ML PROJECT – SYSTEM THREAT FORECASTER

Disclaimer

This project is being shared solely for the academic learning and reference of Hitesh, Kavin Kishore, Meghana, and Shyam Sharan. I do not claim ownership of this project. The intellectual property rights belong to IIT Madras, as this work was originally done as part of my Diploma in Programming and Data Science coursework. This material should not be reproduced, distributed, or used outside the intended purpose.

Problem Statement:

The goal of this project is to predict a system's probability of getting infected by various families of malware, based on different properties of that system. The telemetry data containing these properties and the system infections was generated by threat reports collected by system's antivirus software.

Dataset Description:

Each row in this dataset corresponds to a machine, uniquely identified by a MachineID. target 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 target for each machine in test.csv.

Submission:

For each **id** in the test set, you must predict a class for the **target** variable. The file should contain a header and have the following format:

id	target
0	0
1	1
2	1
4	0
5	1

I will try to submit in my Kaggle portal to see if I still get the score. If not I will come up with a different way to submit it.

Submissions are evaluated on **accuracy_score()** between the predicted classes and the True target.

Dataset Links:

train.xlsx: https://docs.google.com/spreadsheets/d/1Cou4ja_yVyi--0kDiwhXQyUAWvdOdtmZ/edit?usp=sharing&ouid=106515426988878113606&rtpof=true&sd=true

test.xlsx:
https://docs.google.com/spreadsheets/d/1DbOtaoUiavR6b4cb3c2dhbFT5JaMbfh2/edit?usp=sharing&ouid=106515426988878113606&rtpof=true&sd=true

Metadata:

- 'MachineID' : Unique Identifier for Each Machine
- 'ProductName': Name of the Installed Antivirus Product
- 'EngineVersion': Version of the Antivirus Engine
- 'AppVersion' : Version of the Antivirus Application
- 'SignatureVersion': Version of the Antivirus Signatures
- 'IsBetaUser': Whether the User is on a Beta Version
- 'RealTimeProtectionState': Status of Real-Time Protection
- 'IsPassiveModeEnabled' : Whether Passive Mode is Enabled
- 'AntivirusConfigID' : Identifier for Antivirus Configuration
- 'NumAntivirusProductsInstalled' : Number of Installed Antivirus Products
- 'NumAntivirusProductsEnabled' : Number of Enabled Antivirus Products
- 'HasTpm' : Whether the Machine has a Trusted Platform Module (TPM)
- 'CountryID': Identifier for the Country of the Machine
- 'CityID' : Identifier for the City of the Machine
- 'GeoRegionID' : Identifier for the Machine's Organization or Industry
- 'LocaleEnglishNameID' : English Locale Name ID of the Current User
- 'PlatformType' : Platform Type Derived from OS and Processor Information
- 'Processor' : Processor Architecture of the Installed OS
- 'OSVersion' : Operating System Version
- 'OSBuildNumber' : OS Build Number
- 'OSProductSuite' : Product Suite Mask for the Operating System
- 'OsPlatformSubRelease' : Sub-release of the Operating System

- 'OSBuildLab' : Detailed OS Build Information
- 'SKUEditionName' : SKU Edition of the Operating System
- 'IsSystemProtected' : Whether the System has Active Protection
- 'AutoSampleSubmissionEnabled' : Auto Sample Submission Setting
- 'SMode' : Whether the Device is Running in S Mode
  'IEVersionID' : Internet Explorer Version Identifier
- 'FirewallEnabled' : Whether Windows Firewall is Enabled
- 'EnableLUA',
- 'MDC2FormFactor',
- 'DeviceFamily',
- 'OEMNameID',
- 'OEMModelID',
- 'ProcessorCoreCount',
- 'ProcessorManufacturerID',
- 'ProcessorModelID',
    - 'PrimaryDiskCapacityMB',
- 'PrimaryDiskType',
- 'SystemVolumeCapacityMB',
    - 'HasOpticalDiskDrive',
- 'TotalPhysicalRAMMB',
- 'ChassisType',
    - 'PrimaryDisplayDiagonalInches',
- 'PrimaryDisplayResolutionHorizontal',
    - 'PrimaryDisplayResolutionVertical',
- 'PowerPlatformRole',
    - 'InternalBatteryNumberOfCharges',
- 'NumericOSVersion',
- 'OSArchitecture',
    - 'OSBranch',

- 'OSBuildNumberOnly',
- 'OSBuildRevisionOnly',
- 'OSEdition',
    - 'OSSkuFriendlyName',
- 'OSInstallType',
- 'OSInstallLanguageID',
    - 'OSUILocaleID',
- 'AutoUpdateOptionsName',
- 'IsPortableOS',
    - 'OSGenuineState',
- 'LicenseActivationChannel',
- 'IsFlightsDisabled',
    - 'FlightRing',
- 'FirmwareManufacturerID',
- 'FirmwareVersionID',
    - 'IsSecureBootEnabled',
- 'IsVirtualDevice',
- 'IsTouchEnabled',
  'IsPenCapable',
- 'IsAlwaysOnAlwaysConnectedCapable',
- 'IsGamer',
    - 'RegionIdentifier',
- 'DateAS' : Malware signature dates ,
- 'DateOS' : timestamps for OSVersion which gives the time that the OS was last updated
- 'target'

Objectives:

Learn and perform following on the dataset:

1) Importing Libraries and Data Loading
2) Exploratory Data Analysis

3) Data Pre-Processing
    a. Feature Engineering
    b. Imputation
    c. Pipelines and Column Transformer
4) Model Pipeline
    a. Logistic Regression Model
    b. KNN Classifier
    c. Support Vector Classifier
    d. Decision Tree Classifier
    e. Random Forest Classifier
    f. XGBoost Classifier
    g. LightGBM Classifier
    h. HistGradientBoosting Classifier
    i. MLP Classifier
5) Model Training
6) Hyperparameter Tuning
7) Prediction
8) Comparative Analysis and Study of different classifiers output

Evaluation:

Upon successful completion, you are supposed to make a report and undergo a Viva with me.

Tips:

1) Just use one ipynb note book. Learn to write markdown and code with best practices.
2) Add comments wherever needed.
3) Feel free to use any LLMs for learning and coding.
4) If you get stuck, try reading or watching a YouTube video before asking doubts, self-learning is often more powerful than learning directly from me.
5) Enjoy the process and go at your own pace.
6) Do not compare yourself with anyone; just focus on your own progress.

ALL THE BEST