Financial Analytics Project

*Part 1 Report*

Computing the feature importance scores

for a clustered malware dataset.

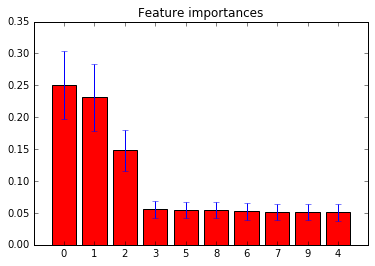
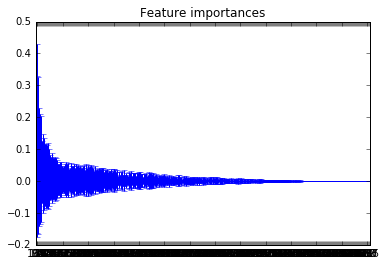
Ezeanaka Kingsley Uchechukwu

**Introduction**

In the area of machine learning and statistics, feature selection is a term commonly used to refer to the process of selecting a subset of relevant features, which could be variables or predictors, for using when developing a training model. Feature selection is also known as variable selection, attribute selection or variable subset selection. The process of grouping the entire set of features in some order of relevance is termed **feature ranking.** The values associated with each rank are called feature importance values or feature ranking values. The idea is based on the fact that datasets are bound to contain many features that are either redundant or irrelevant, and thus can be removed without incurring much of information.

Feature selection techniques simplify models for easier interpretation by researchers and users, reduce model training times, avoid unnecessary complications associated with dimensionality and also to enhance generalization of dataset by reducing overfitting.

There are a number of algorithms that are used to compute the importance scores of various features of dataset. The most common ones which are utilized in the project include: Linear Regression, Ridge method, Lasso algorithm, MIC model, and Random forest.

1. (b)

Fig.1 – Feature importance graph for (a) 10-feature dataset, (b) 350-feature dataset

Figure 1 above illustrates a visual picture of the relative importances of features in a dataset. The vertical axis represents the range of importance scores and the horizontal axis represent the labels for the features. The features are sorted with the most important features coming first.

The error bars are reflective of the OOB or out of bag error estimate which calculates the mean prediction error on each data point of the dataset being fit into the random forest and is averaged over the entire forest. The importance score for the j-th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences.

**Related Work**

Feature ranking and selection is a fairly popular concept in the field of machine learning, and as such, there are a number of other algorithms in the industry that also compute the importance scores of dataset features. A number of them worthy of note include the recursive feature elimination algorithm,

**Program Pipeline**

The GUI takes a pair of datasets as inputs the first input is the raw dataset with as many features and any of the label or target dataset containing the clusters of output. The user is prompted to choose any of the listed algorithms with which to train the dataset. Choosing any of the algorithm buttons activates the corresponding algorithm, subsequently trains the dataset and produces the list of importance values sorted in descending order.

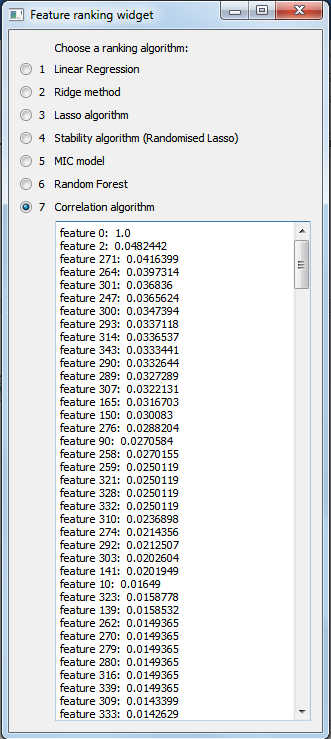


Figure 2. Screenshot of the widget executing correlation ranking algorithm

As shown in the figure above, the feature ranking values are shown to be sorted from the important for the selected algorithm, down to the list important, hence the most important features are easily spotted at the top of the list.

**Limitations and challenges**

The feature ranking GUI is currently takes a while to train the data when some algorithms are utilized. For instance, the stability algorithm doesn’t work well with our dataset as it produces 0’s across for all the features. The GUI currently doesn’t distinguish between a clustered dataset and a raw untampered dataset.

The code was developed in python 3.5 interpretation while the code of my team mates was implemented in python 2.7. So we are not able to easily combine both codes for this part of the project.

**Future plan**

The plan for the part 2 of the project is, among other things, to incorporate the clustering algorithms into the GUI. We intend to reconcile the differences in the python interpretations i.e to have everyone on the same page as regards to the python version to adopt for the project.

**References**