Microsoft Malware Prediction

Daniel Omeh

Gopal Masilamani

Nugzar Nebieridze

W207 - Machine Learning - Fall 2019

Dec 09, 2019

Problem Description

- Problem origin - Kaggle Microsoft competition
- Project goal - achieve similar results as the competition’s top winning teams
- 8 million machine identifiers, 82 features and 8GB of data
- https://www.kaggle.com/c/microsoft-malware-prediction/overview

Approach

- Explore the existing data
- Data Normalization
- Use Multiple Machine Learning Models
- Perform Parameter Tuning
- Finalize the best scoring algorithm

Data Processing

- Removed the fields that in documentation were mentioned as obsolete
- Converted classifiers/strings to numbers
- Ran the dataset through StandardScaler for some models that perform better
- Removed the features that were causing overfitting (build number in versions)
- Removed the features that had NaN values most of the time 90%+
- Removed the features that had mostly garbage in them (Battery type)

Feature Engineering

- Principal component analysis (PCA)
- Selecting most relevant features (AV installed, Firewalled, Version…)
- Checking skewness of data and unifying values
- Generating new features (RAM by core number, Screen resolution…)
- Splitting version numbers into separate features
- Using good old brute force ;) (Dropping features one by one)

Models

- KNeighborsClassifier
- ExtraTreesClassifier
- DecisionTreeClassifier
- RandomForestClassifier
- GaussianNB
- GradientBoostingClassifier
- HistGradientBoostingClassifier
- AdaBoostClassifier
- Neural Network

Parameter Tuning

Area Under the ROC Curve

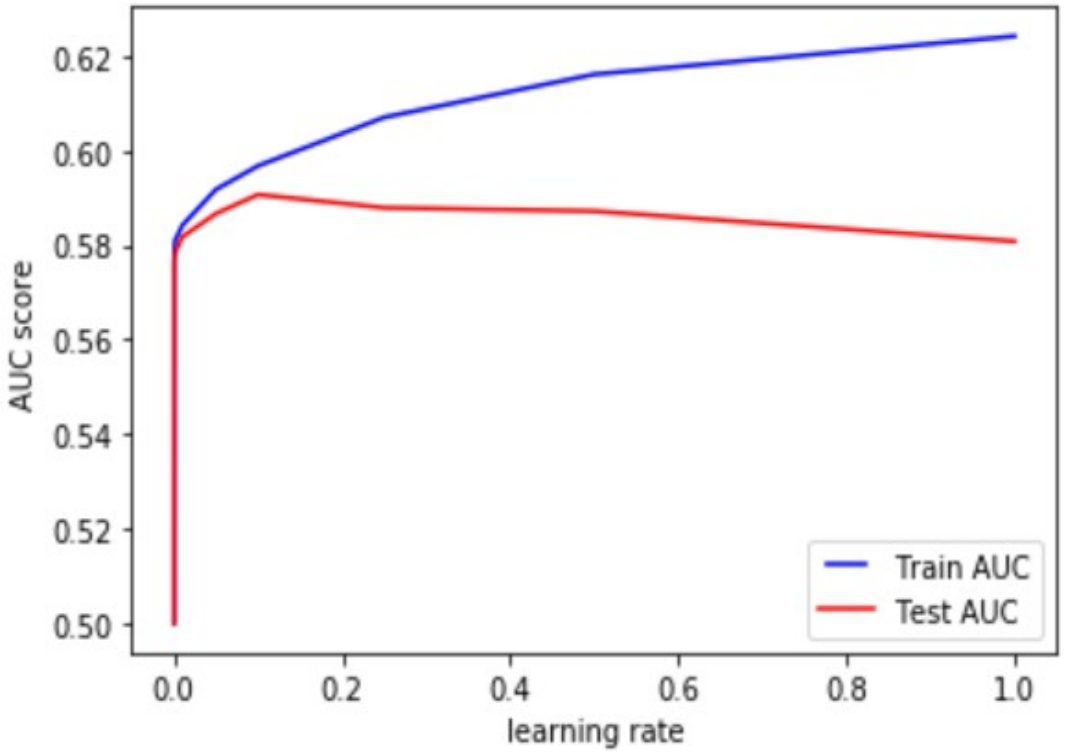
- Tree-Specific Parameters
- Boosting Parameters
- Miscellaneous Parameters

Parameters

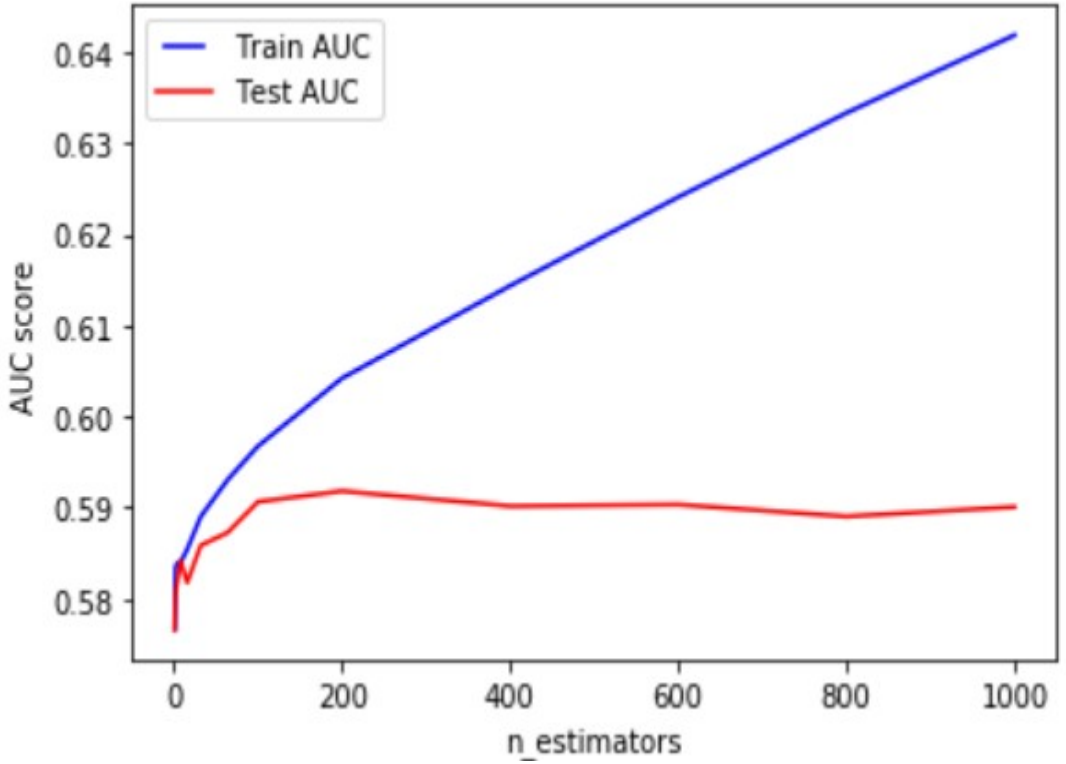
- n_estimators
- learning_rates
- max_depths
- min_samples_splits
- min_samples_leafs

Parameter Tuning

Learning Rate



n_estimator



Overall Results

- KNeighborsClassifier - 54.65%
- ExtraTreesClassifier - 58.61%
- DecisionTreeClassifier - 61.03%
- RandomForestClassifier - 62.08%
- GaussianNB - 50.24%
- GradientBoostingClassifier - 61.62%
- **HistGradientBoostingClassifier - 63.87%**
- AdaBoostClassifier - 62.23%
- Neural Network - 61.36%

The End

THANK YOU!

GitHub link:
https://github.com/nugzar/mics-w207/tree/master/final