DSO 530
Microsoft Malware Prediction

Level-II

Mengtian (Monica) Hu mengtiah@usc.edu

Problem Understanding & Data Description

Goal: Predict a Windows machine's probability of getting infected by malwares

Output: Probability of a machine attacked by malwares

Response variable: *HasDetections* indicating whether malware was detected

Predictors: 83 input fields including producer info, hardware info...

Data Description:

- Our training and testing data contains 100,000 records each
- The training data is balanced (50:50 Attacked: Unattacked), no resampling is needed

Data Cleaning

- Remove fields with high ratio of missing values (i.e. PuaMode) , extremely imbalanced categorical fields (i.e. IsBeta), and strange numbers (i.e. Census_InternalBatteryNumberOfCharge)
- Categorical fields:

  Missing values: treat missing categories as new groups
  Merge small groups (<100 items) into one group
  Mean-encode all categorical fields
- Numerical fields:

  Build supervised models using Bagging(DT) to predict missing values

Variable Creation

- **Create all interactive variables X_iX_j**

Created about **2000** interactive variables out of all individual variables cleaned

- **Create expert variables**

1. **primary_drive_c_ratio**=Census_SystemVolumeTotalCapacity/Census_PrimaryDiskTotalCapacity
2. **non_primary_drive_MB**=Census_PrimaryDiskTotalCapacity-Census_SystemVolumeTotalCapacit
3. **aspect_ratio**=Census_InternalPrimaryDisplayResolutionHorizontal/Census_InternalPrimaryDisplayResolutionVertical
4. **ram_per_processor**=Census_TotalPhysicalRAM/Census_ProcessorCoreCount
5. **new_num_0**=Census_InternalPrimaryDiagonalDisplaySizeInInches/Census_ProcessorCoreCount
6. **new_num_1**=Census_ProcessorCoreCount*Census_InternalPrimaryDiagonalDisplaySizeInInches

Variable Creation

- **Create all interactive variables X_iX_j**

Created about **2000** interactive variables out of all individual variables cleaned

- **Create expert variables**

1. **primary_drive_c_ratio**=Census_SystemVolumeTotalCapacity/Census_PrimaryDiskTotalCapacity
2. **non_primary_drive_MB**=Census_PrimaryDiskTotalCapacity-Census_SystemVolumeTotalCapacity
3. **aspect_ratio**=Census_InternalPrimaryDisplayResolutionHorizontal/Census_InternalPrimaryDisplayResolutionVertical
4. **ram_per_processor**=Census_TotalPhysicalRAM/Census_ProcessorCoreCount
5. **new_num_0**=Census_InternalPrimaryDiagonalDisplaySizeInInches/Census_ProcessorCoreCount
6. **new_num_1**=Census_ProcessorCoreCount*Census_InternalPrimaryDiagonalDisplaySizeInInches

Feature Selection

Applied 2 different feature selection approaches

- **The Lasso**

  We applied Lasso to reduce some coefficient to zero

  Results: 72 → **65** important variables

- **Best Subset Selection**

  We searched through 2^p models to find the best subset

  Results: 72 → **44** important variables

66 Lasso Selection

HasTpm_risk
Firewall_risk
SmartScreen_risk
Census_FlightRing_risk
Census_IsVirtualDevice_risk
AVProductStatesIdentifier_risk
Census_DeviceFamily_risk
EngineVersion_risk
RtpStateBitfield_risk
Wdft_IsGamer_risk
CountryIdentifier_risk
Census_MDC2FormFactor_risk
AppVersion_risk
Census_OSBranch_risk
Census_OSVersion_risk
Census_ProcessorModelIdentifier_risk
Census_OSEdition_risk
Census_ChassisTypeName_risk
Processor_risk
Census_PrimaryDiskTypeName_risk
leVerIdentifier
ram_per_processor
non_primary_drive_MB
Census_PrimaryDiskTotalCapacity
Census_IsTouchEnabled_risk
SMode_risk
Census_OEMNameIdentifier_risk
LocaleEnglishNameIdentifier_risk
Census_IsAlwaysOnAlwaysConnectedCapable_risk
Census_OSInstallTypeName_risk
GeoNameIdentifier_risk
SkuEdition_risk
OrganizationIdentifier_risk
UacLuaenable_risk
Census_PowerPlatformRoleName_risk
Census_GenuineStateName_risk
Census_OSArchitecture_risk
aspect_ratio
Census_HasOpticalDiskDrive
Census_OSSkuName_risk
Census_ActivationChannel_risk
AVProductsEnabled
Census_OSUILocaleIdentifier_risk
Census_InternalPrimaryDiagonalDisplaySizeInInches
Census_ProcessorCoreCount
MegaPixels
Census_OSBuildRevision
OsBuild
Census_InternalPrimaryDisplayResolutionHorizonta
new_num_1
new_num_0
primary_drive_c_ratio
IsProtected_risk
Census_FirmwareManufacturerIdentifier_risk
AVProductsInstalled
Census_IsPortableOperatingSystem_risk
OsVer_risk
IsSxsPassiveMode_risk
OsPlatformSubRelease_risk
Wdft_RegionIdentifier_risk
Census_IsPenCapable_risk
Census_OSWUAutoUpdateOptionsName_risk
OsSuite_risk
ProductName_risk
Census_IsSecureBootEnabled_risk

44 Best Subset Selection

AVProductsInstalled
AVProductsEnabled
OsBuild
IeVerIdentifier
Census_PrimaryDiskTotalCapacity
Census_TotalPhysicalRAM
Census_InternalPrimaryDiagonalDisplaySizeInInches
Census_OSBuildNumber
EngineVersion_risk
AppVersion_risk
RtpStateBitfield_risk
AVProductStatesIdentifier_risk
CountryIdentifier_risk
GeoNameIdentifier_risk
LocaleEnglishNameIdentifier_risk
Processor_risk
OsPlatformSubRelease_risk
IsProtected_risk
SMode_risk
SmartScreen_risk
Firewall_risk
Census_OEMNameIdentifier_risk
Census_ProcessorModelIdentifier_risk
Census_PrimaryDiskTypeName_risk
Census_ChassisTypeName_risk
Census_PowerPlatformRoleName_risk
Census_OSVersion_risk
Census_OSArchitecture_risk
Census_OSBranch_risk
Census_OSEdition_risk
Census_OSSkuName_risk
Census_OSInstallTypeName_risk
Census_OSUILocaleIdentifier_risk
Census_OSWUAutoUpdateOptionsName_risk
Census_GenuineStateName_risk
Census_ActivationChannel_risk
Census_FirmwareManufacturerIdentifier_risk
Census_IsVirtualDevice_risk
Census_IsTouchEnabled_risk
Wdft_IsGamer_risk
Wdft_RegionIdentifier_risk
primary_drive_c_ratio
non_primary_drive_MB
new_num_0

Modeling

We applied 7 different models

- Lasso Logistic Regression
- Logistic Regression
- Extreme Gradient Boosting
- Ada Boosting
- Random Forest
- K Nearest Neighbors
- Deep Neural Network

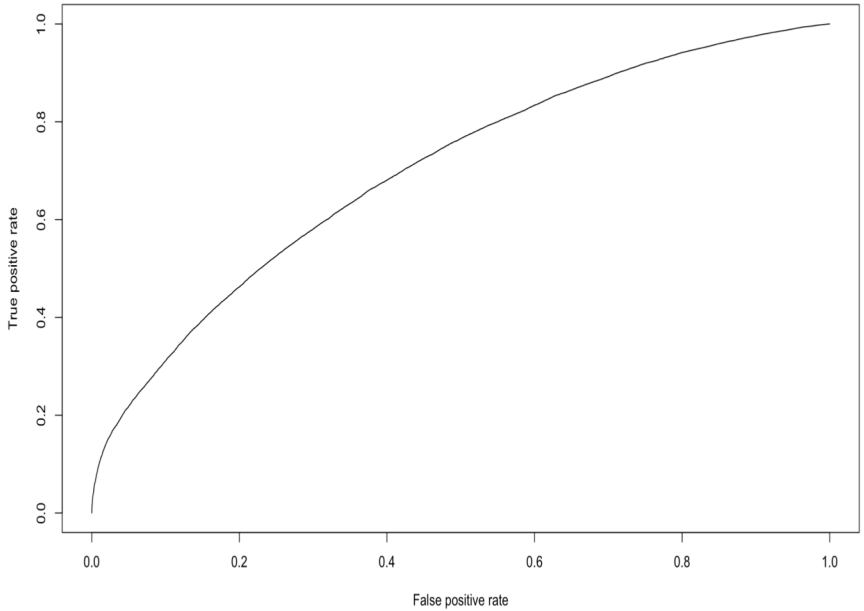
Performance Comparison

The training/testing accuracy rate and AUC value of 7 models

Model	Training Accuracy	Testing Accuracy	AUC (Test)
Lasso Regression	63.11%	63.22%	0.69
Logistic Regression	53.53%	53.69%	0.56
Ada Boosting	64.07%	63.27%	0.63
XGBoost	**66.75%**	**64.11%**	**0.70**
Random Forest	64.27%	63.03%	0.63
6 Nearest Neighbors	67.07%	52.15%	0.52
Deep Neural Network	50.16%	49.98%	0.50

Final Approach

	XGBoost
objective	“binary::logistic”
max_depth	8
nround	30
early_stopping	3
num_parallel_tree	10
subsample	0.8
colsample_bytree	0.8
AUC	0.7017



Business Interpretation

We summarize characteristics of the top two most risky groups:

- Machines with the following Version of Defender:

  "4.10.14393.0", "4.10.14393.1198", "4.10.14393.1593", "4.10.14393.1794", "4.10.209.0"

  Have higher probability to be detected with malware.

- If a Virtual Machine is installed, the computer has higher probability be detected with malware.
- Computers with Monitor Size larger than 11 inches have higher probability be detected with malware.

Business Interpretation

- Machines located in the following Country:

  "104", "190", "95", "214", "100"

  Have higher probability to be detected with malware.

- Computers with Census_OSArchitecture as “amd64” have higher probability to be detected with malware.

Useful Statistical Learning Insights and Extended Question

- Creating new variables based on business insights is better than creating possible interactive variables
- Building models to predict missing values such as screening resolution is better than using median value
- Before using mean encoding, it's better to take a look at the categorical variables and merge low frequent categories
- Simple model (logistic regression or KNN) and complex model (deep neural network) are not good choices in this case

Which kinds of Windows machines are safer to use?

Q & A