A Machine Learning Framework for Studying User Behaviors in Phishing Email Processing

Yi Li[1], Kaiqi Xiong[1], and Xiangyang Li[2]

[1] University of South Florida, Tampa FL 33620, USA
yli13@mail.usf.edu, xiongk@usf.edu
[2] Johns Hopkins University, Baltimore MD 21218, USA
xyli@jhu.edu

Abstract. Phishing has posed serious problems in security and it is important to understand user behaviors in phishing email processing. In this paper, we propose a machine learning framework to predict the performance of users when phishing attacks occur. Specifically, we first focus on studying how users will behave when they read emails including phishing emails. Then, we design an email sorting task to mimic the email reading action in our daily life and recruit participants through Amazon Mechanical Turk for evaluating the proposed framework.

Keywords: User behaviors, phishing email, machine learning, MTurk

1 Introduction

Phishing attacks are usually considered as an online identity theft to deceive users to provide their personal informations, such as login credentials [5]. Phishing can be disguised in emails, website links, or other forms of messages. Many researchers have studied how to detect and prevent phishing attacks in different ways [1], however, there are only a very few studies on understanding the user behavior related with phishing attacks. Users are susceptible to phishing attacks at different degrees due to their background of network security [4].

Dhamija et al. [3] analyzed some hypotheses about the reasons of phishing attack feasibility and assessed those hypotheses by showing 20 web sites to 22 participants and asked them to determine which ones were deceptive. Their results showed that 23% of the participants were not aware of security indicators, leading to incorrect choices 40% of the time. As phishing becomes a more and more popular attack vector, email has been the most common way to conduct phishing attacks [8]. Some machine learning techniques have been applied to detect phishing emails [9]. Supriya et al. [7] has studied user behaviors in phishing emails with incentive and intervention. They designed a three-round experiment to let users distinguish the phishing emails from the normal emails.

In this research, we aim at studying how users behave when encounter with phishing attacks based on their personal profile and behavior. Specifically, in our experiments, we recruit participants to conduct email sorting tasks. The emails used in the research consist of both phishing emails and normal emails.

Performance of each participant, such as sorting correctness and time, is recorded in each experiment. To understand the collected data, we propose and develop a machine learning framework to predict the performance score of each participant based on his/her profile. The proposed machine learning framework consists of four different models that are developed with a 10-fold cross-validation and cross-validation based feature selection. We also perform attribute reduction by analyzing the data obtained from participants' performance as well as the participants' basic information from the survey to select the best attributes for our machine learning framework.

2 Study Design

Nowadays, emails have been widely used throughout the world via the Internet. Many people, especially employees in a work environment and students at colleges, read and respond emails daily. Emails become an integral part in daily life for most of people. Thus, it is very likely that many people might have experience to wrongly click on a request link seemingly to be legitimate, but actually a phishing link. To understand user behaviors related to phishing attacks thoroughly, we present a study design to mimic the email opening, reading, and decision task like what we usually do in our daily life. The participants will perform an email sorting task to decide if an email is phishing or not by moving the preloaded emails to either a "phishing" or "non-phishing" folder. The emails used in the study are obtained from the real world with some necessary modification. We derive some phishing emails from the "Phish Bowl" database [2].

2.1 Phishing Types

1. Suspicious sender's email address: The scammers utilize the phenomena that people usually pay less attention to a sender's email address. For example, the scammers can use the number '0' to replace the letter 'o' because they are very similar. Therefore, the scammers can fake the the domain name 'wellsfarg0' rather than 'wellsfargo'.

2. Suspicious links or attachments: The suspicious link is very similar to a suspicious sender's email address. The scammers will try to deceive a user by using similar characters or misspelling words in the links. The suspicious attachments are usually disguised as an exe file, a pdf file, or other types of files.

3. Malicious Email Contents: It is a trick type of phishing. Everything seems legitimate, but when we examine the content of the email carefully, we will find that the email has some issues, such as grammar issues or enclosed with the faked icon of popular social networks. This type of attack is very hard to identify if people are not familiar with the icons or weak in grammar.

2.2 Study Design

To sufficiently understand user behaviors regarding phishing attacks, we proposed a study design to collect data from participants and proposed a machine learning framework to predict the performance based on user behaviors.

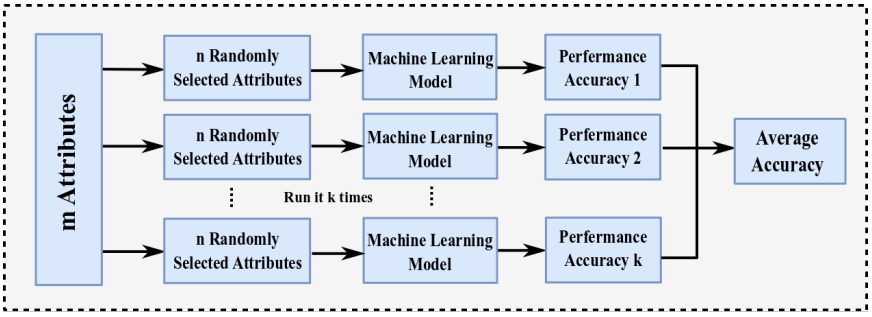

Fig. 1. The proposed Machine Learning Framework

Environmental Setup We utilize a RoundCube email client as an interface for users to preview emails and perform the email sorting task. The RoundCube email client is a browser-based IMAP client. We use a JavaScript-based Data Capture to collect a user's input and a AJAX-based Data Sender to communicate the captured data to the server. In the RoundCube email client interface, we add a rating module for the participants to rate their confident level of sorting each email from 1 to 10.

Participant Recruitment The IRB had been approved before we started to recruit participants (the approval number is: Pro00026240). To have more demographic diversity, we recruited participants through Amazon Mechanical Turk (MTurk) [6]. 90 Participants are recruited with different backgrounds for this study. The ages of the participants range from 20 to 61 and the average age is about 34 years old. Among 90 participants, 35 are female and 55 are male.

We introduce a monetary incentive mechanism in our study design. We divide the participants into two groups, a monetary incentive group and a control group (non-incentive group). Each participants in the non-incentive group will get $4 payments regardless of his/her performance. The participants in the incentive group will have a chance to earn more payments (up to $8) if their performance of sorting emails is higher than 75% accuracy.

Experimental Task and Performance Score In this study, we preload 40 emails in the RoundCube email client and participants are asked to sort these 40 emails into either a "phishing" or "non-phishing" folder. Among those 40 emails, 20 emails are legitimate and 20 are phishing. Participants are not aware of this distribution when they perform their tasks. The participants have 30 minutes to finish this sorting task and rate their confidence level for each email. The performance score is calculated based on the correctness of moving the emails into correct folders. For example, the participant will get 1 point if he/she moves the email to the right folder, otherwise, the participant will get 0 point on that email. Therefore, the maximal performance score a user can get is 40 points.

Survey We have designed two types of surveys, pre-survey and post-survey. We use the pre-survey to investigate the basic information and background of participants, such as age, gender, education background, cybersecurity background,

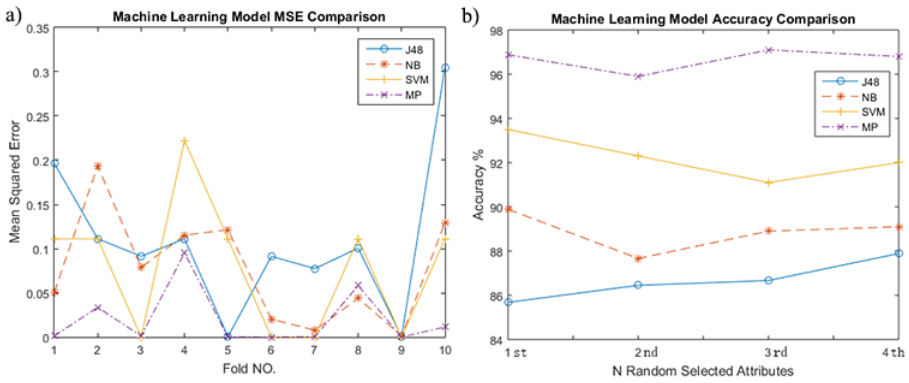

Fig. 2. Evaluation of the Proposed Machine Learning Framework

and habits of using social media. In this study, we include the Informed Consent Form and email sorting instructions in the pre-survey. The post-survey asked questions related to the email sorting task they perform.

2.3 Machine Learning Framework

To determine if a user performs well or poorly when encountering phishing attacks, we propose a machine learning framework to predict a user's performance. We divide the performance of a user into two classes, *Good* and *Poor* based on the average performance score of the participants. We extracted 119 features from the data we collected from 90 participants. After feature reduction, we choose 16 features to be used in our machine learning framework.

We built four different machine learning models, Decision Tree-J48, Naive Bayes (NB), Support Vector Machine (SVM), and Multilayer Perceptron (MP). We also proposed to use the method of 10 fold cross-validation to precisely predict the performance. We applied the same idea of cross-validation on the features we have selected. As shown in Figure 1 we randomly choose n attributes to do the cross-validation training by applying our machine learning model. The next step is to calculate performance accuracy. This process can be running k times. These k performance accuracies are averaged to form one final accuracy. The final performance accuracy is calculated by averaging all the accuracies.

3 Evaluation

We first evaluated the Mean Squared Error (MSE) of each fold for four different machine learning models, as shown in Figure 2 (a). Among them, J48 of fold 10 has the highest MSE because accuracy of J48 with fold 10 is the lowest. Then, we evaluated the performance accuracies as presented in our machine learning framework in Figure 1. As we described in section 2.3, we also applied the similar idea of cross-validation to attributes. Figure 2 (b) shows the accuracy result of random selected attributes. In our study, we chose $N = 4$, so we have each time, where there are 4 attributes used for testing and rest are used for training, and

this process is done for 4 times, as shown in the x-axis, 1st, 2nd, 3rd, and 4th. The accuracy of each time is the performance accuracy after we do the 10-fold cross-validation and the final accuracy is the average of the four performance accuracies. The accuracies change because the different attributes are chosen each time. We can see from the figure that the final accuracy for J48, NB, SVM and MP are 86.67%, 88.89%, 92.22%, and 96.67%, respectively.

4 Conclusions and Future Work

We have introduced a study design and a machine learning framework to understand how well users behave in phishing emails. In the machine learning framework, we have tested four different models and applied 10-fold cross-validation with randomly selected feature cross-validation to analyze the data collected from both survey and from the experiments the participants did. The performance score in sorting phishing emails is predicted based on the user's background information. We also achieved the user performance prediction accuracies of 86.67%, 88.89%, 92.22%, and 96.67% for J48, NB, SVM, and MP, respectively.

In the future, we plan to conduct more experiments, recruit more a large number of participants to perform the experiments, and collect more data for evaluating our proposed machine learning framework. We will further introduce intervention in the study design and carefully analyze how some of the factors, such as phishing types, intervention, and incentive, will affect user behavior.

5 Acknowledgement

We would like to thank NSF for partially sponsoring the work under grants #1620868, #1620871, #1620862, and #1651280. We also thank the JHU team that provided the data used in this project.

References

1. Chin, T.J., Xiong, K., Hu, C.: Phishlimiter: A phishing detection and mitigation approach using software-defined networking. In: IEEE Access (2018)
2. Database, P.B.: [online]. Available: https://it.cornell.edu/phish-bowl. (Accessed Sept 2018)
3. Dhamija, R., Tygar, J.D., Hearst, M.: Why phishing works. In: Proceedings of the SIGCHI conference on Human Factors in computing systems. ACM (2006)
4. Goel, S., Williams, K., Dincelli, E.: Got phished? Internet security and human vulnerability. Journal of the Association for Information Systems **18**(1), 22 (2017)
5. Gupta, S., Singhal, A., Kapoor, A.: A literature survey on social engineering attacks: Phishing attack. In: ICCCA. pp. 537–540. IEEE (2016)
6. MTurk, A.M.T.W.: [Online]. Available: https://www.mturk.com/mturk/welcome (Accessed Sept, 2018)
7. Muthal, S., Li, S., Huang, Y., Li, X., Dahbura, A., Bos, N., Molinaro, K.: A phishing study of user behavior with incentive and informed intervention. In: Proceedings of the National Cyber Summit (2017)
8. Pande, D.N., Voditel, P.S.: Spear phishing: Diagnosing attack paradigm. In: WiSPNET. pp. 2720–2724. IEEE (2017)
9. Smadi, S., Aslam, N., Zhang, L., Alasem, R., Hossain, M.: Detection of phishing emails using data mining algorithms. In: SKIMA. pp. 1–8. IEEE (2015)