Feature Selection Using Quantum Annealing

Team:
Nipun Ramagiri
Swapnil Bhosale

What is Feature Selection?

- Important concept in Statistics and ML
- Select set of input variables (features)
- Mutual information at rescue for above task
- N choose K problem.

Mutual Information

- Measure of the mutual dependence between the two variables
- Quantifies "amount of information" of one variable with respect to other variable
- Shannon Entropy and Mutual Information.

The method is based on the 2014 paper, **Effective Global Approaches for Mutual Information Based Feature Selection**, by Nguyen, Chan, Romano, and Bailey published in the Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining.

Forming a QUBO

	Formula	Optimization	Linear Terms	Quadratic Terms
Feature Selection	$\sum_{i=1}^{n} \left\{ I(X_i; Y) + \sum_{j \in k \mid i} I(X_j; Y \mid X_i) \right\}$	Maximize	$I(X_i; Y)$	$I(X_j; Y \mid X_i)$
QUBO	$\sum_{i}^{N} q_i x_i + \sum_{i<j}^{N} q_{i,j} x_i x_j$	Minimize	$q_i x_i$	$q_{i,j} x_i x_j$

- Diagonal elements (linear coefficients) represent MI: Qii ← - I(Xi; Y)
- non-diagonal elements (quadratic elements) represent CMI: Qij ← -I(Xj ; Y jXi)

Penalizing invalid states

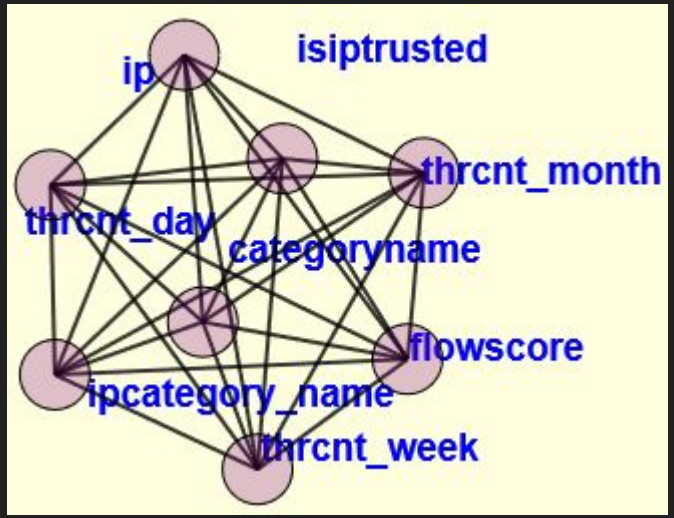
$$P = \alpha \sum_{i=1}^{N} (x_i - k)^2$$

P is positive whenever the number of 1s in solution x1, x2, . . . xN is not k

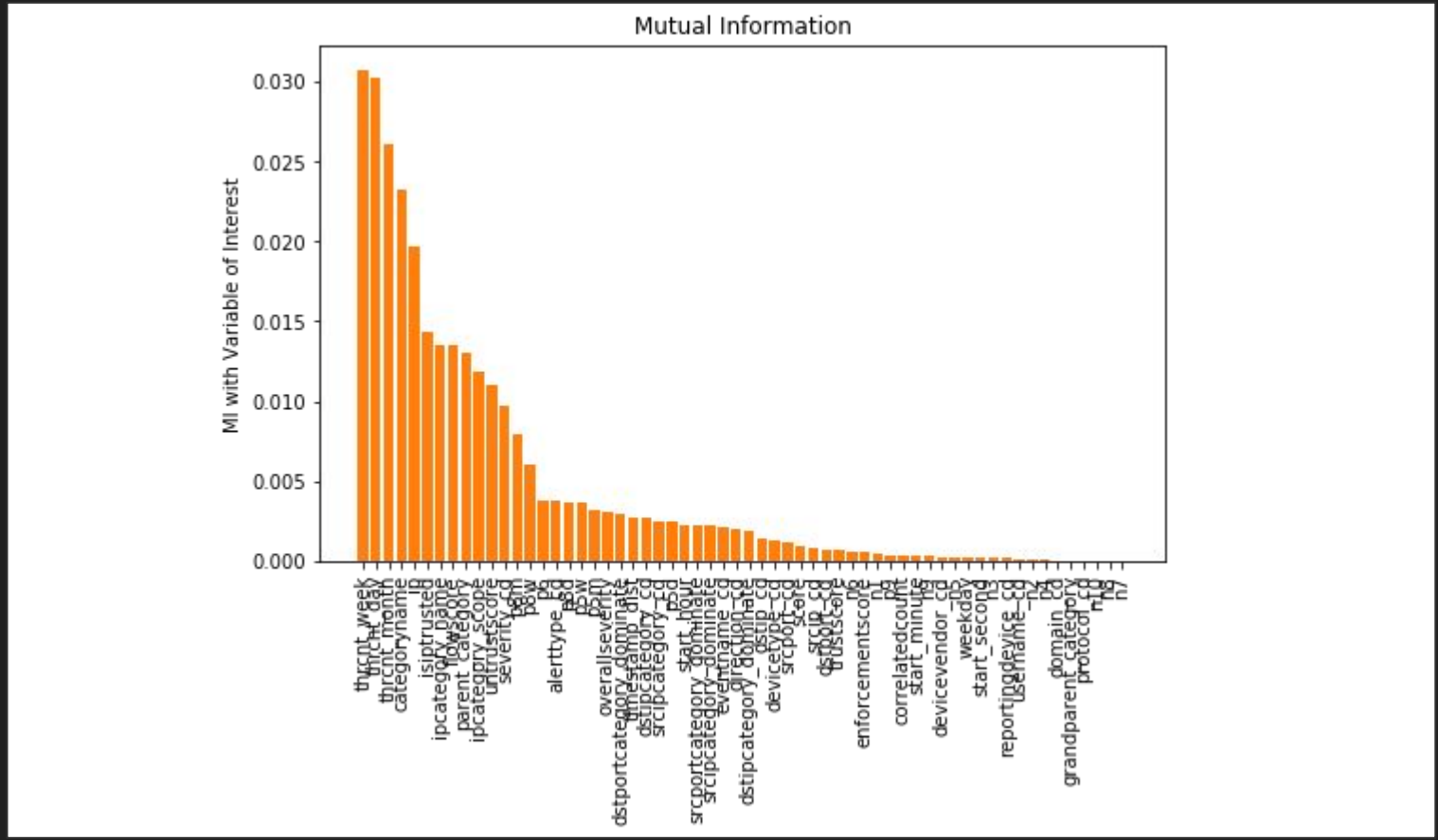
Sample Dataset used

Dataset: IEEE BigData 2019 Cup(Suspicious Network Event Recognition) with more than 60 features

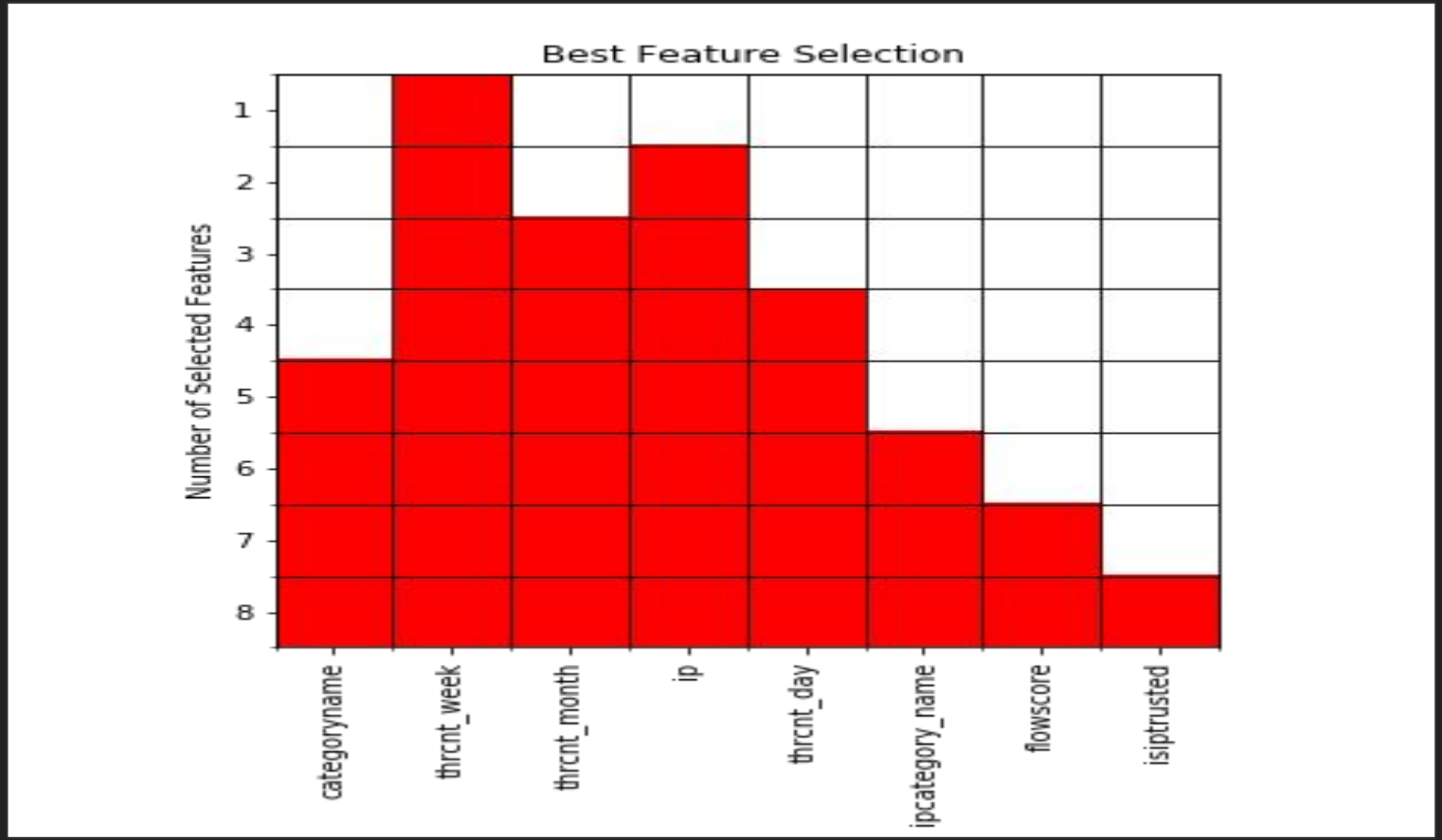
BQM with 8 nodes and 28 edges



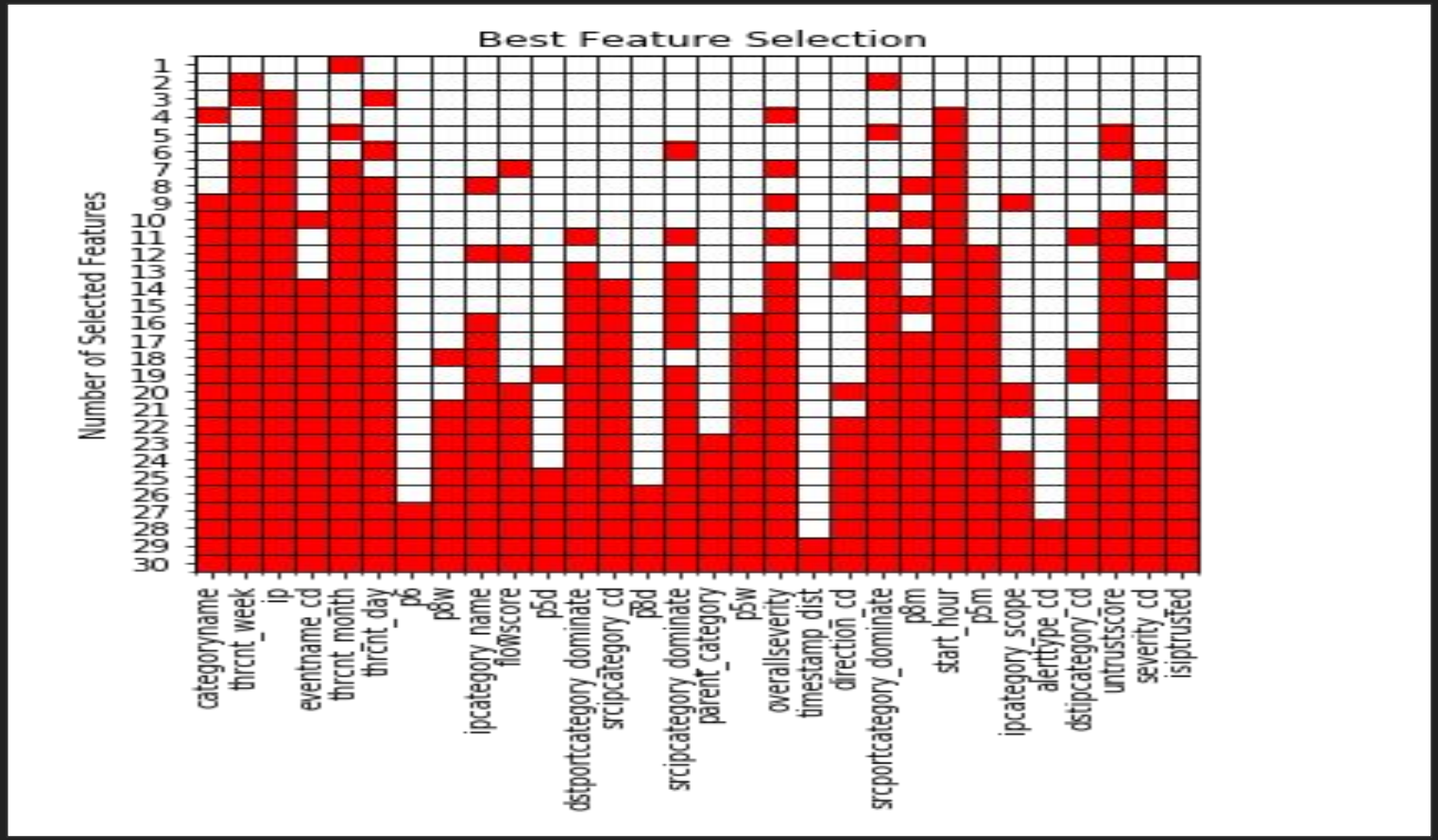
Output:


Mutual Information
MI with Variable of Interest
0.030
0.025
0.020
0.015
0.010
0.005
0.000
trncnt_week
trncnt_day
trncnt_month
categoryname
ip
isiptrusted
ipcategory_name
parent category
ipcategory_scope
untrustscore
severity_id
pow
pa
alerttype_id
pod
pow
pa
overallseverity
dstportcategory_domainrate
dstipcategory_dst
dstipcategory_cd
srcipcategory_cd
pow
start_hour
srcportcategory_dominate
srcipcategory_dominate
eventname_cd
direction_cd
dstipcategory_dominate
dstip_cd
devicetype_cd
srcport_cd
score
srcip_cd
dstport_cd
trustscore
ng
enforcementscore
ni
pa
correlatedcount
start_minute
ng
devicevendor_cd
pa
weekday
start_second
ng
reportingdevice_cd
username_cd
ni
pa
domain
grandparent_category
protocol_cd
ng
ni

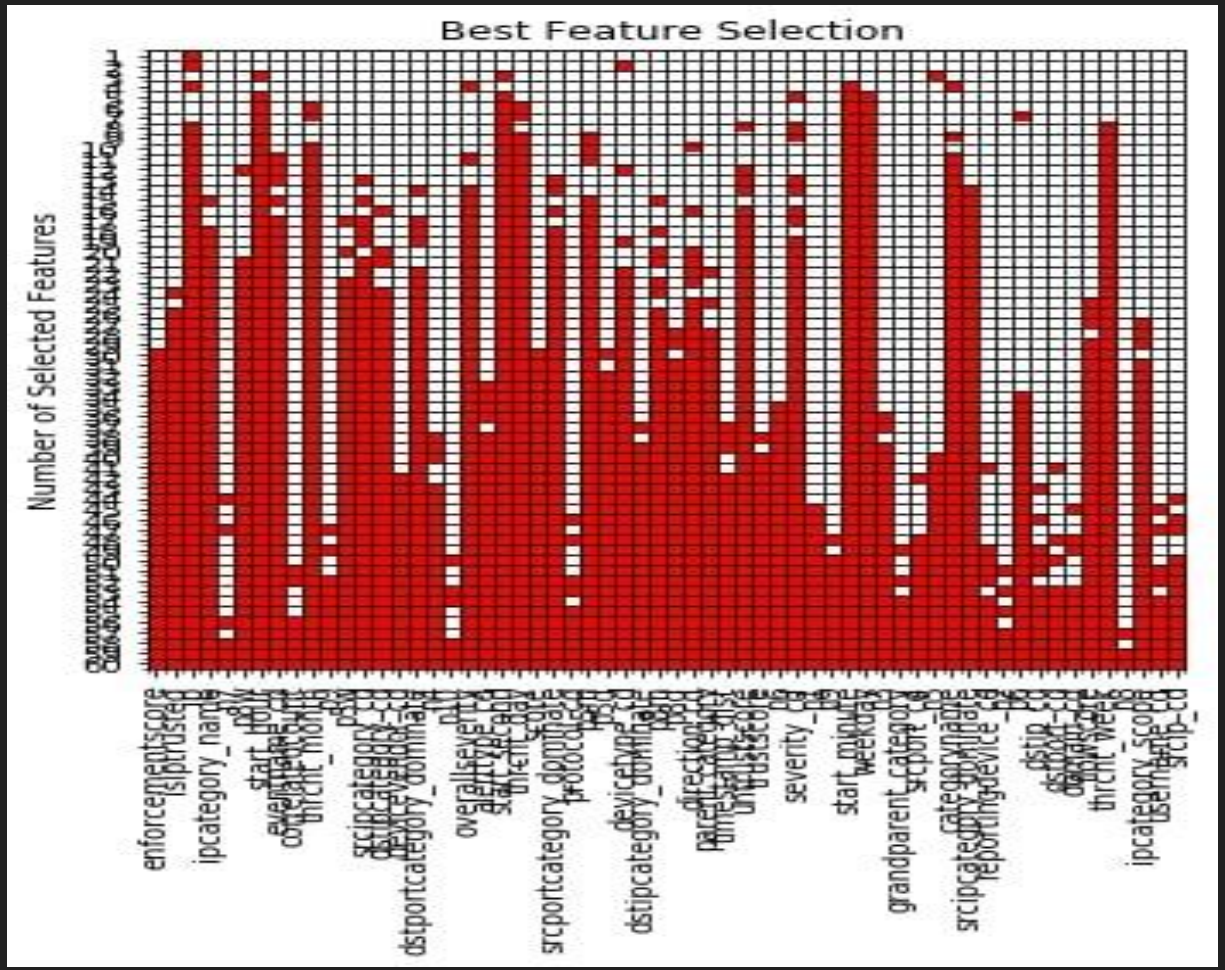
Output:


Best Feature Selection
Number of Selected Features
1
2
3
4
5
6
7
8
categoryname
thrcnt_week
thrcnt_month
ip
thrcnt_day
ipcategory_name
flowscore
isiptrusted

Output:


Best Feature Selection
Number of Selected Features
categoryname
thrcnt_week
thrcnt_ip
eventname_cd
thrcnt_month
thrcnt_day
p6
p8w
ipcategory_name
flowscore
p5d
dstportcategory_dominate
srcipcategory_cd
p8d
srcipcategory_dominate
parent_category
p5w
overallseverity
timestamp_dist
direction_cd
srcportcategory_dominate
p8m
start_hour
p5m
ipcategory_scope
alerttype_cd
dstipcategory_cd
untrustscore
severity_cd
isiptrusted

Output


Best Feature Selection
Number of Selected Features
enforcementscore
isptrusted
ipcategory_name
dstportcategory_domainname
overallsexemity
srcportcategory_domainname
dstipcategory_domainname
grandparent_category
srcipcategory_domainname
ipcategory_scoreName

Runtimes:

No. of Features	Tabu Sampler	Neal Annealer	Dwave 2000Q
8	4.86 s	0.147 s	3.272 s
15	NA	0.392 s	5.865 s
30	NA	1.85 s	12.297 s
60	NA	10.4 s	24.602 s

References:

1. Ieee bigdata 2019 cup: Suspicious network event. https://knowledgepit.ml/suspicious-network-event-recognition/

2. Introduction to quantum annealing. https://docs.dwavesys.com/docs/latest/c_gs_2.html

3. Application of high-dimensional feature selection: evaluation for genomic prediction in man. https://www.nature.com/articles/srep10312 (2003

4. Isabelle Guyon, A.E.: An introduction to variable and feature selection. http://jmlr.csail.mit.edu/papers/v3/guyon03a.html (2015)

Thank You!!!