A Framework for Comprehensive Fraud Management using Actuarial Techniques

Rohan Yashraj Gupta, Satya Sai Mudigonda, Phani Krishna Kandala, Dr. Pallav Kumar Baruah

Abstract— Fraud acts as a major deterrent to a company's growth if uncontrolled. It challenges the fundamental value of "Trust" in an Insurance business. This concern must be addressed on priority else, it brings down the castle of the insurance business. The regulation provides powers to authorities to act on fraud. Currently, this effort within most organizations happens discretely, involving those unconnected with actuarial or technology. In fact, actuarial techniques are powerful tools that help to bring efficiency and to target the right areas to deploy the right level of resources for fraud investigation.

An effective solution approach to tackle this challenging problem is provided in this work which is empirically tested.

In this work, we propose comprehensive fraud management (CFM) framework using actuarial techniques and AI Technology that helps increase fraud detection rate in comparison with other proposed models available in the literature. This framework includes three stages:

- Stage 1: Automate Fraud identification using triggers specific to individual LoB and rule engine. This is a prevention stage.
- Stage 2: Integrate Statistical/ Actuarial Techniques and Technology to identify fraud. Statistical/ Actuarial techniques for fraud detection include techniques such as classification trees, logistic regression, suspicious scoring, significance testing, random sampling, clustering, linear regression, peak analysis, extreme value theory etc. Technologies that are effective in detecting fraud include machine learning, deep learning, blockchain and distributed systems etc. This is a fraud detection stage.
- Stage 3: Further analyse results from stage 2 to create a new set of fraud identification triggers. This adds on to the existing set of triggers in Stage 1 and increases the fraud detection rate in the subsequent runs of CFM.

Proof of concept presented here on Motor line of business can be tested and extended to other lines of business or industries. This should encourage companies to explore new opportunities in comprehensive fraud management by utilising actuarial skillset "carclaims.txt."

Index Terms—Comprehensive fraud management, Emerging experience, Extreme value theory, Behavioural finance, Classification trees, Logistic regression, Suspicious scoring, Spectral clustering, Peak analysis, Machine learning, Blockchain and Distributed computing

1. Introduction

Fraud is malpractice, an act of using a dishonest method that is done in order to gain some financial benefits, which are not otherwise entitled. Fraud is a major problem in many financial and non-financial sectors. This could include providing wrong (misleading) information, raising a false claim, etc.

Today, economies all over the world are plagued with fraud that is affecting various aspects of organizations ranging from financial performance to organizational morale.

Insurance fraud is not new to this world. This came into existence ever since insurance business took the form of a commercial enterprise. Amount summing up to billions are lost every year due to insurance fraud.

The insurance sector today is growing rapidly and in the process, this growth has also generated a humongous amount of data. Unfortunately, a majority of companies have legacy systems that do not capture sufficient details to identify and combat fraud. In the process, companies identify very few cases of fraud and often it is years later that these cases come into light. Some companies, on the other hand, have leveraged this data to improve their fraud management mechanisms, thereby gaining a competitive advantage over their peers.

This research paper is organised into ten sections. Section 2 explains the motivation behind this research and presents potential benefits that can be gained by executing an effective fraud management strategy. Section 3 discusses regulation specific to insurance fraud in India. Section 4 explains the approach to this research effort and reveals the concept behind the developed framework. Section 5 goes into details of each aspect of the Comprehensive Fraud Management (CFM). Section 6 explains the Business and the Technical view of the CFM framework. Section 7 explains the fraud detection methodologies. Section 8 tests a working CFM framework based on motor insurance as proof of concept using machine learning. Section 9 lists the steps involved to move in the direction of CFM for an organisation. Section 10 states conclusions based on results achieved. Section 11 provides a preview of current and future work that is being undertaken in this area. Section 12 acknowledges experts, contributors and infrastructure provided by SSSIHL. Section 13 provides all the references used in this paper.

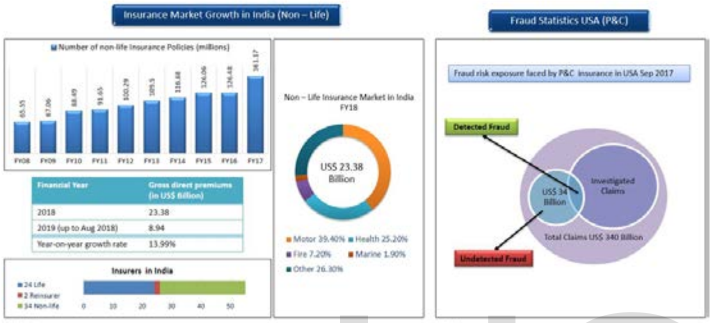
2. Motivation

Fraud can have a significant impact on the working of a

- *Rohan Yashraj Gupta, Tech Actuarial researcher, M.Sc. Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India, PH- +91 9593256368. E-mail: rohanyashraj@gmail.com*
- *Satya Sai Mudigonda, Senior Tech Actuarial Consultant and Hon. Professor in Department of Mathematics and Computer Science in Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India, PH- +91 9603573032. E-mail: satyasaibabamudigonda@sssihl.edu.in*
- *Phani Krishna Kandala, Assistant Vice President, Actuarial and visiting faculty, Sri Sathya Sai Institute of Higher Learning. PH:+91 9182472136, Email: kandala.phanikrishna@gmail.com*
- *Pallav Kumar Baruah, Head of Department, Department of Mathematics and Computer Science in Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India. PH: +91 9440699887. Email: pkbaruah@sssihl.edu.in*

business. It is not easy to put a number to the amount lost in a company due to fraudulent activities. The main reason being that it is not visible, thus making it very difficult to detect. The number of cases that are detected as fraudulent is very low compared to the actual figures. Fraud has the potential to disrupt the activities within the business, be it small or big. The direct impact of fraud in most of the cases is financial losses. In extreme cases, this can even lead to the bankruptcy of a company.

The "2014 Report to the Nations on Occupational Fraud and Abuse" estimate that the "typical organization loses 5 percent of its revenues to fraud each year. At it's extreme, fraud can destroy entire companies — Enron, Arthur Anderson and WorldCom are just a few examples".[36]



Financial Year	Gross direct premiums (in US$ Billion)
2018	23.38
2019 (up to Aug 2018)	8.94
Year-on-year growth rate	13.99%

Sources –
General Insurance Council, Financial Express, IRDAI & Research paper (Ernst $ Young)

Sources –
Fraud USA Statistics from Coalition Against Insurance Fraud

Figure 1 - Introduction to Insurance Fraud

"The Global Fraud Report 2015-16 by risk mitigation consultancy Kroll, with the aid of the Economist Intelligence Unit, found that the perceived prevalence of fraud in India is the third highest (80 per cent) among all countries and regions surveyed across six continents. Only Colombia (83 per cent) and Sub-Saharan Africa (84 per cent) surpass India".[1]

"An overwhelming 80 per cent of companies polled in India said that they had been victims of fraud in 2015-16, up from 69 per cent in 2013-14", according to a survey report.

An estimate of "Coalition Against Insurance Fraud Predictably" says that across all lines of business in insurance losses of nearly $80 billion a year is seen in the USA alone. In property and casualty sector in the USA, there is nearly $340 billion amount of claims of which approximately 10 percent is fraudulent amounting to $34 billion a year. See Figure 1.

Fraud is one of the most expensive crime and the effect of it is faced directly by customers and various other stakeholders. Due to fraudulent activities, companies face a huge amount of losses. This loss will invariably result in an increased premium for future customers. Fraudsters employ various types of techniques, strategies and tools to commit fraud. Some of the most common types of fraud include Health care fraud, Debit and Credit card fraud, Identity Theft, Health Insurance fraud, etc.

Many stakeholders get affected either directly or indirectly because of fraud. A direct impact that is faced by the insured is the increase in premium; this could lead to a knock-on effect for the insurance companies by a loss in business. It thus becomes very important for a corporate entity to have an effective fraud management process in order to ensure a healthy financial future. A key element that is required to identify and combat fraud is access to data and systems. If one has access to the right set of data fields and has systems equipped with sufficient controls, fraud can definitely be managed better.

Fraud has become a cause of anxiety for many organizations. In many companies, fraud is, in most cases, identified only after it has occurred. Ideally, we should be able to identify fraud before the damage is done (i.e. identifying proactively).

Fraud detection and thereby prevention will help save organizations many of their earnings. This will also increase the confidence of the organization. A strong fraud prevention system will increase the confidence of all the stakeholders including the investor and customer towards the company.

3. Insurance Regulation in India Specific to Fraud

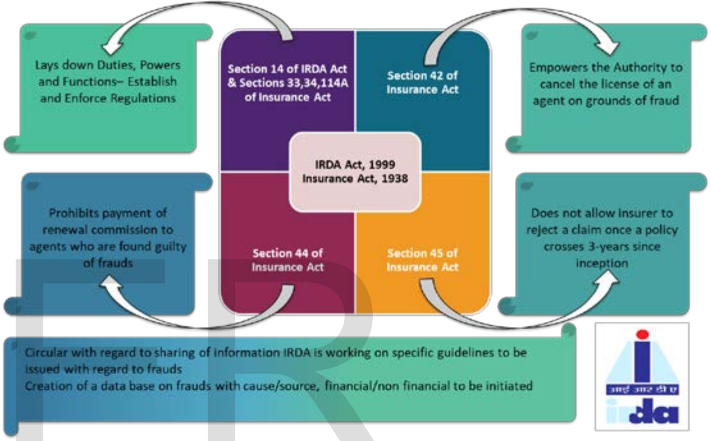

Lays down Duties, Powers and Functions– Establish and Enforce Regulations
Section 14 of IRDA Act & Sections 33,34,114A of Insurance Act
Section 42 of Insurance Act
Empowers the Authority to cancel the license of an agent on grounds of fraud
IRDA Act, 1999
Insurance Act, 1938
Prohibits payment of renewal commission to agents who are found guilty of frauds
Section 44 of Insurance Act
Section 45 of Insurance Act
Does not allow insurer to reject a claim once a policy crosses 3-years since inception
Circular with regard to sharing of information IRDA is working on specific guidelines to be issued with regard to frauds
Creation of a data base on frauds with cause/source, financial/non financial to be initiated
irda

Figure 2 - Regulation Fraud Specific India

The Insurance Regulatory and Development Authority of India (IRDAI) which is an autonomous, statutory body which functions as regulating and promoting the insurance and reinsurance industries in India. The functions of the IRDAI are defined in "Section 14 of the IRDAI Act, 1999". This lays down the duties, powers and functions of IRDAI. Various regulations in India specific to fraud is depicted in Figure 2.

The Insurance Act, 1938 is a law originally passed in 1938 in British India to regulate the insurance sector. It provides a broad legal framework within which the industry operates. Section 33, 34, 144A of Insurance Act, 1938 lays down power to appoint staff, Authority to issue directions and power of authority to make regulations.

Section 42 empowers the Authority to cancel the license of an agent on grounds of fraud. Section 44 of the Act prohibits payment of renewal commission to agents who are found guilty of frauds. Section 45 allows the insurer to reject a claim on grounds of fraud with proper evidence.

IRDAI has come out with a circular on the 8th of December 2010 that lays out the framework with regard to sharing of information. It is working on specific guidelines to be issued with regard to frauds. Creation of a database on frauds with cause/source, financial/non-financial to be initiated.

4. Comprehensive Fraud Management (CFM) -

Concept

Based on the study done by Society of Actuaries [2], wherein 450 research papers and articles were studied and 27 of them were found to be most relevant for their study in "Examining Predictive Modelling–Based Approaches to Characterizing Health Care Fraud". Various methodologies for fraud detection were identified as part of this study. We analysed this further and arrived at a Comprehensive Fraud Management (CFM) framework incorporating both actuarial techniques and technology. Fraud management in this context means both prevention and detection of fraud. The concept is depicted in Figure 3.

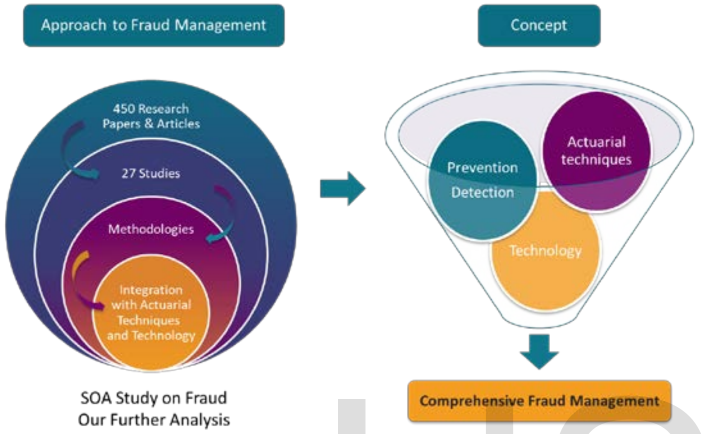

Approach to Fraud Management
Concept
450 Research Papers & Articles
27 Studies
Methodologies
Integration with Actuarial Techniques and Technology
SOA Study on Fraud
Our Further Analysis
Prevention Detection
Actuarial techniques
Technology
Comprehensive Fraud Management

Figure 3 - CFM Concept

5. Comprehensive Fraud Management (CFM) - Framework

The CFM gives a framework for solving fraud management problem for the insurance companies in various Lines of Business. This framework is developed considering the discussions in the following research papers and articles.

Figure 4 depicts the key components of the CFM Framework.

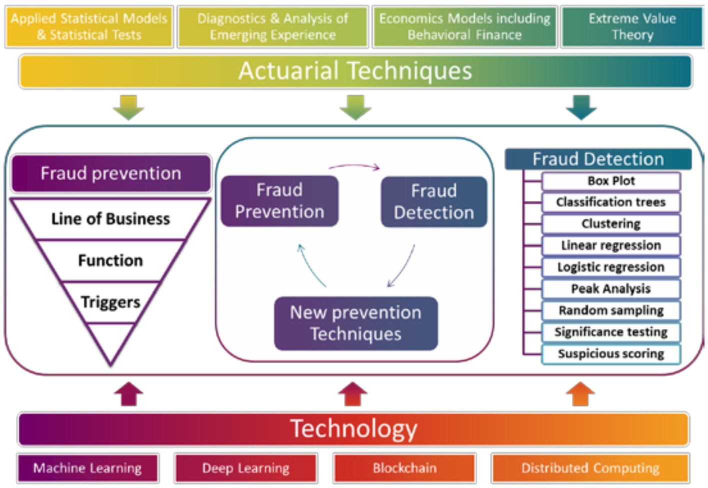


Figure 4 - CFM Framework

The central part of the CFM includes fraud prevention, detection and identification of new prevention techniques supported by both actuarial techniques and technology. There is a feedback mechanism within the cycle, an arrow which goes back from "New Prevention Techniques" to "Prevention". However, this is not an automatic process. The feedback mechanism requires the actuary to exercise judgement.

The following sub-sections discuss various stages of the CFM.

5.1 Stage 1: Fraud Prevention

This is the first stage of CFM in which we come up with preventive techniques for potential fraud. The proposed approach to fraud prevention is a trigger based system.

The first step for developing the model for fraud prevention is to identify the line of business (LoB) viz. Motor, Health, Life, General insurance, etc. Once the LoB is selected we further need to identify the functions (could be multiple functions) within the business that we are looking at, for e.g. underwriting, claims, administration, etc. After having considered the LoB and the function for which we want to develop the fraud prevention model we need to establish triggers for early detection and taking appropriate preventive action.

Triggers are parameters or a set of parameters, which can help, identify and raise alerts for suspicious activities. These can be managed through an automated system or manually. Identification of triggers in itself is a herculean task and requires a thorough understanding of the business processes. Triggers identified could differ based on the stage in which the fraud is committed.

Fraud can be committed at various levels, which can be broadly classified into three parts.

Internal Fraud – Fraud/misappropriation against the insurer by an employee e.g. deliberate incorrect data entry into the system.

Intermediary Fraud – This is not a direct, but indirect fraud. Involvement of the doctor, the agent, or another outsider which could involve misrepresentation of information relating to the income, pre-existing medical conditions, education qualification, current occupation, etc. are some examples of intermediary fraud.

Customer Fraud – Fraud done by the customers against the insurer in the purchase and/or execution of an insurance product, including fraud at the time of making a claim.

As part of our work we have identified 100+ triggers for various LoB viz. health, motor and general. Some of the fraud triggers in the insurance business are the following:

- Claims made shortly after the issue of the policy or just before the end of the policy terms.
- Very vague or misleading information provided by the policyholder. E.g. Incorrect health history details, like date of treatment, place of treatment, doctors name, diagnosis is done, etc.

Fraud prevention in its early stage is the first step in proactive fraud identification. This will not only help the fraud identification to be faster but also save the company of losses which it would have incurred had the fraud been committed. This will give the company real-time information about the activities within the company which are fraud-prone and required measures can be taken accordingly. However, it may not always be possible to prevent fraud in its early stages.

This brings us to the next step of the CFM, which is fraud detection. Thereby, explaining the arrow joining from fraud prevention to detection.

5.2 Stage 2: Fraud Detection

The important step in fraud detection is the identification of suspicious activities that have a higher probability of being fraudulent. Detecting an insurance fraud and abuse requires an in-depth knowledge of the insurance industry. Many insurance systems have experts who manually review each of the transactions to check for the suspicious ones. However, it becomes almost impossible for a human to detect anomalies in the trends, given that there is so much data to look at. With the advancement in technology, there are models and methodology, which takes advantage of the highly powerful computers to do the required investigation. The key focus is to identify things that do not appear to be normal. Very often, it is seen that it is this abnormality that are the key indicators of fraud. However, identifying these key indicators is a herculean task and would require a thorough understanding of the business. We would require to calculate various statistical parameters to look for outliers or the values which are way above the average behaviour as seen in the data. We look at both the extreme values both high and low and find and anomalies present there. We study the classification of data into specific groups and analyse or check for a number of instances that are occurring outside of statistical parameters.

There various methodologies that are being used in insurance fraud detection in recent times. The details of each of these methodologies which comprise of their pros and cons, the technology needed to implement these methodologies and data requirements are described in detail in the later sections.

5.3 Actuarial Techniques

One very well-known framework in the actuarial domain is the ACC (Actuarial Control Cycle) which gives a framework for solving any actuarial problems. The actuarial control cycle which comprises of three main components viz. specifying the problem, developing the solution and monitoring the results, is a model that can be applied to many aspects of actuarial work to find a solution. With further analysis into the framework the following Actuarial techniques have been identified which can be used in Fraud detection:

- Applied Statistical Models & Statistical Tests
- Diagnostics & Analysis of Emerging Experience
- Economics Models including Behavioural Finance
- Extreme Value Theory

5.3.1 Applied Statistical Models & Statistical Tests

A lot of research underwent on the applied statistical models such as GLM, GBM, GAM and others to arrive at a direction or solution for the existing uncertainties. We study the features of the given data and apply the statistical and mathematical models such as exponential family for GLM in order to generalize the model structure for the existing problem of Fraud detection and study the impact of each feature on to the final output such as the probability of fraud or severity impact of fraud. For example, when age (feature) increased from 51 to 54, the premium increased by 20% (final output).

Statistical tests are used to determine the optimal set of features. This helps us to assess whether the addition of any new feature would have the desired improvement in output or would be neutral.

5.3.2 Diagnostics & Analysis of Emerging Experience

Diagnostics are metrics which help us to interpret data or results and verify underlying methodologies and assumptions. For example: in a typical quota share arrangement, we need to have a RI to gross ratio to be consistent across all contracts, we can identify those contracts which do not exhibit this behaviour and investigate further. Here, the diagnostic used by us is RI to Gross ratio. Interpretation of diagnostics is one of the most important constituents where care needs to be taken.

A direct application of the actuarial control cycle framework is seen in the analysis of emerging experience. Here we monitor the impact of deviation from the expected results both in the short term as well in the long term. This is a very useful tool to monitor the current methodology used for fraud detection and quantification of fraud amount. It would capture the following aspects:

- Change in methodology
- Change in assumptions
- Movement solely due to experience.

5.3.3 Economics Models including Behavioural Finance

Behavioural Finance as an economic model is generally an important constituent to understand the nature of an individual profile which includes both monetary and non-monetary transactions. Machine learning is used to understand, predict or anticipate behaviours at the most granular level for each transaction. Features for non-monetary transactions generally include a change of address, request for a duplicate identity card or a request for password reset are used in order to understand the behaviour type. These features have more explanatory power about the existing mental bias or rational/irrational behaviour which we use to detect fraudulent behaviour.

Monetary transactions also help us to understand the behavioural aspects such as steady/sudden increase in wealth, spend velocity and number of days between transactions of similar type. These directly help us to understand the quantitative aspect of the fraud.

5.3.4 Extreme Value Theory

After performing trigger functionality on the given data, we would have obtained a subset of events which we would be interested to investigate. To perform this investigation, a suitable class of models would be those models that have low frequency and high severity impact. These are generally present in the tails of the distribution which are best described by generalized extreme value and generalized Pareto type of distributions.

From the above methodology, it helps us to obtain the quantum of loss due to a particular type of fraud/suspicion.

5.4 Technology

5.4.1 Machine Learning

It is an Artificial Intelligence (AI) application that provides computers with the ability to learn and improve from the provided dataset without programming it explicitly. The process of "learning" starts from the observation of data in order to look for patterns in the data provided. There are three broad categories of Machine learning methods: Supervised learning, Unsupervised learning and Semi-supervised learning.

Every insurance company possess a large amount of data. It becomes challenging to analyse the pattern for fraudulent claims manually. Often, it takes more time and money whenever there is human intervention. Machine learning models provide us with a solution to help us tackle both these issues efficiently.

Many models were built using machine learning algorithms wherein new claims are given suspicious scores. Based on the score, further decisions to investigate or not being taken.
Classification models are built to identify the nature of the claims. These models can further be used for fraud management for future claims.

5.4.2 Deep Learning

Deep learning is a subset of machine learning that helps us to build a deep network for the problem at hand. Deep models are used to identify complex patterns hidden in large datasets. These models are used as a pre-processing technique for getting the proper feature representation for building Deep models.

In insurance companies, claims data contains varied features. All of which may not be equally important for building fraud management tools. In such cases, Deep models can provide a way of identifying only the important features. Models like Auto-Encoders are used for such purposes.

5.4.3 Blockchain

It is a distributed digital ledger, secured through cryptography. Data provided is sequentially recorded in "Blocks" and are permanent (Immutable). Each new block is linked to the immediately previous block with a cryptographic signature, forming a 'chain'. This tamperproof validation of the data is done without any central authority. The ledger is not hosted in one location or managed by a single owner but is shared and accessed by anyone with the appropriate permissions.[3]

Blockchain technology aims to provide a fraud-free solution for insurance business because of its in-built features such as transparency, immutability and security. Many data sets may contain sample-bias due to incompleteness; with Blockchain, this problem can be handled effectively.

In any insurance company, there are many functions which are multi-faceted. This involves the authentication of data from multiple sources. There is a possibility of error and/or fraud happening at each and every stage of the process of validation. It is a fact that insurance companies are incurring huge losses due to such frauds or errors.

5.4.4 Distributed computing

Processing a huge amount of data in a single system can take a long time. Distributed computing provides a solution for processing data in real-time. High-speed computing is achieved by distributing the work to more than one system.

This technology is used to boost the speed of processes within the organizations. For comprehensive fraud management system, an organization would require the models to run on a real-time basis. So integrating distributed computing within the models opens avenues for companies for building frameworks which are complex and would otherwise require a very long time for running them.

6. Business and Technical view of the CFM Framework

This section discusses the business and the technical view of the framework. This is the representation of the process behind the working of the framework, demonstrating the various requirements at each stage of the framework process.

6.1 Business view

At the very first level data needs to be provided which could belong to different functions. For e.g. sales data, policy data, claims data, reporting data, underwriting data, etc. Once the data is selected it then goes through the prevention stage where depending on the line of business and the data chosen the triggers are chosen, this could also be n in number. Depending on the triggers found the appropriate model is created using which some of the potential fraud cases are prevented. Remaining passes through the next stage which is the detection stage where appropriate actuarial techniques and technology is selected for creating the model for fraud detection. This is done by identifying the methods. For e.g. classification trees, clustering, random sampling, etc. At this stage, we may further detect some more fraud cases. The results obtained are analysed further to improve the existing model. With every run of the model, the existing model improves and provides a better solution than the existing model. The business view is depicted in Figure 5

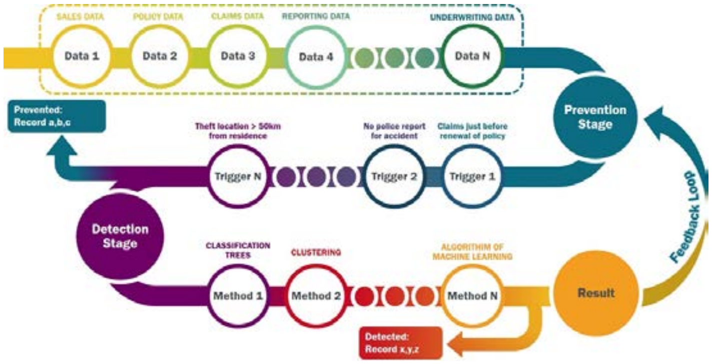


Figure 5 - CFM framework business view

6.2 Technical view

At the very first level data provided would come from various databases that the organization has e.g. HQ, MySQL, CSV, Postgres, OracleDB, etc. it is from these databases that we get various data like sales data, policy data, claims data, reporting data, underwriting data, etc. Once the data is selected it then goes through the rule engine where depending on the line of business and the data chosen rules are generated which is used to create the model. Rules could be a combination of triggers are chosen, this could also be n in number. Depending on the rules found, an appropriate model is created using which some of the potential fraud cases are prevented. Remaining passes through the next stage where the data is passed through the API's (Application Interface) which has their own libraries and are created using various technology, actuarial techniques and methodologies. At this stage, we may further detect some more fraud cases. The results obtained are analysed further to improve the existing model. With every run of the model, the existing model improves and provides a better solution than the existing model. The technical view is depicted in Figure 6

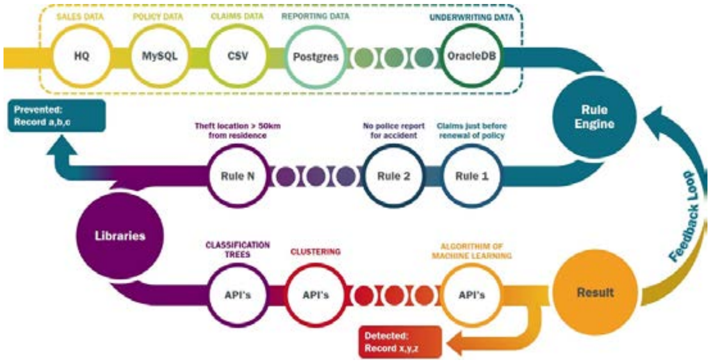


Figure 6 - CFM framework technical view

7. Fraud Detection Methodologies

7.1 Box plot

7.1.1 Description of the methods

Box plot is a graphical way of representing data. This method helps us graphically identify outliers. For e.g., if we are plotting the frequency of claims submitted by the insured and we see that there is a very high number of claims (outliers) by some of the members then this could be an indication to suspicious activities. Figure 7 depicts this methodology.

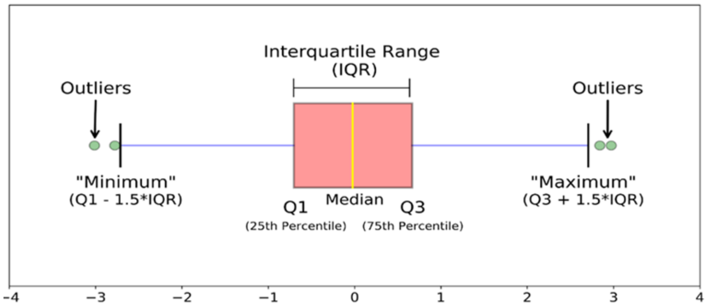


Figure 7 - Box Plot

7.1.2 Pros and cons

<u>Pros:</u>

- The box plot shows outliers and is unaffected by it
- Very easily handles and summarises large data sets
- They are good for comparing datasets
- A very effective way of visually showing the summary of the data

<u>Cons:</u>

- Not good for summarizing small datasets
- Box plot does not show the original data
- It is easy to identify mean and mode
- One of the biggest limitations is that it can only be used with numerical datasets
- Can be skewed

7.1.3 Technology needed

Excel, R, Python

7.1.4 Data requirements

If we are trying to find the outliers in the claims data for an e.g. excessive amount of claim, unusually high claims frequency, etc. then we would require claims data which would contain the information about the amount of claim submitted by the individual along claims submission date. [4],[5].

7.2 Classification trees

7.2.1 Description of the methods

Classification tree or decision tree is an effective way of making a decision. In case of fraud management, the final decision that we need to arrive at will be to classify the claims as fraudulent or not. This method provides us with a way to classify the outcome and the probability of achieving them.

Classification tree can be understood to be like an inverted tree with "root" node being the first node and is at the topmost level. Further, the data is divided into two or more subbranches which are called as "branch-node". The bottom-most node is called the "leaf node". For the data to be partitioned each of the nodes has certain conditions under which the data is portioned. Whenever any model is built the dataset is initially divided into training data and test data based on proportions like 70:30, 60:40, etc. The tree is usually built using the training dataset, the rules and patterns in the data that is produced can be implemented in making a detection algorithm. For reaching each of the leaf nodes there is a certain path that needs to be followed this path represents a sequence of classification rules. For e.g., two of the rules that could be used "fraud" leaf node could be as follows, by way of illustration (based on health insurance).[6]

(1) Rule 1:

IF "distance between the hospital and the patients address" < 50km,

THEN "claim status" = clean → move to the next node

IF "distance between the hospital and the patients address" > 50km,

THEN "claim status" = suspicious → stop at this node

(2) Rule 2:

IF "average medical expenditure" < $50,

THEN "claim status" = clean → move to the next node

IF "average medical expenditure" > $50,

THEN "claim status" = suspicious → stop at this node

:

:

The iteration is carried forward till the optimal results are obtained

7.2.2 Pros and cons

<u>Pros:</u>

- Easy to interpret and explain
- Requires very little data preparation
- Does not require data to follow a certain distribution
- No need to worry about outliers or if the data cannot be linearly separable

- Can be used for both categorical and numerical data

<u>Cons:</u>

- It can easily overfit
- Non-numerical data not easy to handle

7.2.3 Technology needed

MatLab, Python, R

7.2.4 Data requirements

The data required for using the classification tree depends on the application and would be specific to the case. In the case of fraud detection, the data point should contain an attribute that is of relevance for the purpose of running the algorithm that has been created. For an e.g. thing like claims amount, gender, locality, number of claims, type of vehicle, the term of the policy, etc. [7],[8]

7.3 Clustering

7.3.1 Description of the methods

Clustering is a way of grouping together the observations with similar features. This is another way of finding outliers. Outlier's detection based on clustering methodology helps organizations see for any outliers in the input data. Further analysis can be done on the clustered data to arrive at suspicious or fraudulent activities.

The classification of various types of clustering is given in Figure 8.

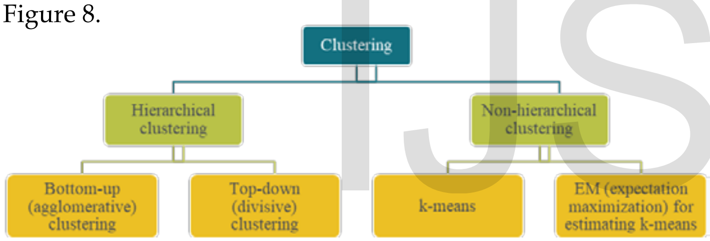

Clustering
Hierarchical clustering
Non-hierarchical clustering
Bottom-up (agglomerative) clustering
Top-down (divisive) clustering
k-means
EM (expectation maximization) for estimating k-means

Figure 8 - Clustering Classification

7.3.2 Pros and cons

<u>Pros:</u>

Non-hierarchical clustering | K-means

- Conceptually k-means is the simplest method and is one of the first methods used on a new data set.
- It is easy to implement k-means clustering and interpret the results produced.
- Where there is a large number of variables present in the data, k-means clustering is computed much faster than hierarchical clustering (for a small value of K).
- Clusters produced by k-means is tighter than hierarchical clustering, especially with globular clusters.
- K-means algorithms are very flexible as the can easily adjust to changes made in the cluster segments.

Hierarchical Clustering

- This technique of clustering provides more information than non-hierarchical techniques there making it more preferred for detailed data analysis
- It is also easy to implement.
- Hierarchical clustering outputs a hierarchy

<u>Cons:</u>

Non-hierarchical clustering | K-means

- K-means clustering can be performed in numerical data only and cannot be used in nominal data like colours.
- It gives different results for different runs of the algorithm thereby lacking consistency.
- The clusters produced by this method are of uniform size even though the input data is of different sizes.
- The final result of the data sets is affected by the manner in which the data is ordered while building the algorithm.
- The final results are sensitive to changing or rescaling of the dataset.
- Predicting the k-values or the number of clusters is not easy.
- For the effective functioning of K-means clustering, the K-value needs to be specified at the beginning of the algorithm.
- K-means clustering operates on various assumptions.

Hierarchical Clustering

- Undoing the previous step is not possible in hierarchical clustering
- When it comes to large dataset this method becomes unsuitable
- There is a very high significance of initial seed on the final result
- The order of the data has an impact on the final results.
- This model is very sensitive to outliers
- It is low in efficiency in the sense that it requires the user to compute at least n x n similarity coefficients and update them during the clustering process. [4],[9],[10],[11],[13]

7.3.4 Technology needed

Clustering is unsupervised learning as the test data provided for the purpose is not labelled, categorized or classified.

Python, R, MatLab

7.3.5 Data requirements

Irrespective of whether the data is categorical or numerical in nature this technique can be applied to form clusters (groups with similar features). Thus the various insurance datasets on which this technique can be performed are as follows:

- Amount of premiums written, claims and expenses of business both in-force and new according to the lines of business
- Pricing and underwriting: Claims amount, total sum insured, the total number of insured individuals
- Profit & Loss, balance sheets, asset allocations, reserves, affiliated reinsurance transactions
- Sales by distribution channels, sales force headcount

7.4 Linear regression

7.4.1 Description of the methods

An approach which explains the relationship between a response variable (output) and one or more explanatory variables which helps us to identify the features of existing frauds. These can be used for identification of future fraud cases which have similar features. If you have a single explanatory variable then it is known to be simple linear regression, else it would be called as linear regression. This has the advantage of identifying features of the existing type of fraud but may not be of help in the identification of new types of fraud.

7.4.2 Pros and cons

Pros:

- Linear regression is a simple algorithm and works very well if the data has a linear trend.

Cons:

- Fail if the data does not have a linear trend.
- Data must be independent i.e. there should not be any correlation between claims submitted by two individuals.

7.4.3 Technology needed

Python, R, MatLab

7.4.4 Data requirements

This technique is used to find the relation between various fields in the dataset. The data is required to have numerical or categorical fields. E.g. fields like claims amount, member's age, gender, the premium charged, sum assured, the term of the policy, etc. [14],[15],[16],[17],[18]

7.5 Logistic regression

7.5.1 Description of the methods

Regression analysis is the methodology which uses statistical techniques for finding the relationship between explanatory and response variables in the given dataset. Regression becomes more complicated as the variables (attributes) and the dataset size increases. Thus using logistic regression as fraud detection methodology is quite challenging as the dataset is generally huge and the number of variables are no less. However, this is a very powerful tool and can help one to understand the significance and variable or a combination of variables would have on determining the predictive power of the comprehensive fraud strategy. In this methodology, the genuine instances of claims are compared with the fraudulent ones to build the algorithm which would help to determine whether any new transactions are fraudulent or not.

7.5.2 Pros and cons

Pros:

- It is easy and quick to implement and very efficient to train
- Easy to interpret
- Does not too much of computational power

Cons:

- Non-linear problems cannot be solved using logistic regression since its decision surface is linear
- Results highly depend on the proper representation of the data.
- Can only be used to predict categorical outcomes

7.5.3 Technology needed

Python, R, MatLab

7.5.4 Data requirements

Logistic regression is a data classification techniques which uses linear boundary to separate input into regions. Thus, for this method to be used one would require data to which can be linearly separable. Like given in

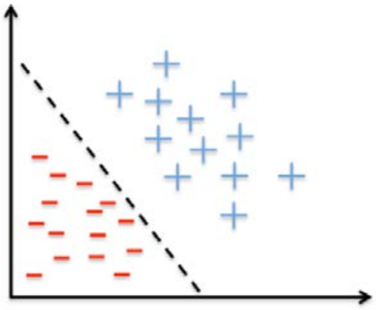


Figure 9 - Linearly Separable Data

Data is linearly separable or not, cannot be seen right away. We will require the data to be plotted.

However, the data required for performing such analysis would be claim amount, claims frequency, sum assured, the premium charged, etc. [7],[19]

7.6 Peak analysis

7.6.1 Description of the methods

Any peak in the data is analysed and based on this analysis further inferences are made. The following example is from (Capelleveen, Poel, Mueller, & Hillegersberg, 2016) it shows in detail the idea of peak analysis.

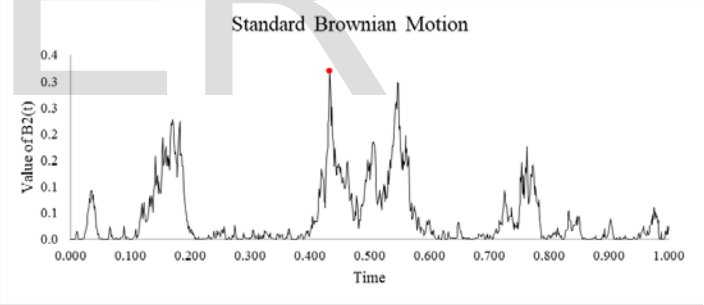


Figure 10 - Peak Analysis: Brownian motion

The illustration of how peak analysis could be used to detect any outliers is shown in Figure 10 (this is a hypothetical graph which is being used only for the explanation of the methodology). The figure shows the plot of the share price movement which is captured using the Brownian motion. B(t) represents the standard Brownian motion. The sample paths of B(t) for 0 ≤□t ≤□T is modelled by discretizing time in equal time steps of length ∆t and then considering the increment process ∆B(t), where ∆B(t) ~N(0,∆t). Here ∆t = 0.001 and T=1. The graph shows the sample path of B2(t) for 0 ≤□t.

An unusual increase or decrease in the value of B2(t) identified is searched for in peak analysis. The peaks are selected when the value is at least twice or half the value of the previous time step. These outliers are marked with a red dot in the figure.

7.6.2 Pros and cons

Pros:

- This method is easy to implement
- Examining by eye can give a good indication for peak

Cons:

- This method can only be used for identifying a peak which is relative to the neighbouring data points
- In some cases, the peaks may not actually represent an outlier
- Identifying peaks could be difficult for the datasets which cannot be plotted

7.6.3 Technology needed

Python, R

7.6.4 Data requirements

If we are trying to find the outliers in the dataset for the purpose of peak analysis for an e.g. excessive amount of claim, unusually high claims frequency, etc. then we would require dataset to have the relevant information. [4]

7.7 Random sampling

7.7.1 Description of the methods

Random sampling is the process of collecting samples from distribution. Sampling is a necessary step in while building any model. This is because data may not be available for some situations thus we may require to upscale the sample by randomly sampling data point and then adding them back to the dataset. Or in some cases, the dataset may be so huge that it is not possible to use all the data point to perform the calculation thus it requires random sampling which would ensure that the property of the parent population is maintained. Sampling is mandatory for certain stages in fraud detection. Sampling techniques heavily rely on the probability distribution of the given population thus it shows better results when the data point is a data set is large in number.

7.7.2 Pros and cons

Pros:

- It is the simplest form of data collection
- Sample collected using this method represents the distribution of the parent population
- The results obtained by using the random samples can be applied to the entire parent population

Cons:

- The sample size is required to be very large
- The samples obtained are only a subset of the population thus capturing the exact behaviour of the population may not always be possible.

7.7.3 Technology needed

Python, MatLab, R, C++

7.7.4 Data requirements

The population from which samples are to be obtained must be very large and complete in order to get an unbiased sample. [5],[10],[20],[13]

7.8 Significance testing

7.8.1 Description of the methods

(ACL, Detecting and Preventing, 2013)This is a statistical method which can be used for finding the answer (truth value) for a given hypothesis. So, given a claim, it would be possible to test the hypothesis that the given transaction is fraudulent or not.

Consider that we have a null hypothesis that there is at least one witness at the time of a given accident claim against the alternative hypothesis that there are no witnesses. We can then go and look for all the past claims information for a similar instance and compare the data. If we find something of greater significance then this could indicative of the fact the null hypothesis is true. So, if the claimant tells that there was no witness at the time of accident then this could be indicative of the fact that the claimant is lying and could be doing a fraud. If that sort of anomaly seems to be relatively prevalent or there is certain exposure to risk that we are not comfortable with, maybe we would want to investigate on a recurring basis.

7.8.2 Pros and cons

Pros:

- This is a good statistical way of finding the truth value of the hypothesis for a given confidence level

Cons:

- It may not always be possible to find enough sample point to carry out the test
- Finding the appropriate statistic to carry out the test may not be possible

7.8.3 Technology needed

R, Python, Java, MatLab, C, C++

7.8.4 Data requirements

Whenever carrying out the significance testing we should have enough data available to carry out the test. We should also require to know the statistics which would be required to carry out the test. [15],[21],[20],[22],[23],[19],[17]

7.9 Suspicion scoring

7.9.1 Description of the methods

A scoring metric used to identify the cases which are fuzzy (cases which can neither be classified as fraud or not), this metric would help us to identify the intensity of the event and further decide on action whether to deeply investigate the event or to conduct a high-level investigation in the respective event.

7.9.2 Pros and cons

Pros:

- Helps us identify the intensity of event

Cons:

- Its may not always be possible to create a scoring metric for the given dataset

7.9.3 Technology needed

R, Python, Java, MatLab, C, C++

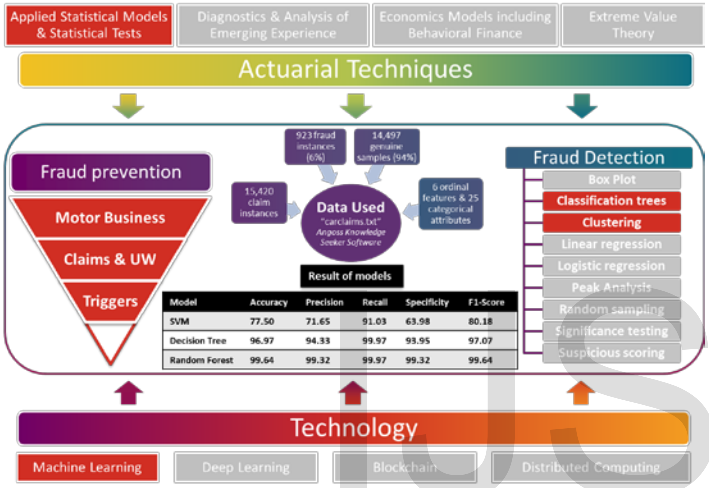
7.9.4 Data requirements

We require the data to be in the form for which we can generate the scoring metric.[24],[25],[26],[8],[18],[27]

8. Proof of Concept

In order to demonstrate, we have applied the CFM framework to Motor line of business on the internationally used data “carclaims.txt” which is provided by Angoss Knowledge Seeker Software. "carclaims.txt" dataset is the only publicly available automobile insurance dataset and is taken from [28]. It consists of 15,420 instances of claim from January 1994 to December 1996. There are a total of 14,497 genuine samples (94%) and 923 fraud cases (6%). Hence the dataset is highly imbalanced. The dataset has “6 ordinal features and 25 categorical attributes”.

The framework in highlighted which represents the actuarial techniques and technologies used. It also highlights which LoB was considered and the functions identified, methodologies used have also been used. The figure is as follows:



Model	Accuracy	Precision	Recall	Specificity	F1-Score
SVM	77.50	71.65	91.03	63.98	80.18
Decision Tree	96.97	94.33	99.97	93.95	97.07
Random Forest	99.64	99.32	99.97	99.32	99.64

Figure 11 - Proof of Concept

8.1 PoC – Fraud Prevention through trigger

In this exercise, we have identified and applied 50 triggers relating to claims and underwriting functions. These 50 triggers have been coded on the data to identify the claims and policies which needs investigation.

8.2 PoC – Fraud Detection using Machine Learning

We used one-hot encoding and binary encoding for preprocessing the data such as the representation of categorical attributes in the data.

With respect to actuarial techniques, we used applied statistical models. Since it is a class imbalance problem, we have used MWMOTE (Majority Weighted Minority Oversampling Technique) to enhance the sample. This technique involves:

- Identification of sample observations, which are hard-to-learn and identification of most important minority samples.
- For each of the hard-to-learn minority sample, a weight is given based on its importance in the data. These weights are based on the majority of samples.
- Generate new synthetic minority samples following a similar strategy to SMOTE (Synthetic Minority Oversampling Technique).

For fraud detection, we have trained and used three different models namely Random Forest (RF), Decision Tree (DT) and Support Vector Machine (SVM). [29]

We used Tenfold Cross-Validation, Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC) for statistical significance. The results are shown in the Figure 11. We can see all three different models have good accuracy. Recall has increased significantly by using class imbalance techniques to identify fraudulent cases more appropriately.

9. Moving in the direction of CFM

Figure 12 lists the steps involved to move in the direction of CFM for an organisation. This will potentially lead to increased detected fraud through investigation of claims.

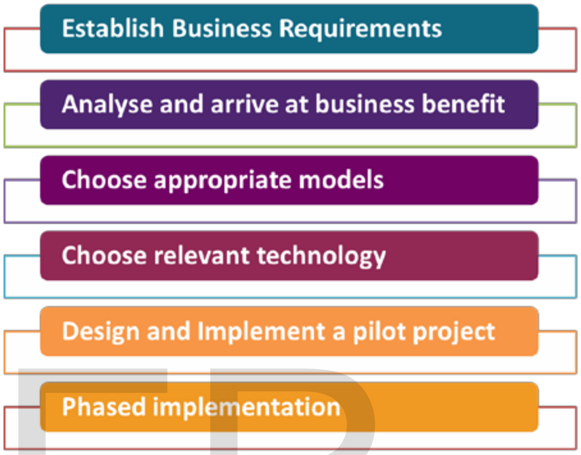

Establish Business Requirements
Analyse and arrive at business benefit
Choose appropriate models
Choose relevant technology
Design and Implement a pilot project
Phased implementation

Figure 12 - Moving in the direction of CFM

10. Conclusion

In this paper, we developed a framework for Comprehensive Fraud Management, which integrates both actuarial techniques and technology. An emphasis on the use of various actuarial techniques in the process of CFM was discussed and the necessity of using technologies was highlighted. We have demonstrated this by using applied statistical models and machine learning applied to motor insurance data. Results indicate that this can be used to arrive at increased detected fraud in a given scenario.

11. Future Work

Current and future work in fraud management include identifying triggers based on a line of business, building algorithms, Unsupervised learning using Auto-Encoders, Spectral Clustering Techniques and Deep Learning.

Our future work involves integrating Blockchain and Deep Learning to create “Deep Chains”. They ensure that the model receives the appropriate data from authenticated sources, which can be used for training and making the right level of business decisions. Interoperability is one of the essential features of any technology and we are in the process of publishing our work in this area.

The CFM proof of concept on motor business presented in this paper can be tested and extended to other organizations in any industry. This is currently being explored within Insurance for different lines of business and also with other financial and non-financial organizations. This should encourage actuaries to explore new opportunities in comprehensive fraud management.

12. Acknowledgements

We express our deep sense of gratitude to Sri Sathya Sai Institute of Higher Learning (Deemed to be University) SSSIHL - for providing IT expertise, IT Labs, Previous publications and Actuarial consultants for building a Comprehensive Fraud Management Framework and to demonstrate it using proof of concept.

We acknowledge Harold D Singh, FSA, Sai Aditya T V S, FIA, Nikhil Rai, M.Tech, Sravan Viswanadha, M.Tech, for their inputs.

13. References

[1] VASUDEVA, D. P. (2016, june 27). India Ranks 3rd in the World in Global Fraud, Just Behind Colombia and Sub-Saharan Africa. Retrieved from The Citizen: https://www.thecitizen.in/index.php/en/NewsDetail/index/1/8084/India-Ranks-3rd-in-the-World-in-Global-Fraud-Just-Behind-Colombia-and-Sub-Saharan-Africa

[2] Ai, J., Lieberthal, R. D., Smith, S. D., & Wojciechowski, R. L. (2018). Examining Predictive Modeling–Based Approaches to Characterizing Health Care Fraud. Society of Actuaries (SOA).

[3] Sravan, N. P., Baruah, P. K., Mudigonda, S. S., & K, P. K. (2018, October). Use of Blockchain Technology in integrating Heath Insurance Company and Hospital. IJSER, 9(10). Retrieved from https://www.ijser.org/researchpaper/Use-of-Blockchain-Technology-in-integrating-Heath-Insurance-Company-and-Hospital.pdf

[4] Capelleveen, G. v., Poel, M., Mueller, R. M., & Hillegersberg, D. T. (2016). Outlier detection in healthcare fraud: A case study in the. International Journal of Accounting Information.

[5] Gilliland, D., & Feng, W. (2010). An adaptation of the Minimum Sum Method. Health Services and Outcomes Research Methodology.

[6] Liou, F.-M., Tang, Y.-C., & Chen, J.-Y. (2008). Detecting hospital fraud and claim abuse through diabetic. Health Care Manage Sci, 6.

[7] Fen-May, L., Ying-Chan, T., & Jean-Yi, C. (2008). Detecting hospital fraud and claim abuse through diabetic outpatient services. Health Care Management Science.

[8] Shin, H., Park, H., Lee, J., & Jhee, W. C. (2012). A scoring model to detect abusive billing patterns in health insurance claims. Expert Systems with Applications.

[9] San(ni, M. (2016, Ded 12). Advantages & Disadvantages of k-Means and Hierarchical clustering (Unsupervised Learning). Retrieved from http://stp.lingfil.uu.se/~santinim/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf

[10] Yang, W.-S., & Hwang, S.-Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications, 13.

[11] Lieberthal, R. D., & Comer, D. M. (n.d.). What are the characteristics that explain hospital quality? A longitudinal PRIDIT approach. Risk Management and Insurance Review.

[12] Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Improving fraud and abuse detection in general physician claims: A data mining study. International Journal of Health Policy and Management.

[13] Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Applied Soft Computing.

[14] Kang, H., Hong, J., Lee, K., & Kim, S. (2010). The effects of the fraud and abuse enforcement program under the National Health Insurance program in Korea. Health Policy.

[15] Pande, V., & Mass, W. (n.d.). Physician Medicare fraud: characteristics and consequences. International Journal of Pharmaceutical and Healthcare Marketing.

[16] Srinivasan, U., & Arunasalam, B. (2013). Leveraging big data analytics to reduce healthcare costs. IT Professional.

[17] Becker, D., Kessler, D., & McClellan, M. (2005). Detecting Medicare abuse. Journal of Health Economics.

[18] Fang, H., & Gong, Q. (2017). Detecting potential overbilling in medicare reimbursement via hours worked. American Economic Review.

[19] Mesa, F. R., Raineri, A., Maturana, S., & Kaempffer, A. M. (2009). Fraud in the health systems of Chile: a detection model [original written in Spanish]. Pan American Journal of Public Health.

[20] Edwards, D. (2011). On stratified sampling and ratio estimation in Medicare and Medicaid benefit integrity investigations. Health Services and Outcomes Research Methodology.

[21] Dietz, D. K., & Snyder, H. (2007). Internal control differences between community health centers that did or did not experience fraud. Research in Healthcare Financial Management.

[22] Ekin, T., Fulton, L. V., & Musal, R. M. (2015). Overpayment models for medical audits: multiple scenarios. Journal of Applied Statistics.

[23] Major, J. A., & Riedinger, D. R. (2002). EFD: A hybrid knowledge/statistical-based system for the detection of fraud. Journal of Risk and Insurance.

[24] Victorri-Vigneau, C., Larour, K., Simon, D., Pivette, J., & Jolliet, P. (n.d.). Creating and validating a tool able to detect fraud by prescription falsification from health insurance administration databases [original written in French]. Thérapie.

[25] Iyengar, V. S., Hermiz, K. B., & Natarajan, R. (2014). Computer-aided auditing of prescription drug claims. Health Care Management Science.

[26] Lu, F., & Boritz, J. E. (2005). Detecting fraud in health insurance data: Learning to model incomplete Benford's law distributions. Machine Learning: ECML 2005, Proceedings.

[27] Johnson, M. E., & Nagarur, N. (2015). Multi-stage methodology to detect health insurance claim. Health Care Management Science.

[28] Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. New York: ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets.

[29] Rai, N., Baruah, P. K., Mudigonda, S. S., & Kandala, P. K. (2018, November). Fraud Detection Supervised Machine Learning. IJSER, 9(11). Retrieved from https://www.ijser.org/researchpaper/Fraud-Detection-Supervised-Machine-Learning-Models-for-an-Automobile-Insurance.pdf

[30] ACL. (2009). Fraud Detection: Using Data Analytics in the Insurance Industry. Society of Actuaries in Ireland.

[31] ACL. (2013). Detecting and Preventing. Retrieved from ACL: https://www.acl.com/pdfs/ACL_fraud_ebook.pdf

[32] Ai, J., Brockett, P. L., & Golden, L. L. (2009). Assessing Consumer Fraud Risk in Insurance Claims: An Unsupervised Learning Technique Using Discrete and Continuous Predictor Variables. North American Actuarial Journal, 13(4).

[33] Aral, K. D., Güvenir, H. A., Sabuncuoğlu, İ., & Akar, A. R. (2012). A prescription fraud detection model. Computer Methods and Programs in Biomedicine.

[34] Barton, C. (2009). Outwitting the fraudsters. The story so far. General Insurance, IFoA. IFoA .

[35] Cressey, D. (1973). Other people's money, 30. Montclair, NJ: Patterson Smith.

[36] Examiners, A. o. (2018). Report to the Nations. Global study done on occupa-

tional fraud and abuse. Association of Certified Fraud Examiners.

[37] Group, F. (n.d.). Fraud Examination Expert Witnesses. Retrieved from Forensis Group: https://www.forensisgroup.com/expert-witness/fraud-examination/

[38] Kirlidog, M., & Asuk, C. (2012). A fraud detection approach with data mining in health insurance. Procedia—Social and Behavioral Sciences.

[39] Konijn, R. M., Duivesteijn, W., Meeng, M., & Knobbe, A. (2015). Cost-based quality measures in subgroup discovery. Journal of Intelligent Information Systems.

[40] Labs, M. T. (n.d.). HOW MACHINE LEARNING FACILITATES FRAUD DETECTION? Retrieved from Maruti Tech Labs: https://www.marutitech.com/machine-learning-fraud-detection/

[41] Lægreid, I. (2007, Sep). Automatic Fraud Detection — Does it Work? Annals of Actuarial Science.

[42] Li, J., Huang, K.-Y., Jin, J., & Shi, J. (2008). A survey on statistical methods for health care fraud detection. Health Care Management Science.

[43] Mangaldas, C. A. (2018, January). Insurance regulation in India. Retrieved from Norton Rose Fulbright: http://www.nortonrosefulbright.com/knowledge/publications/163131/insurance-regulation-in-india

[44] Marr, B. (2018, May 21). How Much Data Do We Create Every Day? . Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#4064922760ba

[45] School, B. B. (2019, Jan 9). Centre for Actuarial and Big Data Analytics. Retrieved from Bond University: https://research.bond.edu.au/en/organisations/centre-for-actuarial-and-big-data-analytics/publications/?type=%2Fdk%2Fatira%2Fpure%2Fresearchoutput%2Fresearchoutputtypes%2Fcontributiontoconference%2Fabstract

[46] Thornton, D., Mueller, R. M., Schoutsen, P., & Hillegersberg, v. J. (2013). Predicting health care fraud in Medicaid: A multidimensional data model and analysis techniques for fraud detection. Procedia Technology.

[47] Wolfe, D. T., & Hermanson, D. R. (2004). The Fraud Diamond: Considering the Four Elements of Fraud. CPA Journal, 3-6.