FEATURE SELECTION FOR NETWORK TRAFFIC DATA WHERE NUMERICAL AND CATEGORICAL FEATURES CO-EXIST.

Research problem statement.

- Yash Shah
ys1300@rit.edu

CATEGORICAL FEATURES

- **Protocol**

A network protocol is a set of established rules that dictate how to format, transmit and receive data so that computer network devices - from servers and routers to endpoints - can communicate, regardless of the differences in their underlying infrastructures, designs or standards.

Protocol ID	Protocol Type
0	IPv6 Hop-by-Hop Option (HOPOPT)
6	Transmission Control Protocol (TCP)
17	User Datagram Protocol (UDP)

One Hot Encoding – Makes n number of columns for n categories. Represents the value.

PROTOCOL 0	PROTOCOL 6	PROTOCOL 17
0.0	1.0	0.0
0.0	1.0	0.0
0.0	1.0	0.0
0.0	1.0	0.0
0.0	1.0	0.0
0.0	1.0	0.0
...	...	...
0.0	0.0	1.0
0.0	0.0	1.0
0.0	1.0	0.0
0.0	0.0	1.0
0.0	0.0	1.0

- **Destination port**

- Server applications
- 53805 unique dest. Port.

Two feature engineering/ encoding methods used

1. Freq Encoding - replaces the ports with their frequency throughout the data.
2. Aggregating function - keeps the high frequency values as it is and replaces the others with a new category.

APPROACH

Jupyter Notebook with Scikit learn, matplotlib, seaborn.

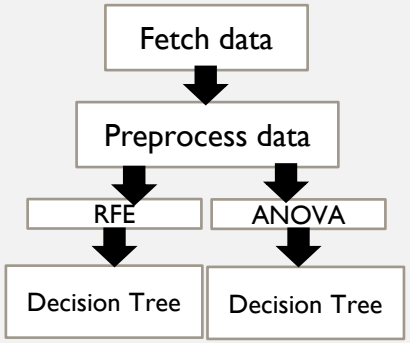
PREPROCESSING DATA

- CIC IDS 2017 - 2.8 million samples – 14 classes – data over 5 days.
- Dropped columns which has only 1 value throughout the dataset. 10 columns.
- Removed the rows with nan and inf values – because its less than 0.01 as compared to whole data.
- Robust scaling - value = (value – median) / (p75 – p25). (p=percentile) – hence does not get affected by outliers and at the same time keeps them in the data.

FEATURE SELECTION

- Removed highly correlated features using the correlation plot – Only removes 11 features hence not a viable option.
- Recursive feature elimination – uses 100 random forests to rate features by importance (how often they were used to split the data.) – selected top 6 features – selected with 0.4 threshold from importance scale – gives good results even with 6 features – only issue is that since these features are highly correlated, it will give us a different set of features when its run again. But it doesn't make much difference because we get similar results from classification.
- ANOVA(***An****alysis* ***o****f* ***Va****riance)* - If there is equal variance between groups/classes, it means this feature has no impact on response and it can not be considered for model training. – selected about top 20 features.

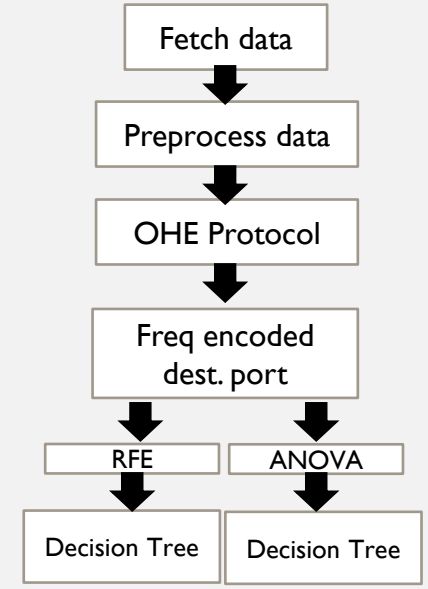
PIPELINE (ONLY NUMERICAL DATA)



RFE ~~ 6 features – above 0.4 importance threshold
Packet Length Variance
Init_Win_bytes_forward
Bwd Packet Length Mean
Bwd Packet Length Std
Packet Length Std
'Destination Port_freq_encode

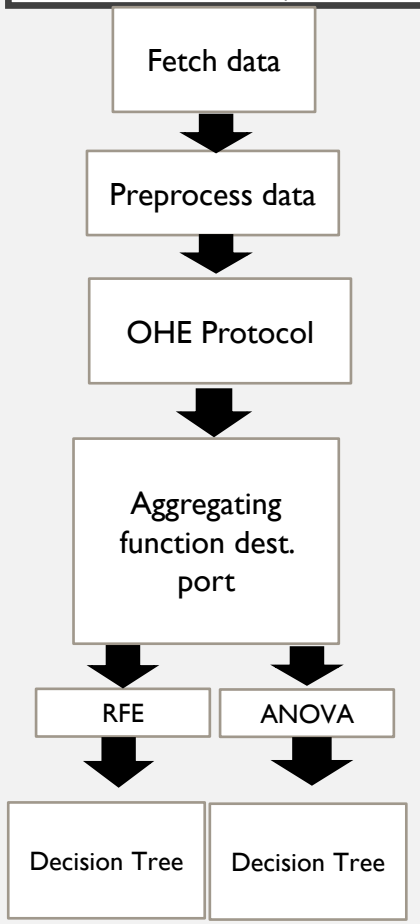
ANOVA for 2nd pipeline~~ 20 features -
Flow Duration
Flow Duration, Bwd Packet Length Max, Bwd Packet Length Mean,
Bwd Packet Length Std, Flow IAT Std, Flow IAT Max,
Fwd IAT Total, Fwd IAT Std, Fwd IAT Max, Max Packet Length,
Packet Length Mean, Packet Length Std, Packet Length Variance,
PSH Flag Count, Average Packet Size, Avg Bwd Segment Size,
Idle Mean, Idle Max, Idle Min, Destination Port_freq_encode.

PIPELINE (FREQ ENCODED DEST. PORT)



ANOVA for 3nd pipeline~~ 20 features -
Flow Duration
Flow Duration, Bwd Packet Length Max, Bwd Packet Length Mean,
Bwd Packet Length Std, Flow IAT Std, Flow IAT Max,
Fwd IAT Total, Fwd IAT Std, Fwd IAT Max, Max Packet Length,
Packet Length Mean, Packet Length Std, Packet Length Variance,
PSH Flag Count, Average Packet Size, Avg Bwd Segment Size,
Idle Mean, Idle Max, Idle Min, 'PROTOCOL 6'.

PIPELINE (FREQ ENCODED DEST. PORT)



Different strategies I have tried over the semester

1. Decision tree with only numerical data without feature selection (robust model).
2. Decision tree with numerical as well as Protocol feature without feature selection (robust model).
3. Decision tree with numerical data, unsupervised feature selection (correlation plot).
4. Decision tree after adding protocol feature to the 3rd strategy.

RESULTS

DATA	Feature Selection	DT – Accuracy (weighted) %	DT – Macro Avg Accuracy (Unweighted) %
Only numerical	RFE*	99	67
	ANOVA	99	80
Numerical + OHE protocol + freq_encoded dest.	RFE (selects freq_encoded dest. But none of the protocol features)	99	73
	ANOVA (20 features - Dest. Port one of them)	99	83
Numerical + OHE protocol + agg function dest. Port with 90% threshold**	RFE (selects agg function Dest port as a feature but none of the protocol features)	85	33
	ANOVA (20 features – protocol 6 one of them)	98	80

* Increase in number of features from RFE would increase the accuracies but I was looking for the minimum I could go for.
** Accuracies decrease as the threshold for agg. Function decreases. i.e 85-80-75

Was hoping for better results from RFE (since it is based on the decision trees itself).
Anova gets better results as compared to RFE, but that may simply because its using more number of features than RFE.

Generally bad results for labels with less samples – Web Attack Brute force, Web attack sql Injection, Web attack XSS, Heartbleed, Infiltration.

Tried to implement svm and knn, to tackle the disadvantages of Decision Tree, but both of them were taking a lot of time.

TAKEAWAYS

- Learned how to build an ML project from scratch.
- Preprocessing and Importance of Scaling/normalization.
- Encoding techniques and their Impact on data.
- Feature selection and their impact on model.
- Disadvantages of Decision trees for this dataset (if Class imbalance problem is not tackled).

GITHUB REPO

- Link.
- Includes – 3 notebooks for the 3 pipelines.
  - A readme file explaining everything I have done.
  - This PPT as well.

THANKYOU ☺