Securing Indonesia's Telecommunications: A Machine Learning Approach to Network Attack Classification

Multiclass Classification using Light Gredient Booster Machine

The rapid evolution of information technology has propelled the telecommunications sector in Indonesia, fostering connectivity, information accessibility, and community empowerment. With the government's initiatives aligning with the vision of "Indonesia Emas 2045," advancements in internet connectivity and data transfer speed are transforming global communication.

1 Data preprocessing

Data Conversion

Convert the data type from object to float so that null values can be identified

Missing Value Handling

Handle missing values with mean for each target class, then drop or remove the data that cannot be filled by mean.

Data Transformation

For numeric data, Transformation is done using min-max scaler so that the data are range between 0 to 1 and for categorical data, transformation are done using one hot encoding.

Numeric Data

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

(Min-Max Scaler)

Before Scaling	After Scaling
min(duration) = 0.00	min(duration) = 0.00
max(duration) = 99999	max(duration) = 1.00

Categoric Data

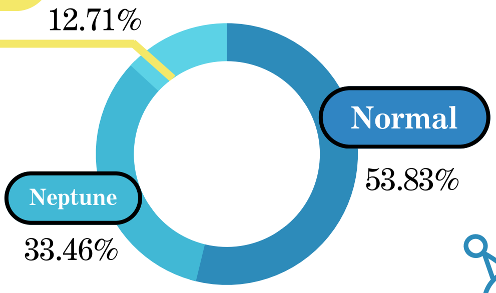
service
private
http
smtp

service_private	service_http	service_smtp
1	0	0
0	1	0
0	0	1

(One-Hot Encoder)

2 Exploratory Data Analysis

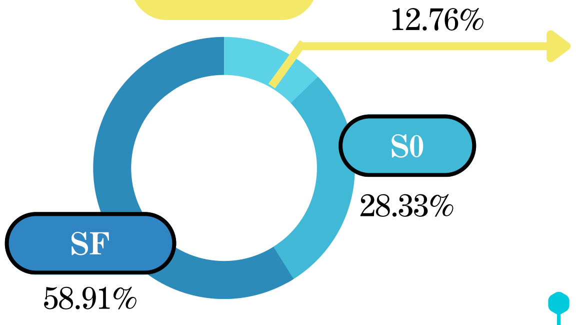
Type of Attack



Others

- Ipsweep
- Satan
- Portsweep
- Smurf
- Nmap
- Denial of Service Attack

Flag



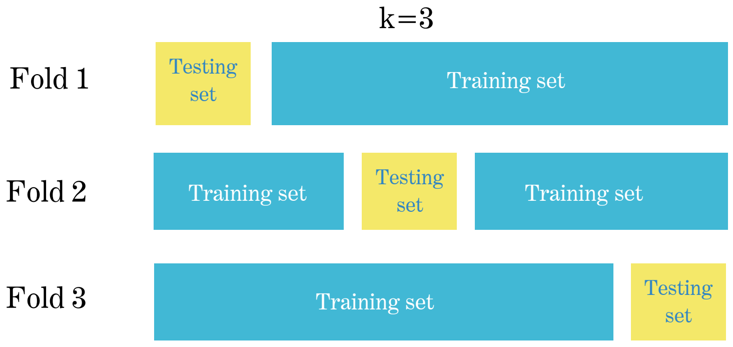
Others

- SH
- S1
- S2
- S3
- OTH

- REJ
- RSTO
- RSTOS0
- RSTR

3 Dataset Splitting

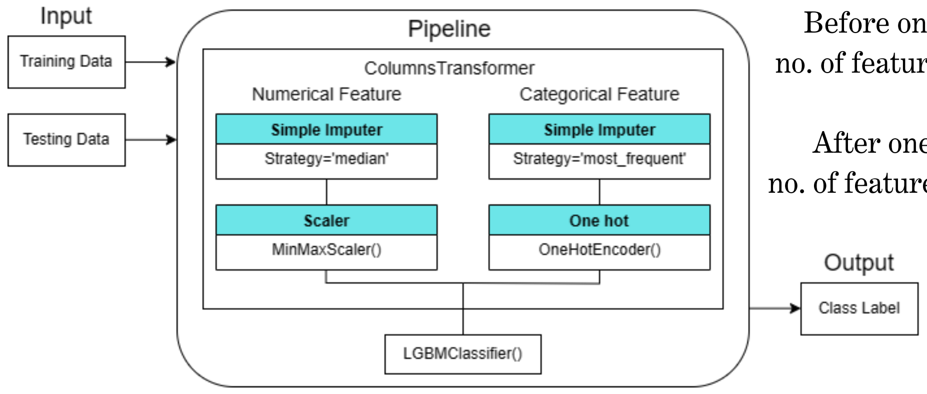
Before training the ML model, data were split into training and testing with ratio of 80% and 20% and also using 3-fold cross validation



4 Model Training

LGBM model shown to perform very well in multiclass classification, especially with al lot of categorical features. In this scenario the model were contruct inside scikit-pipeline

Pipeline



Before onehot
no. of feature = 41

After onehot
no. of feature = 121

Hyperparamater tuned

n_estimators:	100, 200
reg_alpha:	0.0, 0.1
reg_lambda:	0.0, 0.1

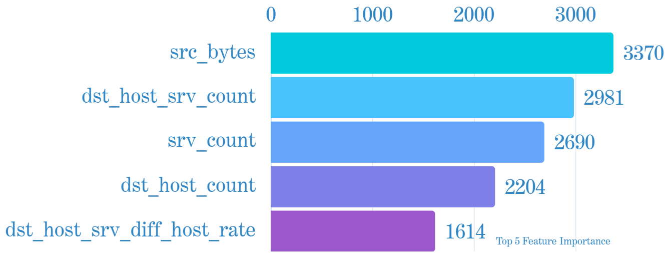
■ best_params

F-1 Score

99.58%	99.96%	99.86%
CV Score	Training Score	testing Score

5 Feature Importance

Top 5 Feature Importance



Least Important Feature

- service_urh_i
- service_tim_i
- serive_tftp_u
- service_systat

and more, with total of 61 features

Conclusion

Light Gradient Boosting Machine can effectively predict a data accurately with of **0.995.**
From Feature Importance, we can conclude that the most influential feature is **src_bytes.** And also we can conclude that there are **61 features** not having contribution to the model (getting feature_impotance score equal 0), which mostly are from **onehotencoded "service" columns.**

Recomendation

Managing the network services used by users, by limiting services that could be sources of attacks, network provider could also consider the used type of service and protocol.
In order to get better results, it is recommended to try changing number of k fold, number of parameter tuned, or even trying another Machine learning model.

DAC-01-0117 | Masih Pemula