Microsoft Malware Prediction

Tsen Tai (contact) tsentai@usc.edu
April Ying-Chieh Chiu
I-Ting Wang
Yiping Liu
Yu-Heng Chiang

Agenda

Problem Overview

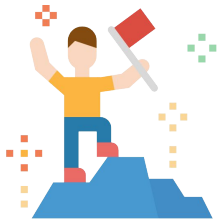
Approach

Results

Business Interpretations

Problem & Data

- Everyone hates it when their **devices becomes infected** and important information like passwords or credit card numbers are potentially stolen.
- We decided to choose the **Microsoft Malware Prediction Challenge** to better understand what variables make good predictors.
- **Predictors** - Machine Properties / **Response** - HasDetections



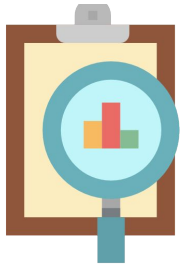
Approach

- Missing Value
- Subcategories
- Encoding

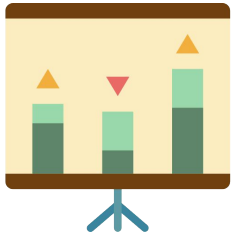
- Subset selection
- Classification model building

- Variable Importance
- Accuracy Rate
- AUC

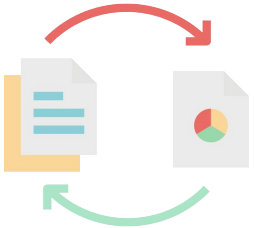
Data Preparation



Address data with low variation



Combine levels with smaller quantities

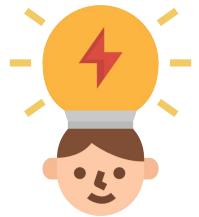


Dummy Encoding & Mean Encoding

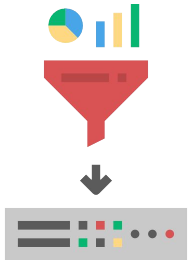
Modeling

What we tried:

- **Selecting subset** - Best Subset Selection, Forward/Backward Stepwise Selection, The Lasso
- **Building classification model** - XG Boosting, Random Forest, Logistic Regression, LDA



Final Modeling



Selecting subset:

The Lasso

Building model:

Random Forest

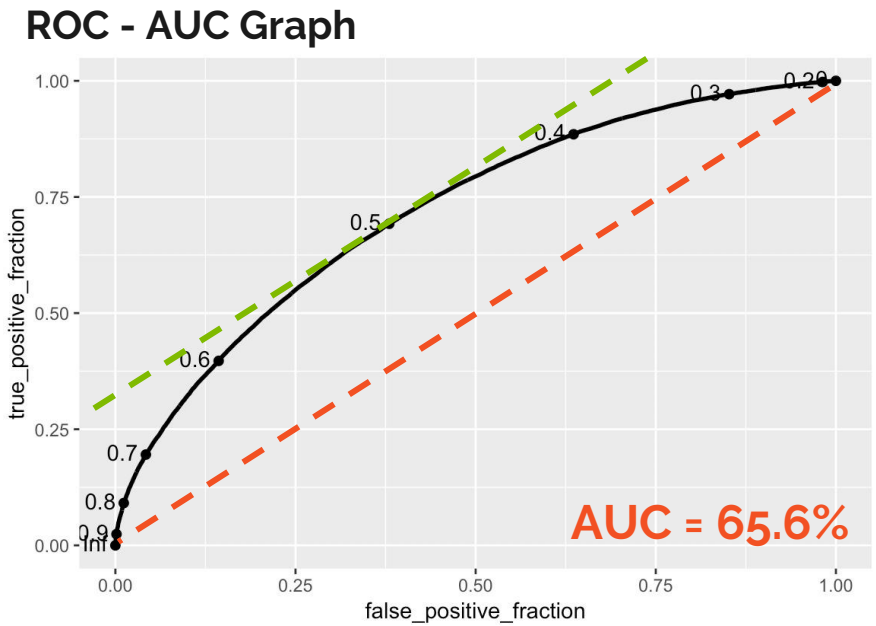
Results

Accuracy Rate

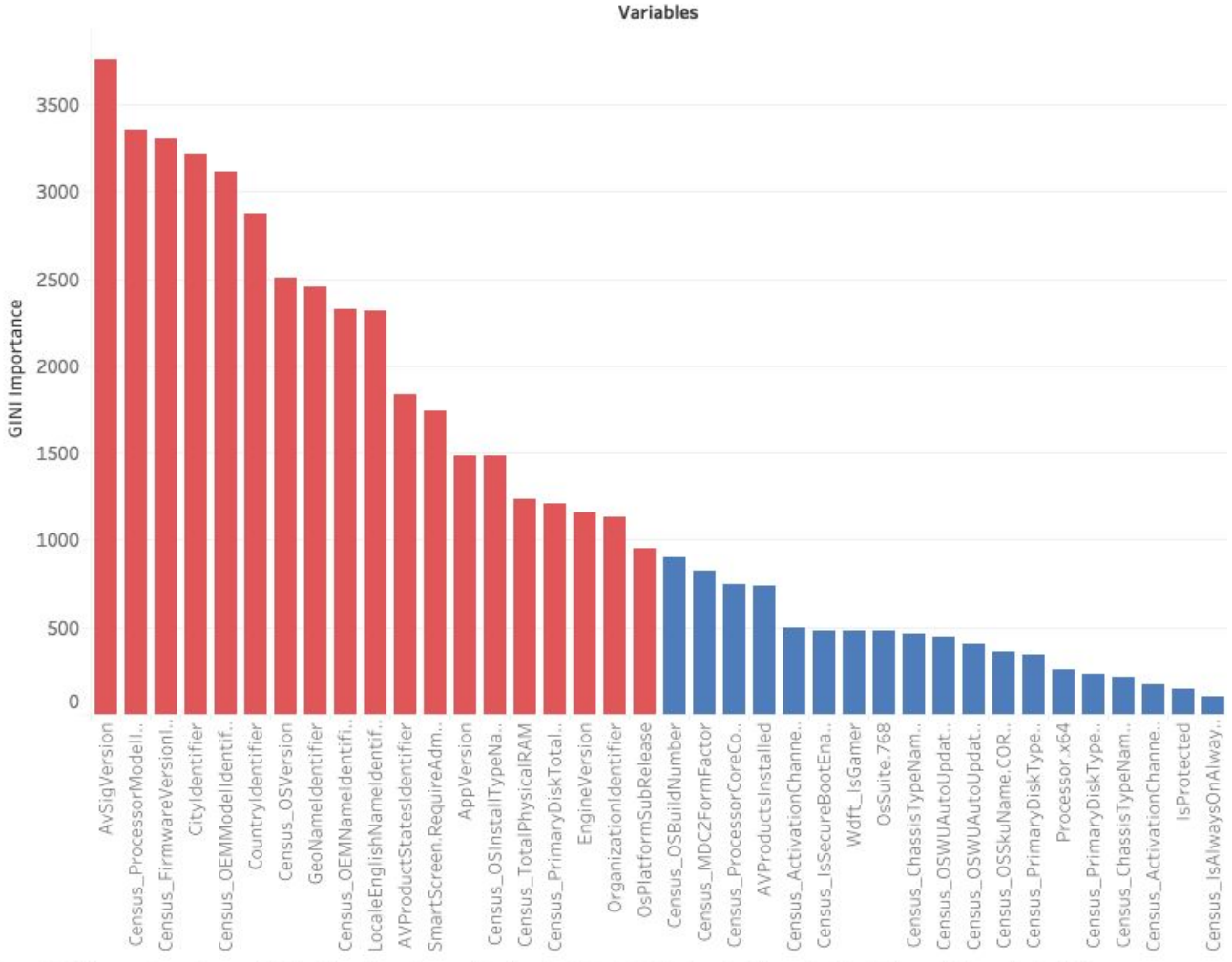
65.7%

Test Sensitivity

65.2%



Importance for each variable

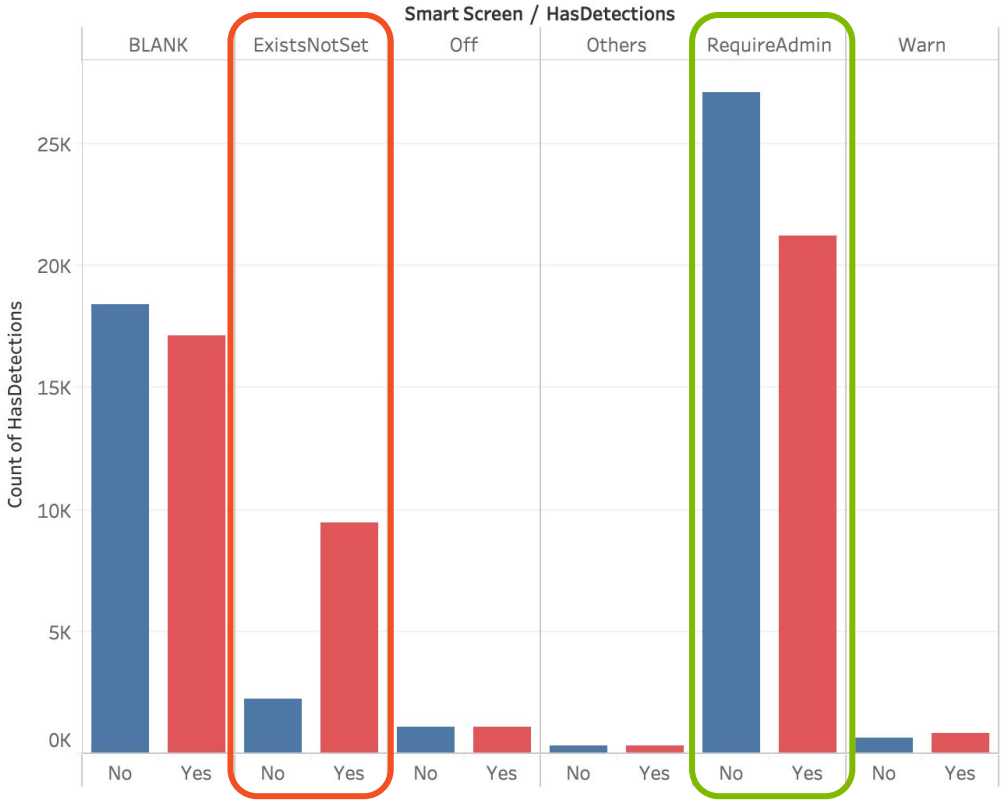


Business Interpretation

AvSigVersion,
AVProductStatesIdentifier,
SmartScreen.RequireAdmin,
AppVersion

Problem:
PC users may be **unaware**

Recommendation:
Educate users on already built-in options and **OPT-IN** to anti-malware protection programs.



Business Interpretation

CityIdentifier,
GeoNameIdentifier

Problem:
Location - some cities are targeted more than others.

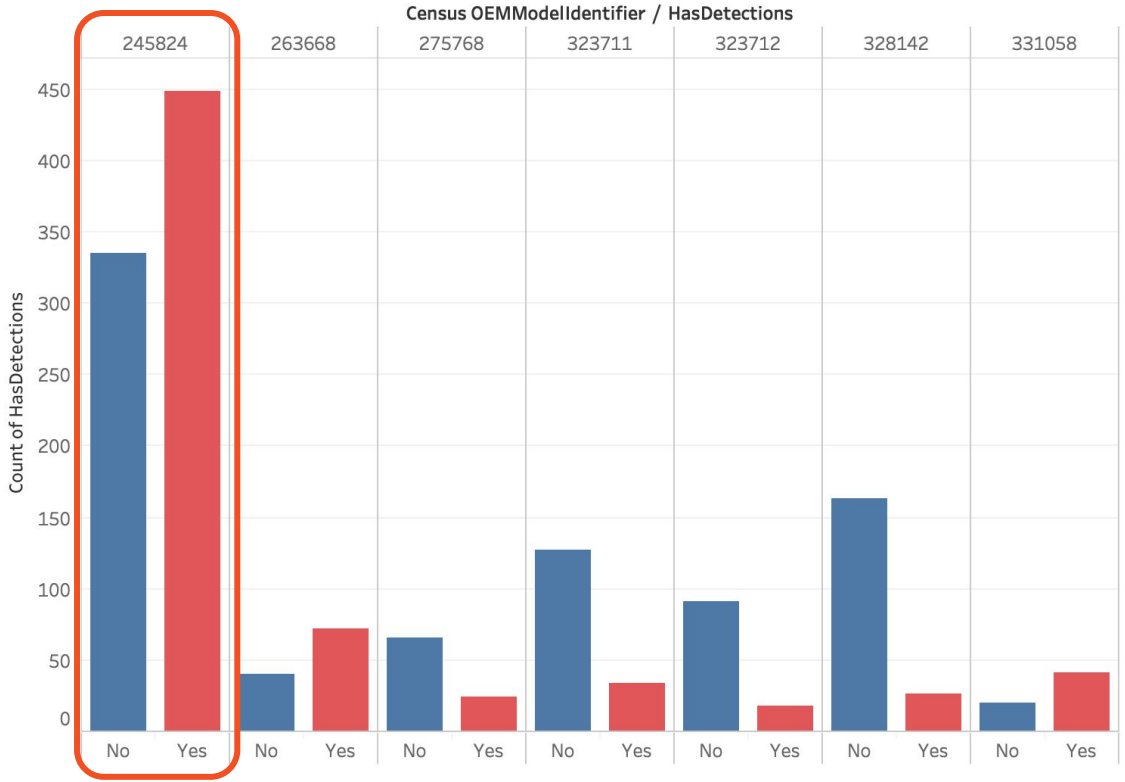
Recommendation:
Automatically opt-in to antivirus protection.

Business Interpretation

ProcessorModelIdentifer, OEMModelIdentifier, FirmwareVersionIdentifier

Problem:
OEM model, Primary disk capacity may be susceptible to malware

Recommendation:
Keep updating anti-malware protection services.



Thank you!