MICROSOFT MALWARE PREDICTION

Introduction

Industries of malware and hacking are very powerful and growth sharply. It works to harm millions of victims and companies. These industries have great funding from a variety of organizations and people to meet their purposes. However, large global companies like Microsoft work hard to defense their customer from any attack.
Microsoft provides its dataset ‘ https://www.kaggle.com/c/microsoft-malware-prediction/overview ’ that includes many features about its consumer’s machine devices that collected by by Microsoft's endpoint protection solution and Windows Defender. It makes open source in Kaggle to challenge the data scientist to apply various machine learning models to predict malware before occurring.

What is the problem

There are two datasets, each row into that refer to machine device via ‘MachineIdentifier’ column. The ‘train.csv’ dataset that provides ‘HasDetections’ as a target. Therefore, the goal of this project to predict the value for ‘HasDetections’ for each machine in the test.csv dataset after training the model on previous dataset.
The prediction value is categorical that includes ‘0 as hasn’t been detected or 1 as has been detected’. So, classification models will be used such as logistic regression, random forest, and RNN.

Describe the data

The data contains two datasets ‘train.csv’ and ‘test.csv’. It's very large data; each dataset includ es 8921482 rows and 82 columns.

It required heavy cleansing to remove the empty 10 columns which are unavailable value or have more than 25% of missing value. So, deleting these columns is best solution to improve the accuracy of prediction and fill-in the missing values in 35 columns. Other missing values don’t exceed 3%, therefore replacing the missing value with the mean and the median would be good.
The features in the dataset so large so, using the function ‘. corr()’ helps to extract only the features that have a positive relationship with the target. Also, some features may need to be implemented ‘. get_dummy’ function on it before starting modeling
‘train.csv’ dataset will be split to train and test with cross_validation, then it’ll be applied to the model in ‘test.csv’ dataset.

What are the tools

There are many tools that use to achieve the project.

- Main libraries
    1. Pandas
    2. NumPy

- EDA library
    1. ProfileReport

- Visualizing libraries
    1. Matplotlib
    2. Seaborn
    3. Plotly

- Modeling libraries
    1. sklearn.metrics
    2. train_test_split
    3. cross_val_score
    4. LogisticRegression
    5. RandomForestClassifier
    6. keras