Name: Tan Nguyen

CS 422

INTRODUCTION TO MACHINE LEAERNING

FALL 2023

ASSIGNMENT 3

Machine Learning Report: Malware Detection using Naïve Bayes Gaussian on Network Traffic Dataset

Dataset Source:

Stratosphere Laboratory. A labeled dataset with malicious and benign IoT network traffic. January 22th. Agustin Parmisano, Sebastian Garcia, Maria Jose Erquiaga (Source: Kaggle)

About the Dataset: The dataset provides comprehensive insight into the realm of network malware, offering detailed labels that shed light on the relationship between network flows and malicious activities. The dataset comprises approximately 1 million records. All of the non-integer input features have been removed from the original dataset to reduce size and fit the Gaussian model.

Field name	Description	Type
duration	The duration of the connection.	Int
orig_bytes	The number of bytes sent from the source to the destination.	Int
resp_bytes	The number of bytes sent from the destination to the source.	Int
local_orig	Indicates whether the connection is considered local or not.	Bool
local_resp	Indicates whether the connection is considered local or not.	Bool
missed_bytes	The number of missed bytes in the connection.	Int
orig_pkts	The number of packets sent from the source to the destination.	Int
orig_ip_bytes	The number of IP bytes sent from the source to the destination.	Int
resp_pkts	The number of packets sent from the destination to the source.	Int
resp_ip_bytes	The number of IP bytes sent from the destination to the source.	Int
label	A label associated with the connection (e.g., 'Malicious' or 'Benign').	String

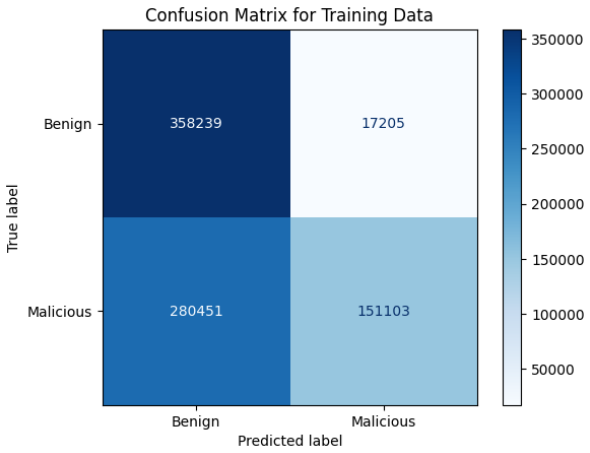
Data Processing:

- The dataset was split into an 80/20 ratio for training and testing, respectively.
- Labels (y) were encoded to make them compatible for the log_loss function.

Evaluation Metrics:

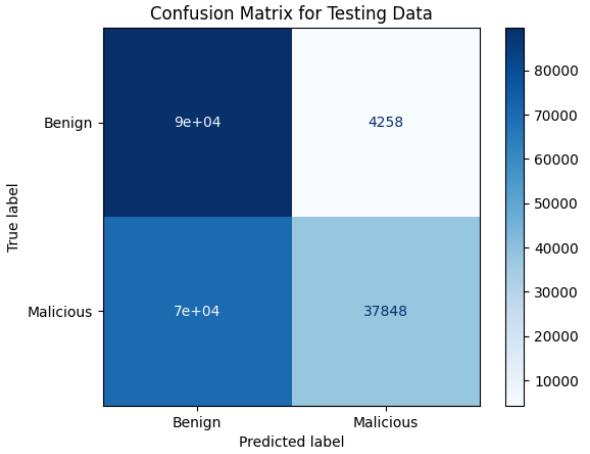
Training Data:

- **Accuracy**: 0.63
- **Precision**: 0.90
- **Recall**: 0.35
- **F1 Score**: 0.50
- **Log Loss**: 3.64
- **Selectivity**: 0.95



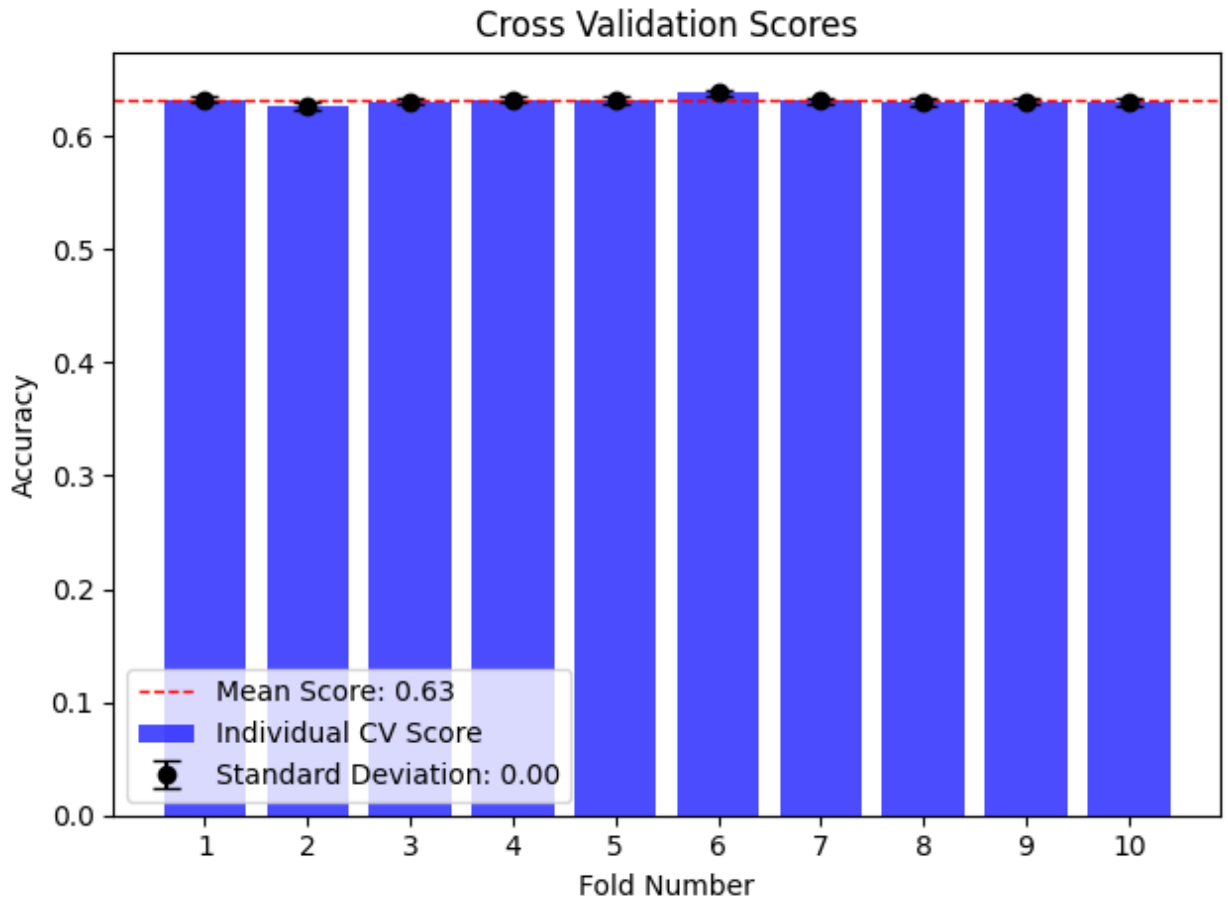
Testing Data:

- **Accuracy**: 0.63
- **Precision**: 0.90
- **Recall**: 0.35
- **F1 Score**: 0.50
- **Log Loss**: 3.64
- **Selectivity**: 0.95



Cross Validation

Cross validation technique was experimented, and it gave the same accuracy rate across the iterations with standard deviation of 0.



Conclusion

The large size of the dataset ensured consistency in the performance of the GaussianNB() model for both the training and testing sets

For broader implications, the metrics of Precision and Recall are more important. High precision ensures the detected threats are indeed true threats. While high recall rate implies that the majority of threats are detected. Given the potential harm of malicious access, false negative rate should be minimized as much as possible.

To improve the model's performance, there are a few suggestions:

- Add more input features
- Change model
- Ensemble with other model