International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Healthcare Insurance Fraud Detection Leveraging Big Data Analytics

Prajna Dora[1], Dr. G. Hari Sekharan[2]

[1]Department of Information Technology, Database systems, SRM University, Katankulathur, Chennai, India

[2]Professor, Department of Information Technology, Database systems, SRM University, Katankulathur, Chennai, India

Abstract: ***Health Insurance fraud is a major crime that imposes significant financial and personal costs on individuals, businesses, government and society as a whole. So there is a growing concern among the insurance industry about the increasing incidence of abuse and fraud in health insurance. Health Insurance frauds are driving up the overall costs of insurers, premiums for policyholders, providers and then intern countries finance system. It encompasses a wide range of illicit practices and illegal acts. This paper provides an approach to detect and predict potential frauds by applying big data, hadoop environment and analytic methods which can lead to rapid detection of claim anomalies. The solution is based on a high volume of historical data from various insurance company data and hospital data of a specific geographical area. Such sources are typically voluminous, diverse, and vary significantly over the time. Therefore, distributed and parallel computing tools collectively termed big data have to be developed. Paper demonstrate the effectiveness and efficiency of the open-source predictive modeling framework we used, describe the results from various predictive modeling techniques .The platform is able to detect erroneous or suspicious records in submitted health care data sets and gives an approach of how the hospital and other health care data is helpful for the detecting health care insurance fraud by implementing various data analytic module such as decision tree, clustering and naive Bayesian classification. Aim is to build a model that can identify the claim is a fraudulent or not by relating data from hospitals and insurance company to make health insurance more efficient and to ensure that the money is spent on legitimate causes. Critical objectives included the development of a fraud detection engine with an aim to help those in the health insurance business and minimize the loss of funds to fraud.***

Keywords: Big Data, Hadoop, RHadoop, Decision tree, Naive Bayesian classification, Clustering

1. Introduction

Health care has become a major expenditure .The size of the health care sector and the enormous volume of money involved make it an attractive fraud target. Healthcare fraud, based on the definition of the National Health Care Anti-fraud Association is an intentional deception or misrepresentation made by a person or an entity, with the knowledge that the deception could result in some kinds of unauthorized benefits to that person or entity. Not only is the financial loss a great concern, fraud also severely hinders the health care system from providing quality and safe care to legitimate patients. Therefore, effective fraud detection is important for improving the quality and reducing the cost of health care services.

Frauds committed by a Policyholder could consist of members that are not eligible, concealment of age, concealment of pre-existing diseases, failure to report any vital information, providing false information regarding self or any other family member, failure in disclosing previously settled or rejected claims, frauds in physician's prescriptions, false documents, false bills, exaggerated claims, etc.

An important question being asked today in health informatics is how big data healthcare implementations can help correlate and collates insights across various heterogeneous data sources to enable a better understanding of issues. This method studies how a big data framework can be leveraged to extract and preprocess data. They leveraged on Hadoop as our big data framework to archive performance, scalability and fault tolerance. Hadoop is a popular open source map-reduce implementation, which is being used as an alternative to store and process extremely large data sets on commodity hardware. To achieve this goal, they use Hive as an open-source data warehousing solution built on top of Hadoop. In addition, we use Hive as an open-source data warehousing solution built on top of Hadoop. Hive supports queries expressed in a SQL like declarative language – HiveQL. RHadoop is a bridge between R, a language and environment to statistically explore data sets, and Hadoop, a framework that allows for the distributed processing of large data sets across clusters of computers. RHadoop is built out of three components which are R packages: rmr, rhdfs and rhbase. The rmr package offers Hadoop Map Reduce functionalities in R. The rhdfs package offers basic connectivity to the Hadoop Distributed File System. It comes with convenient functions to browse, read, write, and modify files stored in HDFS. The rhbase package offers basic connectivity to HBase. It comes with convenient functions to browse, read, write, and modify tables stored in HBASE. Such tools enable important quality of care metrics to be developed across hundreds of dimensions. In this work the focus is on demonstrating how our implementation using current big data tools leveraged this ability to manage millions of events efficiently to develop accurate predictive models for detecting health insurance fraud. In proposed system we build a predictive model that can identify the erroneous or suspicious insurance claims.

a) First the tables are loaded into Hadoop file system.
b) All the selected predictor variables are retrieved from different tables and combined into a single coherent data set which is used for modeling using Hive queries.
c) By selecting and applying various classification prediction techniques using RHadoop and R Studio the actual model building task is done.

In our proposed system we use classification techniques such as decision tree, Naive Bayes classifier and clustering because they are some of the most popular and effective techniques .

2. Data: Sources, Characteristics and Preprocessing

Sources of data:

This phase involves exploring the raw data in order to gain initial insights and discover interesting actionable patterns. Getting familiar with the data is extremely important because knowledge of the data is useful in data preprocessing, a prerequisite for building a predictive model. Real world data that collected from different hospital and insurance company data is noisy and heterogeneous in nature and in order to get the data ready for modeling it is necessary to closely examine the attributes and their values. As the data comes from different sources are of different format such as

ORACLE	XML	MY SQL	CSV	MR	TEXT	ODS	DAT .....

need to be handled. Gaining insights into the data, such as the type of attributes present, the kind of data values for each attribute and their distribution, help with subsequent analysis. The data set is prepared from the hospital database which consists of detail description about the patient's medical history, diagnosis history, addiction history, length of stay, illness of patient and many factors and the insurance company data consist of insurance claim details such as insurerid , age, disease type, diagnosis, hospitalization details etc . This phase involves creating of data set from the data that is collected from different sources that is necessary for predictive modelling.

Fig.1 Example of one hospital patient database

patientid	gender	age	admission_date	discharge_date	.....

Diseae type	Diagnosis	physician_id	surgeries	anestesia

Fig. 2 Example of one insurance company claim database

Insurance id	Insurer name	Health id card	Disease type	gender	age	Name of the hospital

Date of admission	Date of discharge	diagnosis	Name of treating doctor	.....

Data Preprocessing

Once we have a good understanding of the data we need to prepare it for modeling. Data preprocessing is a precursory step to the actual modeling and helps in constructing the final homogeneous data set suitable for training predictive models. This phase poses several challenges due to the presence of heterogeneity in the real time health care data and prevalence of missing values and inconsistencies. The missing values in a dataset of hospital and insurance company databases will lead to an inaccurate result of an analytics of predictive modeling.

Designing of data for domain specific or Attribute Selection:

This step aims to define new features out of the original attributes, to maximize the discrimination power of the statistical method in separating fraudulent and legitimate cases. Health insurance is a complex phenomenon governed by multiple variables because there is no universal factor that can be used to predict the fraud. One of our major challenges is to determine the factors that have a significant impact on detection of fraudulent claim present in the data set. The number of these features needs to be further reduced, i.e., only those with certain discrimination power will be kept for statistical analysis. For example the attributes that selected for fraud detection in insurance data sets are

Insurance id	healthcard no.	Disease type	region	non_capture_of_procedure	diagnosis_procedure_coding_error

Delay in submission of document	Lack of pre authorization or precertification	Lenth of stay	no. of surgery	.....	Fraud or not

Figure 3: Insurance company insurance claim dataset

And then in order to apply factor analysis have to make the data to numeric value and to apply decision tree we have to make the class column (fraud or not) to binary. Such way to apply various analytical algorithms we need to make the data to domain specific according the requirement.

3. Analysis Methods for Detecting Fraud in Health Insurance

Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in four observed variables mainly reflect the variations in two unobserved variables. For finding factors in insurance claim database which will contribute more for the decision are extracted by using factor analysis. For factor analysis first need to convert the data into numerical form. And then by using the correlation matrix find the Eigen values and cumulative variance which will intern help to find how many variables has more impact. Then by using factor loading matrix we find the error values and then adding one more factor to minimize the error. This process goes on till factors will suffi-

cient to describe the class.

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
SS loadings	2.411	1.920	1.603	1.501	1.074	0.986
Proportion	0.219	0.175	0.146	0.136	0.098	0.090
Cumulative	0.219	0.394	0.539	0.676	0.774	0.863

Test of the hypothesis that 6 factors are sufficient.
The chi square statistic is 5.11 on 4 degrees of freedom.
The p-value is 0.276

In paper first we have applied factor analytics in the final dataset of patient and insurance claim. And the factors or variable with higher impact were found out. Then further analytical methods are applied on those data.

Decision tree

Decision tree is the method to find the target value and check the possibility of the trends with the different branches. In the decision tree all are instances are represented as the attribute values and it automatically perform the reduction of the complexity and selection of the features and regarding the predictive analysis its structure is vary understandable and interpretable.
This generates a tree from the given training samples. Each of the interior nodes of the tree is labeled by an attribute, while branches that load from the node are labeled by the values of the attribute. The tree construction process is heuristically guided by choosing the most informative attribute at each step, aimed at minimizing the expected number of tests needed for classification. Let E be the current set of training examples, and C1.... Cn the decision classes. A decision tree is constructed by repeatedly calling a tree construction algorithm in each generated node of the tree. Tree construction stops when all examples in a node are of the same class. Leaf node is labeled by a value of the class variable. Otherwise the most informative attribute say Ai is selected as the root of the sub tree, and the current training set E is split into subsets Ei according to the values of the most informative attribute. Recursively, a sub tree Ti is built for each Ei. Ideally, each leaf is labeled by exactly one class name. This is an effective method for decision making. because of below features

- Good interpret-ability of results
- Ability to generate rules from tree
- Ability to handle missing values

In paper we have applied decision tree in insurance claim database and hospital patient database to find out the rules which lead to a decision. A decision tree was built for each group and then converted into a set of rules. An example in the patient db the rule is: "if Age is between 18 and 25 and disease type= diabetes, and no. of procedure covered is >3 and length of stay is >3then it found that as an abnormal case." And same as in insurance claim dataset if insurer is from region 'A 'and have disease heart-surgery and have diagnosis procedure coding error =yes and delay in submitting the document =yes and so on then it's a possibility that his claim going to be rejected. This lead to a further investigation. Finally, each rule was evaluated by establishing a mapping from the rule to a measure of its significance using simple summary statistics made in that cluster and the average size of the claims; then would be signaled for further investigation. By relating those type of case from the decision tree from the insurance claim dataset with the patient details of the same insurer, we can find out the claim is fraudulent or not.

4. Performance Evaluation of Supervised Method

Confusion matrix

This section summarizes and compares different ways of evaluating the performance of binary classifiers, in which cases are labeled as either fraudulent or legitimate. In the performance evaluation of a binary classifier, an important initial step is to construct a confusion matrix based on the testing dataset.

		Actual classes	
		Fraudulent (+)	Legitimate (-)
Predicted classes by classifier	Fraudulent	True positive (TP)	False positive (FP)
	Legitimate	False negative (FN)	True negative (TN)

Confusion matrix

The cell labeled with "TP" records the number of actual fraudulent cases that are correctly predicted by the classifier; other cells, FP, FN, and TN, are defined in similar ways. Among error-based methods, an ROC curve is commonly used which plots the true positive against the false positive rates at different decision making thresholds of a classifier. The curve can be used to select a decision threshold that leads to a satisfactory sensitivity (i.e., true positive rate) and specificity (i.e., 1–false positive rate). In addition, the area under the curve, called AUC (Area under Curve), indicates the discriminating power of the classifier. Specially, if a classifier performs perfectly, AUC=1; if it performs randomly, AUC=0.5; the closer the AUC to 1, the higher the discriminating power of the classifier.

5. Visualization

Logistic regression

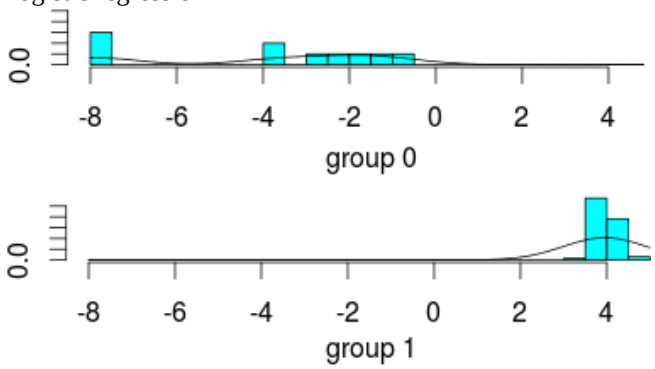

Figure indicates group1 as fraudulent claims and group 0 as non fraudulent claims decision tree

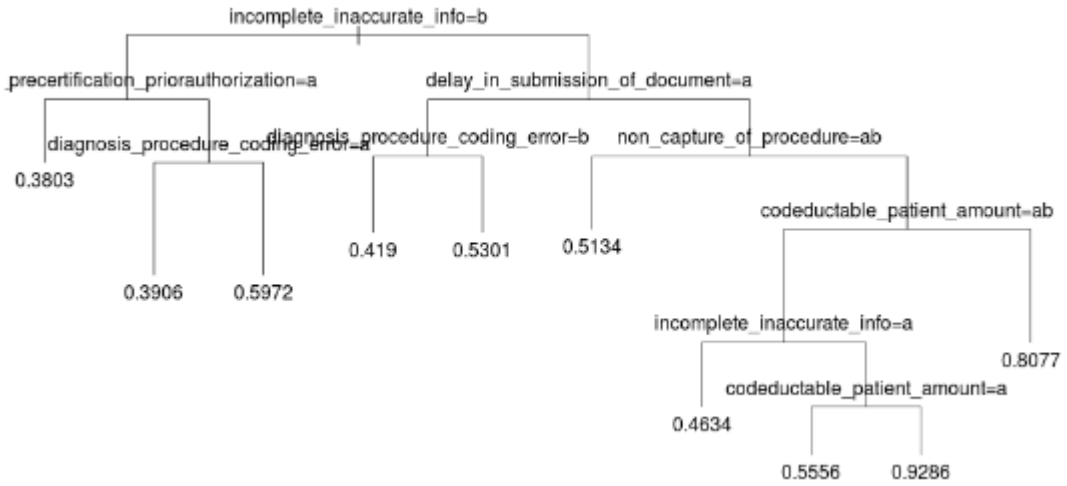

Figure shows a decision tree for insurance claim acceptance and rejection decision dataset.

6. Conclusion

In conclusion, this paper introduces some preliminary knowledge of health care system and its fraudulent behaviors, analyzes the characteristics of health care data. For future research, several directions have been pointed out and different other methods can be experimented to improve analytics and accuracy of detection of fraud However, to identify and eliminated the causes of fraud is the ultimate goal, so that fraud can be prevented in the future. Second, because both fraudulent and legitimate patterns in health care data may change over time, health care fraud detection method have to be dynamic enough to adapt these changes. Hence, future researches can attempt to develop self-evolving fraud detection methods.Applying Big Data analysis methods can lead to rapid detection of abnormal claims,and then creates a new set of tests to narrow the segment potentially fraudulent applications or to detect new patterns of fraud, previously unknown. An analysis of Big Data technology demonstrates its huge potential, but it shows that native tools for data analysis are still immature. The analysis methods applied in the field of health insurance were briefly described, each of them being effective for a particular type of fraud or a particular stage of the fraud detection process. All this leads to the conclusion that the best solution for detecting fraud in the health insurance system is, at present, a decision tree and naive Bayesian classification, both in terms of technologies and in terms of models of analysis.

References

[1] "Health Expenditure Australia 2011–12," Australian Inst. Health and Welfare, 25 Sept. 2014; www.aihw.gov.au/publication-detail/?id=60129544658.
[2] Private Health Insurance Australia: Quarterly Statistics, Australian Government, June 2013; http://phiac.gov.au/wp-content/uploads/2013/08/Qtr-Stats-Jun13.pdf.
[3] S. Barrette, "Insurance Fraud and Abuse: A Very Serious Problem," Quackwatch, 10 Jan. 2013www.quackwatch.org/02ConsumerProtection/insfraud.html.
[4] H. Chen, R.H.L. Chiang, and V.C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," MIS Quarterly, vol. 36, no. 4, 2012;http://ai.arizona.edu/mis510/other/MISQ%20BI%20Special%20Issue%20Introduction%20Chen-ChiangStorey%20December%2013.pdf.
[5] "Hospital Casemix Protocol (HCP)," Australian Government—Department of Health, May2012;www.health.gov.au/internet/main/publishing.nsf/Content/health-casemix-data-collections-about-HCP.
[6] "Round 12 (2007-08) Cost Report—Public Version 5.1, Private Version 5.1 and Private Day Hospital Facilities (Standalone) Version 5.1," Australian Government—Department of Health, Dec. 2010;
[7] Loshin, D., "Business Data Suited to Big Data Analytics", October 18, 2010,
[8] Halevi, G., & Moed, H., "The evolution of big data as a research and scientific topic: overview of the literature. Research Trends", Special Issue on Big Data, 30, 3-6, 2010.
[9] Hüsemann, S., Schäfer, M., "Building Flexible eHealth Processes using Business Rules", ECEH, volume 91 of LNI, page 25-36. GI, 2006
[10] Fawcett, T., "AI Approaches to Fraud Detection and Risk Management", Papers from the 2006AAAIWorkshop, Technical Report WS-97- 07. AAAI Press;
[11] Gill, K. M., Woolley, K. A., & Gill, M., "Insurance fraud: The business as a victim", in M. Gill (Ed.), Crime at work, Vol 1. (pp. 73-82), Leicester: Perpetuity Press, 2000