Framework for Analysis and Detection of Fraud in Health Insurance

Nirmal Rayan

Medi Assist Healthcare Services Pvt. Ltd., Bangalore 560029, India
nirmal_rayan@yahoo.com

Abstract: The health insurance industry generates a wide range of data from patients' information to provider payment and claims report. The impact of fraud, waste, and abuse (FWA) in medical management is on the rise and contributes significantly to the increase in cost. Traditional methods of handling fraud include human inspection and heuristic rules. They are time-consuming, impractical and insufficient. Data Mining and Machine Learning play a dominant role in detecting and preventing fraud. We explore the use of statistical methods to create a rule-based heuristic engine that works with self-learning Decision Trees. This paper introduces a hybrid framework that combines domain expertise (Rule Engine), supervised learning (Decision Trees & Averaged Perceptron) and unsupervised learning (outlier analysis, k-means Clustering) techniques to identify fraudulent claims from a given set of outstanding claims. The investigation team is intimated with a weighted priority queue of outstanding claims listing the most-likely fraudulent claims with remarks for proactive and retrospective analysis. Our initial case study with one insurer demonstrates an increase in hit-rate by 209.4%.

Keywords: Fraud Detection; Healthcare fraud; Supervised Methods; Unsupervised Methods; Rule-Based Engine; Data Mining; Machine Learning; Decision Trees; Neural Network; Averaged Perceptron; k-means Clustering

1 Introduction

The health insurance industry, a foundation in modern-day society, serves the purpose of providing individuals with affordable healthcare [1]. As both medical expenses and mortality rate persist to rise, more economic pressure is applied to government and private organizations which assist in funding hospitalizations and treatments [21] [22]. The objective of the health insurance system must therefore be to offer adequate and appropriate coverage to individuals at a reasonable price to both patients and practitioners [16]. One reason for the increase in cost of healthcare can be ascribed to the funds lost in the health insurance system through fraud [1]. To help manage costs, we need to curb fraud, waste, and abuse (FWA). This quote of Christopher McDougall "Every morning in Africa, a gazelle wakes up, it knows it must outrun the fastest lion or it will be killed. Every morning in Africa, a lion wakes up. It knows it must run faster than the slowest gazelle, or it will starve. It doesn't matter whether you're the lion or a gazelle-when the sun comes up, you'd better be running." [23], reflects the role of a fraud analyst and a fraudster. Both must compete well to prevent being each other's victim. A fraud analyst must study the fraudsters' mode of operation to contest with them. In brief, an ideal professional in fraud analytics must be capable of thinking like a crook in order to prevent falling for their deception [17].

1.1 Challenges with Traditional Methods

Substantial medical knowledge is required to identify fraud and abuse in health care. Many health insurance systems trust human experts to manually evaluate health insurance claims and determine suspicious claims. This results in time-consuming system growth and claim processing, particularly for the large national insurance programs in nations such as India [12]. Health insurance companies have traditionally used rules to detect fraud. These heuristic rules have been summarized from past instances of fraud and are used to detect fraud from occurring in the future but it is not feasible to detect fraud manually for three primary reasons. First, by manual evaluation of large databases, it is not possible to discover all healthcare fraud. Second, new kinds of healthcare fraud are continually emerging. Third, to represent the present world situation, regular updates to existing laws are essential [5],[6].

1.2 Proposed Methodology & Solution

According to latest advances in insurance policies, fraud detection system needs more advanced data mining, machine learning (ML) and statistical modeling techniques capable of automatically understanding fraud patterns from experience [24],[7]. The insurance claim fraud detection system integrates three mining approaches as discussed below.

Rule-Engine (with Anomaly Detection): To create a fraud detection framework that works in the existing world, it is necessary to integrate available domain knowledge using data mining techniques. For instance, the model maintains a watch-list consisting of providers, beneficiaries, agents/brokers, and corporates/policy-holders exhibiting suspicious behavior. Based on severity, the watch-list is categorized as blacklisted or suspicious entities. Part of the watch-list generally comes as ad-hoc input from underwriters belonging to insurance companies.

978-1-7281-3863-3/19/$31.00 ©2019 IEEE

Unsupervised Learning: We use anomaly detection (for example, policy-level risk scoring, outlier analysis based on procedure cost, etc.) and k-means clustering (where k=5) to group data based on their similarities, patterns, and differences.

Supervised Learning: decision trees (with and without under-sampling), deep neural network and averaged perceptron are the artificial intelligence (AI) algorithms used for supervised machine learning.

This paper seeks to detect fraudulent practices in health care, evaluate the properties of health care data, compare and review currently proposed fraud detection approaches as well as their corresponding data pre-process with traditional methods and discuss future research directions. More specifically, this paper starts with a literature review, an introduction to the Indian health care system, different types of health care fraud and our methods for detecting them. Section III analyzes the features of health insurance data used in our academic research. In Section IV, we discuss the experimental design, approach, and operationalization of the fraud detection engine. We then compare and review the currently suggested approaches to fraud detection with traditional methods in Section V. Sections VI and VII discuss the scope and future directions for research and draws some conclusions.

2 Literature Review

2.1 Related researches

Survey: Obodoekwe, and van der Haar [1], [18] compared multiple machine learning techniques used to detect fraud in healthcare claim. They used dental medicare (which is a national healthcare scheme in the US) dataset between 2012 and 2015. Several models were built using Bayesian classifier, random forest, logistic regression, gradient boosted tree and neural network. A comparative analysis between the evaluated models suggested that the gradient boosted tree performed best for the medicare dataset followed by random forest and neural network.

Duman et. al. [13] surveyed to analyze the different fraud detection methodologies researched, implemented and published over the last decade in health care. They reviewed 31 individual research publication of which 18 paper publications used unsupervised techniques, 11 papers used supervised techniques and 2 papers discussed semi-supervised machine learning techniques.

Another important survey paper by Dua et. al. [19] compared the advantages and disadvantages of different supervised methods and identified which method worked best for healthcare datasets. The algorithms considered for this survey include neural network, Bayesian belief network, fuzzy Bayesian classifier, logistic regression, classification tree, genetic algorithm, k-nearest neighbor and association rules. The neural network and association rules had highest performance among all the supervised data-mining methods in contention and are thus used more frequently by researchers to identify fraudulent patterns in healthcare data.

Supervised Machine Learning: Bauder et. al. [15] used 2012-2015 medicare data to build a random forest model on class imbalanced big data. The imbalance was handled using a random under-sampling (RUS) technique on the majority class. Multiple random forests were constructed with different class distribution configurations. It was identified that the class distribution of 90:10 (genuine:fraud) gave the best outcomes, while the two extremely imbalanced distributions yielded the worst performance in fraud detection. It also suggests that the balanced class distribution does not produce the best results for medicare fraud identification. The same authors have also published another research paper [16] to detect anomalous provider or doctor practice behavior with that of their peers using multinomial naïve bayesian (MNB) classifier.

Unsupervised Machine Learning: Liu et. al. [12] used a geolocation clustering model to identify beneficiaries who availed provider services from long distance. Two measures were primarily used – incurred cost and the geographic proximity between beneficiary and service provider to generate clusters for the top 3 most prevalent diagnoses. Outliers were identified to be potentially fraudulent or abusive.

Peng et. al. [5] applied two clustering methodologies, SAS Enterprise Miner (EM) and open source CLUTO on health insurance claims dataset to understand the data and also detect fraud. Experimental findings show that CLUTO is quicker than SAS EM, however SAS EM offers more helpful clusters.

Peng et. al. [9] used network-based analysis to detect outliers by - creating heterogeneous information network, calculating correlation scores (between patients, medicines and diseases) and identifying the lowest score to be possibly fraudulent. Their experimental results confirmed that their method was accurate and effective.

Hybrid Machine Learning: Verma et. al. [7] used a rule-based engine with association rule mining (ARM) and clustering for detecting if a claim is fraudulent in the context of the Indian healthcare industry. They classified fraudulent behavior into two segments - disease-based anomalies and time-based claim anomalies. Association rule mining was applied on disease-based anomalies for outlier detection and k-means clustering and statistical decision rules were applied on period-based claim anomalies.

Another research from India by Rawte et. al. [8] used a hybrid modeling approach by harnessing the advantages of supervised and unsupervised machine learning models. They used clustering (using the evolving clustering method) and classification (using support vector machine) for fraud identification.

Similar research conducted by Kareem et. al. [6] aimed

to propose a hybrid approach in identifying frequently correlated items for detecting fraudulent health insurance claims using clustering (evolving clustering method), association rule mining (apriori algorithm) and supervised classification (support vector machine).

Joint Fraud Analysis: Very few research had been conducted to unearth joint fraud (or conspiracy fraud). One such research by Sun et. al. [3] on abnormal group based joint medical fraud detection methods claimed to differentiate suspicious fraudsters from ordinary people with unusual behaviors. The experimental method (which used adjacency graph analysis) enhanced the accuracy of joint fraud detection by more than 10% compared to standard approaches.

2.2 Impact of Fraud in India

According to the Federation of Indian Chambers of Commerce and Industry (FICCI), it is a matter of concern that the Indian Insurance Act does not define 'insurance fraud'. IRDA cited the definition provided by the International Association of Insurance Supervisors (IAIS) that describes fraud as "an act or omission intended to gain dishonest or unlawful advantage for a party committing the fraud or for other related parties [4], [20]". According to a latest study, the number of false claims in the industry is estimated to be roughly 15% of total intimated claims. The study indicates that every year, the Indian healthcare industry loses roughly Rs.600-Rs.800 crores (in INR) on fraudulent allegations. Health Insurance is usually regarded to be a bleeding industry with a very elevated claims ratio. Hence, to ensure that the health insurance sector remains viable, it is crucial to focus on the minimization or elimination of illicit claims arriving through health insurance [8].

2.3 Fraud, Waste, and Abuse

Health insurance abuse: is the billing of practices that are either directly or indirectly inconsistent with the objectives of - supporting patients with medically necessary services, satisfying professionally established standards and being reasonably priced [25].

Health insurance waste: in healthcare is mostly unrelated to fraud, as it essentially deals with the provision of health services that is not necessary. Waste from health insurance can only be seen as fraud and abuse when the act is deliberate. Waste can occur when services are overused, resulting in unnecessary expenses [25].

Health insurance fraud: is intentionally charging for undelivered services or unreceived supplies, medically unnecessary services and modifying claims for more reimbursement than the service provided [25].

2.4 Types of Health Insurance Frauds

Healthcare fraud behaviors can be segregated into four kinds, depending on the type of party committing fraud, namely -

Medical Service Providers:

- Forging a patient diagnosis to legitimize treatments and operations that are not clinically necessary [26], [10],
- Higher invoicing for insurance companies by ignoring patient co-payments, co-insurance, or deductibles [26],
- Charging for services that have not been conducted using real patient records, assisting in identity theft, or altering claims with additional fees for the services or operations that have never been conducted [26],
- Invoicing for unwanted procedures or services, like regular check-ups rather than monthly check-ups, only to generate insurance payments [19], [26],
- Billing for costly services instead of billing for low cost services or treatments which were conducted, known as **upcoding** or coding a patient's diagnosis to a more critical and costly charge and applying charges with incorrect CPT codes, such as charging a 30 minutes group therapy as a 50 minutes personal treatment [19], [26],
- Billing each phase of a process as if it were a discrete method, known as **unbundling** [26], for instance, billing tests within test ''sessions'' as if they were different sessions [19],
- Misrepresentation of non-covered treatments as medically necessary covered treatments to obtain insurance payments [10].

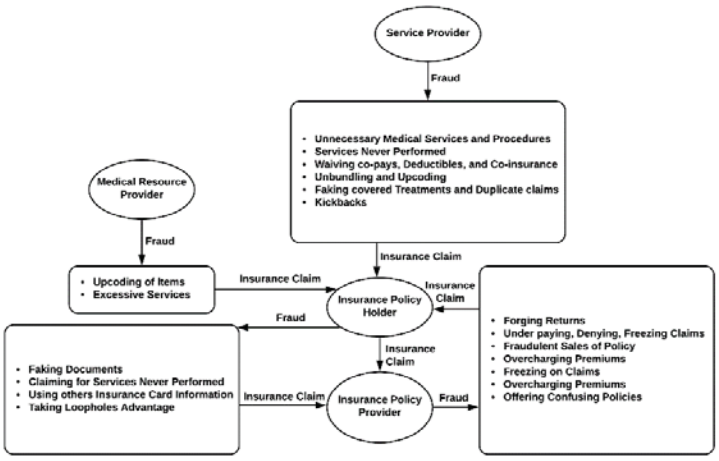

Figure 1 Types of health insurance frauds.*

Insurance Policy Holders:

- Forging records pertaining to eligibility or employment to receive low premium prices and better benefits [27],
- Unlawfully claiming insurance advantages using someone else's coverage or insurance data [27],
- Claiming for procedures or medical services that have never taken place [27], and
- Taking advantage of insurance benefits by finding policy loopholes.

Insurance Policy Providers:

- Underpaying claims for forging returns and benefit/service statements [27],
- Overcharging the policyholders' premiums by misinterpreting the customer's data and charging them for the coverage they do not actually have [19],
- Freezing claims without reviewing the legitimacy of claims [19],
- Insurance company wrongly rejects valid claims to attempt to deter the policyholder and hopes the patient will eventually give up [19],
- Making fraudulent sales of fabricated policies that are of no use to the policyholders and are primarily designed to receive high premiums from them.

Conspiracy Fraud:

- It involves more than one party fraud; for instance, the fraudulent activity involves a patient and a physician (or insurance firm).

Note: * *Fraud committed by Medical Resource Provider is currently out of the scope of our study.*

2.5 Our methods of detecting fraud

Statistical Rule Engine (Anomaly Detection): The anomaly detection method computes the likelihood of a claim to be fraudulent by analyzing historically reported claims based on defined outlier thresholds. The field investigation officers further investigate the instances flagged by the data mining model.

Decision Trees: A decision tree is structured like a flowchart in which each interior node depicts a "test" on a feature (For example, whether a claim investigation outcome ends up genuine or fraudulent), each branch reflects result of the test and each leaf node represents a class label (decision made after computing all features). The root-to-leaf path represents classification criteria.

A decision tree generates a segmentation of the input data hierarchically based on a series of conditions applied to every observation. Based on the value of one predictor, each rule assigns an observation to a section. Rules are imposed sequentially, resulting in segment hierarchy within sections. The hierarchy is a tree, and each segment is called a node. The initial section includes the entire information set and is called the root node. Each node and all of its successors form a branch. The last nodes are called leaves. A choice is taken on the response variable for each leaf and this decision is applied to all observations in that leaf. The actual decision relies on the response variable.

Deep Neural Network: A neural network is a collection of layers that are interconnected. The first layer is called the input layer and is linked to an output layer by an acyclic graph consisting of nodes and weighted edges.

You can insert several hidden layers between the input and output layers. Most predictive tasks can be achieved effortlessly with just one or a few hidden layers. Recent study, however, has shown that deep neural networks (DNN) with multiple layers can be highly efficient in complicated functions such as image recognition or speech recognition. The consecutive layers are used to model growing levels of semantic complexity and depth.

The neural network learns the relationship between inputs and outputs through input data training. The graph direction starts from the inputs through the hidden layer to the output layer. Every node in a layer is linked to nodes in the next layer by the weighted edges.

In the hidden layers and in the output layer, a value is calculated at each node to compute the network output for a specific input. The value is set by calculating the weighted sum of the previous layer's node values. Finally, an activation function is implemented to the weighted sum.

Averaged Perceptron: The averaged perceptron method is an early and very simple version of a neural network. In this method, inputs are classified into numerous possible outputs based on a linear function, and then combined with a set of weights obtained from the feature vector — hence the name "perceptron."

Simple perceptron models are generally used for learning linearly separable patterns, whereas neural networks (especially deep neural networks) may be appropriate for modeling more complex class boundaries. However, perceptrons are faster and can be used with continuous training because they process instances serially.

k-Means Clustering: Clustering is a mechanism for grouping observations based on their similarity with the purpose of managing them in groups. k-means clustering divides n observations into k (user-defined) clusters. Each observation belongs to the cluster with the closest mean, acting as the cluster's representation or prototype.

3 Healthcare Data

Medi Assist is the largest third-party administrator (TPA) in India servicing for over 15% of the Indian population. In terms of our reach, we work with over 6,500 organizations who collectively employ about 15 million individuals, spread across the country. Each day we process more than 5,000 claims. In India, TPA is in the epicenter of healthcare data exchange ecosystem which consists of stakeholders including:

- Insured – Individual or Group,
- Insurer,
- Provider (Hospital, Diagnostic Center, Pharmacy, etc.),
- Agents/Brokers and
- The Government of India.

We obtain **policy-related information** from Insurer, **hospitalization information** from Provider, **demographic information** from Insured, **customer acquisition information** from Agents/Brokers and **regulatory information** from the government –

Insurance Regulatory and Development Authority of India (IRDAI) to fully realize every outstanding claim.

We use over 160 features for our fraud analytics engine. The dataset used for decision tree model building is 481,207 historically investigated claims. The under-sampled Decision Tree uses 262,982 investigated records for training. The neural network uses 383,298 investigated claims for training and testing.

3.1 Feature Selection

The Fraud Detection Engine uses over 160 features for both statistical rule engine and supervised machine learning at the claim level. The claim level data is appended with relevant information from other levels including policy, beneficiary, provider, agent/broker, etc. For example, we calculate the loss ratio, earned premium, earned members, incidence ratio, hospitalization costs (actual, pre-hospitalization and post-hospitalization amounts) to determine the risk at the policy level and map it back to claims data using unique beneficiary ID.

An obstacle to fraud control initiatives is the expurgation of public discourse on new methods of fraud detection to avoid alerting fraudsters [11]. If fraudsters know how detection systems function, this could obstruct the effectiveness of fresh ideas before the opportunity to detect fraud arises. Academic literature thus rarely reveals the characteristics used for isolation of fraud [20]. The features created for this study were substantially original and cannot be disclosed owing to an agreement with Medi Assist authorities.

3.2 Data pre-processing

Some scientists indicate that preparation of data expends 80% of the time in any data analysis project [11]. Electronic health documents and database structures of raw claims are intended to support health care delivery and financial transactions, rather than query creation or fraud detection, and therefore need to be redesigned to facilitate data analysis activities. A data library was developed from the data files which could be used to help various inquiries. It was time-consuming to prepare data for this project. We estimate that the preparation of data took about 80% of the entire project time. On the other hand, once the data was loaded in a data warehouse, it could be accessed in a multitude of ways for various analytical applications so we anticipate that future data pre-process time will be considerably less than the initial development.

Pre-processing data involves data cleaning, data integration, data reduction, data transformation, and feature engineering. We create two types of masters:

- Blacklist Master: This list consists of entities (providers, beneficiaries, policies, corporates, agents/brokers, etc.) who were historically proven to be involved in fraudulent activities.
- Watch-list Master: This list consists of suspicious entities (providers) who seem to exhibit abnormal behavior in the recent past.

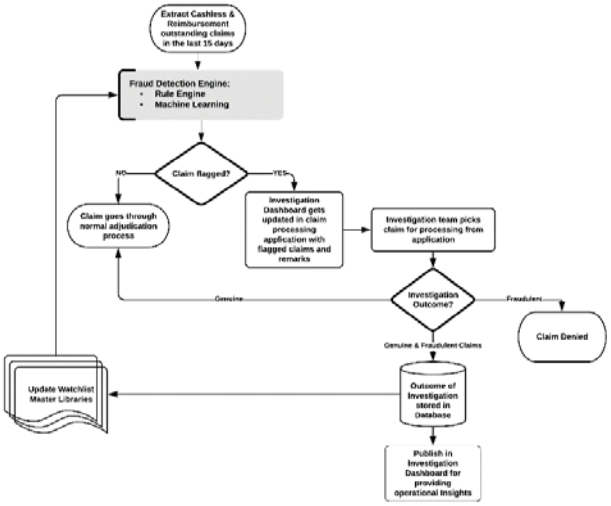

Figure 2 Fraud Analytics Workflow.

In addition to the masters, other feature engineered variables include chronic disease flagging, claim-months-tagging (which determines the duration from the start of policy period till the date of hospitalization), disease grouping where claims get segregated into 3 groups based on International Classification of Diseases (ICD 10) ailment coding.

- Disease Group 1: Consists of chronic conditions where the average cost of treatment and the average length of stay are normally high and repetitive. For example, circulatory, urology, neoplasm, endocrine disorders, etc.
- Disease Group 2: Consists of unpredictable or epidemic diseases where the average cost of treatment and the average length of stay are normally at the lower end of the cost spectrum. This includes injuries, infections, clinical findings, accidents, etc.
- Disease Group 3: Consists of the remaining ICD chapters including blood diseases, mental disorders, respiratory disorders, etc.

3.3 Handling class imbalance in dataset

Areas like medical insurance fraud experience class imbalance because there are significantly fewer instances of fraud vs. normal behavior [15].

Two sampling techniques are usually deployed to adjust the class distribution of a dataset. These are:

- Under-sampling: In supervised machine learning, we have created two decision trees, of which one is built based on under-sampled data. We use simple random under sampling (RUS) technique to remove samples from the majority class.
- Over-sampling: We have oversampled the externalized machine learning models built using neural network and averaged perceptron. The over-sampling algorithm used is called Synthetic Minority Oversampling Technique (SMOTE).

3.4 Data update frequency

For an effective fraud detection model, we have to ensure that the model is updated regularly to incorporate recent investigation findings. The rule engine, machine learning models and all the associated libraries are updated every day. After updating, the model flags outstanding claims for the day for further investigation to be conducted by field investigation officers.

4 Experimental Design

The fraud detection engine is designed to produce a subset of outstanding flagged claims every day for the investigation team. Our current model produces 12 binary flags against every outstanding claim. The first 10 flags use the statistical rule engine and the remaining 2 flags use machine learning (Decision Trees) for predicting the probability and classification of an outstanding claim. Figure (2) shows the process flow of the fraud analytics framework.

4.1 Rule-Based Method

In this system, fraud patterns are identified as violation rules. Rules may consist of a composite set of conditions derived by domain experts. An alert will be raised when all conditions are met [4]. The fraud detection engine creates 10 flags against every claim record. The claim gets evaluated based on:

a) *Beneficiary's current and historical claim experience*: For example, as suppression of pre-existing condition, early/late claimer, etc.
b) *Admitted hospital and hospitalization experience*: For example, suspicious/blacklisted hospital, average length of stay, network type, etc.
c) *Ailment group analysis*: For example, disease grouping, chronic disease, etc.
d) *Riskiness of policy:* For example, high incidence ratio, burnout ratio, etc.
e) *Demographic analysis:* For example, fraud-prone areas, age-band, etc.

It is important to note that the rule engine uses outlier analysis and central tendencies of variables to define threshold values. It is far more effective to use statistical methods than to analyze individual claims [11]. For example, one such defined rule could be - *the occurrence of a* ***high-value*** *claim being reported from a non-network hospital and also has a high length of stay*. In this, the definition of "high value" could either be hard-coded (such as incurred claim amount = Rs. 75,000 in INR or above) or statistically derived based on the distribution of dataset (such as the third-quartile value of claimed amount) and we use the latter. Because both fraud and genuine trends in healthcare data may transform over time, techniques for detecting fraud in health care need to be dynamic enough to adapt to these modifications. Moreover, this method forces the engine to make data-driven decisions instead of expert-driven decisions. Our objective is to develop self-evolving solutions for fraud detection [12].

4.2 Supervised Learning Methods

We have used Decision Trees (a modified C4.5 algorithm) for the implementation of a binary classifier. Decision Tree 1 (for Flag 11) is created using the full investigated dataset, however, Decision Tree 2 (for Flag 12) is created using the under-sampled investigated dataset.

Also, an externalization of the fraud detection engine for other stakeholders to consume has been provisioned using deep neural networks. This proactive model is created using 3,83,298 investigated claims and is operationalized as a web-service so users can send data to our model using the web service REST API and receive back the results in real-time. The API service can be consumed in 2 modes:

1) Request-Response: Every claim to be sent individually for scoring and classification.

2) Batch: Scores a batch of claims.

4.3 Operationalizing the engine

Each outstanding claim goes through the fraud detection engine for flagging. If even one of the 12 binary flags is set, the claim gets shortlisted for actual field investigation. The team is notified about these yet-to-be-investigated claims with relevant remarks for flagging to assist in the investigation.

After analyzing the performance of each flag based on their respective hit-rates, we have further defined a weighted scoring mechanism that considers the severity of every flag (based on their historical performance) along with time-decay and sorts the flagged outstanding claims into a single priority queue. This method can calculate each fraud detection factors' contribution rate of fraud [14]. The claims featured on top of the list are considered more severe than the ones afterward. If a claim gets picked for investigation, it gets removed from this unified queue.

Measures have been taken to ensure that the rate of flagged claims does not exceed the rate of current investigations. This calibration is necessary to ensure sufficient field investigation officers are available to handle the additional workload for investigating claims triggered by fraud detection engine. Currently, the fraud detection engine averages at 477 flagged claims per day and the average number of investigations conducted per day are 573 claims (averaged over one year). Besides, the stakeholders are provided with an interactive dashboard to track the adherence (of the team), performance (i.e. hit-rate of the engine), savings, analyze fraud claims not identified by fraud detection engine but from other methods, the progress of fraud detection engine and its flags, etc.

5 Results & Discussion

We had deployed the fraud detection engine on 8th

August 2018 and the model has been assisting the investigation team with claims that are likely to be fraudulent. Based on our pilot case study for one insurer, the statistics before the pilot was as follows:

- Duration: 1st April, 2017 to 25th April, 2018
- Claims Investigated: 2,190
- Genuine: 1,999
- Fraudulent: 191
- Hit-Rate: 8.72%

Statistics post pilot:

- Duration: 25th April, 2018 to 25th April, 2019
- Triggered & Investigation: 2,827
- Investigation & Fraudulent: 763
- Hit-Rate: 26.98%

The percentage of increase in hit-rate was at 209.40% from the earlier hit-rate before deploying the fraud detection engine.

The model was eventually deployed on 8th August 2018 for all 4 public and 31 private sector insurance companies for which Medi Assist is the TPA. It is operating with an overall hit-rate of 10.3% (as of 13th June 2019) and has helped in saving Rs. 6,54,65,257/- (in INR) for Medi Assist's public and private insurer partners.

The tables I and II summarized the evaluation statistics for two-class neural network and two class averaged perceptron respectively. The averaged perceptron has better AUC but comes at a cost of a slightly reduced F1 Score for the same threshold. The ROC (or receiver operating characteristic) curve, is a graphical representation of the diagnostic capability of a binary classifier system as its threshold of discrimination is varied. It determines how much a model can differentiate between classes. The model is better at distinguishing between fraudulent and genuine claims for higher AUC.

Table I Summary Statistics for Neural Network

True Positive	**False Negative**	**Accuracy**	**Precision**	**Threshold**
97,207	771	0.973	0.980	0.5
False Positive	**True Negative**	**Recall**	**F1 Score**	**AUC**
1,943	796	0.992	0.986	**0.48**

Table II Summary Statistics for Averaged Perceptron

True Positive	**False Negative**	**Accuracy**	**Precision**	**Threshold**
1,07,653	3,589	0.946	0.963	0.5
False Positive	**True Negative**	**Recall**	**F1 Score**	**AUC**
4,098	27,929	0.968	0.966	**0.982**

The ROC curve is plotted with True Positive Rate (TPR or sensitivity) against the False Positive Rate (FPR or 1-specificity) where TPR is on y-axis and FPR is on the x-axis.

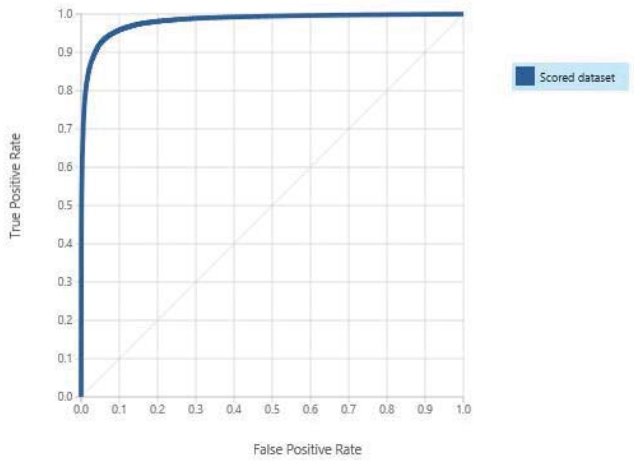

Figure 3 ROC curve of two class averaged perceptron.

We used the ROC curve of Decision Tree to estimate that the lowest possible threshold with the highest gain is 0.03. This modification will increase the number of false positives, however, it also drastically reduces the possibility of fraudulent claims slipping through our detection system. Figure (5) below demonstrates the misclassification assessment after updating the threshold to 0.03.

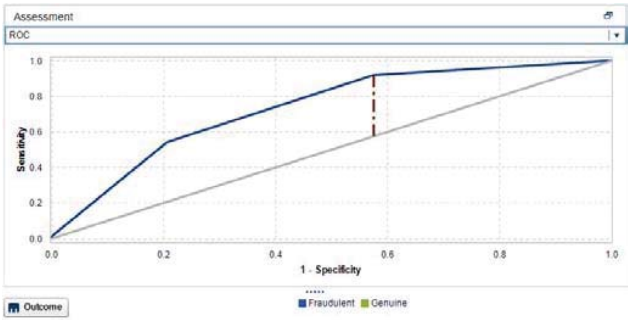

Figure 4 ROC curve of Decision Tree (down-sampled).

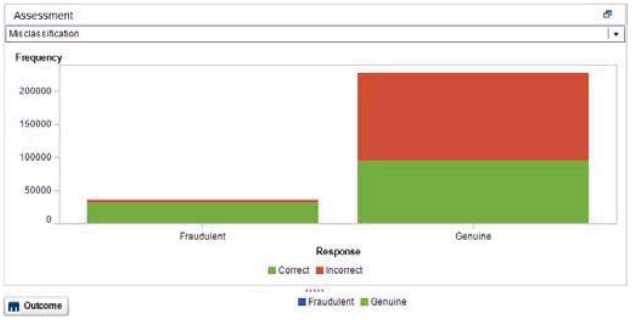

Figure 5 Misclassification assessment of Decision Tree (down-sampled).

6 Scope & Future of Research

Perhaps the areas we described are those where statistical and other techniques of data analytics have had the greatest impact on fraud detection. A comparative analysis was performed between two-class

classifier models including Averaged Perceptron, Bayes Point Machine, Decision Forest, Boosted Decision Tree, Decision Jungle, Logistic Regression, Locally-Deep Support Vector Machine, Neural Network, and Support Vector Machine. We have identified that averaged perceptron model outperforms all other models and takes a fraction of the time consumed by the neural network for model building (training, scoring, and evaluation). The externalized fraud detection model will, therefore, be updated with Averaged Perceptron for better performance in real-time.

6.1 Hybrid Model

It is discovered that detection of fraud using supervised learning is more effective and precise [2]. A hybrid model is created by using the concepts of supervised as well as unsupervised learning techniques [4].

A supervised machine learning model can only predict the possible investigation outcome of a claim based on historical experiences involving fraudulent activities and patterns. It cannot classify new types of fraudulent claims. However, it is essential for the engine to address aberrations and abnormal claim behavior as anomalies.

1) *Unsupervised Learning:* We use k-means clustering on bill level details of claims to create 5 clusters (k=5). Each claim gets assigned to a cluster and gets a Cluster ID. We use 11 bill level cost breakup features to create 5 clusters with each feature containing 16 bin levels. Some of the features are Operation Theater Charges, Consultant Charges, Investigation Charges, Room Charges, Operation Theater Charges, Surgeon and Surgery Charges, number of bill items, etc.
2) *Supervised Learning:* A new decision tree is trained with one of the predictor input variables being Cluster ID. Therefore the output of unsupervised clustering is factored into the decision tree as an additional categorical feature.
3) *Implementation:* First, the outstanding claims will iterate through the k-means clustering algorithm and get assigned to a cluster. The Decision Tree model is applied to the resulting dataset to predict and classify investigation outcomes based on claim characteristics.

6.2 Usual, Customary and Reasonable (UCR) Validation

It is defined as being "The amount paid for a medical service in a geographic area based on what providers in the area usually charge for the same or similar medical service. The UCR amount is then used as a reference to determine the allowed amount."

We will define UCR costs by procedures for every city in India. Later, the fraud detection engine can use the reference UCR master to cross-validate the claimed amount with UCR amount for the same procedure within the city. This process should highlight cost-based outliers in claims.

6.3 External Blacklist Master

Blacklisted entities are generally published by stakeholders such as private insurers and government authorities.

a) The General Insurance Council (GIC) of India publishes a report on fraudsters (which includes hospitals, beneficiaries, agents/brokers, investigators, etc.) in their Fraud Risk Mitigation Portal (FRMP). As a member of GIC, Medi Assist is provided with access to this portal.

b) Insurance Regulatory and Development Authority (IRDA) of India publishes a report on fraudulent agents for health, life, and general insurance companies. The data can be used to cross-verify if these agents are involved in claim submission activities in Medi Assist.

c) Medical Council of India (MCI) publishes a list of blacklisted doctors from 1966 to 2015. The data consists of the registration number, state medical council, and duration of suspension of license for each blacklisted practitioner.

d) Most private insurers (such as Aditya Birla, Bajaj Allianz, Future Generali, etc.) publish blacklisted hospitals on their public websites. This data can be consumed to create a geographical heatmap to further scrutinize claims getting reported from fraud-prone areas.

6.4 Other parallel developments

Provider scoring engine – This is scored at the provider level based on TPA, beneficiary and insurer's experience with the provider. Some parameters used to evaluate each provider include:

a) Beneficiary's hospitalization experience (ratings).

b) Claim processing experience – ratio between number of investigated claims identified as fraudulent and total claims reported by provider, number of loss control measure communications conducted with the provider, number of reimbursement claim vs. number of cashless claim submission ratio, percentage of claims identified as fraudulent, average cost of treatment, UCR adherence percentage, etc.

A provider ranking system can provide insights to our provider contracting team, audit team and fraud detection model to increase or decrease the level of claim scrutiny based on our historic relationship with every networked provider.

7 Scope & Future of Research

In this research work, the concepts of data mining and machine learning have been applied for building an engine that can assist in identifying fraudulent claims filed by beneficiaries. Various factors were taken into consideration to ensure that the model is updated and evolving every day. Later, the engine was deployed within our production pipeline to improve the

investigation team's operational efficiency.

Research can be undertaken in the future to determine which traditional fraud indicators contain useful claim sorting information and dynamically weigh every flag based on two important criteria - the achieved hit-rate and the performance of individual flags. Furthermore, unsupervised methods or the hybrid of supervised and unsupervised methods can be used by leveraging the advantages from both the approaches and constructing a hybrid or semi-supervised model. Also, a model can be externalized to process real-time claims, detect frauds and flag them for further investigation proactively.

Acknowledgements

My sincere gratitude to our Chairman Dr. Vikram Chhatwal, Chief Executive Officer Mr. Satish Gidugu and Chief Technical Officer Mr. Himanshu Rastogi for their expert guidance, technology assistance, insights and valuable suggestions for the formulation of the engine. I would like to include a special note of thanks to Dr. Pradeep G. Siddheshwar for his feedback and assistance that greatly improved this manuscript. I thank my colleagues, underwriters, doctors, investigation team and all my well-wishers for their patience and assistance. Finally, I would like to acknowledge with gratitude, the support, and love of my family.

References

[1] N. Obodoekwe, and D. T. van der Haar, ``A Comparison of Machine Learning Methods Applicable to Healthcare Claims Fraud Detection," Springer Nature Switzerland AG, A. Rocha et al. (Eds.): ICITS 2019, AISC 918, pp. 548–557, 2019. https://doi.org/10.1007/978-3-030-11890-7_53

[2] A. Sheshasaayee and S. S. Thomas, ``A Purview of the Impact of Supervised Learning Methodologies on Health Insurance Fraud Detection," in Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing 672, https://doi.org/10.1007/978-981-10-7512-4_98. Springer Nature Singapore Pte Ltd., V. Bhateja et al. (eds.).

[3] C. Sun, Z. Yan et. al., ``Abnormal Group based Joint Medical Fraud Detection," IEEE Access, 2018. DOI 10.1109/ACCESS.2018.2887119, IEEE Access.

[4] P. Pandey, A. Saroliya and R. Kumar, ``Analyses and Detection of Health Insurance Fraud using Data Mining and Predictive Modeling Techniques," in Soft Computing: Theories and Applications, Advances in Intelligent Systems and Computing 584, https://doi.org/10.1007/978-981-10-5699-4_5

[5] Yi Peng, G. Kou et. al., ``Application of Clustering Methods to Health Insurance Fraud Detection," IEEE 2006.[DOI 1-4244-0451-7/06].

[6] S. Kareem, Dr. R. B. Ahmad, and Dr. A. B. Sarlan, ``Framework for the Identification of Fraudulent Health Insurance Claims using Association Rule Mining," in IEEE Conference on Big Data and Analytics (ICBDA), 2017. [DOI 978-1-5386-0790-9/17].

[7] A. Verma, A. Taneja, and A. Arora, ``Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques," in Proceedings of 2017 Tenth International Conference on Contemporary Computing (IC3), 10-12 August 2017, Noida, India.

[8] V. Rawte, G. Anuradha, ``Fraud Detection in Health Insurance using Data Mining Techniques," in International Conference on Communication, Information \& Computing Technology (ICCICT), Jan. 16-17, Mumbai, India [DOI 1-4244-0451-7/06].

[9] J. Peng, Q. Li et. al., ``Fraud Detection of Medical Insurance Employing Outlier Analysis," in Proceedings of the 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design. [DOI 978-1-5386-1482-2/18].

[10] Dr. M.S. Anbarasi and S. Dhivya, ``Fraud Detection Using Outlier Predictor in Health Insurance Data," in International Conference On Information, Communication \& Embedded Systems (ICICES 2017) [DOI 978-1-5090-6135-8/17].

[11] L. Copeland, D. Edberg et. al., ``Applying Business Intelligence Concepts to Medicaid Claim Fraud Detection," in Journal of Information Systems Applied Research (JISAR) 2012.

[12] Q. Liu and M. Vasarhelyi, ``Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information," in 29th World Continuous Auditing and Reporting Symposium (29WCARS), November 21-22, 2013, Brisbane, Australia.

[13] E. A. Duman, and S. Sa˘gıro˘glu, ``Health Care Fraud Detection Methods \& New Approaches," in 2nd International Conference on Computer Science and Engineering (UMBK'17) 2017 [DOI 978-1-5386-0930-9/17].

[14] H. Peng and M. You, ``The Health Care Fraud Detection Using the Pharmacopoeia Spectrum Tree and Neural Network Analytic Contribution Hierarchy Process," in 2016 IEEE TrustCom-BigDataSE-ISPA [DOI 10.1109/TrustCom/BigDataSE]

[15] R. A. Bauder and T. M. Khoshgoftaar, ``Medicare Fraud Detection using Random Forest with Class Imbalanced Big Data," in 2018 IEEE International Conference on Information Reuse and Integration for Data Science [DOI 10.1109/IRI.2018.00019].

[16] M. Herland, R. A. Bauder and T. M. Khoshgoftaar, ``Medical Provider Specialty Predictions for the Detection of Anomalous Medicare Insurance Claims," in 2017 IEEE International Conference on Information Reuse and Integration [DOI 10.1109/IRI.2017.29]

[17] A. Sheshasaayee and S. S. Thomas, ``Usage of R Programming in Data Analytics with Implications on Insurance Fraud Detection," in Springer Nature Switzerland AG 2019. ICICI 2018, LNDECT 26, pp. 416–421, 2019. https://doi.org/10.1007/978-3-030-03146-6_46

[18] N. Obodoekwe, and D. T. van der Haar, ``A Critical Analysis of the Application of Data Mining Methods to Detect Healthcare Claim Fraud in the Medical Billing Process," in Springer Nature Switzerland AG 2018. UNet 2018, LNCS 11277, pp. 320–330, 2018. https://doi.org/10.1007/978-3-030-02849-7_29

[19] N. Obodoekwe, and D. T. van der Haar, ``A Critical Analysis of the Application of Data Mining Methods to Detect Healthcare Claim Fraud in the Medical Billing Process," in Springer Nature Switzerland AG 2018. UNet 2018, LNCS 11277, pp. 320–330, 2018. https://doi.org/10.1007/978-3-030-02849-7_29

[20] Bolton, R. J., \& Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235-249.

[21] P. L. Brockett, R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert, ``Fraud classification using principal component analysis of ridits," Journal of Risk \& Insurance, vol. 69, no. 3, p. 341C371, 2002.

[22] H. Joudaki, A. Rashidian, B. Minaeibidgoli, M.

Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, ``Using data mining to detect health care fraud and abuse: A review of literature," in Global Journal of Health Sciences, vol. 7, no. 1, p. 37879, 2015.

[23] McDougall, C.: Born to Run: A Hidden Tribe, Superathletes, and the Greatest Race the World Has Never Seen. Alfred A. Knopf, New York (2009).

[24] Behl, D., Handa, S., \& Arora, A. (2014, February). A bug mining tool to identify and analyze security bugs using naive bayes and tf-idf. In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on (pp. 294-299). IEEE.

[25] Bush, J., Sandridge, L., Treadway, C., Vance, K., Coustasse, A.: Medicare fraud, waste and abuse. In: Business and Health Administration Association Annual Conference (2017).

[26] NHCAA The problem of health care fraud, consumer alert: the impact of health care fraud on you, report of national health care anti-fraud association (NHCAA).

[27] Li J, Huang K, Jin J, Shi J (2008) A survey on statistical methods for health care fraud detection. Health Care Manage Sci, 275–287.