IBM Project - Malware Detection/Classification

Group 22
Sandeep Rajakrishnan, Sudhay Senthilkumar

Common features

- From the 3 datasets obtained namely UNSW_NB15, IoT Botnet and NSL KDD, we were able to find out the influencing features and common features across these datasets.
- Our observation also includes a multiclass classified output with various types of attacks like DoS, Backdoor, Reconnaissance, etc...
- Hence, using these common features, we understand that when a new data point is provided and asked to classify under a type of attack with these respective columns' data, we can find it's type of attack.

protocol
service
state
duration
category / attack_cat / labels

spkts
dpkts
src_bytes
dest_bytes
state

Dataset Description

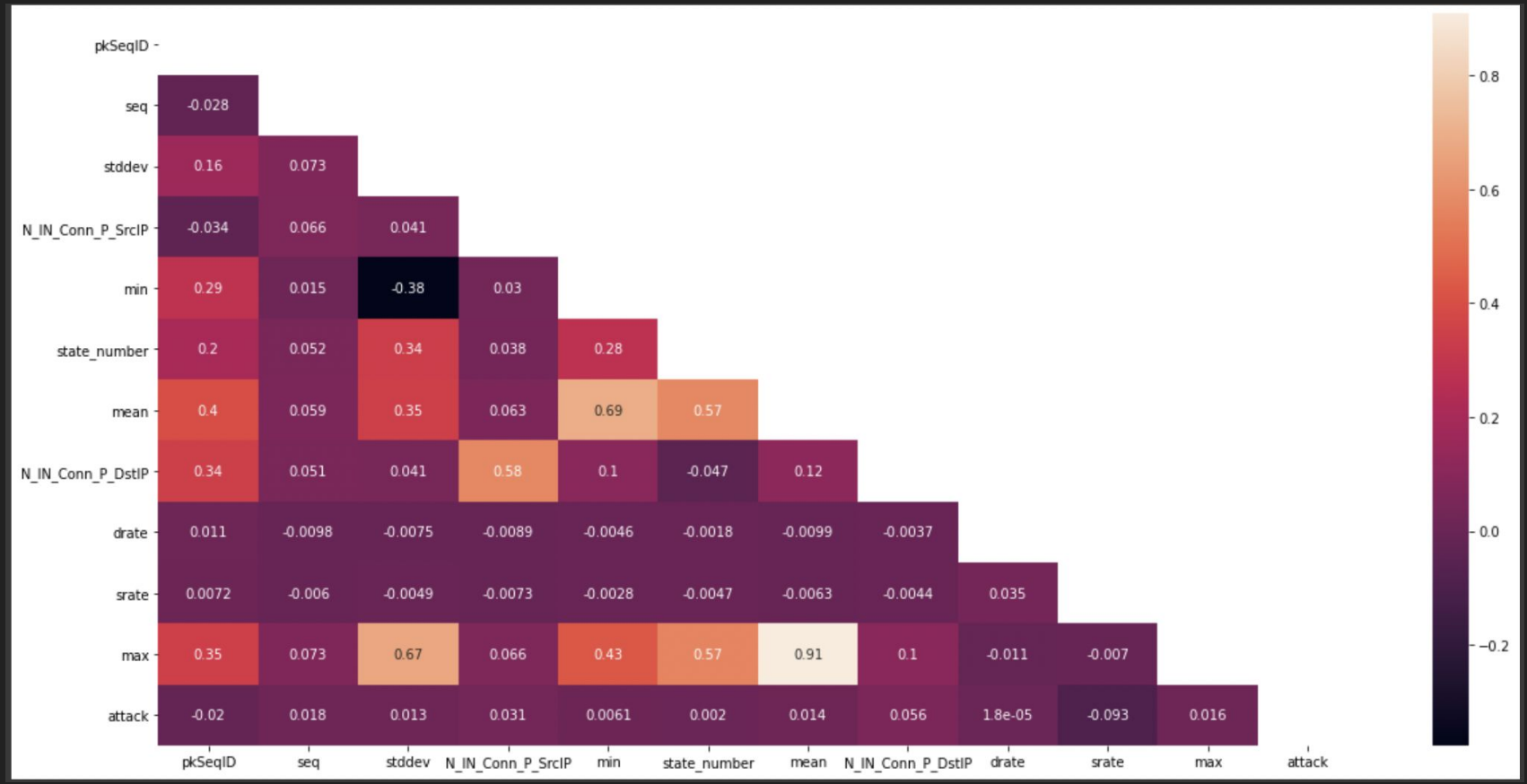
The dataset we have considered is *UNSW_2018_IoT_Botnet_Full5pc_4*

The meaning of each column has been described below.

Feature	Description
pkSeqID	Row Identifier
Stime	Record start time
flgs	Flow state flags seen in transactions
flgs_number	Numerical representation of feature flags
Proto	Textual representation of transaction protocols present in network flow
proto_number	Numerical representation of feature proto
saddr	Source IP address
sport	Source port number
daddr	Destination IP address
dport	Destination port number
pkts	Total count of packets in transaction
bytes	Totan number of bytes in transaction
state	Transaction state
state_number	Numerical representation of feature state
ltime	Record last time
seq	Argus sequence number

dur	Record total duration
mean	Average duration of aggregated records
stddev	Standard deviation of aggregated records
sum	Total duration of aggregated records
min	Minimum duration of aggregated records
max	Maximum duration of aggregated records
spkts	Source-to-destination packet count
dpkts	Destination-to-source packet count
sbytes	Source-to-destination byte count
dbytes	Destination-to-source byte count
rate	Total packets per second in transaction
srate	Source-to-destination packets per second
drate	Destination-to-source packets per second
attack	Class label: 0 for Normal traffic, 1 for Attack Traffic
category	Traffic category
subcategory	Traffic subcategory

Data Correlation



Algorithm Used and Approach

- We have used the Support Vector Machine, ANN, LR Algorithm on the dataset to obtain the confusion matrix and a classification report.
- To begin, we have performed :
  - Encoding - To give integer representations to strings - Target Encoding performs best
  - Scaling - Used StandardScaler() - to normalize all values
  - Obtained correlation between columns to find influential attributes
  - Split dataset into Test and Train datasets
  - Applied SVM algorithm and obtained a predicted dataset with test set
  - Compared with actual test dataset so as to obtain performance metrics

Performance Metrics

Accuracy Score = 0.888275
Algorithm Used : SVM

	precision	recall	f1-score	support
1	0.86	0.90	0.88	17891
2	0.91	0.87	0.89	21084
3	1.00	1.00	1.00	4
4	1.00	1.00	1.00	1
5	1.00	1.00	1.00	1020
accuracy			0.89	40000
macro avg	0.95	0.95	0.95	40000
weighted avg	0.89	0.89	0.89	40000

Final accuracy score=0.888275

Artificial Neural Network using Keras

Trained on a dataset with 700,000 data points

```
model = Sequential()
model.add(Dense(16,input_dim=26,activation='relu'))
model.add(Dense(12,activation='relu'))
model.add(Dense(5,activation='softmax'))
```

```
[ ] model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

```
[ ] history = model.fit(X_train, y_train, epochs=100, batch_size=64)
```

```
from sklearn.metrics import accuracy_score
a = accuracy_score(pred,test)
print('Accuracy is:', a*100)
```

```
Accuracy is: 99.40371130086343
```

Multinomial Logistic Regression

Trained on a dataset with 700,000 data points

```
[ ]
    X_train,X_test,y_train,y_test = train_test_split(X,y,train_size=0.8,random_state=109)
```

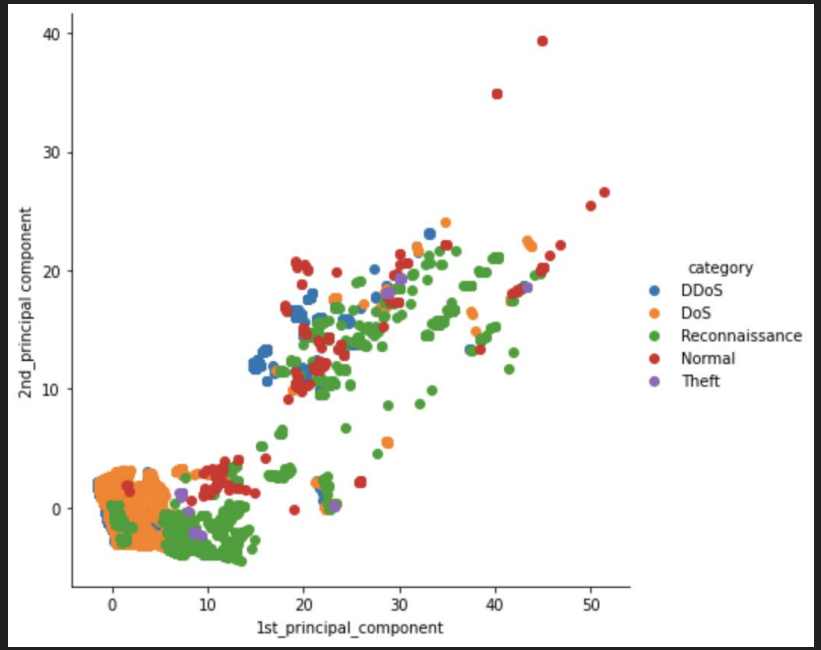
```
[ ] clf = LogisticRegression(random_state=0,multi_class='multinomial').fit(X_train, y_train)
```

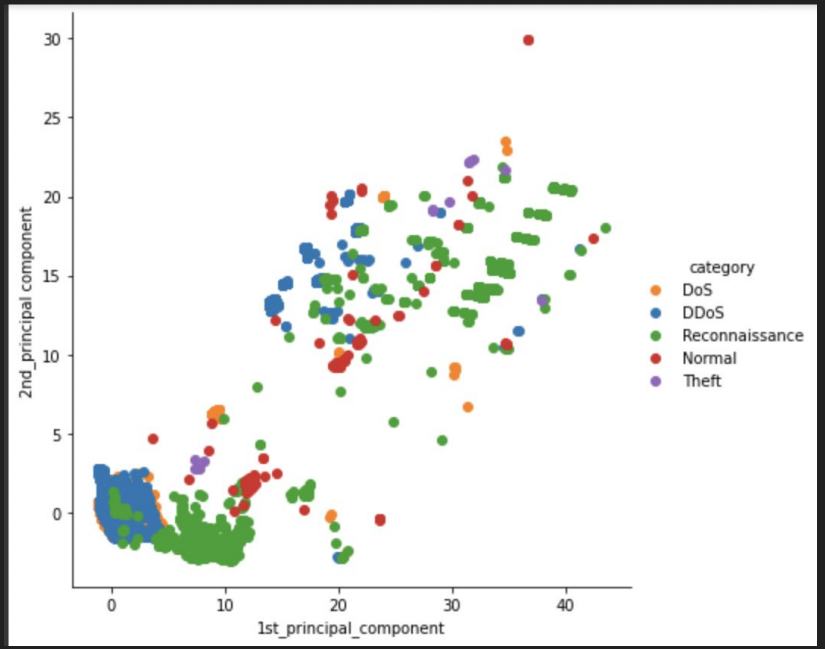
```
[ ] predictions = clf.predict(X_test)
    print(accuracy_score(y_test,predictions))

    0.9853347053652354
```

PCA Results

Two principal components (Features reduced to two)

Data is not easily separable





References

- All ipynb notebooks can be found at : https://github.com/IBM-ML-PROJECT/GROUP-22-ML-project
- Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset

Link to dataset:

- https://raw.githubusercontent.com/defcom17/NSL_KDD/master/KDDTrain%2B.csv
- https://cloudstor.aarnet.edu.au/plus/s/umT99TnxvbpkkoE?path=%2FCSV%2FTraning%20and%20Testing%20Tets%20(5%25%20of%20the%20entier%20dataset)%2FAll%20features