Microsoft Malware Prediction problem statement(Vinay Dalu)

Microsoft wants to predict the Windows machine's probability of getting infected by various families of malware, based on different properties of that machine given in the dataset by at least an accuracy of 80%. within next year.

H

1 Context

The malware industry continues to be a well-organized, well-funded market dedicated to evading traditional security measures. Once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways.Microsoft takes this problem very seriously and is deeply invested in improving security.So we have to predict if a machine will soon be hit with malware.

2 Criteria for success

Success for this project = To predict the malware (HashDetections=1 (or) True) in both train and test data and built a model with accuracy at least 80%

3 Scope of solution space

By analysing the training and test data and making it more refined when we are trying to make predictions.

4 Constraints within solution space

No constraints as per description given on kaggle .

5 Stakeholders to provide key insight

- **Microsoft analytics team.**
- **Arihant Jain**

6 Key data sources

Dataset from kaggle provided by Microsoft.

H D E I P

Initial Approach

- We should clean the dataset after we load into the jupyter notebook.

  1. Finding columns with high percentage of missing values.

  2. Drop the columns which have more than 30% of missing values.

  3. Impute with either (mean,median or mode) for the columns which have below 30% and highest correlation with HasDetection column.

  4. Use some visualization like (scatter plot,box plot,swarm plot , kde plot) to find the distribution of data and its correlation with respect to the target variable(HasDetections) and between themselves .

  5. Getting meaningful insights from the Visualization and move on to Feature Engineering.

  6.Applying Machine Learning Models which best fit the training data and providing generalized model for the unseen data.