The goal of this competition is to predict a Windows machine’s probability of getting infected by various families of malware, based on different properties of that machine. The telemetry data containing these properties and the machine infections was generated by combining heartbeat and threat reports collected by Microsoft's endpoint protection solution, Windows Defender.

Each row in this dataset corresponds to a machine, uniquely identified by a MachineIdentifier. HasDetections is the ground truth and indicates that Malware was detected on the machine. Using the information and labels in train.csv, you must predict the value for HasDetections for each machine in test.csv.

The sampling methodology used to create this dataset was designed to meet certain business constraints, both in regards to user privacy as well as the time period during which the machine was running. Malware detection is inherently a time-series problem, but it is made complicated by the introduction of new machines, machines that come online and offline, machines that receive patches, machines that receive new operating systems, etc. While the dataset provided here has been roughly split by time, the complications and sampling requirements mentioned above may mean you may see imperfect agreement between your cross validation, public, and private scores! Additionally, this dataset is not representative of Microsoft customers’ machines in the wild; it has been sampled to include a much larger proportion of malware machines.

Columns

Unavailable or self-documenting column names are marked with an "NA".

* MachineIdentifier - Individual machine ID
* ProductName - Defender state information e.g. win8defender
* EngineVersion - Defender state information e.g. 1.1.12603.0
* AppVersion - Defender state information e.g. 4.9.10586.0
* AvSigVersion - Defender state information e.g. 1.217.1014.0
* IsBeta - Defender state information e.g. false
* RtpStateBitfield - NA
* IsSxsPassiveMode - NA
* DefaultBrowsersIdentifier - ID for the machine's default browser
* AVProductStatesIdentifier - ID for the specific configuration of a user's antivirus software
* AVProductsInstalled - NA
* AVProductsEnabled - NA
* 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
* OrganizationIdentifier - ID for the organization the machine belongs in, organization ID is mapped to both specific companies and broad industries
* GeoNameIdentifier - ID for the geographic region a machine is located in
* LocaleEnglishNameIdentifier - English name of Locale ID of the current user
* 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 (Windows Vista, Windows 7, Windows 8, TH1, TH2)
* OsBuildLab - Build lab that generated the current OS. Example: 9600.17630.amd64fre.winblue\_r7.150109-2022
* SkuEdition - The goal of this feature is to use the Product Type defined in the MSDN to map to a 'SKU-Edition' name that is useful in population reporting. The valid Product Type are defined in %sdxroot%\data\windowseditions.xml. This API has been used since Vista and Server 2008, so there are many Product Types that do not apply to Windows 10. The 'SKU-Edition' is a string value that is in one of three classes of results. The design must hand each class.
* IsProtected - This is a calculated field derived from the Spynet Report's AV Products field. Returns: a. TRUE if there is at least one active and up-to-date antivirus product running on this machine. b. FALSE if there is no active AV product on this machine, or if the AV is active, but is not receiving the latest updates. c. null if there are no Anti Virus Products in the report. Returns: Whether a machine is protected.
* AutoSampleOptIn - This is the SubmitSamplesConsent value passed in from the service, available on CAMP 9+
* PuaMode - Pua Enabled mode from the service
* SMode - This field is set to true when the device is known to be in 'S Mode', as in, Windows 10 S mode, where only Microsoft Store apps can be installed
* IeVerIdentifier - NA
* SmartScreen - This is the SmartScreen enabled string value from registry. This is obtained by checking in order, HKLM\SOFTWARE\Policies\Microsoft\Windows\System\SmartScreenEnabled and HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Explorer\SmartScreenEnabled. If the value exists but is blank, the value "ExistsNotSet" is sent in telemetry.
* Firewall - This attribute is true (1) for Windows 8.1 and above if windows firewall is enabled, as reported by the service.
* UacLuaenable - This attribute reports whether or not the "administrator in Admin Approval Mode" user type is disabled or enabled in UAC. The value reported is obtained by reading the regkey HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Policies\System\EnableLUA.
* Census\_MDC2FormFactor - A grouping based on a combination of Device Census level hardware characteristics. The logic used to define Form Factor is rooted in business and industry standards and aligns with how people think about their device. (Examples: Smartphone, Small Tablet, All in One, Convertible...)
* Census\_DeviceFamily - AKA DeviceClass. Indicates the type of device that an edition of the OS is intended for. Example values: Windows.Desktop, Windows.Mobile, and iOS.Phone
* Census\_OEMNameIdentifier - NA
* Census\_OEMModelIdentifier - NA
* Census\_ProcessorCoreCount - Number of logical cores in the processor
* Census\_ProcessorManufacturerIdentifier - NA
* Census\_ProcessorModelIdentifier - NA
* Census\_ProcessorClass - A classification of processors into high/medium/low. Initially used for Pricing Level SKU. No longer maintained and updated
* 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
* Census\_SystemVolumeTotalCapacity - The size of the partition that the System volume is installed on in MB
* Census\_HasOpticalDiskDrive - True indicates that the machine has an optical disk drive (CD/DVD)
* Census\_TotalPhysicalRAM - Retrieves the physical RAM in MB
* Census\_ChassisTypeName - Retrieves a numeric representation of what type of chassis the machine has. A value of 0 means xx
* Census\_InternalPrimaryDiagonalDisplaySizeInInches - Retrieves the physical diagonal length in inches of the primary display
* Census\_InternalPrimaryDisplayResolutionHorizontal - Retrieves the number of pixels in the horizontal direction of the internal display.
* Census\_InternalPrimaryDisplayResolutionVertical - Retrieves the number of pixels in the vertical direction of the internal display
* Census\_PowerPlatformRoleName - Indicates the OEM preferred power management profile. This value helps identify the basic form factor of the device
* Census\_InternalBatteryType - NA
* Census\_InternalBatteryNumberOfCharges - NA
* Census\_OSVersion - Numeric OS version Example - 10.0.10130.0
* Census\_OSArchitecture - Architecture on which the OS is based. Derived from OSVersionFull. Example - amd64
* Census\_OSBranch - Branch of the OS extracted from the OsVersionFull. Example - OsBranch = fbl\_partner\_eeap where OsVersion = 6.4.9813.0.amd64fre.fbl\_partner\_eeap.140810-0005
* Census\_OSBuildNumber - OS Build number extracted from the OsVersionFull. Example - OsBuildNumber = 10512 or 10240
* Census\_OSBuildRevision - OS Build revision extracted from the OsVersionFull. Example - OsBuildRevision = 1000 or 16458
* Census\_OSEdition - Edition of the current OS. Sourced from HKLM\Software\Microsoft\Windows NT\CurrentVersion@EditionID in registry. Example: Enterprise
* Census\_OSSkuName - OS edition friendly name (currently Windows only)
* Census\_OSInstallTypeName - Friendly description of what install was used on the machine i.e. clean
* Census\_OSInstallLanguageIdentifier - NA
* Census\_OSUILocaleIdentifier - NA
* Census\_OSWUAutoUpdateOptionsName - Friendly name of the WindowsUpdate auto-update settings on the machine.
* Census\_IsPortableOperatingSystem - Indicates whether OS is booted up and running via Windows-To-Go on a USB stick.
* Census\_GenuineStateName - Friendly name of OSGenuineStateID. 0 = Genuine
* Census\_ActivationChannel - Retail license key or Volume license key for a machine.
* Census\_IsFlightingInternal - NA
* Census\_IsFlightsDisabled - Indicates if the machine is participating in flighting.
* Census\_FlightRing - The ring that the device user would like to receive flights for. This might be different from the ring of the OS which is currently installed if the user changes the ring after getting a flight from a different ring.
* Census\_ThresholdOptIn - NA
* Census\_FirmwareManufacturerIdentifier - NA
* Census\_FirmwareVersionIdentifier - NA
* Census\_IsSecureBootEnabled - Indicates if Secure Boot mode is enabled.
* Census\_IsWIMBootEnabled - NA
* Census\_IsVirtualDevice - Identifies a Virtual Machine (machine learning model)
* Census\_IsTouchEnabled - Is this a touch device ?
* Census\_IsPenCapable - Is the device capable of pen input ?
* Census\_IsAlwaysOnAlwaysConnectedCapable - Retreives information about whether the battery enables the device to be AlwaysOnAlwaysConnected .
* Wdft\_IsGamer - Indicates whether the device is a gamer device or not based on its hardware combination.
* Wdft\_RegionIdentifier - NA