CHAPTER 1
INTRODUCTION

1.1 Background

Nowadays, digital transaction is widely used by people all around the world for variety of purposes such as online purchases, or payment of services. This is made possible by ease of internet access, almost every device can connect to internet and businesses starts to implement more online based payment. Instead of going to the physical store, customer can now pay with the comfort of home using said online payment services to pay for their daily needs such as water, phone, and electric bills to grocery shopping or any product at all. The growing field of E-commerce gives advantage that the customer is possible to pay online for their needs rather than having to go to the physical store or service provider that may be far from their area. But this ease of access also leads to many opportunities for fraudster to take advantage of.

To give an illustration of how big of a problem fraud is, the following is a chart of reported identity theft cases in US in 2018.

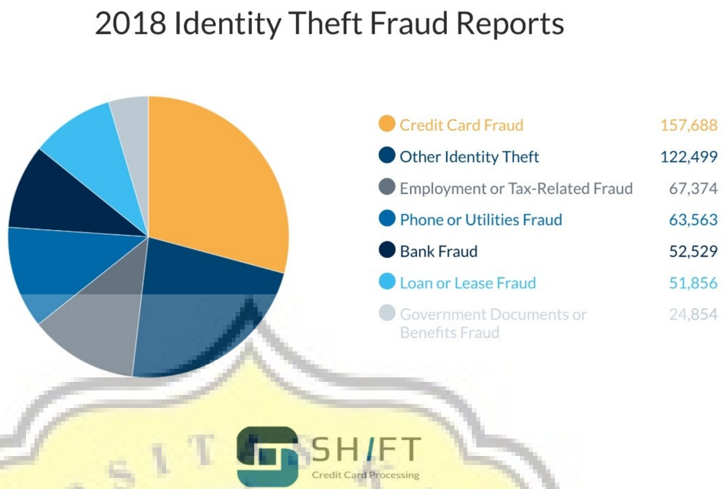


Illustration 1.1: Reported fraud cases occurred in U.S. in 2018, taken from https://shiftprocessing.com/credit-card-fraud-statistics/

Keeping in mind that the above chart shows only reported cases, the actual number of fraud cases is higher than what is shown in above chart.

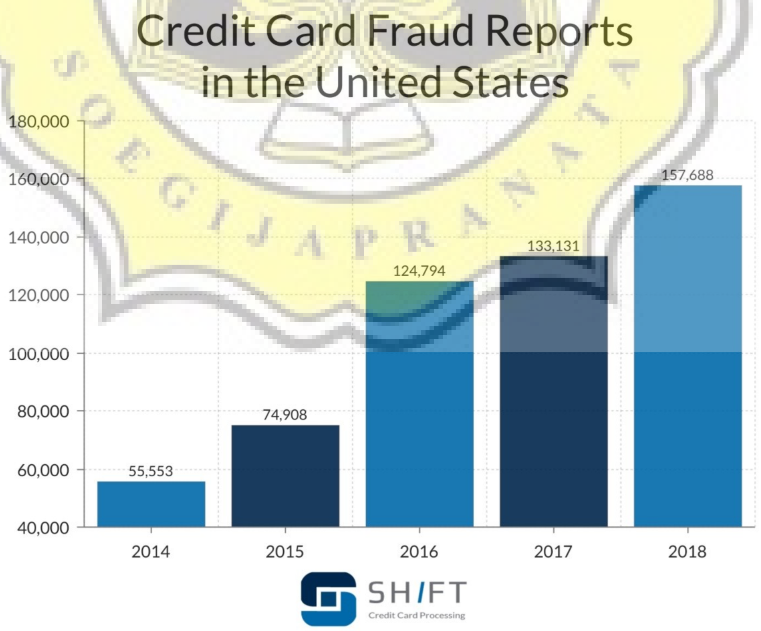


Illustration 1.2: Reported credit card fraud cases in U.S. from 2014 to 2018, taken from https://shiftprocessing.com/credit-card-fraud-statistics/

As much as 157 thousand of such cases are credit card fraud, which makes up for more than 25% of all reported identity theft fraud. The number of fraud cases continues to increase, as shown in above chart, in 2014 the number of credit card fraud cases is recorded at 55 thousand and continues to increase as each year passes.

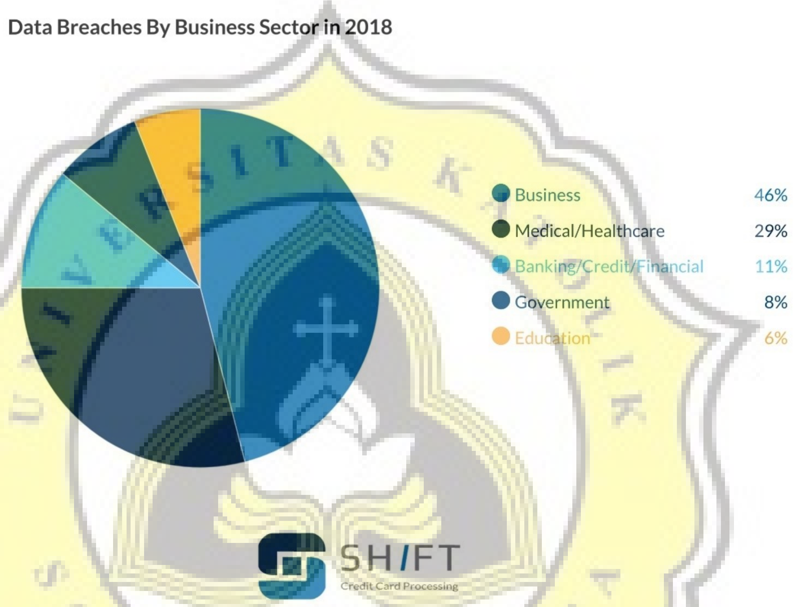


Illustration 1.3: Data breaches by sector in 2018, taken from https://shiftprocessing.com/credit-card-fraud-statistics/

In 2018, data breaches mostly occurred in the business section with Facebook and Cambridge Analytica data breach and Google in March, Reddit in August, Quora in December being the most well-known ones [2].

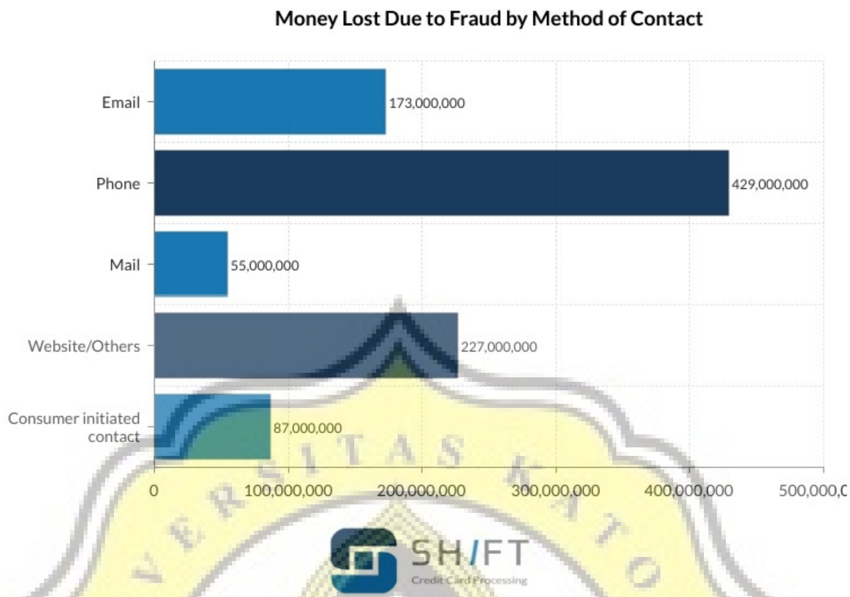


Illustration 1.4: Losses suffered by method of contact, taken from https://shiftprocessing.com/credit-card-fraud-statistics/

Fraudsters will likely use many ways to get victims information, above chart shows the losses suffered by methods of contact from abovementioned reported fraud cases in 2018. Phone being the most used method of communication by scammers mostly pretending to be someone else or a business entity offering some products that requires the victim to tell sensitive information such as credit card number or personal ID.

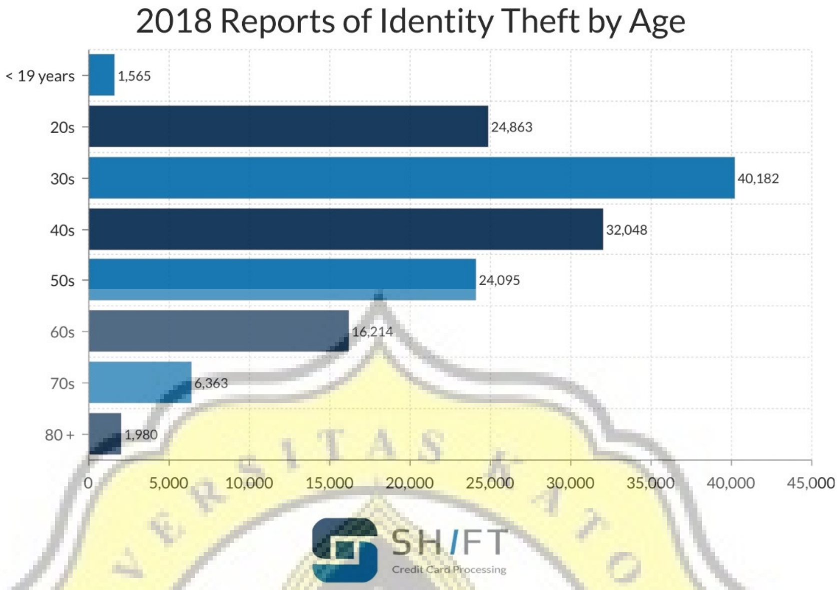


Illustration 1.5: Age of identity theft victims, taken from https://shiftprocessing.com/credit-card-fraud-statistics/

Scammers targets many age groups but most commonly target individuals from 30s and 40s age group since the majority of prominent figure are of those ages. Above chart shows identity theft victims by age from reported cases of 2018.

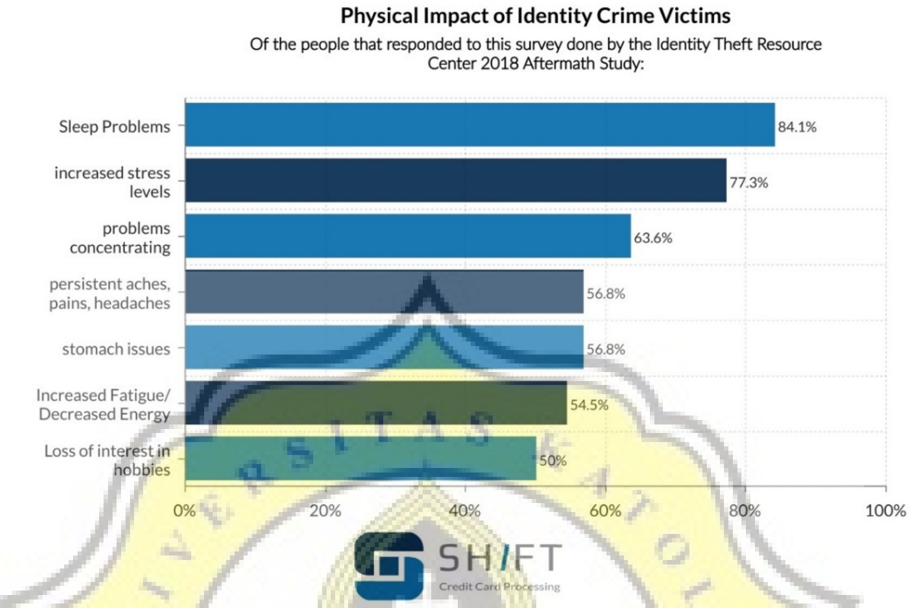


Illustration 1.6: Physical impact to victims, taken from https://shiftprocessing.com/credit-card-fraud-statistics/

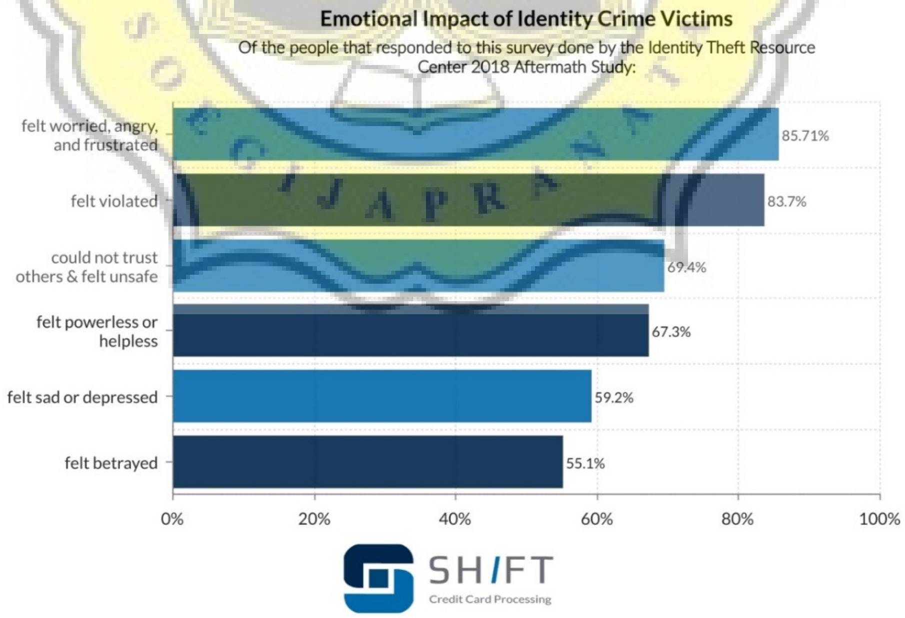
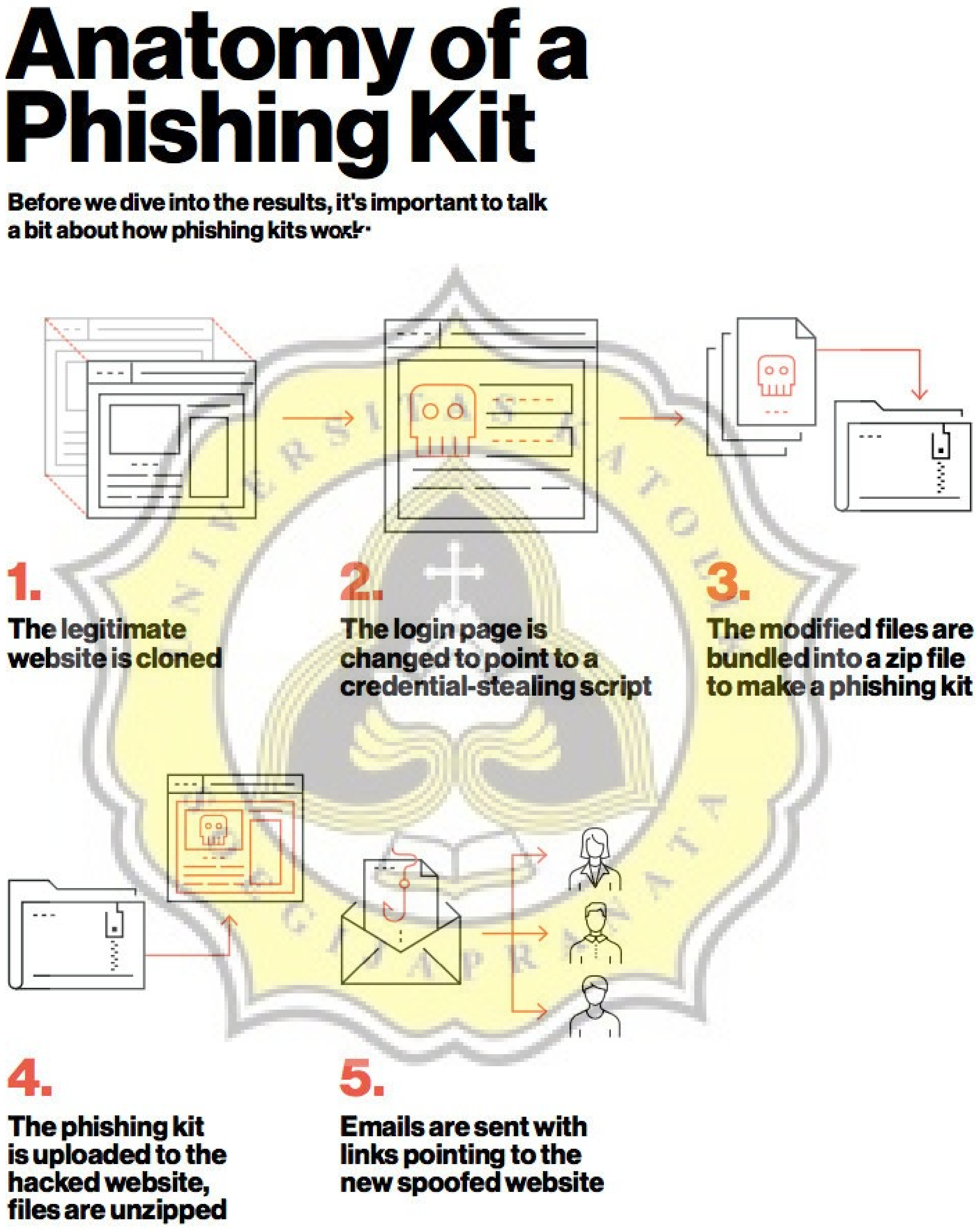


Illustration 1.7: Emotional impact to victims, taken from https://shiftprocessing.com/credit-card-fraud-statistics/

Besides the immediate monetary or data loss, the victims also suffers plethora of emotional and physical problems ranging from sleep problem to depression. Below chart shows physical and emotional impacts the victims most commonly suffer after such experience.

Types of fraud occurring in the field of E-commerce includes: phishing which can result in the fraudster getting the victim's identity or credit card information. If the fraudster does get the credit card information, he can easily make purchases and bill it to the victim's credit card which causes a loss to the victim and possibly the service provider too if the victim decides to seek reimbursement. This kind of method can be avoided if the would-be victim is attentive; If the fraudster creates fake website to gather user information even if the website's UI is exactly the same as the real one, there will be differences in the website's URL. Since URL is unique to each website, the fake website will never have same URL as the real website and as such must resort to changing parts of domain, lengthen it or shorten it. e.g. www.mybonk.com as opposed to www.mybank.com, another example being www.ebay.net which is different from www.ebay.com. Phishing also comes in the form of email containing a link to fraudulent websites to try and get user's information. Not unlike fake website, the user can avoid this if the user pay attention to the sender of the email. Often, official company email does not use @gmail.com or @yahoo.com but instead @arpatech.com for Arpatech company for example.



Anatomy of a Phishing Kit

Before we dive into the results, it's important to talk a bit about how phishing kits work:

1. The legitimate website is cloned

2. The login page is changed to point to a credential-stealing script

3. The modified files are bundled into a zip file to make a phishing kit

4. The phishing kit is uploaded to the hacked website, files are unzipped

5. Emails are sent with links pointing to the new spoofed website

Illustration 1.8: General steps of phishing, taken from https://www.csoonline.com/article/2117843/what-is-phishing-how-this-cyber-attack-works-and-how-to-prevent-it.html

Phishing generally hinges on the fact that the user does not notice the scam, if the user realizes before it's too late, the attempt failed. However, should the scam succeed, depending on the countermeasure the company placed to prevent frauds the company can prevent the fraudulent transaction from being completed, thus preventing the incoming loss.

There are several patterns of suspicious activity indicating a fraud[3] that a company can use to detect fraudulent transactions on their platform, those being:

1. **Orders from same account but different credit card**. In this case, the fraudster likely has many stolen credit card information already and while there are some customers who has several credit card information, the company is sure to keep track of what credit card this particular customer usually uses.
2. **Multiple transaction in single credit card in a short span of time**. Some customers do place few orders closely together, they may forget an item in their first transaction, or have someone to split the bill with. But when there are more than 5 or 6 transaction with only small items on them, it's most likely a sign of fraudulent activity.
3. **Multiple transaction with different account/ credit card with same IP**. A regular customer may have more than 1 credit card and account, so the exact number may differ in each company, But, when the number of different combinations goes over the usual, the fraudster likely has many credit card and account information and is trying to make purchases using them.
4. **Too many declined transactions**. Customers declines a lot of transaction, they may forget an item, or orders an item with the wrong color. However, when the decline count is too much, it may be a sign of a fraudulent activity.

5. **Sudden change in customer address**. This can be either the customer is changing address physically or is on a trip or a fraud attempt, in this scenario it’s best to contact the customer first to check their identity.
6. **Phone number does not match billing address area or is far away**. This is usually the case when the fraudster sends the goods to someone in their ring or to a customer
7. **Large order from unlikely customer**. Companies usually keep track of their customer’s spending habits and/ or have contact with new customer before they make a big purchase. When a customer who usually buys 100$ worth of goods at most suddenly starts spending 10.000$ it’s definitely suspicious.
8. **Unusually large order with expedited shipping**. Expedited shipping is costly, a real customer would not want to pay several times their purchases just to send it quicker unless they’re extremely impatient and have too much money to spare. A fraudster on the other hand without second thought will pay for it since it is not their money, and the faster they receive the goods, the more likely they can get away with the transaction

When the company spots these kinds of suspicious activities, they may stop the transaction and try to contact their customer to confirm their identity. Another way is to rely on AI to spot and halt the transaction automatically, it can be either a NN or a classification algorithm, their choices depends on how they want to tackle the problem.

1.2 Problem Formulation

As discussed before, fraud prevention can be done by AI automatically by halting the transaction if the transaction is flagged as such. There are several ways to implement this, one way is to tally points based on what rule the transaction broke. For each rule broken the transaction’s suspicious score goes up, and after passing certain threshold the transaction is then halted. The other way is to scan the transaction as a whole then classify it as a legitimate or fraudulent transaction based on past fraudulent transaction or user activity. Both of these requires a classification algorithm to work, classification algorithm works by examining the features of the data then try to classify it to one of predefined set of classes [4]. In this paper, the author intends to compare few classification algorithms within several scenario, the scenarios being: original data, no duplicate phone number, no product type, no date related feature. The questions to the problems above are as such:

1. Which algorithm is the most accurate on imbalanced dataset.
2. What feature is important or is a common denominator that separates legitimate and fraudulent transactions.
3. Is ML a good approach to prevent frauds.

1.3 Scope

In this research, the author has set these limitations:

1. Dataset is contained in a csv file
2. Dataset rows totaling at 40761
3. Fraudulent transaction count is 54; 27 is the original, and the other 27 is a duplicated data made with pattern closely resemble the original fraudulent transactions.
4. Python is the programming language used.
5. This research tests the accuracy of 4 different models:
    a) RF
    b) RF + PCA
    c) SVM
    d) SVM RBF

6. Data in csv is unformatted for ease of reading, the labelling is done when the program starts before being fed to the algorithm.
7. Accuracy is not used as the evaluation; recall and precision is.

1.4 Objective

1. To give alternative fraud prevention and identification method.
2. To give additional option for fraud prevention system in XYZ company.
3. To shed light on how accurate a classification algorithm can classify on particular fraud scenario in XYZ company.