HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Network Attacks Detection

Presentation by Group 16

Group members:
Doan Minh Viet 20210933
Nguyen Viet Trung 20214934
Do Hoang Tuan 20214939
Dau Van Can 20214879
Ngo Viet Anh 20214875

Instructor Assoc Prof.Linh Giang Nguyen

Contents

I.Introduction

- Internet is a global system of interconnected computer networks.
- There is always a chance of getting attacked, whether by DDOS, Website Defacement, Directory Traversal, etc
- Several models have been proposed and implemented
- In this project, our group will try to build software to detect network intrusions and protect a computer network from unauthorized users
- The intrusion detector learning task is to build a predictive model capable of distinguishing between “bad” connections, called intrusions or attacks, and “good” normal connection

Exploratory Data Analysis

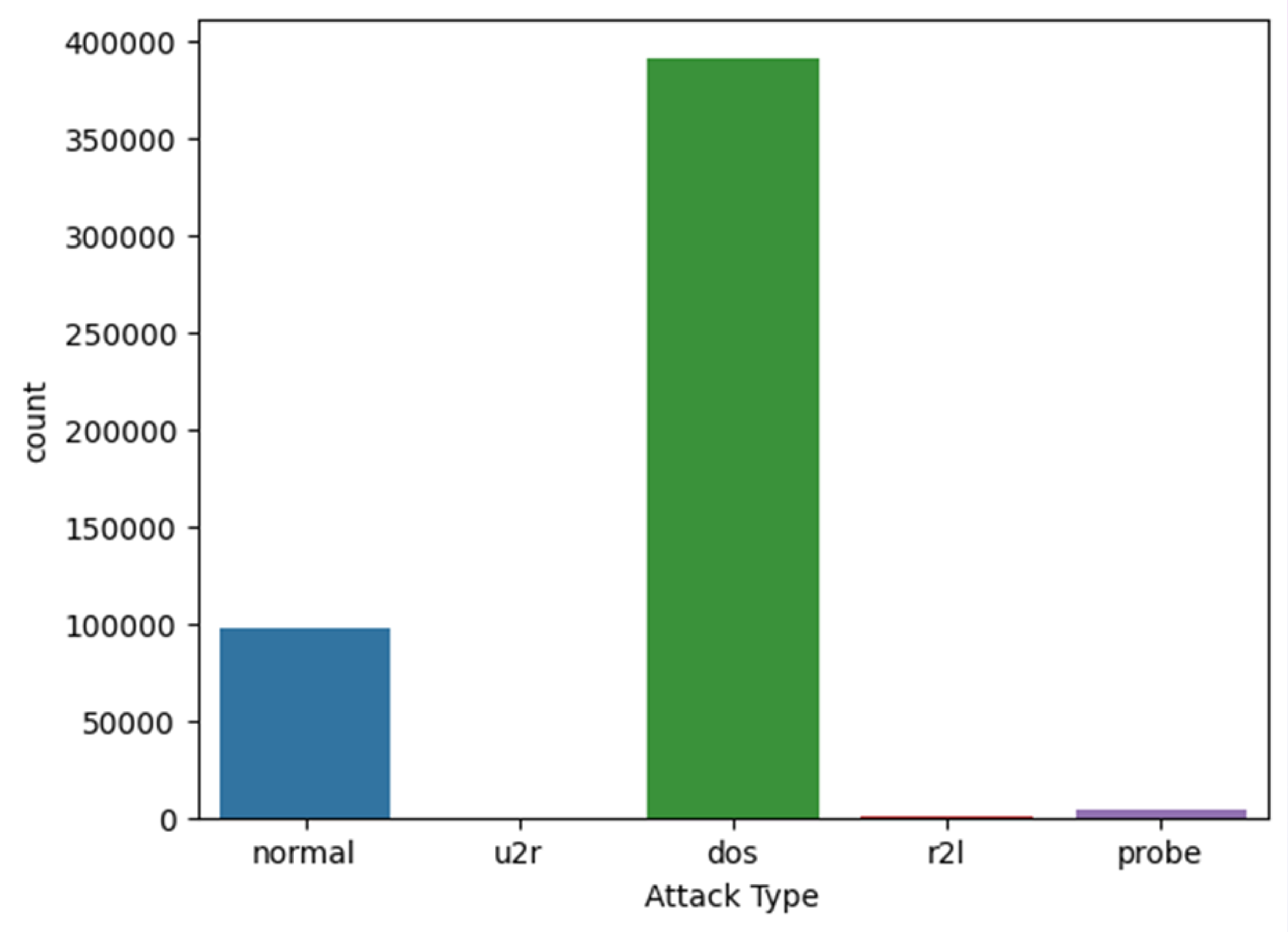
Data Understanding & Visualization

The simulated attacks are categorized into one of four categories:

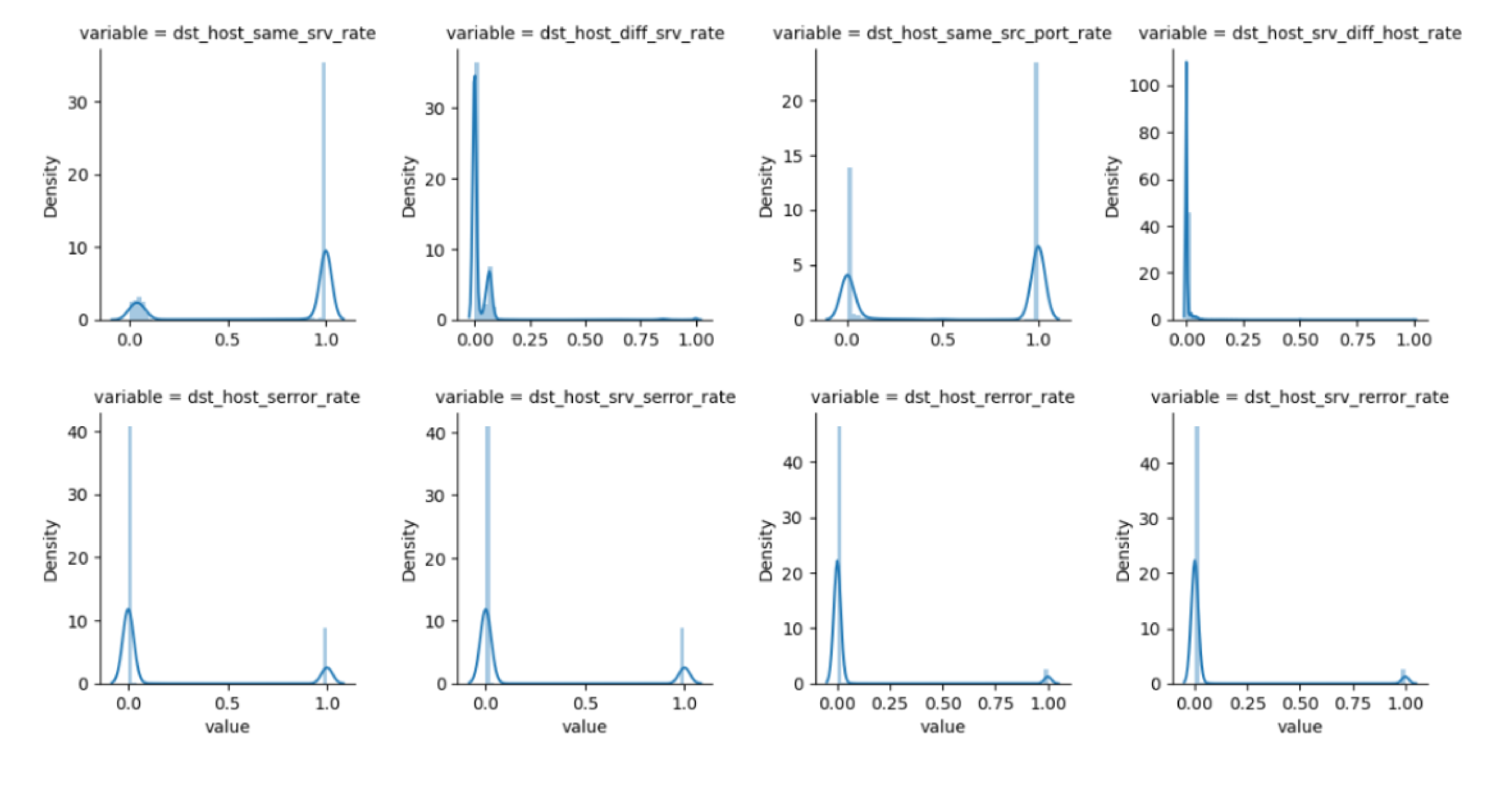
- Denial of Service Attack (DoS)
- User to Root Attack (U2R)
- Remote to Local Attack (R2L)
- Probing Attack

Besides the target variable, KDD Cup 99 features can be classified into three groups (two derived feature categories):

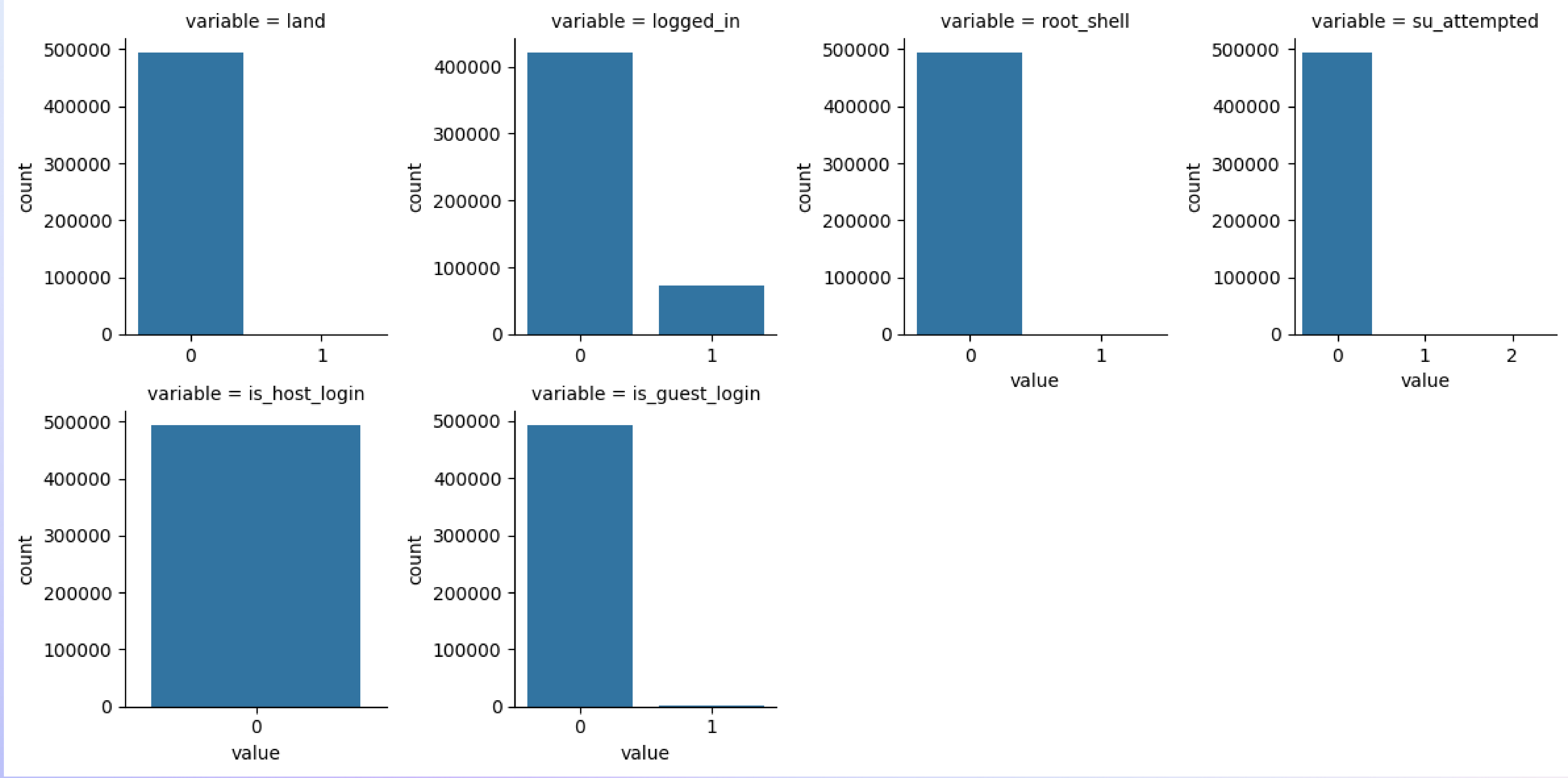
- Basic features
- Content features
- Traffic features



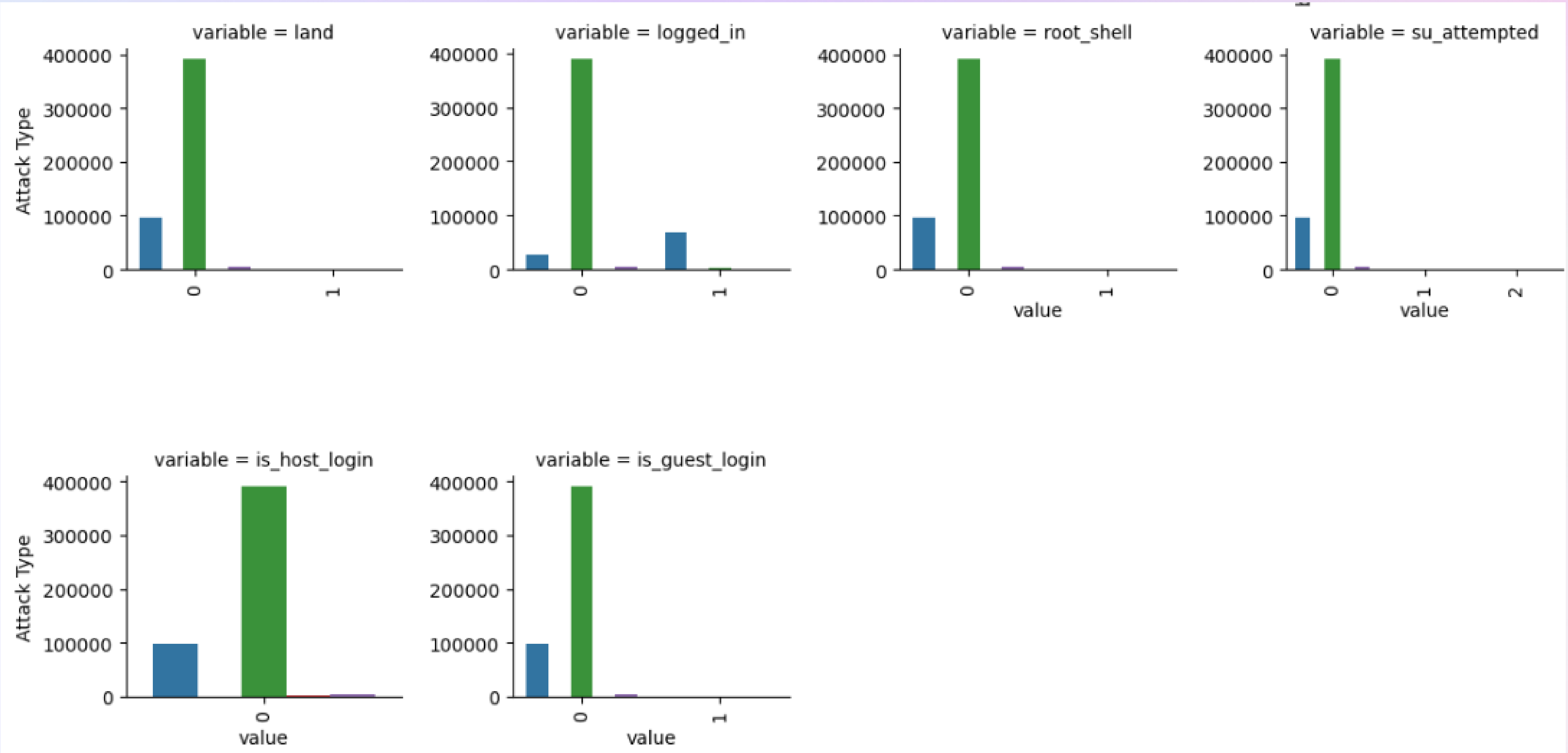
Count Histogram of Attack Type(output)



Distribution of Some Continuous Variables



Count Histogram of Discrete Variables



Count Histogram of Some Categorical Variables

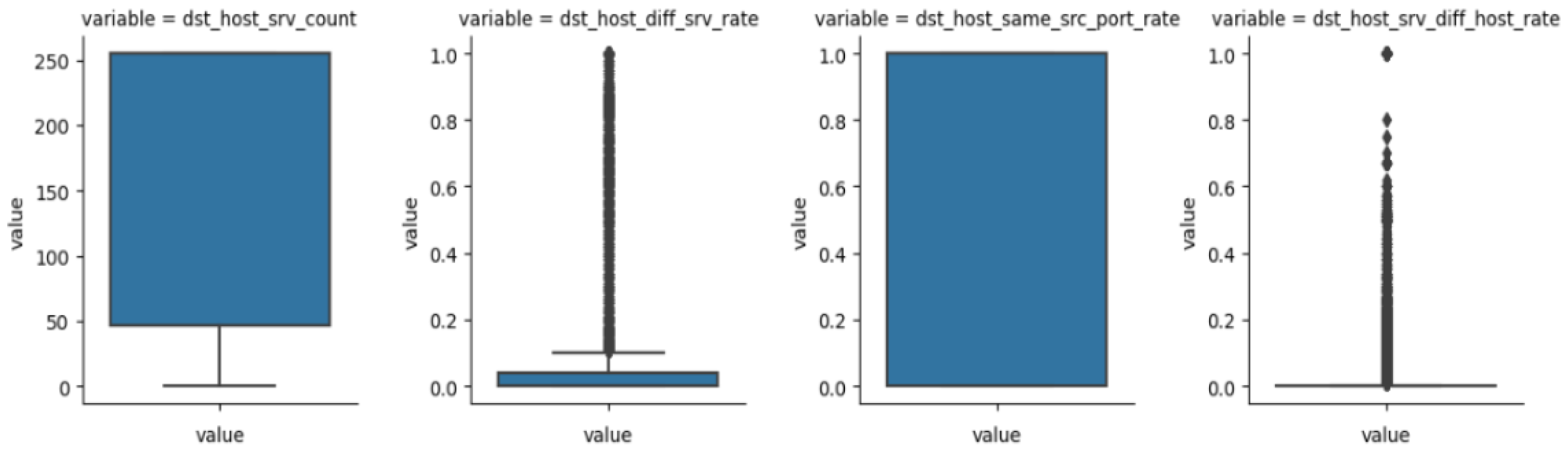
Data Cleaning

a. Missing Value

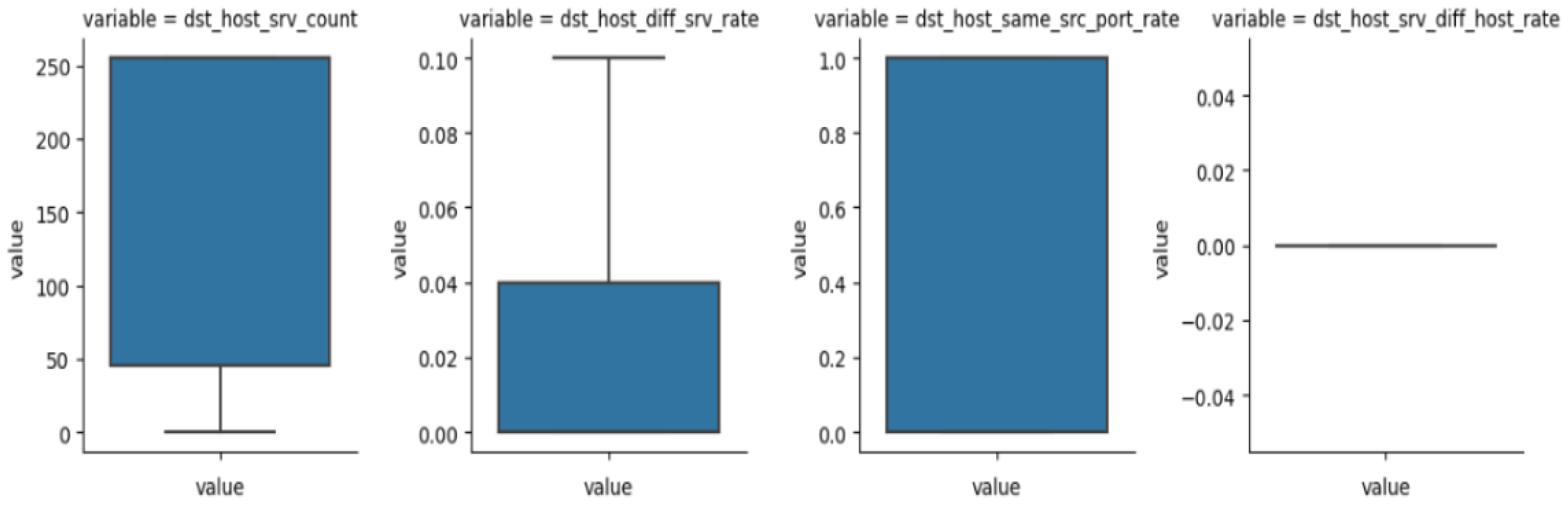
We depicted the heatmap of null values for each attribute in. We find that there are no null values for all attributes so we decided to not drop a feature or delete any instances yet

b. Outliers

According to statistical theory, almost 99% of the value of a random variable is between Q1-1.5IQR(lower) and Q3+1.5IQR(upper), points outside this range are outliers and we need to deal with these points. I selected clips of outlier points about lower and upper.



Outliers before treating



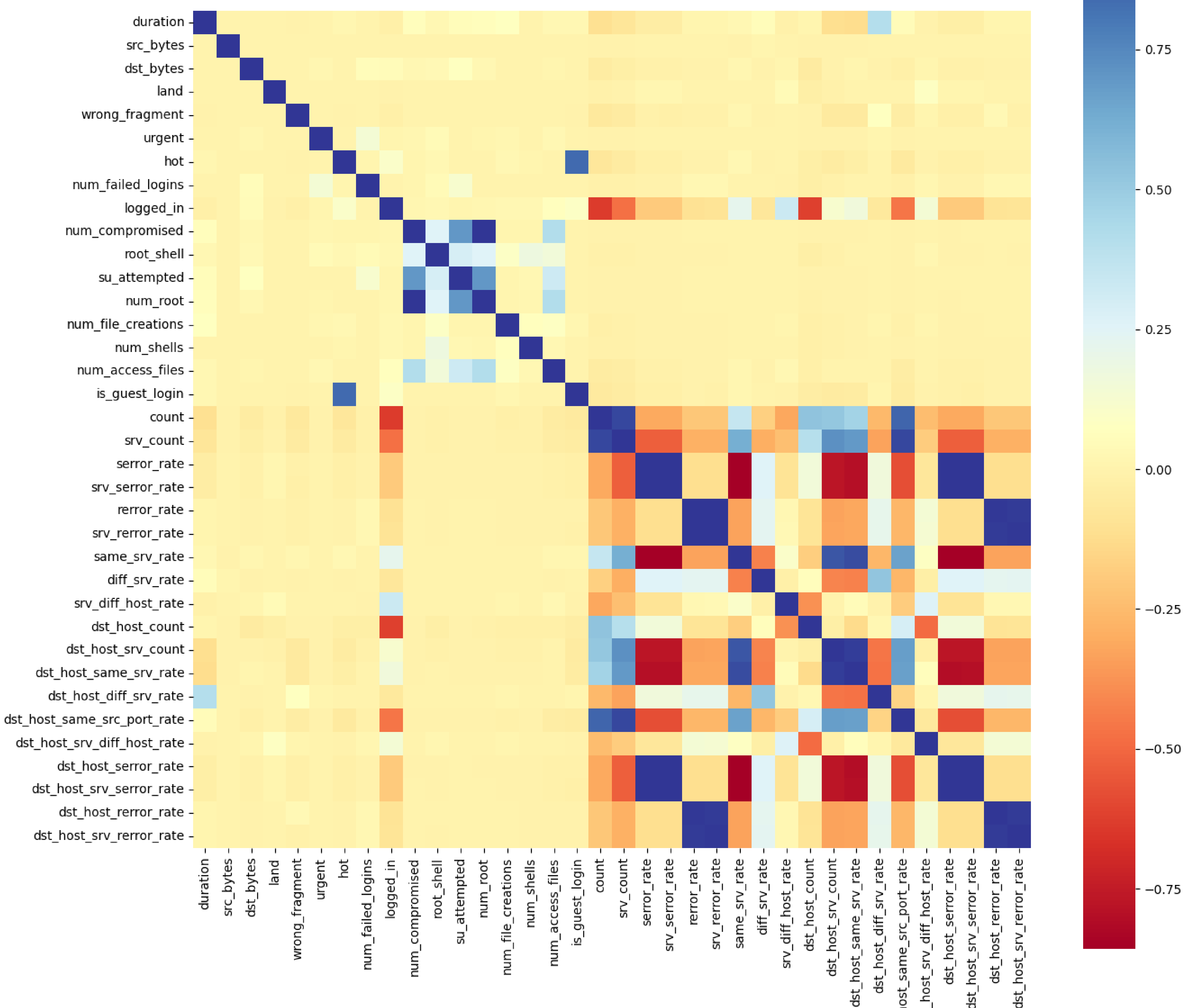
Outliers after treating

Data Cleaning

c. Redundant Variables

- Only one unique value
- Have high correlation

→ **Remove 9 features**



Variable Transformations

Encoder

- Create a feature target in binary form is "label_transform"
- encode "normal" = 1 and other type of attack = 0
- Use JamesSteinEncoder's theory directly on categorical variable and feature "label_transform"

For feature value i, estimator returns a weighted averange of:

- The mean target value for the observed feature value i
- The mean target value (regardless of the feature value)

JS_i = (1-B)*mean(y_i) + B*mean(y)

B = var(y_i) / (var(y_i)+var(y))

Variable Transformations

Normalizing and Scaling

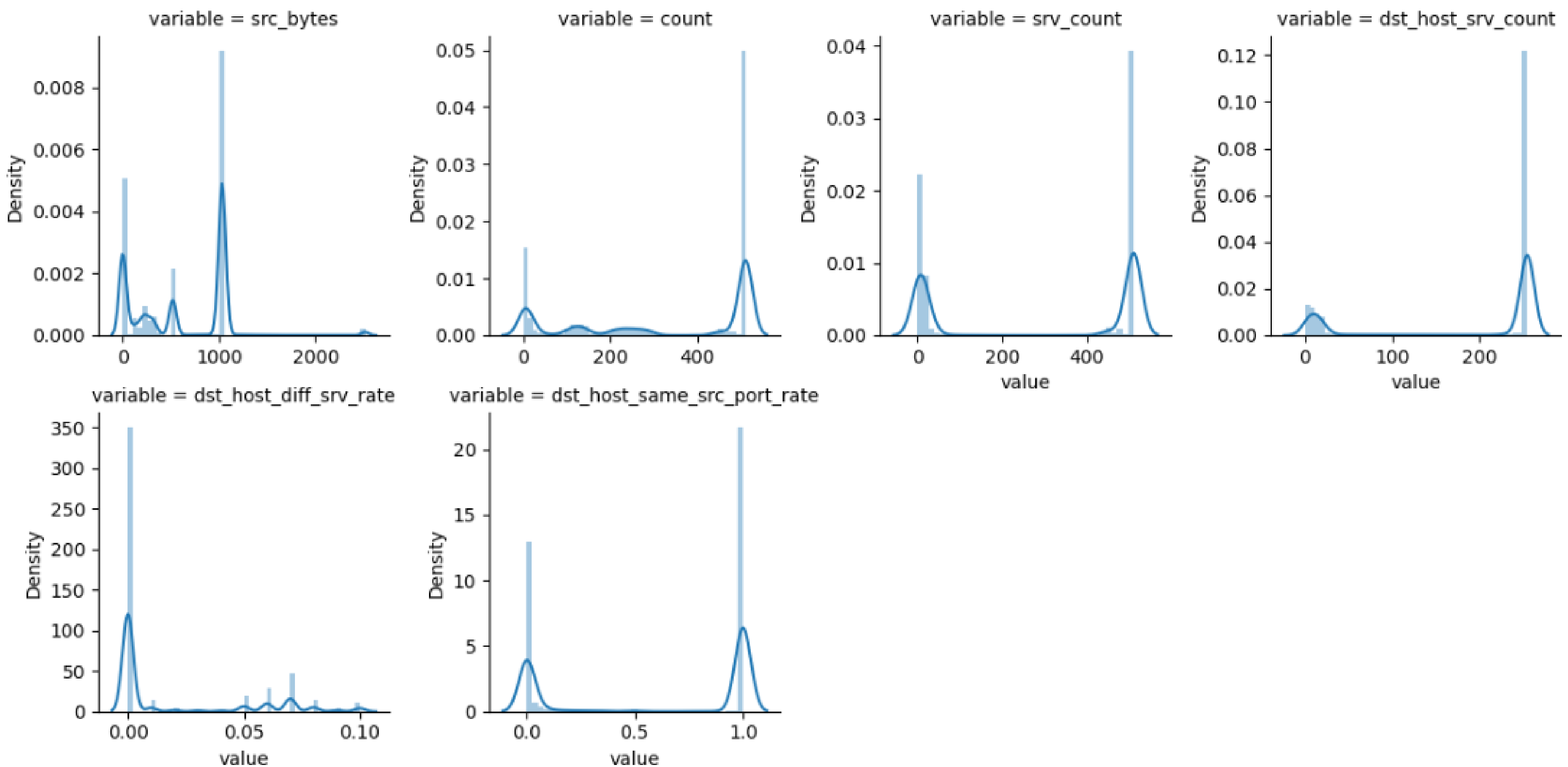
- There are many fields with quite large different scales (such as fields ranging from 0 -> 1 or hundred to thousand)
- Not only that the distribution of continuous variables don't fit with 1 normal distribution is skewed or can be seen by mixing 2 or more different normal distributions.
- From the above 2 reasons, we use Min-Max scale to normalize data

Scale value of different features(before scaling)

	protocol_type	service	flag	src_bytes	logged_in	root_shell	su_attempted	is_guest_login	count	srv_count	dst_host_srv_count
317921	0.004439	0.002657	0.223207	1032	0	0	0	0	511	511	255
171422	0.004439	0.002657	0.223207	1032	0	0	0	0	511	511	255
312181	0.004439	0.002657	0.223207	1032	0	0	0	0	511	511	255
87346	0.404179	0.825340	0.223207	345	1	0	0	0	6	6	255
57449	0.404179	0.102271	0.001055	0	0	0	0	0	260	2	2
...	...	...	...	...	...	...	...	...	...	...	...
367818	0.404179	0.102271	0.001055	0	0	0	0	0	128	11	7
82157	0.404179	0.825340	0.223207	303	1	0	0	0	15	19	255
26246	0.404179	0.825340	0.223207	306	1	0	0	0	10	10	255
303821	0.004439	0.002657	0.223207	1032	0	0	0	0	511	511	255
18458	0.404179	0.825340	0.223207	316	1	0	0	0	8	8	255

98805 rows × 13 columns

Distribution of some continuous variables(before scaling)



Model

III. Model

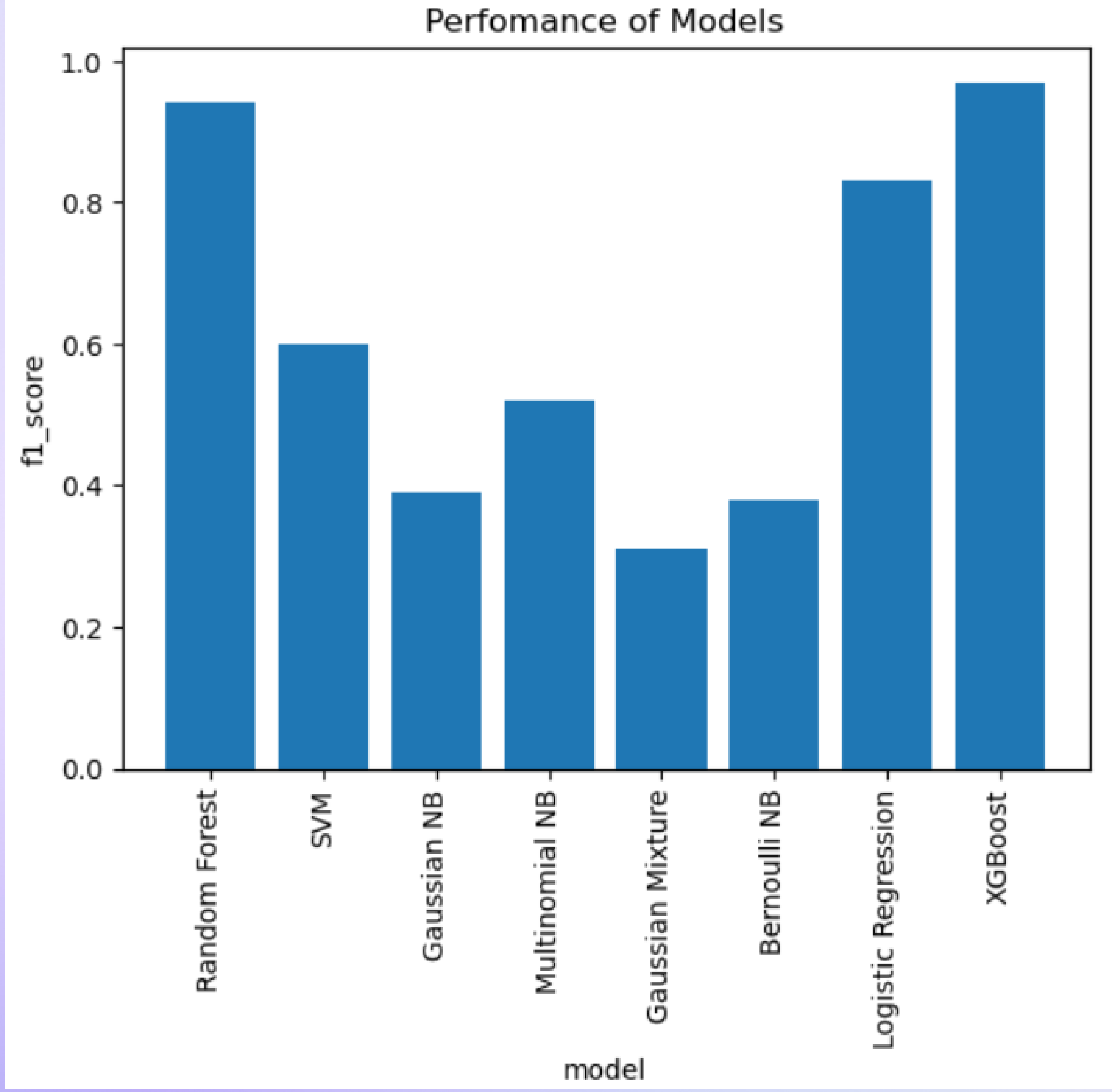
- **Probabilistic models**
    - Gaussian Naive Bayes
    - Multinomial Naive Bayes
    - Gaussian Mixture Model
    - Bernoulli Naive Bayes
- **Machine learning Model**
    - Logistic Regression
    - Support Vector Machine
    - XGBoost
    - Random Forest

Result

IV. Result

Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-score
Gaussian Naive Bayes	0.72	0.44	0.72	0.39
Multinomial Naive Bayes	0.97	0.66	0.51	0.52
Gaussian Mixture Model	0.68	0.40	0.52	0.31
Bernoulli Naive Bayes	0.93	0.61	0.36	0.38
Logistic Regression	0.99	0.90	0.79	0.83
Support Vector Machine	1.00	0.93	0.88	0.60
Random Forest	1.00	0.99	0.91	0.94
XGBoost	1.00	0.99	0.96	0.97

Result of different models in terms of different metrics



F1_Score of Different Models

Conclusion

V. Conclusion

- So far, we have successfully solved network attacks detection by using different techniques from exploratory data analysis to modeling and achieved the best result on the XGBoost which has overall accuracy 100%, macro average f1_score 97% using the KDD Cup 99 dataset.

- Although the dataset has been considered as a benchmark for network attacks detection problems for decades, it is quite outdated as it was released in the year 1999. It is also a suitable reason why we could achieve such a surprising result with some state-of-the-art techniques and models.

- For future development, we would like to use a problem-related dataset involving time series as now the network attacks are hardly spotted by not using the dependence on time.

References

[1] KDD Cup 1999 Data; 1999. Available from: http://kdd.ics.uci.edu/databases/kddcup99/ kddcup99.html.

[2] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. 2009:1-6.

[3] The Importance of Data Preparation for Business Analytics; 2019. Available from: https://www. ironsidegroup.com/2019/07/16/data-preparation-business-analytics/.

[4] Tour of Evaluation Metrics for Imbalanced Classification; 2020. Available from: https:// machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/.

[5] Failure of Classification Accuracy for Imbalanced Class Distributions; 2020. Available from: https:// machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/.

[6] Scikit-learn;. Available from: https://scikit-learn.org/stable/.

Thank You