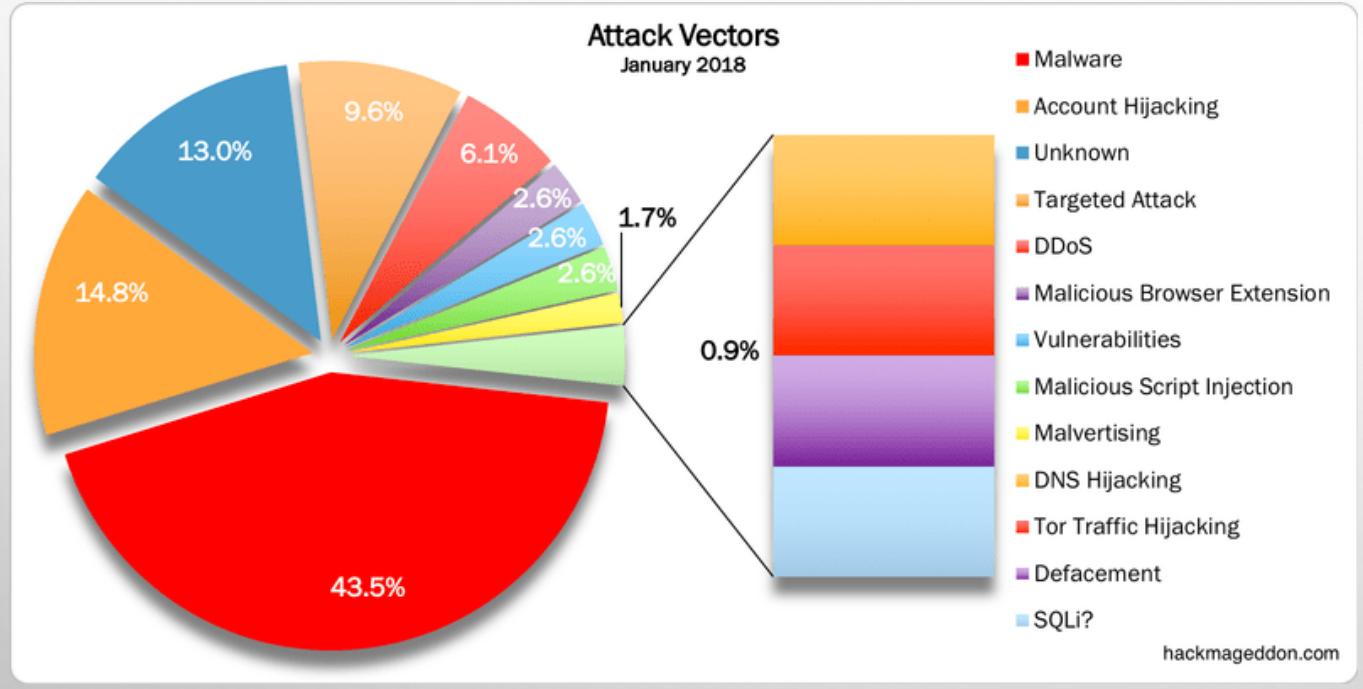
MICROSOFT MALWARE PREDICTION

By,
Victor
James
Priya

TABLE OF CONTENTS

MALWARE

$1.1M IS LOST TO CYBERCRIME EVERY MINUTE OF EVERY DAY



RANSOMWARE
costs to organizations
$8 BILLION/year[5]
$15,221/minute[5]
1.5 organizations/minute fall victim to ransomware attacks[6]

MALWARE
1,274 new malware variants/minute[7]

PHISHING EMAILS
22.9 attacks/minute[8]

RECORDS LEAKED
from publicly disclosed incidents
2.9 BILLION/year[9]
5,518/minute

https://www.hackmageddon.com/2012-cyber-attacks-statistics-master-index/
https://threatpost.com/threatlist-1-1m-is-lost-to-cybercrime-every-minute-of-every-day/136871/

DATA DESCRIPTION

THE GOAL OF THIS PROJECT IS TO PREDICT WINDOWS MACHINE'S PROBABILITY OF GETTING INFECTED BY VARIOUS FAMILIES OF MALWARE, BASED ON DIFFERENT PROPERTIES OF THAT MACHINE. BELOW ARE SOME OF THE DETAILS OF THE AVAILABLE DATA FOR PREDICTING THE MALWARE

Predictors (82)		Response (1)	Observations
74 categorical variables	*8 numeric*	*1 binary*	*8.9 M*
OSEdition	*SystemVolumeTotal Capacity*	*HasDetections*	
ProcessorClass	*ProcessorCoreCount*		

DATA PREPROCESSING

As part of data Preprocessing, the first step was to check the check the variables for missing values, NA and the number of categories and the below variables were deleted

Variables	Deletion Reason
Census_OEMNameIdentifier	*Too Many NA values*
Census_OEMModelIdentifier	*Too Many NA values*
Census_ProcessorModelIdentifier	*Too Many NA values*
PuaMode	*High no of Blank Values (99%)*
AutoSampleOptIn	*99% of values are 0's*
Census_InternalBatteryType	*71% values were blank*
Census_IsVirtualDevice	*Too many missing values*
Census_IsAlwaysOnAlwaysConnectedCapable	*Too many missing values*
Census_OSInstallLanguageIdentifier	*Too many missing values*
Census_DeviceFamily	*More than 90% of values are of same category*
IsBeta	*99.999% False*
DefaultBrowsersIdentifier	*It's an ID*
CityIdentifier	*Too many levels*

DATA PREPROCESSING

MISSING VALUES

- IMPUTATION OF MISSING VALUES USING THE MOST FREQUENT CLASS (BINARY VARIABLES)
- NO ACTION FOR REMAINING CATEGORICAL VARIABLES (WOE TAKES CARE OF IT)

INITIAL APPROACH:

- GROUPING BY FREQUENCY AND GROUPING BY RESPONSE RATE, THIS HAD TOO MANY VARIABLES DUE TO DUMMY CREATION AND COULD NOT BE HANDLED FOR LARGE NUMBER OF CATEGORIES

FINAL APPROACH:

- CATEGORICAL VARIABLES WITH < = 3 LEVELS WERE CHECKED FOR THE FREQUENCY AND COMBINED TOGETHER FOR ANY FACTOR WITH VERY LESS VALUES (LESS THAN 1%)
- WEIGHT OF EVIDENCE FOR REMAINING CATEGORICAL VARIABLES

FEATURE ENGINEERING

- MOST OF THE VARIABLES WERE CATEGORICAL AND VERY FEW CONTINUOUS VARIABLES

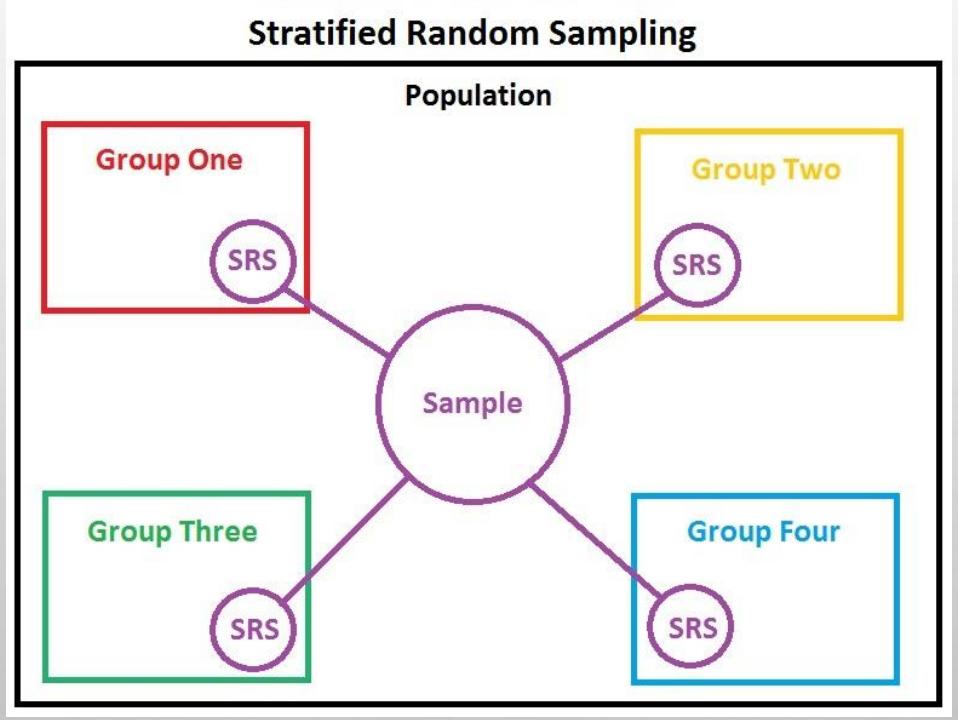
- HIGH CARDINALITY IN MANY CATEGORICAL VARIABLES
  - WOE : A NUMERICAL REPRESENTATION OF THE PREDICTIVE POWER OF A VARIABLE
    - REPRESENT CATEGORICAL DATA IN A CONTINUOUS WAY (AVOID DUMMIES)
    - NAS ARE TAKEN CARE OF IN THE PROCESS

$$WOE = \ln\left(\frac{\%\text{ of non-events}}{\%\text{ of events}}\right)$$

DATA SAMPLING

AFTER THE COMPLETION OF FEATURE ENGINEERING AND BASETABLE COMPLETION(8,921,483 OBS AND 65 VARIABLES), WE HAD TO SUBSET THE DATA DUE TO ITS HUGE VOLUME FOR FURTHER STEPS OF DATA PROCESSING.

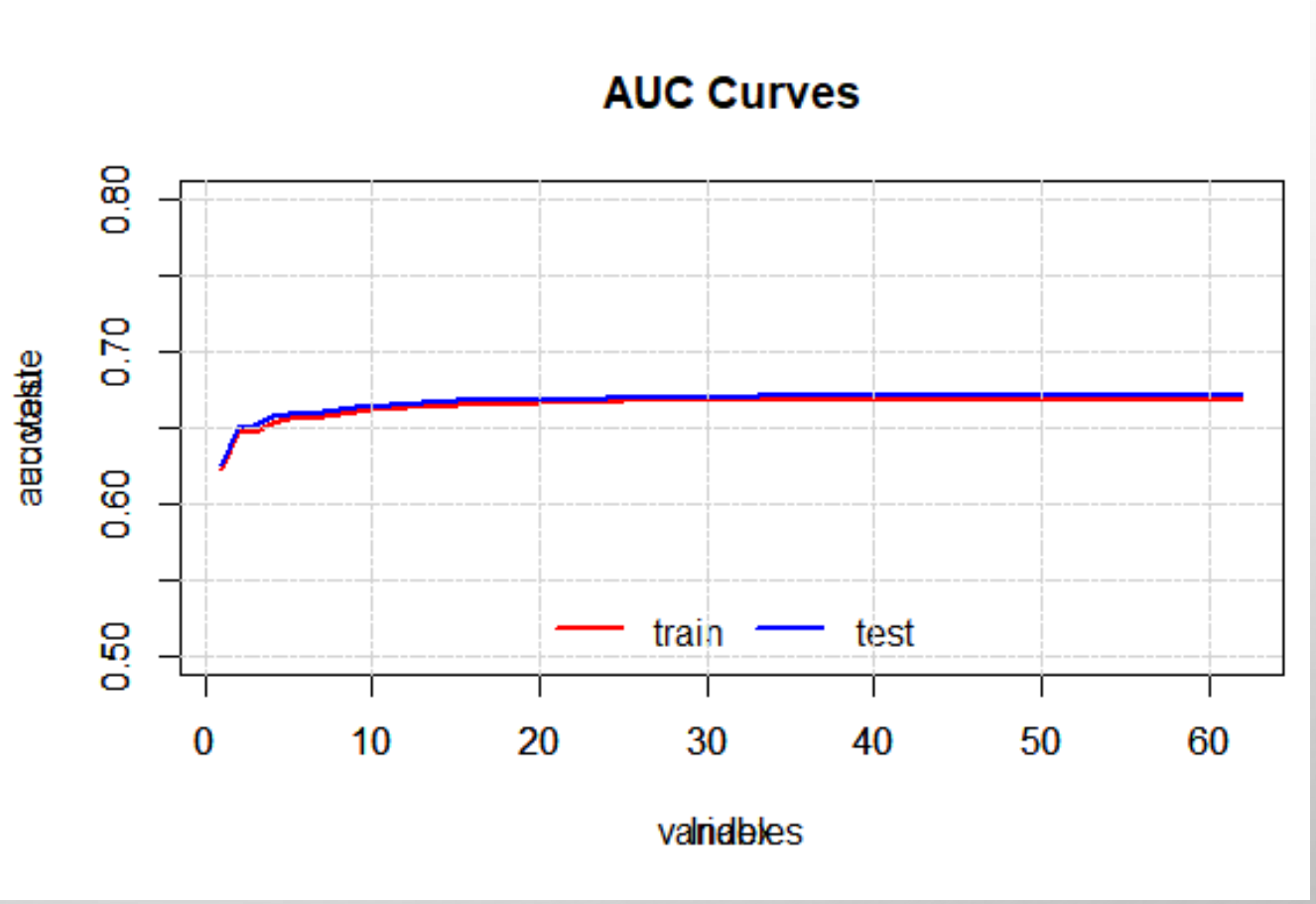
STRATIFIED SAMPLING OF DATA – 1%(178428 OBS AND 65 VARIABLES)



VARIABLE SELECTION

ONCE THE DATA WAS CLEANED VARIABLE SELECTION WAS DONE USING THE FOLLOWING TWO METHODS (178428 OBS AND 35 VARIABLES)

- AUC BASED VARIABLE SELECTION
- IV (INFORMATION VALUE OF VARIABLES)



VARIABLE SELECTION

INFORMATION VALUE - $$IV = \sum(DistributionGood_i - DistributionBad_i) \times WOE_i$$

Variable	IV
woe_SmartScreen	2.391564e-01
woe_AVProductStatesIdentifier	1.440958e-01
woe_Census_SystemVolumeTotalCapacity	1.433386e-01
woe_AvSigVersion	9.022682e-02
woe_AVProductsInstalled	8.852625e-02
woe_EngineVersion	6.727927e-02
woe_Census_FirmwareVersionIdentifier	5.284957e-02
woe_AppVersion	4.068608e-02
woe_Census_TotalPhysicalRAM	3.904400e-02
woe_Census_PrimaryDiskTotalCapacity	3.739415e-02
ıs_InternalPrimaryDiagonalDisplaySizeInInches	3.759078e-02
woe_OsBuildLab	3.149699e-02
woe_CountryIdentifier	2.692521e-02
woe_Census_OSBuildRevision	2.231698e-02
Processor	2.025062e-02
woe_Census_OSArchitecture	1.991372e-02

MODELING

5 algorithms to bench mark the data

LOGISTIC REGRESSION

SIMPLE TRAIN-TEST SPLIT

BASIC MODEL USED FOR CLASSIFICATION ,PREDICTION DONE FOR THE TEST SET WITH TRAIN AND TEST SPLIT.

ACCURACY = 62.05 , AUC = 0.65

CROSS VALIDATION – 10 FOLD CV

ACCURACY = 62.05, AUC = 0.66

SVM (RBF)

CROSS VALIDATION AND TUNING – RADIAL KERNEL(1% FROM THE SAMPLE)

10 FOLD, COST = 0.01 AND GAMMA = 0.5

SIMPLE TRAIN AND TEST SPLIT

ACCURACY = 63.69, AUC = 0.6396

MODELING

LDA (LINEAR DISCRIMINANT ANALYSIS)

- **GOALS:**
  - FIND A LINE FUNCTION THAT ENABLE TO CLASSIFY THE RESPONSE
  - SEE WHICH VARIABLES HAVE MORE IMPACT IN THE LINE FUNCTION.
- **EVALUATION:**

  SIMPLE SPLIT TRAIN(70%)/TEST
  - ACCURACY: 0.6232
  - AUC: 0.67

  CROSS VALIDATION: 10 FOLDS AND 3 REPETITIONS
  - ACCURACY: 0.6183
  - AUC: 0.66

Variable	LD1
woe_Platform	5.2542685
woe_Census_FlightRing	1.7546908
OsSuite_LFL	1.4936013
Census_ProcessorClass_low	1.2931017
woe_AvSigVersion	1.1897438
woe_AVProductsInstalled	0.9876676
woe_CountryIdentifier	0.9636418
woe_OsBuildLab	0.9322967
woe_Census_OSBuildNumber	0.844999
woe_Census_OSBuildRevision	0.7187281
woe_Census_InternalPrimaryDiagonalDisplaySizeInInches	0.685628
woe_Census_ActivationChannel	0.6815406
woe_Census_OSEdition	0.6335684
woe_Census_MDC2FormFactor	0.6212193
woe_AVProductStatesIdentifier	0.6124042
woe_Census_OSInstallTypeName	0.5552379
Census_ProcessorClass_high	0.5513473
woe_Census_ChassisTypeName	0.537942
woe_LocaleEnglishNameIdentifier	0.4698756
woe_AppVersion	0.4650536
woe_Census_SystemVolumeTotalCapacity	0.3406834
woe_SmartScreen	0.269284
woe_Census_ProcessorCoreCount	0.2465109
woe_Census_FirmwareVersionIdentifier	0.2388534
woe_Census_TotalPhysicalRAM	0.1244191
woe_Census_PrimaryDiskTotalCapacity	-0.1578829
Census_IsSecureBootEnabled_0	-0.1635044
OsSuite_256	-0.2456146
AVProductsEnabled_2	-0.2657403
RtpStateBitfield_LFL	-0.2940224
RtpStateBitfield_0	-0.3154163
AVProductsEnabled_LFL	-0.3172237
Census_HasOpticalDiskDrive_0	-0.3649296
woe_Census_OSArchitecture	-0.4276157
woe_Census_FirmwareManufacturerIdentifier	-0.4693872
woe_Census_OSWUAutoUpdateOptionsName	-1.2727505
Census_IsVirtualDevice_1	-1.6880367
SMode_1	-2.1475205
woe_OsPlatformSubRelease	-2.6511357

MODELING

RANDOM FOREST

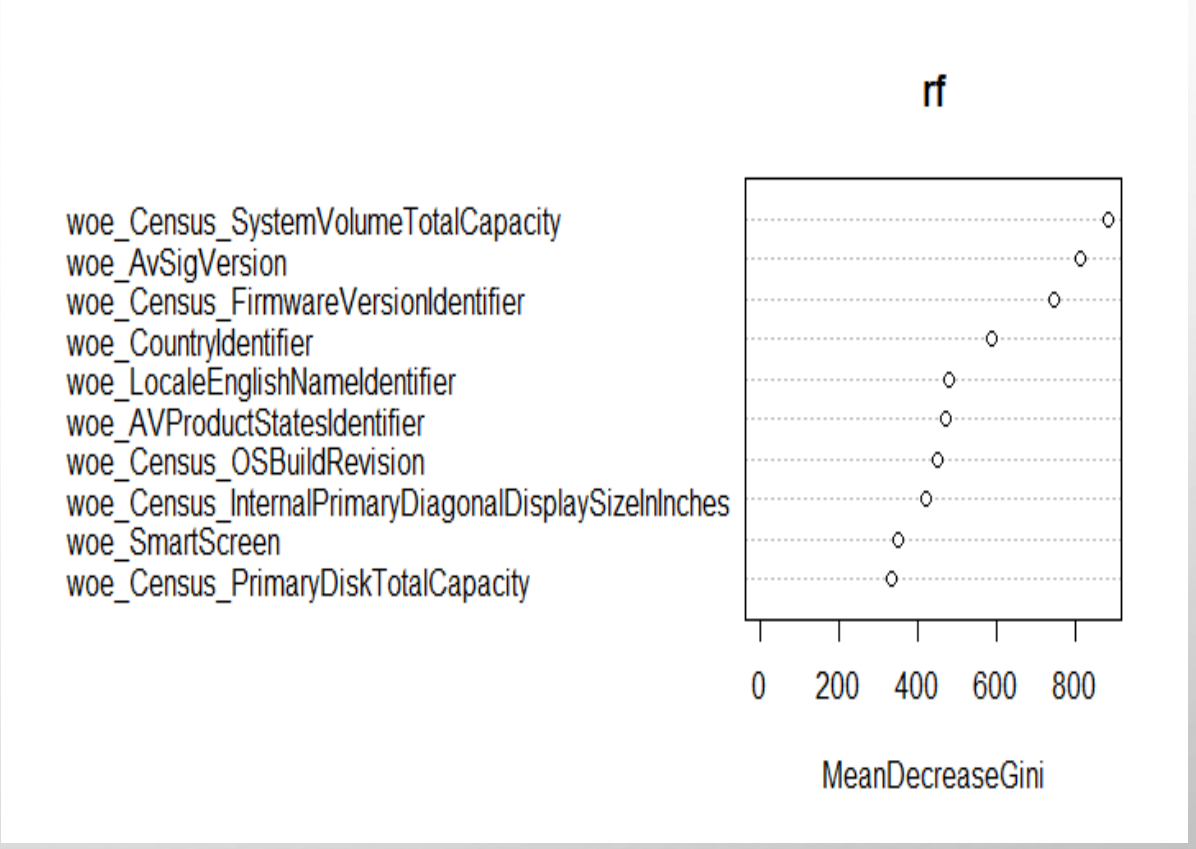
DATA USED IS A STRATIFIED SUBSET OF THE INITIAL SUBSET

GRID-SEARCH FOR HYPERPARAMETER TUNING (SIMPLE TRAIN – TEST SET)

- TUNED MTRY & NTREE
- 25 PARAMETER COMBINATIONS CONSIDERED

CROSS-VALIDATION WITH THE BEST PARAMETER COMBINATION (AUC-WISE)

- MTRY = 10, NTREE = 1000, 5-FOLDS
- AUC : 0.665



MODELING

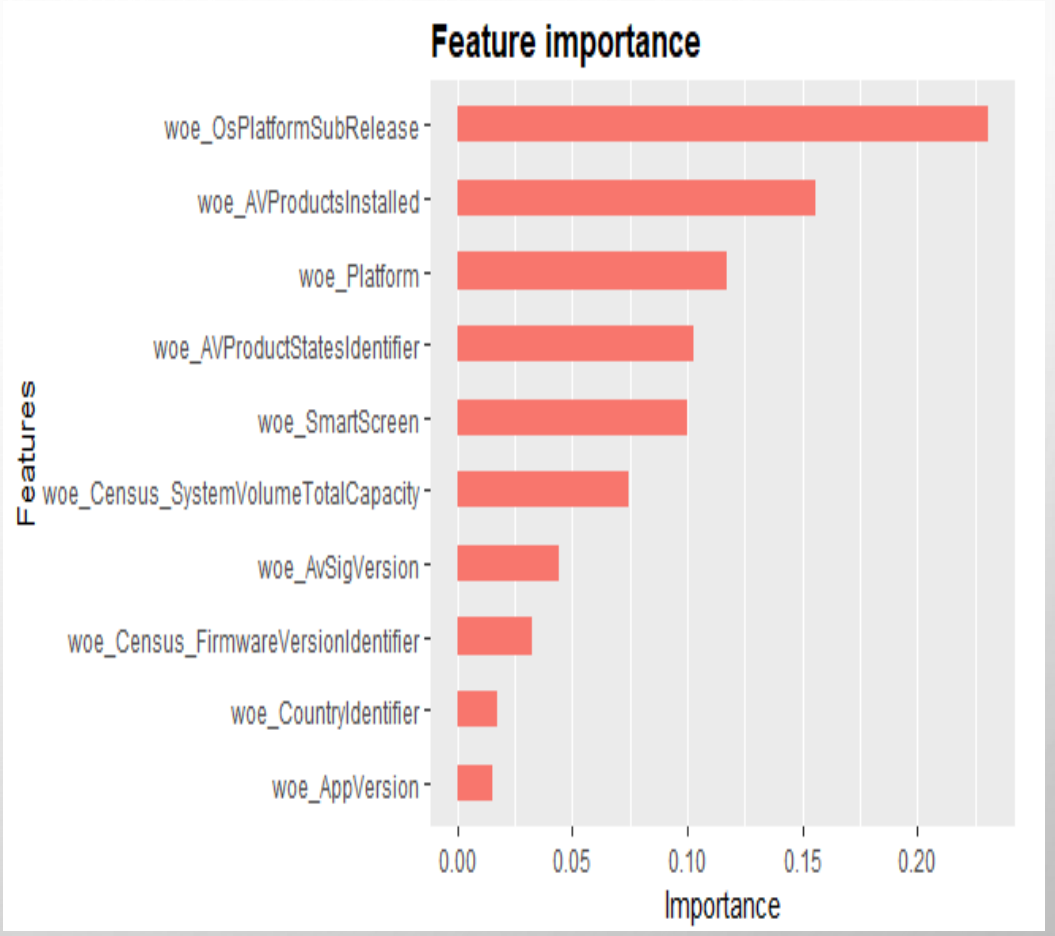
XGBOOST

GRID-SEARCH FOR HYPERPARAMETER TUNING (SIMPLE TRAIN – TEST SET)

- CONSIDERED NUMBER OF ITERATIONS, LEARNING RATE, GAMMA, SAMPLE PROPORTION
- 54 PARAMETER COMBINATIONS TRIED

CROSS-VALIDATION WITH THE BEST SET OF PARAMETERS

- NROUND=200, ETA=0.3, GAMMA=10, COLSAMPLE_BYTREE=0.5, SUBSAMPLE=0.5
- 10-FOLDS
- AUC : 0.678



RESULTS

Model	Tuned parameters	Accuracy	AUC
Logistic Regression	*Cross Validation*	*0.62*	*0.665*
LDA	*Cross Validation (10 folds,3 rep)*	*0.62*	*0.66*
SVM	*Cross Validation, Cost = 0.01, gamma = 0.5*	*0.636*	*0.64*
Random Forest	*(mtry=10, ntree=1000)*	*0.664*	*0.664*
XGBoost	*(nround=200, eta=0.3, gamma=10, subsample=0.5, colsample_bytree=0.5)*	*0.678*	*0.678*

CONCLUSION

MODEL SELECTED : XGBOOST

- BEST PERFORMANCE OVERALL
- VERY QUICK TRAINING TIME
- RATHER INTERPRETABLE MODEL

LIMITATIONS AND FURTHER RESEARCH :

- RUNNING A MORE ELABORATE BENCHMARK WITH SYSTEMATIC CROSS-VALIDATION
- HUGE DATA VOLUME, DUE TO WHICH MANY OF THE PROCESSING AND MODELING COULD NOT BE DONE ON THE ENTIRE DATA SET
- WOE COMPUTED WITHOUT A SMOOTHING PARAMETER (RESULTS IN NA'S)
- WOE WAS COMPUTED BEFORE SPLITTING INTO TEST AND TRAIN, WHICH HAD TO BE COMPUTED ONLY ON TRAINING SET

REFERENCES

EFFICIENCY OF IMPORTING LARGE CSV FILES IN R

HTTPS://WWW.R-BLOGGERS.COM/EFFICIENCY-OF-IMPORTING-LARGE-CSV-FILES-IN-R/

MICROSOFT MALWARE PREDICTION - PREPROCESSING TRAIN

HTTPS://WWW.KAGGLE.COM/FLUBBER/MICROSOFT-MALWARE-PREDICTION-PREPROCESSING-TRAIN

6 DIFFERENT WAYS TO COMPENSATE FOR MISSING DATA

HTTPS://TOWARDSDATASCIENCE.COM/6-DIFFERENT-WAYS-TO-COMPENSATE-FOR-MISSING-VALUES-DATA-IMPUTATION-WITH-EXAMPLES-6022D9CA0779

SVM TUNING

HTTPS://WWW.R-BLOGGERS.COM/RADIAL-KERNEL-SUPPORT-VECTOR-CLASSIFIER/

STRATIFICATION

HTTPS://RDRR.IO/CRAN/SPLITSTACKSHAPE/MAN/STRATIFIED.HTML

AN XGBOOST EXPLANATION

HTTP://BLOG.KAGGLE.COM/2017/01/23/A-KAGGLE-MASTER-EXPLAINS-GRADIENT-BOOSTING/

WEIGHT OF EVIDENCE

HTTPS://MEDIUM.COM/@SUNDARSTYLES89/WEIGHT-OF-EVIDENCE-AND-INFORMATION-VALUE-USING-PYTHON-6F05072E83EB