# COMBATING FRAUD BY LEVERAGING MACHINE LEARNING AND HUMAN BEHAVIOR

Casey Astiz

Adviser: Professor Michael Linderman

A Thesis

Presented to the Faculty of the Computer Science Department

of Middlebury College

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Arts

December 2018

# ABSTRACT

I am writing about fraud and how to detect it! Wooh!

# ACKNOWLEDGEMENTS

I would like to thank Professor Linderman and Professor Scharstein first and foremost for their help in this Computer Science Thesis. Without them, this thesis would cease to exist. I would also like to thank the Computer Science department at Middlebury College for giving me the opportunity to learn and develop as a computer scientist. Last, I would like to thank all my unnamed interviewees who were willing to discuss their expertise about fraud detection with me. I truly appreciate all who helped create this thesis, and add to my knowledge!

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

CHAPTER 1

# INTRODUCTION

In today's world, there are more and more ways to pay for goods and services. Cash is a way of the past; mobile, peer to peer, and online payments are the way of the future. While these convenient payment methods are very exciting, the additional methods of payment lead to more gaps in security, in particular fraud. Fraudsters have taken advantage of every system in some form or another, and it is the financial institution's job to block or detect these fraudsters or else deal with the payments. For reference, credit card fraud through online payments in Australia hit $476 million in 2017, rising $58 million from the year before [5]. This level of fraud is happening all over the world, and it is necessary to find more efficient and effective methods to catch fraudsters to keep everyday financial transactions safe. In addition, when banks lose money to fraudsters, card holders partially or entirely pay for this loss through higher interest rates, fees, or reduced benefits. Without automatic fraud detection, a significant amount of overhead is created because of the need to investigate these transactions [2]. It is important to limit the amount of fraud because consumers pay for this fraud.

The main purpose of this thesis is to investigate current methods of fraud detection and implement one or more of those methods in a financial context. I will discuss the most prevalent methods for fraud detection are currently, as well as past fraud detection methods. Fraud detection is a difficult field of study because it falls in the subset of anomaly detection and has very skewed classes as more transactions are not fraud than are fraud, and the classification of fraud is always changing. In order to build an effective fraud detector, it must be adaptive to new patterns and categories of fraud that are developing as fraudsters find new methods to deceive the system.

The main focus of this thesis will be a review of different methods used in the industry in the past and today for fraud detection, as well as my own implementation using some dataset of transactions to be determined. This background section will include information from published research papers as well as information I learned from interviewing people at different companies in this field to provide some context for the literature review. Comparing business interactions with fraud to published research in the field will allow me to somewhat analyze the performance of different machine learning methods in real life versus a research context.

Chapter 2 discusses past and current research into the field of fraud detection. Chapter 3 gives context for the problem at hand, synthesizing the information I learned from interviewing professionals in the fraud world. In Chapter 4, I describe the methodology and experiments I ran.

CHAPTER 2

## MODERN DAY FRAUD

What exactly is fraud? Fraud is defined as "wrongful deception with the intent to gain personally or financially," or intentionally deceiving another person to obtain something they have [1]. Fraud can either be a civil or criminal offense, depending on the state. There are a multitude of types of frauds, from credit card fraud and website misdirection to pyramid schemes and insurance fraud. There has been much research into the best and most efficient ways to detect different kinds of fraud. The goal of this research is to detect the most amount of fraud with the least error. This may seem obvious, but is particularly important because this research can be applied to real scenarios. It is important for business to balance finding all fraud and limiting their customers. A company may choose to accept a small level of fraudulent transactions in order to include as many real transactions as possible. For example, it is less expensive for a company to let through a new customer that is potentially fraudulent occasionally, and maximize the number of new customers that are not fraudsters. The opportunity cost of rejecting every potential fraudster is too high for most companies to risk. For that reason, fraud detection out in the wild is never going to be at 100 percent because the false positives are too big a cost.

However, in a research context, there are no such limits. The previous fraud research has covered many different types of fraud, from broad to narrow scenarios. Researchers are able to manipulate datasets to have ideal fraud to valid transaction ratios, and find the place their algorithm performs the best. Since researchers do not need to make a split second decision for whether or not to deny a transaction, they have time to find the best algorithm for the specific problem they are working on. By examining past and current research in this field, I have created a synopsis of information about fraud detection, and will implement a similar approach to one of the papers discussed in Chapter 4.

## 2.1 Credit Card Fraud Research

An early example of credit card fraud detection comes from a study by Chan and Stolfo from 1998 [2]. This paper, titled “Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection” seeks to find a solution to classifying skewed classes. Here, the number of fraudulent transactions is relatively small compared to the legitimate ones, and the amount of financial loss for each fraudulent transaction depends on the amount of the transaction and other factors. However, millions of transactions occur every day, and only a small portion of these transactions are fraudulent. Skewed classes are also the norm for cellular phone fraud detection as well as natural language processing problems.

In order to solve the problem, Chan and Stolfo implement a cost-sensitive approach instead of a classic error rate. They use a multi-classifier meta-learning approach, where they create data subsets with appropriate class distributions and apply learning algorithms to the subsets. Then, they integrate and optimize the cost performance of the classifiers by learning from their classification behavior. The benefit of this approach is that it handles non-uniform cost per error and is cost sensitive during the learning process. Chan and Stolfo use a dataset from Chase Manhattan Bank containing half a million transactions, from which 20 % are fraudulent, and implemented a cost model based on information from a bank representative.

Given this data and information, the authors experimented with the effects of training class distributions on the credit card cost model. They used the first 10 months as training data and the 12 month for testing, and randomly sampled transactions to get varied fraud amounts. The multi-classifier meta learning approach works by randomly dividing the dataset into 4 subsets, and then learns on the subsets using one of four algorithms: CART, C4.5, RIPPER, or BAYES. The meta learning then learns from that class combiner behavior and uses that as a class-combiner strategy. The authors found that the training class distribution can affect the performance of the learned classifier, which is an issue given that natural distributions can be different than desired training distribution. However, they found their method to be efficient and scalable for large datasets. Since thieves learn and fraud patterns evolve over time, and adaptive classifier selection method is essential. This modular multi-classifier approach facilitates adaptive over time and removal of out of date learned behavior.

### 2.1.1 Online Payments Fraud Research

## 2.2 Other Fraud Research

While this thesis mainly focuses on fraud research in a credit card and online payments format, there has also been much research into fraud in other forms. However, the broader goals of this research align with the goals of credit card fraud detection, and the results are widely applicable. An example of this is a paper by Fawcett and Provost from 1996. This paper, "Combining Data Mining and Machine Learning for Effective User Profiling," describes automatic methods for fraud detection based on profiling customer behavior [3]. Instead of focusing on credit card data, this paper uses cellphone cloning data, which is a particularly expensive type of fraud. This is when a customer's Mobile Identification Number and Electronic Serial Number are cloned and programmed into a cellphone that does not belong to that customer.

In this paper, the authors present a framework for automatically generating fraud detectors. Here, the model uses a lot of calls to determine fraudulent patterns then use these broader patterns and apply them to individual accounts. Some of the standard methods used to detect fraud in these calls are to search for collisions or overlapping calls between the original user and the fraudster, or to search for calls in terminal proximity that could not have possibly been placed by the same user. The profilers the paper describe capture the typical behavior of an account and calculate how far an account is from typical behavior. Given this information, the authors use rule learning program, which searches for rules with certainty factors avoe a user-defined threshold. Here, each account has its own set of rules, and each call has 313 attributes that allow for partitions of the calls. Through their data mining process, 3630 rules were created for detecting fraud, which were then narrowed down to 99 rules. The conclusions of this paper were that the authors had to sacrifice accuracy in order to reduce the total cost of the system, but that their method was overall effective [3]. Like other papers, the authors highlight the need to build adaptive systems to account for ever changing fraud.

## 2.3 The Human Element

In this section I will discuss some game theory and behavioral elements of fraud and ways of eliminating frauds and fraudsters.

CHAPTER 3

# FRAUD IN CONTEXT

## 3.1 Interviews from Professionals

This section will be about what professionals are doing in the fraud field.

## 3.2 Research Versus Reality

This section I will compare whatever I find out from professionals to the most modern research on this topic. I assume research will be slightly behind what these professionals are dealing with and how they are solving these problems.

CHAPTER 4

# IMPLEMENTING A FRAUD DETECTION SYSTEM

## 4.1 Model

This section is where I will talk about the overall system I am planning on implementing.

## 4.2 Experiment

### 4.2.1 Hypothesis

### 4.2.2 Data

The data used from this project comes from a public source on Kaggle.com. I used the “Synthetic Financial Datasets for Fraud Detection” which was generated by the PaySim mobile money simulator [4]. This dataset comes from the paper “PaySim: A financial mobile money simulator for fraud detection” and is scaled down to a quarter of the original dataset. The dataset includes information about type of the transaction, amount, the customer that started the transaction, the initial balance of the sender, the balance after the transaction of the sender, whether its fraud, and other variables. The summary statistics are shown in Table 4.1.

Table 4.1: Summary Statistics

| | 2009 to 2015 | | | |
|---|---|---|---|---|
| Variable | mean | s.d. | min | max |
| *Measures of abortion access* | | | | |
| Distance (hundreds of miles) | 0.967 | 0.72 | 0.003 | 3.12 |
| I(distance<50 miles) | 0.28 | 0.45 | 0 | 1 |
| I(50< Distance ≤ 100) | 0.37 | 0.48 | 0 | 1 |
| I(100< Distance ≤ 150) | 0.16 | 0.37 | 0 | 1 |
| I(150< Distance ≤ 200) | 0.06 | 0.24 | 0 | 1 |
| I(200 < Distance) | 0.12 | 0.32 | 0 | 1 |
| Clinics | 0.144 | 0.817 | 0.0 | 10 |
| Average Service Population (100,000s) | 2.01 | 1.19 | 0.144 | 5.73 |
| *Measures of family planning access* | | | | |
| Distance (hundreds of miles) | 0.39 | 0.401 | 0.002 | 2.63 |
| I(distance<50 miles) | 0.75 | 0.43 | 0 | 1 |
| I(50< Distance ≤ 100) | 0.16 | 0.37 | 0 | 1 |
| I(100< Distance ≤ 150) | 0.05 | 0.22 | 0 | 1 |
| I(150< Distance ≤ 200) | 0.02 | 0.15 | 0 | 1 |
| I(200 < Distance) | 0.01 | 0.08 | 0 | 1 |
| Clinics | 0.75 | 2.33 | 0.0 | 29 |
| *Race* | | | | |
| White | 53.85 | 21.17 | 2.45 | 91.53 |
| Black | 6.66 | 7.53 | 0 | 40.12 |
| Hispanic | 37.65 | 23.57 | 2.79 | 97.03 |
| Other | 1.82 | 2.048 | 0 | 21.23 |
| *Economic conditions* | | | | |
| Median Income | 4.88 | 1.12 | 2.33 | 9.15 |
| Poverty Rate | 17.54 | 6.32 | 0 | 42.73 |
| Population in Urban Areas | 13.71 | 30.1 | 0 | 99.31 |
| Unemployment rate | 6.435 | 2.27 | 2 | 18.5 |
| *Abuse Rates per 1000 Children* | | | | |
| Unconfirmed | 37.87 | 14.78 | 0 | 105.26 |
| Confirmed | 13.10 | 8.34 | 0 | 122.66 |
| Total Victims | 50.97 | 20.11 | 0 | 226.61 |
| *Population Statistics* | | | | |
| Child Population | 27,759.88 | 101,386.3 | 19 | 1,224,413 |
| Highschool Degree (Over 25) | 77.99 | 8.186 | 44.877 | 94.02 |
| College Degree (Over 25) | 17.54 | 7.09 | 3.66 | 6671 |

### 4.2.3 Methods

LOTS of Machine Learning incorporated into bigger fraud detection and rule based learning systems! Overall system will be discussed in Model section.

### 4.2.4 Results

This section will be full of tables and figures. There will probably also be an appendix with the entire set of experiments run. The most interesting or a subset of the experimental runs will be in a table in this section.

### 4.2.5 Discussion

This section will compare my hypothesis to the results, as well as my results to the model I'm basing it off of. I also plan on adding any notes I learn along the way of this implementation.

CHAPTER 5

# FURTHER APPLICATIONS

## 5.1 Anomoly Detection

### 5.1.1 Cell Phone Fraud

### 5.1.2 Biometric Fraud

### 5.1.3 Bioinformatics

CHAPTER 6

# CONCLUSIONS

# BIBLIOGRAPHY

[1] Fraud - Definition, Types, Examples and Processes.

[2] Philip K Chan and Salvatore J Stolfo. Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. Technical report.

[3] Tom Fawcett and Foster Provost. Combining Data Mining and Machine Learning for Effective User Profiling. Technical report, 1996.

[4] Paysim. Synthetic Financial Datasets For Fraud Detection — Kaggle.

[5] Mac Wang. How to protect your business from online payments fraud - CSO — The Resource for Data Security Executives, 2018.