Springboard - Data Science Track
Jacqueline (Xuan) Guo

Capstone Project 2 - Milestone Report

Title:

Microsoft Malware Prediction (Kaggle Link)

Introduction:

Right now we are in a computer based society. Microsoft Windows is one of the most popular computers for people to use. However, once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways. Thus we could predict if the computer will soon be hit with malware, we can reduce the risk of malware infection.

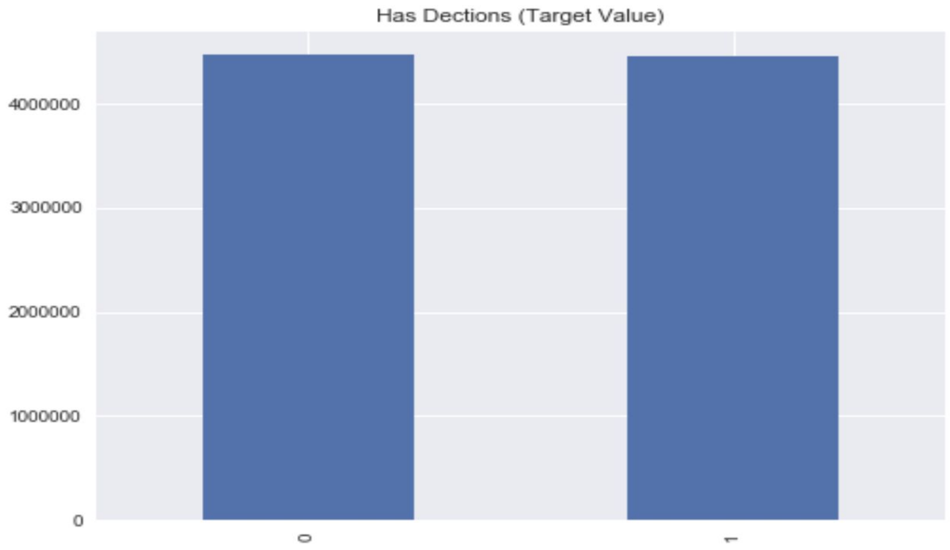
Computer services companies such as Microsoft or any windows users would like to apply this prediction to their system to keep their user's using experience safe. Most people or companies who use computers would care as well, for example, institutions like financial service companies or banks would love to keep their company information safe.

Dataset:

Dataset is from Microsoft of Kaggle competition. The size of the data is 7.89 GB with 167 columns and 8921483 rows. Microsoft provides one training dataset and one testing dataset which include the information about product name, different system versions, or different engine versions. This information could help us to predict a Windows machine's probability of getting infected by various families of malware, based on different properties of that machine.

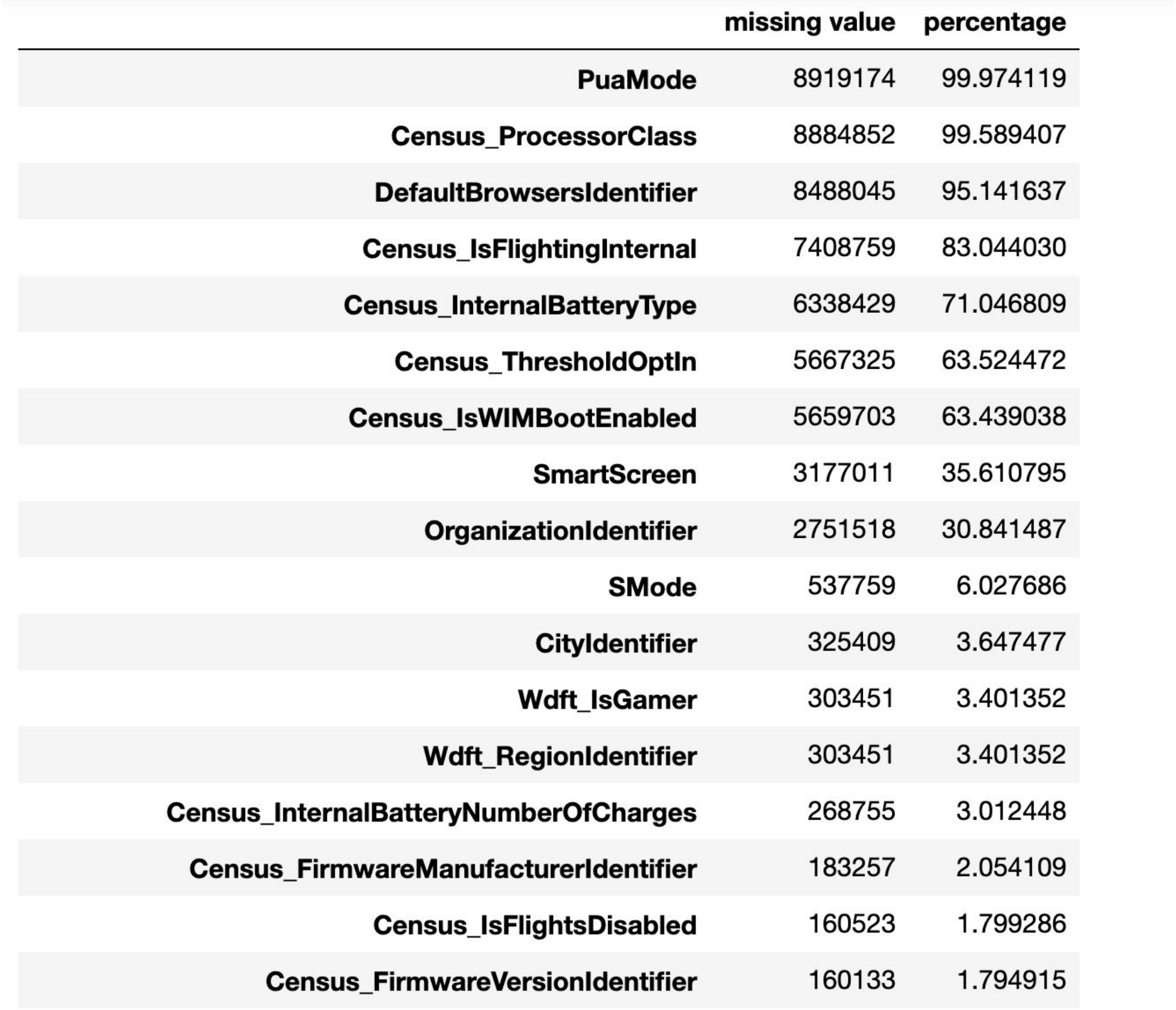
Preprocessing:

My second capstone project is available to view at: Capstone Project 2. First, I've checked if our dataset has a balanced target value distribution. I've drawn a histogram as following:



As the graph states, our labeled value is pretty balanced.

Then I want to check if there are a lot missing values in our dataset:

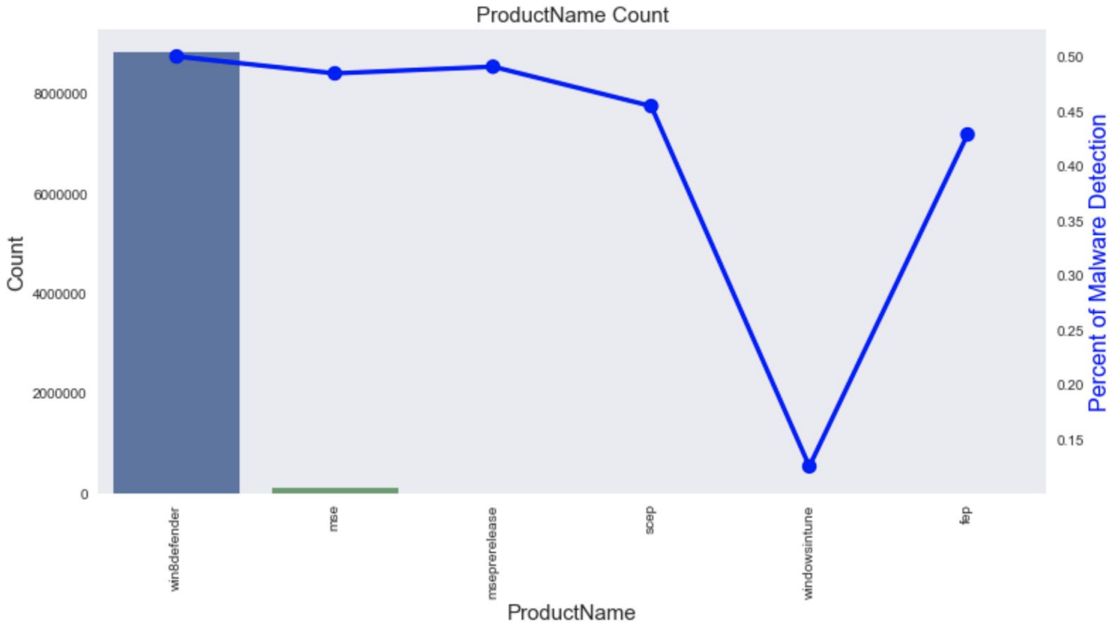


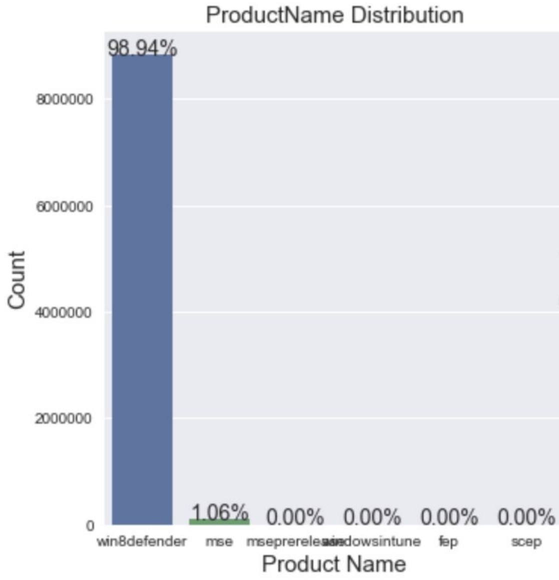
	missing value	percentage
PuaMode	8919174	99.974119
Census_ProcessorClass	8884852	99.589407
DefaultBrowsersIdentifier	8488045	95.141637
Census_IsFlightingInternal	7408759	83.044030
Census_InternalBatteryType	6338429	71.046809
Census_ThresholdOptIn	5667325	63.524472
Census_IsWIMBootEnabled	5659703	63.439038
SmartScreen	3177011	35.610795
OrganizationIdentifier	2751518	30.841487
SMode	537759	6.027686
CityIdentifier	325409	3.647477
Wdft_IsGamer	303451	3.401352
Wdft_RegionIdentifier	303451	3.401352
Census_InternalBatteryNumberOfCharges	268755	3.012448
Census_FirmwareManufacturerIdentifier	183257	2.054109
Census_IsFlightsDisabled	160523	1.799286
Census_FirmwareVersionIdentifier	160133	1.794915

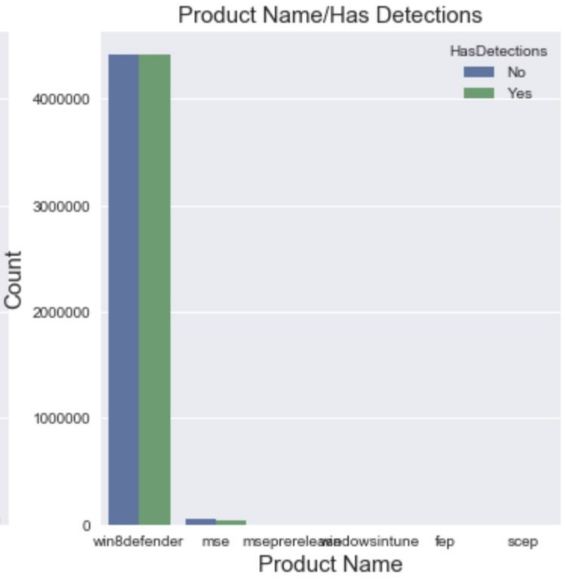
I found that I could drop four features by analysis which include: 'DefaultBrowsersIdentifier', 'PuaMode', 'Census_IsFlightingInternal', and 'Census_InternalBatteryType'.

Next step is to do the Exploratory Analysis focusing on categorical features.

I. Product Name

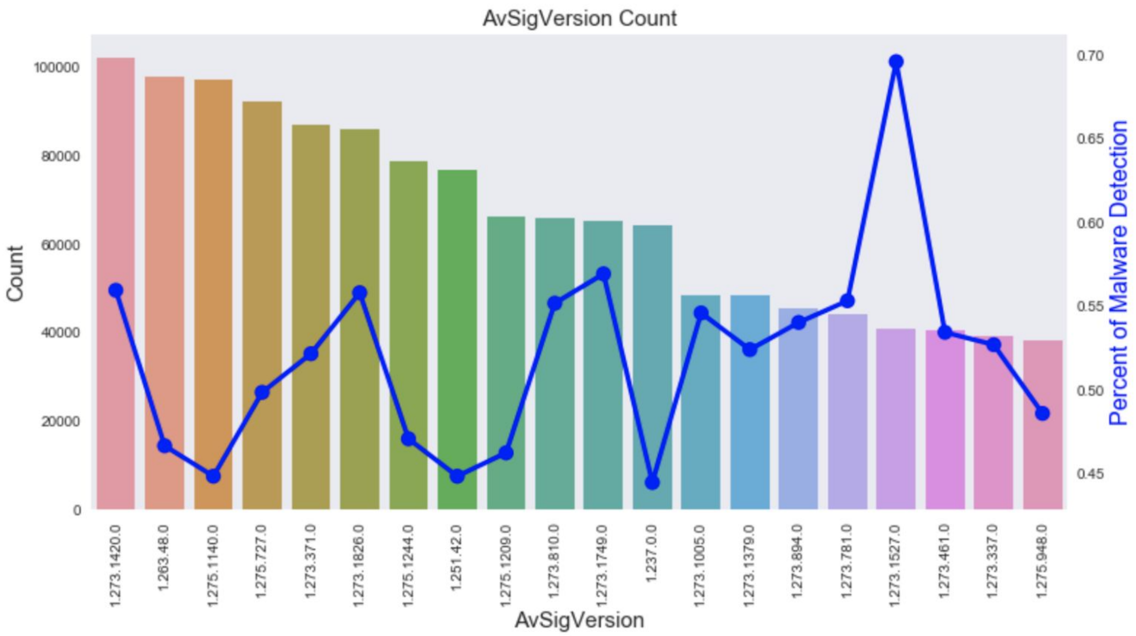






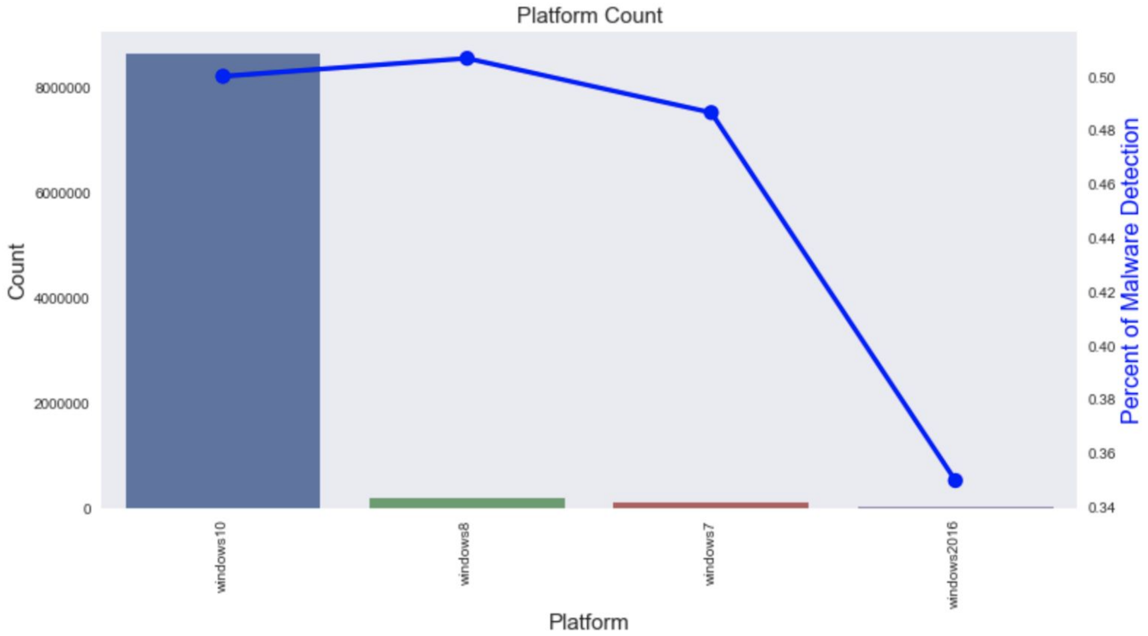
And above graphs we can see that windows8defender is the most common product and it has detection rate of around 50 %. We can also see that windowsintune has a very low detection rate. It's possible that malware is less likely to be detected in windowsintune.

II. AvSigVersion



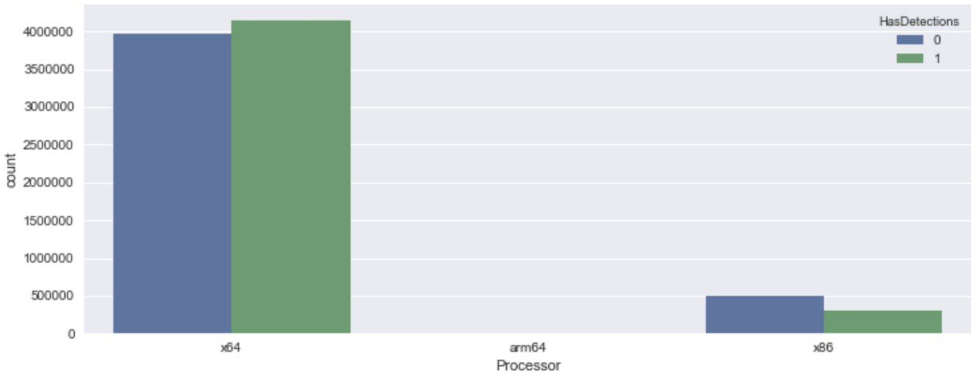
According to the graph, we can notice that there's an unusual peak of detection rate at version '1273.337'. Thus we should pay more attention to this feature.

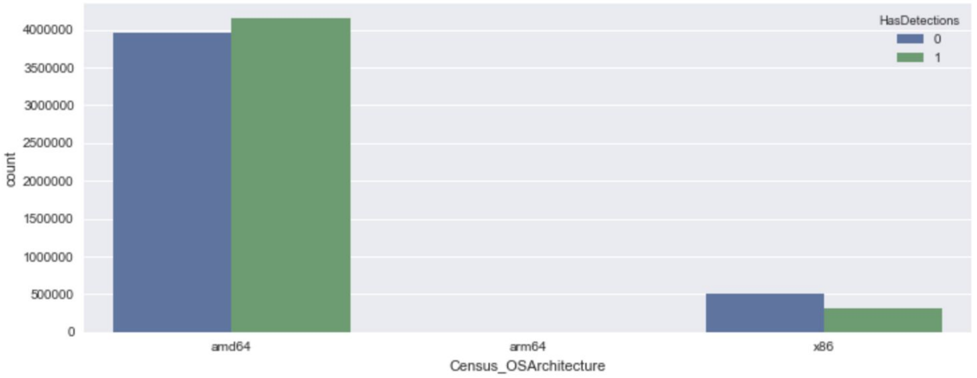
III. Platform



We can see that windows 10 is the most common platform and it has a detection rate of around 50%. And windows 2016 has a low detection rate.

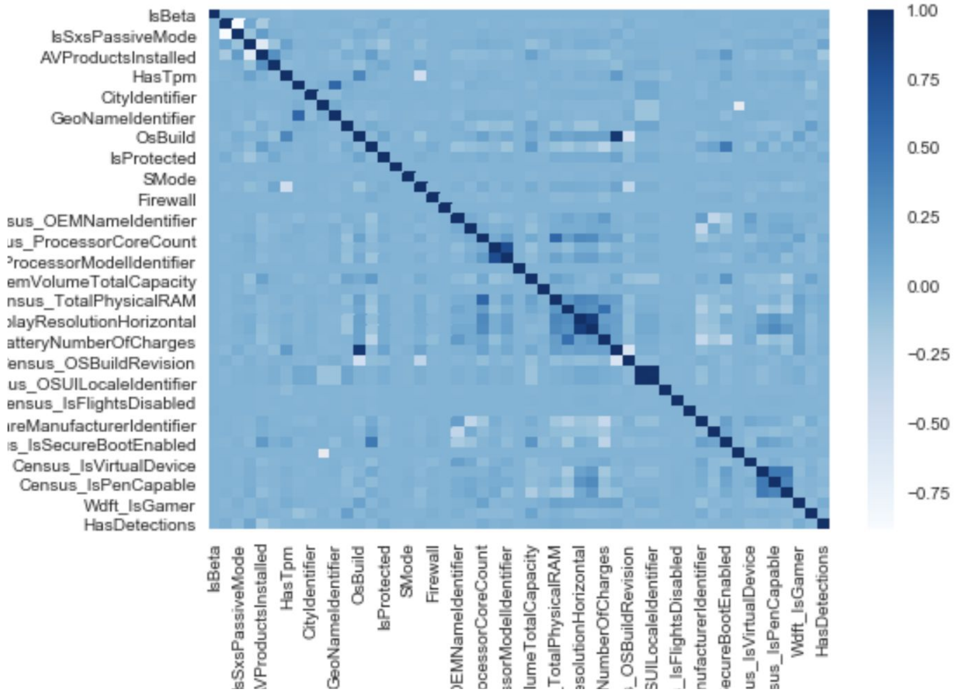
IV. Processor and Census OSArchitecture





We can see that for x86 processors, the detection rate is lower than other processors.

I draw a heatmap to measure the correlations between numerical features as following:



We could see from the heatmap that some of the features highly related to each other, thus we could apply feature selection to produce higher accuracy when fitting the model.

Deliverables:

1. Code notebooks
2. Report on the capstone project
3. Presentation or poster on the capstone project