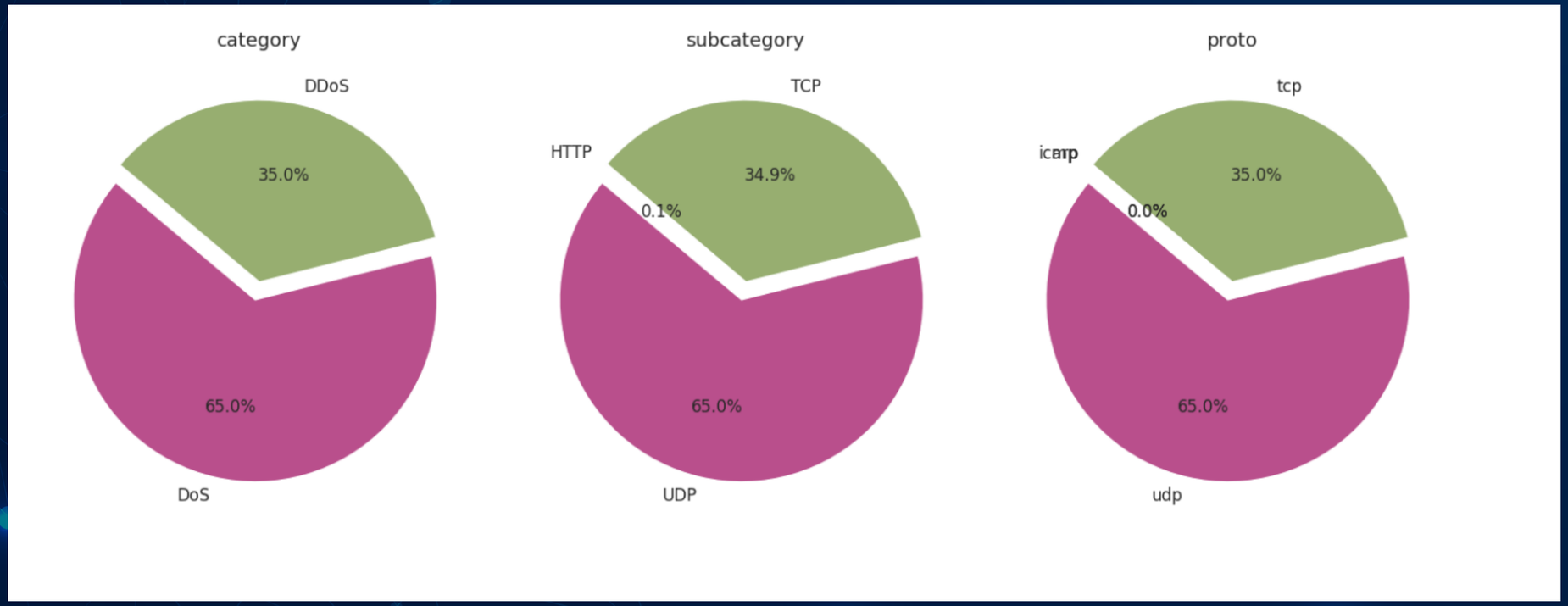
- We selected the Dataset from kaggle name - **BoT-IoT - All Features - 5% sample**
- The dataset comprises of 46 columns and around 1 million rows and after preprocessing we were left with 40 columns.
- The main columns are Category, Subcategory, State, flags, proto

EDA

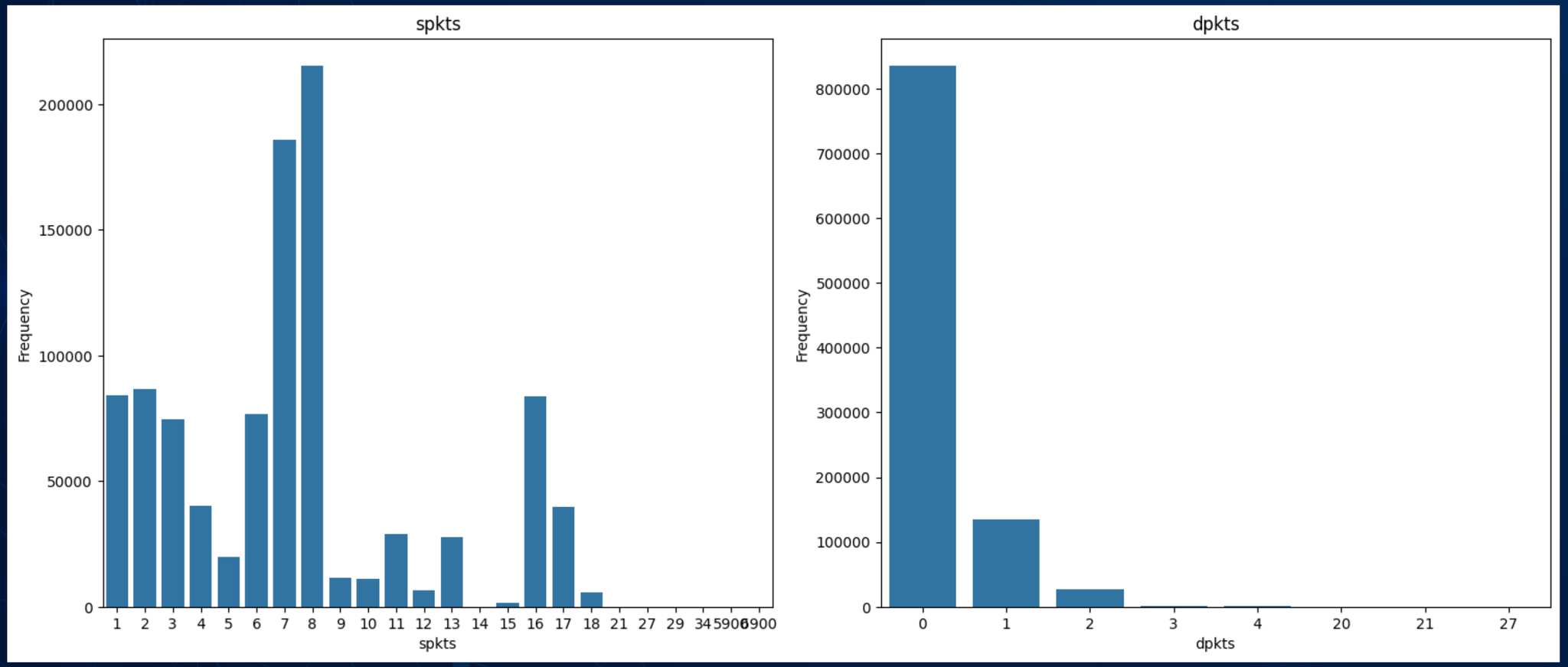
Results


Correlation Matrix
pkSeqID
stime
_number
_number
pkts
bytes
_number
ltime
seq
dur
mean
stddev
sum
min
max
spkts
dpkts
sbytes
dbytes
rate
srate
drate
TnBPSrcIP
TnBPDstIP
nP_PSrcIP
nP_PDstIP
_PerProto
Per_Dport
o_P_SrcIP
o_P_DstIP
n_P_DstIP
n_P_SrcIP
o_P_Sport
_P_Dport
_P_DestIP
l_P_SrcIP
category
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

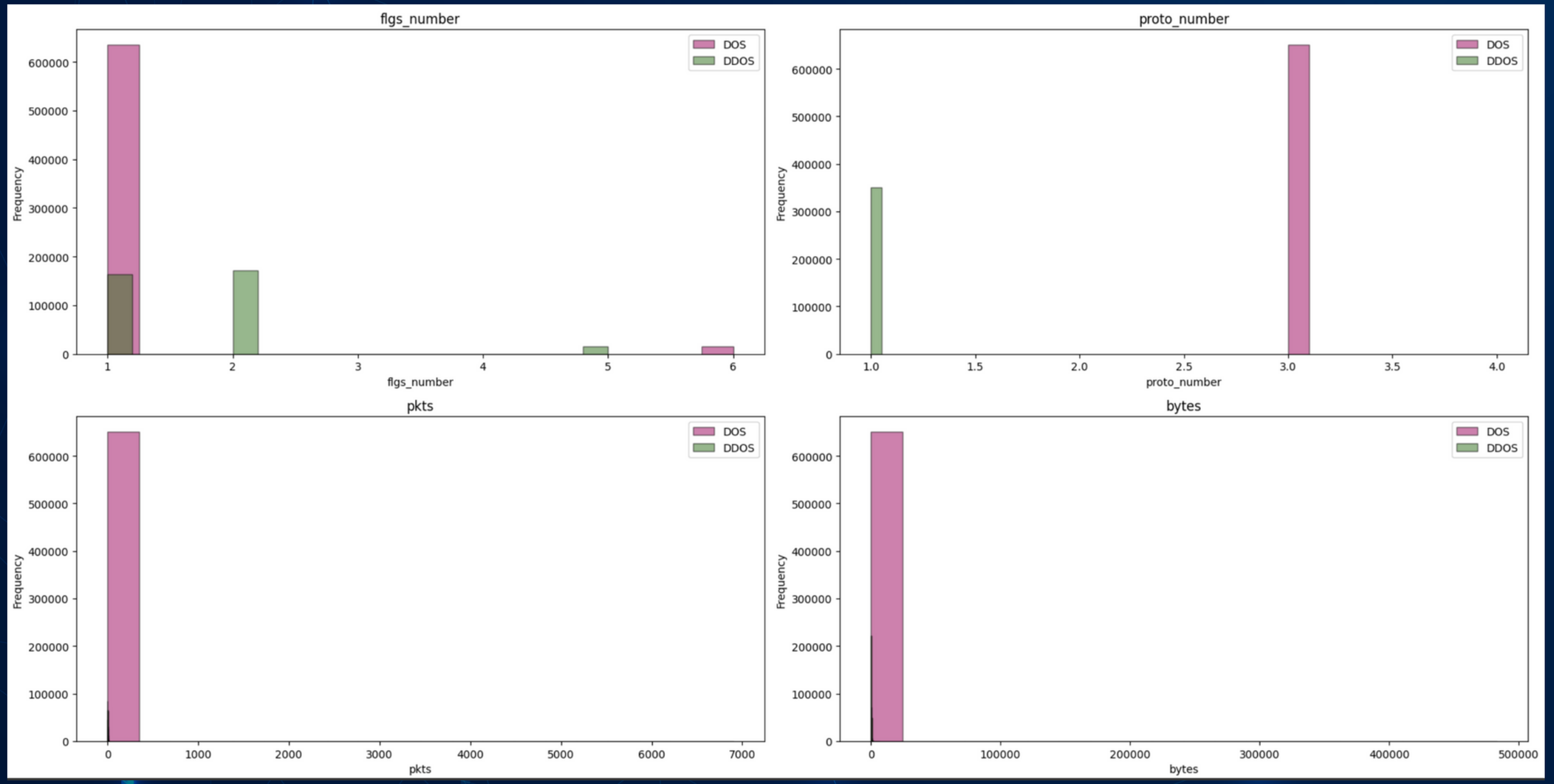
Results


category
DDoS
35.0%
65.0%
DoS
subcategory
TCP
HTTP
34.9%
0.1%
65.0%
UDP
proto
tcp
icmp
35.0%
0.0%
65.0%
udp

Results


spkts
Frequency
200000
150000
100000
50000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 21 27 29 34 5906 6900
spkts
dpkts
Frequency
800000
700000
600000
500000
400000
300000
200000
100000
0
0 1 2 3 4 20 21 27
dpkts

Results


flgs_number
DOS
DDOS
Frequency
flgs_number
proto_number
DOS
DDOS
Frequency
proto_number
pkts
DOS
DDOS
Frequency
pkts
bytes
DOS
DDOS
Frequency
bytes

3. Model Planning

Comparitive Analysis

Aspect	Logistic Regression	Random Forest	Neural Networks
Model Complexity	Linear model	Ensemble of decision trees	Composed of interconnected layers
Interpretability	High	Moderate	Low
Nonlinearity	Limited (linear decision boundary)	Can capture nonlinearity	Can capture complex nonlinearity
Feature Importance	Coefficients indicate importance	Feature importance from trees	Not as straightforward
Overfitting	Less likely	Moderate (can be controlled)	Prone, but regularization helps
Performance	Works well with linear relationships	Effective for various data types	Powerful for complex patterns
Training Time	Fast	Moderate	Slow (especially for deep nets)
Scalability	Scales well	Can be slow for large datasets	Can be resource-intensive


Thank You