Group 12

組員：309551*** 朱〇毅
309551*** 王〇憲
309551*** 張玟棋
309553*** 陳〇安

Outline

- 問題描述
- 資料集說明
- 分析流程
- 分析結果
- 討論

問題描述

DDoS

- DoS：一種嘗試對目標系統（如網站、應用程式等）進行惡意攻擊的行為，通常，攻擊者會產生大量的封包或請求，最終使得目標系統無法負荷
- DDoS：第一個D意指「分散式」，即攻擊者會使用多個盜用或受控的來源來產生DoS攻擊

如何防禦DDoS

1. 預留頻寬或系統資源，並在達門檻值時對流量進行整體的限制
   缺點：會影響正常的使用者

2. 找出有問題的IP，並加以過濾
   問題：要如何找出來？

資料集説明

資料來源

- 來源網站：Kaggle
- 資料集：DDoS Dataset
  - https://www.kaggle.com/devendra416/ddos-datasets

資料集簡述

- 資料大小：3.84GB
- column數：85，最後1個column為label
  - Source/Destination IP/Port、Protocol、Total Forward Packet…
  - 除了IP和Timestamp之外皆為純數字

分析流程

Preprocessing	Analysis	Model Training	Test & Evaluation

• Missing value
• Outliers
• One-hot encoding
• Timestamp
• Normalization
	• Correlation matrix
	• Feature selection
		• Naive Bayes
		• Logistic Regression
		• SVM
		• Random Forest
			• Confusion matrix
			  ○ Accuracy
			  ○ Precision
			  ○ Recall
			  ○ F1-score

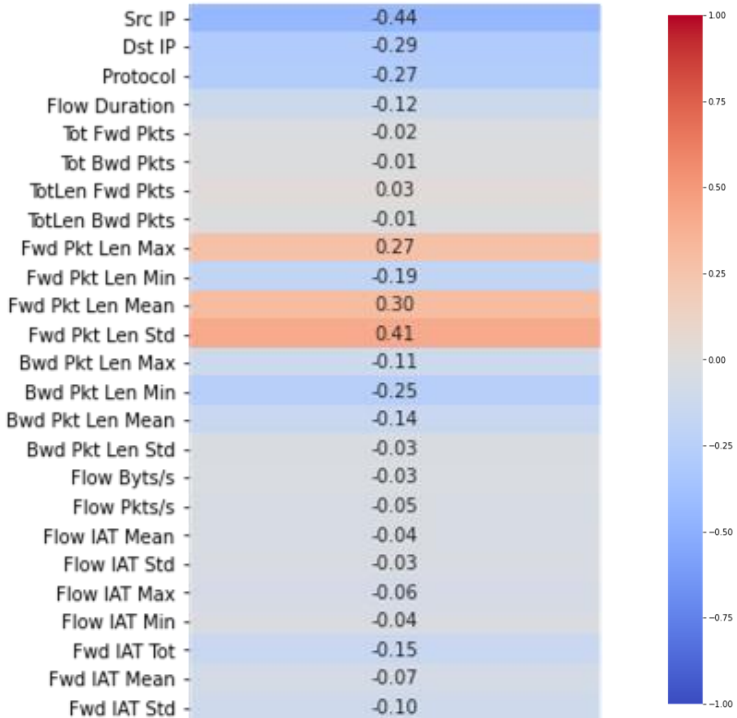
Preprocessing | Analysis | Model Training | Test & Evaluation

- **Missing value**
- **Timestamp** 切割
- **IP** 數值化
- **port** 作 one-hot encoding
- **Normalization**

Preprocessing | **Analysis** | Model Training | Test & Evaluation

- **Correlation matrix** 分析
- 選擇相關係數高的作為 **feature**

Src IP	Dst IP	Protocol
Fwd Pkt Len Max	Fwd Pkt Len Mean	Fwd Pkt Len Std
Bwd Pkt Len Min	Pkt Len Min	PSH Flag Cnt
ACK Flag Cnt	CWE Flag Count	Fwd Seg Size Avg
Fwd Seg Size Min	tHour	Src Port_80
Src Port_1025	Dst Port_53	Dst Port_80



Preprocessing | Analysis | **Model Training** | Test & Evaluation

- **Train : Test = 8 : 2**
- 根據文獻，選擇效果較好的 model 來訓練
  - **Logistic Regression**
  - **SVM**
  - **Random Forest**
- 另取 **Naive Bayes** 作為對照組

Preprocessing | Analysis | Model Training | **Test & Evaluation**

- **Confusion matrix**
  - Accuracy
  - Precision
  - Recall
  - F1-score

如何應對3V的問題

- VOLUME – 資料量
  - 先以1/1000的資料當作input
  - 結果高於目標時才改為使用全部的資料
- VARIETY – 資料多元性
  - 取相關係數高的feature，將84個feature降成18個feature
- VELOCITY – 資料即時性
  - 固定資料集無即時性

分析結果

Naive Bayes

- 訓練model的時間：43分7秒

confusion matrix：

TP：242971	FP：15775
FN：122217	TN：1132225

testing result：

```
Summary Stats
Accuracy =  0.9087044288379122
Precision =  0.9377000262836469
Recall =  0.6652133855417722
F1 Score =  0.7782960757641865
```

SVM

- 訓練model的時間：5分鐘29秒

confusion matrix：

TP：119810	FP：13517
FN：9276	TN：615197

testing result：

```
Summary Stats
Accuracy = 0.9699221430456585
Precision = 0.8986176843400061
Recall = 0.9281409293029453
F1 Score = 0.9131407361677966
```

Logistic Regression

- 訓練model的時間：8分鐘24秒

confusion matrix：

TP：256418	FP：2906
FN：4293	TN：1249696

testing result：

```
Summary Stats
Accuracy =  0.9952428876247016
Precision =  0.9887939411701192
Recall =  0.9835334911070112
F1 Score =  0.9861567009912794
```

Random Forest

- 訓練model的時間：39分鐘45秒

confusion matrix：

TP：255431	FP：3096
FN：26	TN：1254134

testing result：

```
Summary Stats
Accuracy =  0.9979361229388499
Precision =  0.9880244616616446
Recall =  0.9998982216185112
F1 Score =  0.9939258809612751
```

總比較

	NB	SVM	LR	RF	目標
Accuracy	0.909	0.970	0.995	0.998	0.85
Precision	0.939	0.899	0.988	0.988	0.80
Recall	0.665	0.928	0.983	0.999	0.85
F1-score	0.779	0.913	0.986	0.994	X

討論

如何防禦DDoS

- 利用model進行預測
- 在流量達門檻值時進行預測

資料集

- 是否特徵過於明顯？
  - 證據：根據文獻，Naive Bayes對DDoS問題的處理結果很差

Q & A