DETECTION OF PHISHING E-MAILS USING DATA MINING

Krina Ashwin Shah[1], Kanchan Manohar Kamble[2], Poonam Shatrughan Sawant[3]

*[4]**Prof. Parul Choudhary**, Dept. of Information Technology, Usha Mittal Institute of Technology college, Maharashtra, India.*

ABSTRACT –

In This paper, we have discussed about four different algorithms used for detection of phishing E-mails using Data Mining. Also, we have analysing the methodologies and implement the algorithms using Python language. We examine the absolute most well-known machine learning strategies (Decision Tree Classification, ADA Boost, Logistic Regression, Random Forest Algorithm) and of their relevance to the issue of a spam E-mails classification. E-mail filtering job relies upon data classification approach. Descriptions of the algorithms are introduced; alongside the differentiation appear on the Ling Spam corpus data set.

Keywords: Phishing, Data Mining, Link Corpus, Python programming.

1. INTRODUCTION

Phishing is a cybercrime attack wherein sends the social messages to a victim of stealing sensitive information such as: Bank account number, Credit card number etc. Now a day's emails are the most widely used communication purpose. As a result, phishing emails are increasing day by day because of fraud. The phishing problem is every difficult for number of users.

Most hard problem is to stem the fact that is easy for attacker to create the duplicate good sites which look very convince to users. Spam emails are the same messages send to a lot of users. Spam emails have unusual functions. A few of them provide for promotion issues, others are conscientious of dispersal computer viruses as well as there be present spam messages intended to appropriate the user economic identities.

As of the late an unplanned business/mass email or else called spam turn into a most important difficulty over the web. In late in sights, 40 of all messages are spam which around 15.4 billion email for every day and that cost web clients about 355 million every year. Programmed email filtering is by all accounts the best technique for countering spam right now and a tight rivalry amongst spammers and spam filter strategies is removal on. Just quite a long while back the vast majority of the spam could be dependably managed by blocking messages originating from specific locations or filtering out messages with certain subject lines.

2. PROBLEM STATEMENT

A survey is carried out in 2018 it revealed 80% of the cyber-attacks raised against the emails. Phishing attack deliver the malicious link or URL, which clicked by a user system with malware. Phishing emails provide an entry for an attacker to raise further attacks. Phishing attack target the weakest link in security, which is known as a Human factor. Credit card and many other companies become a huge part for an attacker. They cause many finical losses because of these attacks.

Nowadays, phishing attacks pose a real threat to secure email communication. The attacks are getting sophisticated and continuously adapting to existing defence mechanisms. Having said that, phishing email detection is a research area gaining much attention day by day. There is a need to develop an intelligent solution for detecting and combating phishing attacks. Efficient solutions that involve machine learning techniques should be used to enhance email security and make it less dependent on user awareness.

3. LITERATURE SURVEY

1. Detection of Phishing Emails:

- **Key concept:**

In this paper they studied three data mining algorithms used for automatic detection of phishing emails. After carefully analysing this methodology they implement it using R programming language and give comparison study of these algorithms by testing them on publicly available emails corpus.

- **Methodologies Used**:

There are mainly three steps: Data pre-processing, Data mining and Data Postprocessing. In Data pre-processing they accomplished two task File conversion and feature. In Data mining step they implement following algorithms:

1) Random Forest –

Random Forest is a classification algorithm based on the decision trees. In the training phase, it creates a set for a decision trees where in an individual tree operates on a randomly chosen set of attributes. The classification results are determined by majority-based voting from individual trees.

2) C5.0 –

C5.0 is the most advanced form of decision trees algorithms developed by Quinlan.It solves the classification problem by reducing the overall entropy of the dataset. In order to accomplish this, it selects the" best" attributes on the basis of Information Gain to create the decision tree.

3) Logistic Regression -

Logistic Regression is an algorithm is usually applied to a binary dependent feature. There are two possible dependent features usually labelled in binary values i.e., "0" and "1".

• **Results**:

1) C5.0 has the highest Precision (98.9%) 2) Random Forest Precision (97.7%), 3) Logistic Regression Precision (94.7%), 4) C5.0 has the highest Accuracy (99.4%), 5) Random Forest Accuracy (98.4%), 6) Logistic Regression Accuracy (97.5%).

2. A Comparison of Machine Learning Techniques for Phishing Detection.

This paper compares the predictive accuracy of several machine learning methods which includes a Logistic Regression, Classification and Regression Trees, Bayesian Additive Regression Trees, Support Vector Machines, Random Forest and Neural Networks for predicting phishing emails.

• **Key concept**:

This paper compares the predictive accuracy of several machine learning methods which includes a Logistic Regression, Classification and Regression Trees, Bayesian Additive Regression Trees, Support Vector Machines, Random Forest and Neural Networks for predicting phishing emails.

• **Methodologies Used**:

Most of the machine learning algorithms discussed in this paper are categorized as supervised machine learning. They construct the testing data set from the raw phishing emails. In addition, they describe the evaluation metrics use in the comparison. Following algorithms are used in implementation:

1) Logistic Regression (LR) 2) Classification and Regression Trees (CART) 3) Bayesian Additive Regression Trees (BART) 4) Support Vector Machines (SVM) 5) Random Forests (RF) 6) Neural Networks (NNet)

• **Results**:

1) Logistic Regression has highest Precision (95.11%) and Recall (82.96%) 2) CART (92.32%) 3) BART (92.08%) 4) SVM (94.15%) 5) RF (94.17%) 6) NNet (91.77%).

3. Phishing Detection using Classier Ensembles.

• **Key concept**:

This paper classifies emails into Phish or ham categories. In this paper C5.0 algorithm is implemented which achieve very high precision and an ensemble of other classifiers that achieve high recall. Approximately 8,000 emails were used for training model and half of which were phishing emails and the remainder legitimate, are presented. These results give importance of using this recall boosting technique.

• **Methodologies Used:**

For the purposes of experimentation, they used freely for available pre-classified datasets. The next step was to select the features used to represent emails. The individual techniques that were considered in order to construct a successful classier ensemble are: 1) C5.0 decision tree learning algorithm, 2) K-Nearest Neighbour algorithm, 3) Support Vector Machines, 4) Naive Bayes, 5) Linear Regression.

• **Results**:

1) C5.0 has the Accuracy (97.15%) and Precision (98.56%)., 2) K-Nearest neighbour Accuracy (87.21%), Precision (86.48%)., 3) Linear Regression Accuracy (83.03%), Precision (95.12%)., 4) SVKM Accuracy (97.11%), Precision (98.12%).

4. Phishing emails detection using improved RCNN Model with multilevel vectors and attention mechanism.

• **Key concept**:

In this paper, they first analysed the emails structure. Based on the improved recurrent convolutional neural networks (RCNN) model with multilevel vectors and the attention mechanism, they proposed a new phishing emails for and detection model named THEMIS. The THEMIS is used to model emails at the email header, email body, the character level and the level simultaneously. They used an unbalanced dataset that has realistic ratios of phishing and legitimate emails to evaluate the effectiveness of THEMIS.

• **Methodologies Used:**

In this paper, emails are divided into two categories, first is legitimate emails and phishing emails. Naturally, the detection for phishing emails is also a binary classification problem. They define a binary variable y to represent the attributes of an email; that is, y = 1 means that the email is a phishing email and y = 0 means that the email is legitimate. In other words, y is the label of an email. To determine whether the email is a phishing email following are the steps:

we first calculate the probability that the email is a phishing email, that is, P (y = 1). Then, the probability value is compared with the classification threshold, and if it is greater than the classification threshold, it is judged as a phishing email. The goal is to detect whether the target email is legitimate or phishing quickly and accurately.

1.RCNN:

RCNN is a new deep learning algorithm proposed by Lai et al. In 2015.In the RCNN model, S. Lai et al. use the bidirectional Recurrent Neural Network (BRNN) to capture the contexts. However, the RNN has a long-term dependency problem, which may further cause the gradient exploding and vanishing problems. In this paper, the RCNN model is used to learn the text features of email, which is a long text sequence.

2.THEMIS:

Based on the multilevel embedding and improved RCNN Attention model mentioned earlier, they formally put forward the THEMIS model, a phishing email detection model, by combining these two parts. They vectorize the email according to its text structure: header, body, characters, and words, using Word2Vec to output char-level embedding and word-level embedding. They combine the text structure of the email into the char-level of the email that under certain circumstances, the email header content and the email body content have varying degrees of impact on phishing email detection. Based on the situation as mentioned above, the\newline attention mechanism is used to obtain the weighted sum of the email header and the email body, which is the representation of the whole final email. CLASSIFICATION THRESHOL MOVING:

In the binary classification experiment, when we classify a sample x, we are actually comparing the predicted probability value y with the classification threshold value p. The value of the classification threshold p is equal to: Y=M/M+N.

• **Results**:

The experimental results show that the overall accuracy of THEMIS reaches 99.848%. Meanwhile, the false positive rate (FPR) is 0.043%.

High accuracy and low FPR ensure that the filter can identify phishing emails with high probability and filter out legitimate emails as little as possible. This promising result is superior to the existing detection methods and verifies the effectiveness of THEMIS in the detecting phishing emails.

5. Learning to Detect Phishing Emails.

• **Key concept**:

In this paper they present a PILFER method for detecting these attacks, which in its most general form is an application of machine learning on a feature set designed for highlight user targeted deception in electronic communication. This method is applicable, with slight modification, to detection of phishing websites, or the emails used to direct victims to these sites.\newline In his paper they discusses for previous approaches to the filtering phishing attacks, then the overview of machine learning and how we apply it to the task of classifying an phishing emails, and how it could be used in a browser tool bar. Then results of empirical evaluation, as well as some challenges presented their in.

• **Methodologies Used**:

1.PILFER:

PILFER, is a machine-learning based approach to the classification. In a general sense, whether it is designed to trick the user into believing they are communicating with a trusted source, when in reality the communication is from an attacker. Some spam filters use hundreds of features to detect unwanted emails. In this they tested a number of for different features, and present in this paper a list of the ten features that are used in PILFER, which are either binary or continuous numeric features. As the nature of phishing attacks changes, additional features may become more powerful, and PILFER can easily be adapted by providing such new features to the classifier. It is important to note that mis classifying a phishing email may have a different impact than misclassifying a good email, so we report separately the rate of false positives and false negatives. The false positive rate corresponds to the proportion of ham emails classified as phishing emails, and false negative rate corresponds to the proportion of phishing emails classified as ham.

• Results:

PILFER achieves an overall accuracy of 99.5 with a false positive rate fp of approximately 0.13. PILFER's false negative rate of on the dataset is approximately 0.035, which is almost one fourth the false negative rate of the spam filter by itself.

4. PROPOSED SYSTEM

Machine taking in field is a subfield from the expansive field of artificial intelligence, these plans to make machines ready to learn like human. Learning here means comprehended, observe and converse to data about some statistical phenomenon. In unsupervised learning one tries to reveal shrouded regularities (bunches) or to identify anomalies in the information like spam messages or system interruption. In email filtering errand a few features could be the pack of words or the subject line analysis.

Thus, the contribution to email classification assignment can be seen as a two-dimensional matrix, whose axes are the messages along with the features. Email classification assignments are frequently separated into a few sub-undertakings. To start with, Data accumulation and representation are for the most part issue particular (i.e. email messages), second, email feature choice and feature diminishment endeavour to decrease the dimensionality (i.e. the quantity of features) for the rest of the means of the task. At long last, the email classification period of the procedure finds the genuine mapping between training.

In this work we have indicated how classification calculations take a shot at informational index. We had taken is ling spam corpus which is very huge informational collection and it comprises of different sends and these sends are classified into prepare emails and test emails are explain through in given figure 1. In this segment we will first talk about how ling spam corpus workings within steps.

Here we will compare the classification calculation based on disarray network and accuracy. These classification calculations have been applied on the dataset ling-spam which for the most part comprises of huge number of sends for training and for testing purpose. At the same time, we in introduced one more approach that combine the classification calculation whose accuracy could conceivably be more than the previous one, it depends on the dataset and what type of value it contains. Initial step is to organize the data. In this process we have part the downloaded information into training set and test set.

Dictionary will be shaped for every word It can be seen that the primary line of the mail is subject and the third line contains the body of the email. We will just perform text investigation on the content to detect the spam sends. As an initial step, we need to create a lexicon of words and their frequency. For this job, training set of 700 send is exploit. This python work generates the lexicon designed for you.

Feature extraction process Once the lexicon is arranged; we can obtain word count up vector (our feature now) of 3000 dimensions intended for every email of training set. All word check vector holds the frequency of 3000 words within the training file. Of course, you might have guessed at this position a huge portion of them resolve be zero. Let us get an instance, assume we have 500 words within our lexicon. All word check vector encloses the frequency of 500 lexicon words within the training file. Assume text inside training case was "Get the work done, work done" then it will be prearranged as [0,0,0,0,0,… … .0,0,2,0,0,0,… … ,0,0,1,0,0,… 0,0,1,0,0,… … 2,0,0,0,0,0]. Here, everyone the word tallies are located at 296th, 359th, 415th, 495th catalogic of 500 length word tally vector in addition to the rest are zero.

5. IMPLEMENTATION

Our approach is in line with the Knowledge Discovery in Databases (KDD) process. The flow chart below illustrates the key steps in the implementation.

• Data Pre-processing: -

In this work we have indicated how classification calculations take a shot at informational index. We had taken is ling spam corpus which is very huge informational collection and it comprises of different sends and these sends are classified into prepare emails and test emails are explain.

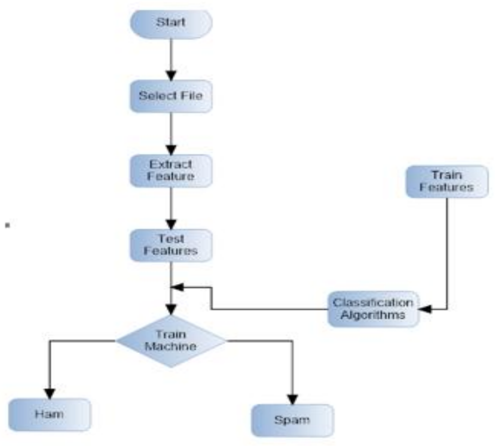


Fig: 1 Proposed Method for Classification

In this segment we will talk about how ling spam corpus workings with its steps. Here we will compare the classification calculation based on disarray network and its accuracy.

These classification calculations applied on the data set ling spam which for the most part comprises of huge number of sends for training and for testing purpose. At the same time, we in introduced one more approach that combine the classification calculation whose accuracy could conceivably be more than the previous one, it depends on the dataset and what type of value it contains. The steps are involved in this process are as follows:

1.Initial step is to organize the data.

2.Dictionary will be shaped for every word.

3.Feature extraction.

4.Training the classifier.

• **Generate Training and Testing Dataset**:

In this project we have dataset which is divided into Train dataset and Test dataset. Initial step is to organize the data. In this process we have part the downloaded information into training set and test set. Here we have taken ling corpus informational collection which for the most part contains 702 training emails and 260 test sends means we have aggregate of around 962 sends. Dictionary will be shaped for every word It can be seen that the primary line of the mail is subject and the third line contains the body of the email.

We will just perform text investigation on the content to\newline detect the spam sends. As an initial step, we need to create a lexicon of words and their frequency. For this job, training set of 700sendisexploit.This python work generates the lexicon designed for you.

• **Algorithms Used: -**

1. Random Forest:

A classification algorithm based on decision trees. In the training phase, it creates a set of decision trees where in an individual tree operates on a randomly chosen set of an attribute. The classification results are determined by the majority-based voting from individual trees.

2. Logistic Regression:

Logistic Regression is an algorithm usually applied to a binary dependent feature. The logistic regression model has a dependent variable with the most two possible values, such as pass/fail which is usually labelled in binary values i.e.”0” and ”1”.

3. Decision tree Classifier:

Decision tree builds classification or regression models in the form of tree structure. It breaks down a data set into smaller smaller subsets. The final result is a tree with a decision nodes and leaf nodes.

4. ADA Boost:

The Ada boosting algorithm creates a strong classifier from a number of weak classifiers. It builds a model from the training data, then it creating a second model that attempts to correct the errors from the first model.

• **Building classifier models:**

Training the classifiers, we have trained 4 models here namely Logistic Regression, Decision Tree classier, Random forest and ADA-Boost.

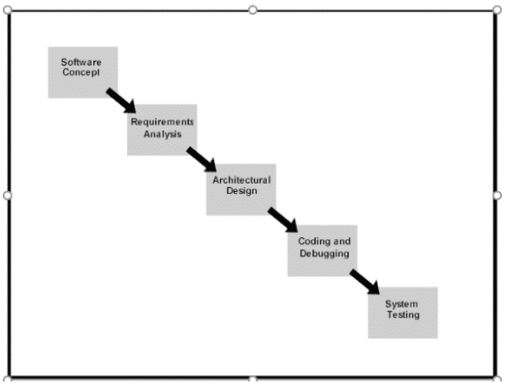


Fig: 2. System Development Life Cycle

• **Testing classifier models:**

We predicted class labels for emails in the test dataset using corresponding classifier model created in each iteration. We have use Random forest algorithm for detection of phishing emails.

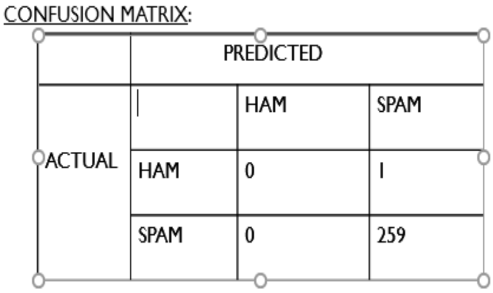
• **Capture test result:**

We captured the confusion matrix for each iteration which is an input for the next phase.

Formulas:

1. FP Rate = $\frac{n\rightarrow P}{nH}$

2. FN Rate = $\frac{nP\rightarrow H}{H}$



CONFUSION MATRIX:

		PREDICTED	
ACTUAL		HAM	SPAM
	HAM	0	1
	SPAM	0	259

Fig: 3. Confusion Matrix.

3. Precision = $\frac{TP}{TP+FP}$

4. Recall = $\frac{TP}{TP+FN}$

5. Accuracy = $\frac{TP+TN}{TP+FN+TP+FP}$

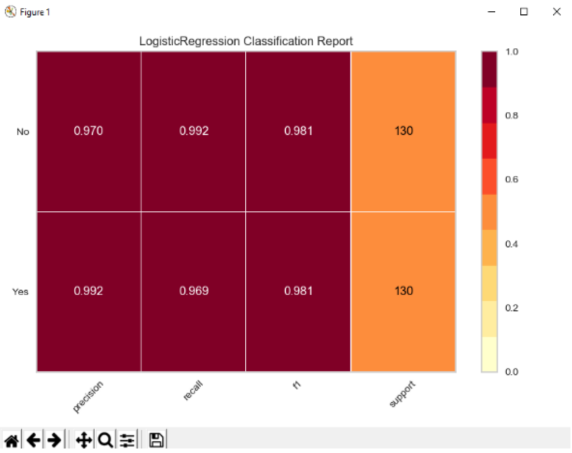


Fig: 4. Random Forest Classification Report.

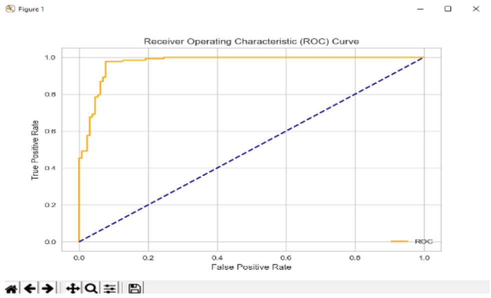


Fig: 5 Random Forest ROC curve.

- **Accuracy of each algorithm**:

– Random Forest = 97.30%

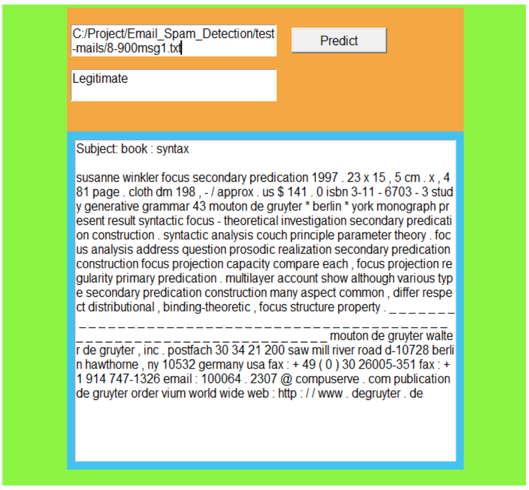
– Logistic Regression = 98.07%

– Decision Tree = 97.25%

– ADA Boost = 95%

- **Graphical User Interface**:

We provide facility to user to check whether the email is spam or ham by using GUI.



C:/Project/Email_Spam_Detection/test-mails/8-900msg1.txt

Legitimate

```
Subject: book : syntax

susanne winkler focus secondary predication 1997 . 23 x 15 , 5 cm . x , 4
81 page . cloth dm 198 , - / approx . us $ 141 . 0 isbn 3-11 - 6703 - 3 stud
y generative grammar 43 mouton de gruyter * berlin * york monograph pr
esent result syntactic focus - theoretical investigation secondary predicati
on construction . syntactic analysis couch principle parameter theory . foc
us analysis address question prosodic realization secondary predication
construction focus projection capacity compare each , focus projection re
gularity primary predication . multilayer account show although various typ
e secondary predication construction many aspect common , differ respe
ct distributional , binding-theoretic , focus structure property . _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ mouton de gruyter walte
r de gruyter , inc . postfach 30 34 21 200 saw mill river road d-10728 berli
n hawthorne , ny 10532 germany usa fax : + 49 ( 0 ) 30 26005-351 fax : +
1 914 747-1326 email : 100064 . 2307 @ compuserve . com publication
de gruyter order vium world wide web : http : / / www . degruyter . de
```

Fig: 6. Graphical User Interface

6. CONCLUSION

This paper examined the task of Detecting Phishing emails using four different classification algorithms viz. Radom Forest, Decision Tree and Logistic Regression, ADA Boost and provide GUI for detection of phishing emails. Descriptions of the calculations are presented, and the correlation of their performance on the Ling corpus Spam Dataset is presented, the experiment demonstrating a very encouraging results specially in the calculations that isn't well known in the commercial e-mail filtering packages, spam recall percentage in the five methods has the accuracy values, while in term of accuracy we can find that the Decision Tree and Logistic Regression methods has a very fulfilling performance amongst the other technique, more research should be done to rise the performance of the Random Forest either through hybrid system or else by decide the feature dependence issue

7. FUTURE SCOPE

The future efforts would be extended towards an Achieving accurate classification, with zero percent (0%) of a misclassification of Ham E-mail as Spam and Spam Email as Ham. The efforts would be applied to square Phishing Sends, which carries the phishing attacks and now-days which is more matter of concern. Use of hybrid of ensemble algorithms for an email spam detection.

8. REFERENCES

[1] Issam dagher, Rima Antoun," Ham- Spam Filtering Using DIFFERENT PCA SCENARIOS", 2016 IEEE International Conference on Computational Science and

Engineering, IEEE International Conference on Embedded and Ubiquitous Computing, and International Symposium on Distributed Computing and Applications to Business, Engineering and Science

[2] Scholkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)

[3] Ali, S., Smith-Miles, K.A.: A meta-learning approach to automatic kernel selection for support vector machines. Neurocomputing 20(1-3), 173–186 (2006).

[4] S. Whittaker, V. Bellotti and P. Moody, "Introduction to this special issue on revisiting and reinventing e-mail", Human-Computer Interaction, 20(1), 1-9, 2005.

[5] M. N. Marsono, M. W. El-Kharashi, and F. Gebali,

"Binary LNS-based naive Bayes inference engine for

spam control: Noise analysis and FPGA synthesis",

IET Computers Digital Techniques, 2008

BIOGRAPHIES:

I am Krina Shah, Currently Pursuing in Information Technology from Usha Mittal Institute of Technology, SNDT, I have done project on Credit Card Fraud Detection, Apprize and Detection of Phishing E-mails. In this paper my Contribution is making Decision Tree & Ada Boost Algorithm.

I am Poonam Sawant, Currently Pursuing in Information Technology from Usha Mittal Institute of Technology, SNDT, I have done project on Credit Card Fraud Detection, Shopping and Detection of Phishing E-mails. In this paper my Contribution is making GUI, Heat map and Roc Curve.

I am Kanchan Kamble, Currently Pursuing in Information Technology from Usha Mittal Institute of Technology, SNDT, I have done project on Credit Card Fraud Detection, Billing Software and Detection of Phishing E-mails. In this paper my Contribution is making Random Forest & Logistic.

Ms. Parul Choudhary Assistant Professor in Usha Mittal Institute of Technology, SNDT, Mumbai, I have 10years of experience in Teaching. Published more than 12 research paper on National and International Journals. Also had attended workshops on Networking/ programming.