DATS 6501 Capstone

Sayra J Moore

The George Washington University

DATS 6501 Capstone

## Abstract

As our lives continue to become indulged into the little screens that we hold in our hands, people forget the power that these devices have, and how easily it they can leave us vulnerable to falling victim to issues like cybercrime. A device so small that’s supposed to be so convenient, causing so much devastation. Think of all the small tasks that we are able to do on said devices on a daily basis: ordering your daily coffee run ahead of time, hopping on Amazon to restock on some of your favorite items, logging into your account to make online banking easy, and so much more! Next thing you know, someone is using your credit card information on the other side of the world, or someone is stealing even more personal information from you by hacking into personal accounts, etc. How is this happening? Well, SentinelOne listed the top two of the most common bad habits as being ‘reused passwords’ and ‘weak passwords’ in their article, “*11 Bad Habits That Destroy Your Cybersecurity Efforts*” (Kedem, 2021). This is common because who wants to remember a new and complex password for every single online account they create – it’s so much work! Though in reality, by simply following such a simple best practice like using complex and unique passwords can prevent so much from happening in the future, like preventing hackers from gaining access to your accounts. Because of this, I have always been interested in looking further into this concept of determining what habits and behaviors influence the likelihood of someone falling victim to a cybercrime like credit card fraud or identity fraud.

## Introduction

In the early months of 2021, Joseph Johnson wrote an article in Statista on the U.S.’s 2020 most reported cybercrimes. These cybercrimes included credit card fraud, harassment, spoofing, identify theft, and the number one cybercrime: Phishing (Johnson, 2020). As Cisco defines it, “Phishing attacks are the practice of sending fraudulent communications that appear to come from a reputable source” (2021). This type of attack happens so often because it is completely reliant on an individual’s lack of attention. We as humans are so easily and often distracted, that these types of things never even cross our radar until is it too late. But could this be prevented? What if studying an individual’s actions around technology could help bring awareness to people about how easily these types of attacks could occur, and in the numerous different ways it could present itself? Better yet, how about working to predict and identify fraud before it goes too far!

## Literature Review

Steve Morgan from CSO wrote the Cybersecurity Business Report on the “*5 worst cybersecurity habits with catastrophic consequences*”. In this, he states that the first habit is “lax attitude” – or the lack of taking security seriously and/or neglecting cyber safety (Morgan, 2017). Another habit he lists is not having any email protection, namely using two-step verification, which requires an extra code when a user logs in, that is texted to the user (Morgan, 2017). This can prevent cyber attacks via email because it’s highly likely that the hacker does not have the device that received the code required to log-in, therefore causing the hacker to not gain access to the account. Another bad habit is clicking on hyperlinks in emails, also known as appear phishing attack – or fake email – that is created to look like it’s coming from a trusted source, but then when clicked on, leads to a fraudulent link (Morgan, 2017). The fourth bad habit that Morgan listed is individuals having poor password practices (Morgan, 2017), as we mentioned earlier. Again, individuals tend to not only use weak passwords, but also reuse them, making it immensely easier for hackers to gain access to accounts. The final bad habit that Morgan talks about is individuals not having data backups. Not having backups can be detrimental especially when something like ransomware is introduced. Ransomware is “a malware that infects computers and restricts their access to files, often threatening permanent data destruction unless a ransom is paid (Morgan, 2017). Not only can this lead to permanent loss of data, but it can cost anywhere from millions to billions of dollars. After my interest and some research, I chose my topic for this study to go along with this concept, in trying to determine if an individual is more likely to be a victim of a cybercrime due to their habits and behaviors.

## Research Methodology

In this study, R was used to complete the data cleansing along with a portion of the Exploratory Data Analysis (EDA) and visualizations, with the other portion being completed in Tableau. R was also used for building out all the different models including simple linear regression, multiple linear regression, logistical regression, clustering, and classification tree. For this deliverable, we will focus on the logistic regression model.

## Data

The dataset that is being used for this study holds 28 variables and approximately 3000 observations. The dependent variable in this dataset is VICTIM, identifies if a person has fallen victim to cybercrime or not, depicted as 1 and 0 respectively. The independent variables include CC\_FRAUD, OTHER\_ID\_THEFT, VICTIM\_CAT, VICTIM\_CAT\_CODE, DATE\_OF\_ATTACK, AGE, DWELLING, CITY\_RURAL, COMPUTER, TABLET, CELL\_PHONE, HH\_INCOME, NO\_OF\_CC, SOCIAL\_NET, SOCIAL\_NET\_ACCT, OL\_BANK, FIREWALL, MALWARE\_PRESENT, FREQUENT\_TRAVELER, SECURITY\_CLEARANCE, KIDS\_IN\_HH, TOT\_KIDS, SECURE\_HNET, SESS\_CNT\_1MOS, CYBER\_ATKS, and NO\_CYBER\_ATKS\_1MO. Some data cleansing needed to be completed in order to make sure that all the independent variables were the correct data types, as well as assuring that its observations were the appropriate data types as well. For example, the variable MALWARE\_PRESENT not only had binary data values (1 and 0), but also had “Yes” and “No” values, therefore miss-categorizing it as a character (CHR) data type and not integer (INT). We saw this several more times with the data including COMPUTER, TABLET, CELL\_PHONE, and SOCIAL\_NET. Once those values were corrected using gsub(), their data types were changed from character (CHR) to integer (INT).

The other data type that needed to be corrected was the date. This was done by using as.date() then setting the date format to “%m%d%Y”. Lastly, there were some nulls in the dataset that needed to be dealt with. Since there were very few, the solution to this was to carry forward the last observation for the null using na.locf().

## Analysis

After studying the dataset, it’s observations and variables, it was made clear that the most logical analysis to complete would be logistic regressions. This assumption was made because the target variable is binary (1 and 0), but so are almost all the remaining variables (except for six!). Therefore, the goal here is to try and determine the likelihood that an individual would fall victim to a cybercrime while looking at the other variables, which more importantly depict habits – which can often be predicted.

## Key Findings

While looking at the different variables, assumptions could be made based on what we know and experience every day. For example, the average person would look at this dataset and assume that as age increases, so does the number of monthly cyber attacks, as well as the likelihood that an individual could fall victim to a cybercrime. This assumption is made because of the rapid advances and changes in technology. Although a study by Emily A. Vogels of a Pew Research Center survey recently showed that approximately 68% of Baby Boomers own smart phones, as well as 40% of the Silent Generation (Vogels, 2020). Although the ability to keep up with such rapid changes tends to become more challenging as time goes by, with this assumption in mind, we can look at a Pearson Correlation test between AGE and NO\_CYBER\_ATKS\_1MO in order to truly determine the correlation. After running the rest, we received a correlation coefficient (r) = 0.0003, which clearly means that there is no correlation between the two variables.

Chart, scatter chart

Description automatically generated

*Figure 1*

Another assumption made between variables in this dataset was the correlation between SESS\_CNT\_1MOS and NO\_CYBER\_ATKS\_1MO. The Pearson Correlation test output a correlation coefficient (r) = 0.83, revealing that those variables are indeed positively correlated, meaning as one increases, so does the other.

Chart

Description automatically generated

*Figure 2*

In Figure 3 below, we can see the correlation matrix more broadly between all of the continued independent variables.

A picture containing chart

Description automatically generated

*Figure 3*

## Conclusion

As mentioned earlier, the focus model in this study is logistic regression. Building this type of model for our data will allow us to determine the impact that the various independent variables have on whether an individual is a victim of fraud or not. In order to successfully complete this, the model was built using glm(), where we received the following coefficients:

**A picture containing text, plaque

Description automatically generated**

*Figure 4*

With logistic regression, once you have generated a model and have an output with coefficients, you then have the ability to create a logistic function in the form of:

*ln(pi/1-pi) = b0 + b1x1 + b2x2 + … + bnxn*.

In looking at the coefficients that were generated with the output, we get the following function:

ln(pi/1-pi) = - 3.06 + 6.11(CC\_Fraud) + 6.11(Other\_ID\_Theft) – 1.94(Computer) – 7.14(Tablet) + 2.13(Cell\_Phone) – 3.52(Social\_Net) – 3.24(OL\_Bank) + 1.43(Firewall) + 2.33(Malware.Present) + 6.24(Frequent.Traveler) – 1.38(Security\_Clearance) – 4.51(Kids\_in\_HH) – 1.32(Secure\_Hnet) + 1.01(Cyber\_Atks)

The way we interpret these coefficients in this model are as the following: if an individual has experienced credit card fraud or any other type of identity theft, the dependent variable (DV) or VICTIM log-odds increases by 6.11. If an individual has a computer, tablet, or a social network account, the DV log-odds decreases by 1.94, 7.14, and 3.52 respectively, but if an individual has a cell phone, then the DV log-odds increases by 2.13. We also see that where an individual has a firewall and malware present, the DV log-odds increases by 1.43 and 2.33 respectively. If an individual is a frequent traveler, the DV log-odds increases by 6.24. If an individual has a security clearance, the DV log-odds decreases by 1.38. Lastly, if an individual has children in their household, the DV logs-odd decreases by 4.51.

In order to do deeper research and improve the model by building a more effective one, I believe could be accomplished if there were a greater number of data points in this data source, as well as having additional independent variables. I feel that although there were 28 variables, there are many that could’ve potentially influence the model in a better way, such as password complexity, knowing how many accounts shared passwords, and knowing if individuals ever clicked on phishing emails.

## Biographical Information

Sayra J. Moore is a student in the Data Science Master’s Program at The Geroge Washington University. As she finishes her final semester in August 2021 while knocking out the last 4 courses in her program, she also works full-time as a Data Analyst at Booz Allen Hamilton, while supporting the government sector. Although she currently does not have free time due to her rigorous schedule, she looks forward to completing her degree and having time to spend with her family again, as well as picking back up some old hobbies like playing piano, camping, binge watching Netflix, and exploring her new city, Washington D.C.!

## References

Cisco. (2021, July 9). *What is phishing? Examples and phishing quiz*. Cisco. https://www.cisco.com/c/en/us/products/security/email-security/what-is-phishing.html.

Johnson, J. (2021, March 18). Most frequently reported types of cyber crime 2020. Statista. https://www.statista.com/statistics/184083/commonly-reported-types-of-cyber-crime/.

Kedem, M. (2021, January 11). *11 bad habits that destroy your cybersecurity efforts*. SentinelOne. Retrieved September 10, 2021, from https://www.sentinelone.com/blog/11-bad-habits-destroy-cybersecurity-efforts/.

Morgan, S. (2017, October 9). *5 worst cybersecurity habits with catastrophic consequences*. CSO Online. Retrieved September 10, 2021, from https://www.csoonline.com/article/3231669/5-worst-cybersecurity-habits-with-catastrophic-consequences.html.

Vogels, E. A. (2020, September 23). Millennials stand out for their technology use, but older generations also embrace digital life. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/09/09/us-generations-technology-use/.

## Appendix

The variables in this dataset are listed with their original data type (prior to data cleansing) as follows:

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| ID | INT | The user ID |
| \*VICTIM | INT | If an individual fell victim to a cybercrime |
| CC\_FRAUD | CHR | If an individual experienced credit card fraud |
| OTHER\_ID\_THEFT | INT | If an individual experienced any other identity fraud |
| VICTIM\_CAT | CHR | The victim category |
| VICTIM\_CAT\_CODE | INT | A code that corresponds to the victim category |
| DATE\_OF\_ATTACK | CHR | The data that the cybercrime occurred |
| AGE | INT | The age of the victim |
| DWELLING | CHR | The type of residential living space an individual has |
| CITY\_RURAL | CHR | Where an individual resides |
| COMPUTER | CHR | If an individual owns a computer |
| TABLET | CHR | If an individual owns a tablet |
| CELL\_PHONE | CHR | If an individual owns a cell phone |
| HH\_INCOME | CHR | The level of income a household has |
| NO\_OF\_CC | INT | The number of credit cards an individual owns |
| SOCIAL\_NET | CHR | If an individual has social networking accounts |
| SOCIAL\_NET\_ACCT | CHR | The type of social network account an individual has |
| OL\_BANK | INT | If an individual does on-line banking |
| FIREWALL | INT | If an individual has a firewall |
| MALWARE\_PRESENT | CHR | If there was malware present |
| FREQUENT\_TRAVELER | INT | If an individual travels frequently |
| SECURITY\_CLEARANCE | INT | If an induvial holds a security clearance |
| KIDS\_IN\_HH | INT | If the household include any children |
| TOT\_KIDS | INT | The total number of children in the household |
| SECURE\_HNET | INT | If the home network of an individual is secure |
| SESS\_CNT\_1MOS | INT | The number of monthly network sessions an individual has |
| CYBER\_ATKS | INT | If an individual experience a cyber attack |
| NO\_CYBER\_ATKS\_1MO | INT | The number of cyber attacks monthly |

*\*Variable is the dependent variable (DV), or target value.*