Big Data Applications in Fraud Detection in Insurance

Mrunal L Chaudhary
Indiana University
Bloomington, Indiana
mchaudh@iu.edu

ABSTRACT

Insurance companies today are incurring a loss in billions of dollars every year because of frauds happening in filing claims, paying premiums, filling applications etc. Detecting frauds manually or by other traditional means is an impossible task since magnanimous amounts of data is getting generated every day and fraudsters change their strategies very quickly. For handling such a humongous amount of data, performing real time analysis on it, and getting accurate outputs; it is imperative that a robust, flexible and scalable technology be used which can detect frauds on the fly. Big data provides just the platform needed to perform analysis of such high complexity.

KEYWORDS

i523, HID205, Insurance, Fraud detection in insurance, Predictive analysis

1 INTRODUCTION

The fact that technology is evolving at the fastest pace in the history of mankind needs no more of a proof than a mere glance of eyes around our surroundings. But like everything else, it is a double-edged knife and this advancement in technology comes at a cost of benefiting the fraudsters of the society. A fraud is a wrongful or criminal deception intended to result in financial or personal gain. And while they are rampant in almost every field, there is no surprise that the field of insurance too has been affected by them. Traditionally, Insurance industry estimated that frauds account for about 10 percent of the total losses incurred by them which is equivalent to billions of dollars, and these numbers are only rising [4]. As these fraudsters are getting smarter, and well equipped with technology, insurance companies are facing difficulties in preventing and detecting the fraudulent activities with the traditional ways. Though the growth of modern technology aids the fraudsters in generating sophisticated fraud techniques, the advancement in technology has enabled better and smarter approaches in detecting fraud. Today when data is getting generated at a break-neck speed and transactions are digitally documented in some format, evidence is just hidden in the data for aiding the investigators to control the fraudulent activities [5]. The question then that needs to be addressed is, how to find that evidence easily [5].

2 BACKGROUND

Most of the insurance frauds are committed in the fields of Health care, auto insurance and workers' compensation. These frauds can either be categorized as hard or soft. Hard frauds occur when someone on purpose fabricates an accident or makes up non-existing claims. Soft fraud occurs when they inflate the amount of a legitimate claim [3]. The parties that are most affected by frauds are the loyal consumers who have to pay higher insurance premiums for compensating the losses from frauds, and medical professionals who are concerned of tarnishing their reputation [4]. The people who do insurance frauds can either be organized criminals who draw large sums of money from fraudulent claims, professionals who add on to legitimate claims or ordinary people who just want to cover the amount of premiums or deductibles [3].

3 REASONS FOR FAILURE OF TRADITIONAL APPROACHES

The natural question then that comes to the mind is, "Why are the traditional approaches to detect fraud inadequate?" The rate at which this voluminous data is getting generated is just impossible to efficiently and accurately process manually. The data sources that get generated were far too large and were changing far too often so as to be helpful in scoring the fraud the traditional way, since they used batch processing which took hours or even days together to run [9]. Moreover, the insurance firms used red-flag method for suspicious claim detection. Thus, using sampling techniques was bound to skip some frauds and introduce errors [7]. Also, the data generated by insurance companies is ever evolving and changing, and the traditional approaches are not sustainable to process data at real time [10]. Thus tweaking of parameters in the fraud detection algorithm was an impossible task since it requires a lot of processing time. Hence, traditional approaches to fraud detection lacked both- the flexibility and the scalability [7]. And lastly, traditional approaches only looked at the structured data since the fraud detection systems were not equipped to handle unstructured data, thereby eliminating a huge subset of data, and making the critical decisions on incomplete information [1].

4 CHALLENGES FACED IN FRAUD DETECTION

It is widely recognized that the volume of data is increasing at breakneck speed. In fact experts believe that the amount of data that has been amassed in the last two years- a zettabyte- is more than the data that has been generated since the dawn of the human civilization [8]. But more astonishing than this is that the volume is only one part of the entire equation. With volume, big data deals with velocity and variety as well. Insurers can access a variety of information that is ever increasing and can be accessed through social media and customer feedback and reports in the form of unstructured data. Also photos and videos are another rich source of visual media information that the insurers can lay their hands over. And this information is ever changing and increasing. Thus,

dealing with the velocity and variety of the voluminous data are challenges in themselves [7].

5 ADVANTAGES OF USING BIG DATA IN FRAUD DETECTION

One of the biggest advantage that high-performance analytics allows is the ability to use the ever-changing and increasing data sources previously ignored because of the lack of sophisticated tools for handling big data [7]. With the advent of high-performance analytics, billions of rows of data can be processed in a matter of seconds. Thus, insurers can determine fraud scoring in real time. Insurers can now test certain approaches in real time and constantly tweak the parameters for maximizing the output of the fraud detection algorithms. Since high performance analytics can process a magnanimous amount of data in mere seconds, sampling of data is no more needed. Hence the error introduced by sampling can be completely avoided by using big data technologies [7]. With the help of high performance analytics, advanced models like supervised predictive modeling, data mining, social network analysis, social customer relationship management, etc. can be implemented to improve the process of analysis [10]. With the arrival of big data, the process of fraud detection can be completely revolutionized. 'High-performance analytics showcases ways in which organizations can now capture data and use it to their benefits. It has revolutionalized the ways in which companies manage data, especially in fraud [7].

6 ROLE OF ANALYTICS IN FRAUD DETECTION

Traditionally, insurance companies have used Statistical models for the identification of frauds in claiming insurance policies. The problem with the traditional way was that it worked in silos, and hence are not scalable to handle the rapidly growing information from different sources [10]. Analytics therefore play an important role in fraud detection by addressing these issues. The key benefits are

(1) Analytics help in the detection of low key incidence events
(2) Analytics helps in effective integration of data.
(3) Since most of the data related to fraudsters like third party reports, is available in unstructured format, text analysis can play an important role in providing valuable insights in fraud detection [10].

7 INNOVATIVE FRAUD DETECTION METHODS

7.1 SNA (Social Network Analysis)

"SNA allows the company to proactively look through large amounts of data to show relationships via links and nodes" [6]. SNA tool combines many analytical methods such as business rules, statistical methods, pattern analysis, and network linkage analysis for uncovering the data to show these relationships [10]. Take a straight forward case of an accident claims made by a victim. Though it may look simple in the beginning, SNA can reveal that the address given by the victim or the car involved in the accident was in fact used in multiple other claims, suggesting a fraudulent activity [6].

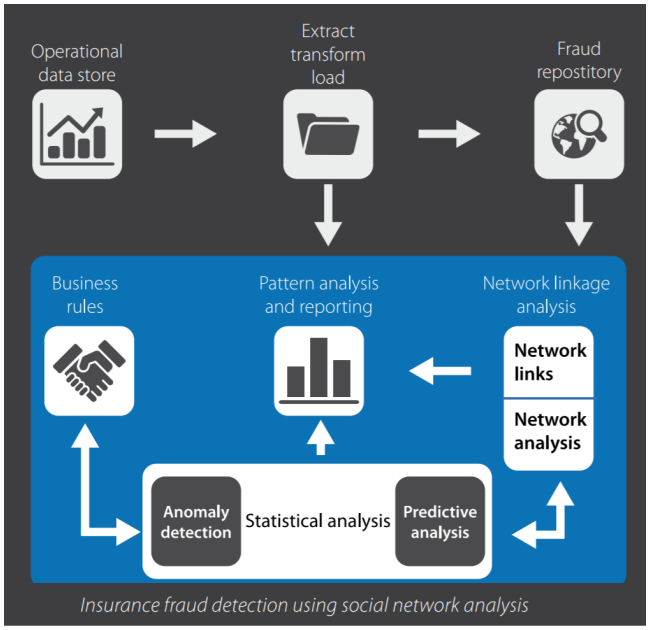

Figure 1: Social Network Analysis Flow chart [10].

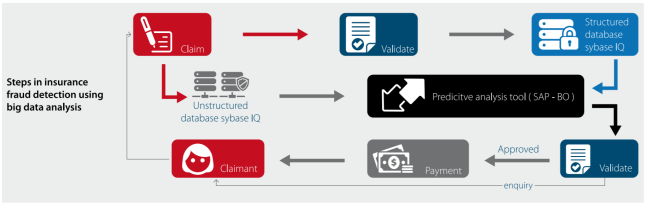

Figure 2: Predictive Analysis Flow chart [10].

SNA works like this: After feeding the data into an ETL (extract, transform and load) tool, the Analytics team scores the risk of fraud by prioritizing the likelihood based on the history of the claimant, relationship with other fraudsters, multiple claims, etc. The Fraud Identification and Predictive Modeling process is an integrated framework of technologies like sentiment analysis, text mining, content categorization, social network analysis. Thus, by doing this, the insurer can rate each claim. A fraudulent claim can therefore be indicated by a high rating. The correctly identified frauds are then added into the business use case system [6].

7.2 Predictive Analysis for Big Data

Predictive analysis makes use of text and sentiment analysis to go through big data for fraud detection [6]. Big data helps in proactively detecting the fraudulent cases by quickly sifting through the large claim reports which are unstructured in nature. An important point to note is that people committing frauds mostly alter their story with time. And these clues are hidden in the log reports submitted by the claim adjusters [10]. The computing system based on the business rules can spot the evidence of the fraudulent claims easily with the help of text and sentiment analysis [6].

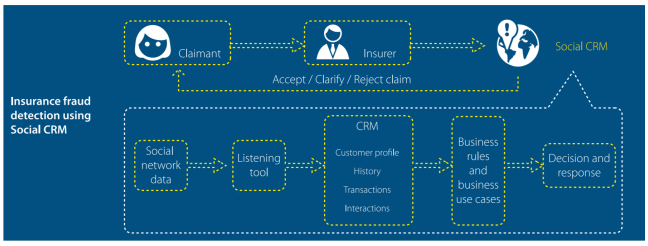

Figure 3: Social Customer Relationship Management Flow chart [10].

7.3 Social Customer Relationship Management

The SCRM is a process that Companies should follow to link their CRM with social media for better understanding of customer behaviour and demands and general trends [6]. The reference data which the company extracts from social media chatter using a 'listening tool', along with information stored in the company's existing CRM is fed into a case management system. This system then sends a response about whether the claim is fraudulent or not, which is confirmed by the investigators [10].

Most of the Insurance Fraud detection tools build a framework around the claim management vertical, but for building a more robust system, a holistic framework is needed, one which can identify potential areas for frauds like in application, premium, claims, etc. Following are the 10 steps for implementing the analytics of fraud detection [10].

8 10-STEP IMPLEMENTATION OF ANALYTICS IN FRAUD DETECTION

8.1 Performing SWOT analysis

Due to the increasing awareness of the fraud detection systems, before going for a solution, the insurance company should perform a SWOT analysis of the existing solutions and choose the most suitable one.

8.2 Building a team dedicated for fraud detection

It is important to form a dedicated team for fraud detection which can handle the responsibilities of fraudulent claims going unnoticed otherwise.

8.3 Building a Solution versus buying it

In case the insurance company decides to go for building a solution, they need to be sure that they have the required skill set for building an in-house analytics product. And if not, then it should find an analytics solution that best fits its requirements.

8.4 Data cleaning

Databases should be integrated and redundancies should be removed from data.

8.5 Coming up with relevant business rules

Certain types of frauds are very industry or company specific and this knowledge can only be gained through experience. The insurance companies should use the existing in-house expertise to define the business rules.

8.6 Defining the thresholds for Anomaly Detection

The insurance companies need to carefully set the threshold for Anomaly detection after considering factors like type of insurance, rate of fraudulent claims, time available etc. If set to a high value there is a chance that many fraudulent claims might go unnoticed and if set to a very low value, it might culminate into wastage of time.

8.7 Using Predictive Modelling

Data mining tools should be utilized for building models that can produce scores for fraud propensity in regard with unidentified metrics. The result thus can then be given for further analysis.

8.8 Using Social Network Analysis

SNA models relationships between various entities involved in claims and therefore, has proved to be very helpful in identifying the fraudulent activities. These entities can be anything from geographic location, car in case of car accidents, age groups, financial status, phone numbers, etc. Through SNA it can be found out that the linkage between some entities is higher than the average connection numbers, thus indicating fraudulent activities.

8.9 Building a Social Customer Relationship Model

The integrated case management systems allows the insurance companies to capture relevant findings to claims data and ensures efficiency and proper assessment of investigations.

8.10 Forward looking Analytic solutions

For building a truly robust system the insurance companies should keep looking for additional third-party sources of data and their integration with existing solutions for increasing the efficiency of the fraud detection system. Also, they should always keep in mind the issue of scalability. With the ever-increasing data, the system should be such that it can handle the size of the data [10].

9 CONCERNS RELATED TO FRAUD DETECTION SYSTEMS

Though Fraud detection system have become pretty robust and sophisticated, there are still come issues that need to be addressed which are the following:

- The data that is collected by the insurers can be properly utilized. The biggest hurdle though in their way is legislative barriers and privacy protection laws that hinder the analysis of the insurance companies [1].
- No matter how advanced the fraud detection tools becomes, there will always be a dependency on humans for converting the reports into actionable intelligence [1].

- Most of the modelling techniques are highly dependent on the past behaviors of the fraudsters. But their behavior changes so quickly that it makes the whole analysis worthless. Evaluating the quality of the data therefore is a huge struggle [2].
- The loses incurred by the insurance firms are somewhat compensated by charging a higher amount for premiums and taking more time, thus the insurance firms may lose out on loyal customers [1].

10 CONCLUSIONS

The upsurge of analytics presents a world of limitless potential for insurance companies which have long held a foundation of information. With the advent of big data, high-performance analytics technology represents an opportunity to completely revolutionize the way fraud is detected. Though Big Data applications in Insurance are still in the early stages, they have proved to be powerful to easily handle and process the velocity with which variety of voluminous data is getting generated. In the years to come, Big Data Analytics has showcases the potential to find widespread applications in the field of insurance.

REFERENCES

[1] Insurance fraud EU. 2016. The Role of Data and Analytics in Insurance Fraud Detection. (June 2016). https://www.insurancenexus.com/fraud/role-data-and-analytics-insurance-fraud-detection

[2] Friss. 2017. The 8 Biggest Fraud Challenges for Insurers. (Feb. 2017). https://www.friss.com/en/news/the-8-biggest-fraud-challenges-for-insurers/

[3] Insurance Information Insitute. 2017. Background on: Insurance fraud. (Sept. 2017). https://www.iii.org/article/background-on-insurance-fraud?cm_mc_uid=96455616751215069824528&cm_mc_sid_50200000=1507445074&cm_mc_sid_52640000=1507445074

[4] Kim Minor. 2013. Improving Claims Fraud Detection in Insurance. (May 2013). http://www.ibmbigdatahub.com/blog/improving-claims-fraud-detection-insurance

[5] Paul Nelson. 2017. Fraud Detection Powered by Big Data - An Insurance Agency's Case Story. (2017). https://www.searchtechnologies.com/blog/fraud-detection-big-data

[6] Sachin Pandhare. 2017. Big data Analytics: new whistleblower on insurance fraud. (2017). https://www.infosys.com/industries/insurance/white-papers/Documents/new-whistleblower-insurance-fraud.pdf

[7] James Ruotolo. 2013. Big Data for Fraud Detection. (May 2013). http://www.insurancetech.com/big-data-for-fraud-detection/a/d-id/1314553?

[8] Jonathan shaw. 2014. Why 'Big Data' Is a Big Deal. (March 2014). http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal

[9] Ikanow Editorial Team. 2014. HOW CAN I USE BIG DATA ANALYTICS FOR FRAUD DETECTION? (Feb. 2014). http://www.ikanow.com/how-can-i-use-big-data-analytics-for-fraud-detection/

[10] Ruchi Verma and Sathyan Ramakrishna Mani. 2013. Using Analytics for Insurance Fraud Detection. (Dec. 2013). https://www.the-digital-insurer.com/wp-content/uploads/2013/12/53-insurance-fraud-detection.pdf