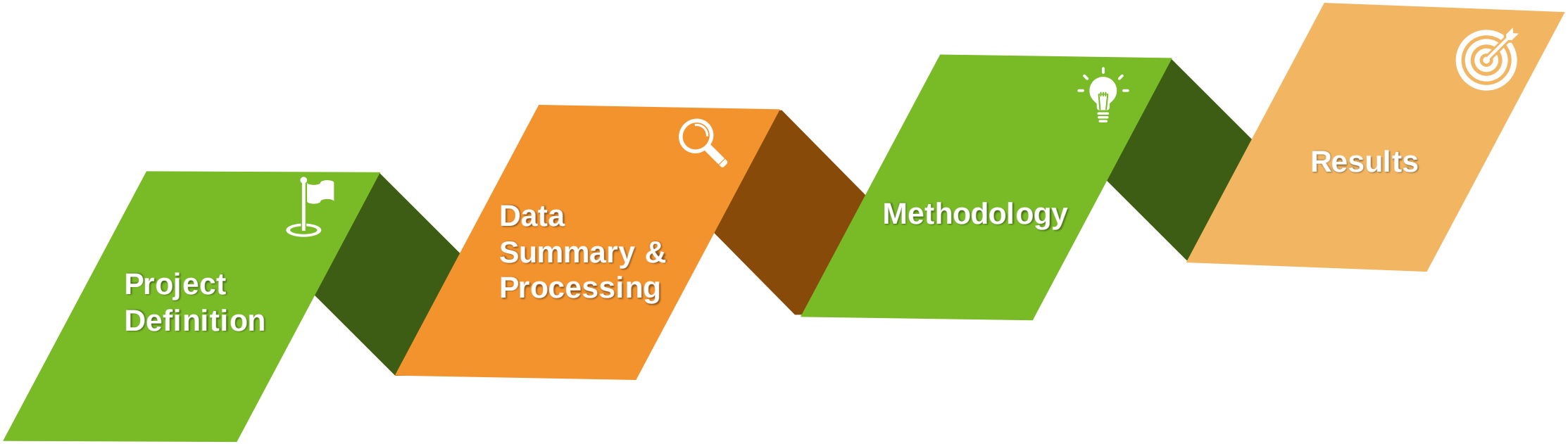
Microsoft Malware Prediction

Group 10: Alejandra Zambrano, Chenxin Xie, Manoj Kumar Purushothaman

Agenda


Project Definition
Data Summary & Processing
Methodology
Results

1. Project definition

Business problem:

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

Data science problem:

Using classification model to predict the probability of a machine being infected by malware, based on different properties of that machine.

2. Data summary and processing

The telemetry data containing machines' properties and the machine infections was generated by combining heartbeat and threat reports collected by Microsoft's endpoint protection solution, Windows Defender.

01 Data Summary

02 Data Processing

03 Variable Selection

2.1 Data Summary

Identifier:
MachineIdentifier - Individual machine ID

Predictor:
81 variables with machine properties

Response:
HasDetections (0/1)

Variables

Missing values:

newColName	missing	percentage
DefaultBrowsersIdentifier	848658	0.951251417
Census_IsFlightingInternal	740995	0.830573144
Census_ThresholdOptIn	566716	0.635225730
Census_IsWIMBootEnabled	565968	0.634387305
OrganizationIdentifier	275613	0.308931580
SMode	53977	0.060502226
CityIdentifier	32763	0.036723686
Wdft_IsGamer	30196	0.033846364

8921483 Machines in total

Data errors

Predictor: Unbalanced

Response: Balanced

	unique_values	perc_biggest_category
IsBeta	2	1.00
AutoSampleOptIn	2	1.00
PuaMode	2	1.00
Census_DeviceFamily	3	1.00
Census_ProcessorClass	4	1.00
Census_IsPortableOperatingSystem	2	1.00
ProductName	3	0.99
HasTpm	2	0.99
UacLuaenable	6	0.99
Census_IsVirtualDevice	3	0.99
Census_IsFlightsDisabled	3	0.98
IsSxsPassiveMode	2	0.98
Firewall	3	0.97

Distribution

2.2 Data Processing

Remove variables

- Variables with more than **50%** of missing values
- Variables with more than **90%** of their values in one category
- Variables with more than **400** levels (computationally expensive convert them into dummy variables)

Dummy variables

- 39 variables were converted into dummy variables
- Remove variables with less than 1000 observations in level 1

Convert to data type

- Convert all variables as factor, except for “MachineIdentifier”
- Convert some variables from numeric or integer into category type, for it should be category type by checking the observations
- Fill na in those converted categorical variables with “no_info”

Special variables

- Census_InternalBatteryType
- Census_PowerPlatformRoleName
- Census_PrimaryDiskTypeName
- Census_ChassisTypeName
- Census_ActivationChannel
- SmartScreen

2.3 Variable Selection

with

FisherScore

	IV	fisher_score
607	Census_IsTouchEnabled	0.057314103
609	Census_SystemVolumeTotalCapacity	0.018983754
147	AppVersion_4_11_15063_1155	0.007683996
50	leVerIdentifier_114	0.007216679
334	LocaleEnglishNameIdentifier_262	0.007209489
383	GeoNameIdentifier_29	0.006595798

	IV	fisher_score
499	Census_OSBuildRevision_909	3.817744e-05
174	SmartScreen_Warn	3.426673e-05
70	Census_OSInstallLanguageIdentifier_15	2.664136e-05
39	Census_FirmwareManufacturerIdentifier_803	2.629565e-05
87	Census_OSInstallLanguageIdentifier_39	1.954703e-05
215	EngineVersion_1_1_14600_4	9.351969e-06

After data processing, train dataset contains over **600** variable.
We use FisherScore to select the most important variables that we could use to train the model.
We took **50** best variables at last, and apply on train, valid and test.

3. Methodology

Logistic Regression

Resample: 5-fold CV
Hyper parameter tuning

AUC : 0.518

Random Forest

Resample: 5-fold CV
Hyper parameter tuning:
ntree
mtry
nodesize

AUC : 0.532

Gradient Boosted

Resample: 10-fold CV
Hyper parameter tuning:
n_rounds
max_depth
eta
lambda

AUC : 0.544

4. Results

- **Results on our models**

Model	Logistic Regression	Random Forest	Gradient Boosted
AUC on valid	0.518	0.532	0.544
AUC on test	0.518	0.530	0.541

- **Private Leaderboard Result on Kaggle**

In the money Gold Silver Bronze

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲1208	abuurista			0.67585	31	1y
2	▲1063	Confiniti			0.66535	6	1y
3	▲1081	ken10ML			0.66523	40	1y
4	▲1352	John DiMarco			0.66474	15	1y
5	▲1523	khas_ccip			0.66403	14	1y

Thank You

Please check the presentation video at the link below:
https://web.microsoftstream.com/video/958b34e0-3af6-4fe2-bda1-41cb65fd0013

Please check the Jupyter Notebook at the GitHub link below:
https://github.com/xiechenxin/SML_GroupProject_Microsoft_Malware_Prediction