>> **Final Project:**

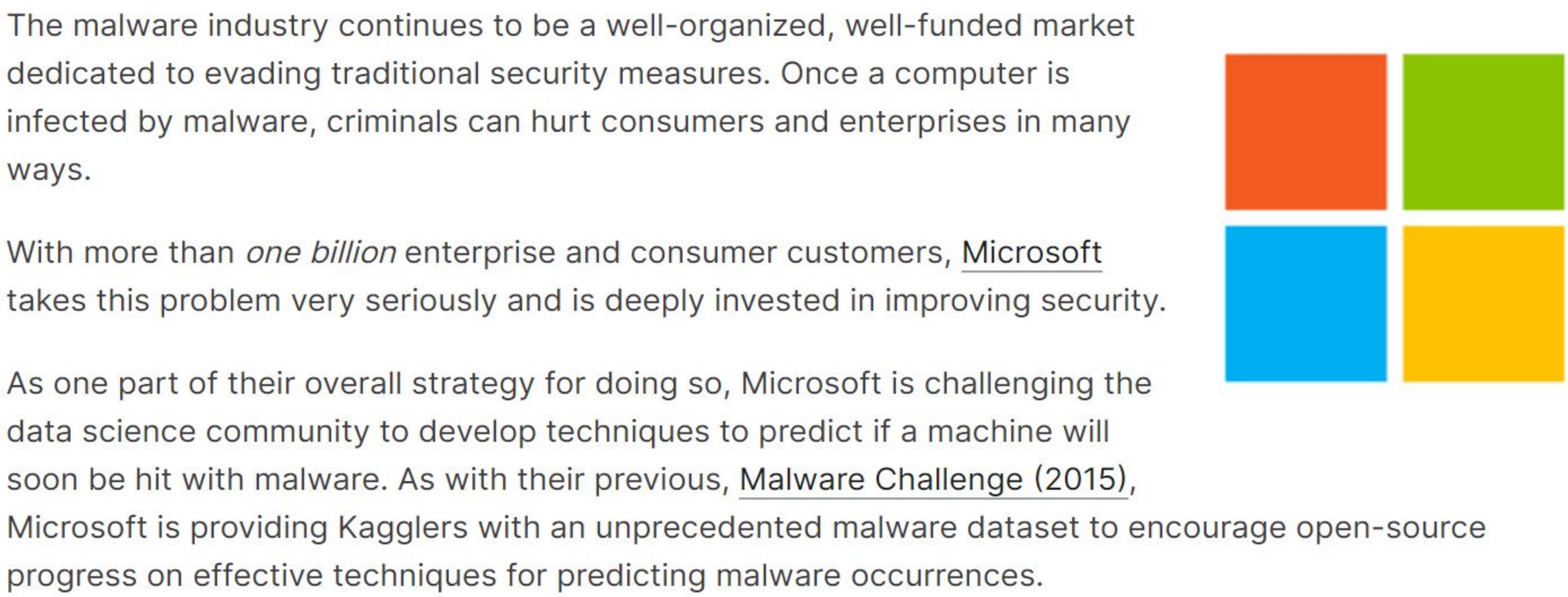
>> Microsoft Malware Prediction

>> Justin Ross and Abhinav Vittal

>> Abstract

This project is an attempt at Microsoft's Malware Prediction Kaggle Competition from 2018. Using a specific train and test datasets, the goal of this project is to predict whether a machine learning model can predict whether a computer has malware given various information about the model. The test and train datasets provided contain information about real Microsoft machines and their malware status. We evaluate performance using area under ROC curve.

>> Background: The Problem



The malware industry continues to be a well-organized, well-funded market dedicated to evading traditional security measures. Once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways.

With more than *one billion* enterprise and consumer customers, Microsoft takes this problem very seriously and is deeply invested in improving security.

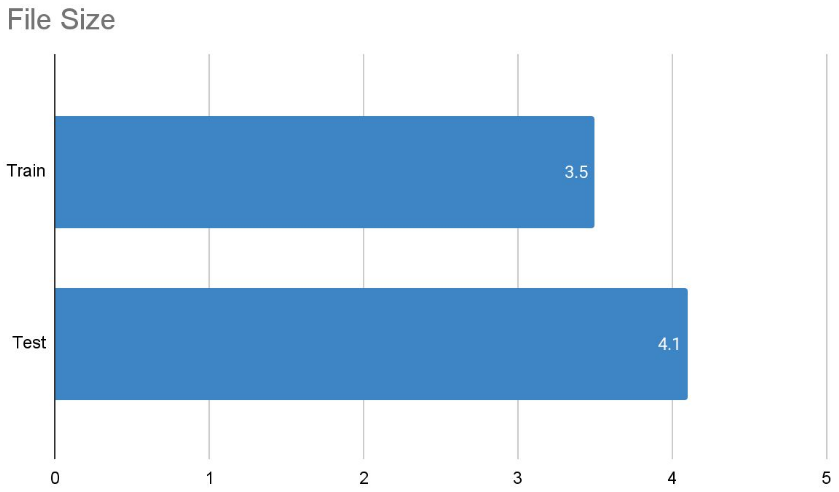
As one part of their overall strategy for doing so, Microsoft is challenging the data science community to develop techniques to predict if a machine will soon be hit with malware. As with their previous, Malware Challenge (2015), Microsoft is providing Kagglers with an unprecedented malware dataset to encourage open-source progress on effective techniques for predicting malware occurrences.

Can you help protect more than one billion machines from damage BEFORE it happens?

>> Why does this matter?

- Cybercrimes to cause $10.5T in damages by 2025
- Personal PCs house sensitive data like SSNs, addresses, and credit card numbers, leaving innocent citizens at the mercy of identity thieves
- Business PCs hold customer data, trade secrets, and unreleased, non-public information that could sway markets

>> The Data



Train: 7 million rows

Test: 9 million rows

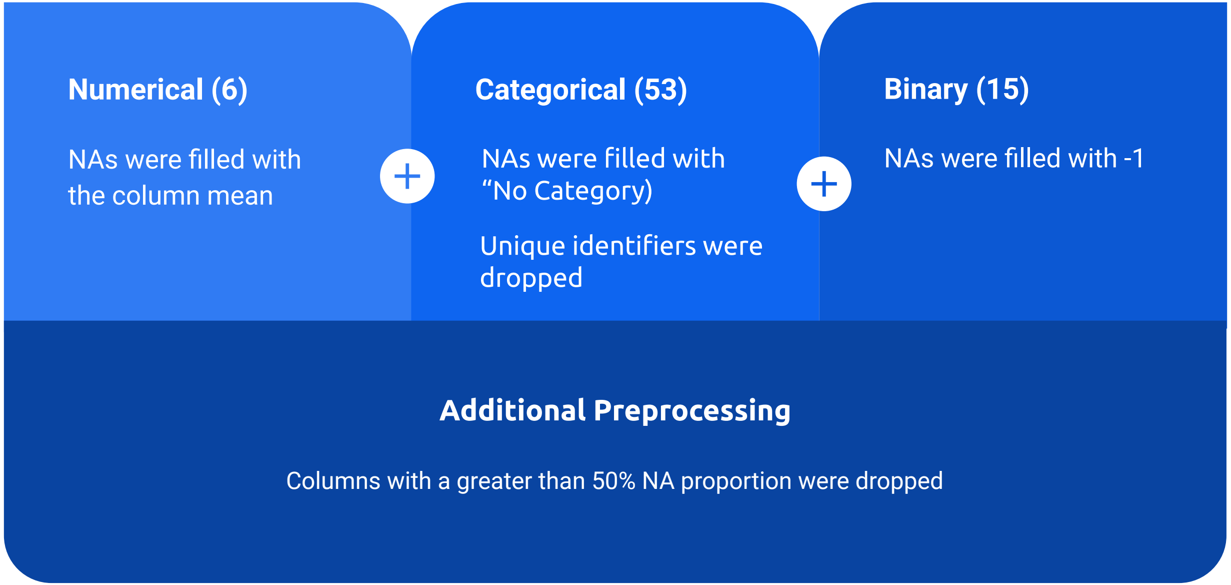
Data Facts:

- Each row corresponds to a machine uniquely identified by a "MachineIdentifier"
- Feature candidates consist of numerical, categorical, and binary columns
- "HasDetections" is the ground truth and indicates malware detection
- 50% of machines in the train data set have detected malware

>> Data Cleaning

Data Subset:

- Random stratified sample of 50,000 rows
- Maintains 50-50 split in target “HasDetection” column
- Necessary due to large data size and limited RAM



Numerical (6)

NAs were filled with the column mean

+

Categorical (53)

NAs were filled with “No Category)

Unique identifiers were dropped

+

Binary (15)

NAs were filled with -1

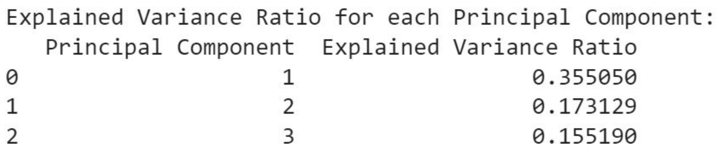
Additional Preprocessing

Columns with a greater than 50% NA proportion were dropped

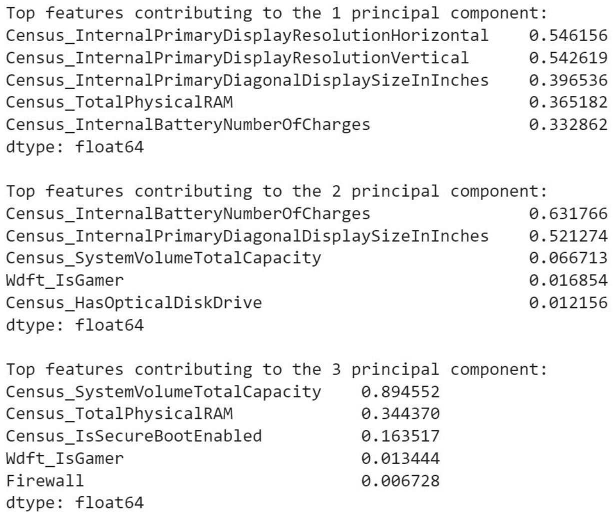
>> Feature Selection: PCA

-Scaled numerical columns to have mean of 0 and st dev of 1

-Performed PCA with numeric and binary columns



```
Explained Variance Ratio for each Principal Component:
   Principal Component  Explained Variance Ratio
0                    1                  0.355050
1                    2                  0.173129
2                    3                  0.155190
```



```
Top features contributing to the 1 principal component:
Census_InternalPrimaryDisplayResolutionHorizontal    0.546156
Census_InternalPrimaryDisplayResolutionVertical      0.542619
Census_InternalPrimaryDiagonalDisplaySizeInInches    0.396536
Census_TotalPhysicalRAM                              0.365182
Census_InternalBatteryNumberOfCharges                0.332862
dtype: float64

Top features contributing to the 2 principal component:
Census_InternalBatteryNumberOfCharges                0.631766
Census_InternalPrimaryDiagonalDisplaySizeInInches    0.521274
Census_SystemVolumeTotalCapacity                     0.066713
Wdft_IsGamer                                         0.016854
Census_HasOpticalDiskDrive                           0.012156
dtype: float64

Top features contributing to the 3 principal component:
Census_SystemVolumeTotalCapacity    0.894552
Census_TotalPhysicalRAM             0.344370
Census_IsSecureBootEnabled          0.163517
Wdft_IsGamer                        0.013444
Firewall                            0.006728
dtype: float64
```

>> Feature Selection: PCA

Process:

- Numeric columns were scaled to have mean 0 and standard deviation 1
- PCA was performed with the numeric and binary columns
- The top three PCs, their explained variance ratios, and the top five features contributing to each PC were obtained

Principal Component	Explained Variance
1	0.355
2	0.173
3	0.155

36 **Principal Component 1**

DisplayResolutionHorizontal
DisplayResolutionVertical
DiagonalDisplayInInches
PhysicalRAM
NumberOfCharges

17 **Principal Component 2**

DisplayResolutionHorizontal
DisplayResolutionVertical
DiagonalDisplayInInches
PhysicalRAM
NumberOfCharges

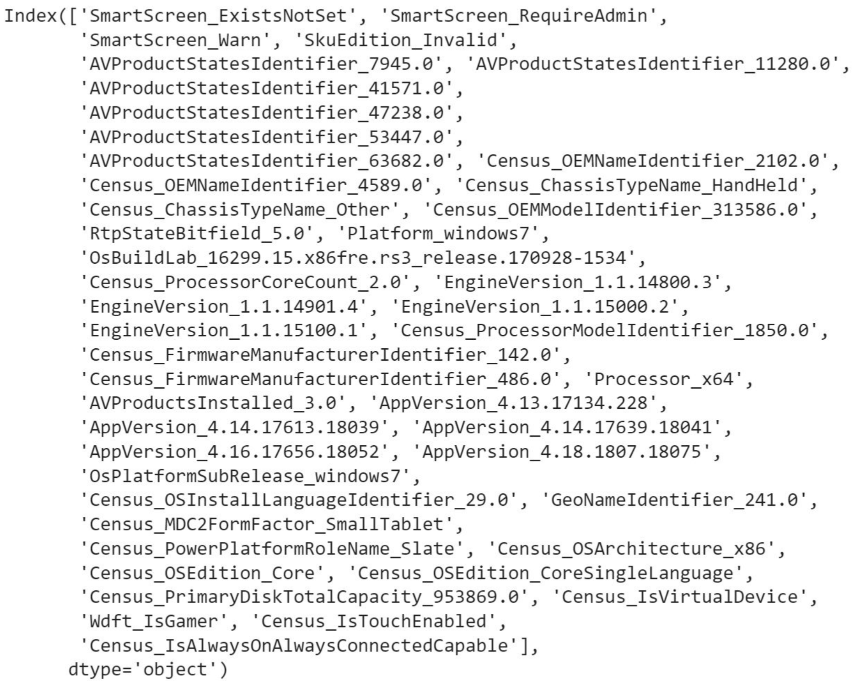
16 **Principal Component 3**

VolumeTotalCapacity
PhysicalRAM
SecureBootEnabled
IsGamer
Firewall

>> Feature Selection: LASSO Regression

Process:

- One-hot encoded categorical columns
- Scaled binary columns
- Set alpha to 0.01 and returned features with a non-zero coefficient
- Returned 46 features

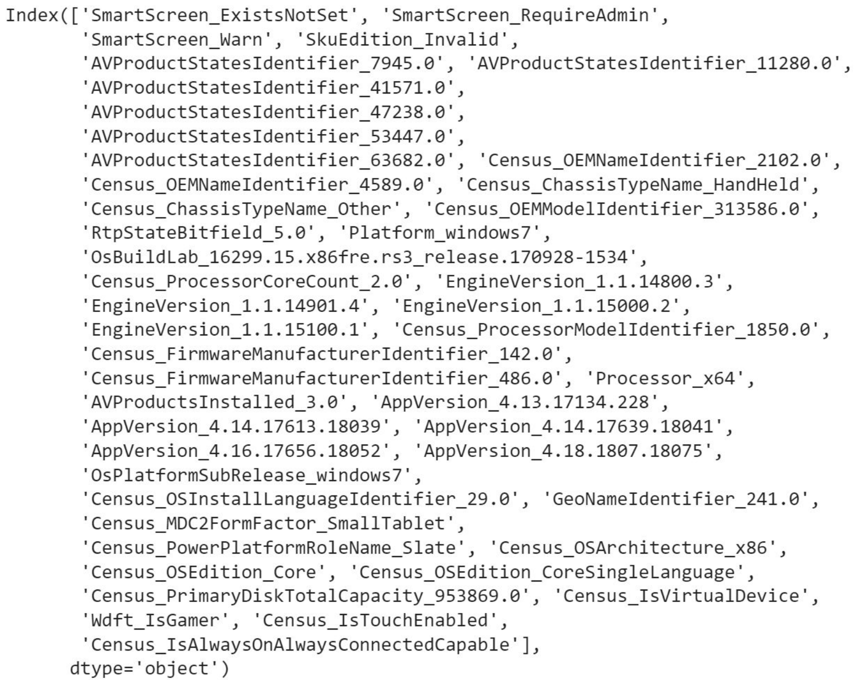


```
Index(['SmartScreen_ExistsNotSet', 'SmartScreen_RequireAdmin',
       'SmartScreen_Warn', 'SkuEdition_Invalid',
       'AVProductStatesIdentifier_7945.0', 'AVProductStatesIdentifier_11280.0',
       'AVProductStatesIdentifier_41571.0',
       'AVProductStatesIdentifier_47238.0',
       'AVProductStatesIdentifier_53447.0',
       'AVProductStatesIdentifier_63682.0', 'Census_OEMNameIdentifier_2102.0',
       'Census_OEMNameIdentifier_4589.0', 'Census_ChassisTypeName_HandHeld',
       'Census_ChassisTypeName_Other', 'Census_OEMModelIdentifier_313586.0',
       'RtpStateBitfield_5.0', 'Platform_windows7',
       'OsBuildLab_16299.15.x86fre.rs3_release.170928-1534',
       'Census_ProcessorCoreCount_2.0', 'EngineVersion_1.1.14800.3',
       'EngineVersion_1.1.14901.4', 'EngineVersion_1.1.15000.2',
       'EngineVersion_1.1.15100.1', 'Census_ProcessorModelIdentifier_1850.0',
       'Census_FirmwareManufacturerIdentifier_142.0',
       'Census_FirmwareManufacturerIdentifier_486.0', 'Processor_x64',
       'AVProductsInstalled_3.0', 'AppVersion_4.13.17134.228',
       'AppVersion_4.14.17613.18039', 'AppVersion_4.14.17639.18041',
       'AppVersion_4.16.17656.18052', 'AppVersion_4.18.1807.18075',
       'OsPlatformSubRelease_windows7',
       'Census_OSInstallLanguageIdentifier_29.0', 'GeoNameIdentifier_241.0',
       'Census_MDC2FormFactor_SmallTablet',
       'Census_PowerPlatformRoleName_Slate', 'Census_OSArchitecture_x86',
       'Census_OSEdition_Core', 'Census_OSEdition_CoreSingleLanguage',
       'Census_PrimaryDiskTotalCapacity_953869.0', 'Census_IsVirtualDevice',
       'Wdft_IsGamer', 'Census_IsTouchEnabled',
       'Census_IsAlwaysOnAlwaysConnectedCapable'],
      dtype='object')
```

>> Final Feature Selection

Findings:

- Numerical data had the strongest effect on PCs in PCA
- Concatenated categorical features to numerical features via LASSO
- Total of 52 features



```
Index(['SmartScreen_ExistsNotSet', 'SmartScreen_RequireAdmin',
       'SmartScreen_Warn', 'SkuEdition_Invalid',
       'AVProductStatesIdentifier_7945.0', 'AVProductStatesIdentifier_11280.0',
       'AVProductStatesIdentifier_41571.0',
       'AVProductStatesIdentifier_47238.0',
       'AVProductStatesIdentifier_53447.0',
       'AVProductStatesIdentifier_63682.0', 'Census_OEMNameIdentifier_2102.0',
       'Census_OEMNameIdentifier_4589.0', 'Census_ChassisTypeName_HandHeld',
       'Census_ChassisTypeName_Other', 'Census_OEMModelIdentifier_313586.0',
       'RtpStateBitfield_5.0', 'Platform_windows7',
       'OsBuildLab_16299.15.x86fre.rs3_release.170928-1534',
       'Census_ProcessorCoreCount_2.0', 'EngineVersion_1.1.14800.3',
       'EngineVersion_1.1.14901.4', 'EngineVersion_1.1.15000.2',
       'EngineVersion_1.1.15100.1', 'Census_ProcessorModelIdentifier_1850.0',
       'Census_FirmwareManufacturerIdentifier_142.0',
       'Census_FirmwareManufacturerIdentifier_486.0', 'Processor_x64',
       'AVProductsInstalled_3.0', 'AppVersion_4.13.17134.228',
       'AppVersion_4.14.17613.18039', 'AppVersion_4.14.17639.18041',
       'AppVersion_4.16.17656.18052', 'AppVersion_4.18.1807.18075',
       'OsPlatformSubRelease_windows7',
       'Census_OSInstallLanguageIdentifier_29.0', 'GeoNameIdentifier_241.0',
       'Census_MDC2FormFactor_SmallTablet',
       'Census_PowerPlatformRoleName_Slate', 'Census_OSArchitecture_x86',
       'Census_OSEdition_Core', 'Census_OSEdition_CoreSingleLanguage',
       'Census_PrimaryDiskTotalCapacity_953869.0', 'Census_IsVirtualDevice',
       'Wdft_IsGamer', 'Census_IsTouchEnabled',
       'Census_IsAlwaysOnAlwaysConnectedCapable'],
      dtype='object')
```

>> Model Selection: Logistic Regression

Process:

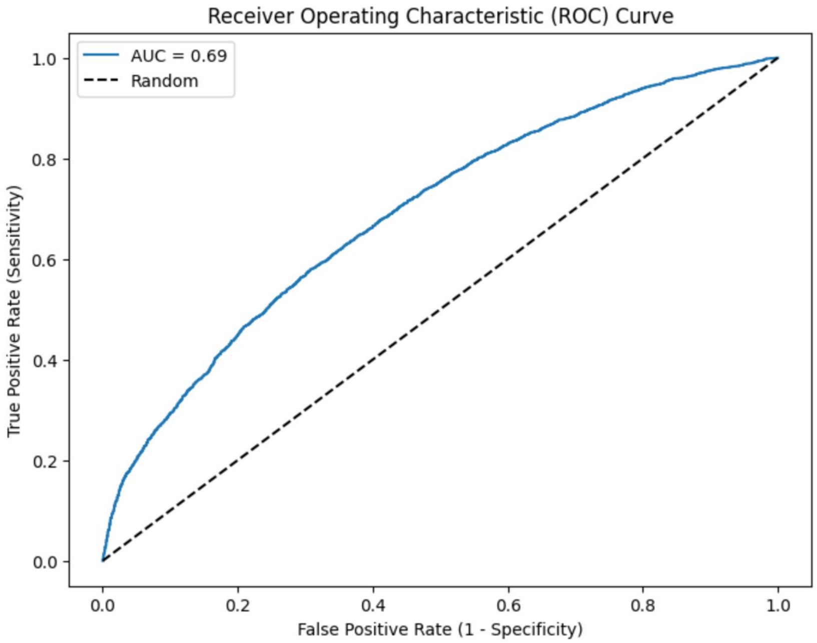
- 70-30 train-test split
- CV with k=5, scored on accuracy

Results:

- Accuracy: 0.6334
- Precision: 0.6362
- Recall: 0.6188
- F1: 0.6274

Confusion Matrix:

3248	1765
1901	3086



Area Under the ROC Curve (AOC): 0.6909

>> Model Selection: KNNClassifier

Process:

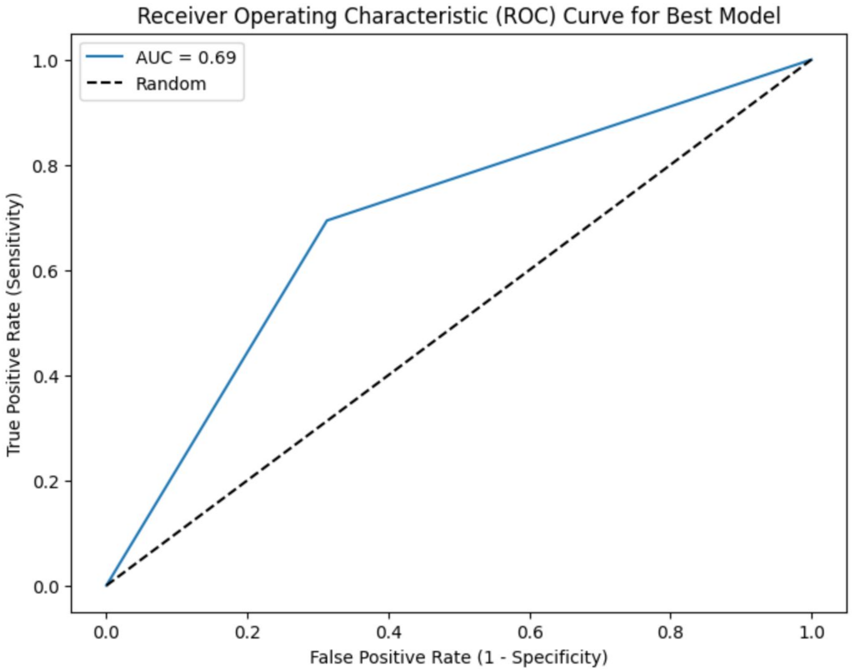
- 70-30 train-test split
- CV with k=5, scored on accuracy

Results:

- Accuracy: 0.6906
- Precision: 0.6882
- Recall: 0.6940
- F1: 0.6911

Confusion Matrix:

3445	1568
1526	3461



Area Under the ROC Curve (AOC): 0.6906

>> Model Selection: GradientBoostingClassifier

Process:

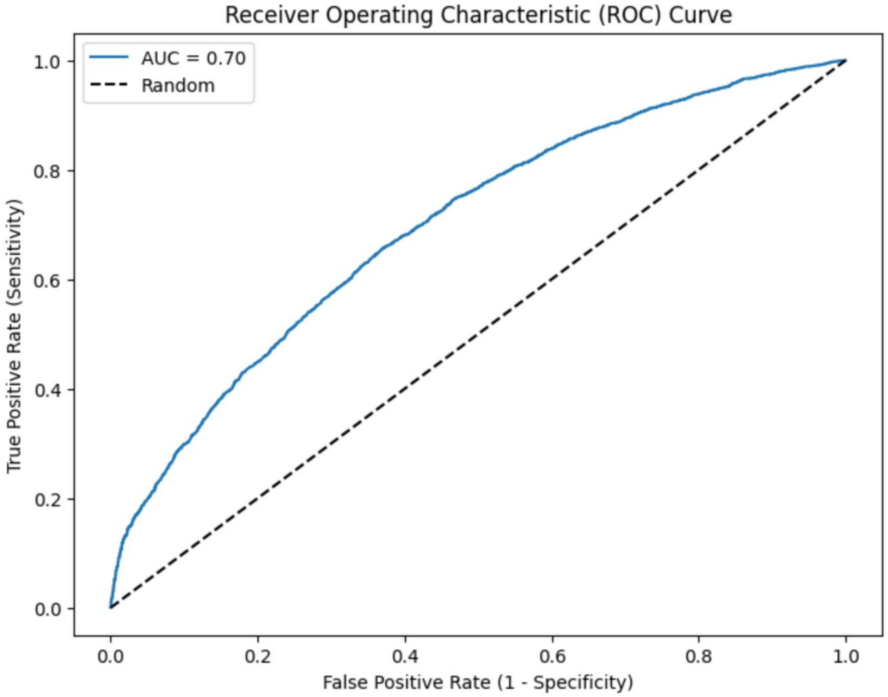
- 70-30 train-test split
- CV with k=5, scored on accuracy
- No hyperparameter optimization due to a lack of compute

Results:

- Accuracy: 0.6414
- Precision: 0.6454
- Recall: 0.6236
- F1: 0.6343

Confusion Matrix:

3304	1709
1877	3110



Area Under the ROC Curve (AOC): 0.6975

>> Model Selection: Random Forest

Process:

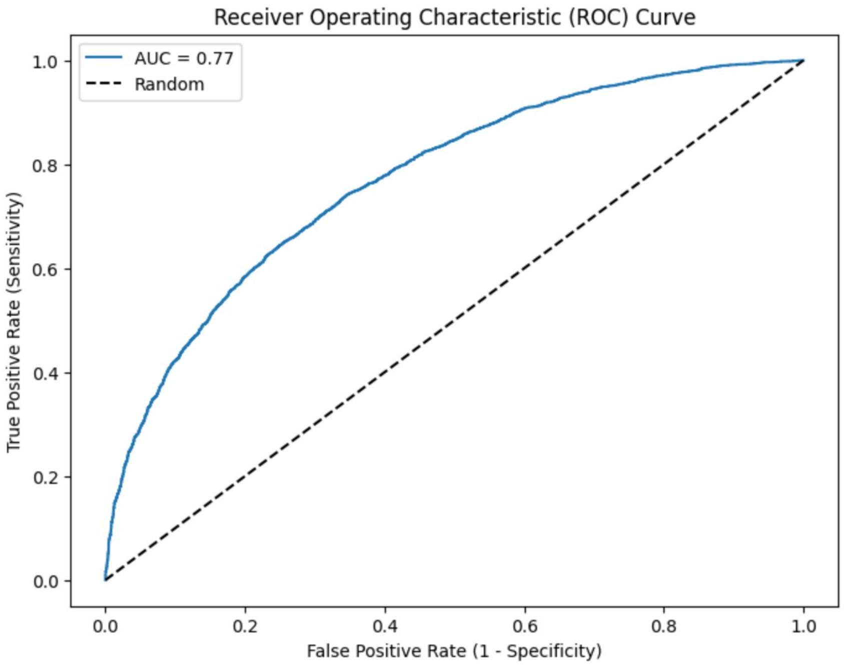
- 70-30 train-test split
- CV with k=5, scored on accuract
- Bayesian hyperparameter optimization
    - max_depth=20
    - min_samples_leaf=1
    - min_samples_split=2
    - n_estimators=100

Results:

- Accuracy: 0.6965
- Precision: 0.6933
- Recall: 0.7018
- F1: 0.6976

Confusion Matrix:

3465	1548
1487	3500



Area Under the ROC Curve (AOC): 0.7715

>> Model Selection: XGBoost

Process:

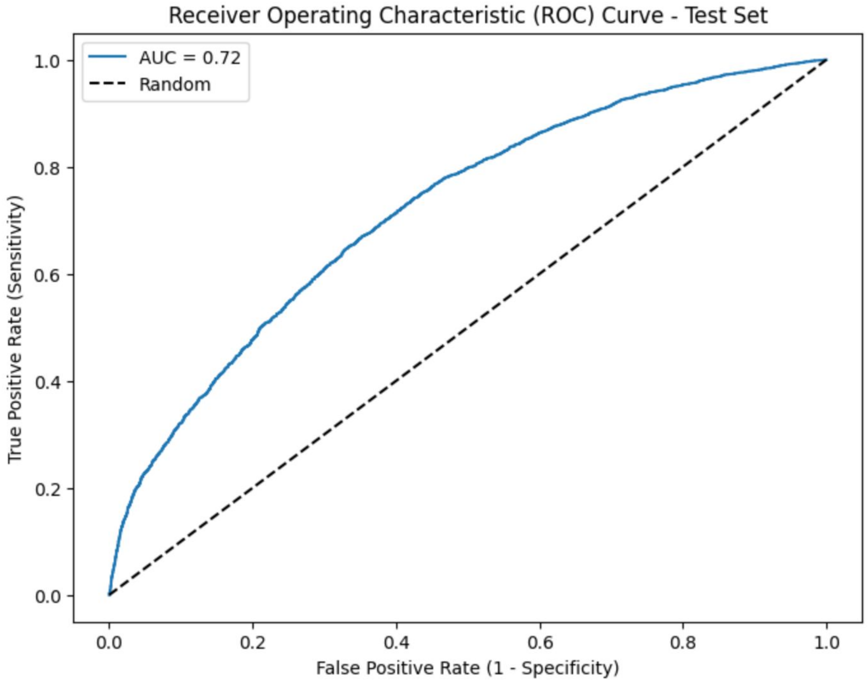
- 70-30 train-test split
- CV with k=5, scored on accuracy
- No hyperparameter optimization due to a lack of compute

Results:

- Accuracy: 0.6582
- Precision: 0.6598
- Recall: 0.6499
- F1: 0.6547

Confusion Matrix:

3341	1672
1746	3241



Area Under the ROC Curve (AOC): 0.7194

>> Model Selection: Best Model

Selection Criteria: AUC-ROC Score

- Singular scalar value that represents the probability will rank a randomly chosen positive instance higher than a randomly chosen negative instance
- Scaled from 0 (no discrimination) to 1 (perfect discrimination)

Selection Criteria: F1 Score

- Harmonic mean of precision and recall scores
- Measure of model accuracy

Model	Area Under ROC Curve	F1 Score
Random Forest	0.7715	0.6976
XGBoost	0.7194	0.6468
GradientBoostingClassifier	0.6975	0.6343
Logistic Regression	0.6908	0.6273
KNN Classifier	0.6906	0.6911

>> Model Validation

Limitations:

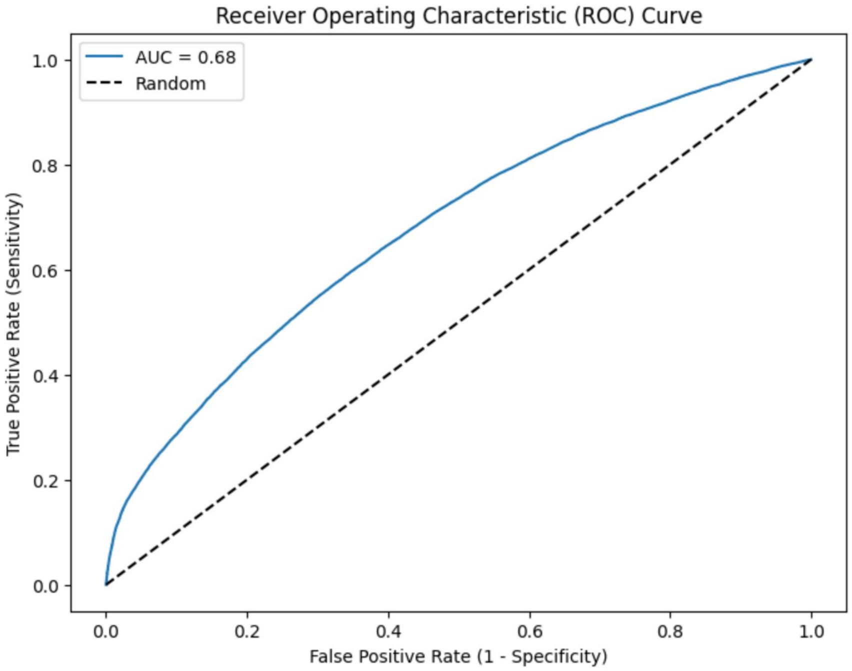
- Low RAM capacity prevented testing on the entire dataset
- Stratified sample of 100,000 rows used to test

Results:

- Accuracy: 0.6243
- Precision: 0.6340
- Recall: 0.5874
- F1: 0.6098

Confusion Matrix:

33004	16919
20589	29311



Area Under the ROC Curve (AUC): 0.6774

>> Challenges

Challenges:

- 3.5GB train dataset and 4.1GB test dataset; high dimensionality = BIG DATA
- Limited RAM and compute power lead to crashes as RAM was maximized

Solutions:

- Utilized PySpark for dataframe computations
  - Lazy evaluation
  - Worker parallelization
- Purchased Google Colab Pro
  - 50 GB of RAM (still not enough)
  - Highlights the importance of cloud computing in big data
- Employed a stratified random subset based on “HasDetections”
  - Concerns over subset being representative of population data

>> Conclusion & Thoughts

Findings:

- Random Forest was the best model using the F1 and ROC-AUC metrics
- Model on par with other top competition entrants
    - Highlights importance of feature selection and model tuning
- XGBoost is second best best model
    - Ensemble, tree methods handle non-linearity well
- Non-machine features must also be considered (computer users, type of use, etc.)

Looking Ahead:

- Migrate to AWS for more storage and faster compute
- Include more training data
- Predict on competition test dataset to see how model compares with other submissions
- Prune decision tree
- Optimize probability threshold

Thank you! Any questions?

>> References

- https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/