What if your Machine is compromised and perpetuator is demanding ransom to give you access?

It can be avoided if we could predict them in advance.

PROBLEM STATEMENT

Microsoft wants to predict a windows machine's probability of getting infected by various families of malware with the help of provided properties by atleast an accuracy of 80%.

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.
With more than one billion enterprise and consumer customers, Microsoft takes this problem very seriously and is deeply invested in improving security.

2 Criteria to Success

Predict the malware ('HasDetactions') in the test data correctly for atleast 80% of the test data.

3 Scope of solution space

A dataset of 167 columns will be provided in which we have

102 - Integers	57 - String
4 - Decimal	4- Others

And our scope will be the above dataset.

4 Constraints within Solution Space.

None can be mentioned as of now

5 Stake Holders

- Microsoft analytics team head
- Arihant jain - Mentor

6 Key Data Sources

- Dataset from Kaggle.

Initial Approach

- Dataset needs to be cleaned after loading into the notebook. This includes

  1.) Searching columns with high cardinality and drop them.

  2.) Handling missing values and outliers.

  3.) We are provided with a lot of columns which needs observation and decide on truncating few columns like columns with more missing values, high cardinality etc.,

  4.) Analyse the distributions of numerical columns and find alike columns to reduce dimension furthermore by drop one among them.

- As we all looking for probability rather than decision. It is better to approach this problem with classification model.