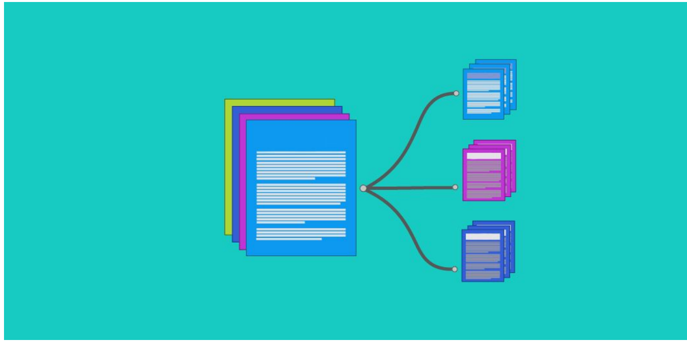
IP Network Traffic Classification

A Machine Learning Approach by Venkit

Overview

- It is important to understand the correlation between network traffic and its causal application.
- It allows organizations to ensure quality of service. Do more with less !!
- Prevent malicious applications and cyber attacks
- Provide lawful intercept
- Saves money and prevents loss of reputation



Intro

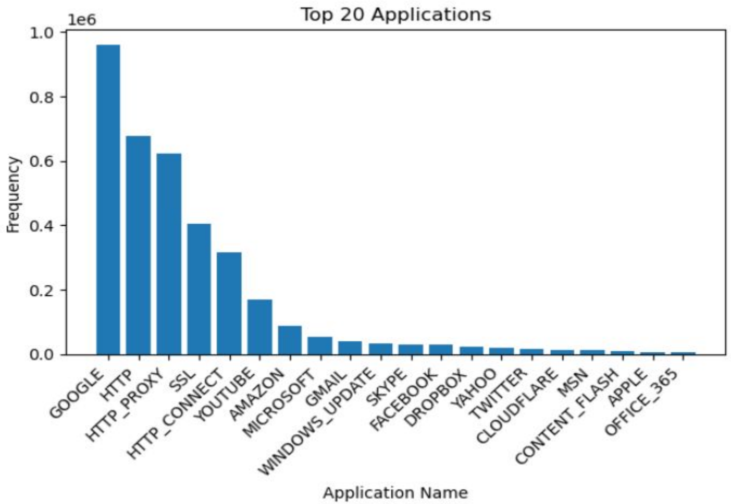
Can we train a ML model on a time series data that has packet captures and other meta information to identify the application??

- **Identify**
  Identify applications traffic as it flows through the network of the organization. Example: who is browsing facebook.com at work?
- **Report**
  Be able to generate reports of different application traffic.
- **Improve**
  Prevent malicious applications. Improve on utilization by redirecting application traffic to different carriers and save money

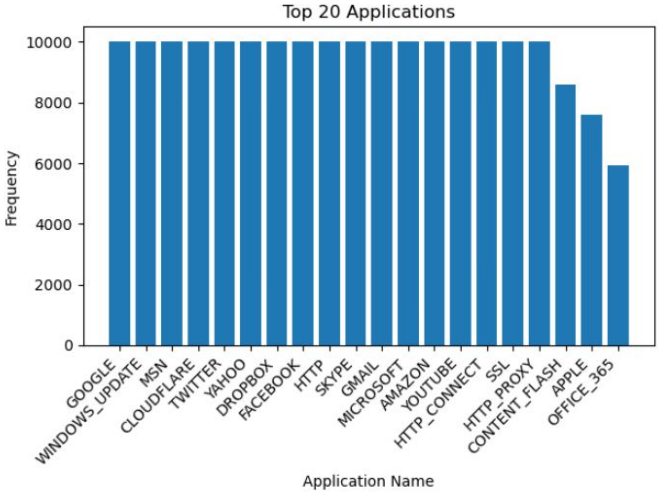
Data

- The data was collected in a network section from Universidad Del Cauca, Popayán, Colombia by performing packet captures at different hours, during morning and afternoon, over six days (April 26, 27, 28 and May 9, 11 and 15) of 2017. A total of 3.577.296 instances were collected and are currently stored in a CSV (Comma Separated Values) file.
- The flow statistics (IP addresses, ports, inter-arrival times, etc) were obtained using CICFlowmeter (github).
- The application layer protocol was obtained by performing a DPI (Deep Packet Inspection) processing on the flows with ntopng (github).

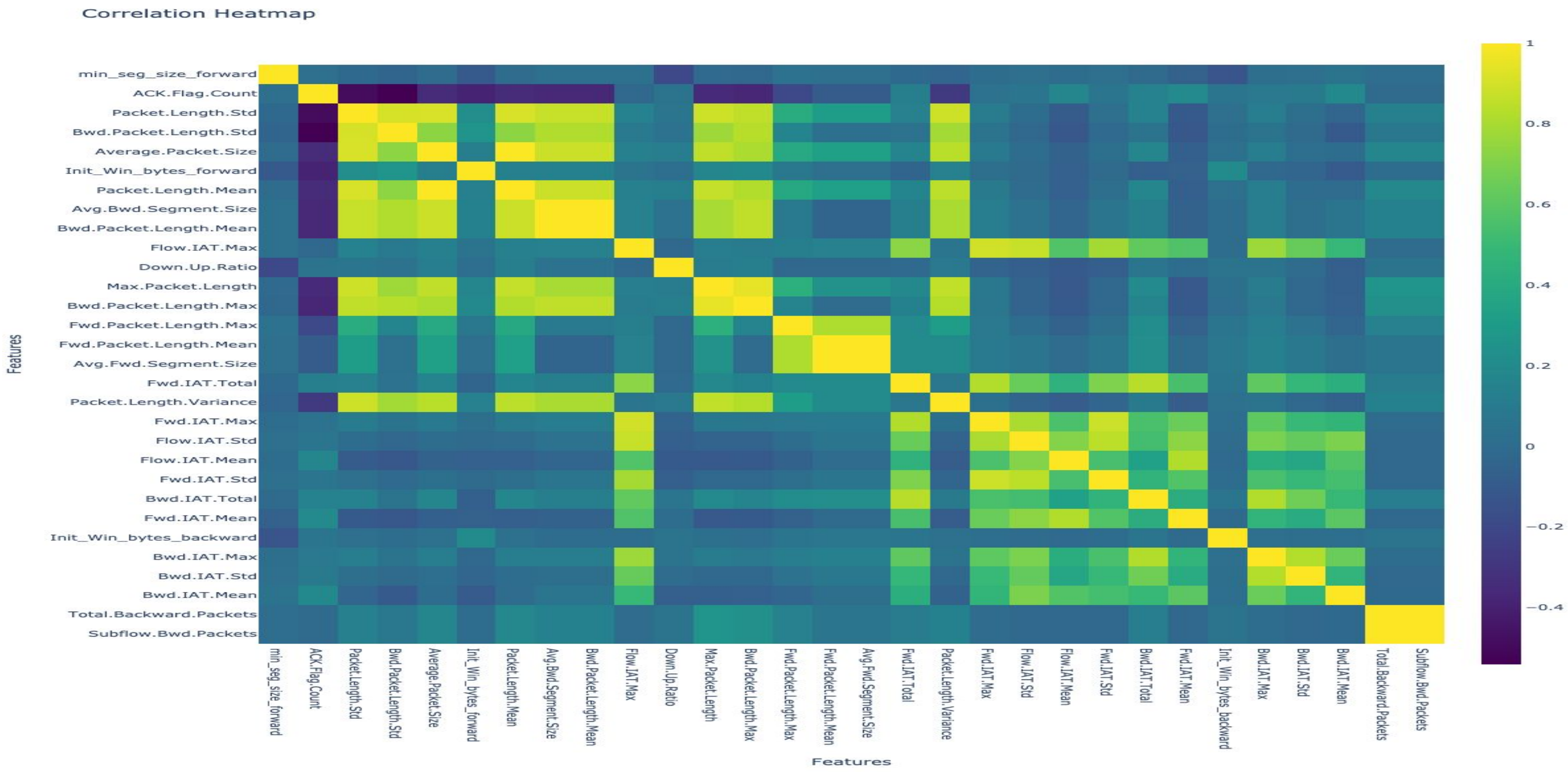
Data Preparation



- Selection top 20 applications
- Balance the dataset



Data Analysis



Data Analysis continued

Observation from correlation heatmap

Some of the features are obviously correlated

- For example most of the packet length related features correlate highly. This is not surprising.
- Similarly some of the IAT values (Bwd and Flow IAT.Max) correlate highly. This is also expected.

Some features are highly negatively correlated

- One example is the ACK.Flag.Count to various packet lengths. Here also, there is no surprise, because as packet length increases, the number of acks published for each packet decreases and vice-versa.

NOTE: Outside of the above obvious observations, there was nothing much to note.

Feature Engineering

The following two feature engineering was performed

1. Using Pricipal Component Analysis (PCA) the dimension was reduced to 10 features
2. Feature selection by sorting the values of Inter Quartile Range (IQR) for all the features. The top 30 features were taken.

A training and holdout test set was obtained for both of the above feature engineering techniques.

Modeling

The training datasets generated in the previous step was used to build models that classifies the network flows into their causal applications/protocols. The following ML algorithms were evaluated as classifiers.

1. Logistic Regression
2. Decision Tree
3. Support Vector Machines (SVM)
4. K-Nearest Neighbors (KNN)
5. Gaussian Naive Bayes
6. Random Forest Classifier
7. XGBoost Classifier
8. Dummy Classifier

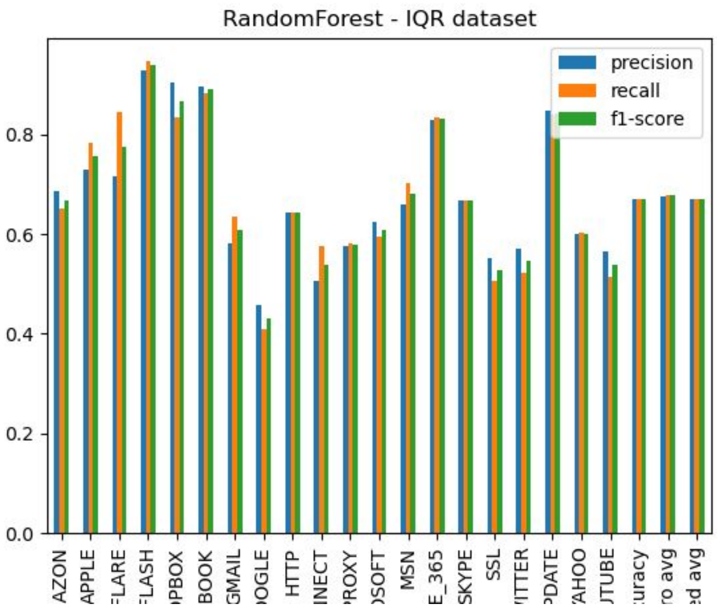
Evaluation

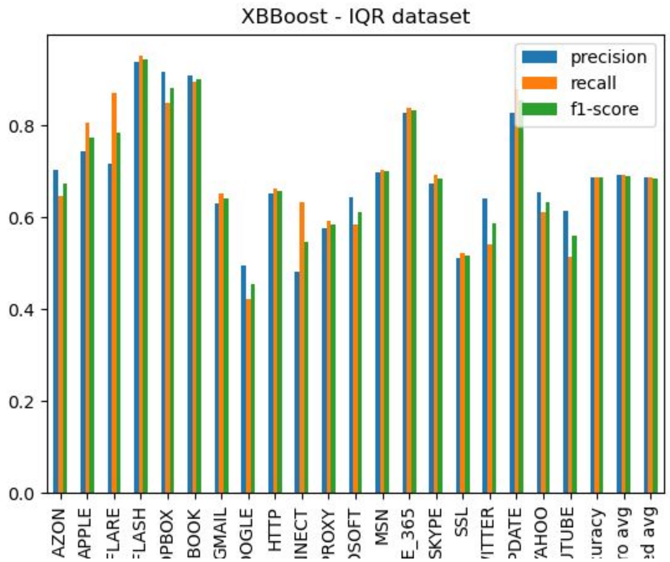
- *#TP = No. of True Positives*
- *#TN = No. of True Negatives*
- *#FP = No. of False Positives*
- *#FN = No. of False Negatives*

- `Accuracy  =  ( #TP + #TN ) /  ( #TP + #TN + #FP + #FN )`
- `Precision =  ( #TP ) / ( #TP + #FP )`
- `Recall    =  ( #TP ) / ( #TP + #FN )`
- `F-measure =  ( 2 x Precision x Recall ) / ( Precision + Recall )`

Accuracy is important !!

Results





Random Forest and XGBoost on the IQR dataset performed the best !!

Results continued

Confusion Matrix for RandomForest - IQR dataset

Actual	AMAZON	APPLE	CLOUDFLARE	CONTENT_FLASH	DROPBOX	FACEBOOK	GMAIL	GOOGLE	HTTP	HTTP_CONNECT	HTTP_PROXY	MICROSOFT	MSN	OFFICE_365	SKYPE	SSL	TWITTER	WINDOWS_UPDATE	YAHOO	YOUTUBE
YOUTUBE	13	27	27	1	7	12	115	224	35	107	80	32	36	7	45	85	76	7	38	1031
YAHOO	41	22	36	2	17	7	70	37	30	158	67	29	30	6	43	56	76	6	1184	44
WINDOWS_UPDATE	7	3	30	7	7	10	6	2	26	19	14	46	21	52	14	13	10	1631	15	8
TWITTER	41	40	48	9	10	14	74	100	40	129	78	49	39	6	63	85	1060	9	67	67
SSL	51	56	60	5	24	11	93	103	57	13	9	101	32	2	86	1041	84	12	126	92
SKYPE	17	10	13	2	11	6	36	40	32	100	37	109	57	9	1335	68	44	6	36	29
OFFICE_365	17	8	24	1	1	2	3	6	15	2	6	12	8	918	6	11	10	45	1	3
MSN	38	11	59	5	10	14	17	15	32	65	42	80	1336	6	47	19	21	27	42	18
MICROSOFT	49	24	34	4	11	12	19	17	38	97	39	1207	119	17	109	69	43	38	57	21
HTTP_PROXY	17	23	7	0	10	27	108	69	164	99	1185	28	50	16	22	10	54	43	48	54
HTTP_CONNECT	15	43	9	0	17	22	27	56	27	1138	112	65	57	3	70	9	99	8	135	67
HTTP	97	14	114	7	5	9	17	63	1300	33	121	21	56	18	16	45	30	23	16	19
GOOGLE	29	37	41	2	8	6	273	831	48	92	93	38	29	3	39	95	79	6	45	235
GMAIL	14	24	20	0	9	6	1275	188	19	69	90	19	28	1	30	39	50	6	39	79
FACEBOOK	11	18	19	1	1	1807	4	8	18	31	33	10	15	7	6	9	9	23	8	6
DROPBOX	15	14	20	1	1714	2	23	23	20	33	15	18	7	5	19	39	25	2	52	9
CONTENT_FLASH	58	10	6	1604	0	0	0	3	2	0	0	0	0	2	0	5	0	1	3	0
CLOUDFLARE	45	23	1727	5	7	23	10	6	30	3	5	13	33	9	13	46	19	6	12	10
APPLE	20	1170	21	8	7	17	9	14	30	18	5	27	14	5	10	50	21	13	19	15
AMAZON	1308	25	101	62	22	9	15	12	57	36	30	32	57	15	26	93	45	14	32	18

Confusion Matrix for XGBoost - IQR dataset

Actual	AMAZON	APPLE	CLOUDFLARE	CONTENT_FLASH	DROPBOX	FACEBOOK	GMAIL	GOOGLE	HTTP	HTTP_CONNECT	HTTP_PROXY	MICROSOFT	MSN	OFFICE_365	SKYPE	SSL	TWITTER	WINDOWS_UPDATE	YAHOO	YOUTUBE
YOUTUBE	9	28	31	1	1	7	114	237	29	131	94	24	30	12	44	92	52	5	32	1032
YAHOO	37	11	40	1	16	6	39	32	32	184	67	30	27	3	52	84	65	6	1200	29
WINDOWS_UPDATE	11	2	29	4	1	9	2	3	23	29	9	22	11	38	4	18	4	1709	6	7
TWITTER	31	47	54	9	14	17	48	60	37	171	80	47	23	7	61	104	1096	15	71	36
SSL	41	68	51	4	29	7	59	100	61	4	11	89	32	12	97	1076	89	9	120	99
SKYPE	18	5	17	0	5	5	16	37	40	114	25	97	59	8	1385	76	43	9	22	16
OFFICE_365	15	4	27	1	1	5	2	2	14	5	6	5	5	923	7	7	2	61	1	6
MSN	51	11	53	5	4	7	16	10	28	73	46	76	1341	12	48	31	17	31	31	13
MICROSOFT	48	21	32	5	11	7	22	9	37	116	45	1182	96	29	127	96	32	50	39	20
HTTP_PROXY	10	23	6	0	8	26	109	64	159	121	1204	34	33	10	20	14	40	57	35	61
HTTP_CONNECT	6	55	0	0	5	30	38	44	23	1252	97	60	51	3	64	1	86	15	108	41
HTTP	108	16	117	3	7	9	14	46	1341	34	114	22	53	12	12	57	18	19	8	14
GOOGLE	22	40	52	1	12	5	247	857	49	120	108	34	31	5	35	111	54	4	33	209
GMAIL	3	16	19	0	4	6	1306	181	15	97	98	27	24	1	33	55	32	5	26	57
FACEBOOK	13	19	22	0	5	1830	2	4	20	35	23	7	9	5	5	8	4	30	1	2
DROPBOX	10	12	19	0	1748	3	16	18	24	43	19	12	6	3	4	45	19	4	44	7
CONTENT_FLASH	62	1	8	1613	0	0	0	0	4	0	0	1	0	0	0	4	1	0	0	0
CLOUDFLARE	34	15	1781	8	4	15	10	4	24	0	0	12	30	10	11	49	8	12	7	11
APPLE	15	1204	23	4	4	8	5	8	30	25	3	32	12	4	17	60	8	10	14	7
AMAZON	1299	22	105	58	25	11	9	13	65	49	38	24	46	19	28	111	35	10	33	9

Conclusions

Looking at the visual representations and observations, here are some high level conclusions:

- In general, across all the models, the feature selection using IQR has performed much better than the PCA mechanism
- The best models are
    i. XGBoost (Accuracy = 0.68)
    ii. Random Forest (Accuracy = 0.67)
- Both operating on the features that were selected using the IQR method.
- DecisionTree performed very poorly.
- SVM was the slowest model

Interpretability

Based on techniques like Feature Importance and SHAP (SHapley Additive exPlanations), the following features were found to be more impactful in determining the applications

- Init_Win_bytes_backward
- Init_Win_bytes_forward
- min_seg_size_forward
- Init_Win_bytes_forward
- Fwd.Packet.Length.Max
- Flow.IAT.Max
- Flow.IAT.Mean

Business Impact

Misclassifying applications may lead to real money loss. It is hard to quantify the amount of loss since it may involve anything ranging from not able to detect malicious packets to not providing adequate level of service level agreements to the customer. To this end, reducing both False-Positives and False-Negatives are important. Hence focus should be on Accuracy. Both precision and recall needs to be maximized.
The models described here could be deployed in the following types of business scenarios:

1. Non critical
2. Best effort

Future

The best models above were still not good enough. I am sure, it can be improved with a GridSearch on parameters and some of the following:

- More advanced feature selection techniques could be used in the future, like gain ration (GR) based techniques.
- Also better cross-validation techniques (like N-fold) and Grid search could be employed to tune the model much better.
- Optimize on Multiclass Receiver Operating Characteristics (multi-class RoC)
- Deep learning algorithms, such as Convolution Neural Networks (CNN), have proven their efficiency through the unnecessity of extracting any statistical feature and through their reliance on the employment of the raw network traffic as their input. A future direction could be using deel learning techniques to classify traffic and identify causal applications.

Further information and Contact

- Link to download data
- https://github.com/1kit/Berkeley-Capstone-Project

Venkit Kasiviswanathan
Email: venkitk@gmail.com
LinkedIn