Detection of DNS attacks using Hidden Markov Model

Student Researcher: Balaji Bharatwaj M (CB.EN.U4CSE16607), Aditya Reddy M (CB.EN.U4CSE17431)
Research Supervisor: Dr. Senthil Kumar T (Associate Professor)
Lead: Sulakshan Vajipayajula, Architect - CTO, IBM security, Bangalore

Keywords

- DDoS - Distributed Denial of Service
- DoS - Denial of Service
- HMM - Hidden Markov Model
- KNN - K-Nearest Neighbours

Problem Statement

The aim of this project is to detect attacks based on DNS (Ex. DDoS, DoS etc.) using HMM. The results shows how Hidden Markov Model performs with other clustering/classification algorithm like Logistic Regression and KNN.

Dataset Features

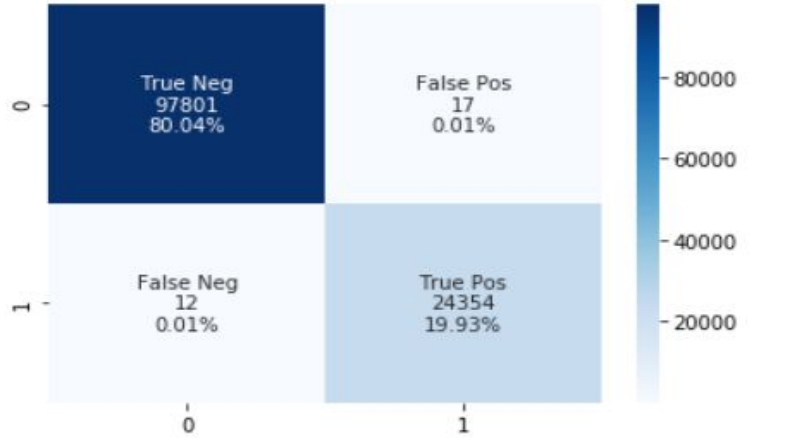
feature name	*description*	*type*
duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol, e.g. tcp, udp, etc.	discrete
service	network service on the destination, e.g., http, telnet, etc.	discrete
src_bytes	number of data bytes from source to destination	continuous
dst_bytes	number of data bytes from destination to source	continuous
flag	normal or error status of the connection	discrete
land	1 if connection is from/to the same host/port; 0 otherwise	discrete
wrong_fragment	number of ``wrong'' fragments	continuous
urgent	number of urgent packets	continuous

Performance with Logistic Regression

```
Number of fraud connections : 396743 Number of fraud prediction connections : 392258
Number of normal connections : 97278 Number of normal prediction connections : 95501
accuracy on fraud connection: 392258 / 396743 = 0.9886954527238035
accuracy on normal connection: 95501 / 97278 = 0.9817327658874566
TP - True Negative 392258
FP - False Positive 4485
FN - False Negative 1777
TP - True Positive 95501
Accuracy Rate: 0.9873244254798885
Misclassification Rate: 0.012675574520111494
```

Performance with KNN

```
TP - True Negative 97801
FP - False Positive 17
FN - False Negative 12
TP - True Positive 24354
Accuracy Rate: 0.9997626530478623
Misclassification Rate: 0.00023734695213775944
```



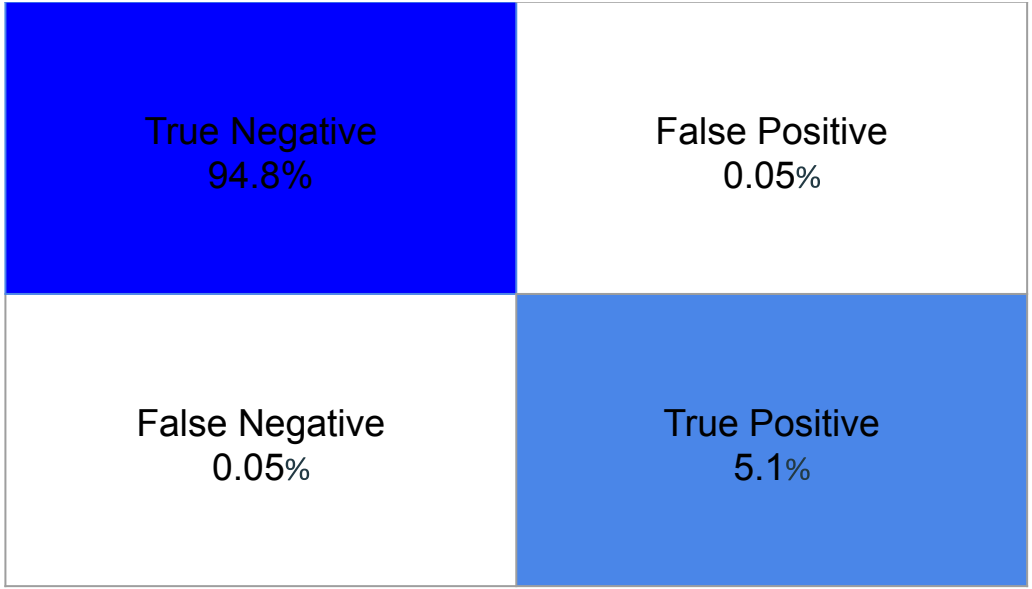
Performance with HMM

```
In [55]: print("The True Positive Rate for the predicted dataset is : ", (totalTwoPred / (totalZeroPred + totalTwoPred)) * 100)
```

```
The True Positive Rate for the predicted dataset is :  5.189210044047198
```

```
In [56]: print("The True Negative Rate for the predicted dataset is : ", (totalZeroPred / (totalZeroPred + totalTwoPred)) * 100)
```

```
The True Negative Rate for the predicted dataset is :  94.8107899559528
```



Final Conclusion

It’s evident that the Hidden Markov Model (HMM) performs better than the Logistic Regression Algorithm, and KNN Algorithm. The True Negative Rate (Negative value in the dataset and the predicted data is also negative) is 94.8% whereas the Logistic Regression and KNN has a True Negative Rate of 79.4% and 80.04% respectively.