Predicting Data Exfiltration via DNS

Zep analytics Internship task

Use Case:

- Create a static machine learning model based on batch data. The dataset that is used is from top secret files obtained from our allies Ring Canada (RC) and the Cyber Threat Intelligence (CTI). The dataset provided to you has DNS traffic generated by exfiltrating various filetypes ranging from small to large sizes.

- The aim of the task is to implement a binary classifier aiming at predicting data exfiltration via DNS.

Work Agenda

- Exploratory Data Analysis (EDA)
- Data cleaning
- Feature engineering
- Model Training and Model evaluation

Exploratory Data Analysis (EDA)

- **Using the file called "static_dataset.csv"**
  1. checked using plots and statistical tools the distribution of each feature and the target variable
  2. checked any type of data skewed pattern.
  3. Validated if your dataset is imbalanced

Data Information:

The dataset is of shape (268074, 16) consists of 16 features 2 float numbers and 3 of object type and the remaining features are integer type.

```
S_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 268074 entries, 0 to 268073
Data columns (total 16 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   timestamp         268074 non-null  object
 1   FQDN_count        268074 non-null  int64
 2   subdomain_length  268074 non-null  int64
 3   upper             268074 non-null  int64
 4   lower             268074 non-null  int64
 5   numeric           268074 non-null  int64
 6   entropy           268074 non-null  float64
 7   special           268074 non-null  int64
 8   labels            268074 non-null  int64
 9   labels_max        268074 non-null  int64
 10  labels_average    268074 non-null  float64
 11  longest_word      268066 non-null  object
 12  sld               268074 non-null  object
 13  len               268074 non-null  int64
 14  subdomain         268074 non-null  int64
 15  Target Attack     268074 non-null  int64
dtypes: float64(2), int64(11), object(3)
memory usage: 32.7+ MB
```

Exploratory Data Analysis (EDA)

Data Description

index	FQDN_count	subdomain_length	upper	lower	numeric	entropy	special	labels	
count	268074.0	268074.0	268074.0	268074.0	268074.0	268074.0	268074.0	268074.0	
mean	22.286596238352097	6.059021016584973	0.8454195483336691	10.410013652946574	6.497586487313205	2.4857352066636893	4.533576549758648	4.788823235375307	8.2
std	6.001204805059592	3.8995053843891636	4.941928624743008	3.20772541446823	4.499865991578234	0.4077094931953377	2.1876833846359314	1.8032564817038876	4.41
min	2.0	0.0	0.0	0.0	0.0	0.219195338	0.0	1.0	
25%	18.0	3.0	0.0	10.0	0.0	2.054028744	2.0	3.0	
50%	24.0	7.0	0.0	10.0	8.0	2.57041707	6.0	6.0	
75%	27.0	10.0	0.0	10.0	10.0	2.767194749	6.0	6.0	
max	36.0	23.0	32.0	34.0	12.0	4.216846949	7.0	7.0	

String Columns

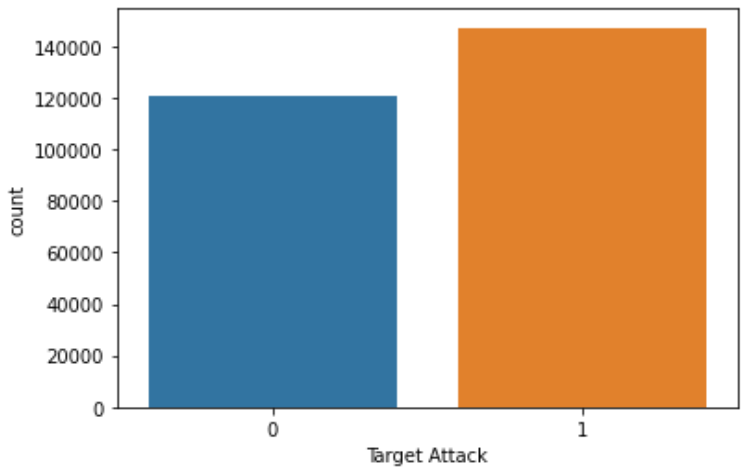
```
S_dataset['longest_word'].value_counts()

2          109981
4           70188
N            4498
C            2969
9            1906
            ...
yaa             1
queue           1
kit             1
airdrop         1
mal             1
Name: longest_word, Length: 6224, dtype: int64
```

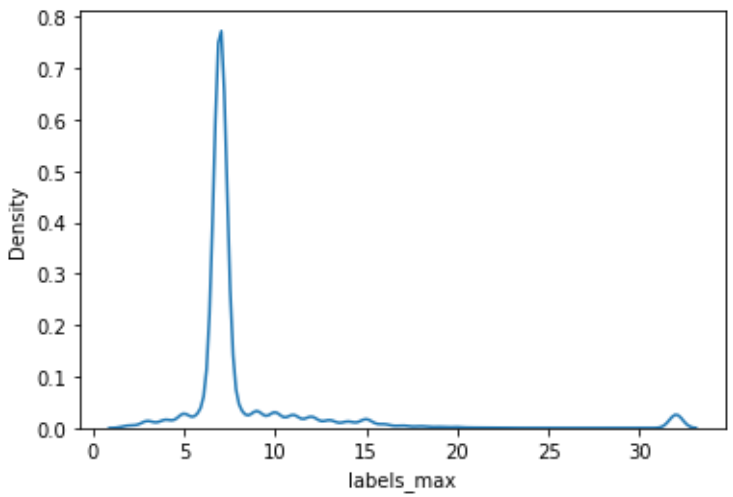
```
S_dataset['sld'].value_counts()

192                                109517
224                                 70188
FHEPFCELEHFCEPFFFACACACACACACABN     4498
DESKTOP-3JF04TC                      1961
239                                  1906
                                    ...
freesgift                               1
secureserver                            1
airdropalert                            1
queue-it                                1
lahemal                                 1
Name: sld, Length: 11112, dtype: int64
```

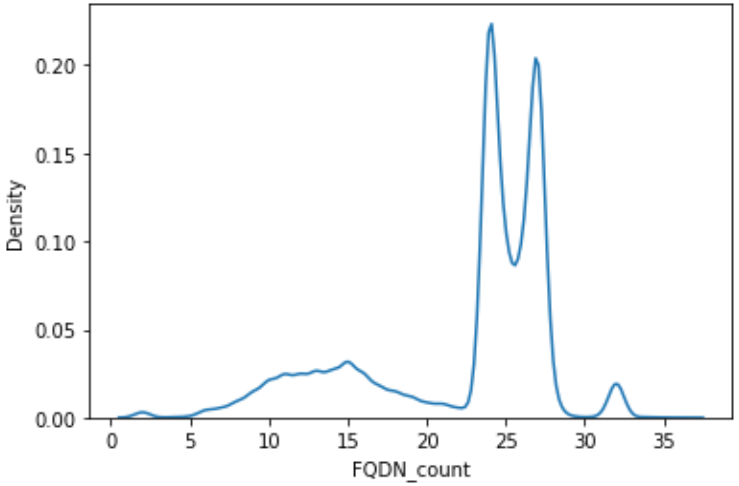
Exploratory Data Analysis (EDA)



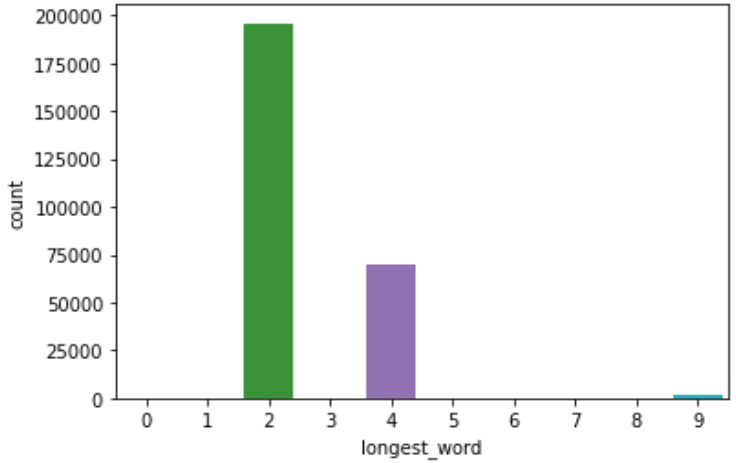
Count the Target Attacks



Skewness of labels max

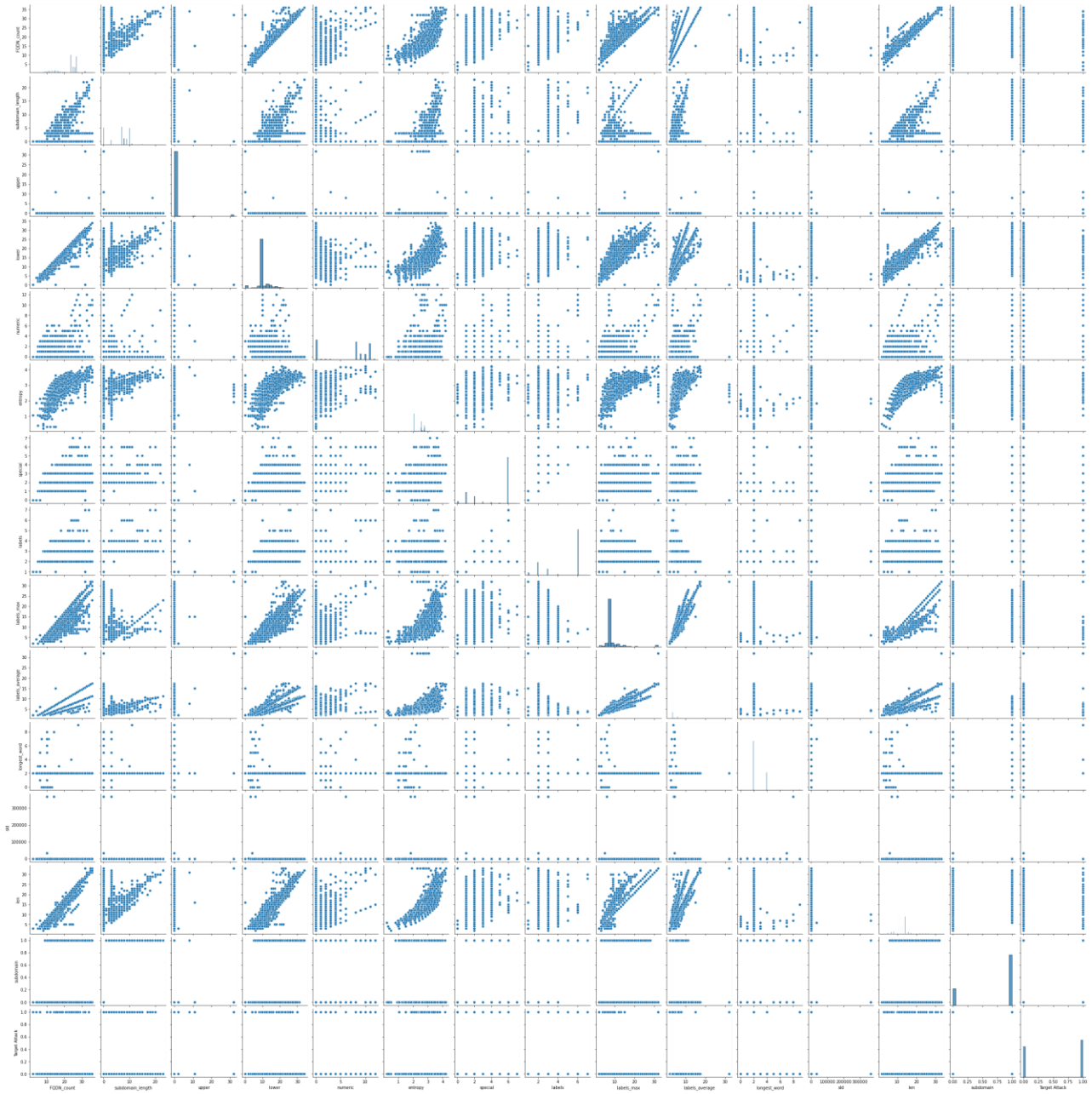


Skewness of FQDN



Count of longest word

Feature Correlation



Data cleaning

- **String columns is converted to integers**

```
S_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 268074 entries, 0 to 268073
Data columns (total 15 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   FQDN_count        268074 non-null  int64
 1   subdomain_length  268074 non-null  int64
 2   upper             268074 non-null  int64
 3   lower             268074 non-null  int64
 4   numeric           268074 non-null  int64
 5   entropy           268074 non-null  float64
 6   special           268074 non-null  int64
 7   labels            268074 non-null  int64
 8   labels_max        268074 non-null  int64
 9   labels_average    268074 non-null  float64
 10  longest_word      268074 non-null  int64
 11  sld               268074 non-null  int64
 12  len               268074 non-null  int64
 13  subdomain         268074 non-null  int64
 14  Target Attack     268074 non-null  int64
dtypes: float64(2), int64(13)
memory usage: 30.7 MB
```

- **Check Skewness of the features**

```
S_dataset.skew()
```

```
FQDN_count          -1.101731
subdomain_length    -0.590480
upper                5.988737
lower                0.343449
numeric             -0.594384
entropy             -0.140156
special             -0.902972
labels              -0.903680
labels_max           3.979910
labels_average       5.087081
longest_word         2.269378
sld                180.987411
len                  2.634801
subdomain           -1.176397
Target Attack       -0.197046
dtype: float64
```

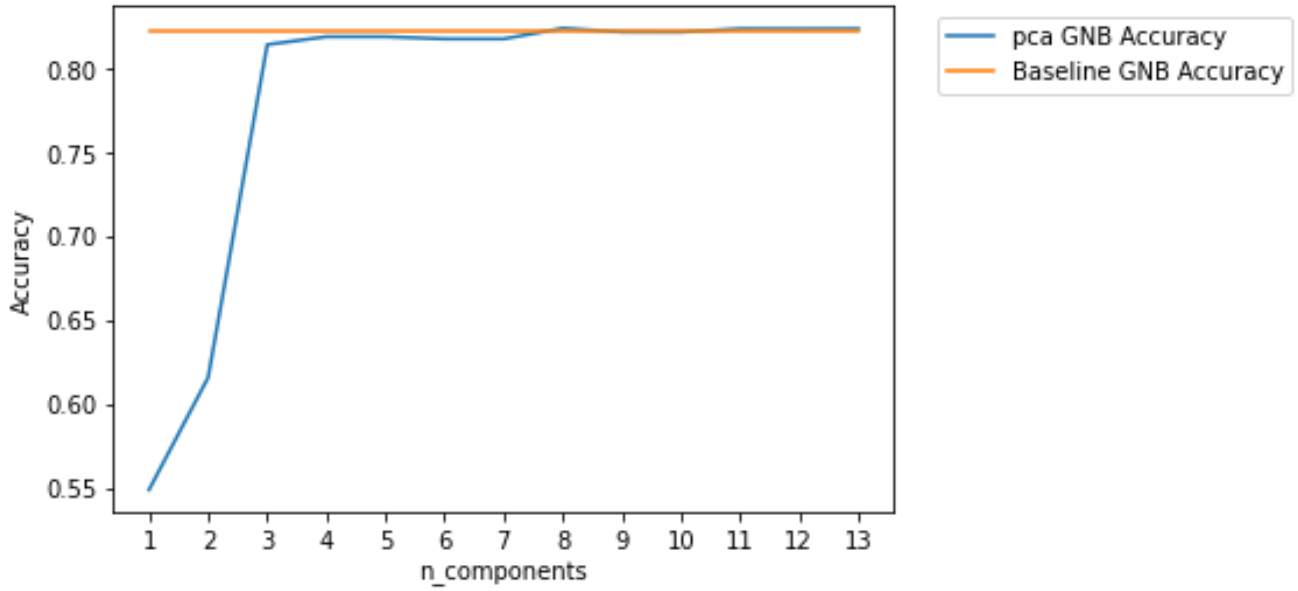
- **Check there is no null values**

```
S_dataset.isnull().sum()
```

```
FQDN_count          0
subdomain_length    0
upper               0
lower               0
numeric             0
entropy             0
special             0
labels              0
labels_max          0
labels_average      0
longest_word        0
sld                 0
len                 0
subdomain           0
Target Attack       0
dtype: int64
```

Feature engineering

- Applied PCA dimensionality reduction on the dataset and found that best component is 13



Model Training and Evaluation

- Logistic Regression Model is used for binary classification problem

```
Classification Report:

              precision    recall  f1-score   support

           0       0.99      0.61      0.76     30224
           1       0.76      1.00      0.86     36795

    accuracy                           0.82     67019
   macro avg       0.88      0.81      0.81     67019
weighted avg       0.86      0.82      0.82     67019


Confusion Matrix:

[[18583 11641]
 [  141 36654]]

Accuracy Score:

0.8241991077157224
```