sustainability

MDPI

Article

An Exploratory Study on Sustaining Cyber Security Protection through SETA Implementation

Guangxu Wang, Daniel Tse *, Yuanshuo Cui and Hantao Jiang

College of Business, City University of Hong Kong, Hong Kong 999077, China; guawang9-c@my.cityu.edu.hk (G.W.); yscui4-c@my.cityu.edu.hk (Y.C.); hantjiang2-c@my.cityu.edu.hk (H.J.)
* Correspondence: iswktse@cityu.edu.hk

Abstract: It is undeniable that most business organizations rely on the Internet to conduct their highly competitive businesses nowadays. Cyber security is one of the important elements for companies to guarantee the normal operation of their business activities. However, there is no panacea in cyber security protection. Common security practices used are to deploy hardware and software security protection tools to combat the known security threats which may become more and more powerful later. In fact, the attackers and security practitioners are at war from time to time. As a result, such a tools-based security protection strategy cannot be sustained. On the other hand, the related awareness training for employees is ignored in a number of companies, which has made biased the decisions made by staff when facing cyber security breaches. In this study, in order to find ways to sustain such protection, we conduct a quantitative analysis to explore the key elements contributing to the SETA implementation of the companies and organizations. We evaluate the performances of eight supervised learning models in a dataset collected from cyber security breach surveys on UK businesses to perform a fundamental analysis. The detailed analysis is performed via the feature importance of features generated in the model with better performance in the task of detecting the companies and organizations with SETA implementation. The experiment result shows that the awareness related factors play the most significant role in the SETA implementation decision-making for the businesses, and most of the businesses are lacking the awareness to prevent the potential cyber security risks in the stuff using externally-hosted web services and products as well as services depending on online services.

Keywords: cyber security; SETA; sustainability; supervised learning; awareness related factors; e-Commerce

check for updates

Citation: Wang, G.; Tse, D.; Cui, Y.; Jiang, H. An Exploratory Study on Sustaining Cyber Security Protection through SETA Implementation. *Sustainability* **2022**, *14*, 8319. https://doi.org/10.3390/su14148319

Academic Editor: Zubair Baig

Received: 14 June 2022
Accepted: 4 July 2022
Published: 7 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cyber security protection has become a constantly high-profile topic for businesses since e-commerce became an essential part worldwide. It is also the foundation of strategy making and investment making for a company [1]. Besides, computer viruses and hacker technology are multiplying as the importance of data increases, which directly impacts a company's core competencies and reputation. Within this background, some media exaggerate the seriousness of reports related to cyber security problems, and consumers also react negatively to data breaches [2]. More senior managers in various industries are concerned about maintaining cyber security to ensure the regular operation of the business and to not just regard cyber security as an IT problem. In fact, there is no panacea in cyber security protection, the best way we can do is to mitigate the impact caused through proper security management in order to sustain such protection in our business operations.

Although it is known for most companies to protect cyber security, many do not take enough measures to prevent cyber security threats [3]. Even under the urging of government and policy, the proportion of companies conducting staff training is not high. Some companies may mistakenly believe that staff with IT skills can prevent cyber security

threats, so there is no need to conduct training on cyber security for staff. Although cyber security protection and IT skills have certain relations, there is still a huge difference between them. All the staff lacking the understanding and awareness of cyber security is likely to lead to internal vulnerabilities and losses. Actually, cyber security's social engineering is one of the biggest threats, and staff awareness plays the most important role in controlling this threat [4]. To prevent this kind of attack, one of the best methods is to conduct education and training for the workforce of companies and further boost their related knowledge and awareness [4].

Security education, training, and awareness (SETA) training cover many aspects, which can comprehensively improve staff's ability of cyber security. It involves common sense, culture, and awareness training, as well as staff training to normally operate organizations' websites, systems, accounts, email, and social media. Besides, through SETA implementation, staff can learn how to prevent and deal with cyber security vulnerabilities and threats and consciously protect the organization's data and confidential information [5].

Some research studies have proven that SETA programs are the most effective way to improve employee information security protection behavior [6,7]. In contrast, there are also some studies showing that without the intervention of other factors, SETA programs alone do not have a significant effect on the improvement of employee safety behaviors [8,9] and that the failure of employees to take timely protective measures is not mainly due to the lack of safety training and safety awareness [10]. Research pointed out that SETA could play the most significant role in safety behavior through monitoring, followed by employee relations, and then accountability [11]. Thus, while some studies have shown that SETA implementation alone has no substantial impact on corporate cyber security protection, SETA implementation can influence employee behavior in other ways, such as monitoring, and further indirectly protecting the company's cyber security. It has a positive effect on the protection of corporate cyber security.

Due to the above current situation of people's attitude towards cyber security, we found the need to dig out the factors affecting the companies' focus on cyber security and explain how such factors affect the companies' decision on effective SETA implementation. Within the factors, the governments are efficient in taking additional measures to carry out the corresponding publicity related to the importance of SETA implementation, boosting the information security sustainability of businesses.

In this paper, we have five sections, which respectively demonstrate the motivation for this study; current research related to SETA implementation; the experimental process and outputs to explore the causes leading businesses to conduct SETA implementation; some discussion of past research, and our findings; and the conclusions of this study.

2. Literature Review

The existing papers related to SETA implementation can be mainly divided into three perspectives: the difficulties in SETA implementation, the human behavior toward SETA implementation, and the innovative design of the SETA implementation system.

2.1. Difficulties in SETA Implementation

SETA implementation is a common and necessary measure to maintain information security. Due to its low cost, most training courses are designed to improve employees' security awareness and reduce human errors. Aoyama et al. summarized that most teams have experienced similar management challenges, and the management challenges observed in incident handling training are possible challenges in real-world cyber incident management [12]. The demand for SETA implementation in industries is increasing, and some authoritative organizations are providing training in cyber incident handling. Traditional network security training mainly raises employees' awareness of vigilance to respond to network security attacks. These training items mainly include on-site training and awareness training through screen savers, posters, program reminders, and online courses [13]. Ghafir et al. pointed out that the shortage of corporate training budgets has

adversely affected employees' awareness training [14]. The company managers generally tend to minimize their budgets. This also provides opportunities for hackers who carry out cyber-attacks. If network security implementation is not sufficient, physical access threats, including information leakage and theft of items, may cause significant economic losses for companies and individuals. Furthermore, we also learned that due to the difference of employees in educational backgrounds and cultural levels, etc., giving on-site training and awareness training to all employees from top to bottom is also a huge problem faced by companies in SETA implementation [15,16].

2.2. Human Behavior toward SETA Implementation

Some scholars take human behavior into consideration and describe the changes in human behavior before and after conducting SETA Implementation. McCrohan, et al. claimed that when users have accepted proper cyber security training, they change to enhance security and tend to be more sensible to cyber security issues [17]. Puhakainen and Siponen found that some employees are not in compliance with security training, which causes security problems, so leaders need to motivate employees to cogitative processing the information they trained before [18]. Furman, et al. conducted an interview with users and finds that users are aware of cyber security, but they do not have enough skills to prevent cyber-attacks, so it is very important for users to accept training and obtain relevant skills [19]. In these articles, through investigating human behavior related to cyber security, the scholars find ways to enhance people's ability and training effectiveness when facing a cyber security attack.

2.3. Innovative Design of SETA Implementation System

Some scholars focus on the innovative design of SETA implementation systems to boost the efficiency of existing training systems. Two general directions are to design new types of SETA implementation systems or take some measures to boost the efficiency of current training systems. Cone et al. concluded that though cyber security training approaches are very universal, most of them are lacking security concepts, there is a new training tool called CyberCIEGE that can successfully and efficiently raise users' security awareness [20]. Abbott et al. found that better structuring of the education and training of cyber security is very significant, so using the technology mining the resulting data logs for relevant human performance variables can improve the quality of the cyber security process [21]. Hatzivasilis et al. used pedagogical practices and a cyber-security model to design a dynamic training program that can provide contentious adaption to users' performance [22]. In fact, SETA implementation is becoming more and more important due to the rapidly developed cyber threat. In conclusion, these articles summarize that the current SETA implementation system is not suitable for schools and companies' needs, there should be a new training system that is more effective than the current one.

Overall, the existing studies on SETA implementation are mainly focused on the difficulties of companies to hold this kind of training, the effects of training on employees, and how to build a more efficient training system for companies. However, there is not much work on the factors affecting the companies' focus on cyber security and explaining how such factors affect the companies' decision on effective SETA implementation, thus we focus on the elements leading to SETA implementation of companies and try to dig out the real cares of companies in cyber security, which can help the government better understand the companies' demand and further publish related policies. It is helpful for companies to reduce the risk of financial issues and privacy issues.

3. Methodology

The objective of this study is to investigate the elements that contribute to a company's SETA implementation. To achieve that, we regard the problem of whether the companies and organizations are willing to conduct SETA implementation as a binary classification task, and the supervised learning method is used to conduct the experiment. Supervised

learning is a common method used in the classification task [23]. In this case, we initially generate representative features from a cyber security survey result collected by the UK government and Ipsos MORI Social Research Institute and evaluate the performances of various supervised learning models in detecting the companies that have held SETA implementation to perform preliminary analysis. Following that, the detailed analysis is conducted by generating feature importance of features. The whole process of methodology is mainly separated into three subsections: Dataset and Feature Generation, Experiment Design, and Performance Comparison.

3.1. Dataset and Feature Generation

In this research, we mainly emphasize the industries in the UK. The dataset used in this study is collected from the survey designed by Ipsos MORI Social Research Institute. Ipsos MORI Social Research Institute associated with the UK government conducted a telephone survey of 1008 UK businesses and 30 interviews during the year 2016 to figure out the cyber security issues and actions that needed to be processed in the UK industry [3]. As the UK's economy has become stronger, there is an increasing number of business operations in the UK. In order to make the UK become one of the most suitable places to run business, Ipsos MORI Social Research Institute and the UK government regularly take this kind of survey. Based on the task of exploring the causes leading companies to conduct SETA training, we extracted and generated various features from the survey result. Initially, several attributes hypothesized to contribute to the companies' SETA implementation decision-making are extracted for the feature generation. After the preparation, we perform several actions, such as reassignment and mergence, to generate the features used to conduct the following analysis. The details of generating rules and generated features are shown as follows:

"Update": The "Update" describes the frequency of antivirals software updates for a company, which is the answer to the question "how often does the company update the antivirals software?" in the survey. More frequent software updates indicate greater cybersecurity awareness. From this perspective, this feature reflects the awareness of corporate leadership in preventing cyber security threats. For this feature, we converted it into the ordinal type.

"Sizec": The "Sizec" is one of the features describing the sizes of the company in the dataset. Different sizes of companies may have different considerations and attitudes toward cyber security issues and SETA implementation. The "Sizec" transformed the companies with varying numbers of employees into different scales of companies. The companies with 2–9 employees, 10–49 employees, 50–249 employees, and equal to or more than 250 employees are respectively regarded as micro, small, medium, and large companies. Based on the characteristics of the "Sizec", it is also converted into the ordinal type.

"Freq": The "Freq" describes the frequency of attacks and breaches that the company encountered during the past year of the survey. Different companies have encountered quite different cyber-attack situations during the past year of the survey. For example, some companies encountered cyber-attacks almost once a day or even several times a day, while some other companies did not encounter any attack in a year, which might be a key element affecting the companies' decision-making on SETA implementation. For the "Freq", we converted it into the ordinal type.

"Priority": The "Priority" describes the level of importance placed on cybersecurity by organizations' directors or senior managers. As the managers consider more cyber security protections, more measures and strategies might be taken to prevent cyber breach loss. This feature is a direct expression of the company management's awareness regarding cyber security, and it is converted into the ordinal type.

"Numbb": The "Numbb" is one of the features describing the total number of attacks and breaches that the company encountered during the past year of the survey in the original dataset. Different companies also encountered a different number of cyber-attacks and breaches in a fixed period. Similar to the "Freq", the "Numbb" also describes the companies' exposure to cyber-attacks. Compared to the "Freq" describing the attack

frequency, the "Numbb" describes the order of magnitudes of cyber-attacks. The "Numbb" is also converted into the ordinal type.

"Core": Th "Core" describes the dependence of products and services provided by an organization on online services. With the rapid advancement of information technology, online services have increasingly become the hard core of products and services in many companies and organizations. One of the goals of cyber-attacks for hackers is financial gain [24], so the high dependence of products and services on online services might lead the managers to care more about cyber security and further conduct SETA training. Based on the dependency level on online services, we convert the "Core" into the ordinal type.

"Insure": The "Insure" represents whether an organization or company has insurance against a cyber security breach or attack. In commercial activities, there is a number of business losses directly related to cyber security accidents. In the case of security accidents, cyber security insurance is an efficient way for organizations to reduce their capital loss, so the organizations with cyber security insurance might ignore many other cyber security protection channels and remedial measures, including SETA training. For the "Insure", we converted it into the nominal type according to the insurance status.

"Factor": The "Factor" describes whether the occurrence of cyber security breaches or attacks is related to staff-related factors. There are many attributes describing the specific reason contributing to the cyber security breaches or attacks in the original dataset. In these types of causes, both external invasion and internal negligence might contribute to cyber security breaches or attacks. It is challenging to prevent cyber security breaches or attacks caused by natural disasters. Besides, targeted external attacks on the organization, politically motivated attacks, and negligence of cyber compliance from the organization's partners such as suppliers, are all external factors that lead to the cyber security breaches or attacks. SETA programs are the most effective way to improve employee information security protection behavior [6,7]. Thus, carrying out SETA training programs is a highly effective way to prevent cyber security breaches or attacks caused by human errors of the staff. In this case, we integrated the staff-related factors in the original dataset into one feature "Factor" to investigate whether the cyber security breaches and attacks related to staff-related factors are the key elements leading to the SETA implementation. Specifically, the feature "Factor" covers eight staff-related factors, including human error, unchanged or unsecure passwords, staff, ex-staff or contractors deliberately abusing their account, staff or ex-staff or contractors not adhering to policies or processes, absent vetting or inadequate vetting of staff, ex-staff or contractors, staff lacking awareness or knowledge, unsecure settings on browsers, software, computers, or accounts, and browsing untrusted or unsafe websites. For the "Factor", we also converted it into the nominal type, same as the "Insure". When there is at least one staff-related factor in the original dataset leading to the occurrence of cyber security breaches or attacks, the value of the "Factor" will be "Yes". When there is no staff-related factor in the original dataset leading to the occurrence of cyber security breaches or attacks, the value of the "Factor" will be "No". When the organization has not been attacked in the past 12 months of the survey, the value of the "Factor" will be "Unable to distinguish".

"Cloud": The "Cloud" describes whether the organization uses cloud or any other type of externally-hosted web services. Externally-hosted web services are efficient for organizations to reduce the difficulty of development, operation, and maintenance, as well as reduce a large number of preliminary IT infrastructure investments, enabling the organizations to focus more on their business operation and innovation. Compared to organizations using traditional servers, organizations using externally-hosted web services will relatively lack some control over some resources and operating material such as datasets, which exist as potential management risks for the organizations and companies. To protect the security of the confidential resources stored in externally-hosted web services, the organizations might conduct SETA implementation to reduce the loss and occurrence of threats such as social engineering. For the "Cloud", we converted it into the nominal type.

"Critical": The "Critical" describes the significance of the externally-hosted web services to an organization when this organization uses externally-hosted web services. The organizations with higher importance on the externally-hosted web services might focus more on the protection of the resources stored in externally-hosted web services. The "Critical" is converted into the ordinal type according to the corresponding significance of externally-hosted web services.

"Train": The "Train" is the class. In this study, we regard the task exploring the causes evoking companies to conduct SETA implementation as a binary classification task. Thus, we merged the situations that anyone from an organization has participated in any type of SETA training programs mentioned in the survey, covering attending seminars or conferences on cyber security, attending any externally provided training on cyber security, and receiving any internal training on cyber security, into one set, while leaving the organizations with nobody participating in any type of SETA training programs as the other set. Following this rule, we converted the "Train" into a binary class type.

Based on the representations of generated features in the real world, we divide the features into four feature groups and evaluate their contribution to SETA implementation in the four corresponding dimensions. The four divided feature groups are respectively companies' and organizations' nature related group, corporate leadership's awareness related group, internal factor related group, and external factor related group.

The feature values and corresponding feature groups of one example company are shown in Table 1.

Table 1. Data Example and Corresponding Feature Groups.

Features	Data	Feature Groups
Update	Weekly	Awareness Related
Sizec	Micro (2 to 9)	Nature Related
Freq	Once only	External Factor Related
Priority	Very high	Awareness Related
Numbb	Fewer than 3	External Factor Related
Core	To some extent	Nature Related
Insure	Yes	Nature Related
Factor	No	Internal Factor Related
Cloud	Yes	Nature Related
Critical	Fairly critical	Nature Related
Train	Yes	Class

3.2. Experiment Design

Compared to the general qualitative methods used in SETA implementation studies, we use the supervised learning method to test the hypotheses that generated features are all contributing to businesses' SETA implementation via feature importance. The framework of the whole experiment is shown in Figure 1.

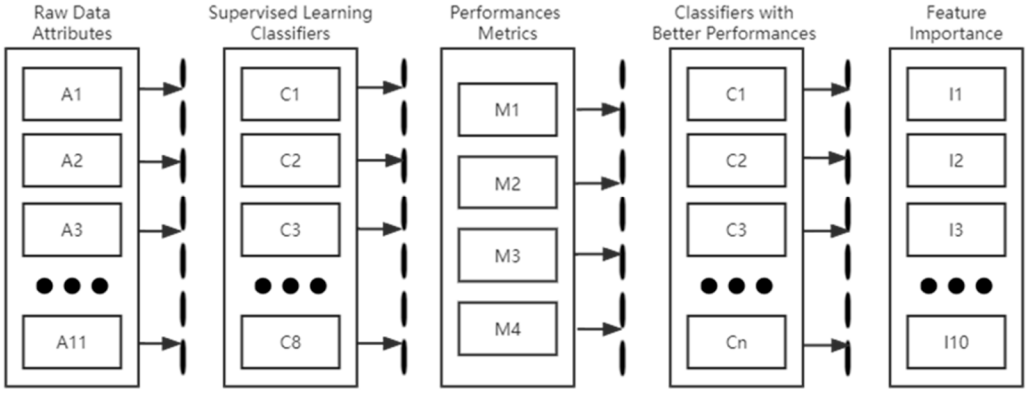

Figure 1. Experimental Framework.

In the experimental part, we use the stratified K-fold cross-validation method to reduce model training bias caused by the possible data splitting contingency. Specifically, we use the stratified 5-fold cross-validation method to split the dataset into training dataset and validation dataset with the proportion of 8:2 in five times. The five split validation datasets could be merged to form the whole dataset. In each splitting, we train the classification models using the training dataset and evaluate the performances of trained models using the validation dataset. In the evaluation step, we calculate the average classification performances of the models among four different evaluation metrics in the five split validation datasets to evaluate the performances of these models in this task.

In order to more accurately investigate the causes evoking the companies to conduct SETA implementation, it is essential to achieve more accurate detection of the companies with SETA implementation using the features. Thus, based on the type of features, we train and evaluate eight representative supervised learning models that are generally effective in the classification of this kind of feature set to select the model with better performances in detecting the companies with SETA implementation. The four models are respectively Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, Decision Tree, Random Forest (RF), Bagging, AdaBoost, and Light Gradient Boosting Machine (LightGBM). Besides, since the dataset is an imbalanced dataset, we choose to use the cost-sensitive method to handle the class imbalanced problem and train the selected classification models except for the Naïve Bayes model. Due to the unique operating principle of the Naïve Bayes model, the cost-sensitive method is not suitable to be applied to this model, which will be specifically demonstrated in the following.

The eight selected models cover four base classification models, SVM, Naïve Bayes, Logistic Regression and Decision Tree, and four ensemble classification models, RF, Bagging, AdaBoost and LightGBM. When training the models, we adopt some methods to avoid overfitting and boost the model performances. For example, we choose to set the maximum depth for the trees generated in the tree-related models to avoid overfitting.

Support Vector Machine is a base classification model that has overall decent performances in many classification tasks [25]. It is a discriminative classifier that is generally used in binary classification. The idea of SVM in binary classification is to find the optimal hyperplane which can separate the m-dimensional data into two classes [26]. In this case, based on the feature type, we employ the Radial Basis Function (RBF) kernel to train the SVM classifier.

Naïve Bayes is a base classification model applying the Bayes' theorem. A Naïve Bayes model-based classifier could be efficiently conducted and trained without any complicated parameter estimation. Thus, it is suitable for the classification task with high dimensional input features.

Logistic Regression is a base classification model, which is a standard method to build prediction models for a binary class outcome. In the binary classification task, the logistic regression uses the sigmoid function to convert the value range into 0–1 and find the decision boundary in the converted hyperplane to classify the input data.

Decision Tree is a base classification model to classify data by generating a treelike graph with a series of rules. It uses a treelike graph to guide the input feature to one of the class labels. A unique rule used to test the input data is embedded in an internal node of a decision tree. The possible results classified by the rules of the internal nodes will generate corresponding numbers of outgoing branches connected to this internal node.

Random Forest is an ensemble classification model. This model constructs a number of decision trees and compares the results from all sub-decision trees to generate the final classification outcome. It is an application of the bagging method. The difference between random forest from the combination of the bagging method and decision tree classifier is that random forest will randomly select a subset of input features when generating nodes of sub-decision trees.

Bagging uses a majority vote to determine the class of ensemble classifiers. In this case, the bagging classifier builds decision tree classifiers on each bootstrap sample and

the generated final output is the majority vote of the built sub-decision tree classifiers classification results.

Compared with the Random Forest model and the combination of the bagging method and decision tree classifier, the AdaBoost model adds weights for each sub-classifier. In determining the final result, the final output is generated by weighted voting. At the beginning of this method, each classifier has the same weight. In each iteration, the weight of the classifier in this iteration might change, and the weight change focuses on the misclassified records in previous iterations.

Light Gradient Boosting Machine is an improved version of the Gradient Boosting Decision Tree. Gradient Boosting Decision Tree is an ensemble model based on Decision Tree, which is a widely used machine learning algorithm [27]. In Gradient Boosting Decision Tree, a series of weak learners are constructed along the gradient, and they are then combined within corresponding weights. The weighted result is the decision made by GBDT. Compared to Gradient Boosting Decision Tree, Light Gradient Boosting Machine could obtain almost the same performances with a much smaller amount of training time [28].

The four metrics that we adopt to evaluate the performances of the classification models are separately precision (P), recall (R), F-measure (F), and accuracy (A) [29]. Equations of these four metrics are shown as follows:

$$P = \frac{tp}{tp + fp} \tag{1}$$

$$R = \frac{tp}{tp + fn} \tag{2}$$

$$F = \frac{2PR}{P + R} \tag{3}$$

$$A = \frac{tp + tn}{tp + tn + fp + fn} \tag{4}$$

In these four equations, tp is true positive, which is the number of correctly detected companies that have held SETA implementation; fp is false positive, which is the number of companies without SETA implementation that are incorrectly detected as companies that have held SETA implementation; fn is false negative, which is the number of companies that have held SETA implementation and are incorrectly detected as companies without SETA implementation; tn is true negative, which is the number of correctly detected companies without SETA implementation. Values of these four metrics range from 0 to 100 percent.

3.3. Performance Comparison

The classification performances of the eight classification models in the four metrics are presented in Table 2.

Table 2. Performances of Models in Validation Dataset.

Classifiers	P	R	F	A
SVM	62.17%	**79.60%**	69.77%	73.98%
Naïve Bayes	**67.29%**	62.80%	64.86%	74.29%
Logistic Regression	60.98%	78.00%	68.39%	72.77%
Decision Tree	58.16%	76.80%	66.17%	70.34%
RF	64.42%	**79.60%**	**71.16%**	**75.64%**
Bagging	63.33%	78.00%	69.84%	74.58%
AdaBoost	56.68%	58.80%	57.62%	67.33%
LightGBM	60.28%	71.60%	65.29%	71.10%

The standard deviations of the model performances among the five different split results of the stratified 5-fold cross-validation method in the four metrics are presented in Table 3.

Table 3. Standard Deviations of Model Performances among the Cross Validation Split Results.

Classifiers	P	R	F	A
SVM	0.0116	0.0445	0.0226	0.0145
Naïve Bayes	0.0365	0.0392	0.0291	0.0223
Logistic Regression	0.0264	**0.0490**	0.0305	0.0248
Decision Tree	0.0251	0.0466	**0.0325**	0.0259
RF	0.0155	0.0427	0.0214	0.0146
Bagging	0.0190	**0.0490**	0.0259	0.0181
AdaBoost	0.0295	0.0412	0.0269	0.0215
LightGBM	**0.0425**	0.0233	0.0177	**0.0298**

As shown in Table 3, the highest standard deviation of the eight classification models in the four metrics are 0.0490, less than 0.05, so these models have stable performances in the four metrics.

Higher outputs in the four metrics, precision, recall, *F*-measure, and accuracy reflect better performances of the model in the classification task. As it is shown in Table 2, the Random Forest classifier obtains the best performances in the recall, *F*-measure, and accuracy, reaching, respectively, 79.60%, 71.16%, and 75.64%, which outperforms all the other classifiers. The Naïve Bayes classifier obtains the best performance in the precision, reaching 67.29%, but is much worse performances in the other three metrics, especially in the recall and *F*-measure. In the precision, the Random Forest classifier also ranks the second-best in all classifiers, reaching 64.42%. Although the SVM classifier also obtains the highest ranking in the recall, such as the RF classifier, its performances in the other three metrics are much worse than the RF classifier.

Overall, the RF model could achieve more stable performances than other models in detecting the companies and organizations that have held SETA training. According to the performances of the RF classifier, it is shown that the RF classifier could obtain high performance in the recall, meaning that this classifier could well detect the companies and organizations that have held SETA implementation. By contrast, the precision values of the RF classifier and other classifiers are all not very high, indicating that there are some companies or organizations without SETA implementation that might have highly similar characteristics to the companies or organizations with SETA implementation in the tested factors.

To better understand the reasons leading to the SETA implementation for companies, we visualize the importance of all features for the RF model in the experiment, shown in Figure 2.

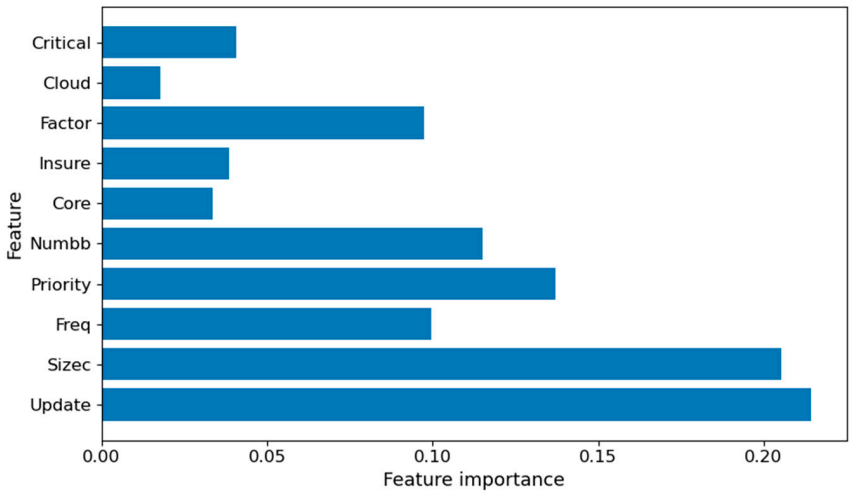

Figure 2. Feature Importance of Random Forest Model.

As it is shown in Figure 2, the features "Update" and "Sizec" play the most important role for the RF model in detecting the companies and organizations with SETA implemen-

tation. Following the top 2 best features, the features, "Priority", "Numbb", "Freq", and "Factor", also have relatively significant impacts on the detection task for the RF classifier. These six features are relatively more useful features in the detection task, and among these useful features, the two features in awareness related feature group respectively obtain the highest and the third highest importance. Thus, the feature group with corporate leadership's awareness related features have a much higher importance than the other three feature groups, reflecting that there is a great willingness for the companies and organizations whose leadership has stable cyber security protection awareness to conduct SETA implementation. As for the companies' and organizations' nature related features, only "Sizec" has a significant impact on the detection, obtaining the second highest importance in the task, which represents that companies and organizations tend to conduct SETA implementation when their scale reaches a certain level. By contrast, the internal and external factor related features have relatively lower importance in the useful features for the classification task.

Compared to the six useful features, the other four features have relatively lower contributions to the classification, so the companies and organizations that participated in the survey are generally not considering much about the potential attacks in the stuff using externally-hosted web services and the potential incremental attack rate brought by the high dependence of products and services on online services.

4. Discussion

The previous research on SETA implementation are mainly qualitative research, through the investigation of institutions and staff to find the corresponding weaknesses and then put forward suggestions. For example, some researchers demonstrated the percentage of employees in Turkish public agencies trapped in social engineering tests and determined the reasons leading them to be trapped through interviews [30], thus suggesting organizations conduct SETA implementation and need to be aware of preventive measures; Applegate summarized the types, psychology, impact, and mitigation of social engineering attacks, and proposed that since hackers like to bypass technical controls through human factors, the most effective mitigation measure for cyber security attacks is to educate staff about the related threats, common techniques used in security attacks, and the potential damage from breaching of critical information [31]. These studies are generally not conducted based on the models and methods of quantitative analysis. Instead, they use qualitative methods such as interviews or apply sociology and psychology theories to state the importance of SETA training or put forward some strategies for conducting training. Based on those shortcomings, this paper uses supervised learning methods to conduct the experiment, which has filled some gaps in previous studies from the perspective of quantitative research and model diversity.

Some scholars also emphasize the designing of suitable training systems. They think that employees are important to handle cyber security issues, so the training process is significant for the company to improve the cyber security issues. Some researchers tried to use a game-based program to train the employees [32]. There are also some researchers who pointed out that due to different kinds of threats that happened in the company, the training program may be more detailed and separate, which means that the flow diagrams needed to be designed for the employees in order to direct them to face the attacks using different approaches [33]. While there are many scholars wanting to design a proper SETA implementation process, they do not consider whether or not the company will spend time and cost to build that kind of training process because the company needs to take lots of things into consideration excluded cyber security. In this article, we figure out the key elements leading businesses to adopt the SETA implementation. Focusing on the issues neglected by companies found in this article, the SETA training designers will be easier to improve the SETA training process and further boost the information security sustainability of businesses.

This paper discusses the factors impacting the conduction of SETA implementation from the perspective of quantitative analysis. Eight supervised learning models are used to explore the critical reasons prompting organizations to conduct SETA implementation. The validity of models is measured by four measures, including precision, recall, *F*-measure, and accuracy. In these four measures, the Random Forest model could achieve relatively better and more stable performances. The feature importance is further generated in the experiment to identify the more important features to conduct SETA implementation. The use of models and methods is one of the main innovations in the task of investigating the factors contributing to companies' and organizations' SETA implementation.

In this study, although most companies that have held SETA implementation could be classified very well, there are still some companies that have held SETA implementation that cannot be classified accurately. From this perspective, this study reveals that the decision-making of conducting SETA implementation in most companies is related to the factors explored in this study. However, even though some companies and organizations have highly similar characteristics to the companies and organizations with SETA implementation, such as the company size and frequency of exposure to cyber-attacks, they are still not conducting any SETA implementation activities, so there might be some bias hindering them to conduct SETA training. According to the feature importance of features in the classification, it can be proved that the corporate leadership's awareness related factors are the key elements leading the companies to conduct SETA implementation. Thus, when the leadership of companies and organizations own a higher degree of awareness of cyber security protection, the probability of holding SETA implementation will be higher. Companies also tend to hold SETA implementation when they realize that their scale has reached certain levels. Besides, most companies and organizations lack the realization that the SETA implementation is beneficial to prevent the potential incremental cyber-attack rate brought by high dependence of products and services on online services and potential cyber-attacks targeted for externally-hosted web services.

In the future, we plan to explore the solutions on how to implement the awareness process for the leadership separately in the small, medium, or high-level organizations and the specific bias hindering the decision-making in conducting SETA implementation for businesses, contributing to their SETA implementation, and further help them to better achieve the sustainability of information security.

5. Conclusions

Due to the insufficiency in the commonly used tools-based security strategy, this study aims to explore the specific causes leading companies and organizations to conduct SETA implementation and further help them to better achieve the sustainability of information security. Based on the dataset collected from a questionnaire designed by Ipsos MORI Social Research Institute, we train and test eight supervised learning model-based classifiers and compare their performances in four metrics. The results show that the Random Forest model could achieve stable performances in task-detecting companies and organizations with SETA implementation. According to the performances and feature importance of the RF model, it is shown that the corporate leadership's awareness related factors have the most significant impact on the decision-making of conducting SETA implementation for the businesses, and there might be some other potential bias hindering part of the businesses to hold SETA training. Besides, most businesses are generally lacking awareness about cyber security protection for the stuff using externally-hosted web services and products as well as services depending on online services. Our findings provide some research foundation for the governments to carry out the corresponding SETA implementation importance publicity and the SETA training designers to improve the SETA training process to boost the development of information security sustainability in businesses.

Author Contributions: Conceptualization, G.W., D.T., Y.C. and H.J.; methodology, G.W.; software, G.W.; validation, G.W.; formal analysis, G.W., Y.C. and H.J.; investigation, G.W.; data curation, G.W., Y.C. and H.J.; writing—original draft preparation, G.W., Y.C. and H.J.; writing—review and editing,

D.T. and G.W.; visualization, G.W.; supervision, D.T.; project administration, D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.gov.uk/government/publications/cyber-security-breaches-survey-2016 (accessed on 29 September 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aldawood, H.; Skinner, G. Reviewing cyber security social engineering training and awareness programs—Pitfalls and ongoing issues. *Future Internet* **2019**, *11*, 73. [CrossRef]
2. Labrecque, L.I.; Markos, E.; Swani, K.; Peña, P. When data security goes wrong: Examining the impact of stress, social contract violation, and data type on consumer coping responses following a data breach. *J. Bus. Res.* **2021**, *135*, 559–571. [CrossRef]
3. Klahr, R.; Amili, S.; Shah, J.N.; Button, M.; Wang, V. *Cyber Security Breaches Survey 2016*; Department for Digital, Culture, Media & Sport: London, UK, 2016.
4. Aldawood, H.A.; Skinner, G. A critical Appraisal of Contemporary Cyber Security Social Engineering Solutions: Measures, Policies, Tools and Applications. In Proceedings of the 2018 26th International Conference on Systems Engineering (ICSEng), Sydney, Australia, 18–20 December 2018; pp. 1–6.
5. Al-Ghamdi, M.I. Effects of knowledge of cyber security on prevention of attacks. *Mater. Today Proc.* **2021**. [CrossRef]
6. Mani, D.; Raymond Choo, K.; Mubarak, S. Information security in the South Australian real estate industry. *Inf. Manag. Comput. Secur.* **2014**, *22*, 24–41. [CrossRef]
7. Kennedy, S.E. The pathway to security–mitigating user negligence. *Inf. Comput. Secur.* **2016**, *24*, 255–264. [CrossRef]
8. Zhang, L.; McDowell, W.C. Am I really at risk? Determinants of online users' intentions to use strong passwords. *J. Internet Commer.* **2009**, *8*, 180–197. [CrossRef]
9. Chin, A.G.; Etudo, U.; Harris, M.A. On Mobile Device Security Practices and Training Efficacy: An Empirical Study. *Inform. Educ.* **2016**, *15*, 235. [CrossRef]
10. Slusky, L.; Partow-Navid, P. Students Information Security Practices and Awareness. *J. Inf. Priv. Secur.* **2012**, *8*, 3–26. [CrossRef]
11. Winfred, Y.; Daniel, O.W.; Peace, K. SETA and Security Behavior: Mediating Role of Employee Relations, Monitoring, and Accountability. *J. Glob. Inf. Manag.* **2019**, *27*, 102–121. [CrossRef]
12. Aoyama, T.; Naruoka, H.; Koshijima, I.; Watanabe, K. How Management Goes Wrong? The Human Factor Lessons Learned from a Cyber Incident Handling Exercise. *Procedia Manuf.* **2015**, *3*, 1082–1087. [CrossRef]
13. Olusegun, O.J.; Ithnin, N.B. People are the answer to security: Establishing a Sustainable Information Security Awareness Training (ISAT) program in organization. *arXiv* **2013**, arXiv:1309.0188.
14. Ghafir, I.; Prenosil, V.; Alhejailan, A.; Hammoudeh, M. Social engineering attack strategies and defence approaches. In Proceedings of the 2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud), Vienna, Austria, 22–24 August 2016; pp. 145–149.
15. Gardner, B.; Thomas, V. *Building an Information Security Awareness Program: Defending Against Social Engineering and Technical Threats*; Elsevier: Amsterdam, The Netherlands, 2014.
16. Kumar, A.; Chaudhary, M.; Kumar, N. Social engineering threats and awareness: A survey. *Eur. J. Adv. Eng. Technol.* **2015**, *2*, 15–19.
17. McCrohan, K.F.; Engel, K.; Harvey, J.W. Influence of Awareness and Training on Cyber Security. *J. Internet Commer.* **2010**, *9*, 23–41. [CrossRef]
18. Puhakainen, P.; Siponen, M. Improving Employees' Compliance Through Information Systems Security Training: An Action Research Study. *Mis. Quart.* **2010**, *34*, 757–778. [CrossRef]
19. Furman, S.; Theofanos, M.F.; Choong, Y.; Stanton, B. Basing cybersecurity training on user perceptions. *IEEE Secur. Priv.* **2011**, *10*, 40–49. [CrossRef]
20. Cone, B.D.; Thompson, M.F.; Irvine, C.E.; Nguyen, T.D. Cyber security training and awareness through game play. In Proceedings of the IFIP International Information Security Conference, Karlstad, Sweden, 22–24 May 2006; pp. 431–436.
21. Abbott, R.G.; McClain, J.; Anderson, B.; Nauer, K.; Silva, A.; Forsythe, C. Log Analysis of Cyber Security Training Exercises. *Procedia Manuf.* **2015**, *3*, 5088–5094. [CrossRef]
22. Hatzivasilis, G.; Ioannidis, S.; Smyrlis, M.; Spanoudakis, G.; Frati, F.; Goeke, L.; Hildebrandt, T.; Tsakirakis, G.; Oikonomou, F.; Leftheriotis, G. Modern aspects of cyber-security training and continuous adaptation of Programmes to trainees. *Appl. Sci.* **2020**, *10*, 5702. [CrossRef]
23. Osisanwo, F.Y.; Akinsola, J.; Awodele, O.; Hinmikaiye, J.O.; Olakanmi, O.; Akinjobi, J. Supervised machine learning algorithms: Classification and comparison. *Int. J. Comput. Trends Technol.* **2017**, *48*, 128–138.

24. Al-Alawi, A.I.; Al-Bassam, S.A.; Mehrotra, A.A. Critical Cybersecurity Threats: Frontline Issues Faced by Bahraini Organizations. In *Implementing Computational Intelligence Techniques for Security Systems Design*; IGI Global: Hershey, PA, USA, 2020; pp. 210–229.
25. Salcedo Sanz, S.; Rojo Álvarez, J.L.; Martínez Ramón, M.; Camps Valls, G. Support vector machines in engineering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2014**, *4*, 234–267. [CrossRef]
26. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1999.
27. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
28. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3149–3157.
29. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
30. Mataracioglu, T.; Ozkan, S. User awareness measurement through social engineering. *arXiv* **2011**, arXiv:1108.2149.
31. Applegate, S.D. Social engineering: Hacking the wetware! *Inf. Secur. J. Glob. Perspect.* **2009**, *18*, 40–46. [CrossRef]
32. Peery, J.G.; Pasalar, C. *Designing the Learning Experiences in Serious Games: The Overt and the Subtle—The Virtual Clinic Learning Environment, Informatics, 2018*; Multidisciplinary Digital Publishing Institute: Basel, Switzerland, 2018; p. 30.
33. Beckers, K.; Pape, S. A serious game for eliciting social engineering security requirements. In Proceedings of the 2016 IEEE 24th International Requirements Engineering Conference (RE), Beijing, China, 12–16 September 2016; pp. 16–25.