A Survey on Fraud Analytics Using Predictive Model in Insurance Claims

K. Ulaga Priya and S. Pushpa

Dept of Computer Science and Engineering, St.Peters University.
ulagapriya@gmail.com

Dept of Computer Science and Engineering, St.Peters University.
pushpasangar96@gmail.com

Abstract

Insurance Industry is a rapidly growing fast industry in terms of large amount of data. The most critical issue in insurance industry is fraudulent claims. Fraud is nothing but wrongful or criminal trick planned to result in financial or personal gains. As the size of data increases, the traditional approach will not work and it will be tedious job to identify the fraudulent claims. Moreover, new types of claim will emerge and hence it will be difficult to predict the fraudulent claims. This paper depicts an overview of Fraud analytics, prediction, and Data Science algorithms based predictions in insurance industry.
1. Introduction

Fraud analytics is a type of data analytics where data analysis is done on the fraudulent behaviour. There are several domains where fraud may happen like Credit card fraud, telecommunication fraud, Insurance fraud, Healthcare fraud, tax evasion etc. Credit card fraud is one of the fraud types which is surveyed widely in the domain of fraud detection.[34],[35],[36]. Due to the popular mode of payment transaction, both online and offline, the fraud associated with it is also increasing. There are multiple techniques to detect credit card fraud like Neural Network [10-11], Group Method of Data Handling [4-5], Bagging[6]. Some other popular models of credit card fraud detection are Hidden Markov Model [2-3], Bayesian learning [7-9], K-means Clustering [1]. The credit card fraud was categorised [35] as two categories namely behavioural frauds and Application frauds. Application frauds happen whenever fraudsters[33] acquire new cards by providing false data to issuing companies[33]. Behavioural frauds include four types: mail theft, fake cards, stolen/lost cards. Several algorithms [43] in credit card fraud prediction were compared and derived that Bagging ensemble classifier is the best method.

Telecommunication fraud is rapidly increasing due to the growth of recent technology and global communication which results in considerable losses in business. There are two categories of telecommunication fraud: subscription fraud and super imposed fraud. Subscription fraud is nothing but claiming false identity for getting service and elude payment. Superimposed fraud happens whenever the service is used without having relevant rights and is usually detected by the appearance of ‘phantom’ call on a bill. Various techniques used in telecommunication fraud detection[12] are Neural Networks, Visualization Methods and Rule-based Approach.

Insurance fraud is defined [37] as fraud in the insurance industry as perceptively creating a fabricated claim, bloating a claim or adding further items to a claim, or being in any way deceitful with the intention of getting more than legitimate privilege. The insurance fraud types include exaggerated claims, fabricated medical history, post-dated policies, faked damage etc. [30] This emphasize the different types of fraud in health insurance sector. There are different techniques for health insurance fraud detection[22]. This paper concentrates on Insurance Fraud and its data analytics. The National Healthcare Anti-Fraud Association (NHCAA) evaluated the health care claims and announced that 10 percent of health care claims contain some element of fraud [38][39]. Insurance protects the customer from monetary loss. Insurance Policy is a legal agreement between the Policy holder and insurance [23] company which specifies the claim amount which the Policy Holder needs to pay. Insurance claim is nothing but, the policy holder request the claim amount from the insurance company based on the insurance policy. Insurance domain can be categorised as (i) Health Insurance. (ii) Travel Insurance (iii) Auto insurance (iv) Life insurance
The section 2 describes about Bigdata analytics in identifying fraudulent claims. Section 3 describes one of the Bigdata Analytics type which is predictive Analytics. This section also discusses the various types of Machine learning algorithms, Section 4 discusses the merits and demerits of the algorithm, it also explains the Fraud analytics process model. Section 5 discusses the performance benchmark of different types of fraud. Section 6 depicts the conclusion and Section 7 holds the references.

2. Big Data Analytics

Fraud detection in insurance is a potential area in insurance where big data plays a major role. However, many insurers remain unknown about the power of data analytics. According to the survey conducted almost 80% of insurer is unaware about the power of Big Data Analytics. Let us examine a few data analytic models that can help insurers strengthen their fraud detection capabilities.

i) Descriptive – Analysing the data on what was already happened. Generally, reports were generated with past data and analysis is done on that data. For example, to identify the sales distribution that has happened in previous year.

![Descriptive Analytics Example]

ii) Diagnostic – Based on the previous data, data analysis will be done on why it is has happened. Identifying and analysing the reason for poor sales in the previous year is an example of diagnostic data analysis.

iii) Predictive – This type of analysis will suggest what will happen in the future. It predicts the futuristic scenario based on past historical data. For example, identifying the area that is likely to perform better sales in the current year based on past data.

iv) Prescriptive – This type of analysis will suggest what action should be taken. Basically, how we can make it happen. It gives recommendation on what needs to be done. For example, how to achieve the best outcome in sales, and strategy to retain key customers.

3. Predictive Analytics

This paper discusses on predictive analytics and the techniques used for prediction. Supervised learning and Unsupervised learning are the techniques used for predictive analytics. Supervised Learning will have a target variable. Target variable is the output that is predicted using other relevant
features. Unsupervised Learning does not have a target variable. Following supervised learning techniques are used for predicting fraudulent claims since it has a target variable.

- Decision tree
- Random Forest
- Support Vector Machines
- Neural Network
- XGBoost

These techniques are used in solving data analytics problems.

**Decision Tree**

Decision tree gives a visualisation view in the form of graph. The sample set is divided into subset of trees which represent choices and their results. Each node of a tree represents a choice and the edges represent the decision. The sample dataset is categorised into training dataset and test dataset. A model is created with training dataset which gives the prediction accuracy. This model is applied on the test dataset and the accuracy of prediction is validated. For each predictor variable, this model can be used to decide on the category (Yes/No, Spam/not spam) of the data. Decision tree can deal with continuous data through various method of decision tree like ID3 method and C4.5.

Decision tree is used in various fraud detection and prediction applications. Some of the fraudulent problem areas where decision tree is used are credit card fraud, Energy fraud etc. Credit card fraud detection [40] uses a cost sensitive decision tree approach. Decision tree is also effectively used in Energy Fraud detection. This technique is widely used for classification and regression. M5P Decision tree is used for energy fraud detection which is a modified version of Quinlan’s [12] M5 algorithm. Following is the general algorithm.

| **Input:** Training dataset |
| **Output:** To create a decision Tree. |

*Step 1:* Identify the best attribute of the dataset which need to be placed at the root of the tree.

*Step 2:* Divide the training set into subsets. Each subset should contain data with the same value such that each subset is created for an attribute.

*Step 3:* Till you find leaf nodes step1 and step2 is to be repeated on each subset in all the branches of the tree.

**Random Forest**

In the random forest technique, multiple decision trees are created. A random
subset of the training data is used to create a single decision tree. [16] The common result of each random subset is taken as the final tree output. A new study is fed into all the trees and majority vote for each classification was taken in this model. Missing values and outliers are taken care in random forest model.

The predictive algorithm which uses this technique will try to imitate the relationship between input and output variable. This algorithm provides excellent accuracy and it runs very effectively on large datasets. This algorithm[14] is widely used for large number of input. Moreover, it has methods for maintaining balance for the unbalanced datasets.

It is identified that for the aggregated model random forest gives better results than Naïve Bayes. Where as in the personalized models Naïve Bayes gives better results. In online shopping [15] when large number of discounts are announced, it paves way for unusual activities in purchasing products and services. This paper uses random forest algorithm to detect faults using R language. Prediction can be done using Random Forest technique to identify customer’s preference regarding the choice of insurance policy options. [12]. Following is the algorithm:

<table>
<thead>
<tr>
<th><strong>Input:</strong> Training dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Output:</strong> To create “n” of Trees</td>
</tr>
<tr>
<td><strong>Step 1:</strong> Randomly select “k” features from total “m” features Where k &lt;&lt; m</td>
</tr>
<tr>
<td><strong>Step 2:</strong> Among the “k” features, calculate the node “d” using the best split point.</td>
</tr>
<tr>
<td><strong>Step 3:</strong> Split the node into daughter nodes using the best split.</td>
</tr>
<tr>
<td><strong>Step 4:</strong> Repeat 1 to 3 steps until “l” number of nodes has been reached.</td>
</tr>
<tr>
<td><strong>Step 5:</strong> Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.</td>
</tr>
</tbody>
</table>

**Neural Networks**

The fundamental element of computation in neural network is the neuron which is also called as node or unit. The input from other nodes is computed and produces an output. Basically, it converts the input from multiple sources to output. Whereas in human brains has a distinct feature of creating transient states through neurons in between sensory organ and brain which is the decision taking unit.
To detect and predict the risk of fraudulent financial reporting, a Multilayer Perceptron (MLP) [17] Artificial Neural Network model was proposed. Weatherford suggests, artificial immune systems, recurrent neural networks, back propagation neural networks for fraud detection. A neural network approach is identified to detect management fraud. The management fraud is detected [18] using the Adaptive Logic Network and generalized adaptive neural network.[42] A three-layer was used with feed-forward Radial Basis Function (RBF) neural network which will produce in every two hours for new credit card transactions. This also propose fuzzy neural networks on parallel machines which rises the rule production for customer-specific credit card fraud detection. Neural Network gave better results for prediction when compared to Logistic Regression ad Decision Tree[18]. A case study was done with 5 strategies to audit the auto insurance claims[9].

<table>
<thead>
<tr>
<th><strong>Input:</strong> Training dataset</th>
<th><strong>Output:</strong> To create data model.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1:</strong> Assign random weights to all the linkages to start the algorithm</td>
<td><strong>Step 2:</strong> Using the inputs and the (input-hidden node) linkages find the activation rate of hidden nodes</td>
</tr>
<tr>
<td><strong>Step 3:</strong> Using the activation rate of hidden nodes and linkages to output, find the activation rate of output nodes</td>
<td><strong>Step 4:</strong> Find the error rate at the output node and recalibrate all linkages between hidden nodes and output nodes</td>
</tr>
<tr>
<td><strong>Step 5:</strong> Using the weights and error found at the output node, cascade down the error to hidden nodes</td>
<td><strong>Step 6:</strong> Recalibrate the weights between hidden node and the input nodes repeat the process till the convergence criterion is met</td>
</tr>
<tr>
<td><strong>Step 7:</strong> Using the final linkage weights score the activation rate of the output nodes</td>
<td></td>
</tr>
</tbody>
</table>

**XGBoost**

XGBoost is a short form for Extreme Gradient Boosting. Boosting is a sequential process. Multiple trees are created and the information of the first tree is fed as input to the second tree so that it improves the prediction in subsequent iterations. Basically it is a additive tree model where it add new trees that complement the already built ones. XGBoost handles missing values and it works only for numeric data.

**Support Vector Machine**

Support Vector Machines (SVM) is also a supervised learning algorithm used for regression and classification problems. In general, it creates a hyper plane in n dimensional space to classify the data based on target class. The SVM separates into different classes through a hyperplane or multiple hyperplane.
The hyperplane separates the data points and sometimes it is difficult to separate the data point through a single hyperplane. The distance between the data point and hyperplane represents a margin.

This enables to perform classification or regression also. Since it has many features SVM becomes a promising technique in prediction. [25] Basically, SVM works on the principle that data points are segregated through hyperplanes. This subsequently maximizes the distance between data points, and the hyper plane is constructed with the help of support vectors. A Turkish insurance company database [19] was taken for research. SVM technique was applied to this data. SVM is basically a classification technique that identifies each record as anomalous or normal record. Subsequently every record is checked with margin and based on that the record is treated as normal or anomalous. SVM is a kernel based [19] algorithm where kernel transmutes the input data points to a high-dimensional space so that the problem is solved. [25] There are different applications which detect fraud through SVM. The top management fraud is detected using SVM, to create the Fraudulent Financial statement. [20]

4. Discussion

A comparative study is done on the Supervised Technique. Each technique has its own merits and demerits. Based on the application area and data technique can be chosen and analytics can be done on that. The merits and demerits are discussed below as follows:

<table>
<thead>
<tr>
<th>Supervised Learning Technique</th>
<th>Merits</th>
<th>Demerits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>Non-Analytical people also can easily understand</td>
<td>This is not suitable for continuous variables</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Can handle large no. of variables and also provides most important variable</td>
<td>Not suitable for classification problem than regression problem</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Breaks very well when the margin for separation is very clear. Otherwise nonlinear classification can be done.</td>
<td>Not Suitable for large data sets</td>
</tr>
<tr>
<td>Neural network</td>
<td>Works well even in underfitted state</td>
<td>Complex network structure</td>
</tr>
<tr>
<td>XGBoost</td>
<td>One tree is built at a time, where each new tree helps to correct errors made by previously trained tree</td>
<td>Generally, takes longer because of the fact that trees are built sequentially</td>
</tr>
</tbody>
</table>

**Fraud Analytics Process Model**

As a first step the business problem must be clearly identified. Next step is to identify the data source which is a very important task in data analysis model. [29] Then subsequently all the data is gathered in one single area which could be a data mart or data warehouse. Then the data is cleaned up re, inconsistent, missing and duplicate values are removed. Additional data transformation is done like data type conversion etc. In analytics phase, data model is built and data is analysed with the newly created model. Once data analytics is done, this
will be examined by functional experts. During the analytics phase, the requirement of additional data may be identified. This triggers the need for another round of data cleaning and transformation. The Pre-processing phase is most time consuming.

5. Performance Benchmark for Different Types of Fraud

The following Scatter plot shows unique fraud types which were discussed and published in various fraud detection papers. These were some of the common fraud types highlighted in the Scatter Plot.

![Different Types of Fraud](image)

The following table provides references with performance metric of different Fraud Types. For better comparison of different types of fraud, the area under Receiver operating characteristic (ROC) curve are only included.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Fraud Type</th>
<th>Dataset Size Used</th>
<th>PERCENTAGE OF Class Distribution</th>
<th>Performance Metric-Area under the curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ortega Figuerora et al. (2006)</td>
<td>Medical Insurance</td>
<td>8,819</td>
<td>5%</td>
<td>AUC 74%</td>
</tr>
<tr>
<td>Sadeghi Farzad et al. (2011)</td>
<td>Automobile Insurance fraud</td>
<td>3,451</td>
<td>1.3%</td>
<td>AUC 71%-92%</td>
</tr>
<tr>
<td>Battacharyya Jha et al. (2011)</td>
<td>Credit card fraud</td>
<td>50 million transactions on about 1 million credit cards from a single country</td>
<td>0.005%</td>
<td>AUC 90.8%-95.3%</td>
</tr>
<tr>
<td>Whitrow, Hand et al. (2009)</td>
<td>Credit Card Fraud</td>
<td>33,000 - 36,000 activity records</td>
<td>0.1%</td>
<td>Gini 85% (~AUC=92.5%)</td>
</tr>
<tr>
<td>Van Vlasselaer Bravo et al. (2015)</td>
<td>Credit Card Fraud</td>
<td>3.3 million transactions</td>
<td>&lt;1%</td>
<td>AUC 98.6%</td>
</tr>
<tr>
<td>Dongshan and Girolami (2017)</td>
<td>Telecommunication Fraud</td>
<td>809,395 calls from 1,067 accounts</td>
<td>0.024</td>
<td>AUC 99.5%</td>
</tr>
<tr>
<td>Van Vlasselaer Meskens et al. (2013)</td>
<td>Social Security Fraud</td>
<td>2000 observations</td>
<td>1%</td>
<td>AUC 80%-85%</td>
</tr>
</tbody>
</table>

6. Conclusion

Like Insurance fraud detection, several fraudulent behaviours are available like Intrusion detection fraud, credit card fraud, telecommunication fraud etc. It is prominent that health insurance fraud is viable since it brings heavy loss overall. By integrating big data technology these claims can be predicted for
large volume of data as well as different variety of data.

References


