Final Project Proposal- Malware Prediction in Microsoft Machines

Srishti Saha (ss1078)

05 November, 2019

Overview

Malware or malicious software is any form of software that is intentionally designed to cause damage to a computer, server, or a computer network. Enterprises like Microsoft take this to be a very serious problem. With over a billion enterprise and consumer customers, they want to develop data-science based techniques to predict if a machine will soon be hit with malware. Microsoft's next-generation detection solutions, like Windows Defender Antivirus uses data science, machine learning, automation, and behavioral analysis to provide 'real-time protection' to customers against threats. However, early detection of the possibility of a malware-attack will strengthen Microsoft's security. The goal of this project is to use data on different properties of a Windows machine to predict the probability of it getting infected by any kind of malware.

Research questions

Although Microsoft has increasingly used machine-learning algorithms along with AI to predict and prevent malware attacks within milliseconds, they are constantly working to improve the efficiency of these algorithms. They want to "to stop malware before it is even seen". The main objective questions being answered in this project are: * What system traits/properties of a machine make it the most vulnerable to a malware attack? * What parameter, if controlled for can provide a Windows machine the maximum security against malware? * Are there geographic factors that make a machine more vulnerable to malware attacks? * Do hardware traits and software traits reflect on a machines' vulnerability to a malware attack differently?

Data

The data is being sourced from Kaggle (linked here) where the data had been put up by Microsoft as a part of a sponsored competition. The telemetry data containing data on the machines' properties and the machine infections was generated by combining heartbeat and threat reports collected by Microsoft's endpoint protection solution, Windows Defender. This data contains anonymized data from 16.8M devices and has over 8,000,000 records of data. Some of the relevant variables in the dataset are:

- MachineIdentifier - Individual machine ID
- IsBeta - Defender state information e.g. false
- HasTpm - True if machine has tpm
- CountryIdentifier - ID for the country the machine is located in
- CityIdentifier - ID for the city the machine is located in
- GeoNameIdentifier - ID for the geographic region a machine is located in
- Platform - Calculates platform name (of OS related properties and processor property)
- Processor - This is the process architecture of the installed operating system
- OsVer - Version of the current operating system
- OsBuild - Build of the current operating system
- OsSuite - Product suite mask for the current operating system.
- OsPlatformSubRelease - Returns the OS Platform sub-release
- Census_PrimaryDiskTotalCapacity - Amount of disk space on primary disk of the machine in MB
- Census_PrimaryDiskTypeName - Friendly name of Primary Disk Type - HDD or SSD

The complete data dictionary are included in the Git repo here.

Project plan

This looks like a binary classification problem. It might be a good decision to use decision trees owing to the high number of variables present in the data and compare its performance to basic logistic regression. An ensemble model can be incorporated to learn and borrow information from various models. Data cleaning, imputation and EDA will be a major chunk of the project and will take 1.5 weeks (10 days). The model building and validation stage will then another 7 days. Validation and final results should take another 3 days to close the project.