Microsoft Malware Detection

Minakshi Mohanty

Introduction

Malware is a collective name for any kind of malicious software, designed to infiltrate and attack systems, servers or gain unauthorized access to networks. Some common types of malwares include:

1. **Virus:** Viruses are designed to damage its target computer. They can cause data corruption, formatting of hard disk, or complete shutdown of system.
2. **Worm:** Worms are one of the most common forms of malware. They are often used to execute a payload. A payload is a piece of code that can delete files on a host system, encrypt data for a ransomware attack, steal information, and create botnets.
3. **Trojan Horse:** Trojan horse enters the host system disguised as a normal harmless file or a program that can trick users into downloading and installing it. Once installed, it gives access to cyber criminals to steal data, install more malware, modify files, monitor user activity and conduct denial of service (DoS).
4. **Spyware:** Spywares are designed to track user's browsing and internet activity. They get installed on host computer, without their knowledge, either by bundling with legitimate software or trojans. They can monitor user activity, collect keystrokes, store login, account information and financial data.
5. **Ransomware:** This is a type of malware that holds user's data captive and demands payments to release data back. They can restrict user access either by encrypting files on hard drive or locking down the system and displaying messages that demand ransom.

Objective:

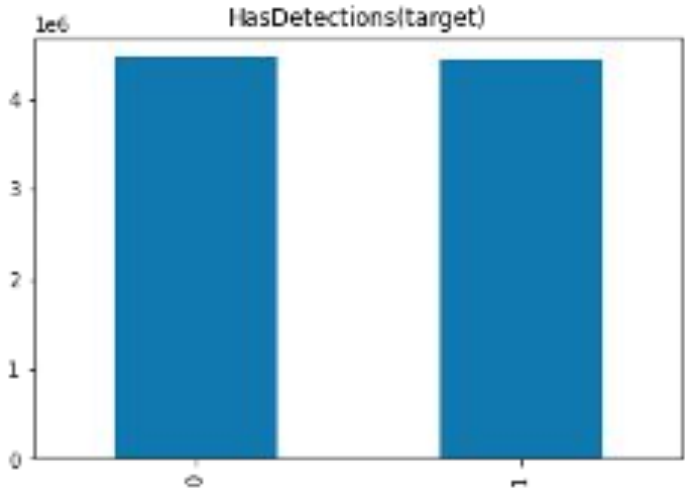
The goal is to predict the presence of a malware on Windows system.

Dataset:

- Microsoft Corporation, Windows Defender ATP Research, Northeastern University College of Computer and Information Science, and Georgia Tech Institute for Information Security & Privacy provides the dataset, through their competition hosted on Kaggle. The link to the competition is: (https://www.kaggle.com/c/microsoft-malware-prediction/data).
- This dataset has been put together by combining heartbeat and threat reports collected by Microsoft's endpoint protection solutions, Windows Defender.
- The data contains properties of the machine and malware infections. There are 82 features in this dataset.
- The target variable is "HasDetections". It indicates that if malware was detected on the machine.
- Each row in the dataset corresponds to a machine that can be uniquely identified by its "MachineIdentifier".

Data Wrangling

- The dataset contains 8921483 entries. The target variable, “HasDetections” has 2 values. 0 corresponds to no malware detection and 1 corresponds to malware detection. There are 4462591entries in class 0 and 4458892 entries in class 1. The dataset is balanced.



Histogram of “HasDetection”

Missing Data

	Total	Percent
PuaMode	8919174	99.974119
Census_ProcessorClass	8884852	99.589407
DefaultBrowsersIdentifier	8488045	95.141637
Census_IsFlightingInternal	7408759	83.044030
Census_InternalBatteryType	6338429	71.046809
Census_ThresholdOptIn	5667325	63.524472
Census_IsWIMBootEnabled	5659703	63.439038
SmartScreen	3177011	35.610795
OrganizationIdentifier	2751518	30.841487
SMode	537759	6.027686

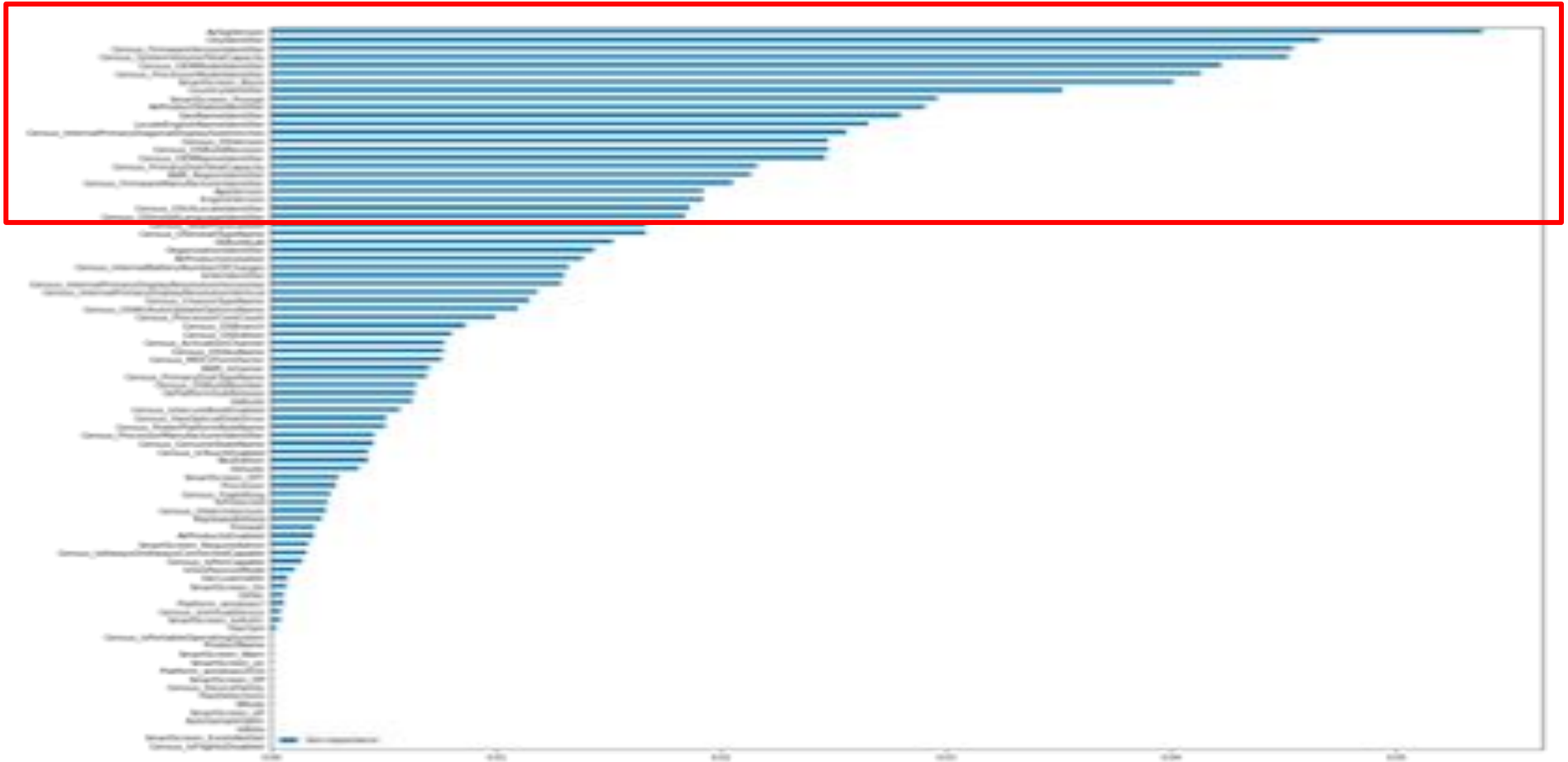
- PuaMode, Census_ProcessorClass, DefaultBrowsersIdentifier, Census_IsFlightingInternal and Census_InternalBatteryTyp e have over 60% missing data. These variables can be removed from the analysis.

Categorical Variables: Encoding

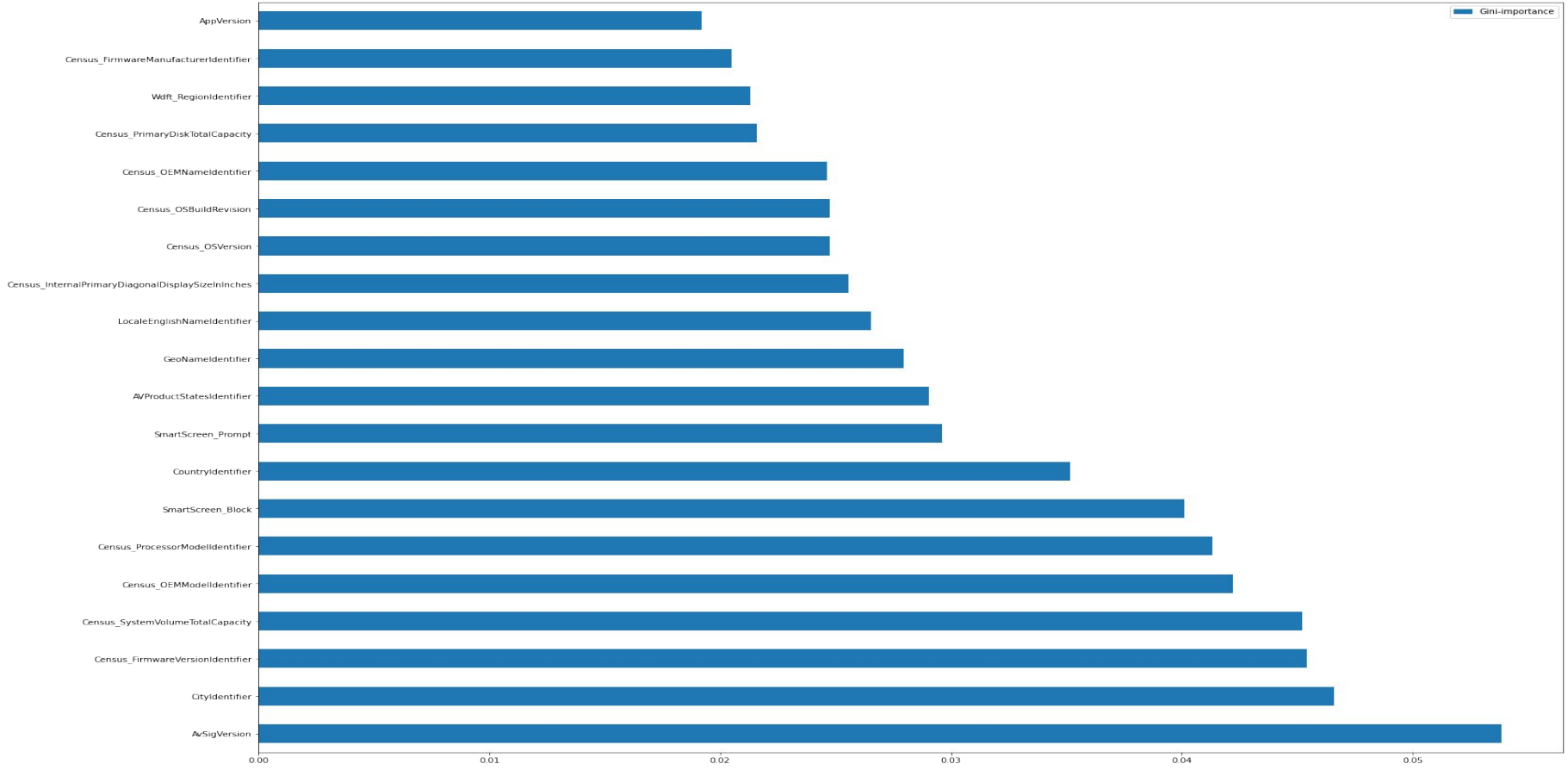
```
ProductName : 6 categories
EngineVersion : 70 categories
AppVersion : 110 categories
AvSigVersion : 8531 categories
Platform : 4 categories
Processor : 3 categories
OsVer : 58 categories
OsPlatformSubRelease : 9 categories
OsBuildLab : 664 categories
SkuEdition : 8 categories
SmartScreen : 22 categories
Census_MDC2FormFactor : 13 categories
Census_DeviceFamily : 3 categories
Census_PrimaryDiskTypeName : 5 categories
Census_ChassisTypeName : 53 categories
Census_PowerPlatformRoleName : 11 categories
Census_OSVersion : 469 categories
Census_OSArchitecture : 3 categories
Census_OSBranch : 32 categories
Census_OSEdition : 33 categories
Census_OSSkuName : 30 categories
Census_OSInstallTypeName : 9 categories
Census_OSWUAutoUpdateOptionsName : 6 categories
Census_GenuineStateName : 5 categories
Census_ActivationChannel : 6 categories
Census_FlightRing : 10 categories
```

- After a careful analysis of each and every variable, it can be seen that all the variables are of category datatype. Most of the variables have very high cardinality. Below, is an image showing the number of categories in some variables.
- For variables with more that 2 unique categories, we use frequency encoding. For others, one-hot encoding is done. The prepared data is saved to a csv file for easier access. Due to hardware limitations, we use on a part of the entire dataset for our modelling.

Feature Engineering - Feature importance using random forest classifier



Top 20 features



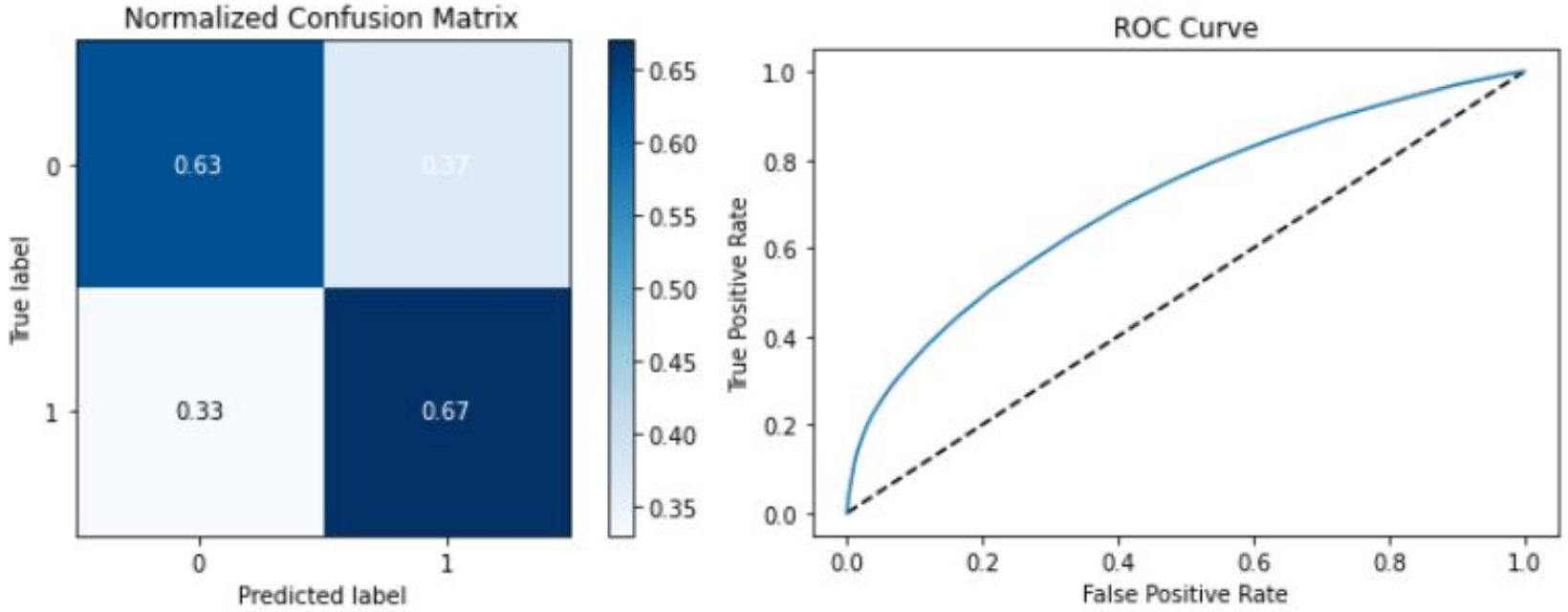
Top 20 features

	Gini-importance
AvSigVersion	0.053834
CityIdentifier	0.046586
Census_FirmwareVersionIdentifier	0.045409
Census_SystemVolumeTotalCapacity	0.045213
Census_OEMModelIdentifier	0.042214
Census_ProcessorModelIdentifier	0.041319
SmartScreen_Block	0.040108
CountryIdentifier	0.035146
SmartScreen_Prompt	0.029595
AVProductStatesIdentifier	0.029032
GeoNameIdentifier	0.027955
LocaleEnglishNameIdentifier	0.026538
Census_InternalPrimaryDiagonalDisplaySizeInInches	0.025540
Census_OSVersion	0.024762
Census_OSBuildRevision	0.024733
Census_OEMNameIdentifier	0.024640
Census_PrimaryDiskTotalCapacity	0.021590
Wdft_RegionIdentifier	0.021293
Census_FirmwareManufacturerIdentifier	0.020493
AppVersion	0.019204

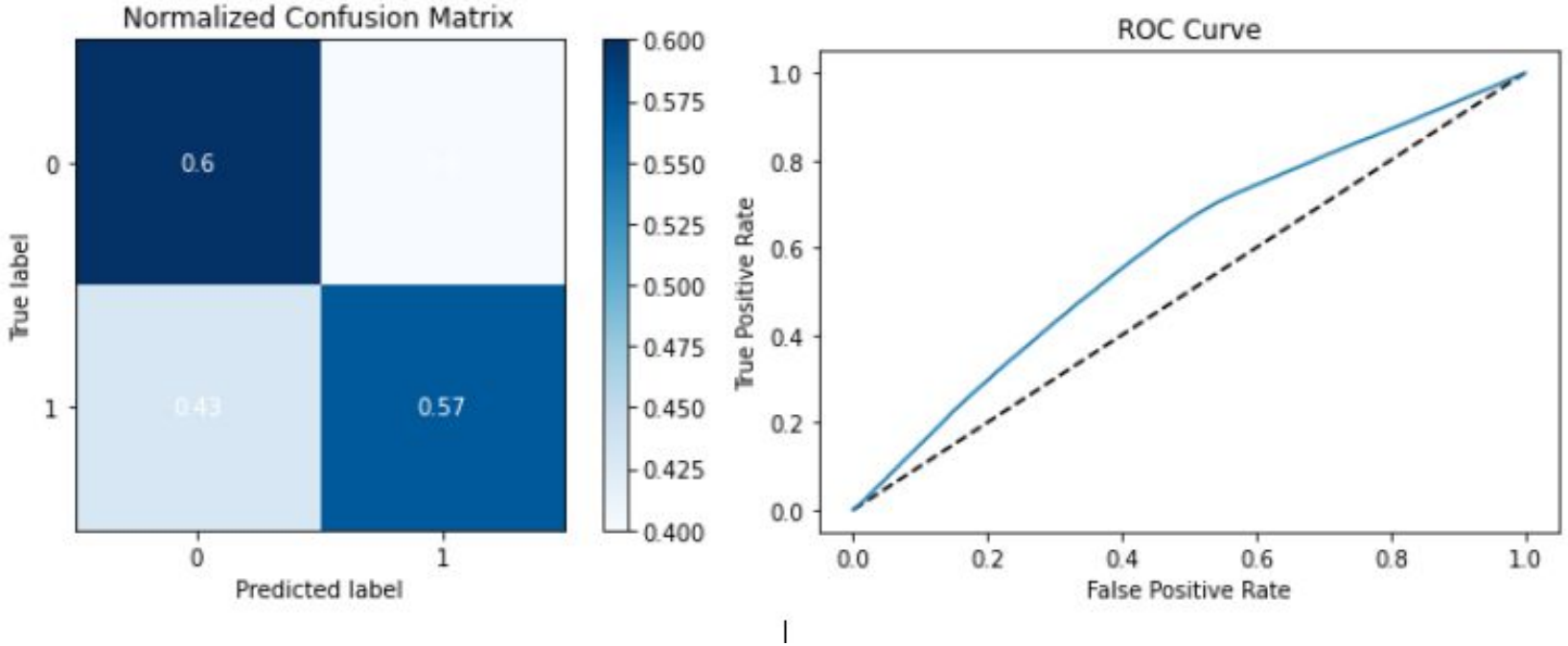
Machine Learning Modeling and Optimization:

1. Random Forest Classifier
2. Logistic Regression
3. AdaBoost Classifier with Decision Tree Classifier as Base Estimator
4. AdaBoost Classifier with Logistic Regression as Base Estimator

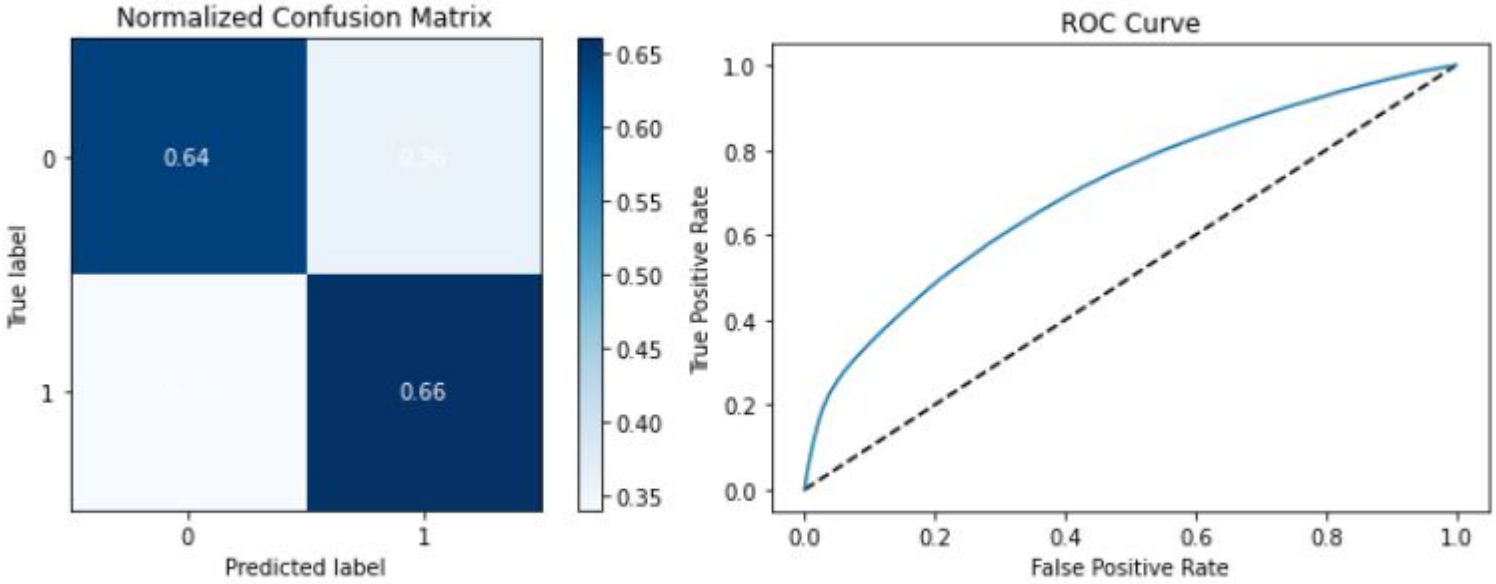
Random Forest Classifier



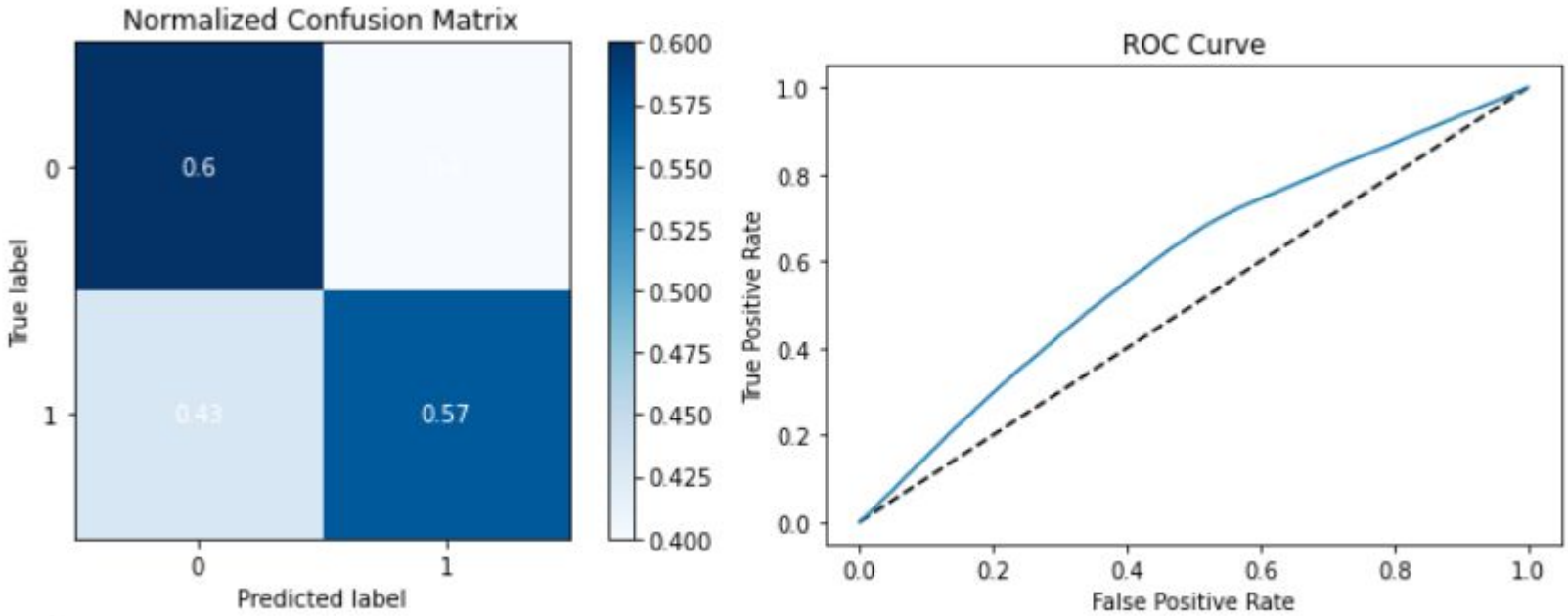
Logistic Regression



AdaBoost Classifier with Decision Trees as Base Estimator



AdaBoost Classifier with Logistic Regression as Base Estimator



Model Evaluation

Random forest classifier and AdaBoost classifier with Decision Tree as base estimator have the same performance. Logistic regression and AdaBoost classifier with Logistic Regression as base estimator have same performance, but lower than that of random forest and AdaBoost.

Model	**Accuracy**	**AUC**	**Precision (weighted avg)**	**Recall (weighted avg)**
Random Forest Classifier	0.65	0.70	0.65	0.65
Logistic Regression	0.58	0.59	0.61	0.58
AdaBoost Classifier (Decision Tree)	0.65	0.70	0.65	0.65
AdaBoost Classifier (Logistic Regression)	0.58	0.59	0.61	0.58

Conclusion

The models showed average performance. For future work, a combination of more feature engineering methods can be used to select the best features. Other boosting models such as XGBoost, Gradient boost classifiers can also be tested. Due to limitations in hardware, only a part of data was used for modeling. This can be addressed in future revisions.

References

1. Cover page image courtesy: https://www.hellotech.com/blog/how-to-remove-malware-from-windows-10
2. Scikit-learn Random Forest Classifier : https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html