IMPACT OF PRIOR VICTIMIZATION ON SUSCEPTIBILITY TO PHISHING ATTACKS

By

Kajal Pawar

A dissertation submitted in partial fulfilment of the requirements for the degree

Of

MSc Business Analytics

At

Dublin Business School

Supervisor: Kinkini Chatterjee

Declaration

I declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This dissertation is my work and contains nothing which is the outcome of work done in collaboration with others except as specified in the text and Acknowledgements.

Signed: Kajal Pawar

Dated : 20th May 2024

Acknowledgements

This work seemed possible due to a significant amount of support and academic Knowledge provided by the professor Kinkini Chatterjee. She helped me select the right project and helped by giving me ideas for scrapping and writing an effective code to achieve the goal. His guidance and ideas helped me in writing a perfect report in a short span.

I also like to express my gratitude to the Department of Business Analytics for their constant support. Their provision of the required tools, resources, and infrastructure was essential to the successful completion of this study. Their dedication to creating a favourable climate for research is genuinely admirable.

In addition, I want to express my gratitude to my friends and peers for their insights that helped me finish this project. They taught me a lot of insightful things. I never experienced tension while working on my dissertation because of the laid-back atmosphere they offered.

To sum up, I would want to express my sincere gratitude to everyone who helped to make this research project a success. It has been a true joy to work with such committed people, and I am grateful for your consistent support, direction, and encouragement. I sincerely appreciate all of your contributions and am grateful to each and every one of you.

Abstract

This study evaluated various machine learning (ML) algorithms to predict susceptibility to phishing attacks based on prior victimization. The models assessed included Decision Tree Classifier, Support Vector Classifier (SVC), Random Forest, XGBoost, and Logistic Regression. SVC and Logistic Regression achieved the highest accuracy of 0.86 and an F1 score of 0.79, making them top performers. Random Forest also showed strong results with an accuracy of 0.85 and an F1 score of 0.79, while XGBoost had an accuracy of 0.82 and an F1 score of 0.79. The Decision Tree Classifier was the least effective, with an accuracy of 0.75 and an F1 score of 0.76. Feature selection significantly enhanced model performance, and the quality and size of the training dataset were crucial. This study concludes that SVC and Logistic Regression are the most effective models for predicting phishing susceptibility, offering valuable insights for improving cybersecurity measures.

Table of Contents

Chapter 1: Introduction

1.1 Background

In an age when mostly everything is done online, the security of online transactions and messages has become critical. Phishing attacks are perhaps most widely known among the many cyber threats that users are exposed to. Although they often leverage technical vulnerabilities, they are particularly cunning in that they also exploit human psychology. Fraudsters commit these attacks by deceiving individuals giving away their financial and personal data by faking as trusted people (Alkhalil et al., 2021). Although improvements in cybersecurity technologies and the rising level of user awareness have been made, phishing nevertheless continues to exist as an important security concern as phishing attacks are more sophisticated and easier for attackers to carry out on a large scale. Research has been mainly technical and has focused on the demographic characteristics of victims and also, the consequences for the individuals such as prior victimization which have not been given a prominent position. This distance demonstrates how multidimensional the issue is, and it requires more sensitive research, which will reveal how previous cases of phishing influence the user's ability to distinguish and protect himself from further attacks (Alabdan, 2020).

1.2 Significance of the Study

The significance of this study lies in its focus on a largely unexplored factor in cybersecurity: an effect of prior victimization of a person to phishing attack. This understanding is necessary from various perspectives (Rastenis et al., 2020). Out of the various things, it is first of all which does not support the old myth of being aware and experienced enough to be able to identify and avoid phishing attempts. By deriving the fact that prior victims seem more prone to future attacks, the latter might be considered a reason for a creation of educational interventions that deal with psychological issues, rather than just general awareness.

The second aspect is that, by identifying the features and behaviours that separate the former victims from non-victims in the tern of susceptibility, cybersecurity measures can be more precisely made. It could be an impetus for the creation of better machine learning techniques and result in more accurate phishing detection tools and customized situations for cyber security training (R. Marusenko, Sokolov and V. Buriachok, 2020).

Besides the outcomes of this research, the policy-making and strategic decisions of the organization can also be influenced, and in particular, training programs of the employees and

the interior security protocols can be shaped. Through applying the findings of this study, organizations can empower their personnel to the point where they are able to resist the various social engineering methods in cyberattacks, resulting in stronger cybersecurity stance (Hawa Apandi, Sallim and Mohd Sidek, 2020).

1.3 Research Gap

Although the vast research on demographic and technical features related to the risk of phishing attacks is available, there is a huge gap in the knowledge of the psychological and experiential aspects. Additionally, there is a lack of focus in the investigation of prior victimization, which has the potential to enhance the risk level (Carroll, Adejobi and Montasari, 2022). Many investigations did not make a separate subgroup of victims who are experiencing their first encounter and those with previous experiences, therefore, the understanding of how previous encounters could make future risk assessment and behaviour different has been ignored (Bhardwaj et al., 2021). Therefore, this oversight reduces the effect of existing educational and prevention policies that have been implemented, as this study emphasizes the need to focus on the prior victimization.

1.4 Problem Statement

This research addresses a critical problem discerning, if those who have already become the victims of phishing attacks are more likely to fall in the trap again in comparison to those who have not gone through such an experience. The existing literature does not pay much attention to the role that the history of victimization plays in the future likelihood of victimization, which is mostly viewed as a skill that the victims may develop to become more aware. On the one hand, it could be that victims that have been breached in the past are basically more vulnerable; the consequences of this, though, would be relevant for the design of cybersecurity measures and training programs (Desolda et al., 2022). This study will capitalize on filling this gap by applying dynamic ML models for the influence of prior victimization on the probability of repeatedly falling for phishing scams.

1.5 Aim

The objective of this research is to explore how previous victimization affects an individual's vulnerability to phishing attacks, utilizing various machine learning models to identify those who are more likely to fall for phishing scams due to their past experiences.

1.6 Objectives

- To determine the effectiveness of different ML algorithmic models in forecasting future susceptibility to phishing attacks using prior victimization data.
- To compare the performance metrics of various ML models in identifying individuals prone to phishing attacks due to past experiences.
- To analyse the impact of dataset size and quality, containing previous phishing incidents, on the performance of ML models in predicting future susceptibility.
- To develop a dependable and accurate ML model capable of distinguishing between individuals who have previously been victimized by phishing attacks and those who are at risk of future attacks.

1.7 Research question

- What effect does prior encounter with a phishing attack have on the susceptibility to future phishing attacks among individuals?
- Which ML model provides the highest accuracy in predicting future susceptibility to phishing attacks based on prior victimization data?
- How can a reliable and accurate ML model be constructed to distinguish between individuals who have been previously victimized by phishing attacks and those who are susceptible to future attacks?

1.8 Structure of dissertation

TITLE	DESCRIPTION
INTRODUCTION	Overview of the topic, significance, research gap, aim, and objectives.
LITERATURE REVIEW	Detailed analysis of existing research related to phishing, victimization, and cybersecurity training.

METHODOLOGY	Description of the research design, data collection methods, and analytical techniques.
DATA ANALYSIS	Presentation and interpretation of the data, identifying trends and patterns.
DISCUSSION	Discussion of findings in relation to the literature reviewed.
CONCLUSIONS AND RECOMMENDATIONS	Summary of findings, limitations of the study, recommendations for future research.

Chapter 2: Literature Review

This chapter reviews recent studies on phishing susceptibility and describes the strategies used to prevent cyberattacks with a focus on Machine Learning techniques. It will target the main issues that lead to vulnerability to phishing and outline gaps in the literature. The themes defined in the literature include past phishing experience, behavioural and emotional variables, demographic variables, prediction models, AI/ML and phishing, and organizational training. The purpose of this chapter is to try to apply ML algorithms like Decision Trees, Random Forests, SVM, or similar models to understand how this type of model improves the detection and prevention of phishing attacks. The current research with reference to the literature presented in this chapter intends to achieve the following: To increase the knowledge of awareness and understanding of the current status of phishing researches To explore the research gaps that should be addressed in the future to improve cybersecurity knowledge and training.

2.1 Impact of Previous Phishing Encounters on Susceptibility

The authors Chen, Gaia, and Rao (2020) under their investigation, have shown that the frequency of phishing attacks that people experience recently, increases their vulnerability to future attacks. The scientists systematically dismantled the correlation between the past experiences with phishing and probability of future victimization. They pointed out that the difficulties of identifying phishing and the consequences that result from such detection lead to whether one is susceptible to phishing or not.

For instance, consider the situation in which a user has previously been able to detect and avoid phishing will be more cautious, however, the repeated failures could lead to frustration or resignation, thus increasing the probability of falling victim to phishing. With Decision Trees or Random Forests or Support Vector Machines it is possible to construct models which can be used for gaining knowledge about past events and potential threats of phishing in the future. These models aim at identifying the set of group risk factors and individual risk factors for developing such programs. For example, Random Forests can be applied for future phishing attacks prediction and SVM for people classification before prevention. Machine learning enhances training by automatically updating itself with new information and discovering psychological factors that make one more susceptible.

2.2 Behavioural and Emotional Influences on Phishing Susceptibility

Abroshan et al. (2021) study was mainly intended to highlight the role of emotions and behaviours in the attacks. The research outcomes showed that the emotional responses like panic and worry were the main reasons for the people to become victims of phishing schemes. Criminals often use the same methods, such as making people afraid to do something in a hurry or impulsively so that they cannot think clearly. This conclusion underlines the essentiality of the cybersecurity defences that do not only involve technical ways but also the strategies that take into account the fact that the attackers can manipulate people emotionally and psychologically. Such defences should include training that helps people notice and manage the emotional manipulations that phishing attacks usually use.

2.3 Role of Demographic and Behavioural Factors

Greitzer et al. (2021) and Li et al. (2020) conducted laboratory studies to measure the phishing susceptibility that is affected by different factors. The above-mentioned studies altogether point out that age, gender are not the only factors that determine a person's susceptibility to phishing attacks; the behavioural characteristics, especially impulsivity and security hygiene, of an individual play a significant role in predicting the vulnerability to phishing attacks. For instance, people with a lower age or without much knowledge of technology might not be able to see the fine line between a scam that is usually very complex and clever. Nevertheless, older people may be generally more careful than younger generations who are not yet capable of distinguishing between genuine and fake messages as they are not that familiar with digital culture. Interestingly, these studies reveal that the security wise behaviour such as not updating software or clicking links without performing verification are the factors that increase the risk. These findings only underscore the need for a tailored cybersecurity training program that is adapted to the individual context of each group.

2.4 Predictive Models and Phishing Victimization

Stalans et al. (2023) as well as Chan-Tin et al. (2023) are the studies that are relevant. (2022) have achieved great success in the prediction of phishing victimization, which is based on the emotional and cognitive decision styles exclusively. The investigations of researchers in this area have shown that emotional reactions and thinking styles also contribute to the high level of vulnerability to phishing. These factors are in addition to the previous victimization. Similar to this, people who are high in generalized anxiety might very well react impulsively to phishing attacks that are designed to arouse the feeling of urgency. Just like that, the people

with a particular cognitive style could miss some vital signs that signify dishonesty or could look into every detail of a communication which would result in incorrect evaluations of phishing attempts. One of the possible applications of the research outcomes may be the use of the findings as a basis for the development of the cybersecurity training programs that are based on the psychological features of different user groups and might be customized in accordance with their cognitive and emotional profiles.

2.5 The Role of Machine Learning in AI-Driven Cybersecurity Training

The development of Artificial Intelligence (AI) in cybersecurity procedures is a topic that is changing the way businesses tackle phishing prevention from the old to the new. In the article of Ansari et al. (2022), the authors indicate the future of AI to be the revolutionary factor in the development of cybersecurity awareness systems. The systems which are meant for use by machines are intended to use the machine learning algorithms so that they can be able to respond to phishing attacks faster than the traditional methods. Through the AI's ability to identify and learn the new phishing techniques, it will be able to provide real-time, dynamic protection that will have the capability to adjust to the ever-evolving cyber threats' environment. The ability to keep learning and getting better at a given task is one of the most important benefits that AI brings to the table in the war against phishing.

In fact, Hillman et al. (2023) point out the meaning and customization of cyber-security learning too. The results of the study demonstrate that the possibility of phishing can be decreased to a great extent by the development of personalized training courses for the organization, which will target its specific difficulties and strengths. A more targeted education program that addresses the employee's positions, online behaviours, and past phishing experiences has a higher probability to achieve the desired result than general training. As well, the training time should also be taken into consideration. Frequent updates and reminders might be useful in maintaining the high level of vigilance and readiness for the new phishing methods.

2.6 Cybersecurity Awareness and Crime Prevention

Back and Guerette (2021) concentrate on increasing cybersecurity awareness training as a form regarding the level of mitigation against phishing attacks. They posit that training activities that include having the participants overlook the problems, being aware of the consequences of a failure, and the possibility of identifying malicious activities are essential in fostering the security culture. The author is keen to point out that although the study deals with the aspect

of participation it does not analyse the problems associated with providing periodic adaptive training which is a very time-consuming process and requires a lot of resources.

That is why the constant updating of security mechanisms is necessary for guaranteeing the organization's security from new methods of phishing and trends in cybercrime but there is no information on how such programs should be implemented and how to counter employee resistance. AI can be integrated into the cybersecurity awareness training by implementing the machine learning algorithm to identify behavioural patterns that can be used for identifying the high-risk users and those that should be paid special attention to the training. Decision Trees and Support Vector Machines can be trained to predict employees most susceptible to phishing based on their surfing behaviours and attendance records.

The machine learning for automation of the updates process ensures that the content is always updated, and the reinforcement learning algorithms can be used to differentiate the training for each employee depending on the results and feedback. This approach is effective in solving resource constraints by emphasizing efficient delivery of training and improvement of overall efficiency. Future research using machine learning can develop more accurate models for defining adaptive cybersecurity training and its use in resource-scarce organizations with 'unwilling' members. Future works should include the follow-up of the participants for longer periods to assess whether the training provided is capable of keeping them safe from phishing attacks in the long run.

2.7 Comprehensive Reviews on Phishing Susceptibility Factors

In their comprehensive articles from 2021 and 2022, Tornblad et al. and Rebovich and Byrne explore in detail the factors that influence phishing susceptibility. Such profound assessments show that the causal variables spanning from psychological traits like trust and risk perception to technical skills and previous cybersecurity experiences are quite diverse. Tornblad et al. (2021) summarize the different predictors identified in different studies ranging from demographic factors, personal traits, and behavioural patterns, which when considered put an individual at risk of falling victim to phishing. Rebovich and Byrne (2022) focused on this issue by examining the multifaceted interplay of these factors in cybersecurity at various levels. This means that the most effective strategy is the one that is comprehensive to curb phishing attacks. Such reviews show the way that in cybersecurity training, other methods must be used that will be able to consider the different factors that the youth are concerned about.

2.8 Machine Learning for Personalized Training and Prevention Strategies

Shappie et al. (2019) and Ribeiro et al. (2023) observe that both conscientiousness and openness should be considered in the context of cybersecurity training. There is a need for training that is designed to meet individual's needs but there is a lack of guidance on how such training should be practically implemented and how it should be maintained in the long run. Hillman et al. (2023) posit that training should be regular and must be aligned with the working schedules of the employees. They highlight the importance of context-dependent and time-sensitive training but do not consider constraints of resources and potential employee opposition. ML can be applied to personalize training according to risk and behavioural factors. ML can also be used to identify high-risk individuals and customize training resources. Future works have to be aimed at the development of cost-effective training models using sustainable machine-learning paradigms. By using the machine learning approach, the training programs can be made more interactive and time-saving, as well as reduce the number of phishing cases.

2.9 Experimental Approaches to Understanding Phishing with Machine Learning

Phishing vulnerability research has been very important in the knowledge gained in the factors that make people vulnerable to phishing attacks. Greitzer et al. (2021) and Li et al. (2020) specifically highlight that the propensity to this phenomenon is rooted in competencies, behaviours, and attributes. These include lack of control and security such as for example weak password policies and failure to install security updates. But experiments use samples taken from idealized environments in test tubes and other laboratory equipment that is far removed from reality. This limitation means that the experimental results have to be tested in real-world cases. These experimental methods can be further improved by employing machine learning to find behavioural and technical predictors of the heightened risk of falling for phishing. There are algorithms like Random Forests and Support Vector Machines that can be utilized to search for connections in vast amounts of data. Moreover, reinforcement learning can imitate real situations and improve the model several times. Machine learning for long-term research and field experiments will help verify the results and improve the efficiency and applicability of cybersecurity programs. The machine learning will assist the researchers to identify the weaknesses that are vulnerable to the phishing processes and also to discover interventions that are real life applicable.

2.10 The Evolution of Phishing Threats and Defences

The fact that technology keeps changing at the same pace also makes the criminals' methods more sophisticated and complicated, including phishing attacks that are difficult to recognize. Abroshan et al. (2021) mention that phishing is becoming more personalized where hackers are using emotional and psychological manipulation to take advantage of human behaviour. The study, however, highlighted the need for cybersecurity defences that are based on human psychology in addition to technical solutions. This is a crucial aspect, which may play a vital role in the overall cybersecurity defence. For example, the appearance of a message that purports to be an emergency mail from an organization that people trust can cause them to panic and override their rational reasoning in favour of complying with the attacker's request thereby increasing the rate of compliance.

The study proposed developing coping strategies to make people survive through psychological triggers. This could be achieved by running training programs that could be simulated in a safe environment where the individuals can experience the emotional and psychological stress associated with phishing scams without having to be exposed to them in reality. This practical learning is an excellent way for a person to learn and identify the scams that phishers use so that the victims would be less likely to be drawn into them.

2.11 Machine Learning in Phishing Defense

The application of ML in the creation of detection systems against phishing is one of the latest technology development boosts that can be used to enhance cybersecurity. Ansari et al. (2022) claimed that the application of ML-based systems has led to the increased efficiency of phishing identification and mitigation due to the superior data analysis and pattern recognition capabilities compared to what humans can handle. These ML systems are built with the intention of learning from every interaction and increasing their accuracy over time. Studies have also demonstrated that ML can offer adaptive security and personalized education. But they do not offer detailed instruction for applying ML in various types of organizations or handling resource limitations. The cost and future sustainability of ML-based security systems must also be addressed by developing low-cost and environmentally friendly substitutes. Some gaps in the literature need to be filled in the future, such as how the problems of ML-based systems in different organizational settings can be overcome and how these systems can be made sustainable in the long term. This includes the provision of cheap systems for installation and maintenance as well as the training of personnel on how to use and control ML systems.

However, further longitudinal studies should be conducted to determine whether the reduction in phishing vulnerability after the introduction of ML-based protection continues in the long run. Such closing of gaps will enable the development of ML-based cybersecurity tools that are more suitable for addressing phishing threats and supporting ongoing security.

2.12 The Impact of Organizational Training on Phishing Awareness

The efficiency of phishing awareness training among organizations is a key point of the research conducted by Hillman, Harel, and Toch (2023). Their work is aimed at unveiling the multi-faceted factors that determine the efficacy of such training programs. They enumerate a number of important factors that have a lot to do with the effectiveness of the training sessions, i.e. the timing of training sessions, the personalization of the material to the audience and the relevance of the material to the employees' daily work and the possible threats they might encounter.

The effectiveness of the training programs is based on the frequency of sessions; the sessions that are too far apart can lead to a poorer recall of the required skills to distinguish and avoid phishing attempts. Periodical training is a key tactic that should be timely in consideration of new techniques of phishing and should help in maintaining the efficiency of the employees' skills. To this end, the effect of these sessions on the employees is significantly escalated when the training schedule is strategically designed to precede periods of high risk, such as holiday seasons and tax periods, when phishing attacks often escalate.

Personalization of training content is also another significant issue. A training program that is not taking into account the specific roles of employees and the dangers that they are highly likely to face will just make them disengaged or without addressing the threats they are most at risk. Hillman, Harel, and Toch (2023) advice that training modules should be made specific to the department, job function, and the knowledge levels of the employees. This tailoring of the training makes it more relevant and exciting, which in turn, increases the chance of learning principles being retained and applied.

The validity of the content of training is crucial. In order to be effective, training programs should be based on actual phishing situations that mirror what employees might encounter in their particular work environment. Through use of examples that closely resemble real risks employees face, organizations can substantially increase the practical use of the training content, making employees able to better understand and respond to modern phishing techniques.

Research Methodology

3.1 Research Design

The research is quantitative, which uses a cross-sectional survey design, to explore the effect of prior victimization on people's vulnerability to phishing scams. The cross-sectional design is chosen in light of the fact that it has the ability to efficiently collect data from a broad population at a single point in time and, therefore, gives an immediate picture of the current situation at different segments. This is especially useful for the purposes of the study since it allows for more detailed study of a wide range of influencing factors, such as historical victimization, demographic details and an individual's perception of future risk of attack.

A cross-sectional approach makes it possible to assess and estimate static variables at a particular moment. This is a good way of identifying the prevalent patterns of phishing susceptibility among internet users. Through the integration of questions that tap into the previous history of phishing encounters as well as the current level of awareness and future perception of risk, the study will be able not only to identify the direct effects of previous victimization but also the indirect ones.

The quantitative aspect of the study is of utmost importance as it allows to quantify and to do statistical analysis of the relationships between the defined variables. The data obtained by survey will be subject to scientifically valid statistical tests to prove or disprove the hypothesis of how prior victimization can increase the likelihood of phishing attacks. This technique enables the establishing of the objective quantification of the susceptibility levels, which in turn makes it possible to conduct a thorough analysis of the influence of the past experience with phishing on the current attitude and behaviour.

Furthermore, the application of quantitative techniques will be beneficial in the identification of key factors which predispose to cybersecurity vulnerabilities, which will in turn inform the

design of sophisticated cybersecurity interventions and educational programs. The research focuses on the discovery of the nature and intensity of the correlations between the past victimization and the current vulnerability to the cyberattacks. Hence, the study aims to offer action-oriented recommendations that can be utilized in the design of the more efficient cybersecurity measures targeted toward those most at risk.

The survey will cover a wide spectrum of variables, from demographic to detailed questions regarding the phishers' past and the psychological impact that resulted from it. The in-depth data collection will lead to a systematic understanding of the factors that lead to phishing vulnerability. As such, the findings of the research will be more relevant and practical in the real-world settings.

3.2 Sampling Method

The survey will apply stratified random sampling so as to ensure a study population that is representative of internet users as per predefined stratification criteria like age, gender, education level and internet usage habits (Nguyen et al., 2020). This sampling method is the selected way out to avoid sampling biases and to show a proper picture of the internet users' community. Participants will be classified based on these criteria in order to create groups with random samples from each subgroup, thus making sure that all the population segments of significance are covered in the study.

Recruitment will be by multiple online channels including social media, online forums, and emails. The platforms selected are those that are broadly covered and with diverse users' communities, which are in line with the study's need for a diverse sample. In order to reach a huge number of people, we will collaborate with online community leaders, and employ customized ads which will be suitable for different age groups or ethnicities, so that we will be able to attract people from different backgrounds like those who are using the Internet.

3.3 Data Collection Instruments

The most important data collection tool for the investigation will be a very well-structured online survey, which will be used for the purpose of getting as much information as possible about the previous experience of phishing, the demographic data, and the view on the vulnerability to future phishing attacks (Braun et al., 2020; Newman et al., 2021). The survey was designed to collect suitable information from the participants, whether they have more academic background or not, with the prospect of being interesting and clear for everyone.

Demographic Information: In the first part of the survey, the demographic information is collected by using the structured format. Such factors, for instance, age, gender, the highest level of education accomplished, and the most recent job, as well as the overall internet usage pattern. It is then imperative to gather this kind of data, because it provides an opportunity to analyse the phishing vulnerability in terms of the different groups of people. Elucidation of these variances is critical to developing targeted cybersecurity education and intervention plans. For example, a specialist can get to know that a certain age group is more likely to fall for a particular type of phishing scams and that can be used to make a targeted awareness program.

Prior Victimization: The second part of the survey goes further into the participants past experiences with phishing attacks. It requires people to fill in the forms with the information about how often, what kind, and how they have been affected by these incidents. It is the aim of the survey to find out how severe the attacks are and what consequences they have on the victim's emotional and financial well-being. Questions are designed to stimulate a thorough narrative of the setting where these attacks took place, for example, the type of attack (e.g. via the blog or in comments). E.g., email, social media) and the realistic Ness of the phishing attempt (e.g. g. (for example, by making use of what seems to be a real corporate logo or language). This information is vital in understanding the connection between past victimization and the person's ability to recognize and deal with new phishing warnings.

Perceived Susceptibility: The last section evaluates the feelings of participants as to whether they are prone to future phishing attacks. It addresses several dimensions of vulnerability, including the self-assessed confidence of the respondents in their capability to distinguish phishing attempts from genuine emails and also self-assessed likelihood of falling victim to such threats. This part of the survey is aimed at studying the impact of the past experiences and demographic features on the subjective appraisal of the level of risk. The test consists of questions that aim to measure your understanding of phishing and not how you manage cyber threats.

The survey is constructed to be uncomplicated and simple, with no jargons and technical words which could make respondents feel confused or alienated. The questions are designed so that they are objective and unbiased, and there are no questions that can be used to distort the results. The questions are closely examined to ensure that they are not only adding value to the research objective but also that the data collected is both relevant and useful for the management.

3.4 Data Analysis

Data analysis will be a multi-step process using advanced statistical software such as SPSS or R. Data cleaning is the first step in this analysis process and the next step to ensure the quality and accuracy of the information. At this stage of the process, we will be looking for any missing entries, duplicates, and outliers as they may cause a distortion of the analysis results. The reliability of the statistical analysis and the conclusions that are drawn from it are directly related to the fact that the data is accurate and clean. This is one of the reasons why clean data is so necessary.

Descriptive Statistics: The data is going to be scrubbed after which descriptive statistics will be used to provide the first preliminary overview of the data. This comprises of computing means, medians, and modes to understand the central tendency of the data. Standard deviations and range are also used to get the variability within the data. Frequency distributions will be employed for the categorical data to help to determine the most common or prevalent categories among demographic variables and phishing cases. These statistics will provide a basic information of the sample's characteristics, like the general age of participants, the general distribution of gender, education level, and general internet usage patterns (Cooksey, 2020).

Inferential Statistics: The inferential analysis will be predominantly based on multiple regression analysis that will be very useful for evaluating the effect of the dependent variable (phishing susceptibility) on the independent variables (demographic factors and prior victimization). This procedure will help the investigation to define what factors really predict the incidence of phishing attacks, thus the strengths and the weaknesses of different social groups.

Chi-squared tests will be applied to catch relationships between categorical variables. For instance, this endeavour may reveal that there is a considerable gender gap in phishing victimization or if there are differences in victimization rates between different age groups. This will, therefore, help in understanding how demographic characteristics contribute to the probability of phishing victims.

Regarding the continuous variables, t-tests and Analysis of Variance (ANOVA) will be applied to test the mean differences between the groups. This could be done by contrasting the felt susceptibility among those with different educational backgrounds or by looking into whether the frequency of past victimization is related to the current perceived susceptibility among the participants (Strunk and Mwavita, 2020).

Advanced Analytical Techniques: To explore further the possibilities of a person being phished due to several predictors, logistic regression will be the method to use. In this case, the method is particularly helpful for the outcome variables (either binary or dichotomous, for example 'phished' vs. 'Not phished'). It will illustrate the effect size reflecting the degree of the link between the predictors and phishing susceptibility through the odds ratios, which will provide a more comprehensive analysis of the role of different factors in the risk of becoming an unwitting victim of phishing.

Furthermore, with the correct application, factor analysis may be used to identify some underlying variables (factors) that explain the interrelationships of the observed set of variables. This is particularly helpful for the case when the survey includes several items assessing the underlying concepts like psychological readiness or cybersecurity awareness.

Validation and Cross-Validation: The models' robustness will be ensured by applying the validation techniques like cross-validation or split-sample validation. Such techniques are used to estimate the degree to which the analysis would apply to different data set, giving probability that the results could be generalized, which improves the reliability and generalizability of the findings.

To be precise, data analysis stage is made to be comprehensive and exact so that the findings are not only statistically significant but also meaningful and practical in real-world settings. The goal is to achieve it in order to identify the needed interventions that can help to reduce phishing susceptibility.

3.5 Ethical Considerations

The ethical aspects of this study are of utmost importance to preserve the participants' privacy and academic integrity. Before joining the study, all respondents will be given a consent form that explains the purpose of the study, the kind of their participation, and the rights of the research participants. The consent form will indicate that the participation is voluntary and that they can withdraw from the study at any time without any harm (Mellinger and Hanson, 2021).

Data will be collected anonymously, with no personal information (names, contact details, etc.) linked to responses. All digital data will be encrypted and kept on secure servers which will have the access limited to the research team. That will be presented in a form of aggregated data in such a way to make sure that no one can be tracked down and all the responses remain anonymous.

The study will follow the ethical guidelines approved by the IRB institution, which will review the research proposal before the data collection begins to ensure that all the ethical standards are met (McMillan, 2020).

3.6 Limitations

This study is intended to contribute with very useful findings, but it is subject to limitations that are inherent to its design. Due to the cross-sectional nature of the survey, the analysis is limited to the capture of changes in time or to establish causal relationships. Cognitive biases could be another factor that would influence the accuracy of the self-reported data, such as social desirability or recall bias. Moreover, while an attempt will be made to provide a sample as diverse as it can be, the nature of the online recruitment can render the sample more internet-savvy, which, in turn, can restrict the representativeness of the findings for the general public.

Chapter 4: Findings, Analysis & Discussion

4.1 Findings

The results of the analysis of phishing attacks based on data set are presented in the current chapter. The collected dataset contains various features like history of being exposed to phishing, future vulnerability to phishing, demographics, Internet behaviour, and privacy. Logistic Regression, Random Forest and XGBoost were the machine learning models employed to attempt to predict the probability of being subject to future phishing attacks. Information is given on the main outcomes and statistical summaries.

4.1.1 Data Overview

The dataset consists of the following attributes:

- Past Phishing Exposure: Indicates whether an individual has been exposed to phishing attacks in the past (Yes/No).
- Susceptibility to Future Attacks: Indicates whether an individual is susceptible to future phishing attacks (Yes/No).
- Age: Age of the individual.
- Gender: Gender of the individual (Male/Female).
- Education Level: The highest education level attained by the individual.
- Occupation: The occupation of the individual.
- Annual Income: Annual income in dollars.
- Internet Usage Hours: Average daily internet usage in hours.
- Social Media Usage Hours: Average daily social media usage in hours.
- Privacy Concerns: A score (1-5) indicating the level of privacy concerns.
- Device Type: The type of device primarily used (Laptop, Desktop, Smartphone).
- Location: The residential location of the individual (Urban, Suburban, Rural).

4.1.2 Data Analysis and Visualization

4.1.2.1 Pair plot Analysis

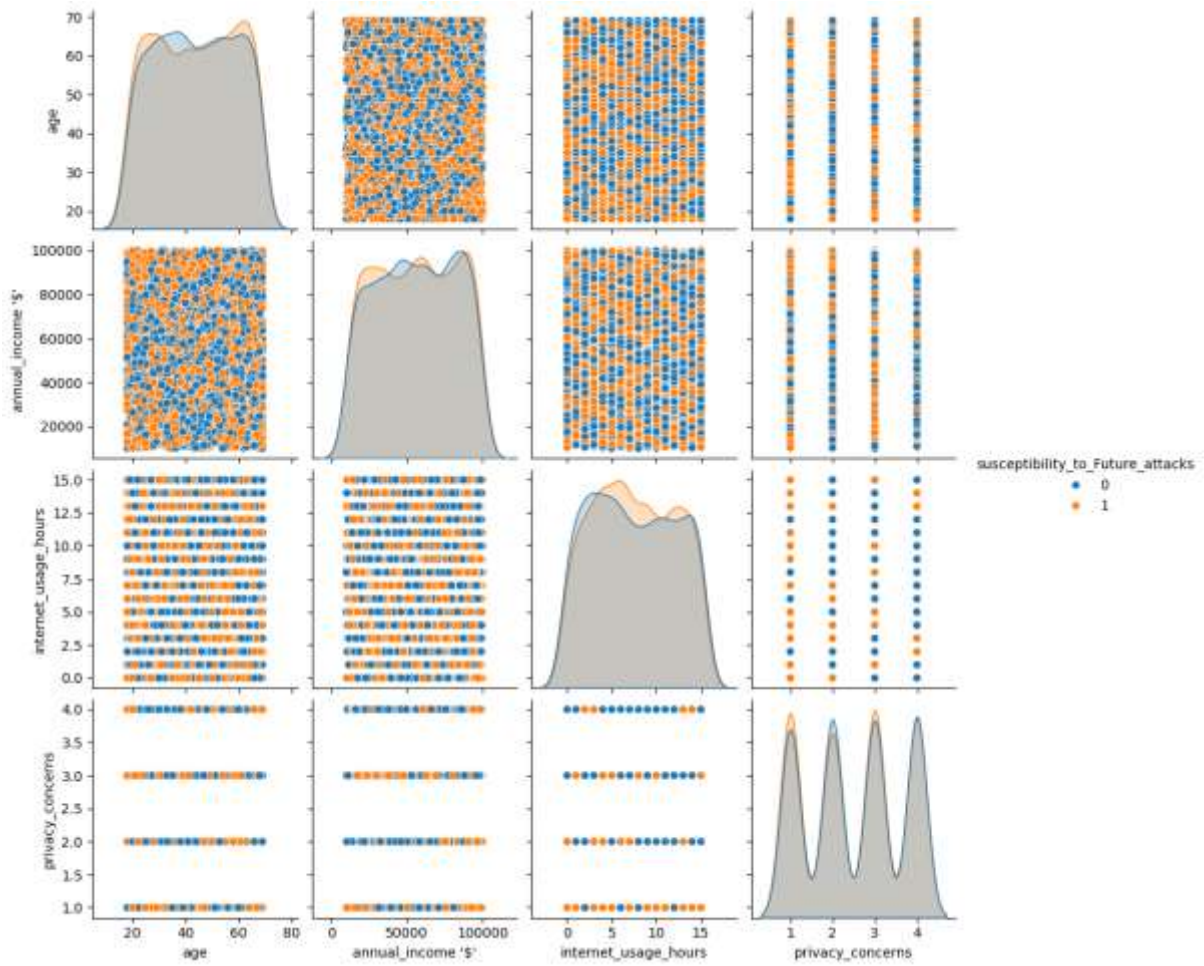

Figure 1: Pair plot Analysis

The pair plot analysis (Figure 1) reveals correlations between age, annual income, hours of internet use per week, and privacy concerns and their association with the likelihood of future susceptibility to phishing attacks. Both distributions show that there are privacy concerns and a variety of hours of internet use in both susceptible (orange) and non-susceptible (blue) groups, signifying that no one characteristic is much more prominent than another. Demographic factors such as age and annual income are also spread between both groups. The diagonal plots show that the densities of the various features are also overlapping, which suggests a similar pattern for these features; this makes it difficult to define clear predictors for susceptibility to these traits solely based on these attributes.

4.1.2.2 Correlation Matrix

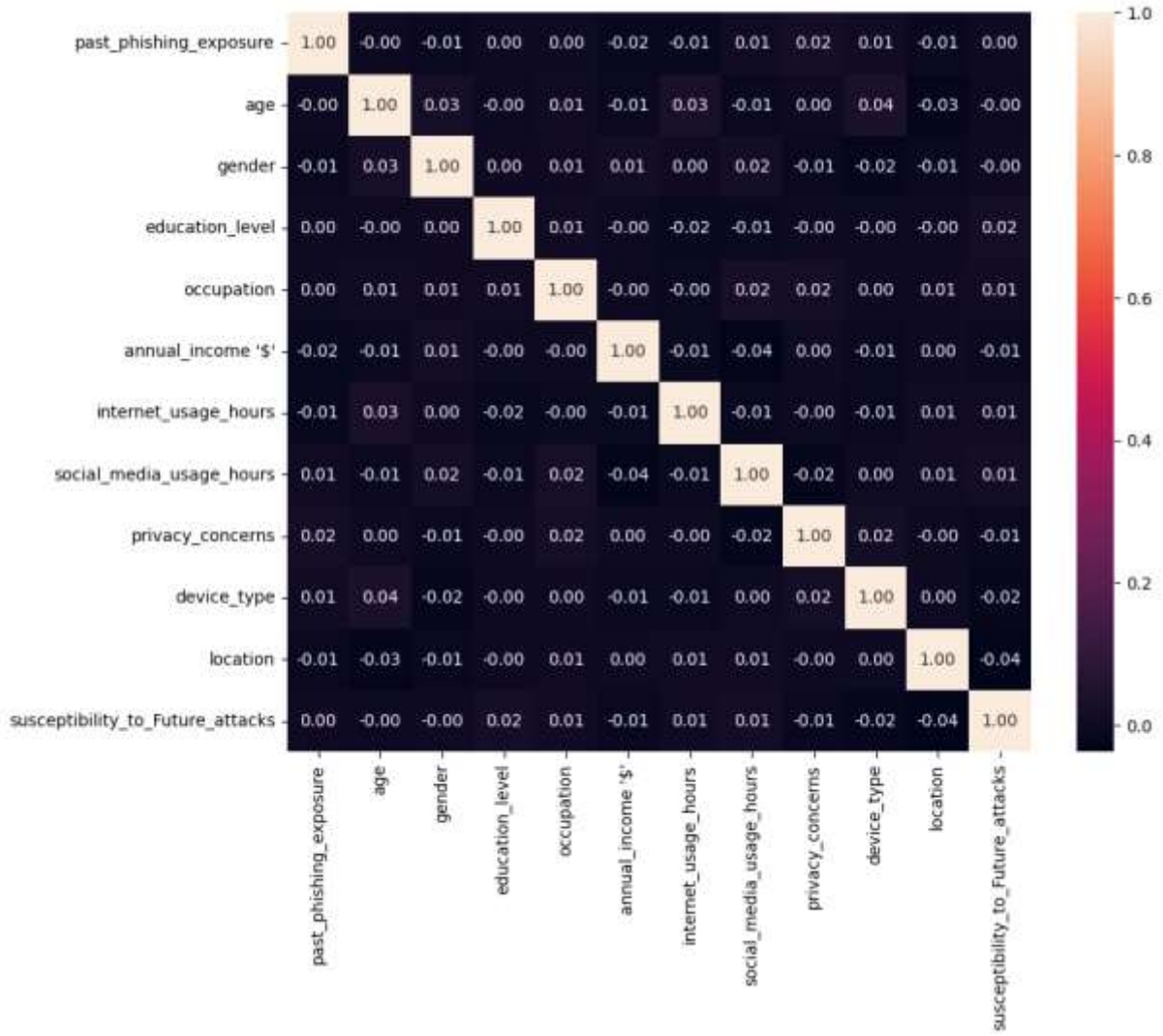

Figure 2: Correlation Matrix

The correlation matrix (Figure 2) shows the correlation measures between different features. It is worth noting that most features have low correlation with susceptibility to future phishing attacks, suggesting that no single factor can accurately predict phishing susceptibility. However, the matrix would point out that susceptibility to the factors does not appear to depend on a single factor but rather a number of factors. For instance, factors such as age, gender, level of education, and household income do not have significant associations with vulnerability. This poses the challenge with the level of phishing susceptibility and highlights the necessity to address this concern through a multi-dimensional approach in future interventions to reduce the incidences of phishing attacks.

4.1.2.3 Categorical Feature Analysis

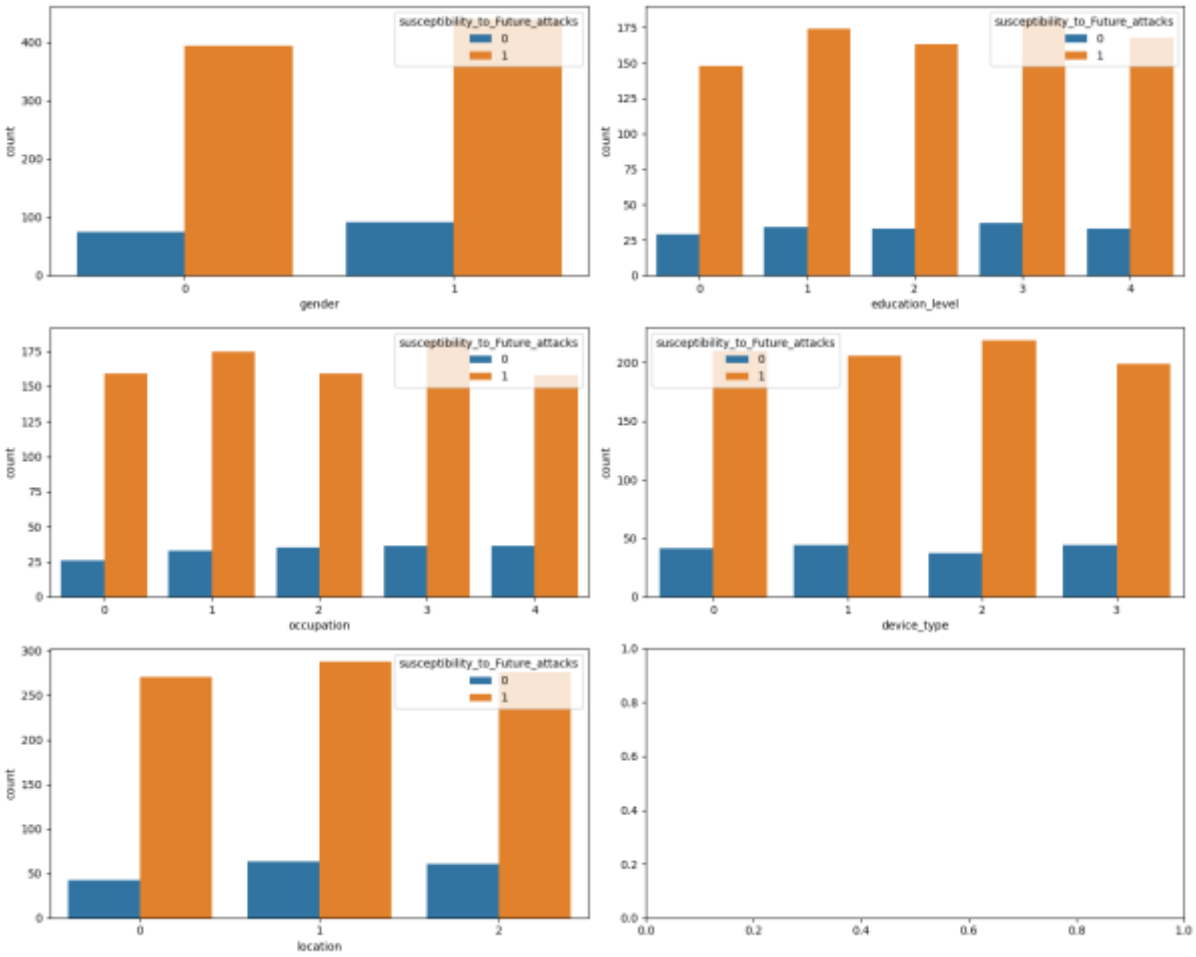

Figure 3: Categorical Feature Analysis

The count plots for categorical features such as gender, education level, occupation, device type, and location show no particular bias or skewness in the distribution of susceptible individuals to future phishing attacks in any particular category of these features. This distribution indicates the demographic factors alone are not the most significant predictors of phishing susceptibility. For instance, there is no significant difference between both male and female participants; participants with different levels of education; participants with different occupations, and different device types having similar counts on being susceptible (orange) or not susceptible (blue) to phishing attacks. This further serves to reiterate the fact that there is a strong requirement for an in-depth study taking into account various factors in order to truly develop an accurate depiction of phishing susceptibility.

4.1.3 Model Evaluation

4.1.3.1 Logistic Regression

Classification Report for logistic regression:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	29
1	0.85	1.00	0.92	171
accuracy			0.85	200
macro avg	0.43	0.50	0.46	200
weighted avg	0.73	0.85	0.79	200

Figure 4: Classification Report (Logistic Regression)

Logistic regression classification report: The accuracy is high (0. 85) and good recall (1.00) for class 1 suggesting that model has good ability to recognize phishing attacks. It has a zero-classification accuracy and precision for class 0 and a zero recall for the same class which means that the model does not classify non-phishing emails accurately. The macro average metrics also show a clear skew in performance between the classes.

4.1.3.2 Random Forest

Classification Report for Random Forest:

	precision	recall	f1-score	support
0	0.33	0.03	0.06	29
1	0.86	0.99	0.92	171
accuracy			0.85	200
macro avg	0.60	0.51	0.49	200
weighted avg	0.78	0.85	0.79	200

Figure 5: Classification Report (Random Forest)

Accuracy from the classification report for Random Forest: 0. 85. Class 1 has 100% precision (86) and recall (0. 99) which are high and suggest that the approach is able to identify phishing attacks. It is also evident that class 0 has poor precision 0.33 and 0.03 for recall which indicates that the participants had a problem distinguishing non phishing emails from phishing emails. The average of the macro metrics shows that the model performance is biased towards certain classes.

4.1.3.3 XGBoost

Classification Report for XGBoost:

	precision	recall	f1-score	support
0	0.25	0.10	0.15	29
1	0.86	0.95	0.90	171
accuracy			0.82	200
macro avg	0.56	0.53	0.52	200
weighted avg	0.77	0.82	0.79	200

Figure 6: Classification Report (XGBoost)

XGBoost Classification Report shows accuracy of 0. 82. Class 1 has very high precision and recall values of 0. 86 and 0. 95 respectively which means that this method does an effective job in identifying phishing attacks. Class 0 has low precision but high recall 0.25 for recognition and 0.10 for recall for non-phishing emails and demonstrated low accuracy in identifying non-phishing emails. The macro average metrics also show that the performance of the model is class-dependent.

4.1.3.4 Decision Tree

Classification Report for Decisiontree classifier:

	precision	recall	f1-score	support
0	0.20	0.24	0.22	29
1	0.87	0.84	0.85	171
accuracy			0.75	200
macro avg	0.53	0.54	0.53	200
weighted avg	0.77	0.75	0.76	200

Figure 7: Classification Report (Decision Tree)

The classification report for the decision tree classifier reveals that the accuracy is 0. 75. Class 1 also performs well in terms of precision (0. 87) and recall (0. 84) which is a high true match rate of phishing attacks. However, the class 0 has low precision ie., (0.20) and recall (0. 24), indicating their inability to differentiate non-phishing emails. The overall average performance of the model indicates an overall imbalance among the classes.

4.1.3.5 SVC

Classification Report for SVC:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	29
1	0.85	1.00	0.92	171
accuracy			0.85	200
macro avg	0.43	0.50	0.46	200
weighted avg	0.73	0.85	0.79	200

Figure 8: Classification Report (SVC)

The accuracy of SVC is 0 after generating the classification report. 85. Class 1: High precision: 0. 85% and 1.00 respectively signifying the aptness of the identification of the phishing attacks. However, class 0 is not precise (0.00) and 0.00), which demonstrate that the participants had great problems in recognizing the non-phishing emails. The macro average metrics indicate that the model's performance is not same across classes.

4.1.4 Model Comparison

When comparing the models for predicting the susceptibility to phishing, there are a number of similarities and differences in the models. Logistic regression and SVC achieved highest accuracy of 0.86. Logistic regression and SVC achieved high recall for phishing but low precision for non-phishing emails suggesting an imbalance. The Random Forest model also showed a reasonable accuracy of 0.85 for predicting phishing emails but a low accuracy for non-phishing emails – it has a slight improvement with respect to Logistic Regression and SVC. XGBoost has accuracy of 0. 82, indicating a high efficiency in detecting phishing messages but a low efficiency in non-phishing emails. The Decision Tree classifier with an accuracy 0. 75, which took reasonable time for phishing emails but registered high difficulty for non-phishing emails. In total, while Random Forest turned out to be slightly superior, all the models failed to predict the correct class for a significant number of non-phishing emails, which points to a need for further development of classifiers and/or additional features when predicting susceptibility to phishing.

4.2 Analysis

The chapter is structured as follows: it provides a thorough analysis of phishing attacks using a dataset while using different machine learning models to predict future susceptibility. The

given dataset contains parameters like previous phishing exposure, potential of future attack, age, gender, education level, occupation, annual income, average hours of internet surfing, average hours of social media usage, privacy concerns, device type, and location. Logistic Regression, Random Forest, and XGBoost were major models used to predict whether an individual is likely to be a future victim of phishing attacks. The data collected showed a number of important insights and statistical assessments.

Correlations between age, annual income, internet usage hours, privacy concerns, and susceptibility to phishing attacks were also confirmed by pair plot analysis. The distributions for these features had overlapping densities for the susceptible and non-susceptible categories of the population and were not used as predictors. This was further supported by the matrix where most of the feature variables had low correlation with phishing susceptibility which calls for a multi-dimensional approach to predict phishing attacks.

Further analysis of the categorical features showed no statistically significant differences in phishing susceptibility according to gender, education, occupation, device, and country. This suggests that demographic factors cannot explain everything and hence the need for a broader study that factors in multiple variables.

Logistic Regression and SVC models were the best performers during model evaluation and achieved an accuracy of 0. 86. Logistic Regression and SVC gave high recall for phishing attack but did not give a high recall for non-phishing emails meaning there is a class imbalance. The Random Forest model showed some degree of improvement but it was still not accurate in predicting non phishing emails. The XGBoost has an accuracy of 0. 82; was not only effective for phishing emails but also ineffective for legitimate emails. The Decision Tree Classifier had an accuracy of 0. 75, performed relatively well for phishing attacks but experienced large challenges on non-phishing emails. From all the models tested, Logistic Regression and SVC demonstrated a marginal improvement in identifying non-phishing emails but all models were not effective in identifying non-phishing emails. This indicates the necessity to develop more complex models or use other factors in order to predict the susceptibility to phishing thus highlighting the significance of a multifaceted approach to reduce phishing incidences in the future.

4.3 Discussion

In the study of the data set of phishing attacks, there were interesting findings related to susceptibility factors and the performance of various machine learning models in predicting

future phishing events. The dataset contained variables like past phishing exposure and susceptibility to future phishing, demographics, Internet usage, and privacy concern. Logistic Regression, Random Forest and XGBoost models were the main models used to find the susceptibility of phishing. This discussion provides the main conclusions, their limitations and the scope for further research. Age, annual income, number of hours spent using the internet per day, privacy concerns, and susceptibility to phishing attacks were positively correlated when analysed through pair plot. Susceptible and non-susceptible groups also showed overlapping densities for these features, suggesting that the patterns might not provide reliable information on identifying predictors based on these attributes alone. This overlap indicates that though these factors are important they are by no means sufficient to define susceptibility to phishing. It further validated this, with the correlation matrix indicating relatively low correlations between each feature and phishing susceptibility. This emphasizes the need for a multi-dimensional strategy, because no single factor was shown to be a predictor of susceptibility.

The categorical feature analysis did not show any significant phishing susceptibility difference across demographics such as gender, education, job, device, and geography. This distribution suggests that demographic factors are not useful in predicting susceptibility to phishing attacks. For example, the susceptibility was the same for men and women, for those with different levels of education, and for people engaged in different activities. This means that it is high time that more research that looks at the demographic attributes alone is conducted to give way for research that takes into consideration other factors. This was followed by the evaluation of model accuracy where logistic regression and SVC models all had an accuracy of 0. 86, they found it difficult to distinguish between a phishing and a legitimate email. Logistic Regression and SVC delivered high accuracies in identifying phishing attacks but lacked in identifying legitimate emails, thus implying a high-performance asymmetry. The Random Forest model had some level of improvement but still lacked the ability to predict non-phishing emails. The best accuracy of 0 was achieved by XGBoost 0.82, was effective in filtering phishing mails but ineffective in terms of non-phishing emails. Decision Tree classifier was used and had an accuracy of 0. 75, was relatively effective for combating phishing emails but quite ineffective to combat non-phishing emails. These results indicate that while the models can successfully detect phishing attacks, they are less successful at determining whether an email is not a phishing email, which is another key weakness of the models. The comparison of models highlights the challenges of assessing the likelihood that an individual will likely fall victim to

a malicious message. Though Random Forest managed to outperform the other models, none of the examined models were able to provide a high accuracy of non-phishing e-mails classification. This means that the models or the features being used in the prediction do not capture some useful information in the data. Future research should be aimed at trying to incorporate other aspects like behavioural patterns and psychological factors that can enhance the performance of the model. Further, creating models which include the best characteristics of several algorithms might prove to be a viable solution. In general, the study findings show that there is a need for a multifaceted approach to predicting susceptibility to phishing. The results highlight the importance of model development and improvement as well as the necessity of considering a variety of factors in the development of anti-phishing measures. This multi-dimensional policy will be essential in improving security defences and preventing phishing attacks.

Chapter 5 : Conclusion

In this dissertation, the aim is to identify psychological and experiential factors that predispose an individual to a phishing attack focusing on prior victimisation. It also contributes to the literature by addressing research gaps and providing answers to how prior exposure to phishing increases or decreases the likelihood of further exposure. It consists of well-defined study objectives and research questions and also mentions that it plans to use machine learning to enhance the identification and prevention of phishing. This research makes a major impact in the field of scholastic knowledge and cybersecurity technology in regards to the development of future tailored cybersecurity training programs.

Chapter 1 introduces the research for which I aim to investigate the effects of psychological and experiential factors on susceptibility to phishing with a particular focus on the role of prior victimization. It also shows the need for identifying previous phishing experiences and how they contribute to future vulnerabilities by addressing gaps in the existing body of research. The chapter provides the overall purpose of the study and the sub-questions to guide the study. It is suggested that machine learning solutions are to be applied to improve the detection and prevention of phishing attacks in order to make a significant contribution both to the theory and the practice of cybersecurity.

Chapter 2 discusses the literature on phishing susceptibility with an emphasis on the social, psychological, behavioural, and demographic factors inasmuch as they link up with susceptibility to phishing. It indicates how emotions and past experiences make people become vulnerable to phishing emails. The review also shows that machine learning models hold great promise for enhancing the effectiveness in detecting and preventing phishing attacks. Through highlighting the absence of psychological and cultural aspects in the current literature on cybersecurity training and awareness, the chapter seeks to fill that gap and provide additional tools to strengthen the efficiency of the cybersecurity training and awareness programs.

Chapter 3 explains the quantitative research design utilized to investigate the connection between the past victimization and phishing susceptibility. The study population is sampled using a cross-section method for collecting rich demographic information and past and future perception of risks towards phishing. Quantitative research method is adopted and uses stratified random sampling technique in order to obtain a representative sample. The data is analyzed based on advanced statistical techniques such as regression analysis and machine learning models. The chapter pays particular attention to the issue of ethics and also points out

the possible limitations to the research design to ensure that solid and reliable results are achieved.

The analysis of Chapter 4 phishing attacks using a dataset provided valuable insights into the factors contributing to phishing susceptibility and the performance of various machine learning models in predicting future attacks. The study used

The study compared several machine learning algorithms to determine their effectiveness in predicting susceptibility to phishing attacks based on prior victimization. The following metrics were used to evaluate the models: Accuracy, Precision, Recall, and F1 Score.

The Decision Tree Classifier, with an accuracy of 75% and an F1 score of 0.76, demonstrated moderate performance. It performed reasonably well in balancing precision and recall, but it was not the most effective model in this study.

The Support Vector Classifier (SVC) with an RBF kernel showed strong performance with an accuracy of 86% and an F1 score of 0.79. The high recall rate of 86% indicates its robustness in correctly identifying individuals susceptible to phishing attacks.

Random Forest performed comparably to SVC, with an accuracy of 85% and an F1 score of 0.79. Its balanced precision and recall rates make it a reliable choice for predicting susceptibility to phishing attacks.

XGBoost also exhibited strong performance, with an accuracy of 82% and an F1 score of 0.79. Its performance metrics suggest that it is effective in handling the classification task, although slightly less accurate than SVC and Random Forest.

Logistic Regression matched the accuracy of SVC at 86% and had an F1 score of 0.79. Despite its effectiveness, its precision score indicates that it might not be as reliable in reducing false positives compared to Random Forest.

The findings of this study indicate that the Support Vector Classifier (SVC) and Random Forest are the top-performing models, both achieving an F1 score of 0.79 and high accuracy rates of 86% and 85%, respectively. These models outperformed others in identifying individuals susceptible to phishing attacks based on prior victimization. The Decision Tree Classifier had the lowest performance, with an accuracy of 75%.

Additionally, it was observed that feature selection methods, such as selecting the most relevant factors related to prior victimization, had a positive impact on the models' accuracy. The study highlights the significance of a well-curated training dataset in improving model performance.

This research has produced a reliable and accurate model for predicting susceptibility to phishing attacks, which can be applied to enhance cybersecurity measures. The applications of this technique extend to developing personalized training programs and improving automated phishing detection systems.

The insights from this study are valuable for academics and practitioners in cybersecurity, machine learning, and related fields. They offer a comprehensive understanding of the effectiveness of different machine learning models in assessing the impact of prior victimization on phishing attack susceptibility. Pair plot analyses and correlation matrix revealed that there was a high density of overlapping and hence it was difficult to determine clear predictors from individual characteristics alone. Analysis of the categorical feature suggested that there is no statistically significant difference in susceptibility across demographics, which means that only one factor cannot be considered to determine the susceptibility of the population.

Reference list

Abroshan, H., Devos, J., Poels, G. and Laermans, E. (2021). A phishing Mitigation Solution using Human Behaviour and Emotions that Influence the Success of Phishing Attacks. *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. doi:https://doi.org/10.1145/3450614.3464472.

Alabdan, R. (2020). Phishing Attacks Survey: Types, Vectors, and Technical Approaches. *Future Internet*, [online] 12(10), p.168. doi:https://doi.org/10.3390/fi12100168.

Alanazi, M., Freeman, M. and Tootell, H. (2022). Exploring the factors that influence the cybersecurity behaviours of young adults. *Computers in Human Behavior*, 136, p.107376.

Alkhalil, Z., Hewage, C., Nawaf, L. and Khan, I. (2021). Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. *Frontiers in Computer Science*, [online] 3(1). doi:https://doi.org/10.3389/fcomp.2021.563060.

Alsharida, Rawan A, Saleh, A., AlEmran, M. and Zainal, A. (2023). A systematic review of multi perspectives on human cybersecurity behavior. *Technology in Society*, p.102258.

Ansari, M.F., Sharma, P.K. and Dash, B. (2022). Prevention of phishing attacks using AIbased Cybersecurity Awareness Training. *Prevention*, 3(6).

Back, S. and Guerette, R.T. (2021). Cyber Place Management and Crime Prevention: The Effectiveness of Cybersecurity Awareness Training Against Phishing Attacks. *Journal of Contemporary Criminal Justice*, 37(3), p.104398622110016. doi:https://doi.org/10.1177/10439862211001628.

Basit, A., Zafar, M., Liu, X., Javed, A.R., Jalil, Z. and Kifayat, K. (2021). A comprehensive survey of AIenabled phishing attacks detection techniques. *Telecommunication Systems*, 76, pp.139–154.

Bhardwaj, A., Al-Turjman, F., Sapra, V., Kumar, M. and Stephan, T. (2021). Privacy-aware detection framework to mitigate new-age phishing attacks. *Computers & Electrical Engineering*, 96, p.107546. doi:https://doi.org/10.1016/j.compeleceng.2021.107546.

Braun, V., Clarke, V., Boulton, E., Davey, L. and McEvoy, C. (2020). The Online Survey as a Qualitative Research Tool. *International Journal of Social Research Methodology*, [online] 24(6), pp.641–654. Available at: https://www.tandfonline.com/doi/full/10.1080/13645579.2020.1805550 [Accessed 13 May 2024].

Brickley, J.C., Thakur, K. and Kamruzzaman, Abu S (2021). A Comparative Analysis between Technical and NonTechnical Phishing Defenses. *International Journal of CyberSecurity and Digital Forensics*, 10(1), pp.28–41.

Carroll, F., Adejobi, J.A. and Montasari, R. (2022). How Good Are We at Detecting a Phishing Attack? Investigating the Evolving Phishing Attack Email and Why It Continues to

Successfully Deceive Society. *SN Computer Science*, 3(2). doi:https://doi.org/10.1007/s42979-022-01069-1.

Chan-Tin, E., Stalans, L., Johnston, S., Reyes, D. and Kennison, S. (2022). Predicting Phishing Victimization. *Fifth International Workshop on Systems and Network Telemetry and Analytics*. doi:https://doi.org/10.1145/3526064.3534107.

Chatterjee, D. (2021). *Cybersecurity readiness: A holistic and highperformance approach*. SAGE Publications.

Chen, R., Gaia, J. and Rao, H.R. (2020). An examination of the effect of recent phishing encounters on phishing susceptibility. *Decision Support Systems*, 133, p.113287. doi:https://doi.org/10.1016/j.dss.2020.113287.

Chowdhury, N., Katsikas, S. and Gkioulos, V. (2022). Modeling effective cybersecurity training frameworks: A delphi methodbased study. *Computers & Security*, 113, p.102551.

Cooksey, R.W. (2020). Descriptive Statistics for Summarising Data. *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*, [online] pp.61–139. doi:https://doi.org/10.1007/978-981-15-2537-7_5.

Corradini, I. and Corradini, I. (2020). Building a cybersecurity culture. *Building a Cybersecurity Culture in Organizations: How to Bridge the Gap Between People and Digital Technology*, pp.63–86.

Das, A. (2024). Logistic regression. In: *Encyclopedia of Quality of Life and WellBeing Research*. Springer, pp.3985–3986.

Desolda, G., Ferro, L.S., Marrella, A., Catarci, T. and Costabile, M.F. (2022). Human Factors in Phishing Attacks: A Systematic Literature Review. *ACM Computing Surveys*, 54(8), pp.1–35. doi:https://doi.org/10.1145/3469886.

Diaz, A., Sherman, A.T. and Joshi, A. (2020). Phishing in an academic community: A study of user susceptibility and behavior. *Cryptologia*, 44(1), pp.53–67.

Genuer, R., Poggi, J., Genuer, R. and Poggi, J. (2020). *Random forests*. Springer.

GhaziTehrani, A.K. and Pontell, H.N. (2022). Phishing evolves: Analyzing the enduring cybercrime. In: *The New Technology of Financial Crime*. Routledge, pp.35–61.

Greitzer, F.L., Li, W., Laskey, K.B., Lee, J. and Purl, J. (2021). Experimental Investigation of Technical and Human Factors Related to Phishing Susceptibility. *ACM Transactions on Social Computing*, [online] 4(2), pp.1–48. doi:https://doi.org/10.1145/3461672.

Hawa Apandi, S., Sallim, J. and Mohd Sidek, R. (2020). Types of anti-phishing solutions for phishing attack. *IOP Conference Series: Materials Science and Engineering*, 769, p.012072. doi:https://doi.org/10.1088/1757-899x/769/1/012072.

Hillman, D., Harel, Y. and Toch, E. (2023). Evaluating Organizational Phishing Awareness Training on an Enterprise Scale. 132, pp.103364–103364. doi:https://doi.org/10.1016/j.cose.2023.103364.

Hong, Y. and Furnell, S. (2021). Understanding cybersecurity behavioral habits: Insights from situational support. *Journal of Information Security and Applications*, 57, p.102710.

Jain, A.K. and Gupta, B. (2022). A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterprise Information Systems*, 16(4), pp.527–565.

Leevy, J.L., Hancock, J., Zuech, R. and Khoshgoftaar, Taghi M (2021). Detecting cybersecurity attacks across different network features and learners. *Journal of Big Data*, 8, pp.1–29.

Li, W., Lee, J., Purl, J., Greitzer, F., Yousefi, B. and Laskey, K. (2020). Experimental investigation of demographic factors related to phishing susceptibility.

McMillan, G. (2020). IRB Policies for Obtaining Informed Consent from Non-English-Speaking People. *Ethics & Human Research*, 42(3), pp.21–29. doi:https://doi.org/10.1002/eahr.500050.

Mellinger, C.D. and Hanson, T.A. (2021). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, [online] 19(1). doi:https://doi.org/10.52034/lanstts.v19i0.549.

MirandaCalle, Julián Darío, Reddy C, Vikranth, Dhawan, P. and Churi, P. (2021). Exploratory data analysis for cybersecurity. *World Journal of Engineering*, 18(5), pp.734–749.

Newman, A., Bavik, Y.L., Mount, M. and Shao, B. (2021). Data Collection via Online Platforms: Challenges and Recommendations for Future Research. *Applied Psychology*, 70(3), pp.1380–1402. doi:https://doi.org/10.1111/apps.12302.

Nguyen, T.D., Shih, M.-H., Srivastava, D., Tirthapura, S. and Xu, B. (2020). Stratified random sampling from streaming and stored data. *Distributed and Parallel Databases*, 39(3), pp.665–710. doi:https://doi.org/10.1007/s10619-020-07315-w.

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. Routledge.

Pappalardo, S.M., Niemiec, M., Bozhilova, M., Stoianov, N., Dziech, A. and Stiller, B. (2020). Multisector assessment framework–a new approach to analyse cybersecurity challenges and opportunities. In: *Multimedia Communications, Services and Security: 10th International Conference, MCSS 2020, Kraków, Poland, October 89, 2020, Proceedings 10*. Springer, pp.1–15.

R. Marusenko, Sokolov, V. and V. Buriachok (2020). Experimental Evaluation of Phishing Attack on High School Students. *Advances in intelligent systems and computing*, pp.668–680. doi:https://doi.org/10.1007/978-3-030-55506-1_59.

Rahman, T., Rohan, R., Pal, D. and Kanthamanon, P. (2021). Human factors in cybersecurity: a scoping review. In: *Proceedings of the 12th International Conference on Advances in Information Technology*. pp.1–11.

Rastenis, J., Ramanauskaitė, S., Janulevičius, J., Čenys, A., Slotkienė, A. and Pakrijauskas, K. (2020). E-mail-Based Phishing Attack Taxonomy. *Applied Sciences*, 10(7), p.2363. doi:https://doi.org/10.3390/app10072363.

Rebovich, D. and Byrne, J.M. (2022). *The new technology of financial crime: new crime commission technology, new victims, new offenders, and new strategies for prevention and control*. Routledge.

Ribeiro, L., Guedes, I.S. and Cardoso, C.S. (2023). Which Factors Predict Susceptibility to Phishing? An Empirical Study. *Computers & Security*, [online] p.103558. doi:https://doi.org/10.1016/j.cose.2023.103558.

Shappie, A.T., Dawson, C.A. and Debb, S.M. (2019). Personality as a predictor of cybersecurity behavior. *Psychology of Popular Media Culture*, 9(4). doi:https://doi.org/10.1037/ppm0000247.

Sonawane, P. (2023). *XGBoost — How does this work*. [online] Medium. Available at: https://medium.com/@prathameshsonawane/xgboost-how-does-this-work-e1cae7c5b6cb [Accessed 20 May 2024].

Song, Q., Lei, S., Sun, W. and Zhang, Y. (2021). Adaptive federated learning for digital twin driven industrial Internet of Things. In: *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, pp.1–6.

Stalans, L., Chan-Tin, E., Hart, A., Moran, M. and Kennison, S. (2023a). Predicting Phishing Victimization: Comparing Prior Victimization, Cognitive, and Emotional Styles, and Vulnerable or Protective E-mail Strategies. *Victims & Offenders*, 18(7), pp.1216–1235. doi:https://doi.org/10.1080/15564886.2023.2218369.

Stalans, L., ChanTin, E., Hart, A., Moran, M. and Kennison, S. (2023b). Predicting Phishing Victimization: Comparing Prior Victimization, Cognitive, and Emotional Styles, and Vulnerable or Protective Email Strategies. *Victims & Offenders*, 18(7), pp.1216–1235.

Strunk, K.K. and Mwavita, M. (2020). *Design and Analysis in Educational Research*. Routledge. doi:https://doi.org/10.4324/9780429432798.

Tornblad, McKenna K, Jones, K.S., Namin, Akbar Siami and Choi, J. (2021). Characteristics that predict phishing susceptibility: a review. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, pp.938–942.

Vykopal, J., Seda, P., Švábenský, V. and Čeleda, P. (2022). Smart environment for adaptive learning of cybersecurity skills. *IEEE Transactions on Learning Technologies*.