



Malware Resistant PCs: A Thing Of The Future?

Name:

Ellie Lew Yi Ting
Wong Xiaoqing
Aloysius Ong Wei Han
Wang Guanlan
Clenlin D'cruz
Tan Yi Heng

Matriculation Number:

U1810880J
U1822123C
U1810928A
U1820588E
U1810393F
U1821247A

A New Level Of Terror

Problem breakdown & analysis
of cybersecurity issues

01

Malware-resistant PCs

Key Influencing factors and
predictive models

04

The Battle Between Giants

Apple vs Microsoft and the key
business question

02

Comparing the models

A comparison of 9-variable vs
32-variable models

05

Under the magnifier

Data exploration and cleaning
of the dataset

03

Conclusion & Summary

How Microsoft can use the
insights developed

06

01

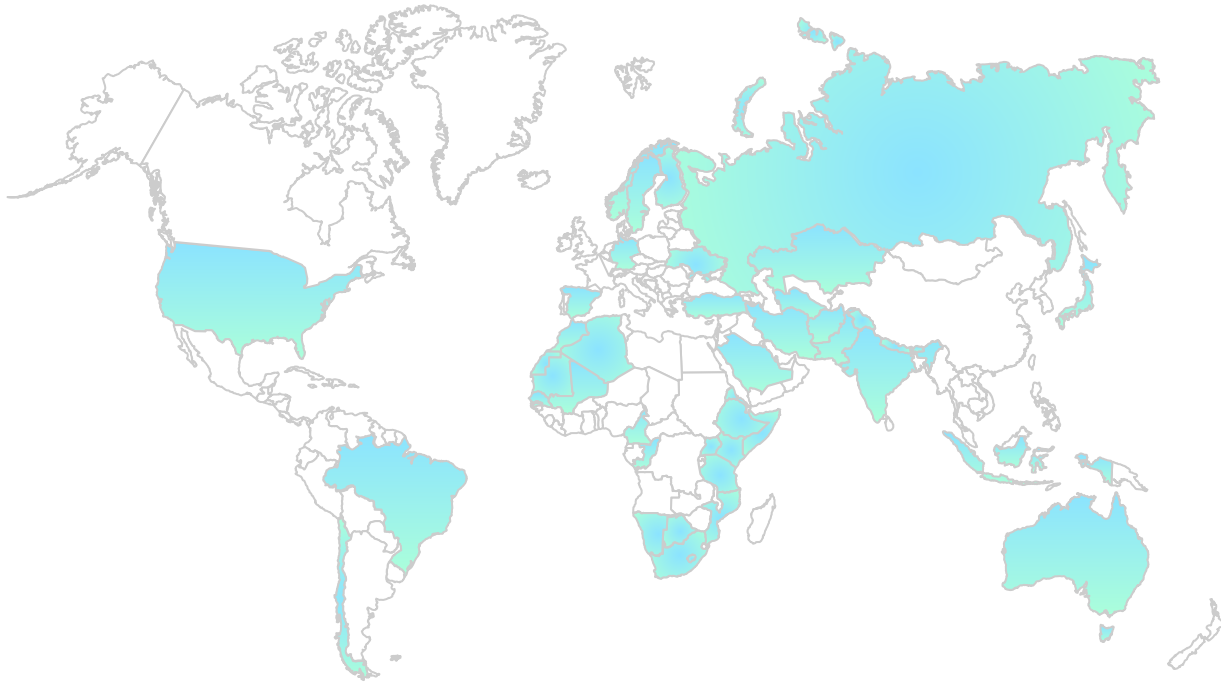
A New Level Of Terror

Problem breakdown & Analysis
of cybersecurity issues



What is malware and why is it bad?

Definition: malicious software that is designed to impair and demolish computers & computer systems



Operation 'Red October'

Victims of a cyber-espionage attack

Examples include:

1. Trojan horse
2. Spyware
3. Adware
4. Ransomware

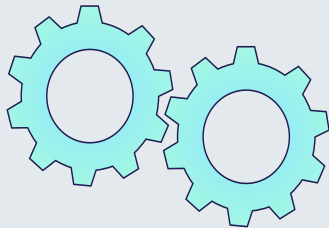
The 4 challenges of malware to businesses

Loss of trade secrets



In a **knowledge-driven** marketplace, the company's IP is especially pivotal to the value of its business & the **loss could threaten its existence**

Loss of productivity



An infection may cause productive processes to stop, & such **productivity losses can amount to US\$112 billion globally**

Loss of reputation



On average, companies lose **\$4.13m** from turnover of customers, reputation losses and diminished goodwill

Changing workplaces



'Anytime, anywhere' endpoint devices results in a no longer a secure perimeter within which business devices, applications and data can be protected

Attacks are increasingly frightening given the high intensity of damage and the high frequency

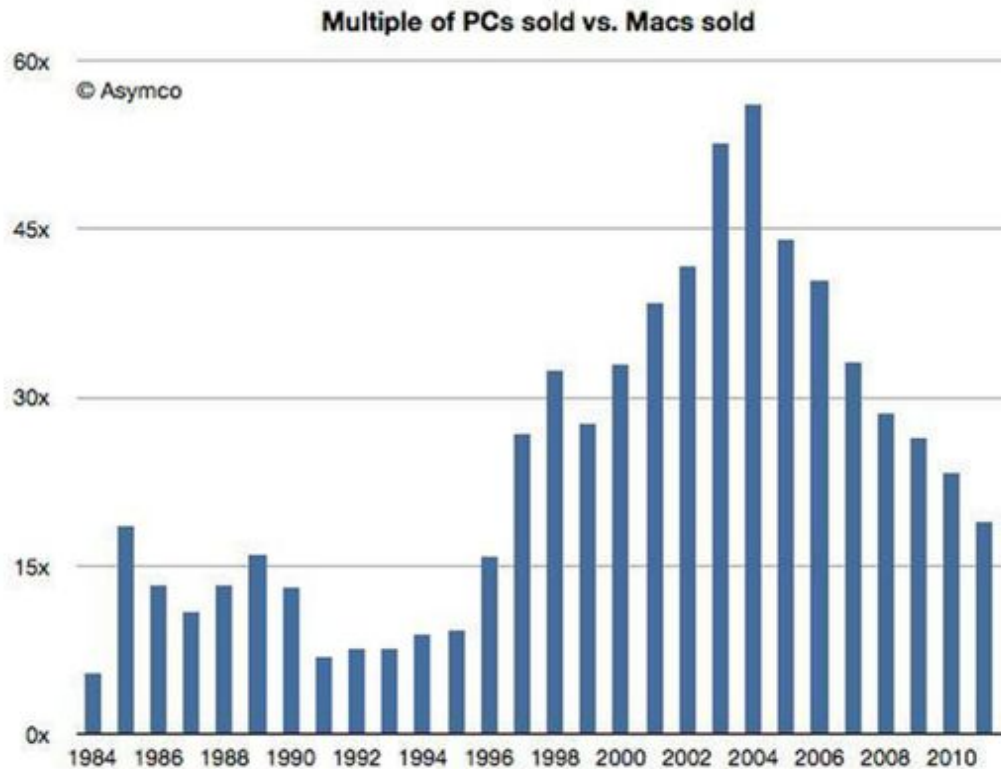
02

The Battle Between Giants

Apple vs Microsoft and the key
business question



The Battle Between Giants



The gap between PC & Mac Sales are narrowing, with one of the reasons being **Microsoft is less safe than Mac**

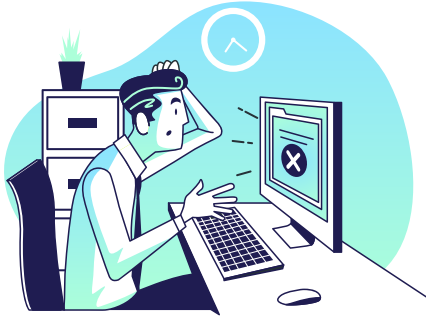
Why is Windows PC not as safe?

1. **Outsource** hardware manufacturing whereas Apple builds its own hardwares
2. **Proprietary software** – bugs can only be fixed by Microsoft team instead of the open source community
3. **Less authorisation** of applications

How to accurately predict malware attacks using features of a Windows machine?

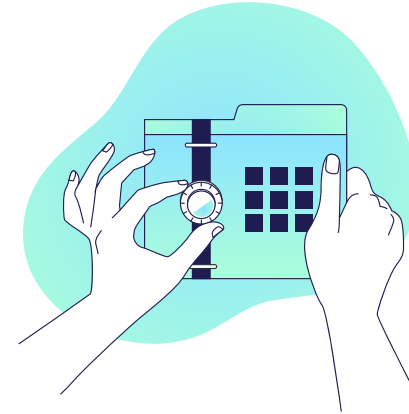


2 - pronged approach is being used



Key Influencing Factors

Analysing the variables within a dataset to sieve out **potential factors that are influential** in affecting malware prediction – one-step closer to malware resistant PCs



Prediction Models

To determine the probability of malware attacks with the identified hardware features – **identify existing computers with high likelihood of attacks**, and develop early interventions to minimise impact



03

Under The Magnifier

Data cleaning and exploration

Data Cleaning

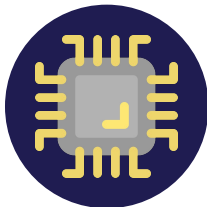
Dataset obtained from Microsoft Malware Prediction Competition from Kaggle.com had 83 columns.

Used **domain knowledge** and information regarding the variables to **remove unnecessary columns**, leaving 33 variables:

Binary dependent variable, HasDetections, and 32 independent variables.



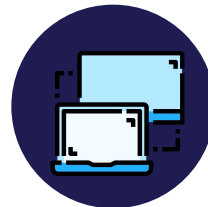
9 out of 32 Independent Variables



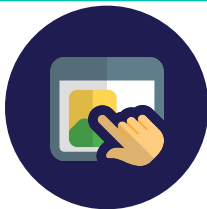
Processor



Census_ProcessorCoreCount



Census_InternalPrimaryDiagonal
DisplaySizeInInches



Census_IsTouchEnabled



Census_SystemVolumeTotalCapacity



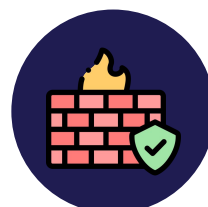
Census_HasOpticalDiskDrive



Census_PrimaryDiskTypeName

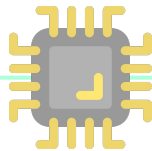


Census_TotalPhysicalRAM



Firewall

Processor

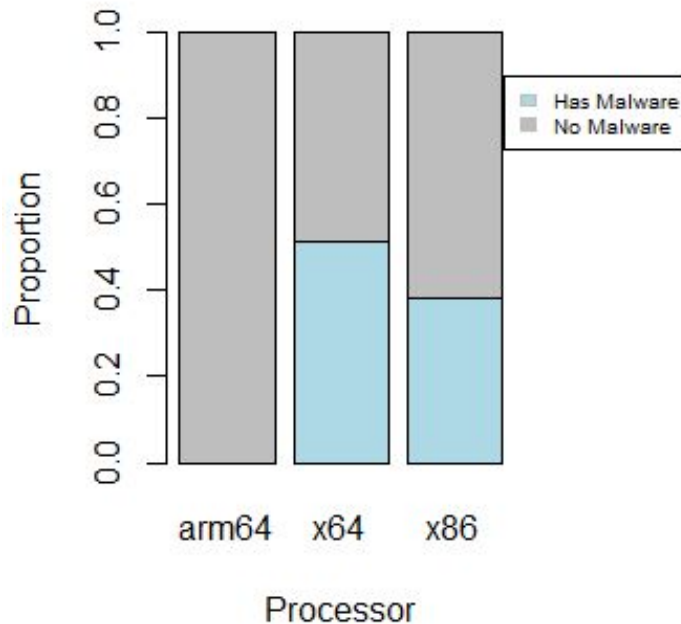


Process architecture of the installed operating system.

3 types of processors in the dataset:
arm64, x84 and x64.

Suggests that **x64** processors are the **most susceptible** to malware attacks while **arm64** are **safe**.

Processor vs HasDetections



Difficult to use processor to predict malware due to **limited number of different types** and most Windows machine uses x86.

Census_ProcessorCoreCount



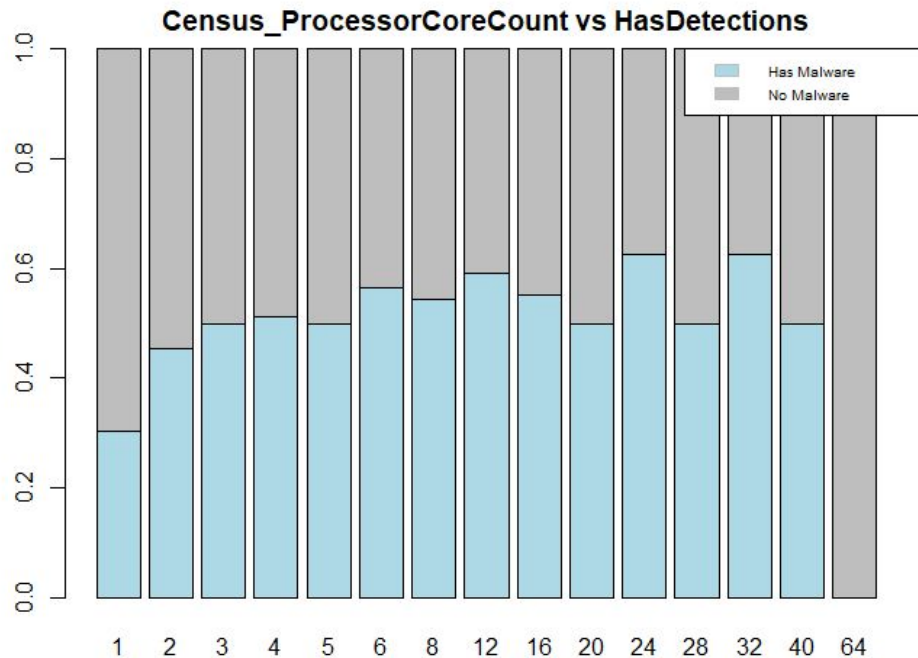
Number of logical cores in processor.

Logical cores determine the processing capabilities of a Windows machine.

Most machines have 4.

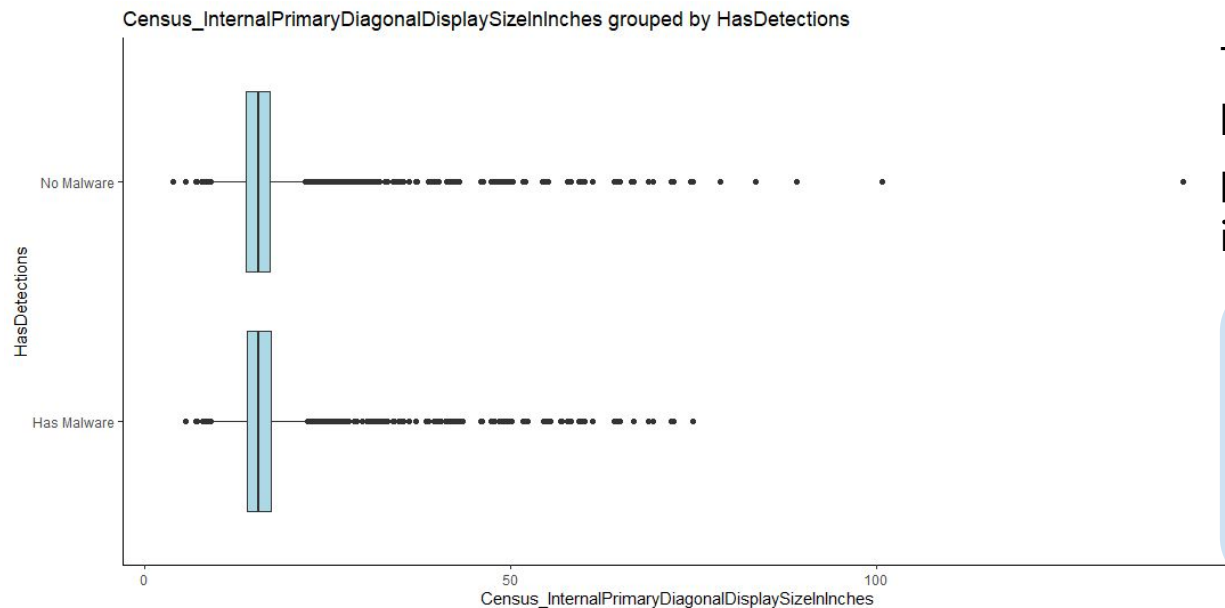
1 core machines are safest with 30% having malware, whereas 45% of 2 cores machines have malware and 51% of 4 cores machines have malware.

Number of cores in a Windows machine is **significant** in predicting malware attacks.





Census_InternalPrimaryDiagonalDisplaySizeInInches



The physical diagonal length of the machine's primary display in inches.

Both seem to have almost the same distribution with the exception of a few outliers.

Primary Diagonal Display Size alone **may not be significant** in predicting malware attacks.

Census_IsTouchEnabled

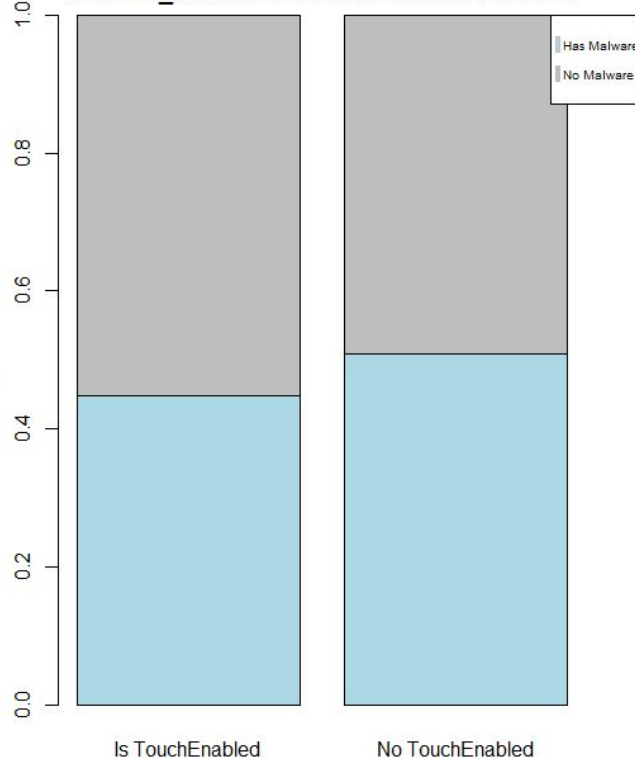


Whether or not the machine has touch capability.

Enables the user to interact with the machine using their fingers or a stylus, instead of using a mouse or a touchpad.

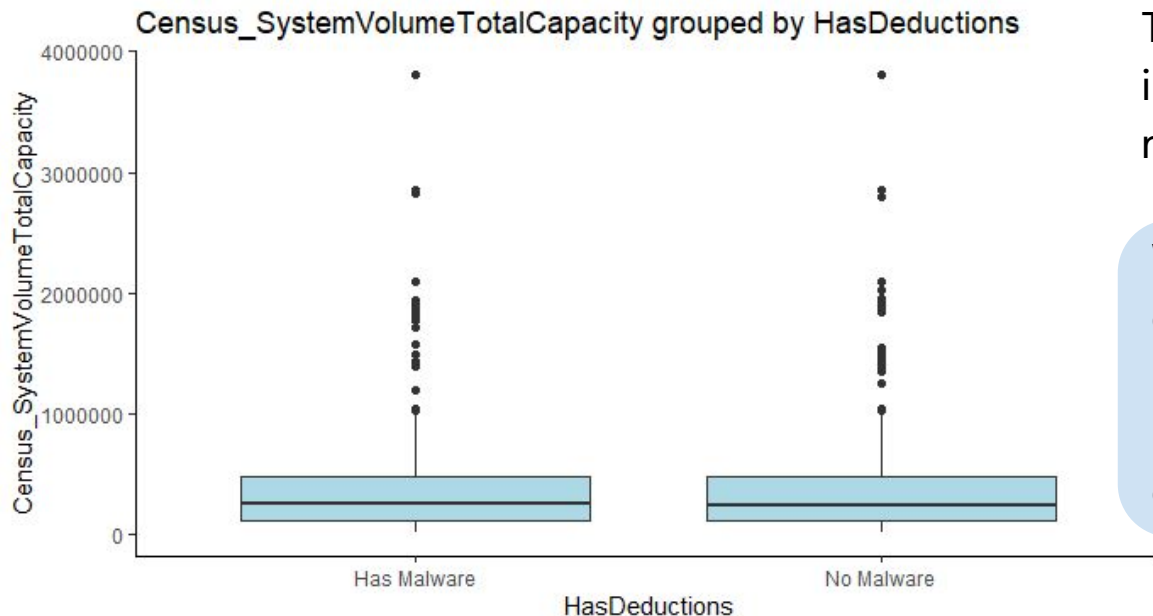
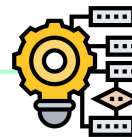
A larger proportion of Windows Machines with no Touch enabled feature have been detected to have malware

Census_IsTouchEnabled vs HasDetections



Touch capability **may be a significant factor** for a computer to be less prone to being attacked by malware.

Census_SystemVolumeTotalCapacity



The total memory space installed in the windows machine (in MB).

Windows machines that detect malware tend to have lesser system volume. However, the mean volume is quite similar.

System Volume Total Capacity alone **may not be significant** in predicting malware attacks.

04 Malware-resistant PCs Association Rule



Association Rules ($X \rightarrow Y$)

Support

Measures how frequently an item occurs and how often a rule is applicable

Confidence

Measures the probability of Y occurring, considering X has occurred

Lift

Measures how useful the rule is based on the context of the dataset and circumstances



Thresholds

To ensure that when the antecedent is present, the consequent is also likely to be present



Support

Minimum to be 1.5%



Confidence

Minimum to be 15%



Lift

Minimum to be 1

Association Rules (HasDetections = Yes)

Antecedent		Consequent	
Census_InternalPrimaryDiagonalDisplaySizeInInches = 21.5		HasDetections = Yes	
Confidence	Support	Lift	
0.5400	0.0169	1.08	



Association Rules (HasDetections = Yes)

Antecedent		Consequent	
Census_Processor CoreCount = 8		HasDetections = Yes	
Confidence	Support	Lift	
0.552	0.0542	1.11	



Spectre

permits location data to be read on random

Meltdown

permits a process to read all the memory in a machine's system

Association Rules (HasDetections = No)

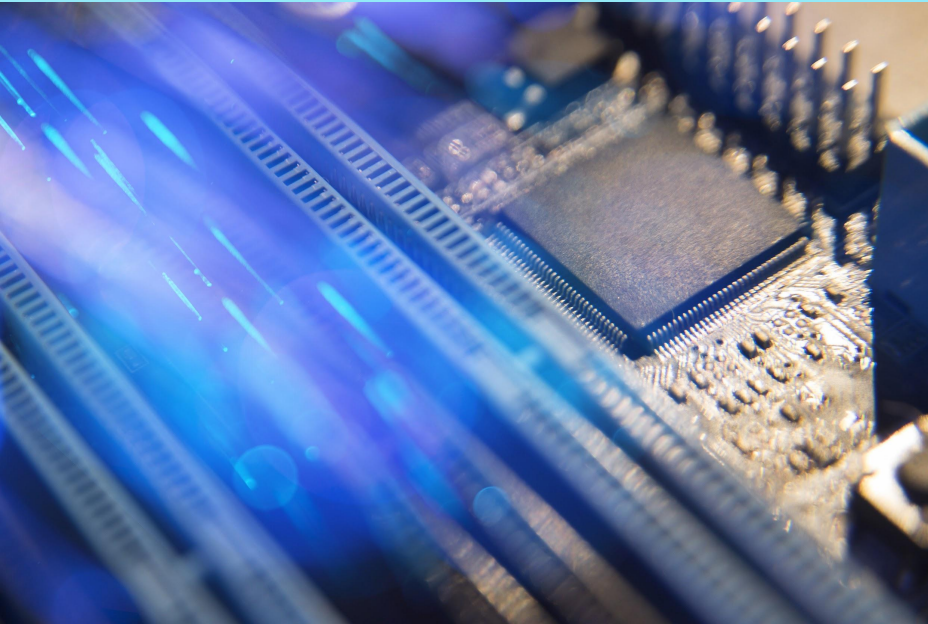
Antecedent		Consequent	
Census_InternalPrimaryDiagonalDisplaySizeInches = 11.6		HasDetections = No	
Confidence	Support	Lift	
0.5487	0.0199	1.10	



Association Rules (HasDetections = No)

Antecedent		Consequent	
Processor = x86		HasDetections = No	
Confidence	Support	Lift	
0.620	0.0557	1.24	



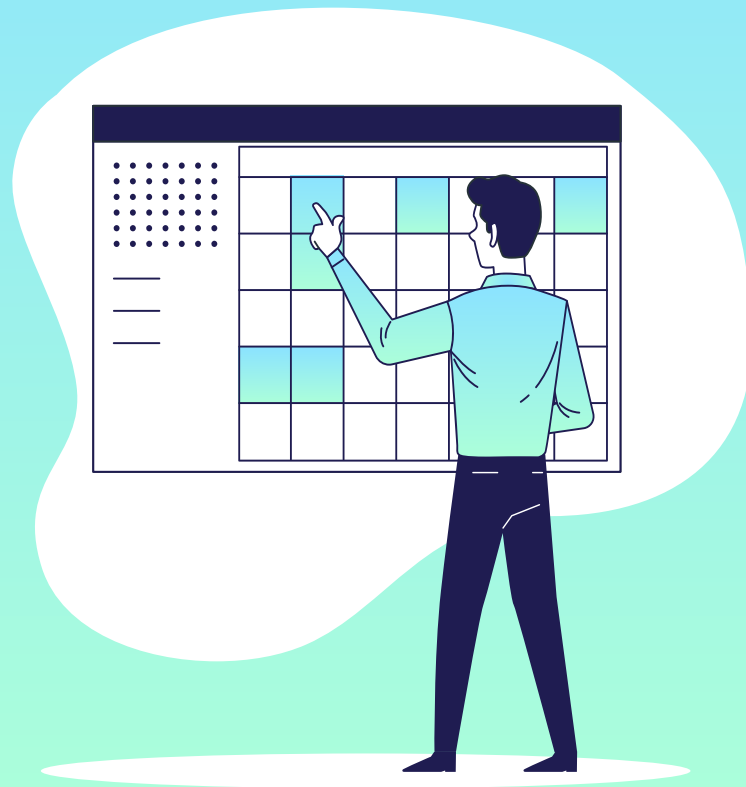


What can the x86 Processor do

- Allows for the assembly of Shellcode and the breaking down of malware
- Malware is often hidden within Shellcode, which could directly alter the programme and cause bugs

04

Malware-resistant PCs: Predictive Models



Predictive Models

Logistic Regression

- Binary dependent variable
- Feature selection

Random Forest

Neural Network



Logistic Regression (9 Variables)

Logistic Regression Variable	Coefficient
Census_TotalPhysicalRAM:4095.0	-1.00632
Census_InternalPrimaryDiagonalDisplaySizeInInches:8.8	-0.99003
Census_TotalPhysicalRAM:2047.0	-0.96696
Census_InternalPrimaryDiagonalDisplaySizeInInches:20.2	-0.88575
Census_InternalPrimaryDiagonalDisplaySizeInInches:8.0	-0.84325
Census_InternalPrimaryDiagonalDisplaySizeInInches:9.8	-0.84302
Census_InternalPrimaryDiagonalDisplaySizeInInches:27.9	0.939313
Census_InternalPrimaryDiagonalDisplaySizeInInches:32.0	0.807833
Census_InternalPrimaryDiagonalDisplaySizeInInches:19.7	0.803696

- Trainset Accuracy: **55.34%**
- Testset Accuracy: **55.08%**



Logistic Regression (32 Variables)

Logistic Regression Variables	Coefficient
AvSigVersion:1.273.1328.0	-1.2622
AvSigVersion:1.275.1718.0	-1.23009
AvSigVersion:1.273.1253.0	-1.22525
AvSigVersion:1.275.461.0	-1.20176
Cesus_PrimaryDiskTotalCapacity:61440.0	-1.19824
AvSigVersion:1.273.1393.0	-1.1628
AvSigVersion:1.275.884.0	-1.15523
AvSigVersion:1.273.761.0	-1.11052
AvSigVersion:1.273.1739.0	-1.05574
Census_InternalPrimaryDiagonalDisplaySizeInInches:10.6	-1.05344



Logistic Regression (32 Variables)

Logistic Regression Variables	Coefficient
AvSigVersion:1.275.132.0	1.362692
EngineVersion:1.1.15300.5	1.321501
SmartScreen:ExistsNotSet	1.264652
AvSigVersion:1.275.845.0	1.198322
AvSigVersion:1.275.164.0	1.152892
AvSigVersion:1.275.617.0	1.11545
AppVersion:4.10.14393.953	1.108201
AvSigVersion:1.275.380.0	1.106163
AvSigVersion:1.275.378.0	1.097013
AvSigVersion:1.275.675.0	1.082119

- Trainset Accuracy: **67.41%**
- Testset Accuracy: **65.02%**



Models

Logistic Regression

- Binary dependent variable
- Feature selection

Random Forest

- Fit classifying decision trees to improve prediction accuracy
- Feature selection

Neural Network



Random Forest (9 Variables)

Random Forest Variable	Importance
Census_SystemVolumeTotalCapacity	0.590201
Census_InternalPrimaryDiagonalDisplaySizeInInches:15.5	0.023095
Census_InternalPrimaryDiagonalDisplaySizeInInches:13.9	0.014776
Census_InternalPrimaryDiagonalDisplaySizeInInches:14.0	0.014442
Census_ProcessorCoreCount:4.0	0.011251
Census_PrimaryDiskTypeName:HDD	0.01106
Census_HasOpticalDiskDrive:1	0.010951
Census_HasOpticalDiskDrive:0	0.010937
Census_TotalPhysicalRAM:8192.0	0.010552
Census_InternalPrimaryDiagonalDisplaySizeInInches:21.5	0.010132

- Trainset Accuracy: **55.64%**
- Testset Accuracy: **55.06%**



Random Forest (32 Variables)

Random Forest Variable	Importance
Census_SystemVolumeTotalCapacity	0.103173
SmartScreen:ExistsNotSet	0.076835
AVProductsInstalled:1.0	0.017754
AppVersion:4.18.1807.18075	0.009518
EngineVersion:1.1.15100.1	0.008101
Census_InternalPrimaryDiagonalDisplaySizeInInches:15.5	0.007995
Census_TotalPhysicalRAM:4096.0	0.00741
Census_OSInstallTypeName:Update	0.006999
Census_OSInstallTypeName:UUPUpgrade	0.00691
Census_OSWUAutoUpdateOptionsName:Notify	0.00691

- Trainset Accuracy: **67.99%**
- Testset Accuracy: **65.22%**



Models

Logistic Regression

- Binary dependent variable
- Feature selection

Random Forest

- Fit classifying decision trees to improve prediction accuracy
- Feature selection

Neural Network

- Recognise hidden patterns and correlations in raw data
- Ability to continuously learn and improve



MLP (9 Variables)

MLP	False Positive Error Rate	False Negative Error Rate	Overall Accuracy
Train Set	37.44%	8.12%	54.44%
Test Set	37.79%	8.12%	54.09%



LSTM (9 Variables)



LSTM	False Positive Error Rate	False Negative Error Rate	Overall Accuracy
Train Set	3.02%	3.43%	93.55%
Test Set	24.53%	23.06%	52.42%

05

Comparing the Models

A comparison of 9-variable vs 32-variable models



Comparing Models

Logistic Regression

- Binary dependent variable
- Feature selection

Random Forest

- Fit classifying decision trees to improve prediction accuracy
- Feature selection

Neural Network

- Recognise hidden patterns and correlations in raw data
- Ability to continuously learn and improve



Comparing Models (9 Variable)

Model	False Negative Error Rate	Overall Accuracy
Logistic	18.81%	55.08%
Random Forest	17.74%	55.06%
MLP	8.12%	54.09%
LSTM	23.06%	52.42%



Comparing Models (32 Variable)

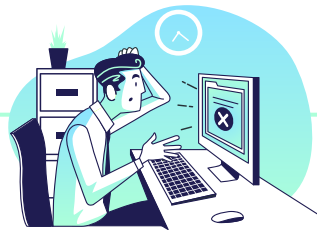


Model	False Negative Error Rate	Overall Accuracy
Logistic	18.72%	65.02%
Random Forest	15.21%	65.22%
MLP	13.04%	64.07%
LSTM	18.84%	64.23%

06

Recommended Solutions





**Key Influencing
Factors**



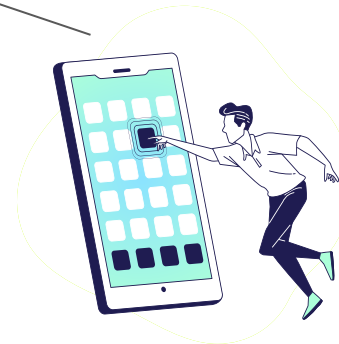
**Prediction
Models**



**Focus on Important
Features**



**Improve Existing
Devices**



**Testing New
Devices**

1. Focus on important features to enhance cybersecurity



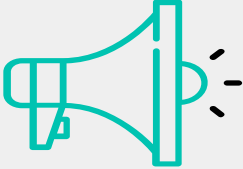
Important Features
Processor
Census_SystemVolumeTotalCapacity
Census_PrimaryDiskTotalCapacity
EngineVersion
AvSigVersion
SmartScreen
AVProductsInstalled



Collaboration with manufacturers to **research and investigate** the relationship between these features with malware attacks

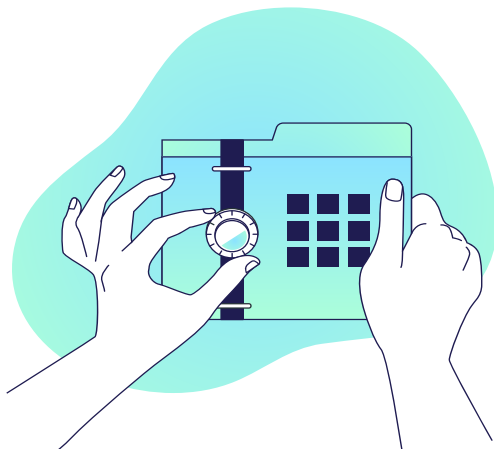
Produce devices with the **necessary hardware features** to become more secure

1. Focus on important features to enhance cybersecurity

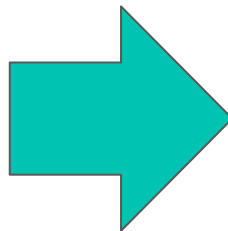
Focus Resources on More Secured Devices	Improve Software Features on Devices	Marketing Campaigns
 <p>Divert more resources to products that meet the hardware requirements and release more secure devices to the market</p>	 <p>Ensure windows defender software and antivirus database are kept up to date. Enable SmartScreen filter and install Antivirus on new devices</p>	 <p>Promoting secured devices that has the important features</p>

Microsoft will be able to change the current impression that their devices are less safe than Apple products.

2. Improving Existing Devices



Prediction Models



- **Test existing Windows machines in the market**
- **Sieve out devices with high vulnerability to malware.**

2. Improving Existing Devices

Recall Unsecured Devices



Limit production and marketing for unsecured devices. Reduce number of unsecured devices in the market

Introduce Software Updates

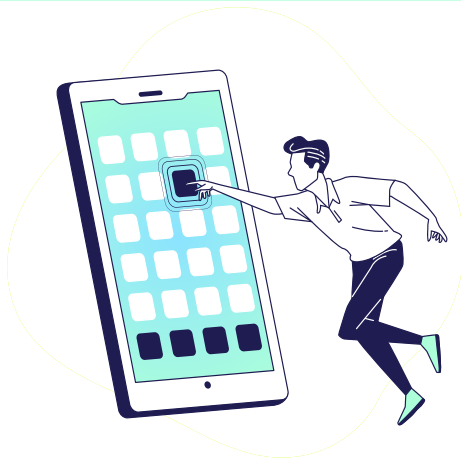
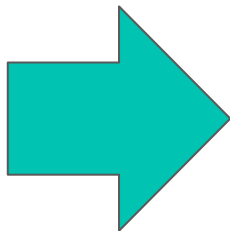


Improve the **Windows Defenders application** and its **Antivirus Signature Version** in these devices.

3. Testing New Devices



Prediction Models



- **Test new windows machines**
- **Prevent insecure devices from entering the market**
- **Introduce specifications for secured devices to important industries**

Limitations of Our analysis



Key influencing factors are derived from only a part of the dataset.

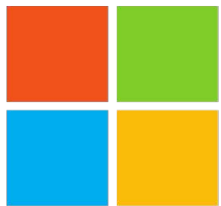
Lack of a powerful enough GPU, we are unable to run the analysis on all 82 variables



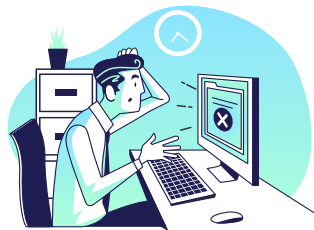
Lack of expertise in terms of computer hardware

Unable to discern the reason why such hardware and software features affects the probability of malware infection

Conclusion

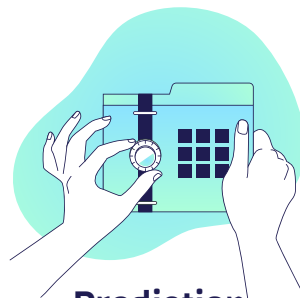


Microsoft



Key Influencing Factors

Consider the key influencing factors during their product development process



Prediction Models

Use prediction model to better gauge whether a change in a particular hardware feature is able to reduce the probability of malware infection

THANK YOU



Appendix

Dashboard of most important findings: Dashboard.jpeg

Overview of Variables to predict Malware Detection

