Malware Classification

A Machine Learning Approach

Capstone 2 Presentation

Presented by: Alex Eldredge

About Me

- Love skiing, hiking, and everything outdoors!
- Previous job working in quality and company analytics
- Led a team of 21 analysts with multi-department focuses

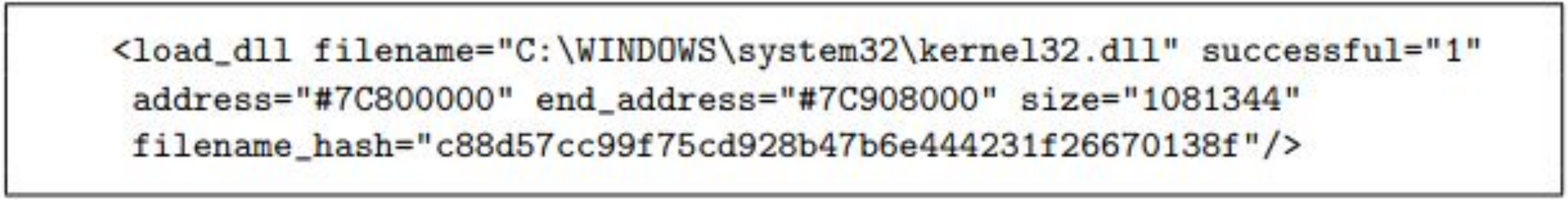
Motivation and Background

Malware and cyber attacks are a persisting concern for businesses, governments, and individuals. One of the most important things that can be done to combat this threat is to stop malicious code and the processes they execute before they can do harm or collect sensitive data. A relatively unexplored way of doing this is classifying malware through machine learning on both static and dynamic features.

Objective:

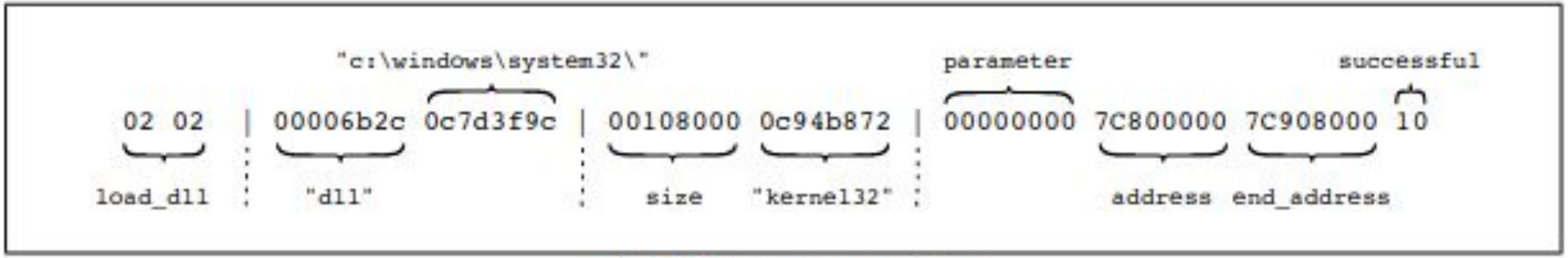
My objective though this project is to see if it is possible to predict malware type using static and dynamic feature based machine learning with a decent degree of accuracy.

Data Source and Intake



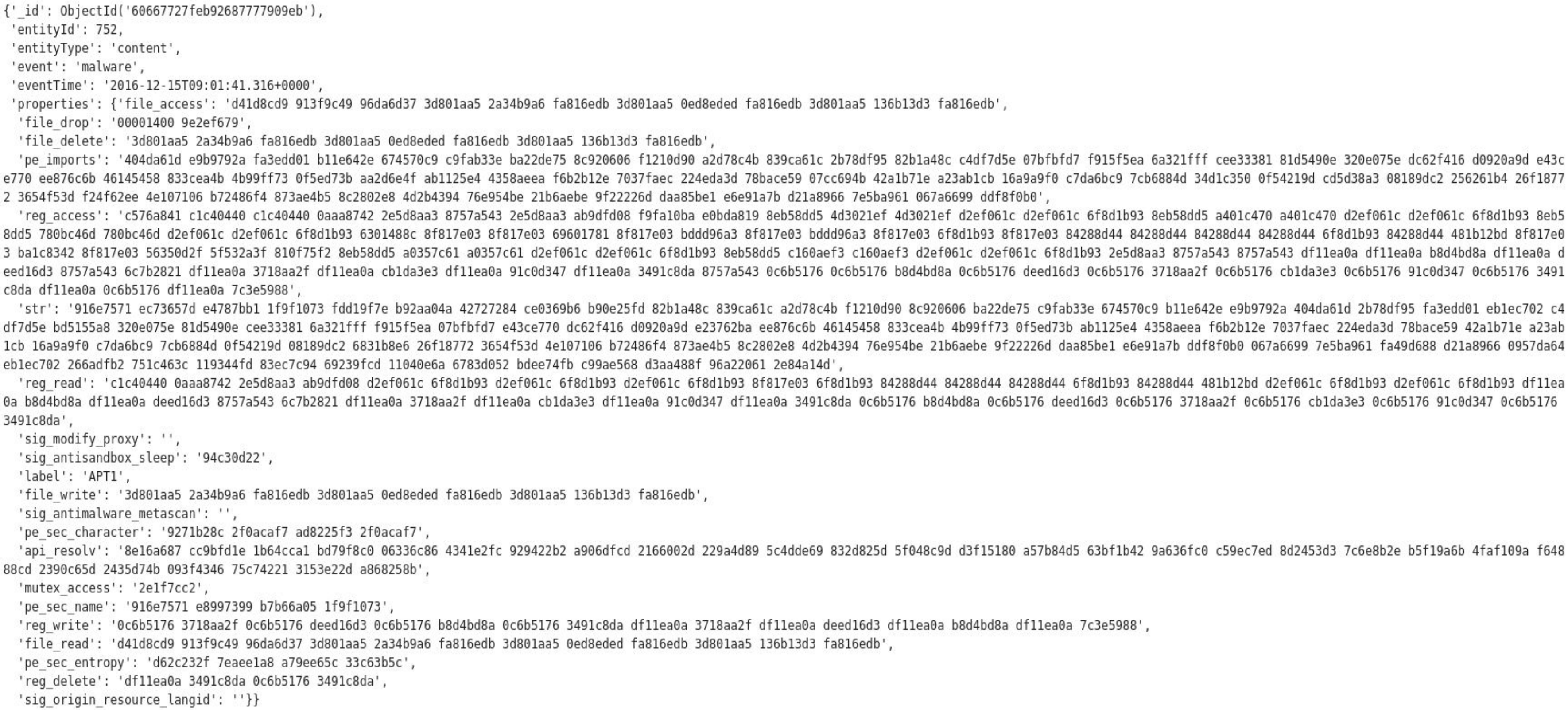
```
<load_dll filename="C:\WINDOWS\system32\kernel32.dll" successful="1"
 address="#7C800000" end_address="#7C908000" size="1081344"
 filename_hash="c88d57cc99f75cd928b47b6e444231f26670138f"/>
```

(a) CWSandbox representation



(b) MIST representation

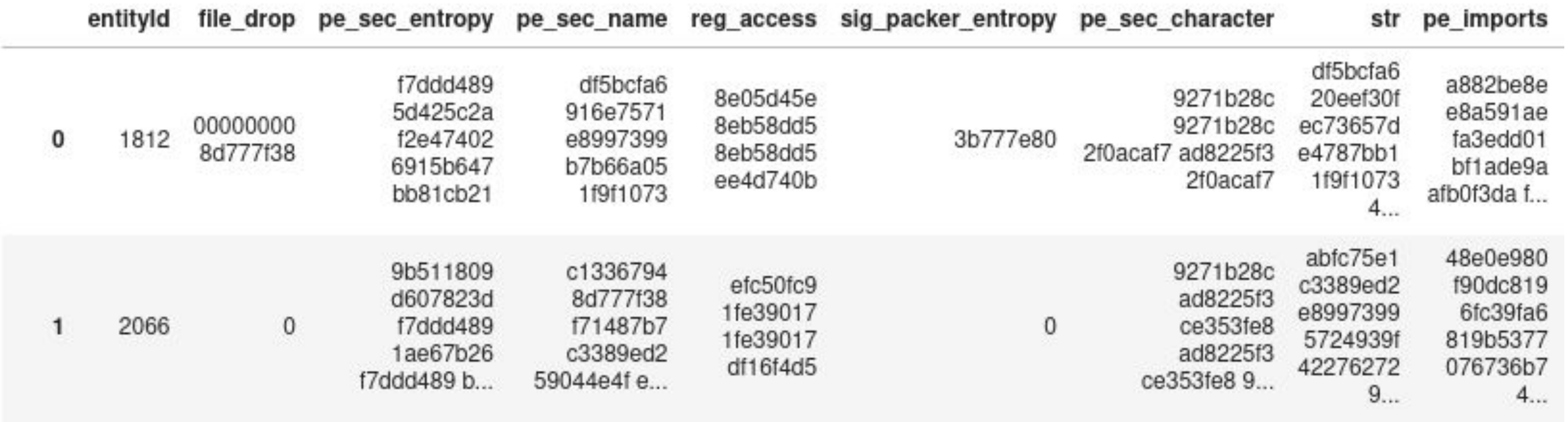
Data Source and Intake



```
{'_id': ObjectId('60667727feb92687777909eb'),
 'entityId': 752,
 'entityType': 'content',
 'event': 'malware',
 'eventTime': '2016-12-15T09:01:41.316+0000',
 'properties': {'file_access': 'd41d8cd9 913f9c49 96da6d37 3d801aa5 2a34b9a6 fa816edb 3d801aa5 0ed8eded fa816edb 3d801aa5 136b13d3 fa816edb',
  'file_drop': '00001400 9e2ef679',
  'file_delete': '3d801aa5 2a34b9a6 fa816edb 3d801aa5 0ed8eded fa816edb 3d801aa5 136b13d3 fa816edb',
  'pe_imports': '404da61d e9b9792a fa3edd01 b11e642e 674570c9 c9fab33e ba22de75 8c920606 f1210d90 a2d78c4b 839ca61c 2b78df95 82b1a48c c4df7d5e 07bfbfd7 f915f5ea 6a321fff cee33381 81d5490e 320e075e dc62f416 d0920a9d e43c
e770 ee876c6b 46145458 833cea4b 4b99ff73 0f5ed73b aa2d6e4f ab1125e4 4358aeea f6b2b12e 7037faec 224eda3d 78bace59 07cc694b 42a1b71e a23ab1cb 16a9a9f0 c7da6bc9 7cb6884d 34d1c350 0f54219d cd5d38a3 08189dc2 256261b4 26f1877
2 3654f53d f24f62ee 4e107106 b72486f4 873ae4b5 8c2802e8 4d2b4394 76e954be 21b6aebe 9f22226d daa85be1 e6e91a7b d21a8966 7e5ba961 067a6699 ddf8f0b0',
  'reg_access': 'c576a841 c1c40440 c1c40440 0aaa8742 2e5d8aa3 8757a543 2e5d8aa3 ab9dfd08 f9fa10ba e0bda819 8eb58dd5 4d3021ef 4d3021ef d2ef061c d2ef061c 6f8d1b93 8eb58dd5 a401c470 a401c470 d2ef061c d2ef061c 6f8d1b93 8eb5
8dd5 780bc46d 780bc46d d2ef061c d2ef061c 6f8d1b93 6301488c 8f817e03 8f817e03 69601781 8f817e03 bddd96a3 8f817e03 bddd96a3 8f817e03 6f8d1b93 8f817e03 84288d44 84288d44 84288d44 84288d44 6f8d1b93 84288d44 481b12bd 8f817e0
3 ba1c8342 8f817e03 56350d2f 5f532a3f 810f75f2 8eb58dd5 a0357c61 a0357c61 d2ef061c d2ef061c 6f8d1b93 8eb58dd5 c160aef3 c160aef3 d2ef061c d2ef061c 6f8d1b93 2e5d8aa3 8757a543 8757a543 df11ea0a df11ea0a b8d4bd8a df11ea0a d
eed16d3 8757a543 6c7b2821 df11ea0a 3718aa2f df11ea0a cb1da3e3 df11ea0a 91c0d347 df11ea0a 3491c8da 8757a543 0c6b5176 0c6b5176 b8d4bd8a 0c6b5176 deed16d3 0c6b5176 3718aa2f 0c6b5176 cb1da3e3 0c6b5176 91c0d347 0c6b5176 3491
c8da df11ea0a 0c6b5176 df11ea0a 7c3e5988',
  'str': '916e7571 ec73657d e4787bb1 1f9f1073 fdd19f7e b92aa04a 42727284 ce0369b6 b90e25fd 82b1a48c 839ca61c a2d78c4b f1210d90 8c920606 ba22de75 c9fab33e 674570c9 b11e642e e9b9792a 404da61d 2b78df95 fa3edd01 eb1ec702 c4
df7d5e bd5155a8 320e075e 81d5490e cee33381 6a321fff f915f5ea 07bfbfd7 e43ce770 dc62f416 d0920a9d e23762ba ee876c6b 46145458 833cea4b 4b99ff73 0f5ed73b ab1125e4 4358aeea f6b2b12e 7037faec 224eda3d 78bace59 42a1b71e a23ab
1cb 16a9a9f0 c7da6bc9 7cb6884d 0f54219d 08189dc2 6831b8e6 26f18772 3654f53d 4e107106 b72486f4 873ae4b5 8c2802e8 4d2b4394 76e954be 21b6aebe 9f22226d daa85be1 e6e91a7b ddf8f0b0 067a6699 7e5ba961 fa49d688 d21a8966 0957da64
eb1ec702 266adfb2 751c463c 119344fd 83ec7c94 69239fcd 11040e6a 6783d052 bdee74fb c99ae568 d3aa488f 96a22061 2e84a14d',
  'reg_read': 'c1c40440 0aaa8742 2e5d8aa3 ab9dfd08 d2ef061c 6f8d1b93 d2ef061c 6f8d1b93 d2ef061c 6f8d1b93 8f817e03 6f8d1b93 84288d44 84288d44 84288d44 6f8d1b93 84288d44 481b12bd d2ef061c 6f8d1b93 d2ef061c 6f8d1b93 df11ea
0a b8d4bd8a df11ea0a deed16d3 8757a543 6c7b2821 df11ea0a 3718aa2f df11ea0a cb1da3e3 df11ea0a 91c0d347 df11ea0a 3491c8da 0c6b5176 b8d4bd8a 0c6b5176 deed16d3 0c6b5176 3718aa2f 0c6b5176 cb1da3e3 0c6b5176 91c0d347 0c6b5176
3491c8da',
  'sig_modify_proxy': '',
  'sig_antisandbox_sleep': '94c30d22',
  'label': 'APT1',
  'file_write': '3d801aa5 2a34b9a6 fa816edb 3d801aa5 0ed8eded fa816edb 3d801aa5 136b13d3 fa816edb',
  'sig_antimalware_metascan': '',
  'pe_sec_character': '9271b28c 2f0acaf7 ad8225f3 2f0acaf7',
  'api_resolv': '8e16a687 cc9bfd1e 1b64cca1 bd79f8c0 06336c86 4341e2fc 929422b2 a906dfcd 2166002d 229a4d89 5c4dde69 832d825d 5f048c9d d3f15180 a57b84d5 63bf1b42 9a636fc0 c59ec7ed 8d2453d3 7c6e8b2e b5f19a6b 4faf109a f648
88cd 2390c65d 2435d74b 093f4346 75c74221 3153e22d a868258b',
  'mutex_access': '2e1f7cc2',
  'pe_sec_name': '916e7571 e8997399 b7b66a05 1f9f1073',
  'reg_write': '0c6b5176 3718aa2f 0c6b5176 deed16d3 0c6b5176 b8d4bd8a 0c6b5176 3491c8da df11ea0a 3718aa2f df11ea0a deed16d3 df11ea0a b8d4bd8a df11ea0a 7c3e5988',
  'file_read': 'd41d8cd9 913f9c49 96da6d37 3d801aa5 2a34b9a6 fa816edb 3d801aa5 0ed8eded fa816edb 3d801aa5 136b13d3 fa816edb',
  'pe_sec_entropy': 'd62c232f 7eaee1a8 a79ee65c 33c63b5c',
  'reg_delete': 'df11ea0a 3491c8da 0c6b5176 3491c8da',
  'sig_origin_resource_langid': ''}}
```

Data Source and Intake

mongoDB
pandas

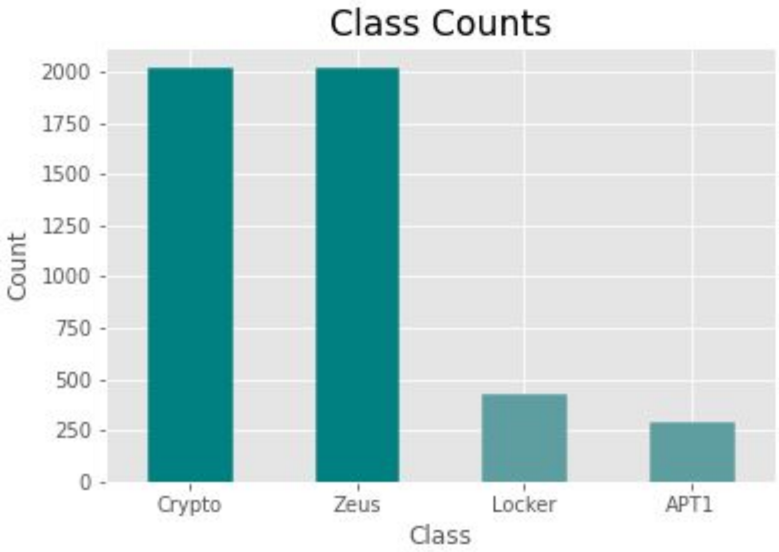


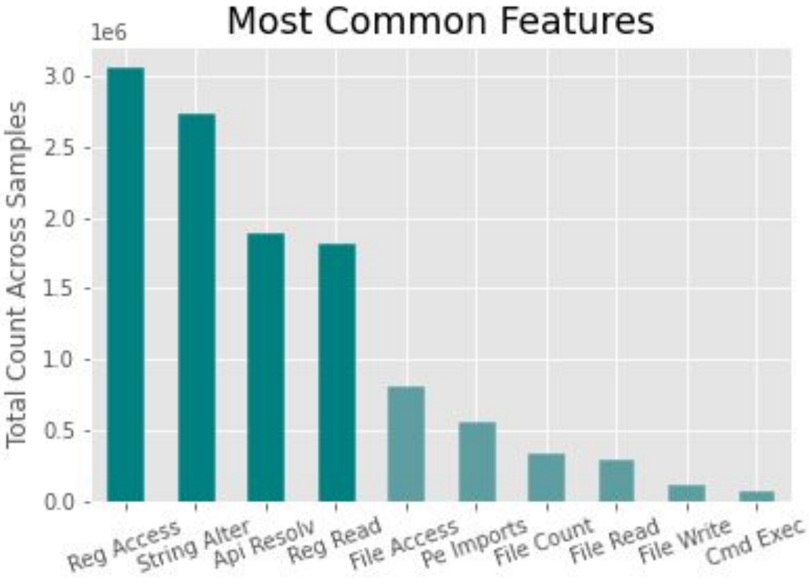
	entityId	file_drop	pe_sec_entropy	pe_sec_name	reg_access	sig_packer_entropy	pe_sec_character	str	pe_imports
0	1812	00000000 8d777f38	f7ddd489 5d425c2a f2e47402 6915b647 bb81cb21	df5bcfa6 916e7571 e8997399 b7b66a05 1f9f1073	8e05d45e 8eb58dd5 8eb58dd5 ee4d740b	3b777e80	9271b28c 9271b28c 2f0acaf7 ad8225f3 2f0acaf7	df5bcfa6 20eef30f ec73657d e4787bb1 1f9f1073 4...	a882be8e e8a591ae fa3edd01 bf1ade9a afb0f3da f...
1	2066	0	9b511809 d607823d f7ddd489 1ae67b26 f7ddd489 b...	c1336794 8d777f38 f71487b7 c3389ed2 59044e4f e...	efc50fc9 1fe39017 1fe39017 df16f4d5	0	9271b28c ad8225f3 ce353fe8 ad8225f3 ce353fe8 9...	abfc75e1 c3389ed2 e8997399 5724939f 42276272 9...	48e0e980 f90dc819 6fc39fa6 819b5377 076736b7 4...

EDA

Initial data observations:

- 4 Target classes
- 4762 samples
- 139 features



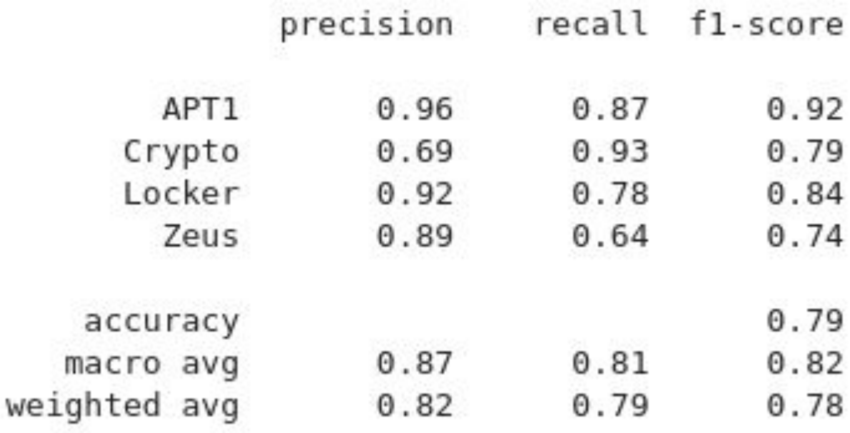


- Unbalanced Classes
- Skewed feature actions/properties

Results

Using a Gradient Boosted Classifier and Random/Halving Cross Validation, best results are as follows:

- Best overall accuracy score ~0.77

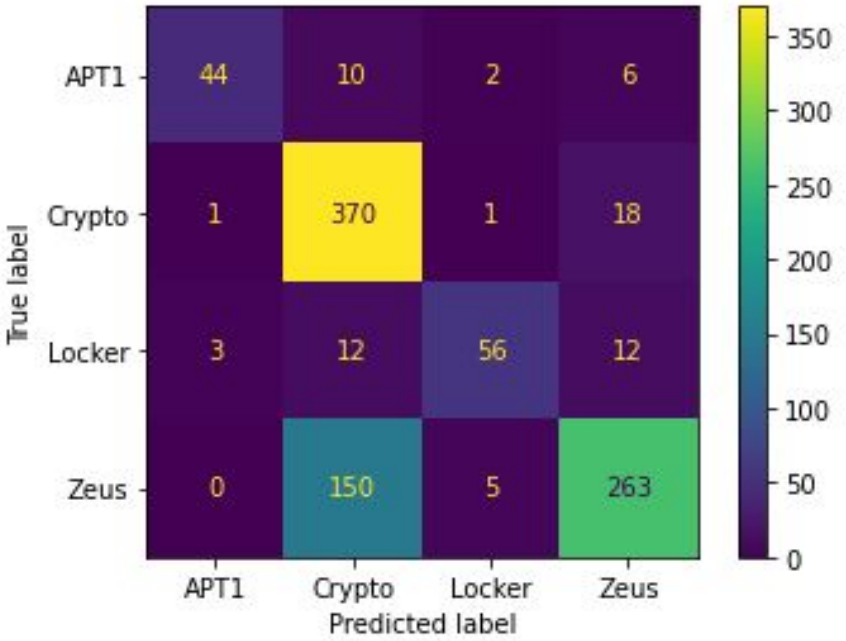


```
              precision    recall  f1-score

        APT1       0.96      0.87      0.92
      Crypto       0.69      0.93      0.79
      Locker       0.92      0.78      0.84
        Zeus       0.89      0.64      0.74

    accuracy                           0.79
   macro avg       0.87      0.81      0.82
weighted avg       0.82      0.79      0.78
```

Confusion Matrix



Conclusions and Next Steps

Overall, the model performs fairly well on the unseen data. It predicts well for advanced persistent threats, crypto ransomware, and locker ransomware. Although this is a good start, I believe there is much optimization to be accomplished with this project.

Future Work

- There is still work that needs to be done to separate the keylogger Zeus
- Adding more balancing to the classes is another avenue to increase performance because the class sizes vary significantly
- Reducing the number of features is also part of the future concept for this project
- In similar concept to the previous bullet, feature manipulation with some of the static features could prove useful in better classification

Questions?

Alexander Eldredge

eldredge.alexander@gmail.com
linkedin.com/in/alexander-eldredge/
github.com/Aeldredge

python

SQL

NumPy

docker

mongoDB

pandas

APACHE Spark

scikit learn

Appendix