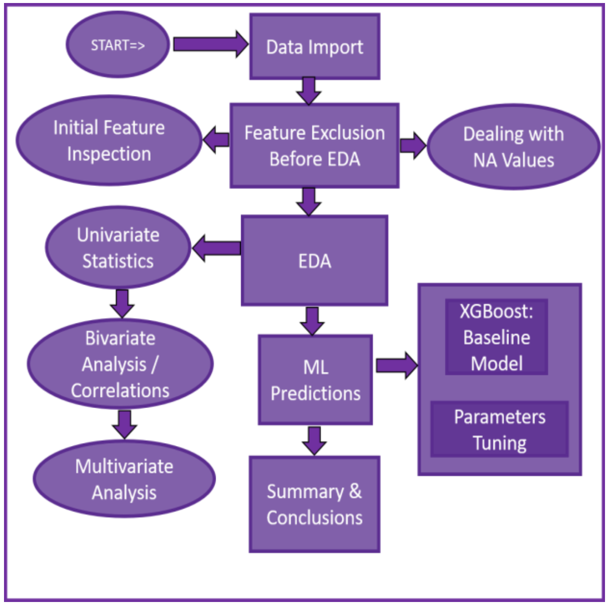
Malware Prediction Using Machine Learning

Priyal Nile, Hitesh Patel
pan303@nyu.edu, hlp276@nyu.edu

Introduction

Abating the risk of malware by using modern machine learning techniques by malware occurrence prediction in the future, based on the system configuration.

Methodology



START=>

Data Import

Feature Exclusion Before EDA

Initial Feature Inspection

Dealing with NA Values

EDA

Univariate Statistics

Bivariate Analysis / Correlations

Multivariate Analysis

ML Predictions

XGBoost: Baseline Model

Parameters Tuning

Summary & Conclusions

Conclusion

The baseline model we used, i.e. XGBoost provides a good accuracy of 69.40%. Creating new features based on the crucial features used in our model the accuracy can be further increased.

Future Work

There is a possibility of creating a probabilistic time series modeling. We can further enhance the performance & accuracy by doing more advanced feature engineering, ensemble modeling & Neural Network implementation.

Dataset

Data	DataSize	Records
Train Data	4.08 GB	9 Million
Test Data	3.54 GB	8 Million

Results

ML Model	Data	Accuracy
XGBoost	Training	71.80%
XGBoost	Testing	69.40%

References

[1] https://www.kaggle.com/c/microsoft-malware-prediction/overview
[2] https://dl-acm-org.proxy.library.nyu.edu/citation.cfm?id=3318448
[3] https://dl-acm-org.proxy.library.nyu.edu/citation.cfm?id=3196515
[4] XGBoost Documentation: https://github.com/dmlc/xgboost
[5] Cornell University: Computer Science: XGBoost: A Scalable Tree Boosting System https://arxiv.org/abs/1603.02754