5-2018

Supervised Machine Learning Techniques, Cybersecurity Habits and Human Generated Password Entropy for Hacking Prediction

Pedro Taveras
Pontificia Universidad Catolica Madre y Maestra, pedrotaveras@pucmm.edu.do

Liliana Hernandez
Pontificia Universidad Catolica Madre y Maestra, lm.hernandez@ce.pucmm.edu.do

Midwest Association for Information Systems
MWAIS 2018 Saint Louis, Missouri

Supervised Machine Learning Techniques, Cybersecurity Habits and Human Generated Password Entropy for Hacking Prediction

Pedro Taveras
Pontificia Universidad Catolica Madre y Maestra
pedrotaveras@pucmm.edu.do

Liliana Hernandez
Pontificia Universidad Catolica Madre y Maestra
lm.hernandez@ce.pucmm.edu.do

ABSTRACT

Attempts to steal information through the hacking of online accounts or passwords violations are becoming more common. The human factor is involved in most cyber-attacks. A way to solve these human oversights is to start using artificial intelligence, delegating some human decisions in the machines, but these innovations also have much to improve. Human judgment is still necessary to fill the gap between the capabilities of technology and our needs. This is where conscious security habits play a differentiating role between being or not the victim of a cyber-attack. This study describes how machine learning techniques can be used to model predictions that allow the anticipation of a hacking event taking into account password entropy and security habits. Prediction models are created and trained using decision tree techniques, multilayer perceptron, and Naïve Bayes. The efficiency of these models is contrasted to determine which of the models is more efficient for the case under study.

Keywords

Password Strength, Cybersecurity Habits, Machine Learning, Neural Networks

INTRODUCTION

Selecting or creating new passwords is probably one of the most annoying processes from the user's perspective. Even more so when the number of services that are used - and, therefore, of passwords - is increasing. In addition, a number of requirements that generally do not facilitate that these are easy to remember (Yan, Blackwell, Anderson and Grant, 2004). This is one of the reasons why solutions such as fingerprint readers in mobile phones have enjoyed such a specular reception by manufacturers and users since the intermediate step of having to enter the password is avoided (Javed, Shehab, and Bello-Ogunu, 2017). Behind all those standards that complicate the choice of a new password (uppercase and lowercase letters, numbers, symbols) the truth is that they are harder to remember for humans, but easy to decipher with current technologies and computational power.

Attempts to steal information and identity theft through the hacking of online accounts or passwords violations are becoming more common. The human factor is involved in most cyber-attacks. At any point in this chain of processes, from an engineer who could inadvertently create a vulnerability in the software, even the end user who clicked on an incorrect link or had a weak password. More recent trends in the field of cybersecurity suggest that the best defense against this type of situation is user itself and the adoption of good security habits. One of the ways to solve these human oversights is to start using artificial intelligence, delegating some human decisions in the machines, but these innovations also have much to improve. Human judgment is still necessary to fill the gap between the capabilities of our technologies and our needs. This is where conscious security habits play a differentiating role between being or not the victim of a cyber-attack.

There are multiple studies related to the field of cybersecurity and data mining taking advantage of the potential of machine learning to solve current problems (Fraley and Cannady, 2017). There are applications of supervised learning techniques such as Multilayer Perceptron, Naive Bayes, Decision Tree, and Support Vector Machine for predictive activities related to online access security and password strength (Vijaya, Jamuna and Karpagavall, 2009; Zheng, Cheng, Zhang, Zhao, and Wang, 2018).

However, most of these efforts have focused on large-scale studies that quantify password composition patterns. These studies focus on quantitative parameters such as the length of the key and the combinatory of the characters that compose it. The present study proposes a different approach by incorporating a set of variables related to users' security habits.

The general objective of the study is to create a supervised classification model capable of determining if an individual is susceptible to being hacked. At the same time determine which data mining algorithm provides a higher level of precision when making the prediction. The model is built from a set of variables associated with users' perception of the relevance of their security habits when selecting, manipulating and managing their passwords. At the same time, the contrast of these habits is generated with the real level of security, using as a parameter the strength of the selected passwords. All this, in order to promote knowledge about the best practices of password habits by end users. This general objective can be broken down into three specific actions:

1. Analyze to which extent password habits of the end users correspond to the probable event of the hacked account.
2. Guide initiatives in the area of cybersecurity to improve individual actions for the mitigation of risks derived from inappropriate password security habits.
    a. Generate knowledge so that organizations can know which habits fail the most, to create habits policies so that strengths are stronger, (less vulnerable).
3. Determine which machine learning methodologies provides a higher level of precision when modeling this type of variables.
    a. Model predictions that allow the anticipation of a hacking event using the described habits as input factors.

SUPERVISED DATA MINING TECHNIQUES

Data Mining is the process by which useful and understandable knowledge-previously unknown, is automatically extracted from databases. Then, the exploitation of information poses two challenges: working with large databases and applying techniques that automatically convert these data into knowledge. Human beings, having cognitive capacity, develop a series of behaviors that can be defined as patterns depending on certain situations. Habits and decisions made at certain times respond to the existence of such patterns (Gunderson, 2002). This information, managed through data mining, allows the development of predictive models in scenarios where specific events could happen. The prediction process is complex since it not only includes the obtaining of models or patterns, but also the evaluation and interpretation of them.

Decision Tree C4.5

Decision trees contain a hierarchically organized model. So that an attribute can be classified, deciding in each node the condition that it fulfills. Describing a path from the root to the leaves. In each node, decisions are exclusive. It is assumed that the classes are disjoint, that is, an instance is of class *a* or class *b*, but can not be at the same time of classes *a* and *b*. The classes are exhaustive (Bouzida and Cuppens, 2006). The decision tree is constructed in the learning phase and then used to predict the instances. The more types of conditions allowed, the more possibilities we will have to find the patterns behind the data. The question in a good decision tree learning algorithm is to find the appropriate balance between expressivity and efficiency. The C4.5 algorithm proposed by Quinlan (1993) was the implemented. C 4.5 is based on the divide and conquer method and partition criteria (GainRatio), including rule based pruning and other more sophisticated mechanisms.

Multilayer Perceptron

Multilayer perceptron is one of the most used supervised algorithms of neural networks. Its architecture is based on an input layer, another output layer, and at least one hidden layer. Activation is propagated in the network through the weights from the input layer to the intermediate layer where some activation function is applied to the incoming inputs. Then activation is propagated through the weights to the output layer (Witten, Frank, Hall and Pal, 2016). The learning capacity of the neural network is determined by the number of hidden layers and the number of units in each layer. The variables to be used must be chosen carefully: the aim is to include in the model the predictor variables that actually predict the dependent or exit variable, but which in turn do not have relations with each other since this may cause unnecessary overfitting in the model.

Naïve Bayes Classification

Probability theory and Bayesian methods are one of the techniques that have been used the most in artificial intelligence problems and, therefore, in machine learning and data mining. They are a practical method to make inferences from the data, inducing probabilistic models that will later be used to reason (formulate hypotheses) about new observed values (Cichosz, 2015). They allow to explicitly calculate the probability associated to each of the possible hypotheses, which constitutes a great advantage over other techniques. In addition, Naïve Bayes provides a useful framework for the understanding and analysis of

numerous data mining and learning techniques that do not work explicitly with probabilities. The Naïve Bayes algorithm is based on the assumption that the attributes are independent. Although this assumption is assumed, it is quite strong in most cases. Based on this assumption, the study performed an exhaustive analysis of the variables that were taken from the instrument.

RESEARCH METHODOLOGY

Target Population

The target population for this study consisted of 368 internet users. Each participant was familiar with the use of internet, systems logins and password selection and creation process. The survey was distributed in United States, China, Poland, Angola, and the Caribbean.

Experiment Design

The experiment was designed by selecting a random population with a sample size of N = 368 individuals with different characteristics. Based on the training data obtained in the survey, it is intended to find models of machine learning that, based on the password strength and user security habits, are able to predict if the user will be hacked.

A password dictionary was created by asking participants to voluntarily provide an example of a password that they would normally use, completely free without specifying any kind of rule but indicating the universe of symbols conforming the password alphabet. A second question asked the user to create a password that they would normally use but based on a specific rule, including password size, a combination of alphanumeric characters and use of special symbols. All participants were asked if they ever had been a victim of any hacking or having an online account compromised. According to the data obtained so far, 30% of users have been victimized, at least once, through hacking, attack or cybersecurity violation.

Lime Survey online tool was used to carry out the survey. It allows to capture an unlimited number of samples and provides a wide set of tools that ease the data processing. The instrument covers a set of variables grouped into three categories: demography, password habits, and password strength. From this instrument, the following variables were obtained:

Academic grade, age range, gender, average range of daily logins, confidence level in current password habits, password selection methodology, average range of online active accounts, average range of password used, frequency passwords change, frequency of password recovery, preferred mechanisms to remember passwords, passwords sharing habits, passwords transportation mechanism, type of connection , percentage of time connected per place, factors for password selection, average length of password, password strength mechanisms, password example without rule, example of password with induced rule and use of two-factor authentication.

Preparation of Data

The data preparation process was carried out taking into account the quality of the data obtained, to guarantee a correct application of the learning algorithms. The survey instrument was designed with all answers mandatory, thus eliminating the null data. Only the surveys that were 100% completed are taken into account to compute the output variable that is sought in the study. Quantitative variables were discretized. Some quantitative variables such as number of online active accounts, number of different passwords, the frequency of password recovery, were taken in forms of intervals making it easier to capture.

On Password Entropy

Shannon (1949) entropy theory was used to determine the strength of the password provided by the users. Password entropy refers to a measure of how unpredictable the password is (NIST, 2013; NIST 2017), and was determined using the following equation:

$$E = Log_2(R^L)$$

where

E = Password entropy
R = Universe of unique possible characters
L = Password length (characters count)

then

R^L= Number of possible passwords

and

$Log_2(R^L)$ = number of bits of entropy

The value of R was established considering the universe of unique possible characters as a base element, according to Table 1. The present study considered complex alphanumeric and usable ASCII chars, for a password composed from a set of 92 possible symbols. The variable password strength was discretized by intervals to apply the supervised classification algorithms. Table 2 present the ranges for bits of entropy and the strength classification of each range.

Type	Possible Characters
Alpha lowercase	26
Alpha lower + uppercase	52
Alphanumeric simple	36
Alphanumeric complex	62
Usable ASCII char	30

Table 1. Universe of Symbols

Password Entropy	bits
Very Weak	< 28
Week	28-35
Good or Fair	36-59
Strong	60-127
Very Strong	> 128

Table 2. Password Entropy

PRELIMINARY RESULTS

Waikato Environment for Knowledge Analysis (WEKA), version 3.8, was used for the preprocessing and construction of data mining classification models and knowledge analysis. WEKA is open-source and allows the implementations of various state-of-the-art data mining algorithms, including those used in the study (Witten, Frank, Trigg, Hall, Holmes and Cunningham, 1999). The models were constructed for the described techniques: C4.5, Multilayer Perceptron, and Naïve Bayes using cross validation to select the training and test partitions by dividing them into 10 intervals (10-fold cross validation). Ensuring that the results are not dependent on the subsets in which the data is partitioned.

	Decision Tree C4.5	Naïve Bayes	Perceptron Multilayer
Positive Instances Correctly Classified (PICC)	39	37	47
Negative Instances Correctly Classified (NICC)	60	69	85
Positive Instances Incorrectly Classified (PIIC)	21	20	9
Negative Instances Incorrectly Classifieds (NIIC)	28	22	7
Percentage of Correctly Classified Instances	66.89%	71.62%	89.19%

Table 3. Results Per Method

In the present research the output variable "hacked" belongs to a binary class, having two states: [Hacked = YES], if the respondent has been a victim of hacking and [Hacked = NO] if the respondent affirms never been hacked. Therefore, the model can give as results four possible cases:

1. *Hacked* instances classified as *Hacked* → Positive Instances Correctly Classified (PICC)
2. *Non-hacked* instances classified as *Non-Hacked* → Negative Instances Correctly Classified (NICC)
3. *Hacked* instances classified as *Non-Hacked* → Positive Instances Incorrectly Classified (PIIC)
4. *Non-Hacked* instances classified a *Hacked* → Negative Instances Incorrectly Classifieds (NIIC)

The precision of each algorithm was calculated taking into account the percentage of instances well classified in the test data set, as shown in Table 3. According to the following equation Percentage of Correctly Classified Instances (PCCI) is equal to:

$$PCCI = \frac{(PICC + NICC)}{(PICC + NICC + PIIC + NIIC)}$$

As shown in Figure 1, the Multilayer Perceptron overcomes Naive Bayes and Decision Tree algorithms, presenting 89.19% of correctly-classified instances. This result can be considered as a good percentage of classification seeing that there is no over adjustment to the training data. With a percentage close to 100, it is inferred that the model is too adjusted to the test data and could tend to incorrectly classify an instance that is not included in the training data. Neural networks tend to show a higher percentage of correct classifications, regardless of the correlation pattern of the input variables. This may be the reason why the Naïve Bayes classifier and the C4.5 classifier are yielding lower percent rankings.

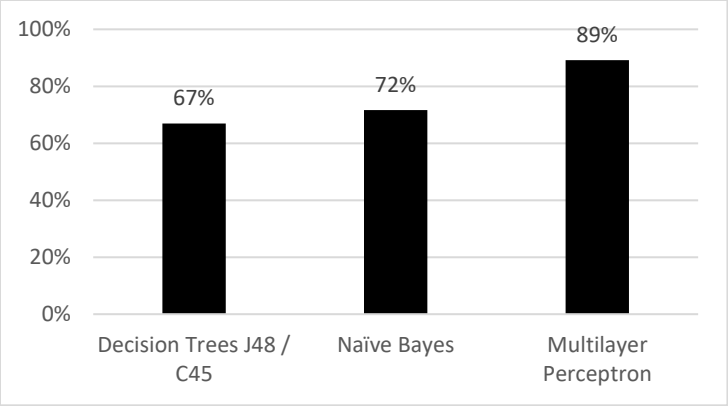

Figure 1. Percentage of Correctly Classified Instances

CONCLUSIONS

This research attempted to analyze to which extent password habits of end users correspond to the probable event of the hacked account and at the same time, model predictions that allow the anticipation of a hacking event using the described habits as input factors. A survey instrument was used to collect data about user security habits and a password dictionary was created by asking participants to select and write down a password of their choice. While the preliminary findings demonstrate that machine learning algorithms, specifically neural networks, can be used to model the prediction, the study is still limited. Incoming phases of this research should attempt to examine the factors through the prism of different instruments and with a greater number of subjects. In this preliminary stage, a major limitation observed is that the results may be skewed due to the potential presence of atypical users. So far 79% of participants come from engineering or technology background.

Finally, as a next step, the present research pretends to analyze the correlation of input variables, to make a better selection in order to adjust/calibrate the model. It is estimated that by increasing the sample size a better mode will be obtained. These are partial results according to the data obtained, so far 148 complete surveys, which represent only 40% of the expected samples. It is estimated that counting with a wider sample could lead to obtaining accurate learning models that predict the vulnerability of users in computer systems based on their password habits and password entropy. In a general way it is concluded that when applying supervised learning techniques, Neural Networks can make a good prediction of the vulnerability of users taking into account password habits and password entropy.

REFERENCES

1. Bonneau, J., Herley, C., Oorschot, P. C. v., & Stajano, F. (2012, 20-23 May 2012). The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. Paper presented at the 2012 IEEE Symposium on Security and Privacy.
2. Bouzida, Y., &, F. e. C. (2006). Neural networks vs. decision trees for intrusion detection. Retrieved from http://www.diadem-firewall.org/workshop06/papers/monam06-paper-29.pdf
3. Cichosz, P. (2015). Naïve Bayes classifier Data Mining Algorithms (pp. 118-133): John Wiley & Sons, Ltd.
4. Fraley, J. B., & Cannady, J. (2017, March 30 2017-April 2 2017). The promise of machine learning in cybersecurity. Paper presented at the SoutheastCon 2017.
5. Gunderson, L. F. (2002, 6-9 Oct. 2002). Using data mining and judgment analysis to construct a predictive model of crime. Paper presented at the IEEE International Conference on Systems, Man and Cybernetics.
6. Javed, Y., Shehab, M., & Bello-Ogunu, E. (2017). Investigating User Comprehension and Risk Perception of Apple's Touch ID Technology. Paper presented at the Proceedings of the 12th International Conference on Availability, Reliability and Security, Reggio Calabria, Italy.
7. John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. Paper presented at the Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, Montreal, Canada.

8. Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., . . . Egelman, S. (2011). Of passwords and people: measuring the effect of password-composition policies. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada.
9. NIST Standards & Technology. (2013). NIST Special Publication 800-63-2 Information Security: CreateSpace.
10. NIST Standards & Technology. (2017). NIST Special Publication 800-63-3 Information Security: CreateSpace.
11. Ur, B., Bees, J., Segreti, S. M., Bauer, L., Christin, N., & Cranor, L. F. (2016). Do Users' Perceptions of Password Security Match Reality? Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, California, USA.
12. Quinlan, J.R. (1992). C4.5 Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann.
13. Shannon, C. E. (1949). The mathematical theory of communication. Urbana: University of Illinois Press.
14. Viyaya, M., januma, K., & Karpagavalli, S. (2009, 28-29 Dec. 2009). Password Strength Prediction Using Supervised Machine Learning Techniques. Paper presented at the 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies.
15. Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S.J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. (Working paper 99/11). Hamilton, New Zealand: University of Waikato, Department of Computer Science.
16. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: practical machine learning tools and techniques. Amsterdam: Morgan Kaufmann.
17. Yan, J., Blackwell, A., Anderson, R., & Grant, A. (2004). Password memorability and security: empirical results. IEEE Security & Privacy, 2(5), 25-31. doi:10.1109/MSP.2004.81
18. Zheng, Z., Cheng, H., Zhang, Z., Zhao, Y., & Wang, P. (2018). An Alternative Method for Understanding User-Chosen Passwords. Security and Communication Networks, 2018, 12. doi:10.1155/2018/6160125