Determining Demograpic and Socio-ecomonic Features of Reported Cybercrime Victims

N. Aardse[1], J. Bakker[2], A. Etaoil[3], D. Stenavtten[4] and R. Verhulst[5]
Faculty IT & Design
The Hague University of Applied Sciences
The Hague, the Netherlands
Email: [1]n.m.aardse@student.hhs.nl, [2]j.j.bakker@student.hhs.nl, [3]a.etaoil@student.hhs.nl,
[4]d.stenavtten@student.hhs.nl, [5]r.h.verhulst@student.hhs.nl

Abstract—Digital safety is beginning to play an increasingly important role in today's Dutch society, cybercrime has a major impact in the field. But which groups are actually more vulnerable to cybercrime? This exploratory study tries to answer this question by applying Machine Learning classification algorithms against police and population datasets. By letting these algorithms search for high scoring socio-economic and demographic features for cybercrime victims, it can be determined which features correlate with cybercrime victims. First, the algorithms will be trained on sampled datasets. These sampled datasets contain cybercrime victims and non-victims. After training the algorithms, the statistical method Chi-square was also applied. Then, all the true positives (correctly predicted cybercrime victims by the algorithms) were taken from all the results of the Machine Learning algorithms. Using these true positives and the same baseline, all the algorithms will be trained again to see if the scores of the results are improving. After training and comparing the results of the models against each other, the varying results from the models limits any claims to be made around any features. But the score of the models do perform better than random, this in turn suggest that there may be something to be found here, maybe with less explainable algorithms like neural networks or by improving the quality of the input data.

Index Terms—Cybercrime victims, classification, supervised machine learning, demographic and socio-economic features

1 Introduction

In this day and age cybercrime has an ever growing impact on society, and the rapid digitalization of society doesn't suggest a change in this trend. Cybercrime has climbed to the top in the National Security Strategy of many EU states [1]. Although as stated by [2] hard statistics on losses from online crime are hard to come by in most countries. Depending on the literature, the figures can heavily vary from billions to trillions. This is caused by many different factors such as security companies inflating the numbers and banks being hesitant to release any information at all. Despite the varying numbers and lack of solid statistics on the subject, the magnitude and impact of cybercrime is undeniable.

Although proposals favouring improvements and higher security, it does not appear that today's societies will choose to make the internet safe at all costs [3]. It is therefore of greater importance to focus on protecting potential victims from the risks specific to virtual environments, and to delve into the causes of the imprudent use of ICT and the internet. "The question lies, therefore, in analysing the vulnerability of victims and correcting shortcomings in their use of ICT—specifically, their inclination not to consider the risks of their actions." [3].

The public is well acquainted with cyberattacks, but how they handle or respond to these threats varies based on the individual. This statement further reinforces the thesis that there are groups in society that are more likely to being victimized by cybercrime [4].

Which motivates action being taken from both ends of cybercrime, where one side is the perpetrator and the other is the victim. Is the victim group completely random, or are there any groups in society who run a higher risk of being victimized by cybercrime? Identifying these groups could be of great significance, and therefore this study tries to determine which socio-economic and demographic features associate most positively with reported cybercrime victimization of Dutch adults in 2016.

Hence, the goal is determining whether victimization is correlated to socio-economic and demographic features. To extract features that correlate positively with whether a person who has reported being a victim of cybercrime, following classification algorithms were trained to classify cybercrime and non-crime victims: *Random Forest, Support Vector Machine, Logistic Regression and Gradient Boosted Decision Tree*. Then, features which these algorithms deemed useful to distinguish between the groups were analyzed. Further experiments were conducted by clustering all the individuals who were correctly classified by all the algorithms with t-SNE. This was done to gain further insight into the individuals who were prone to be classified by the algorithms.

2 Related Works

Previous work related to the application of machine learning for cyber victim classification is still quite limited. [5] conducted a study regarding five different psychological traits of cybercrime victims in the Netherlands using logistic regression. In comparison, this study will focus on the socio-economic and demographic characteristics of the victims. Thus making it hard to find a good comparison for the results of this study in the field of machine learning.

However statistical studies have had similar goals as this study, [6] found statistically significant differences between gender as a demographic feature when it came to password generation, proactive awareness and software updating. All of which can impact vulnerability.

[7] and [8] analysed demographic features and phishing: both found correlation between the age group 18-25 and being susceptible to phishing attacks. In comparison [9] claims demographic features not to be sufficient to predict phishing attack susceptibility.

Other studies points towards respondents to be more susceptible to online purchase fraud, and older respondents were more often affected by online banking fraud. They tie these findings to a difference in online activity between the victims. They found a weak correlation between victimization and gender and economic features [10].

3 Experiments

3.1 Data

The data used in this article is supplied by SN (Statistics Netherlands; Dutch: Centraal Bureau voor de Statistiek) and consists of two different datasets regarding dutch citizens:

1. Social Statistical Database (SSD) of 2016
The SSD is a dataset that contains an, by SN pre-selected, amount of demographic and socio-economic factors of citizens who reside in the Netherlands in 2016. These factors describe: gender, age, level of education, household type, migration background, income, debt rescheduling, social benefit and city density.

2. Police Records Database (PRD) of 2016
The PRD is a dataset that contains all reported crime from the year 2016. The file contains information about the reported crime and if the reporter was victimized by the crime. All victimized individuals, whether they reported the crime themselves or not, are selected.

Traditional crime, cybercrime and non victims
Within the PRD, there is a difference made between those victimized by cyber- and traditional crime. SN previously conducted an experiment where a text classification algorithm predicts if a reported crime is a cybercrime. In this paper, those prediction results are seen as ground truth and are used to derive cybercrime from traditional crime. The algorithm predicts crimes where both IT is both used as a means and as a target of a crime. On the other hand, traditional crime is defined as all reported crime excluding cyber crime prediction and inconclusive prediction. Due to limitation of the algorithm, those inconclusive crime prediction were excluded from the dataset, since the algorithm could not give a verdict. Our last group consists of all the individuals SSD that aren't represent in the PRD.

Linking the two datasets
Every natural person in the Municipal Population Register (MPR) collected by SN is attributed with a Record Identification Number (RIN). All observations in the PRD provided with a RIN are joined with the SSD file. Only the columns indicates that a person has been victimized by cyber- or traditional crime, are used from the PRD, because all other features are duplicates of SSD data or describe the crime itself. This creates a new dataset consisting of unique individuals with features from the SSD and a feature that indicates whether they have/are reported being victimized by cybercrime, traditional crime or neither in 2016.

Data cleaning
All citizens under 18 years were removed due to the fact that they are underrepresented in the data. This can possibly be explained seeing most parents will report the crime for their children. The majority of the current dataset consist of low-cardinality categorical features, therefore one-hot encoding was applied to each categorical variable in the dataset (Dorogush, A. V., 2018). For certain discrete variables with a big range, such as income, bins (partitions) with regular intervals were created. On the one-hot encoded dataset, correlation coefficients were calculated for each encoded feature. From feature pairs that have perfect positive correlation, one of the features was removed due to redundancy (Guyon I., 2003). Although perfect negative correlated features contain the inverse of the information (thus reduced), they were kept because of practical reasons when showing positive feature importance. The after cleaning the dataset contains 100 one-hot encoded features.

Stratified Samples
Since the distinction is made between cybercrime victims and non-cybercrime victims the dataset need a appropriate sample of non-victims. This sample was drawn from the population file (SSD) where it is known the that person doesn't appear in the PDR. The sample is stratified with following features: gender, age, highest achieved level of education, income, household type. This results in 1080 group combination. The stratification labels were selected because they were most descriptive of the population. Stratification allows the sample to reflect the origin dataset in distribution, this is important to be able to make any claims regarding the Dutch population.

3.2 Classification

Following classification models were applied in this experiment: random forest, support vector machine and logistic regression. They were used to classify people either as a cybercrime victim or not. To produce results from the classification, the features deemed most positively correlated with cybercrime victims were extracted. Two different samples of the data was used in the experiments. The first sample was a randomly balanced dataset (n = 85 673) from both classes *(Classified cybercrime victims and non-victims)*. The second was a combination of all correctly predicted cybercrime victims from all the models, with an equal random sample of non victims. To clarify, a person who was correctly predicted by x-models occurred x-times in this dataset (n = 123 149). This was done in an attempt to improve the classification results by only using accurately predicted cybercrime victims, thus increasing the models chances of differentiating between the classes.

3.3 Excluded Models

Apart from the above mentioned models, following models were implemented. However, they did not perform well with the data at hand.

Low scoring classification models.

Tests were conducted using a perceptron and K-nearest neighbour (K-NN) algorithms with different configurations. The models didn't perform well with the data, they barely got an accuracy score above 50 percent which is the score of the average random model and were thus excluded. Aba-Boosted decision trees were also tested, since they performed worse than random forest and gradient boosted trees they were deemed unnecessary.

Neural-networks

Initially some research was done around the implementation of neural networks for classification. Though these models generally perform well in the task of classification, the process of extracting the feature importance of them is very complex due to the "black-box" nature of neural networks. Thus, they were deemed fruitless for this study.

Unsupervised Learning

At first, Principal Component Analysis (PCA) and K-means were used as a method to reduce dimensionality within the data and allocate clusters of cybercrime victims. Sadly, due to the nature of dimensionality reduction, information of features could not be retained. This made these algorithms unsuitable for the purpose of this research

3.4 Model explanation and motivation

Logistic Regression (LR)

Logistic Regression (LR) is a supervised machine learning algorithm, which classifies the given data.This data contains at least one independent label that determines the binary outcome. By calculating the probability of a certain value to be classified as a 1 (true value), and by using a certain threshold value, the model classifies a certain datapoint as a 1 when the probability is higher than the threshold value. Otherwise it will be classified as a 0.

Random Forest (RF)

Random forest is a parallel ensemble learning model, that means the model. It consists of n amount of individual decision trees *(see appendix: A, visualization of one decision tree)*. Then a majority vote approach is used, meaning the majority of the best performing trees are used. This method was applied because of its robustness against overfitting through the use of many estimators, which was necessary with the amount of features the data set contained. [11]Conducted a study to benchmark classification algorithms, and the random forest was one of the top performer in the majority of the tests. The relevant hyperparameters to tune for this model are: n_estimators, max_depth, max_features and max_leaf_nodes.

Gradient Boosted Decision Tree (GBDT)

Just as previous method GBDT is an ensemble learning model, but in comparison to random forest it is sequential instead of parallel. The trees in GBDT can vary in size, the so called "weak learners" can be reduced to just one splitt. [12] Claims it to be one of the best performing algorithms when it comes to both multi-label and binary classification.

Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning approach for solving supervised classification and regression problems, or more by learning from examples. A support vector machine (SVM) uses a kernel method, to map the data with a non linear transformation to a higher dimensional space, to try to find a linear separatic space between the two classes [13]. This method has been chosen because of its ability of capturing more complex relationships between data points, compared SVM with another classification algorithm that is similar (Logistic Regression), and the SVM algorithm performed better in many situations.

3.5 Chi-squared feature selection

Chi-square is a statistical test to describe the relationship between two variables. The test can be performed on either a known or unknown distribution of the variables. First, a null-hypothesis (H0) is formed which states that the two variables are independent. If the statistical significance boundary is exceeded, then the null-hypothesis is rejected and the alternative hypothesis (H1), which states that the two variables are dependent, is accepted. Determining the exceedance of the boundary is done by creating a 2x2 contingency table where the observation (O) of every variable combination (in this case 4) is counted. Through the use of probability expected (E) values are determined after which both O and E are used calculate the chi-square score with to use of the following formula [14]:

$$\tilde{\chi}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

Chi-square test is a common used method for selecting features [15] and has been used in multiple different applications such as: diagnosing diseases [16], text classification [17] and network intrusion detection [18]. Within the context of this study, chi-squared is used to calculate which features are most positively depend toward those reported being victimized by cybercrime. In order to achieve this, features where noted that met the following two conditions: The total expected value of the contingency table is bigger then five [14] The observed amount is bigger than the expected value for those in possession of the feature and reported being victimized by cybercrime.

After the calculation process, a list is generated and used as a means of statical reference towards the list of feature importance. Beside that, the chi-square test will be used in combination with T-SNE. These plots are used identify groups (a section of the plot where multiple features are represented) that fall into the clusters where cybercrime is represented. The chi-squared method is then used on multiple features at once to determine whether those groups are significantly more present as cybercrime victims then non-victims.

3.6 t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a technique that visualizes high dimensional data by giving a location to each datapoint in a multi-dimensional space. The technique is a variation of Stochastic Neighbor Embedding but is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map[19].*"t-SNE is better than existing techniques at creating a single map that reveals structure at many different scales"* [19]. t-SNE is used to cluster different groups of people, based on their most prominent features as predicted by multiple models. By clustering people based on these features, it could be possible to see if they have other features that are similar to each other, that are also clustered by t-SNE.

3.7 Evaluation

Evaluation metrics

The performance of all the models were measured by F1-Score, accuracy, precision and recall which are different performance metrics to evaluate classification models based on *True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)*. Below is the formulas for calculating these scores.

$$Accuracy = TP + TN/(TP + FP + TN + FN)$$

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 = 2 * ((Precision * Recall)/(Precision + Recall))$$

F1 was deemed most useful because the models ability to distinguish between classes whether it was True Positive or True Negative in a balanced fashion was important, to be able to make any claims about feature importance produced by said models.

Hyperparameter tuning

For hyper parameter tuning the random search approach was applied. Random search was chosen over grid search to optimize run-time and use of computational resources. Random search over grid search, have been proven to perform just slightly worse than grid search over several different algorithms on different datasets, and is more efficient because not all hyperparameters are equally important to tune[20]. Since the project was limited by computational power, this approach was the best choice based on computation time contra results.

	Acc.	Prec.	Recall	F1
RF	60%	71%	58%	64%
SVM	60%	67%	59%	63%
LR	60%	63%	59%	61%
DTGB	60%	65	59%	62%

(Table: 1, Model Scores)

4 Results

The models almost have the same scores *(See table: 1, Model Scores)*, the RF model performs slightly better in comparison to the other models. But since it is a minor difference, a critical eye is needed to determine which algorithm is suited the best for determining cybercrime victims. For this, it is important to know which problems are encountered and what the goal of this study is.

For this study, it is essential to know which algorithm is the best to classify cybercrime victims. Therefore, it is important to know which metrics are the most useful. The most relevant scores will come from the combination of the precision and recall metric represented as F1, since they show how many and precise the true positives (correctly cybercrime victims) are predicted. However, since all these scores are quite close to each other, and the data consists of no excluded values, the decision is made to rerun all the models and to run a t-SNE plot. However, this time, the dataset only contains true positive values (correctly predicted cybercrime victims). But since the scores of the models got worse, as it could be seen at the table below, and because t-SNE could not show clear clusters, it is not practical to make further use of the dataset that only contains true positive values. Also, in the table a gap can be found for the results of SVM. The reason for this is the lack of computational power.

	Acc.	Prec.	Recall	F1
RF	57%	44%	60%	51%
SVM				
LR	55%	49%	55%	52%
GBDT	56%	47%	55%	52%

(Table: 2, Model scores with all true positives)

Because there is not a stand out algorithm, it is recommended to look at each model and their most important features for further evaluation. Therefore, a top 5 is made from the positive, negative and standard features. Because both the DT and RF could not (easily) make a distinction between positive and negative correlated features, the decision is made to make distinctive tables between the model groups.

LR	SVM
CEO- Large shareholder	CEO- Large shareholder
Private household with unknown income	Entrepreneur
Single parent household	Social benefit/incapacitated
Entrepreneur	Between 25 nd 45 years old
Social benefit/incapacitated	Single parent household

(Table: 3, Positively correlated features)

When looking at the positive correlated features, it shows that both the LR and SVM contain the same features for the most part: only where the Private household with unknown income' feature is important to LR, 'Between 25 and 45 years old' is that for the SVM.

LR	SVM
Not yet attanding school	Not yet attanding school
65 years or older	65 years or older
Household without observed income	Unknown or Institutional household
Institutional household	Unknown work
Unknown or Institutional household	Household without observed income

(Table: 4, Negatively corrolated features)

While analysing the top negative features, there could be seen that, as with the positive related features, the important features of both models are quite the same.

GBDT	RF
65 years or older	65 years or older
Between 45 and 65 years old	Male
Male	Married couple without children
Single person household	Female
Female	Between 45 and 65 years old

(Table: 5, Feature importance)

Here could be seen that, as with the previous tables, both models almost contain the same variables. Also, it is noticeable that 'Between 45 and 65 years old' is one of the top features on the DT, where the same variable just makes the top 5 for the top features of the RF model.

When taking a look at all the top features, there could be seen that the top features of both the DT and RF, except for the '65 years or older', aren't present in the top features (both positive and negative) of LR and SVM. Therefore, the assumption can be made that it is not clear which features are the most important for determining cybercrime victims.

But to still get a sense of which features could contribute to be a cybercrime victim, it therefore could be useful to implement the chi-squared test. Below an overview of the top 10 features with the chi-squared test:

1) Between 25 and 45 years old.
2) Male.
3) Secondary education.
4) Single parent household.
5) Non-western migration background.
6) Very high population density.
7) Higher education.
8) Second generation migration background.
9) Entrepreneur.
10) Single person household.

While comparing the results of the models to the results of the chi-squared tests, it seems to be that according to the chi-squared test, Secondary education' and the migration background features are important features for determining a cybercrime victim, while these features do not even make the top lists of all the models. Also, it looks like that men between 25 and 45 are more prone to cybercrime victimization when it comes to the chi-squared test. Also, a Decision Tree was created to show more about cybercrime victims *(see appendix: A, visualization of one decision tree)*. In the Decision Tree plot, there could be seen that there is not a big difference between genders when it comes to determining cybercrime victims. Also, the strongest correlated features are female who are not in the top 25% of income level and men who take part in an institutional household that are not in the top 50 to 25 percent income level. When the results of both the models and the chi-squared test are compared to the previous related works, it seems to be that the results differ from each other.Where at the related work chapter was stated that people in the age group of 18-25 are more likely to be victims of phishing attacks, it seems to be that according to the models and test that were used for this study, people between from higher age groups are more likely to be victims of cybercrime.

5 Conclusion

The goal of this study was to find subsequent groups that could describe some features that positively associate with a person becoming a cybercrime victim. Firstly, the results of the classification algorithms accuracy measured around 65%, which doesn't allow for any specific claims to be made towards a concrete profile of victims. Furthermore, the score tells us that there are tendencies for some groups to be victimized. This can be concluded since the models are able to distinguish between victims and non-victims with 15% better accuracy than random guessing.

Therefore, the feature importance of the different models can still provide information towards which features are relevant in the case. The age groups are important for all the models and also the chi-squared test. Sex of the victim is also present in some of the models and is a relevant splitting point for the decision tree *(see figure: 1, decision tree in the appendix)*. Furthermore, the t-SNE results show that it is not capable of clustering correctly classified victims on their features *(see figure: 2, t-SNE in the appendix)*.

As for which socio-economic and demographic features associate positively with reported cybercrime victimization of Dutch citizens in 2016, the varying results from the models limits any claims to be made around any features. But the score of the models do perform better than random, this in turn suggest that there may be something to be found here, maybe with less explainable algorithms like neural networks or by improving the quality of the input data.

6 Discussion

Results

During the exploratory data analysis, it turned out that it was hard to find any differences in characteristics between reported cybercrime and reported non-cybercrime victims in the PRD in combination with the SSD. To make it possible to better discover differences between the two types of victims, some characteristics were clustered and filtered. For example, the characteristic income was clustered into groups, so that there may be found differences between income groups.

After performing the Machine Learning models and the chi-squared test on the reported cybercrime and reported non-cybercrime victims, it turns out that the accuracy is moderately, namely scores around 65%. This indicates that the algorithms had difficulty discovering the differences between the characteristics of reported cybercrime victims and reported non-cybercrime victims in the PRD. However, it could be seen that several algorithms in the last run showed the same characteristics with high scores as a result. It was decided to take the most occuring characteristics with high scores as characteristics for Dutch cybercrime victims.

Limitations

Because sensitive data was used, it was only possible to work on location (at SN) during this study. For this reason, working days had to be planned with the team on which most team members were available. This had to ensure that enough time could be spent on the project. Moreover, the server, which was made available inside SN for this study, did not have sufficient computing power. This server was needed to be able to use the data and apply calculations and Machine Learning algorithms to it. The lack of computing power meant that the execution of code sometimes took a lot of time. For example, just reading in a DataFrame took sometimes up to more than ten minutes. Because of this, we sometimes had to take each other into account when doing complex calculations (including algorithms), so that the rest of the project group was not too limited in computing power and therefore had to wait.

6.1 Validity and reliability

In addition to the PRD that was used during this study, there was data about the safety monitor available as well. The safety monitor is a survey conducted among a part of the Dutch population in 2016. This survey asks questions related to the feel and experience of safety in daily life. Within this survey, there is also a section related to cyber security. This section contains questions about how safe people feel on the Internet and in the digital environment, what actions they take to prevent themselves against cybercrime and whether they have ever been a victim of cybercrime. However, eventually the data was not used in this study because it was found to be not reliable enough. The safety monitor is a dataset that has been assembled based on answers given by respondents. These answers may not always be true, for example, because the respondent is lying or because the person is not sure of the answer he/she has given. In addition, the safety monitor also has to deal with weighting factors; certain groups of respondents (e.g. certain age groups) have given more response than other groups of respondents. This is the reason why, when it comes to cybercrime victims, only the PRD was eventually used. The police data is largely based on facts. Of course, there is still a chance that the respondent is not telling the truth, but this probability may be much lower than in the safety monitor. Because a crime report in the PRD was made on the initiative of the person himself (the reporter) and the answers to the questions in the safety monitor were not, it can be assumed that persons in the PRD see more need to give honest answers. The reports in the police data were created at the time when the crime in question had just taken place, the answers in the safety monitor are based on how a person feels personally and what he or she can remember about certain things in the past. In addition, there is no problem with weighting factors in the police data that there is in the safety monitor.

In order to ensure that the answers given can be considered valid, data cleaning took place in advance. Initially, the SSD was filtered on persons who lived in the Netherlands in 2016. It turned out that the dataset also contained persons who had already passed away before 2016 or who no longer lived in the Netherlands. As a consequence, the results would not be valid when the highest correlated features of reported cybercrime victims were investigated. Moreover, there were also a number of characteristics in the PRD with empty values (NULL values), such as ’classifier_predictions’. The characteristic ’classifier_predictions’ are the cybercrime predictions that SN has previously performed. If the label has a positive value, then the case in question has been predicted as a cybercrime case. All the cases within the PRD data which have a positive value for ’classifier_predictions’ are the cases we used for this study for all reported cybercrime victims. If the label has a negative value, then the case is predicted as a non-cybercrime case. In addition to the fact that most reports had a positive or negative value for ’classifier_predictions’, there were also a number of cases that had a blank value (NULL) here. It was decided not to use these cases, because it cannot be said whether or not they are cybercrime-related.

6.2 Recommendations

The results of this study can be used by SN to produce (more) information on cybercrime victims from 2016. SN can use the features of cybercrime victims (the top features that have emerged from the algorithms) to sketch a profile for these cybercrime victims and to carry out further study into cybercrime victims.

In addition, SN can also apply the implementation (which has been used to perform calculations and algorithms in this study) to datasets (PRD SSD) from previous years, so that it can be investigated whether the top features of cybercrime victims have changed. SN can then use the results to describe the development of the profile of cybercrime victims.

Acknowledgements

We would like to thank SN in general for offering us a workspace and facilities to work on this research. We also want to thank Susan van Dijk, Sigrid van Hoek, Barteld Braaksma, Jeroen Vuurens and Tony Andrioli for helping, coordinating and advising us during this research.

References

[1] J. Armin, B. Thompson, and P. Kijewski, “Cybercrime economic costs: No measure no solution,” in *Combatting cybercrime and cyberterrorism*. Springer, 2016, pp. 135–155.

[2] R. Anderson, C. Barton, R. Böhme, R. Clayton, M. J. Van Eeten, M. Levi, T. Moore, and S. Savage, “Measuring the cost of cybercrime,” in *The economics of information security and privacy*. Springer, 2013, pp. 265–300.

[3] J. R. Agustina, "Understanding cyber victimization: Digital architectures and the disinhibition effect," *International Journal of Cyber Criminology*, vol. 9, no. 1, p. 35, 2015.

[4] V. Benson, G. Saridakis, and A.-M. Mohammed, "Understanding the relationship between cybercrime and human behavior through criminological theories and social networking sites," 2019.

[5] S. G. van de Weijer and E. R. Leukfeldt, "Big five personality traits of cybercrime victims," *Cyberpsychology, Behavior, and Social Networking*, vol. 20, no. 7, pp. 407–412, 2017.

[6] M. Gratian, S. Bandi, M. Cukier, J. Dykstra, and A. Ginther, "Correlating human traits and cyber security behavior intentions," *computers & security*, vol. 73, pp. 345–358, 2018.

[7] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 373–382.

[8] J. L. Parrish Jr, J. L. Bailey, and J. F. Courtney, "A personality based model for determining susceptibility to phishing attacks," *Little Rock: University of Arkansas*, pp. 285–296, 2009.

[9] J. G. Mohebzada, A. El Zarka, A. H. BHojani, and A. Darwish, "Phishing in a university community: Two large scale phishing experiments," in *2012 International Conference on Innovations in Information Technology (IIT)*. IEEE, 2012, pp. 249–254.

[10] M. Junger, L. Montoya, P. Hartel, and M. Heydari, "Towards the normalization of cybercrime victimization: A routine activities analysis of cybercrime in europe," in *2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*. IEEE, 2017, pp. 1–8.

[11] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[12] P. Li, "Robust logitboost and adaptive base class (abc) logitboost," *arXiv preprint arXiv:1203.3491*, 2012.

[13] J. Gualtieri, S. R. Chettri, R. Cromp, and L. Johnson, "Support vector machine classifiers as applied to aviris data," in *Proc. Eighth JPL Airborne Geoscience Workshop*, 1999.

[14] Yates, *Practice of statistics*. W H Freeman, 1998.

[15] A.-M. Bidgoli and M. N. Parsa, "A hybrid feature selection by resampling, chi squared and consistency evaluation techniques," *World Academy of Science, Engineering and Technology*, vol. 68, pp. 276–285, 2012.

[16] A. So, D. Hooshyar, K. Park, and H. Lim, "Early diagnosis of dementia from clinical data by machine learning techniques," *Applied Sciences*, vol. 7, no. 7, p. 651, 2017.

[17] A. Adel, N. Omar, and A. Al-Shabi, "A comparative study of combined feature selection methods for arabic text classification." *JCS*, vol. 10, no. 11, pp. 2232–2239, 2014.

[18] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class svm," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472, 2017.

[19] L. van der Maaten and G. Hinton, "Visualizing data using t-sne. journal of machine learning research 9," *Nov (2008)*, 2008.

[20] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

Appendix

A - Visualization of one decision tree.

B - t-SNE plot.