Demystifying Predictive Analytics with Data Mining to Optimize Fraud Detection in the Insurance Industry

Betelhem Zewdu[1](✉) and Gebeyehu Belay[2]

[1] School of Computing and Informatics, Department of Information Technology, Wachemo University, Hosaena, Ethiopia

[2] Institute of Technology, Computing Faculty, Bahir Dar University, Poly Main Campus, Agri Building 71 First Floor, Bahir Dar, Ethiopia

Abstract. The insurance industry is a company that renders risk management in the form of finance, humans, etc. ensuring contracts. Fraud is one risk, which does for self benefits or interest. In workmen's compensation, insurance fraud is intentional deception for gaining some interest in the form of health expenditures, which is challenging to handle manually. In this study, we proposed and introduced a novel approach to demystifying a predictive analytics approach using data mining techniques. The model can detect and predict fraud suspicious insurance claims with a particular emphasis on Insurance Corporation in the case of Workmen's Compensation. We use ensemble clustering followed by classification techniques for developing the predictive model. The predictive analytics applied to build an analytical model of the known variables' value to build a model that can predict the value of the variable of the unknown value. K-Means clustering algorithm is employed to find the natural grouping of the different insurance claims as fraud and non-fraud. The resulting cluster is employed to develop the classification model. The classification performed using the J48 and JRip algorithm to create the model of classifying fraud suspicious insurance claims using the AdaBoost method JRip as a base classifier, and it scored an accuracy of 98.26% on an 80% split CLAIM_REPORT_LENGTH_DATE is the determinant factor for predict fraud suspicious.

Keywords: Fraud · Detection · Data mining · Optimization · Predictive analytics · Determinant factor

1 Introduction

The insurance industry is a growing industry, which plays a vital role in ensuring the economic aspects of a county. In the encyclopedia, insurance is a sector of protection or risk management from losses. Risk management to hedge against the problem of accidental or uncertain circumstances. Fraud protection is a problem and a concern for many organizations. Fraud in workmen's compensation insurance is an intentional

M. A. Delele et al. (Eds.): ICAST 2020, LNICST 384, pp. 432–442, 2021.
https://doi.org/10.1007/978-3-030-80621-7_31

deception to gain benefits in the form of health expenditures. A false representation of facts in words or conduct, misleading claims, or concealment of what should have been disclosed that cheats and is intended to deceive. It is an illegal action of the individual will act upon someone else or institutes to theft or injury [1]. For example, a dishonest person may be called a fraud. Therefore, the data towards such practices are challenging and huge. The data is sensitive and also sophisticated, which is growing in a multi-dimension.

There are two most common types of claimant workmen's compensation fraud occurring in the insurance company [2]. (i) Abusers fraud: employees might not injure at work, or that can happen elsewhere and claim the injury occurred at work. Some of these may not document fraudulent workers' compensation claims but also file fraudulent automobile or general liability claims. (ii) Opportunists fraud: these employees are injured at work but then take advantage of more benefits than they should. They work to extend their interest beyond a reasonable period; exaggerate their symptoms; are uncooperative with treatment plans, and use delaying techniques when medical providers or employers take positive actions for the claims management process. Claims by opportunists can also become very costly, again deflecting payments from the parties who deserve them. The most common form of workmen's compensation insurance fraud is the exaggeration of claims; this refers to as Opportunistic Fraud.

Insurance companies are collected a large amount of data in their day to day activities. As the company's services and works become more advanced in technology, insurance companies store a large amount of data from their customers every day in a claim process. A very large amount of data is generated from a different database like underwriting and claim departments. These data are invaluable information about their customers' behavior, service, and preferences. The huge amount of data stored by a company it must be further manipulated to get useful information. However, to extract valuable information from this enormous data it takes time and effort for companies in hand. Knowledge extracted from the data can be used to develop new products and services to meet customers' needs. The need to provide more effective and customer-centric service becomes necessary to succeed in the business. Due to their protective regulations, extracting information from the database caused a lot of time. In this situation, the potential applicability of data mining is, therefore, to extract useful knowledge, pattern, and prediction ability from this data.

The fraud problem is the fact of a fraudulent claim, which is a concern of financial burden on insurers and result in the overall insurance costs. The estimated cost of property and casualty fraud each year is billion dollars losses [3]. According to IBM Corporation, insurance companies lose millions of dollars each year through fraudulent claims because they do not have a way to detect which claims are legitimate and which may be fraudulent.

Furthermore, fraud is a major problem for the insurance industry. Although the true cost of fraud for the industry, and subsequently for insurance policyholders who bear this cost through higher premiums cannot be known. The FBI estimates the total cost of insurance fraud is estimated to be more than $40 billion per year [4]. Another research indicates the annual insurance fraud cost for the property and casualty insurance industry is over 25 billion dollars. From this Workers' compensation insurance alone accounts for a sizable portion of this total cost [5]. To explore such complex data, an advanced

analytic tool is pertinent, which is proposed in this study, predictive analytics with data mining techniques and algorithms.

2 Related Works

There are also researches conducted to investigate the application of data mining in the insurance industry. As [6] has tried to apply Data Mining Techniques in Health Fraud Detection, and [7] worked on Detection of Automobile Insurance Fraud [5]. The conducted research is also on Predicting Workers' Compensation insurance fraud use SAS Enterprise Miner 5.1 and SAS Text Miner. The author focused on building predictive models to score an open claim for a propensity to be fraudulent [8]. It also worked Mining Insurance Data for Fraud Detection in the area of motor insurance at Africa Insurance Share Company. The author tried to do the applicability of the data mining technique in developing models that can detect and predict fraud suspicious insurance claims. The same work also has been done [9] An Integration of Prediction Model with Knowledge Base System for Motor Insurance Fraud Detection in the case of Awash Insurance Company S.C. The author tried to integrate the predictive model with a knowledge base system to detect insurance fraud.

Moreover, [10] studied the identification algorithm and model construction of Automobile Insurance Fraud by using data mining technology, which depends on an outlier's data [11]. Also studied in fraud detection in health insurance using data mining techniques. And this paper explores predictive analytics in insurance fraud detection using data mining technique. In addition to this, [12] also researched Data Mining Techniques in Fraud Detection. He focused on present some classification and prediction data mining techniques which is important to handle fraud detection [13]. Works on Modeling Insurance Fraud Detection Using Ensemble Combining Classification. He tried to solve the imbalance dataset problem by applying a proposed novel partitioning-under sampling technique using base-classifiers for insurance fraud detection.

3 Methodology

In this research work, we use the rule to identify red flag variables to scores the claims, which indicates fraud or not. In addition to this, the data analytics approach employs to discover hidden patterns in the data, which support fraud prediction. The predictive model is the one in data analytics, and it needs categorized data about fraud and non-fraud suspicious from the historical data. However, the data, we get from the insurance is not organized. So, to conduct this experimental research first, it needs to develop clustering techniques followed by classification. Unsupervised Clustering technique classifies the data into clusters having similar characteristics using the right set of variables that are indicative of fraudulent claims that should be able to cluster the fraudulent claims and non-fraudulent claims in different clusters. K-mean clustering is one of the algorithms we use for this research work. And the second technique develops a classification model, which helps to predict insurance fraud suspicious claims. For this purpose, the J48 decision tree and JRip rule-based algorithm are applied, as it is shown in Fig. 1.

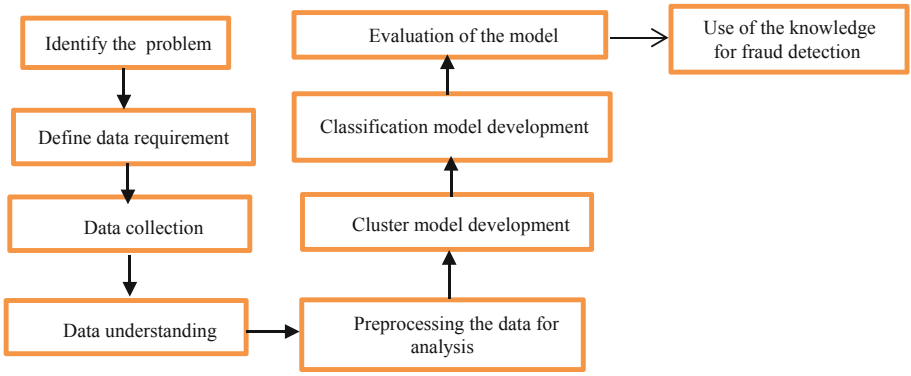

Fig. 1. Model building flowchart

For this thesis work, secondary data are used from the database of the EIC main branch in Addis Ababa. There are different insurances offered by the company, from this, the study is considered in workmen's compensation daily claim registration/process form to be filled by the experts of the insurance company. We use WEKA data mining tool for analysis purpose because it is java based open-source data mining tool which has a collection of data mining algorithms such as lazy, rules-based, decision trees, and so on. The original dataset which gets from the INSIS database is 17296 records, 27 attributes and the file size is a 6.56MB from the year 2011–2015 G.C. The whole dataset, which is gotten from the insurance database, is not important to the data mining task. We took a long time to preprocess the data i.e. the irrelevant or inappropriate data fields/records are removed first discuss with the domain experts in the insurance because they have not any influence on fraud detection. For this study, the object type column missed value filled by WEKA ReplaceMissingValues nearest records mean value (the most common class). CLAIM_REPORT_LENGTH_DATE (this refers to the length of the accident report date since its occurrence) is derived from the EVENT DATE and NOTIFICATION DATE columns of the data set. CLAIM_REPORT_LENGTH_DATE = EVENT_DATE-NOTIFICATION_DATE.

After discarding the irrelevant data, the total records account in the study is 1.92 MB with 17275 records and 10 attributes are prepared for further computational analysis listed in the following table, Table 1.

4 Experiment and Discussion

For the proposed predictive model, we use 17275 records and 10 attributes dataset for clustering and classification algorithms. K-Means clustering-based model use as an input for classification techniques using the J48 decision tree and JRip rules algorithm. Thre experiments using a clustering algorithm to change the default parameter of the K-Means algorithm for labeled the claim as fraud and non-fraud and select. The categories of features as their similarities to the segment of the claim within a cluster. For validating the clustering result, which is the initial model of the empirical analysis of the given instance, the researcher would use the number of iteration within a cluster sum of squared

Table 1. Final selected attribute for the study

No	Attribute name	Data type	Description	Remark
1	CLAIM_REPORT_LENGTH_DATE	Varchar2	The length of the accident report date since the occurrence of an accident	Derived
2	CLAIM_OFFICE	Varchar2	Branch service unit. The data type of this attribute is initially Number	
3	RISK_TYPE	Varchar2	The type of risk has occurred in the accident	
4	CLAIM_STATE	Varchar2	Refers to the status of the claim. The data type of this attribute is initially Number	
5	COVER_TYPE	Varchar2	The type of policy/coverage the insured buy from the insurance	
6	NOTIFICATION_DATE	Varchar2	The event/ accident that has been notified in the insurance. The data type of this attribute is initially date	
7	EVENT_PLACE	Varchar2	The place where the accident has occurred	
8	OBJECT_TYPE	Varchar2	Whether the insurance was bought as an individual or group. The data type of this attribute is initially Number	
9	PROFESSION	Varchar2	The type of job of the insured	
10	CLAIM_AMOUNT	Varchar2	The amount of recovered money	

error and domain expert judgment. For clustering algorithms, we set a threshold value to determine what patterns discover for each subsequent cluster model, which supports to identify and label the cluster based on given data. We set the value of k as 2 for identity fraud and non-fraud from the given data sets. In addition to this, the 10-folds cross-validation and percentage split test options are used for training and testing the classification model. These are the experimental setups that we follow to do the analysis.

4.1 Clustering-Based Model Development

Before experimenting, the threshold value set for numeric attributes, which use build, the clustering model. The threshold value for each attributes determined using Weka's minimum, maximum, and mean values displayed for the attribute shows in Table 2.

Table 2. Threshold values used for a cluster

CLAIM_REPORT_LENGTH_DATE		CLAIM_AMOUNT				
		DEATH	PTD	TTD	MHFO	Threshold
0–5	Fast	<10,000	<2000	≤1500	<500	Low
6–15	Moderate	≤20,000	≤3500	≤1750	<1000	Medium
>15	Slow	>20,000	>3500	>1750	>1000	High

In Table 2, the field annotated as PTD is permanent total disablement. TTD is temporary total disablement. MHFO is Medical, Pharmaceutical, Hospital, Funeral, and Other. The number of clusters chosen based on data sets and its clear visualization towards fraud detection. Among the experiments we did, we selected the last ones. For validating the clustering result that the intracluster similarity measure (within-cluster sum of squared error) is the number of iteration of the experiments for the converge as possible [14]. In addition to this, the domain expert's objectiveness evaluation, which considering the claim notification period, document submission formats, length of the claim report date, and other information. The following table shows the selected cluster experiment for the study in Table 3.

Table 3. Training of the 2nd experiment with seed value 10 to 100 and default distance function

Clustering result of the Second Experiment						
Distance function	K-value	Seed Value	Cluster Distribution		Number of Iteration	Within cluster sum of squared errors
			C1	C2	5	82869.74
Euclidean distance	2	100	4242 (25%)	13033 (75%)		

The second experiment presents for each segment of the cluster, as it shows in Table 4. The ranking is determined based on the fraudulence nature of the insurance claims.

Table 4. The 2nd experiment clustering result and rank of clusters

Cluster	Description	Rank
1	Slow claim reported date, Northern Addis branch, WC_MHFO risk type, Annulled claim state, WRKMEN_CMG cover type, Notification date Friday, event place Others, object type Group, High claim amount, and All Other Estates: Machinists and Drivers Profession	1 (this means this cluster is considered as fraud suspicious)
2	Fast claim reported date, South Addis branch, WC_MHFO risk type, paid claim state, WRKMEN_CMG cover type, Notification date Tuesday, event place Others, object type Group and Low claim amount and Daily Laborers Profession	2 (this means this cluster is considered as non-fraud)

From three experiments that were done in the clustering technique, the one is select with a smaller cluster sum square error and a minimum number of iteration depicts in the following figure, Fig. 2.

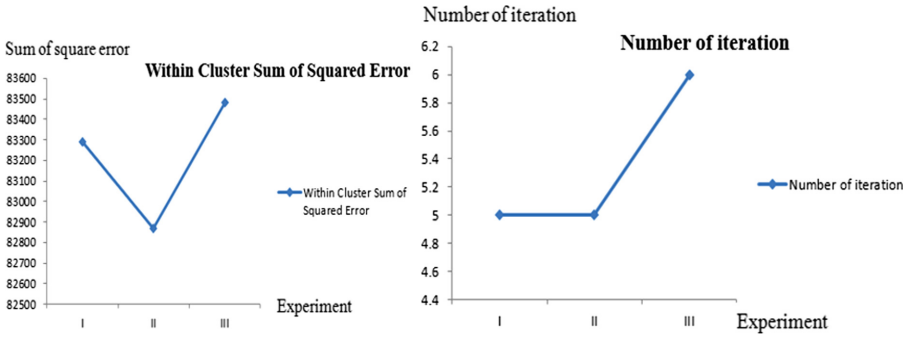

Fig. 2. Cluster model sum of square error clustering model number of iteration

4.2 Classification Based Model Development

After we developed a clustering model, we developed a predictive model using the classification techniques to define the fraud parameters. The predictive model developed using the J48 decision tree, JRip rule algorithm, and the ensemble method. From the

ensemble method, the Meta classifier of Adaboost and Bagging with J48 and JRip has as the base learners, whereas stacking to mix two algorithms J48 and JRip in the classifier of the J48 algorithm. The records are classified based on their values for the given cluster index, and the model is trained by employing the 10-fold cross-validation and the percentage split classification modes (Table 5).

Table 5. 80% split of ensemble method with JRip as base learner

Method	Algorithm	Correctly classified		Incorrectly classified		Time taken/s
		In %	Instance	In %	Instance	
AdaBoost	JRip	98.26	3395	1.73	60	355.43 s
Bagging		95.97	3316	4.02	139	987.45 s
Stacking	JRip & J48	96.23	3325	3.76	130	1181.87 s
80% training and 20% testing percentage split						

From the above table, in an 80% split, the Adaboost Meta classifier for the JRip algorithm scored an accuracy of 98.26%. From total testing datasets of 3455 records, 3395 records are correctly classified, while 60 records are incorrectly classified. On the other hand, bagging classifier scores the accuracy of 95.97% and 3316 records are classified correctly while 139 records are classified incorrectly from a total number of the testing set. Besides this, the stacking method classifier scored 96.23% accuracy and from the total testing set, 3325 records are correctly classified whereas 130 records are incorrectly classified from the total testing set.

4.3 Comparative Analysis

For this study, a number of the experiment was done but we present only the finally selected ones in the above table. However, these are some experiments done for the classification model in the single and ensemble method mention in Table 6 to select the final one.

Figure 3 shows the comparison of ensemble method classifiers in an 80% percentage split.

As the experiment result showed, the pattern that characterizes the length of the claim report date is slow, the claim state is annulled, the notification date is either Monday or Friday and the claim amount is high, a given claim is fraud suspicious. On the other hand, the length of the claim report date is fast, the claim state is paid, the notification date is neither Monday nor Friday, and the claim amount is low, a given claim is a non-fraud suspicious claim. Plus, the proposed work is demystifying the predictive analytics for workmen's insurance fraud suspicious claims by using an ensemble method and defining the determinant factors, which produces higher performance classification accuracy. Therefore, it is conceivable to conclude that the Adaboost ensemble method JRip algorithm as a base classifier is more appropriate to this particular case.

Table 6. Model comparison

Method	Algorithm	Accuracy of Models in %
Single	J48	96.19
	JRip	96.30
AdaBoost	J48	96.29
	JRip	97.74
Bagging	J48	96.59
	JRip	96.61
Stacking	JRip &J48	96.75
10 folds cross-validation		
Single	J48	95.83
	JRip	95.89
AdaBoost	J48	96.43
	JRip	**98.26**
Bagging	J48	96.09
	JRip	95.97
Stacking	JRip & J48	96.23
80% training and 20% testing percentage split		

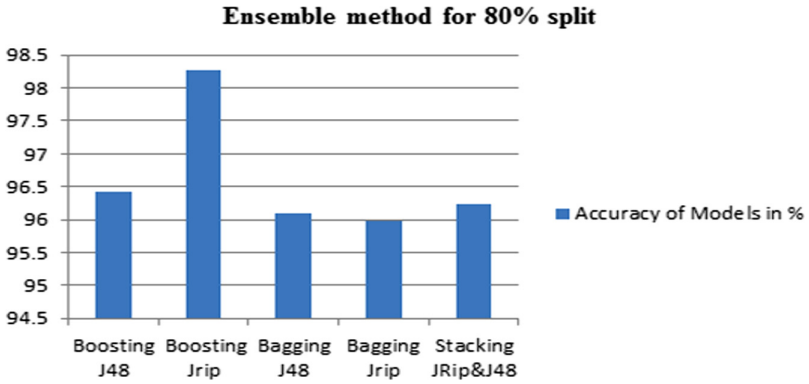

Fig. 3. Ensemble methods for comparison in 80% split

In addition to this, as we see a determinant factor in gain ratio the CLAIM_REPORT _LENGTH _DATE is the most determinant factor that causes predict fraud suspicious claim. Besides claim state and the claim amount is also the determinant factor. Here are some rules that gain from the JRip rule-based classifier algorithm.

Rule1: (CLAIM_REPORT_LENGTH_DATE = Slow) and (CLAIM_STATE = Annulled) and (NOTIFICATION_DATE = Friday) => Cluster = cluster0 (Fraud Suspicious)

If claim report length date = Slow AND claim state = Annulled AND notification date = Friday THEN the probability is to suspect fraud.

Rule2: (CLAIM_AMOUNT = High) and (CLAIM_REPORT_LENGTH_DATE = Slow) and (CLAIM_STATE = Annulled) => Cluster = cluster0 (Fraud Suspicious)

If claim amount = High AND claim report length date = Slow AND claim state = Annulled THEN predicts to suspect fraud.

Rule3: (CLAIM_STATE = Paid) and (OBJECT_TYPE = Individual) and (NOTIFICATION_DATE = Tuesday) => Cluster = cluster1 (Non- Fraud Suspicious)

If claim state = Paid AND object type = Individual AND notification date = Tuesday the probability is non-fraud.

5 Conclusion

In conclusion, insurance companies lose millions of dollars each year through fraudulent claims. And predictive analytics ways used to minimize fraud by detecting the cases. The data we used obtained from the insurance company database. It did not indicate a given claim is whether fraudulent or not, use a clustering technique to find out the natural grouping of data. The data set clustered and used as an input for the classification technique to predict fraud. We use the K-Means clustering algorithm for segmenting the data into the target classes of fraud and non-fraud. By changing the default parameters of the data, the experiments are done to generate a conceptual model that can create different cluster groups of insurance claims. Among the three models, the one with K = 2, Seed value = 100, and Euclidean distance function revealed better segmentation of the insurance claims.

The clustering outputs used for input for the classification model for model building using the J48 decision tree, JRip rule algorithm, and ensemble method. The model developed with the 10-fold cross-validation and 80% percentage split. From this model boosting ensemble method, particularly Adaboost Meta classifier JRip rule as a base classifier scored accuracy 98.26%, which better than the other and the final selected algorithm for this study. And, CLAIM_REPORT_LENGHT_DATE is the determinant factor that predicts fraud.

Acknowledgments. We would like to thanks the anonymous reviewers for their detailed review, valuable comments, and constructive suggestions. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. FBI Insurance Fraud. https://www.fbi.gov/stats-services/publications/insurance-fraud. Accessed 24 Aug 2020

2. Woodfield, T.J.: Predicting Workers' Compensation Insurance Fraud Using SAS® Enterprise Miner™ 5.1 and SAS® Text Miner (2005)
3. Yan, C., Li, Y.: The identification algorithm and model construction of automobile insurance fraud based on data mining. In: Fifth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC) (2015)
4. Rawte, V., Anuradha, G.: Fraud detection in health insurance using data mining techniques. In: International Conference on Communication, Information & Computing Technology (ICCICT) (2015)
5. Bhowmik, R.: Data mining techniques in fraud detection. J. Digit. Forensics Secur. Law **3** (2015)
6. Hassan, A.K.I., Abraham, A.: Modeling insurance fraud detection using ensemble combining classification. Int. J. Comput. Inf. Syst. Ind. Manag. Appl. **8**, 257–265 (2016)
7. Farlex Inc. Fraud (2013). http://legaldictionary.thefreedictionary.com/fraud. Accessed 3 Sept 2016
8. Jennings, G.: The three most common types of workers' comp fraud, and how to prevent them, 11 November 2014. http://www.propertycasualty360.com/2014/11/11/here-are-the-3-most-common-types-of-workers-comp-f. Accessed 31 Mar 2017
9. IBM Financial Crimes Insight for Claims Fraud (2020)
10. Pal, R., Pal, S.: Application of data mining techniques in health fraud detection. Int. J. Eng. Res. Gener. Sci. **3**(5), 129–137 (2015)
11. Subudhi, S., Panigrahi, S.: Detection of automobile insurance fraud using feature selection and data mining techniques. Int. J. Rough Sets Data Anal. **5**(3), 1–20 (2018). https://doi.org/10.4018/IJRSDA.2018070101
12. Tariku, A.: Mining insurance data for fraud detection: the case of Africa insurance share company. Master thesis Addis Abeba University, June 2011
13. Abdi, C.: An integration of prediction model with knowledge base system for motorinsurance fraud detection: the case of awash insurance company S.C. Master thesis Addis Abeba University, February 2016
14. Jain, S., Aalam, M.A., Doja, M.N.: K-means clustering using WEKA interface. In: Computing For Nation Development (2010)