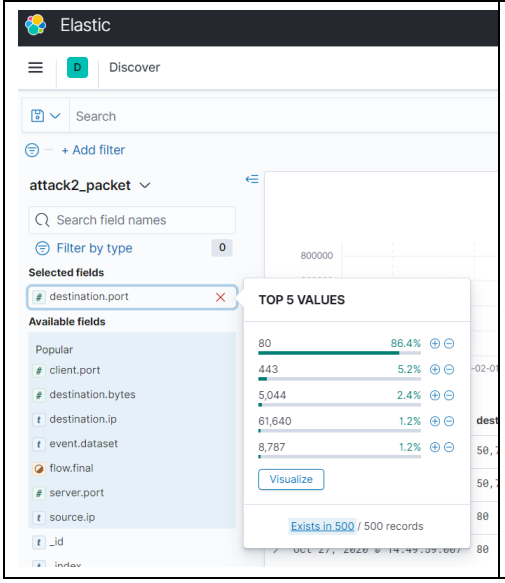
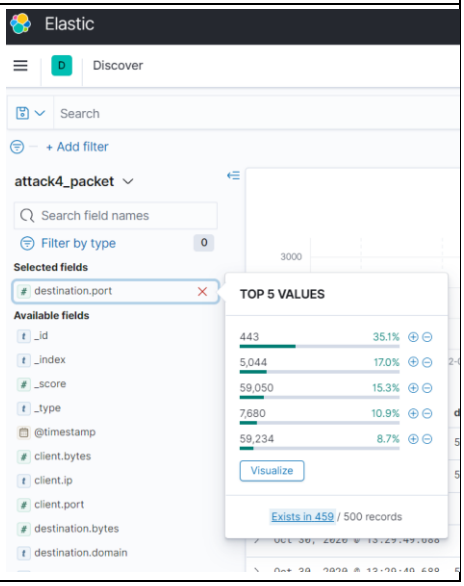
Project2

0616236 趙秉濂

I. Feature select
   A. 先判斷哪些特徵對預測有幫助，觀察 Attack1-5 應該分別屬於哪些攻擊，以及這些攻擊有哪些特徵
   B. 攻擊特徵分析(使用 ELK 觀察資料)
      1. Port Scan
         a. Port 會大量分散
         b. 封包大小集中(因為是由電腦所產生的格式化封包)
      2. SQL Injection
         a. Query 中帶有 SQL 與法的字串(SELECT/FROM)等。
         b. 如果有偷資料出去，其封包大小可能偏大。
      3. Brute-Force attack
         a. Query 中帶有眾多不同的 username/password
      4. DDoS
         a. 封包集中在 port 80。
      5. Phishing Email
         a. 遠端登入啟動 cmd.exe 的情況
   C. 使用 google 提供的 ELK 來上傳資料觀察資料分布的狀況，來尋找線索(如圖是觀察 attack2、4 的 port 分布狀況，雖然是抽樣的狀況，但仍可以參考)，並依照攻擊特徵分析，判斷 Attack 1-5 對應的攻擊。

Attack2-destination.port	Attack4-destination.port
	

D. 依照資料觀察，Attack1-5 對應的攻擊如下
    1. Attack1：Brute-Force
    2. Attack2：DDos
    3. Attack3：Port scan
    4. Attack4：Phishing Email
    5. Attack5：SQL Injection
E. 取樣特徵：依照以上的觀察，決定計算這些資料集的某些特徵來做為需要的資料。
    1. 取樣方式
        a. 計算某欄的某個值的出現頻率，例如：destination.port 是 80 的機率在 DDos 的情況下就會比較高
        b. 計算某欄位的分散情況，參考 entropy，使用以下公式：

$$\sum_{i=1}^{n} p(x_i)^2$$

        P(x)表示 x 值在此欄位出現的機率(即是上面的 a)，此公式在值很集中的時候會越接近 1，而越分散則會越接近 0。
    2. 取樣欄位
        a. packetbeat
            1. Login_query_ratio：url.query 出現 username/password 等字樣的機率。
            2. select_query_ratio：url.query 出現 select/from 等字樣的機率。
            3. dest_byte_avg：封包平均大小
            4. dest_port_80_ratio：destination.port 是 80 的比例
            5. dest_port_dis：使用 1-b 方法計算 destionation.port 的分散程度
        b. winlogbeat
            1. event_out_success：event.outcome 是 success 的比例。
            2. process_name_cmd：cmd.exe 執行的比例
            3. process_name_xampp：xampp-control.exe 執行的比例

II. Model and algorithm
A. 使用 Random forest 做訓練：多個 Decision Tree 的投票結果
    1. 使用預設的 100 個 tree
    2. Bootstrap 抽樣取消，每一個樹都使用整個 dataset 來做
B. Preprocessing
    1. 一開始由於 packetbeat 會使 memory 不夠用的原因，如果超過 1 萬行就抽樣 1 萬行來做訓練。
        a. 此 1 萬行以外會多抽樣最多 0.5 萬包含有 SQL、Login 樣式的行
        b. 取樣訓練集的資料如下圖

	A	B	C	D	E	F	G	H	I	J
1	attacktype	login_query_ratio	select_query_ratio	dest_byte_avg	dest_port_80_ratio	dest_port_dis	event_out_success	process_name_cmd	process_name_xampp	
2	attack 1	0.340577295	6.67E-05	8700.976488	0.976734884	0.954228	0.612687067	0	0.386642841	
3	attack 2	0	0	18241.51705	0.8252	0.681101	0.760531258	0	0.239334585	
4	attack 3	0	0	2661.887357	0.0031	0.000744	0.999508438	0	0	
5	attack 4	0	0	637391.4246	0.007438017	0.229002	0.999420849	0.016023166	0	
6	attack 5	0.000280899	0.07228464	26069.7086	0.544101124	0.386932	0.646092145	0	0.34559327	
7										

2. 之後為了增加訓練資料量，抽 2000 行，然後抽樣 10 次來增加訓練資料，結果如下圖。最後訓練是抽 2000 行，抽樣每個檔案 40 次來訓練。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	attacktype	dest_byte_avg	dest_byte_var	dest_ip_d	dest_port	dest_port	dest_port	login_que	select_qu	event_coc	event_out	process_r	process_r	process_n
2	attack 1	11578.09913	43185794481	0.95652	0.014	0.97734	0.9554	0.34055	0	0.3437	0.61269	0.39748	0	0.38664
3	attack 1	6825.016027	5506115982	0.95132	0.01699	0.97368	0.94834	0.34255	0	0.3437	0.61269	0.39748	0	0.38664
4	attack 1	16464.01403	90117480667	0.95652	0.014	0.97667	0.9541	0.33755	0	0.3437	0.61269	0.39748	0	0.38664
5	attack 1	4900.476302	73107692.03	0.95978	0.01466	0.97801	0.95672	0.33922	0	0.3437	0.61269	0.39748	0	0.38664
6	attack 1	13893.755	76055262858	0.96174	0.015	0.97934	0.95934	0.34155	0	0.3437	0.61269	0.39748	0	0.38664
7	attack 1	13874.27094	75369583233	0.95847	0.014	0.97701	0.95475	0.33955	0	0.3437	0.61269	0.39748	0	0.38664
8	attack 1	14202.6481	76750459831	0.96435	0.01266	0.98001	0.96058	0.34222	0	0.3437	0.61269	0.39748	0	0.38664
9	attack 1	8059.008005	25102616663	0.96043	0.01533	0.97801	0.95674	0.34122	0	0.3437	0.61269	0.39748	0	0.38664
10	attack 1	6248.810874	1274430269	0.95652	0.01599	0.97667	0.95416	0.34122	0	0.3437	0.61269	0.39748	0	0.38664
11	attack 1	15604.13642	79123450830	0.95197	0.01866	0.97468	0.95035	0.34222	0	0.3437	0.61269	0.39748	0	0.38664
12	attack 2	17179.36862	16597379155	0.71673	0.0085	0.8295	0.68826	0	0	0.29011	0.76053	0.54856	0	0.23933
13	attack 2	20964.71481	25084614715	0.69967	0.01	0.8165	0.66691	0	0	0.29011	0.76053	0.54856	0	0.23933
14	attack 2	16897.37598	13225781185	0.70459	0.0075	0.822	0.67587	0	0	0.29011	0.76053	0.54856	0	0.23933
15	attack 2	21326.45016	19443774848	0.70688	0.0065	0.8215	0.67503	0	0	0.29011	0.76053	0.54856	0	0.23933
16	attack 2	23860.91286	22978530407	0.69673	0.0135	0.8155	0.66535	0	0	0.29011	0.76053	0.54856	0	0.23933
17	attack 2	28623.60063	37370603889	0.70369	0.0105	0.823	0.67759	0	0	0.29011	0.76053	0.54856	0	0.23933
18	attack 2	19471.69186	20556872537	0.67879	0.0105	0.7995	0.63949	0	0	0.29011	0.76053	0.54856	0	0.23933
19	attack 2	18290.79832	13761464490	0.70806	0.0085	0.823	0.67753	0	0	0.29011	0.76053	0.54856	0	0.23933
20	attack 2	19154.85193	16603972423	0.71462	0.008	0.827	0.68413	0	0	0.29011	0.76053	0.54856	0	0.23933
21	attack 2	19140.25501	21653902765	0.69364	0.009	0.811	0.65795	0	0	0.29011	0.76053	0.54856	0	0.23933
22	attack 3	1270.618031	205589778.7	0.88772	0.019	0.005	0.00122	0	0	0.25425	0.99951	0.28303	0	0
23	attack 3	1711.602017	541379179.8	0.87839	0.026	0.005	0.00156	0	0	0.25425	0.99951	0.28303	0	0
24	attack 3	16066.76432	1.37434E+11	0.89634	0.016	0.0025	0.00125	0	0	0.25425	0.99951	0.28303	0	0

III. Problem encounter

A. Entropy

1. 原本選擇要觀察的欄位是使用 entropy 資料量來做判斷，公式如下：

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

結果留下來的是與攻擊較無關的欄位，如：host 資訊與時間資訊。

2. 因此從頭判斷攻擊可能產生的特徵來選用觀察的欄位與資訊。

B. File size too large

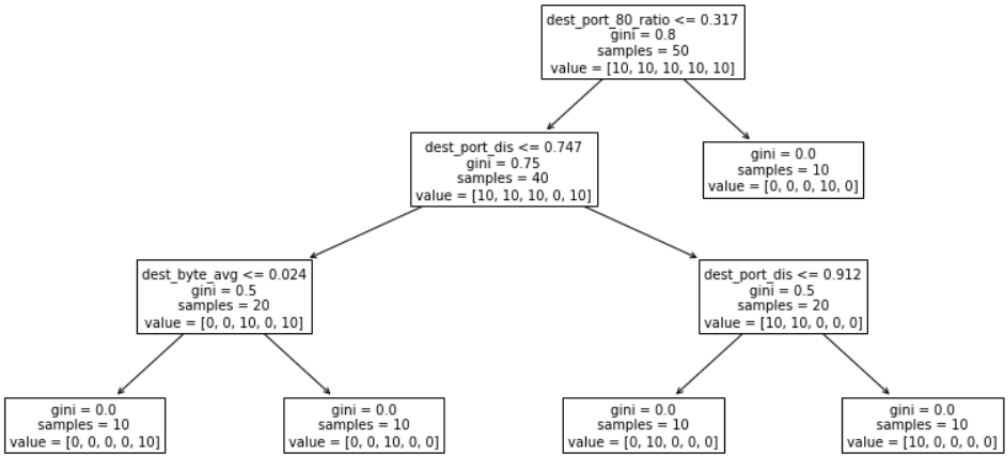
1. Packet beat 檔案過大，時間訓練困難
2. 用抽樣的方式，一方面可以減少一次記憶體所需的量，一方面可以增加訓練的資料。

C. Model judge make no scense

會依據一些沒有根據的 feature 來做判斷，因此 drop 掉一些沒有幫助於判斷的 feature，如下：

a. Dest_port_443_ratio：destination.port 是 443 的比例
b. Dest_ip_dis：計算 destination.ip 的分散程度
c. process_name_browser：使用 chorme、explore、firefox 等瀏覽器程式的比例。

2. 從 Decision tree 改成使用 Random forest，因為我們的 feature 常常沒有用到，就已經判斷出結果了(如下圖，Login_query_ratio 沒有被使用到)，因此使用多個 tree 來確定所有的特徵都是有被考慮到的。



D. File Size to large for ELK
   1. ELK 上傳有限制檔案大小，因此切分上傳並命名同樣的開頭，在用 index pattern match 到一樣的開頭來觀察整筆完整的資料。

IV. 觀察到的其他現象
   A. Phishing Email 平均的封包大小特別大。
   B. xampp-control.exe
      1. 發現是一個可以監控鍵盤與滑鼠輸入的程式，在 attack-1、2、5 中有出現(如下圖中最後一欄)，因此納入觀察的特徵當中，作為判斷的依據，但仍不清楚為什麼是 1、2、5 會特別出現或是只是偶然出現。

	A	B	C	D	E	F	G	H	I	J
1	attacktype	login_query_ratio	select_query_ratio	dest_byte_avg	dest_port_80_ratio	dest_port_dis	event_out_success	process_name_cmd	process_name_xampp	
2	attack 1	0.340577295	6.67E-05	8700.976488	0.976734884	0.954228	0.612687067	0	0.386642841	
3	attack 2	0	0	18241.51705	0.8252	0.681101	0.760531258	0	0.239334585	
4	attack 3	0	0	2661.887357	0.0031	0.000744	0.999508438	0	0	
5	attack 4	0	0	637391.4246	0.007438017	0.229002	0.999420849	0.016023166	0	
6	attack 5	0.000280899	0.07228464	26069.7086	0.544101124	0.386932	0.646092145	0	0.34559327	
7										