Research on Insurance Fraud Detection based on Big Data

Yanrong Wu 1801212952

Insurance fraud is an act or omission intended to gain dishonest or unlawful advantage, either for the party committing the fraud or for other related parties. Broad categories of fraud include:

- ***Policyholder fraud and claims fraud*** — Fraud against the insurer in the purchase and execution of an insurance product, including fraud at the time of making an insurance claim.
- ***Intermediary fraud*** — Fraud perpetuated by an insurance agent, corporate agent, intermediary, or third party-agent against the insurer or the policy holders.
- ***Internal fraud*** — Fraud against the insurer by its director, manager, or any other staff or office member.

Data processing has always been at the very core of insurance business; traditional datasets such as demographic data, exposure data or behavioral data have historically been processed by insurance firms to inform underwriting decisions, price policies, evaluate and settle policyholders' claims and benefits, as well as to detect and prevent fraud. In the era of digitalization, these tradition- al datasets are increasingly combined with new types of data such as Internet of Things (IoT) data, online data, or bank account / credit card data in order to perform more sophisticated and comprehensive analysis, in a process that is commonly known as 'data enrichment.'

1 Big data characteristics (3Vs)

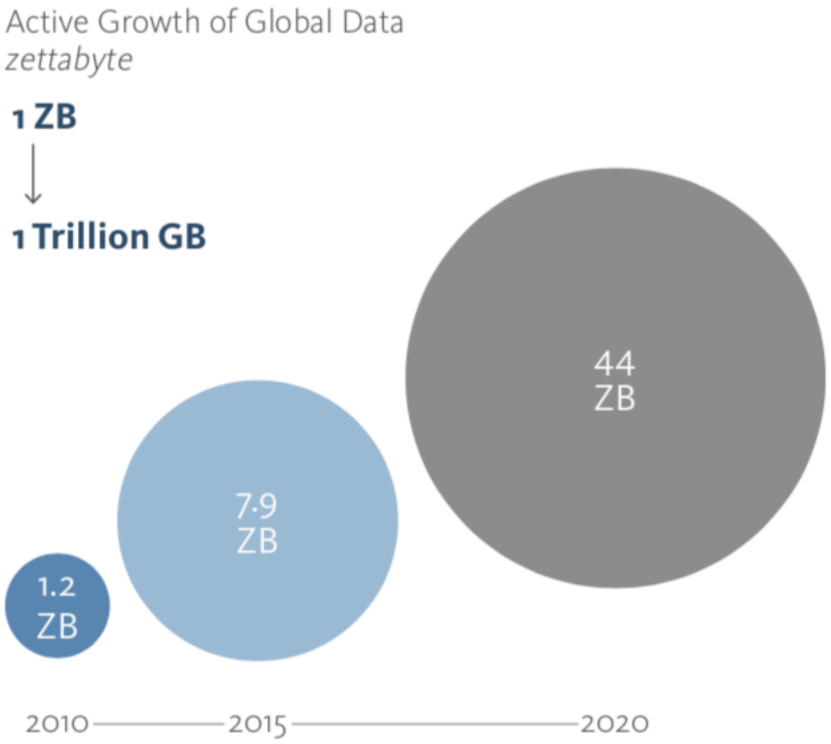
Volume：

The scale of the insurance industry data is very large. First of all, the scale of the transaction data itself is very large.

In China, in 2018, there were 110 million new life insurance policies, 300,000 per day, 13,000 per hour, and 3.5 per second. This is only life insurance. The number of all insurance (health insurance, accident insurance, and property insurance) policies is much larger than life insurance.

The data in the insurance industry is not limited to the transaction data itself, not just the data in various documents filled in for business. There is also data from all user behavior. There is no doubt that the insurance industry data is large enough.

Figure 1: Increasing availability of data

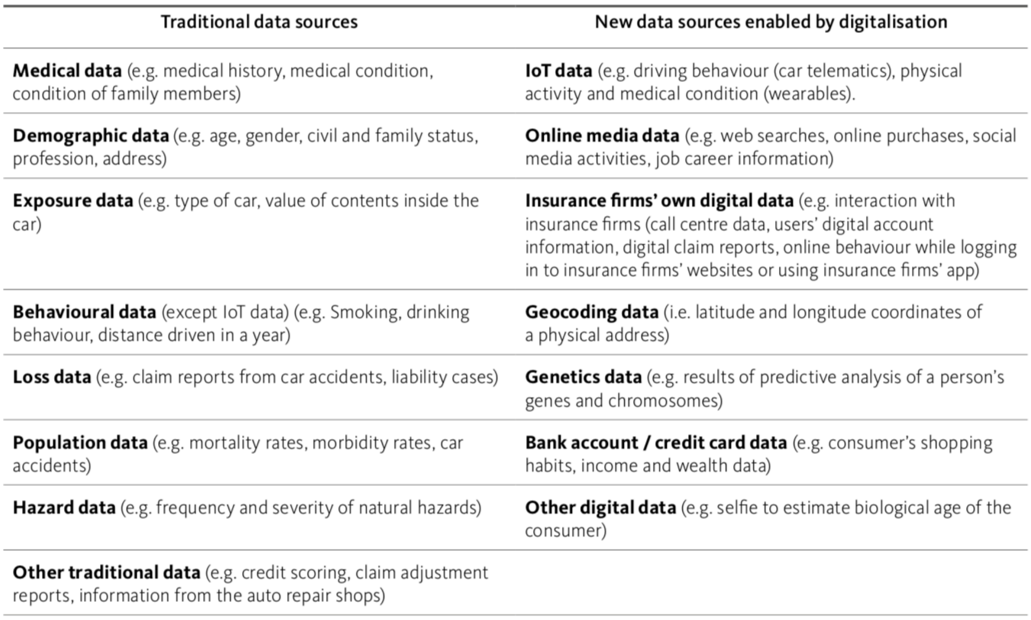


Source: Institute of International Finance

Variety:

The data used by insurance firms in the different stages of the insurance value chain may include personal data(e.g. medical history) as well as non-personal data (e.g. hazard data), and it can be structured (e.g. survey, IoT data) or unstructured (e.g. pictures or e-mails). It can be obtained from internal sources (e.g. provided directly by the consumer to the firm) as well as from external sources (e.g. public databases or private data vendors).

Chart 1: Data Source



Traditional data sources	New data sources enabled by digitalisation
Medical data (e.g. medical history, medical condition, condition of family members)	**IoT data** (e.g. driving behaviour (car telematics), physical activity and medical condition (wearables).
Demographic data (e.g. age, gender, civil and family status, profession, address)	**Online media data** (e.g. web searches, online purchases, social media activities, job career information)
Exposure data (e.g. type of car, value of contents inside the car)	**Insurance firms' own digital data** (e.g. interaction with insurance firms (call centre data, users' digital account information, digital claim reports, online behaviour while logging in to insurance firms' websites or using insurance firms' app)
Behavioural data (except IoT data) (e.g. Smoking, drinking behaviour, distance driven in a year)	**Geocoding data** (i.e. latitude and longitude coordinates of a physical address)
Loss data (e.g. claim reports from car accidents, liability cases)	**Genetics data** (e.g. results of predictive analysis of a person's genes and chromosomes)
Population data (e.g. mortality rates, morbidity rates, car accidents)	**Bank account / credit card data** (e.g. consumer's shopping habits, income and wealth data)
Hazard data (e.g. frequency and severity of natural hazards)	**Other digital data** (e.g. selfie to estimate biological age of the consumer)
Other traditional data (e.g. credit scoring, claim adjustment reports, information from the auto repair shops)	

Source: The Geneva Association

Velocity：

Earlier I have mentioned that life insurance is 3.5 policies per second, this number does not seem to be fast enough to generate data.

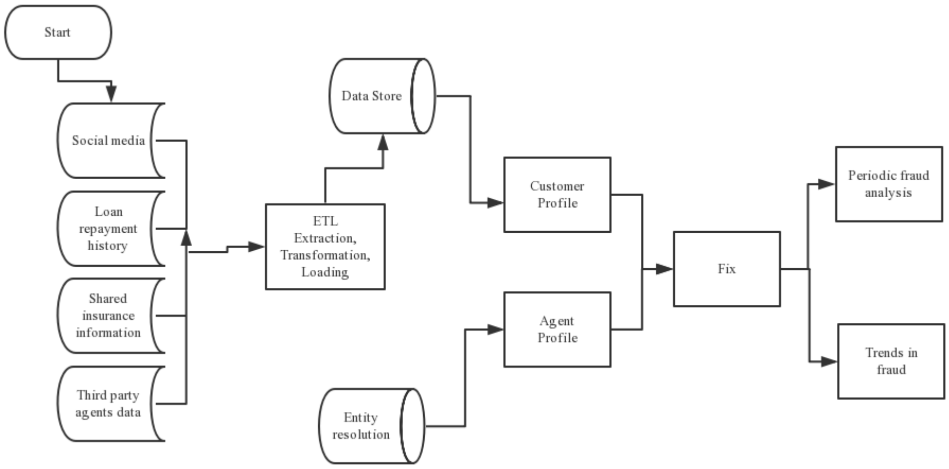
Let's look at the phone sales. A rough estimate is that if a company has 30,000 sales in the life insurance telemarketing industry, it will make 8 hours of phone calls per day and generate 1M audio files based on 3-5 minutes, which is about 300M audio per second.

2 Solution

Because a large volume of varied data from many sources must be collected, stored, and processed, this business challenge is a good candidate for a big data solution.

The following diagram shows the solution pattern, mapped onto the logical architecture.

Figure 2: Workflow



We can use data providers from:

- External data sources
- Structured data storage
- Transformed, structured data
- Entity resolution
- Big data explorer components

The data required for insurance fraud detection can be acquired from various sources and systems such as banks, medical institutions, social media, and Internet agencies. It includes unstructured data from sources such as blogs, social media, news agencies, reports from various agencies. With big data analytics, the information from these varied sources can be correlated and combined, and — with the help of defined rules — analyzed to determine the possibility of fraud.

The required external data is acquired from data providers who contribute preprocessed, unstructured data converted to structured or semi-structured format. This data is stored in the big data stores after initial preprocessing. The next step is to identify possible entities and generate ad-hoc reports from the data.

Entity identification is the task of recognizing named elements in the data. All entities required for analysis must be identified, including loose entities that do not have relationships to other entities. Entity identification is mostly performed by data scientists and business analysts. Entity resolution can be as simple as identifying single entities or complex entities based on the data relationships and contexts. This pattern uses the simple-form entity resolution component.

Structured data can be simply converted into the format most appropriate for analysis and directly stored in big data structured storages.

Ad-hoc queries can be performed on this data to get the information like:

- Overall fraud risk profile for a given customer, region, insurance product, agent, or approving staff in the given period
- Inspection of past claims by certain agents or approvers or by the customer across insurers

I plan to use Neo4J database, a graph database. Unlike relational databases, Neo4j stores interconnected data that is neither purely linear nor purely hierarchical, making it easier to detect rings of fraudulent activity regardless of the depth or the shape of the data.

Reference

[1] Neo4J Official Website: https://neo4j.com/
[2] NoSQL Databases (2017), http://nosql-database.org/
[3] Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. Ieee Access, 5, 16568-16575.
[4] Bologa, A. R., Bologa, R., & Florea, A. (2013). Big data and specific analysis methods for insurance fraud detection. Database Systems Journal, 4(4), 30-39.