**Phishing Email Detection using various Machine Learning Techniques**

**Problem Statement**

Email Communication is one of the most common modes of communication. However,

phishing email one of the most serious Internet phenomenon. Phishing is a form of cybercrime where an attacker imitates a real person/institution by promoting them as an official person or entity through e-mail or other communication mediums. Hence, we need to find out different techniques to detect fraudulent emails.

**Objective**

Building different of ML techniques to identify phishing email and comparing the performance of these ML techniques with respect to different metrics.

**Dataset**

fraud\_email\_final.tsv is a dataset from Kaggle.com. It consists 2 columns. The first column consists of 11881 text emails. The second column tells us whether an email is legitimate(0) or phishing(1).

**Software**

Spyder with python 3.7

**Procedure**

For building a machine learning to predict whether an email is legitimate or phishing, following steps are performed.

1. Cleaning Text data

* Imported the dataset(tsv file) into Spyder(Python 3.7) using Pandas library
* Text cleaning and pre-processing:

1. Removed Punctuation and numbers from text by selecting only text format using re library
2. Converted text to lower case
3. The String of text is converted into list of words
4. Removed stop words using nltk library
5. Lemmatization – It is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.
6. Again joined the list of words with a space in between using function “join”
7. Created a bag of words model. It is nothing but we use the tokenized words for each observation and find out the frequency of each token. Generating a sparse matrix with max 1500 features
8. Dimensionality Reduction

Each observation is represented by thousands of tokens, which make the classification problem very hard for many classifiers. Dimensionality reduction is a typical step in text mining, which makes the dataset short and compact. The new space is easier to handle because of its size, and also to carry the most important part of the information needed to distinguish between emails, allowing for the creation of profiles that describe the data set. Two major classes of dimensionality reduction techniques are feature extraction and feature selection.

Feature extraction techniques – PCA and LDA are used in this project.

1. Prediction model: Built various machine learning models as shown in the following block diagram.

Naive Bayes

PCA

LDA

Naïve Bayes

Clean Text Dataset

Naïve Bayes

Logistic Regression

Random Forest

SVM

**Observation**

* 1. Area Under ROC for various ML algorithms without dimensionality reduction( Number of features = 1500)

* 1. Area Under ROC for a ML algorithm with dimesionality reduction methods - PCA and LDA

**Conclusion**

* Random Forest model outperforms other methods in terms of the area under the ROC curve(0.9911)
* Also, after applying feature extraction techniques like PCA and LDA, we are getting similar results(Area under ROC = 0.9511) with less number of features (900)