Applied Statistics and Experimental Design

Network Attacks Detection

Group 7 - DSAI K65 - HUST

July 22, 2022

Our Team Members

Nguyen Quang Duc

20204876

Le Hong Duc

20204874

Tran Hoang Quoc Anh

20200044

La Dai Lam

20204918

Luu Trong Nghia

20204888

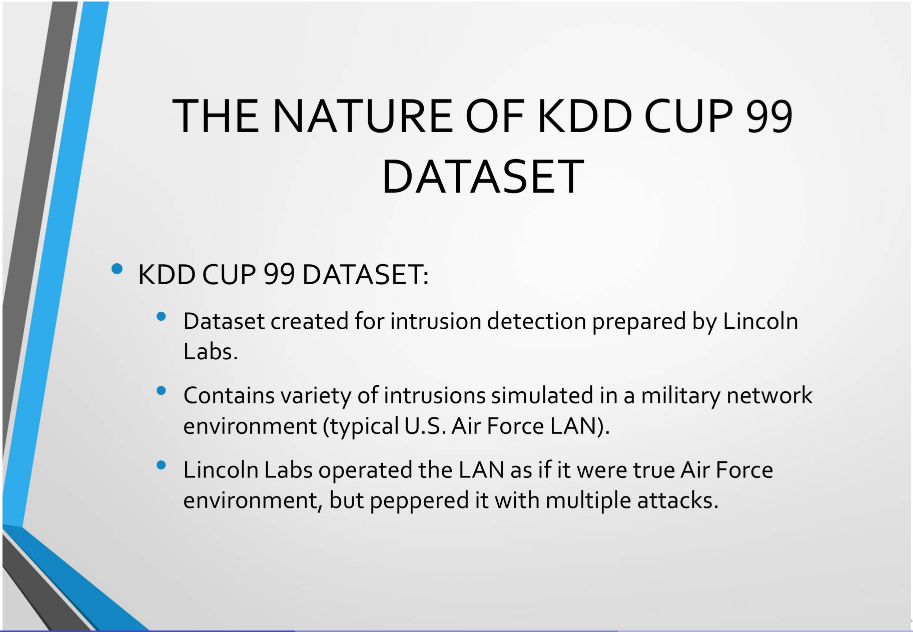
Table of contents

- Introduction
- Datasets
- EDA
- Data preparation
- Modelling
- Practical results

Introduction

- Internet: a global system of interconnected computer networks
- Can be attacked by DDOS, Website Defacement, Directory Traversal, etc
- Build software to detect network attacks protect a computer network

Datasets



THE NATURE OF KDD CUP 99 DATASET

- KDD CUP 99 DATASET:
    - Dataset created for intrusion detection prepared by Lincoln Labs.
    - Contains variety of intrusions simulated in a military network environment (typical U.S. Air Force LAN).
    - Lincoln Labs operated the LAN as if it were true Air Force environment, but peppered it with multiple attacks.

Exploratory data analysis (EDA): Univariate Analysis

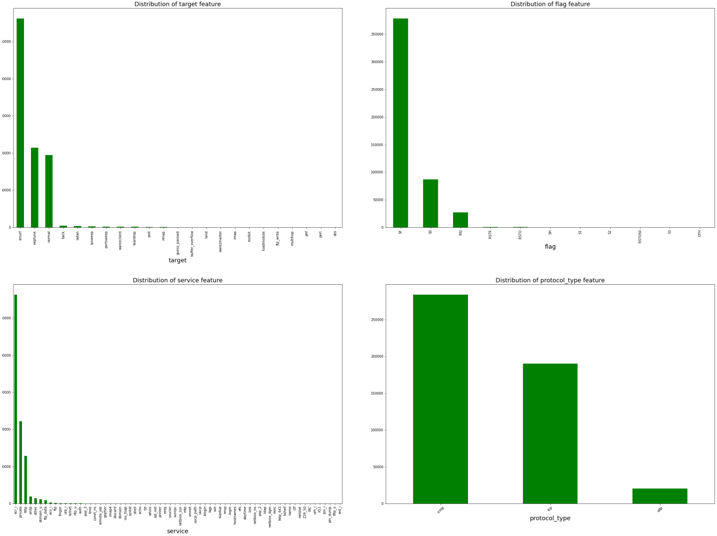


Figure 1. Distribution of categorical features

Exploratory data analysis (EDA): Univariate Analysis

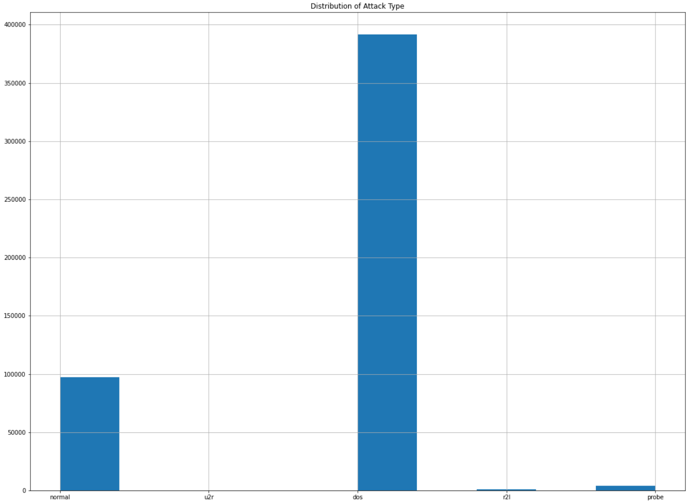


Figure 2. Distribution of target feature - ‘Attack Type’

Exploratory data analysis (EDA): Multivariate Analysis

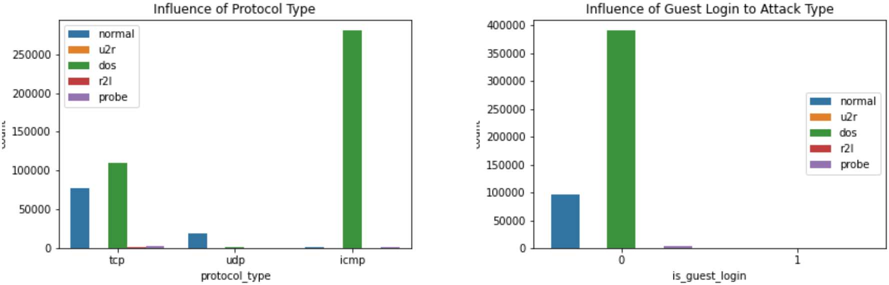


Figure 3. Influence of protocol type and attack type towards dependent variable

Exploratory data analysis (EDA): Multivariate Analysis

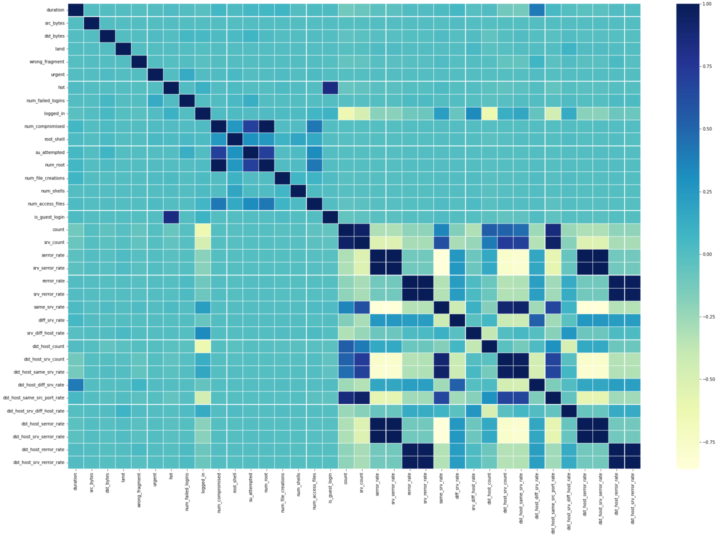


Figure 4. The heatmap representing correlation of independent variables

Data Preparation

In this part, we handle on 3 steps:

- Data Cleaning: Check whether there is a missing value ⇒ no null values ⇒ not drop a feature or delete any instances
- Redundant Variables: Remove variables have 1 unique ,value or have a high correlation value with others
- Variable Transformations: Transform text and categorical to numeric values. We use label encoder for categorical and standard scalar (standardize) for numeric values

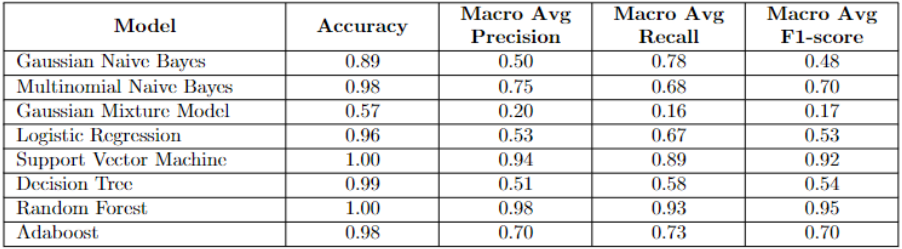
Modelling

- Probabilistic models
    - Gaussian Naive Bayes
    - Multinomial Naive Bayes
    - Gaussian Mixture Model
- Other Machine Learning Models
    - Logistic Regression
    - Support Vector Machine
    - Decision Tree
    - Random Forest
    - AdaBoost

Practical results

Evaluation metrics

- Accuracy score
- Precision score
- Recall score
- F1-score
- Macro average accuracy/precision/recall score



Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-score
Gaussian Naive Bayes	0.89	0.50	0.78	0.48
Multinomial Naive Bayes	0.98	0.75	0.68	0.70
Gaussian Mixture Model	0.57	0.20	0.16	0.17
Logistic Regression	0.96	0.53	0.67	0.53
Support Vector Machine	1.00	0.94	0.89	0.92
Decision Tree	0.99	0.51	0.58	0.54
Random Forest	1.00	0.98	0.93	0.95
Adaboost	0.98	0.70	0.73	0.70

Table 1: Result of different models in terms of different metrics

Conclusion

Summary

- The best result is in the Randon Forest model which: overall accuracy 100%, average recall 93%
- The dataset is quite outdated so we could achieve such a surprising result with some state-of-the-art techniques and models

Future development

- Nowadays the network attacks are hardly spotted by not using the dependence on time
- Use a problem-related dataset involving time series