**Malware Meets Data Science: Exploration of Malware Using K-Nearest Neighbors (kNN)** Stephanie Zhang

Department of Data Science, University of Maryland Baltimore County

DATA 699: Independent Study

Dr. Ajinkya Borle

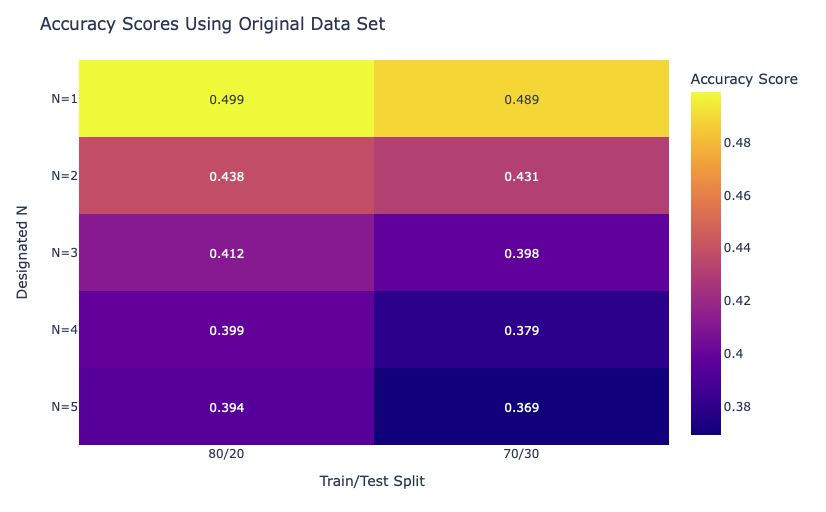
December 20, 2022

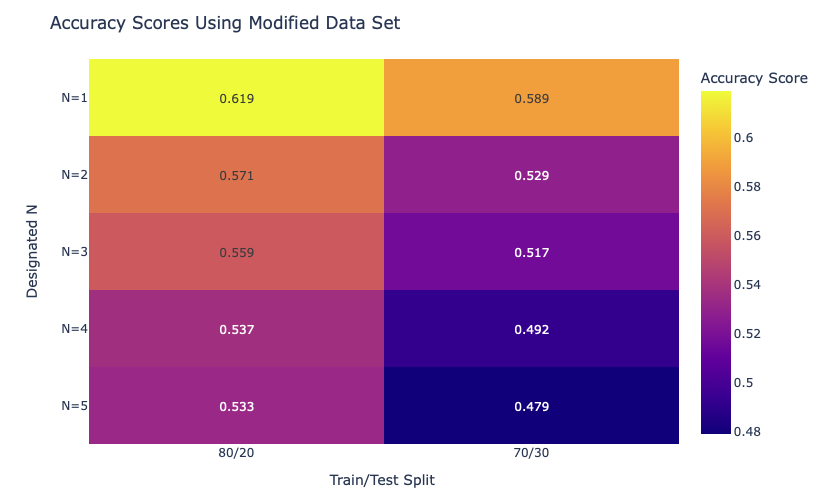
**Introduction**

This research study focuses on the exploration of the Malware Open-source Threat Intelligence Family (MOTIF) data set. This data set has 3,095 malware samples, with the viruses removed. Previously, XG Boost and computational neural networks (MalConv net) were used to set accuracy benchmarks for this dataset. Now we want to explore using kNN as an alternative. K-Nearest Neighbors (kNN) is a supervised machine learning algorithm. Based on a determined number of classifiers k, a data point is predicted from the k points around it.

kNN comes with pros and cons. Cons include decreased speed of the model as k increases, memory inefficiencies, and outlier sensitivity. Pros include easy application, interpretation, and classification needs. Due to the groupings of malware families, we experiment to discover whether kNN is a reliable model for this data set. Note: Each model will have its data written to a csv file for reference purposes.

**Results**





**Conclusion**

K=1 has the best accuracy for both the original and modified data sets. However, the kNN model is not the best model for the MOTIF data set. For good kNN models, it should be that as k increases, accuracy increases, which is the opposite case here. Certain models will perform better (see page 8 of the original white paper contained in GitHub), our model performs better than MalConv2 (accuracy of .487), but worse than LightGBM (accuracy of .724).

**References**

https://github.com/boozallen/MOTIF