Microsoft Malware Prediction

Kavya Sahai
Neil Shah
Priyal Agrawal
Shabari Girish Ganapathy

Problem Statement

Malware is one of the important problems in today's software world as once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways.

Goal

The goal of this project is to predict a Windows machine's/ cloud cluster's probability of getting infected by various families of malware, based on different properties of that machine.

Data Description

Datasource:
https://www.kaggle.com/c/microsoft-malware-prediction/data

- 9.5 GB dataset
- 9M data points
- Input: 83 features (MachineIdentifier, Firewall, OSPlatformSubRelease, EngineVersion etc.)
- Output: HasDetections (True/False)

Challenges

1. Switch from local to cloud (Google Cloud Deep-Instance)
    a. Computational constraints
    b. Time constraints
    c. Memory

2. Model Accuracy

Challenges

3. Data Pre-processing:
    a. Lots of Missing Features
    b. Skewed Features
    c. Text Fields

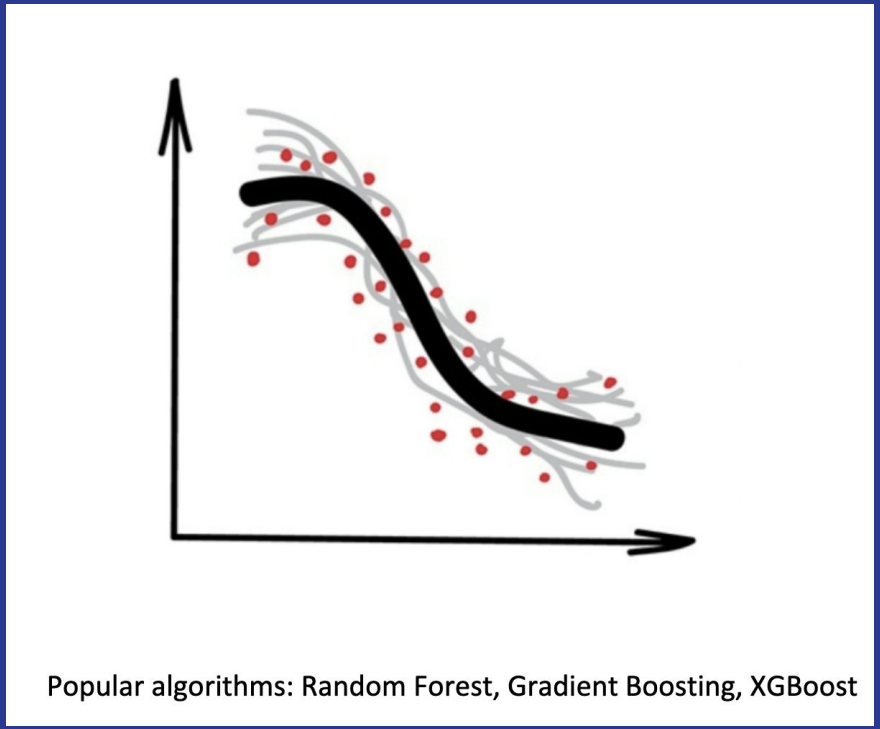
Feature Engineering

- Removed unnecessary features/ missing features
- Removed skewed features
- Transformed categorical data into numeric data using scikit-learn LabelBinarizer
- Used scikit-learn pipeline for numerical and categorical attributes processing
- Combined features using scikit-learn FeatureUnion

Model Selection

- We first started training different models (SVC and Logistic regression).
- Did not get very good accuracy. ~48-49%
- We then resorted to Ensemble Learning Techniques

What is Ensemble Learning



Ensemble Learning techniques used in this project

- Max Voting
- Averaging
- Weighted Averaging
- Boosting (Light GBM)

Results

Kaggle Competition Winner Accuracy- 67.585%

Our Models:

- Logistic Regression - 48.831%
- Simple SVC - 49.276%

After Ensemble Learning:

- Max Voting - 49.99%
- Averaging - 59.87%
- Weighted Averaging - 50.98%
- Boosting - 55.89%